THE UNIVERSITY OF QUEENSLAND

AUSTRALIA

SELF-NONSELF RECOGNITION

GENOMIC AND TRANSCRIPTOMIC INSIGHTS FROM THE SPONGE AGGREGATION FACTORS

LAURA FRANCES ELIZABETH GRICE
B.SC. (HONS)

A thesis submitted for the degree of Doctor of Philosophy at
The University of Queensland in 2015
School of Biological Sciences

Abstract

The multicellular condition cannot be maintained without safeguards protecting the integrity of the individual. Tissue contact and fusion with other conspecific individuals may threaten this integrity, as genetically non-identical cells may shirk their somatic duties and gain disproportionate access to the germ line. As sessile invertebrates that commonly inhabit crowded benthic environments, sponges are particularly reliant on a molecular self-nonself defense system in order to resist loss of habitat space, chimerism and possible germ line parasitism by neighbouring conspecific sponges. Sponge allorecognition appears to be, at least in part, under the control of extracellular proteoglycans called aggregation factors (AFs), which were first discovered based on their role in the species-specific reaggregation of dissociated sponge cells. Although the AFs have been extensively studied for over fifty years, the majority of this work has involved biochemical, rather than genetic approaches, and has focussed on the role of the glycan subunits associated with the AFs. In the present work, I investigate the genetic properties underlying the AF protein backbone, to better understand the functions and evolution of these putative allorecognition molecules.

Using newly-available genomic and transcriptomic data, I surveyed the phylum Porifera for novel putative AF sequences, to explore the evolutionary origins of this gene family. I conclude that the AFs are a demosponge and hexactinellid-specific innovation. I then performed an in-depth characterisation of the six AF genes from the model demosponge species, *Amphimedon queenslandica*. The six genes display a highly modular intron/exon organisation. However, as expected of putative allorecognition genes, the AFs are greatly diversified between individuals, with nucleotide polymorphism (and possible positive selection) and intron retention events distributed across the six genes. The AFs are very highly expressed across sponge development and in response to alloimmune challenge, and undergo a particular spike in gene expression levels after the onset of sponge metamorphosis. The AF genes also exhibit expression patterns across development that are significantly correlated with those of other, developmentally important genes with roles in various cell signalling pathways. I conclude that the AFs play a novel developmental role, in addition to their putative allorecognition capabilities.

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, 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 research higher degree 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.

Publications during candidature

Peer-reviewed paper

Grice LF, Degnan BM. 2015. The origin of the ADAR gene family and animal RNA editing. *BMC Evol Biol* **15**: 4.

Book chapter

Grice LF, Degnan BM. 2015. How to build an allorecognition system: A guide for prospective multicellular organisms. In *Evolutionary Transitions to Multicellular Life* (eds. I. Ruiz-Trillo and A.M. Nedelcu), Springer, Dordrecht Heidelberg New York London.

Conference abstracts

Grice LF, Gauthier MEA, Fernandez-Valverde S, Degnan BM. 2013. Genomic and trancriptomic profiling of the aggregation factor genes of the demosponge *Amphimedon queenslandica*. *Ninth World Sponge Conference.* Fremantle, WA, Australia, 4 – 8 Nov 2013.

Grice LF, Gauthier MEA, Degnan BM. 2012. Seeking diversity: Genomic characterisation of the aggregation factor genes of the demosponge *Amphimedon queenslandica*. *Twelfth Congress of the International Society of Developmental and Comparative Immunology (ISDCI)*. Fukuoka, Japan, 9 – 13 July 2012.

Publications included in this thesis

Grice LF, Degnan BM. 2015. How to build an allorecognition system: A guide for prospective multicellular organisms. In *Evolutionary Transitions to Multicellular Life* (eds. I. Ruiz-Trillo and A.M. Nedelcu), Springer, Dordrecht Heidelberg New York London.

This work was incorporated in the Abstract and Chapter 1

CONTRIBUTOR	STATEMENT OF CONTRIBUTION
Laura F. Grice (Candidate)	Conducted the literature review (100%) Wrote and edited the chapter (90%)
Bernard M. Degnan	Wrote and edited the chapter (10%)

Grice LF, Degnan BM. 2015. The origin of the ADAR gene family and animal RNA editing. *BMC Evol Biol* **15**: 4.

This work was included as Chapter 5

CONTRIBUTOR	STATEMENT OF CONTRIBUTION
Laura F. Grice (Candidate)	Designed study (60%) Conducted analyses (100%) Wrote and edited the paper (90%)
Bernard M. Degnan	Designed study (40%) Wrote and edited the paper (10%)

Contributions by others to the thesis

Bernard M. Degnan contributed to the conception and design of this research, advised on methods and analysis, and provided critical comments on the thesis and on the associated publications in Chapters 1 and 5.

Selene Fernandez Valverde advised on methods and analysis, and produced several of the datasets analysed across this thesis (these contributions are specifically acknowledged in the relevant chapters).

William Hatleberg produced several of the datasets analysed across this thesis (these contributions are specifically acknowledged in the relevant chapters).

Transcriptome sequencing was conducted by Macrogen Inc., South Korea.

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

None.

Acknowledgements

I have been a member of the Degnan labs in various capacities since 2008, so it is with great fondness (and just a tinge of sadness) that I write these acknowledgements at the end of this endeavour.

First, of course, to my remarkable supervisor Bernie Degnan. Your enthusiasm and insight has kept this project rolling, when the craziness that is the AF locus seemed impenetrable. Thank you for your clarity, for pretending not to notice when I felt I hadn't done enough, for the 'Degnan Gold' you can sprinkle on a piece of writing to give it life. To my co-supervisor and joint lab Head, Sandie Degnan. Sandie, your wisdom and integrity - regarding not only my project, but also science as a field and as a lifestyle - have been a source of inspiration. Thank you also to Bernie and Sandie for the opportunity to develop additional skills in the more administrative side of science, in the role of Lab Manager in 2013.

I also wish to thank my other co-supervisor, Andy Barnes, for his advice regarding this project. As the chair of my thesis examining committee, I am grateful to Sassan Asgari for his advice regarding my project at each milestone. Thanks are also due to Gail Walter, postgraduate administrator for the School of Biological Sciences, for her endless patience.

Bernie and Sandie have fostered an ongoing lab culture that is nothing but welcoming and helpful. To all the past and current Deglabbers who have overlapped with my period in the lab: your assistance, advice, and generosity with your time - not to mention friendship, competitive baking skills, milkshakes and Spice Girls puzzles - have made the lab a wonderful environment to both do and avoid science. Particular thanks must go to Andrew Calcino, Bryony Fahey, Melanie Havler and Carmel McDougall for their advice in the lab and the field, and to Melanie Havler and Kerry Roper for running a tight ship during their respective glorious reigns as Lab Manager.

On the computational side of things, thanks to Ben Woodcroft for providing bioinformatics assistance from afar in the early days of my PhD, and Timothy Lamberton for on-call statistical, mathematical and programming consulting. In 2012, I fell down the rabbit hole into 'black screen' (command line) data analysis. Most of this project would not have been possible without Selene Fernandez Valverde, whose help and patience went above and beyond - thank you Selene. Thanks also to Felipe Aguilera, Andrew

Calcino, Federico Gaiti and William Hatleberg for sharing the ride, your advice, and troubleshooting prowess with me. I would also like to acknowledge the support, particularly early on, of Ryan Taft and Michael Pheasant for the lab transcriptome projects.

I wish to thank the past and present staff of Heron island Research Station from the past five years; in particular Kyra Hay and Liz Perkins, who were the serving Scientific Officers for most of my trips to the Island, and also to Maureen Roberts for ensuring the smooth organisation of these field trips.

I am grateful to Xavier Fernàndez-Busquets for his insight into the AFs throughout this project, and for providing us with Clathria prolifera genetic material and sequences. Thanks also to Maja Adamska and Marcin Adamski for early access to the Sycon ciliatum genome.

I am indebted to Bernie Degnan, Selene Fernandez Valverde, Maely Gauthier, William Hatleberg, Simone Higgie and Tahsha Say, for taking the time to read and critique the various chapters of this manuscript, and to Timothy Lamberton for assistance with formatting the final document. Additional thanks to William Hatleberg for advice regarding figure composition and design.

This project was funded by a grant to Bernie Degnan, Sandie Degnan and Anthony De Tomaso by the Australian Research Council. I would also like to acknowledge the Australian Government for the receipt of an Australian Postgraduate Award scholarship, which made undertaking this PhD program possible. I wish to thank the School of Biological Sciences (University of Queensland) and the International Society of Developmental and Comparative Immunology (ISDCI) for generously providing funding which allowed me to attend the 12th ISDCI congress in Fukuoka, Japan. I am also grateful to Bernie Degnan for financially supporting this trip, another to the 9th World Sponge Conference in Fremantle, Western Australia, and several field trips to Heron Island Research Station. The opportunities to share my research with, and meet, members of the wider scientific community (and to visit three very different, but remarkable, parts of the world!) were invaluable experiences.

To my entire family, thank you for your love and support for my nine-plus years of uni student-hood (and beyond, of course). Mum and Dad, you have (a) fostered or (b) foisted an interest in science

in me from an early age. I don't know whether to thank nature or nurture (perhaps, as the resident geneticist, that is my job to find out), but how can you not go from pouring iodine on slices of bread to sticking pins in slices of sponge? And finally to Tim, who put up with me when I wasn't there, and when I was - ~~you really know how to dance~~ you're the best.

Keywords

aggregation factor, allorecognition, domain architecture, evolution, genomics, innate immunity, invertebrates, polymorphism, self-nonself recognition, sponge

Australian and New Zealand Standard Research Classifications (ANZSRC)

060405 Gene Expression (20%)

060406 Genetic Immunology (40%)

060409 Molecular Evolution (40%)

Fields of Research (FoR) Classification

0604 Genetics (70%)

0603 Evolutionary Biology (30%)

TABLE OF CONTENTS

List of Figures

List of Tables

List of Appendices

** Appendices marked with an asterisk are available online; details provided in the Appendices section.*

List of abbreviations

Abbreviation	Definition
aa	Amino acid
AD	Adenosine deaminase domain
ADAR	Adenosine deaminase acting on RNA
ADAT	Adenosine deaminase acting on tRNA
AF	Aggregation factor
AFA	Aggregation factor A
AFB	Aggregation factor B
AFC	Aggregation factor C
AFD	Aggregation factor D
AFE	Aggregation factor E
AFF	Aggregation factor F
AGRF	Australian Genome Research Facility
ANOVA	Analysis of variance
AqAFs	Amphimedon queenslandica AFs
AR	Aggregation receptor
ARC	Allorecognition complex
BCP	Bromochloropropane
BLAST	Basic local alignment search tool
BLIND	Basic linear index determination of transcriptomes
bp	Base pair
BSA	Bovine serum albumin
CCA	Crustose coralline algae
cDNA	Complementary DNA
CEL-Seq	Cell Expression by Linear amplification and Sequencing
CNRQ	Calibrated normalised relative quantities
CpAFs	Clathria prolifera AFs

ABBREVIATION	DEFINITION
cpm	Counts per million reads
DAG	Diacylglycerol
DGE	Differential gene expression
DNA	Deoxyribonucleic acid
dNTP	Deoxyribonucleotide triphosphate
dsRB	Double-stranded RNA binding domain
dsRNA	Double-stranded RNA
DTT	Dithiothreitol
e	Expect
ECH	Enoyl CoA hydratase
FDR	False discovery rate
FrzB	Frizzled
GAP	GTPase activating protein
GAPD	Glyceraldehyde-3 phosphate dehydrogenase
GB	Gigabyte
gDNA	Genomic DNA
GLM	Generalised linear modelling
GO	Gene ontology
HA	Hyaluronan
HAase	Hyaluronidase
HMM	Hidden Markov model
hpe	Hours post emergence
hpf	Hours post fusion
hpg	Hours post grafting
hpr	Hours post resettlement
HPRT	Hypoxanthine phosphoribosyltransferase
hps	Hours post settlement

ABBREVIATION	DEFINITION
HSD	Honest significant difference
IgSF	Immunoglobulin superfamily
ILF2	Interleukin enhancer binding factor 2
IP3	Inositol triphosphate
kb	kilobase pairs
LPS	Lipopolysaccharide
MIDAS	Metal ion-dependent adhesion site
mRNA	Messenger RNA
n	Number
NF-kB	Nuclear factor kappaB
NMD	Nonsense mediated decay
no-RT	Reverse transcriptase-free
nt	Nucleotide
ORF	Open reading frame
p	Probability
PAF	Observed frequency of each intron phase in the AF gene dataset
PASA	Program to Assemble Spliced Alignments
PCA	Principal component analysis
PCalx	Observed frequency of each intron phase in the Calx-beta domain-containing gene set
PCP	Planar cell polarity
PCR	Polymerase chain reaction
PGRP	Peptidoglycan recognition protein
PLAN	Personal BLAST Navigator [software]
Pobs	Observed frequency of each intron phase
Prand	Random frequency of each intron phase
P_{ref}	Observed frequency of each intron phase in the Reference Set
PSI-BLAST	Position-specific iterative basic local alignment search tool

Abbreviation	Definition
PTC	Premature termination codon
qPCR	Real-time quantitative PCR
RACE	Rapid amplification of cDNA ends
RFLP	Restriction fragment length polymorphism
RHAMM	Hyaluronan-mediated motility receptor
RNA	Ribonucleic acid
RNA-Seq	RNA sequencing
RPKM	Reads per kilobase of transcript per million mapped reads
rpm	Revolutions per minute
RT	Reverse transcriptase
SDHA	Succinate dehydrogenase complex subunit A
SLIP	Sponge LPS-interacting protein
SNP	Single nucleotide polymorphism
SP	Signal peptide
SRCR	Scavenger receptor cysteine-rich
ssIII	SuperScript III
TAE	Tris-Acetate-EDTA
TBE	Tris-Borate-Acetate
VERL	Vitelline envelope receptor for lysin
VWA	Von Willebrand type A domain
VWD	Von Willebrand type D domain
wpf	Weeks post fusion
wpm	Weeks post metamorphosis
YWHAZ1	Tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein, zeta polypeptide
ZB	Z-DNA/RNA binding domain

Chapter 1 - General Introduction

1.1 Commonalities and predictors of allorecognition molecules

1.1.1 The importance of allorecognition for the multicellular condition

Transition to the multicellular state is a key step in the evolution of organismal complexity and has occurred independently multiple times across life on Earth (Buss 1987; Bonner 1988; 2000; King 2004; Grosberg and Strathmann 2007). One potential benefit of transition to a multicellular state is the new capacity for the division of labour, whereby different cells within an organism become responsible for producing and sharing different key gene products or performing useful functions (Kirk 2005; Gavrilets 2010; Rossetti et al. 2010; Goldsby et al. 2012; Ispolatov et al. 2012; Ratcliff et al. 2012). The division of labour allows an organism to increase metabolic efficiency by dividing different cellular tasks between specialised cell types (Goldsby et al. 2012), and by partitioning incompatible cellular processes such as motility and cell division (Buss 1987), or nitrogen fixation and photosynthesis (Fay 1992).

Successful multicellularity, particularly in organisms with multiple cell types, requires cooperation between and amongst different cells and cell types, with each cell performing its required role and receiving support in return (Buss 1987). This cooperation requires individual cells to sacrifice their own autonomy to benefit the fitness of the higher-order organismal unit. A clear example of this requirement can be seen in organisms with distinct somatic and germ cell groups, with somatic cells relinquishing the capacity to contribute their genetic material to subsequent generations (Michod 2007). Mechanisms are therefore required to ensure these cells do not abandon their somatic duties in favour of a more individually-advantageous path, for example by unchecked cell replication or neglect of key cellular roles. Such behaviour is termed cheating, that is, exploitative behaviour that benefits an individual unit (in this case, a cell) at the expense of other members of a usually cooperative group (Strassmann and Queller 2011).

Cell cheating typically takes one of two forms, depending on the source of the cheater – either internal or external cheating. Internal cheaters arise when mutations cause cells to exploit otherwise-genetically identical cells within the multicellular body, as occurs in cancers. Multiple mechanisms exist to aid the control of internal cheating. For example, apoptosis, DNA repair and the arrest of cell division can minimise the expression of somatic mutations (Kastan and Bartek 2004), while sequestration of the germ line and a unicellular bottleneck stage of development both limit the potential for transmission of deleterious cheater mutations to the next generation (Grosberg and Strathmann 2007). External cheating occurs when other individuals threaten organismal integrity, for example by tissue or organismal fusion. This is potentially problematic, because the altruism of somatic cellular cooperation and sacrifice of germ line contribution can only be maintained if genetically identical (or at least, closely related) cells are able to contribute genetic material to the next generation (Eberhard 1975). Unrelated cells have no 'motivation' to contribute fairly, and can thus exploit resources provided by the somatic cells, potentially using these resources to increase their own reproductive output at the expense of the host.

Control of external cheating has been well documented in the colonial ascidian *Botryllus schlosseri*. In this species, colonies sharing one or more alleles for the highly polymorphic locus *FuHC* are considered self and will undergo vasculature fusion, while those with disparate *FuHC* alleles reject each other (Oka and Watanabe 1957). As large numbers of *FuHC* alleles are present in *B. schlosseri* populations, fusion is effectively limited to closely related colonies. However, fusion between histocompatible individuals has been observed at relatively high rates (Rinkevich et al. 1998); when this does occur, it tends to be followed by a process of resorption, whereby one fusion partner is partially or entirely eliminated, in a competitive and reproducible fashion (Rinkevich and Weissman 1987). Intriguingly, however, the resorptive winner can experience germ or somatic cell parasitism, which, in extreme cases, may lead to total replacement of winner cells with those from the resorptive loser (Stoner and Weissman 1996; Stoner et al. 1999). This parasitism occurs despite the presence of a complex self-nonself recognition system, which emphasises the importance of restricting fusion, and therefore potential germ line control, to self or close kin. Systems allowing the recognition of and discrimination between self and nonself allow successful multicellular organisms to limit wasted resources and potential loss of reproductive output.

The first half of this chapter focuses on the requirements and features of self-nonself recognition systems that allow the distinction between conspecific members of a single species.

1.1.2 The three-phase model of self-nonself recognition

All self-nonself recognition reactions occur as a three-phase process. The first phase of the process is *detection* – a particular individual unit (e.g. a cell type, organism, etc.) must detect the presence of another biological entity in its vicinity. Phase two is *recognition*, whereby the first unit must then determine the identity of the detected unit as self or nonself. Different systems may recognise the presence (or absence) of self, of nonself, or be able to directly recognise both self and nonself. The simplest, and thus probably most ancient, of these hypothetical systems is one based on self recognition, whereby cells or molecules lacking some label identifying them as self are rejected (Coombe and Ey 1984; Boehm 2006) (Figure 1.1). The final phase of the self-nonself recognition process is *discrimination*, where some action is taken on the basis of the recognition decision. The outcome of this action varies. For example, self could be favoured (or nonself disfavoured) as is the case in immune reactions, whereas nonself may be favoured (or self disfavoured) in mate selection processes. The mechanisms employed to execute this discrimination also vary, and may be passive or aggressive. For the purposes of this thesis, "self-nonself recognition" is taken to refer to the effect of the outcome (i.e. separation of self from nonself) rather than the mechanism (i.e. detection of nonself) of this recognition.

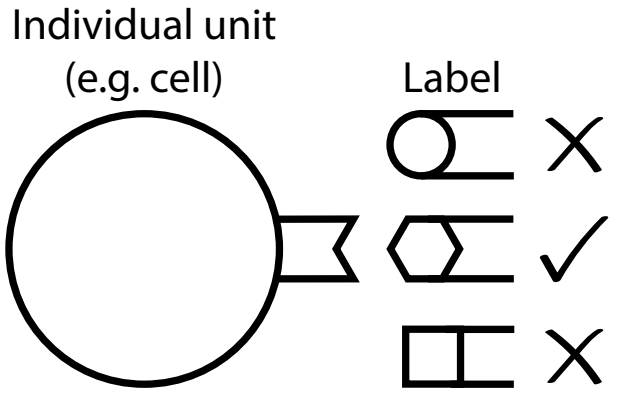


Figure 1.1 A general scheme of self recognition
An individual unit (here, a cell) assesses labels to which it is exposed. In self-only recognition, the cell can recognise only those labels that match its own self template, and all non-matching forms are therefore rejected.

The three phases of self-nonself recognition may not necessarily occur as distinct events. For example, detection and recognition could occur simultaneously in systems where recognition is possible only through the binding of particular homotypic or heterotypic recognition labels. Such binding could trigger activation of downstream pathways in a separate discrimination event, or directly cause,

for example, cellular aggregation, with passive discrimination a direct effect of this binding. Other combinations are also possible. Regardless of the precise mechanisms of action, however, all three phases should occur in some capacity in any self-nonself recognition reaction.

1.1.3 The functional requirements of self-nonself recognition systems predict their underlying molecular features

All allorecognition systems must possess one or more molecules capable of executing the three phases of self-nonself recognition reactions outlined above. Therefore, consideration of the functional requirements of allorecognition systems allows prediction of the expected features of their underlying molecules. These predictions are of practical value, for example acting as useful criteria when attempting to identify putative allorecognition molecules from a set of newly identified candidate genes (for example see Rosa et al. 2010). However, as few allorecognition systems have been thoroughly characterised, these criteria are not likely to apply to all systems.

a. Phase one: Detection

The first phase of self-nonself recognition reactions, detection, involves sensing the presence of other individuals in the nearby environment. This task must be performed by a molecule capable of mediating intercellular interactions, either via direct cellular contact or the binding of secreted molecules. This predicts the existence of an allorecognition molecule with an extracellular region capable of binding molecules attached to, or secreted by, neighbouring cells - although intracellular receptors are known in other signalling pathways (Geuze et al. 1984; Baumann et al. 1999; Meylan et al. 2006) and thus their presence here cannot be excluded. Indeed, the recent identification of a cytosolic gene in the *B. schlosseri FuHC* locus reveals that not all allorecognition factors are on the cell surface or secreted (Voskoboynik et al. 2013).

Proteins fulfilling this requirement are prevalent in the molecular suites of most well-characterised allorecognition systems. These are usually transmembrane or secreted proteins featuring large extracellular regions with tandemly repeated protein domains (Figure 1.2a, Table 1.1). For example, the allodeterminants alr1 (Rosa et al. 2010) and alr2 (Nicotra et al. 2009) from the cnidarian *Hydractinia symbiolongicarpus*, and mFuHC, whose encoding gene resides within the *Botryllus schlosseri* FuHC

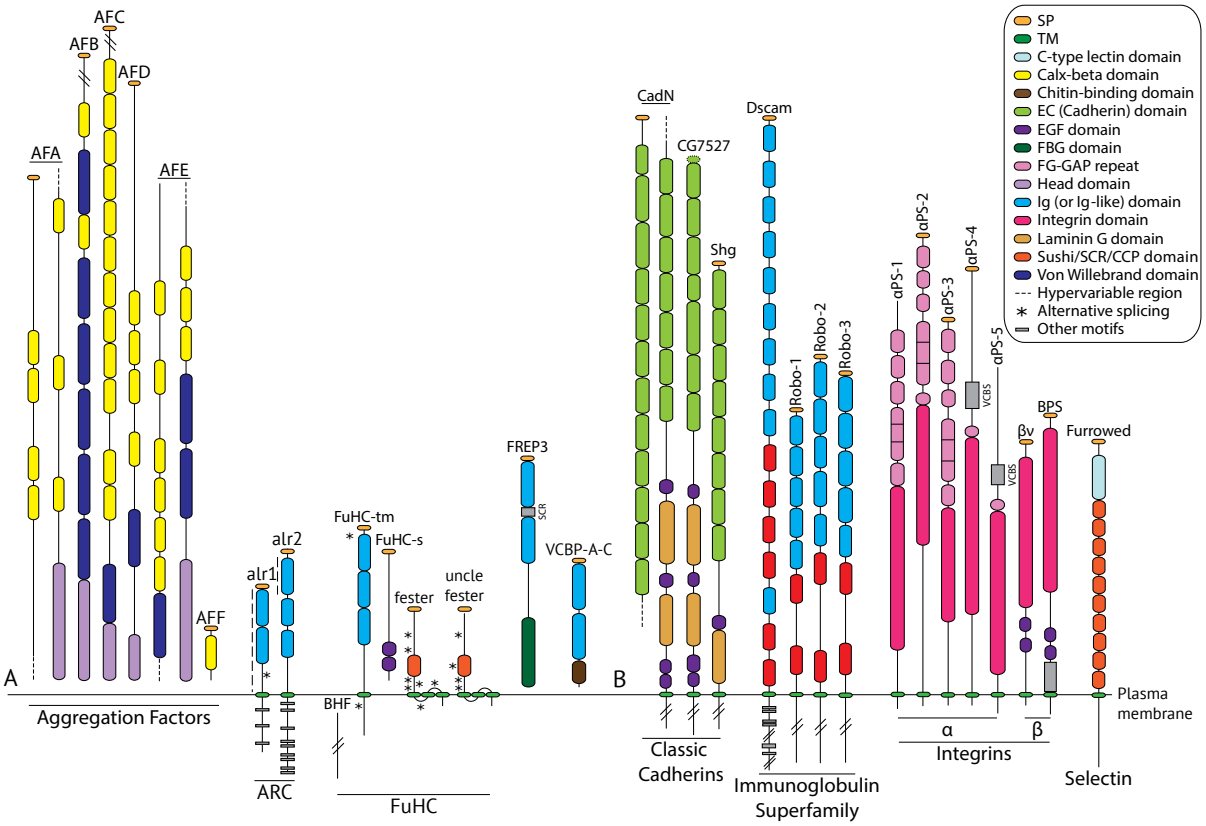


Figure 1.2 Invertebrate allorecognition, self-nonself recognition and cell adhesion proteins

The secondary protein structures of (A) selected invertebrate allorecognition and self-nonself recognition associated molecules and (B) *Drosophila melanogaster* cell adhesion molecules. (A) Featured molecules are the aggregation factors AFA – AFF from *Amphimedon queenslandica* (Gauthier 2009), alr1 and alr2 from the *Hydractinia symbiolongicarpus* allorecognition complex (ARC) (Nicotra et al. 2009; Rosa et al. 2010), *Botryllus schlosseri* FuHC locus proteins BHF, FuHCtm, FuHCs, fester and uncle fester (De Tomaso et al. 2005; McKitrick et al. 2011; Nyholm et al. 2006; Nydam et al. 2013b; Voskoboynik et al. 2013), FREP3 *Biomphalaria glabrata* parasite defense system (Zhang et al. 2001), and a single representative structure of VCBP forms A – C from the anti-pathogen system of the urochordate *Ciona intestinalis* (Dishaw et al. 2011). As FREP3 and the VCBPs are not involved in allorecognition processes, they are here categorised as self-nonself recognition molecules. (B) Members of the key cell adhesion protein families - classic cadherins (Hill et al. 2001), immunoglobulins (Kidd et al. 1998; Schmucker et al. 2000; Simpson et al. 2000), integrins (Narasimha and Brown 2006) and selectins (Leshko-Lindsay and Corces 1997) - from the representative invertebrate species *D. melanogaster* are shown. All identified members for this species of the classic cadherins, integrins and selectins are shown. As the *D. melanogaster* immunoglobulin superfamily is very large, only four members are shown here: the axon guidance receptor molecules Dscam and Robo 1-3. Blocks indicate protein domains and other key features; the linear structures of the proteins are shown. The line symbolises the plasma membrane, with the region above representing the extracellular space, and below representing the cytoplasm. All structures are drawn to scale except where indicated by crossed lines. As AFA, AFB and CadN are very large, these structures have been split in two as represented by dashed lines. SP – signal peptide, TM – transmembrane domain.

Table 1.1 Structural properties of key cell adhesion and self-nonself recognition domains

Domain	Pfam code	Secondary Structure	Additional structural features	Reference
Immunoglobulin	CL0011	β-sandwich	Disulphide bond joins β-strands	Bork et al. 1994 Harpaz & Chothia 1994
EGF	CL0001	Two β-sheets	Three disulphide bridges	Wouters et al. 2005
Calx-beta	PF03160	β-sheet	-	Schwarz & Benzer 1997
Sushi/SCR/CCP	PF00084	β-sandwich	Stabilised by disulphide bridges	Norman et al. 1991
Fibrinogen C	PF00147	α-helices, β-sheets	Two disulphide bridges	Middha & Wang 2008
FG-GAP repeat	PF01839	β-sheet	Seven repeats form β-propeller	Springer 1997
FNIII	PF00041	β-sandwich	-	Leahy et al. 1992
Laminin G	PF00054 PF02210 PF13385	β-sandwich	-	Hohenester et al. 1999
Cadherin	PF00028	β-sandwich	-	Shapiro et al. 1995
C-type lectin	PF00059	Loop-within-a-loop structure with two β-sheets and two α-helices	Two disulphide bridges	Zelensky & Gready 2005
Chitin-binding domain	PF01607	β-sandwich	Three disulphide bridges	Ikegami et al. 2000
Von Willebrand	CL0128	Twisted β-sheet flanked by α-helices	Two disulphide bridges	Edwards & Perkins 1995

locus (De Tomaso et al. 2005; Nydam et al. 2013b; Voskoboynik et al. 2013), are all equipped with multiple immunoglobulin-like domains, while the aggregation factor (AF) proteins from the sponges *Amphimedon queenslandica* and *Clathria* (formerly *Microciona) prolifera* are all predicted to possess numerous tandemly-repeated Calx-beta domains (Fernàndez-Busquets et al. 1996; Gauthier 2009). Such extracellular domains are commonly comprised of β-sheets and related folds such as the β-sandwich structure (Table 1.1). These folds are structurally robust to amino acid change (Wright et al. 2004), which may be of key importance for the maintenance of molecule functionality despite the high levels of intraspecific sequence diversity required of allorecognition molecules (discussed below).

b. Phase two: Recognition

The primary requirement of this phase is a capacity for high-precision recognition decisions, in order to prevent costly self or nonself rejection or acceptance, depending on the circumstance (Tsutsui 2004). Such precision requires an underlying highly polymorphic molecular system, in order to produce unique labels for each individual self unit (Hildemann 1979; Grosberg 1988; Tsutsui 2004). The presence in a population of such levels of polymorphism means that, for recognition reactions between conspecific individuals, there is a strong probability that tags matching an individual's self signature are true representatives of self, rather than random matches due to chance. Mechanistically, this occurs via sequence differences that potentially confer structural changes to allorecognition protein secondary, tertiary and quaternary structure. This in turn affects the binding properties and specificities between mature proteins, allowing self-nonself recognition to occur.

Different strategies may be employed to generate the high levels of polymorphism required by allorecognition systems. Allorecognition genes are often richly allelic. For example, fusion-rejection decisions in *H. symbiolongicarpus* are largely under the control of two tightly-linked, highly polymorphic genes, *alr1* and *alr2* (Rosa et al. 2010); two contacting colonies require at least one shared allele at both *alr1* and *alr2* for recognition as self and subsequent successful fusion. Complementary DNA (cDNA) sequencing has identified around 200 unique *alr2* alleles within a single Connecticut, USA. *H. symbiolongicarpus* population (Gloria-Soria et al. 2012). The rich allelic nature of these genes facilitates only low rates of colony fusion – experimental manipulations of *H. symbiolongicarpus* have demonstrated fusion rates at less than 5% (Rosa et al. 2010). Similarly, fusibility assays in three Israeli

populations of *B. schlosseri* suggest the existence of over 300 FuHC alleles per population (Rinkevich et al. 1995). The putative *B. schlosseri* histocompatibility receptor, *fester*, is also richly allelic, with at least 21 alleles observed in one study (Nyholm et al. 2006).

Although the function of allorecognition proteins predicts that they be equipped with polymorphic extracellular regions, known molecules associated with allorecognition processes vary in their precise localisation and distribution of polymorphisms across their lengths (Figure 1.2a). Sequence polymorphism in *alr1* and *alr2* is largely restricted to particular hypervariable regions (Nicotra et al. 2009; Rosa et al. 2010) (Figure 1.2a). Within the FuHC locus, variation in the new candidate allorecognition gene *BHF* (Voskoboynik et al. 2013) and in *sFuHC* and *mFuHC* (De Tomaso et al. 2005; Nydam et al. 2013b; Voskoboynik et al. 2013) is distributed across each protein's length; in *BHF,* polymorphism is somewhat more prominent within the first 300 nucleotides and is absolutely predictive of fusibility outcomes (Voskoboynik et al. 2013). *fester* polymorphism is restricted to the extracellular region (Nyholm et al. 2006). The recently-characterised *Hsp40-L* also resides within the FuHC locus, and despite being a cytoplasmic protein, is similarly highly polymorphic with diversity localised to the C-terminal region (Nydam et al. 2013a).

In addition to sequence polymorphism, numerous other mechanisms, such as alternative splicing, post-transcriptional modification, recombination and RNA editing, may also be used to create diversity in allorecognition systems, either individually or in combination with one or more other processes (Ghosh et al. 2011).

c. Phase three: Discrimination

The final self-nonself recognition phase, discrimination, may proceed in diverse ways, complicating attempts to make generalisations about the molecular requirements of this stage. System-specific information is required in order to make predictions about the nature of the particular processes occurring therein. For example, systems that utilise differential cell adhesion as a passive discrimination mechanism may be predicted to possess a membrane-bound receptor molecule capable of tethering self cells together. Alternatively, in processes with differential outcomes, where recognition activates or represses a particular cascade or pathway, we can predict the presence of transmembrane receptor

proteins with cytoplasmic tails linking to downstream effector molecules or completely internalised cytoplasmic proteins. The nature of these receptor and effector molecules will vary depending on their precise mechanisms of action. There is, however, evidence of a degree of conservation in the downstream response to allorecognition challenge in marine invertebrates, with particular binding and catalytic proteins, including heat shock proteins, pattern recognition receptors and immunophilins, being implicated in the responses to allorecognition challenge in both cnidarians and ascidians (Oren et al. 2013).

1.1.4 The genomic basis of allorecognition

Self-nonself recognition appears to be a ubiquitous feature of metazoans, however research into the genetic basis of metazoan allorecognition has failed to find preserved evidence of a directly-shared evolutionary history between the 'frontline' allorecognition molecules of different taxa (Table 1.2). Regardless of the evolutionary origins and initial genetic sources of these allorecognition systems, extant systems have and continue to diverge along different evolutionary lineages via mutation, exon (domain) shuffling and molecular tinkering. In conjunction with the shared molecular features that exist between diverse allorecognition systems discussed earlier, it is becoming increasingly clear that allorecognition loci often share commonalities in various genomic properties as well. Here I discuss two trends apparent in the genomic loci encoding diverse allorecognition systems that have already been identified with existing data.

a. Clustering of allorecognition genes

One striking feature of the allorecognition systems characterised to date is that their component genes tend to co-occur in clusters of multiple, usually structurally similar genes (Figure 1.3); but see Voskoboynik (2013) for an exception. The large modular structure of the individual genes, coupled with the tandemly repeated nature of the loci, mean that these regions are often large. The *H. symbiolongicarpus alr1* and *alr2* genes have been mapped to a single genomic interval, the allorecognition complex (ARC) (Cadavid et al. 2004). A 700 kb sub-complex resides within the ARC, in which *alr1* is clustered amongst an additional ten Ig-like domain-encoding genes; at least four of these genes are polymorphic (Rosa et al. 2010). Although the precise role of these genes is unknown, the variable members remain plausible candidates for other currently unidentified allodeterminants within this species. Similarly, AFs, putative

Table 1.2 Selected self-nonself molecular systems in metazoans

Group	Example	Molecular system and putative functions	IM	TLR	CMP	AI
Poriferans	*Amphimedon queenslandica* *Clathria prolifera'*	AFs - histocompatibility, cell adhesion*	+	+	-	-
Cnidarians	*Hydra magnipapillata* *Nematostella vectensis*	ARC (alr1 and alr2) - histocompatibility*	+	+	+	-
Crustaceans	*Daphnia pulex* *Penaeus monodon'*	Crustins - antimicrobial peptides (Smith et al. 2008) Penaeidins - antimicrobial peptides (Destoumieux et al. 1997)	+	+	+	-
Insects	*Drosophila melanogaster*	Dscam - neuronal patterning (Schmucker et al. 2000), pattern recognition receptor function (Dong et al. 2006)	+	+	+	-
Nematodes	*Caenorhabditis elegans*	Various antimicrobial peptides (Bogaerts et al. 2010)	+	+	+	-
Gastropods	*Biomphalaria glabrata'* *Haliotis spp.', Lottia gigantea*	FREPs - parasite defense (Zhang et al. 2004)	+	+	+	-
Echinoderms	*Strongylocentrotus purpuratus*	Extensive expansion of TLR and NLR families (Hibino et al. 2006) RAG1/2-like molecules - possible gene rearrangement role (Fugmann et al. 2006)	+	+	+	-
Cephalochordates	*Branchiostoma floridae*	VCBPs - host-microbe interactions (Cannon et al. 2002)	+	+	+	-
Ascidians	*Botryllus schlosseri* *Ciona intestinalis*	FuHC locus (s/tmFuHC, fester, uncle fester, BHF - ?) – histocompatibility* VCBPs - host-microbe interactions (Dishaw et al. 2011)	+	+	+	-
Jawless vertebrates	*Petromyzon marinus* *Eptatretus burgeri'*	VLRs - adaptive immunity (Pancer et al. 2004)	+	+	+	-
Jawed vertebrates	*Mus musculus* *Danio rerio, Homo sapiens*	MHC, TCR, RAG and Ig molecules - adaptive immunity	+	+	+	+

All example species have a sequenced genome except where otherwise indicated (‘). Molecules listed are either unique or characteristic of the phylogenetic group, or well-studied therein. The far right of the table indicates the presence (+) or absence (-) of major immune pathways; IM – innate immunity, TLR – TLR (Toll-like receptor) pathway, CMP – complement system, AI – ‘true’ adaptive immunity. * - discussed in text, refer for references.

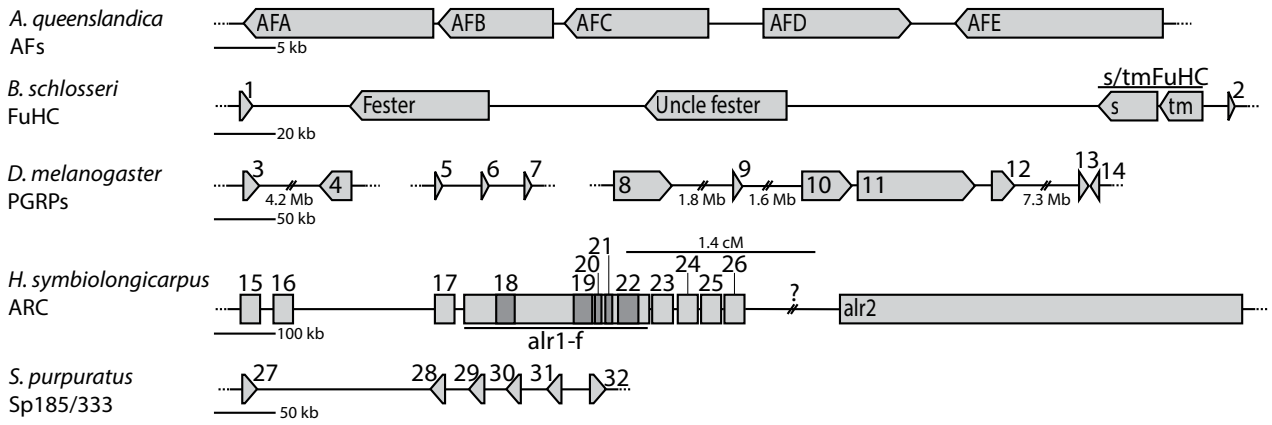


Figure 1.3 Genomic clustering of invertebrate self-nonself recognition genes
Genomic organisation of clustered self-nonself recognition and allorecognition genes, from selected invertebrate species. Shown are the *Amphimedon queenslandica AFs* (Gauthier 2009), various reported *FuHC* locus genes from *Botryllus schlosseri* (De Tomaso et al. 2005; Nyholm et al. 2006; Nydam et al. 2013a; 2013b; Voskoboynik et al. 2013), the peptidoglycan recognition protein genes (*PGRPs*) from *Drosophila melanogaster* (Werner et al. 2000), the *Hydractinia symbiolongicarpus ARC*, including the uncharacterised IgSF-like genes present in the region (Nicotra et al. 2009; Rosa et al. 2010), and the *Sp185/333* gene cluster from *Strongylocentrotus purpuratus* (Miller et al. 2010). The *D. melanogaster PGRP* genes sit in three separate genomic regions, corresponding left to right to the X, 2R and 3L chromosomes, respectively. *PGRP* genetic coordinates are taken from the *D. melanogaster* genomic assembly hosted by Ensembl. The *H. symbiolongicarpus ARC* has not yet been fully mapped beyond linkage analysis, therefore the precise distance between the *alr1* and *alr2* regions is unknown. Five genes (*IgSF-like-1, -4, -7, -X* and *–Y*) sit within the current limits of the *alr1*-containing interval. In all cases, only known, clustered gene family members are shown. For numbered genes, names and Ensembl accession numbers (in brackets, for PGRP genes) are as follows; 1: *BHF*, 2: *HSP40*, 3: *PGRP-SA* (FBgn0030310), 4: *PGRP-LE* (FBgn0030695), 5: *PGRP-SC1A* (FBgn0043576), 6: *PGRP-SC1B* (FBgn0033327), 7: *PGRP-SC2* (FBgn0043575), 8: *PGRP-LD* (FBgn0260458), 9: *PGRP-SD* (FBgn0035806), 10: *PGRP-LA* (FBgn0035975), 11: *PGRP-LC* (FBgn0035976), 12: *PGRP-LF* (FBgn0035977), 13: *PGRP-SB2* (FBgn0043577), 14: *PGRP-SB1* (FBgn0043578), 15: *IgSF-like-F*, 16: *IgSF-like-G*, 17: *IgSF-like-A*, 18: *IgSF-like-7*, 19: *IgSF-like-4*, 20: *IgSF-like-X*, 21: *IgSF-like-Y*, 22: *IgSF-like-1*, 23: *IgSF-like-B*, 24: *IgSF-like-C*, 25: *IgSF-like-D*, 26: *IgSF-like-E*, 27: *Sp185/333-A2*, 28: *Sp185/333-B8*, 29: *Sp185/333-D1y*, 30: *Sp185/333-D1g*, 31: *Sp185/333-D1b*, 32: *Sp185/333-E2*.

allorecognition molecules in sponges (Bonner and Slifkin 1949; Moscona 1968; Humphreys 1970; Henkart et al. 1973; Müller and Zahn 1973; Fernàndez-Busquets et al. 1998; Fernàndez-Busquets and Burger 1999), are also encoded by a set of clustered genes in the *A. queenslandica* genome. Here, five *AF* genes sit within an 80 kb cluster of the genome, with a sixth putative *AF* sitting alone elsewhere in the genome (Gauthier 2009). Finally, while new evidence suggests that *B. schlosseri* histocompatibility may be encoded by a single gene, *BHF* (Voskoboynik et al. 2013), the FuHC locus also contains other genes that appear to contribute to the allorecognition phenotype (De Tomaso et al. 2005; Nyholm et al. 2006; McKitrick and De Tomaso 2010; Nydam et al. 2013b; discussed in theory by Harada 2013). *sFuHC and mFuHC* genes (De Tomaso et al. 2005; Nydam et al. 2013b), which correlate well with

predicted allorecognition properties (De Tomaso et al. 2005; Nydam et al. 2013b) and fusibility outcomes (Voskoboynik et al. 2013), are situated within ~400 kb of other candidate regulators of allorecognition, *fester* and *uncle fester* (Nyholm et al. 2006; McKitrick et al. 2011). Clustered genes have also been reported from the immune or self-nonself recognition systems of other species (Figure 1.3) including *Drosophila melanogaster* (Werner et al. 2000), the purple sea urchin *Strongylocentrotus purpuratus* (Miller et al. 2010), chickens and zebra finches (Hellgren and Ekblom 2010) and the fungus *Neurospora crassa* (Micali and Smith 2006).

The clustering of allorecognition genes in part reflects their origins through tandem duplication, but cluster maintenance appears to have occurred via natural selection. Clustering of allorecognition genes facilitates the transfer of sequence information between regions within an immune locus, which may be executed in a number of ways including gene conversion, recombination and unequal crossing over, alternative splicing and gene inversion (Graham 1995; Ghosh et al. 2011). The clustering of related allorecognition genes may increase the efficiency and precision of co-regulated gene expression if required (Blumenthal 1998), as has been observed in suites of non-allorecognition genes from diverse taxa, such as zebrafish (Ng et al. 2009), *Caenorhabditis elegans* (Spieth et al. 1993), *Saccharomyces cerevisiae* (Zhang and Smith 1998) and *D. melanogaster* (Spellman and Rubin 2002). Clustering can also increase the co-inheritance of particular 'matched set' gene variants (Pál and Hurst 2003), although this hypothesis has not held up in other tests of non-immune ligand-receptor linkage in humans (Hurst and Lercher 2005). Birth and death evolution also can contribute to the maintenance of species-specific features amongst these grouped allorecognition genes (Nei and Rooney 2005). Finally, the primary driving force behind cluster maintenance in allorecognition and other immune systems may be the need to generate high levels of sequence diversity between individuals or species. The mutational divergence of duplicated genes, and the gain or loss of various functional domains, can further increase the rate of diversification within these clusters.

b. Positive selection

Allorecognition molecules are expected to display a high level of diversity within species, to produce different molecular signatures of self for distinct individuals. I have mentioned different methods of gene or transcript rearrangement to facilitate this variation above. However, mutation and

nucleotide-level variants also play a large role in the establishment of allorecognition diversity. Within the expectations of Kimura's neutral theory (Kimura 1968), synonymous mutations are predicted to be selectively neutral and therefore be observed at a higher frequency than non-synonymous mutations when comparing allele sequences within a species (Kimura 1977). Examples where non-synonymous differences are observed at a higher frequency than synonymous changes provide evidence that particular sequences or codons may be under positive selection, whereby amino acid change and protein diversification is selectively favoured (Jensen et al. 2007).

A number of examples of positive selection have been observed in characterised self-nonself recognition systems to date. For example, *Sp185/333* from *S. purpuratus* (Terwilliger et al. 2006), the parasite defense gene *FREP3* from the freshwater snail *Biomphalaria glabrata* (Zhang et al. 2001), the fertilisation genes *lysin* and *VERL* (vitelline envelope receptor for lysin) from the abalone *Haliotis* spp. (Metz et al. 1998; Lyon and Vacquier 1999; Yang et al. 2000; Galindo et al. 2003), the *H. symbiolongicarpus alr1* and *alr2* genes (Nicotra et al. 2009; Rosa et al. 2010), *Dictyostelium discoideum tgrB1* and *tgrC1* (Benabentos et al. 2009) and *het-c* and *pin-c* from the *N. crassa* heterokaryon incompatibility system (Hall et al. 2010) all possess codons which are predicted to be under positive selection. Because of the inherent requirement for self-nonself recognition, immune and allorecognition proteins to generate high levels of diversity, examples of positive selection will certainly be identified at increasing rates as more genome data become available and alleles from a greater number of individuals are surveyed.

1.2 Research introduction

Effective multicellularity requires the constituent cells of an organism to sacrifice their own autonomy and, for most cells, reproductive contribution. The multicellular state is therefore potentially compromised in instances of tissue fusion and cell transfer between conspecific individuals. True cooperation can only be maintained by natural selection if all constituent cells of an organism are genetically identical; nonself invaders of a host do not face the same selective pressures for cooperation. Allorecognition systems, which prevent the invasion of an individual by nonself cells, are therefore widespread amongst metazoans. Despite the apparent lack of directly-shared evolutionary history between the 'frontline' allorecognition molecules of different taxa (Table 1.2), all such systems function

in the same basic way: they must detect the presence of a cell, determine whether the cell is self or nonself, and take some discriminatory action based upon this decision (Chapter 1.1.3). For this reason, many allorecognition systems share similar features. For instance, allorecognition loci are often clusters of multiple allorecognition genes, which encode for large, at least partly extracellular modular proteins that are highly variable within a species.

For my thesis, I investigated the sponge (phylum Porifera) allorecognition system, focussing in particular on the aggregation factor (AF) gene family. A better understanding of sponge allorecognition is valuable for four main reasons. First, sponges are representatives of one of the oldest extant metazoan lineages (it remains undetermined whether sponges or ctenophores are the sister group to the rest of the Metazoa; Ryan et al. 2013), having existed around 800 million years ago (Erwin et al. 2011). Sponges also occupy an ideal phylogenetic position for the study of the forces driving transition to a multicellular state. This is particularly significant here in light of the importance of allorecognition for maintaining multicellular integrity. Second, as sessile invertebrates that commonly inhabit crowded benthic environments, sponges are particularly reliant on allorecognition to resist loss of habitat space, chimerism and possible germline parasitism by neighbouring conspecific sponges in the event of overgrowth. This has contributed to the evolution of a sophisticated allorecognition system capable of recognising, and taking differential action against, self and nonself cells and individuals, so understanding this system is of interest to understanding the ecological forces acting on sponges. Third, sponge self-nonself recognition and the activity of the AFs have been well-studied on phenomenological and biochemical levels (reviewed by Fernàndez-Busquets and Burger 1999; 2003). However, characterisation of the underlying genes and encoded protein sequences has been limited to date. Finally, I am interested in the molecular commonalities that exist between apparently unrelated allorecognition systems, as these features provide insight into the universal selective pressures driving the evolution and function of these divergent systems. Understanding these commonalities, however, first requires an understanding of the features of allorecognition systems from a diverse suite of taxa. Better characterisation of the sponge allorecognition system therefore allows more meaningful comparison with allorecognition systems in other species. For these reasons, I sought to better understand the underlying genetic properties of the *AFs*, to gain a fuller picture of the functions and evolution of these putative sponge allorecognition genes.

1.3 Overview of sponge allorecognition

1.3.1 The sponge allorecognition response

Adult sponge grafting experiments involve artificially bringing pieces of sponge tissue into contact using either the parabiosis or the less reliable (Neigel and Avise 1985) insertion graft technique (reviewed by Müller et al. 1999a). Graft donor tissue can be derived from a single individual (autograft), or from two individuals of the same (allograft) or different (xenograft) species. Graft acceptance, where tissue fusion promotes complete repair of the graft interface to form a single continuous piece of tissue, is limited almost exclusively to autografts (Moscona 1968; Hildemann et al. 1979; Smith and Hildemann 1986; Fernàndez-Busquets and Burger 1997; Gauthier and Degnan 2008). Self and nonself graft responses - the timing of reaction onset and duration (Hildemann et al. 1979; 1980; Bigger et al. 1981; Van de Vyver and Barbieux 1983; Humphreys 1994; Yin and Humphreys 1996; Fernàndez-Busquets and Burger 1997; 1999), level of aggression (Hildemann et al. 1980; Bigger et al. 1981; Van de Vyver and Barbieux 1983; Yin and Humphreys 1996) etc. - differ between species. However, responses to particular graft combinations within or between species are generally repeatable and predictable, and reveal a hierarchical genetic immunological relationship between conspecifics (Hildemann et al. 1979; 1980; Bigger et al. 1981; Kaye and Ortiz 1981; Neigel and Avise 1983; Neigel and Schmahl 1984; Neigel and Avise 1985; Wulff 1986; Fernàndez-Busquets and Burger 1997). This shows that sponges possess a fully functional allorecognition system that is genetically encoded and is capable of recognising and discriminating between self and nonself. The graft response is discussed in greater depth in Chapter 6.

1.3.2 Aggregation factors

Sponges are a classical model system for the study of cell adhesion. In 1907, Wilson demonstrated that sponge cells, dissociated by passing through a fine mesh, segregate species-specifically and re-assemble into small aggregates (Wilson 1907). The existence of individual-specific cell sorting and reaggregation remains ambiguous, but appears to be dependent on the experimental system and species tested (reviewed by Fernàndez-Busquets and Burger 1999). The cellular reaggregation process is calcium ion-dependent (Galtsoff 1925), and chemical dissociation techniques - where tissue is washed in calcium and magnesium free sea water - have also been successfully applied to this system (Humphreys et al. 1960). Reaggregation of dissociated cells is inhibited under a number of conditions, including

the absence of available calcium and magnesium ions (Galtsoff 1925), high salinity (Galtsoff 1925), exposure to antibodies raised against sponge cell suspensions (Spiegel 1954; Conrad et al. 1981), or incubation at low temperatures to reduce cellular metabolic rate and motility (Galtsoff 1925). Inhibition following chemical dissociation can be reversed by addition of the cell-free supernatant derived during the initial dissociation process (Humphreys 1963), indicating that an extracellular product is lost to the supernatant at this step, and that this product mediates species-specific cell adhesion and reaggregation. This product was later isolated (Henkart et al. 1973; Müller and Zahn 1973) and named 'aggregation factor' (AF) (Moscona 1968).

AFs are sponge specific (Srivastava et al. 2010) extracellular proteoglycans (Fernàndez-Busquets and Burger 2003), that are linear in *Halichondria bowerbankii, H. panicea, Haliclona oculata, Suberites ficus* and *Terpios zeteki* (Humphreys et al. 1977; Müller et al. 1978a; Jarchow et al. 2000), but which exhibit a novel 'sunburst'-like confirmation in *C. prolifera* (), *Clathria parthena, Geodia cydonium* and *Oscarella tuberculata* (Cauldwell et al. 1973; Henkart et al. 1973; Müller and Zahn 1973; Humphreys et al. 1975; 1977; Humbert-David and Garrone 1993; Jarchow et al. 2000) (Figure 1.4). Circular proteoglycans, which have not been observed outside the sponges, have thus been named the 'spongicans' (Fernàndez-Busquets and Burger 2003); this circular form is also the best studied AF form to date. The two main protein components of the circular AF complex in *C. prolifera* are MAFp3 and MAFp4. Twenty of each subunit come together to make up the backbone and radiating arms, respectively, of each ring (Jarchow et al. 2000). Attached glycan subunits are an integral mediator of AF binding (Misevic and Finne 1987; Misevic and Burger 1990a; 1990b; 1993), although some binding ability also appears to reside in the AF protein backbone (Jarchow et al. 2000).

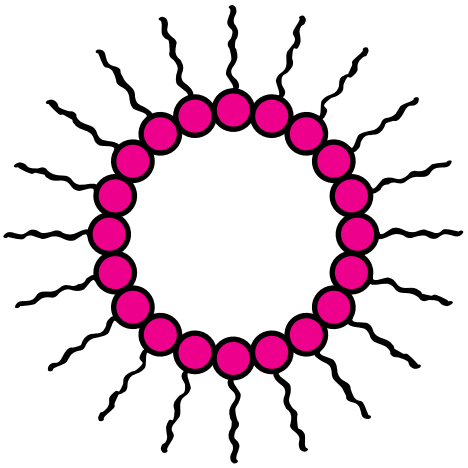


Figure 1.4 Schematic of the *C. prolifera* AF protein core
Twenty head (pink circles, corresponding to the MAFp3/Wreath domain region) and arm (tails, corresponding to the MAFp4 region) come together to form a ring structure that is associated with other proteins and glycan subunits *in vivo*.

C. prolifera AF-mediated cell adhesion occurs when pairs of AFs form bridge-like structures between homologous sponge cells. Here, the head subunits of the two AFs interact via their associated glycans in a calcium ion-dependent manner, while the arm subunits interact with aggregation receptors at the cell surface, again with the assistance of glycan subunits and other associated proteins, but in a calcium ion-independent manner (reviewed by Fernàndez-Busquets and Burger 2003). In the reaggregation model system, AF binding by the cell promotes a host of downstream metabolic changes, including the activation of various cell signalling and regulatory components (Dunham et al. 1983; Müller et al. 1987; Rottmann et al. 1987; Schröder et al. 1988; Pfeifer et al. 1993; Müller et al. 1994; Wimmer et al. 1999a) and the upregulation of DNA, RNA and protein synthesis (Müller et al. 1976a). However, AF-mediated aggregation has also been observed in non-metabolically active contexts. For instance, purified AF has been coupled to beads and found to induce bead aggregation in a species-specific manner (Jumblatt et al. 1980; Popescu and Misevic 1997; Jarchow et al. 2000), while fixed (i.e. killed) cells have also been shown to aggregate after exposure to AFs (Moscona 1963; Jumblatt et al. 1980). These findings reveal that AF-mediated aggregation is, in part, a passive response to physical adhesive forces between cell-bound AFs, but that ligand-receptor binding also promotes a host of other downstream changes and signalling events. This potentially allows for greater diversity or utility of the system, if different AF receptors operate in different biological contexts or cell types (discussed in Chapter 3).

1.4 Aggregation factors as putative allorecognition molecules

Although the AFs have been best-characterised as molecules mediating species-specific cell adhesion, they have also been proposed as candidate allorecognition molecules in the sponge (reviewed by Fernàndez-Busquets and Burger 1999). Although their hypothetical functional role in tissue grafting and other immune challenges remains to be confirmed, the AFs fulfil the requirements and predictions for candidate allorecognition molecules. The evidence supporting this hypothesis is outlined below under the framework of the three essential phases of self-nonself recognition and the predicted features of molecules performing these phases.

1.4.1 Detection: Allorecognition systems rely on evaluator-label (e.g. cell-cell) contact

Profound similarities exist between the functional requirements of the allorecognition detection phase and of cell adhesion processes; it is likely that animal cell adhesion and allorecognition systems

are evolutionarily related (Bodmer 1972; Rothenberg 1978; Curtis 1979; Edelman 1987; Matsunaga and Mori 1987; Fernàndez-Busquets and Burger 1999; Grice and Degnan 2015a). Both systems require the presence of compatible ligands and receptors, which interact specifically to facilitate binding and/or communication between their respective cells. Each ligand or receptor may have multiple possible binding partners. The structural features of each class of molecule are also similar, including the frequent inclusion of transmembrane domains and large extracellular regions comprised of tandemly repeated extracellular protein domains (Figure 1.2, Table 1.1). Examples of these repeated structures can be seen in the various members of the cadherin, immunoglobulin, integrin and selectin cell adhesion families (Figure 1.2b).

Cell adhesion molecules also play roles in cell recognition and sorting events, for example during tissue development and organogenesis (McNeill 2000). Cell aggregation experiments have demonstrated the key role of cadherins in differential cell type-specific adhesion, again showing a clear functional relationship between cell adhesion processes and self-nonself recognition molecules. However, while differential cell interactions in allorecognition are underpinned by highly polymorphic self-nonself recognition molecules, differential cadherin binding is largely mediated by the control of cell surface deployment of invariant molecules (Halbleib and Nelson 2006; Leckband and Prakasam 2006). The link between cell sorting and allorecognition was further highlighted by studies of cell fate in chimeric juvenile sponges (Gauthier and Degnan 2008). Experimental fusion of pairs of fluorescently-labelled sponge postlarvae or juveniles led to an initial period of cellular intermingling. However, the chimeras later underwent near-complete cell sorting, whereby cells from one individual contributed predominantly to the choanocytes, while the cells of the other individual formed the pinacocytes and mesohyl (Gauthier and Degnan 2008). This differential cell sorting process is reminiscent of the cadherin-mediated sorting of cell populations discussed above. Although the molecule/s facilitating this individual-specific cell sorting are unknown, the intriguing strict separation of cell types by individual demonstrates a further link between cell adhesion and migration processes and self-nonself recognition.

The AFs are cell adhesion molecules that fulfil the requirements and expectations of the detection phase of self-nonself recognition. The detection phase requires the presence of molecules that facilitate intercellular interactions, predicting the presence of full or partially extracellular molecules that bind

homologous or heterologous molecules on a neighbouring cell. AFs are known to function in cell-cell interactions by forming bridges between sponge cells, through associations with an aggregation receptor (Weinbaum and Burger 1973; Müller et al. 1976b; Jumblatt et al. 1980; Kuhns et al. 1980; Blumbach et al. 1998). This binding is facilitated in part by the attached glycan subunits (Misevic and Finne 1987; Misevic and Burger 1990a; 1990b; 1993) and with other proteins associated with the AF complex (Schütze et al. 2001). Allorecognition molecules, like those involved in cell adhesion, are often large proteins equipped with tandemly repeated domains. This attribute is also fulfilled by the protein backbone of the AFs, which in *C. prolifera* and *A. queenslandica* are predicted to encode numerous Calx-beta domains in tandem (Fernàndez-Busquets et al. 1998; Gauthier 2009). Finally, the organisation of the AF locus as a large gene cluster means that this locus resembles already-characterised allorecognition loci that are similarly clustered. This clustering may be important for gene co-regulation or diversification (Chapter 1.1.4). The properties of the AF molecules are therefore compatible with a potential role in allorecognition.

1.4.2 Recognition: Allorecognition systems possess a high level of genetic polymorphism

Individual-level self-nonself recognition cannot operate within a population without a polymorphic genetic system capable of producing molecular labels, or combinations thereof, that are unique to each individual. The sponge system is no exception, with large-scale studies demonstrating that tissue contact between different individuals is almost invariably rejected (Hildemann et al. 1979; 1980; Van de Vyver and Barbieux 1983; Fernàndez-Busquets and Burger 1997). This implies that a diverse allorecognition system must be at play. In instances where fusion is seen between different sponge individuals, it occurs at a rate that is inversely proportional to the physical distance (and therefore, genetic relatedness) between the two sponges in the field (Jokiel et al. 1982; Neigel and Avise 1983; Neigel and Schmahl 1984). This further emphasises the role of genetic diversity in promoting self-nonself recognition. The AFs are one sponge gene family that fulfils the requirements of the recognition phase of self-nonself recognition. AFs are sponge specific (Srivastava et al. 2010); various characterised invertebrate allorecognition genes exhibit similar lineage-specificity. The AFs in *C. prolifera* are polymorphic, with five *MAFp3* mRNA isoforms identified, some of which are allelic and others which may represent different genes (Fernàndez-Busquets and Burger 1997). Tests of the genomic DNA (gDNA) sequence diversity of *MAFp3* and *MAFp4* by RFLP (restriction fragment length polymorphism) profile analysis

revealed a complete concordance between fusion/rejection graft outcomes and different RFLP profiles. In addition to polymorphism of the AF protein backbone, the AF-associated glycoprotein p210 (also referred to as S1) (Varner 1996) is also polymorphic; at least some of this polymorphism exists at the glycan level (Fernàndez-Busquets and Burger 1997). These results indicate that the AFs provide the high levels of variation required of an allorecognition system, and that AF molecule diversity can be created by the combinatorial contributions of the protein and glycan components.

1.4.3 Discrimination: Differential action occurs as a result of recognition as self or nonself

Different self-nonself recognition systems will take different paths to discrimination between self and nonself. What this discrimination looks like will also differ; the only requirement of this phase is that nonself or self is rejected in some way as is appropriate. If the AFs are indeed involved in sponge allorecognition, their role in the discrimination phase is likely two-fold. First, AF-AF interactions appear to result at least partially from passive adhesive forces, as demonstrated by aggregation experiments with bead-coupled AFs or fixed cells (Chapter 1.3.2). Therefore, selective adhesion of homologous AFs may promote simultaneous detection, recognition, and passive discrimination between self and nonself cells. Second, AF-receptor binding is coupled to various downstream signalling and regulatory pathways (Chapter 1.3.2); this may stimulate active rejection activity upon exposure to nonself. However, this proposed nonself response remains unexplored, particularly within whole-tissue grafts as opposed to the more artificial single-cell reaggregation model system.

The hypothetical role of the AFs in sponge allorecognition is untested; however, evidence suggests that the AFs do at least have some functional involvement in this process. For example, the *C. prolifera* genes *MAFp3* and *MAFp4* are upregulated in both auto- and allografted tissue, compared with normal tissue (Fernàndez-Busquets et al. 1998). Similarly, the deglycosylated form of the MAFp3 protein (present exclusively in archaeocytes) (Fernàndez-Busquets et al. 2002) is recruited to the site of allogeneic contact (Fernàndez-Busquets et al. 1998). Alone, this provides evidence that the AFs have some role in sponge allorecognition. However, when paired with the other evidence discussed above that demonstrates that the AFs possess other typical allorecognition molecule features, support is gained for the hypothesis that the AFs are not just involved in the allorecognition response, but are in fact the main sponge allodeterminants.

1.5 The *Amphimedon queenslandica* model system

The research I present across this thesis predominantly focuses on the AF gene complement of the haplosclerid demosponge *A. queenslandica* (Porifera, Demospongiae, Haplosclerida, Niphatidae) (Hooper and van Soest 2006). An *A. queenslandica* population is found in Shark Bay, Heron Island Reef (Great Barrier Reef, Australia), although populations are also found elsewhere in the Great Barrier Reef, off One Tree and Magnetic Islands. Related populations have also been observed in Egypt, Japan and the Red Sea, suggesting a wide Indo-Pacific distribution (Hooper and van Soest 2006; Degnan et al. 2008a). On Heron Island Reef, *A. queenslandica* adults are found on the shallow reef flat and crest, generally in rock crevices or in coral rubble, and can be easily collected by snorkelling at low tide (Leys et al. 2008). *A. queenslandica* is a hermaphroditic spermcast spawner which broods embryos year round, allowing easy access to developmental material (Leys et al. 2008; Degnan et al. 2008a; 2008b).

A. queenslandica was the first sponge to have its genome sequenced (Srivastava et al. 2010) and at present this remains the only publically-available demosponge genome. *A. queenslandica* is therefore a valuable model system for the study of the genetics and evolution of key evolutionary and developmental gene families. Three gene model predictions are used across this thesis. The majority of analyses use the Aqu2.1 gene model predictions; these are the most recent and currently best gene model predictions (S. Fernandez-Busquets and B. Degnan, manuscript in preparation). A smaller number of analyses use either the publicly-available Aqu1 gene models (Srivastava et al. 2010) or the in-house Aqu2.0 models (S. Fernandez-Busquets and B. Degnan, personal communication). These older gene models were used for analyses performed prior to the completion of the Aqu2.1 gene models or where other tools relied on the earlier models. The gene models used for each analysis are specified throughout.

1.6 Aims of this study

The general goal of this study was to use genomics and transcriptomics to investigate the AF gene family, in order to better understand the protein backbone of these proposed allorecognition molecules. I examined the AFs at four levels: between species, broadly within a single species, between different conspecifics, and within individuals over time. I pursued four research aims that investigated each of these levels in turn:

Aim 1. To investigate the evolutionary origins of the aggregation factors (between species comparisons)

AFs are not found outside the sponges (Srivastava et al. 2010); however the majority of work on the AFs to date has focussed on particular model demosponge species (e.g. *C. prolifera, Ephydatia muelleri, Geodia cydonium, Suberites domuncula* etc.). In Chapter 2, I developed a set of criteria to identify candidate *AF* sequences based upon the encoded protein domain architecture of characterised AFs. I then used these criteria to probe the genomes and/or transcriptomes of fourteen representative species spanning all four Poriferan classes - Calcarea, Demospongiae, Hexactinellida and Homoscleromorpha - in order to infer whether the *AFs* are ubiquitous to all sponges, or if they evolved after the divergence of the different sponge classes.

Aim 2. To characterise the genomic features of the A. queenslandica AF genes (within species analysis)

The majority of AF research to date has focussed on the proteoglycan AF complex and its components; only preliminary characterisation of the underlying genetic sequence has occurred (Fernàndez-Busquets et al. 1996; Fernàndez-Busquets and Burger 1997; Gauthier 2009). In *C. prolifera*, *AF* gDNA and mRNA sequences have been reported (Fernàndez-Busquets et al. 1996; Fernàndez-Busquets and Burger 1997), but the full *AF* gene complement in this species has not been elucidated. Previous work in this lab has seen the identification of six *AF* genes from the *A. queenslandica* genome (Gauthier 2009). In Chapter 2, I performed a detailed characterisation of the genomic and predicted protein properties of these six genes, to better understand the canonical genomic background of the protein backbone of the AF proteoglycan in *A. queenslandica*. In particular, I focussed on the highly structured genomic architecture of these genes, which contrasts with the low levels of sequence similarity observed between both AF genes and the repeated domains encoded therein.

Aim 3. To investigate how these genomic features might be diversified within and between A. queenslandica individuals to generate polymorphism (between individual comparisons)

As putative allorecognition molecules, the *A. queenslandica* AFs are predicted to exhibit high levels of between-individual diversity to allow precise self or nonself decision making; such diversity has been reported from the *C. prolifera* AF system. Across this thesis, I investigated the potential

contributions of three sources of sequence variation to diversification of the *AF* genes. In Chapters 4 and 6, I catalogued putative instances of alternative splicing of the *AF* genes in new transcriptome datasets spanning development (pre-competent larvae to adults; Chapter 4) and the auto- and allograft response (Chapter 6) in *A. queenslandica*. In Chapter 4, I then examined the amount of nucleotide diversity present in *AF* transcriptome sequencing reads from four *A. queenslandica* adult individuals. Finally, I asked whether RNA editing is a plausible mechanism by which the *AFs* and other genes could be diversified. In Chapter 5, I found that the ADAR (adenosine deaminase acting on RNA) RNA editing family exists in sponges, implying that RNA editing may occur in *A. queenslandica*. I investigated the phylogenetic distribution of the ADARs in early branching metazoans to develop a set of hypotheses regarding the early evolution of the ADAR family.

Aim 4. To examine AF gene expression profiles across A. queenslandica life history and in response to immune challenge (temporal comparisons)

To address the first component of this aim, in Chapter 3 I determined the expression profiles of the six *A. queenslandica* AF genes across sponge development (embryos to adults), to investigate the potential interplay between AF expression and activation of immunological competency in the sponge. I instead found that the AFs are developmentally expressed, particularly in metamorphosis. I subsequently identified a suite of other key developmental genes that display similar developmental expression patterns to the AFs, and used this information to hypothesise about a novel pre-immunological developmental role for the AFs in sponges.

Previous work has demonstrated that MAFp3 and MAFp4 mRNA is upregulated in auto- and allografts in *C. prolifera*. For the second part of this aim, I took a whole-transcriptome approach to investigate AF transcriptional activity in *A. queenslandica* grafts. To do so, I performed a three-day auto- and allograft experiment, before creating transcriptome sequencing libraries for each graft time point. I then analysed the quantitative expression patterns of the AFs and other genes in response to tissue grafting.

Overall, this thesis provides the first in-depth characterisation of the sponge AF gene family to date. This characterisation is broad-ranging, by studying these genes between species across vast evolutionary periods, down to the level of the individual nucleotide.

CHAPTER 2 - CHARACTERISATION OF THE AGGREGATION FACTOR GENES FROM FOURTEEN PORIFERAN SPECIES

2.1 Abstract

Aggregation factors are sponge-specific proteoglycans that are necessary for species-specific reaggregation of dissociated sponge cells, and are also implicated in the allorecognition response to attempted tissue grafts or fusions in conspecific sponge tissue. Aggregation factors have been well characterised biochemically, but knowledge of the genetic background of these proteoglycans is comparatively limited. I have identified novel aggregation factor candidates in the genomes or transcriptomes of thirteen sponge species distributed across the phylum Porifera. A typical aggregation factor sequence encodes numerous Calx-beta domains, a newly defined Wreath domain, and may include other domains such as Von Willebrand (types A or D) domains. In-depth analysis of the *Amphimedon queenslandica* aggregation factor suite reveals that these genes are tightly clustered and comprised of tightly defined exonic and domain structural units. However, these genes show little sequence identity within (i.e. between encoded repeated domains) or between genes. These findings suggest that aggregation factor sequences evolve rapidly, but that the overall integrity of these sequences is maintained by the genomic architecture of the locus.

2.2 Introduction

2.2.1 The macromolecular nature of sponge aggregation factors

Sponge aggregation factors (AFs) were first identified and isolated in the 1960s and 1970s, in studies exploring their role in the species-specific reaggregation of dissociated sponge cells (Chapter 1.3.2). Cellular aggregation is mediated by the formation of molecular bridges between cells, which are assembled through a complex association of several protein and carbohydrate components including the non-integral membrane protein aggregation receptor (AR) (Weinbaum and Burger 1973; Müller et al. 1976b) and the aggregation factor core structure (Henkart et al. 1973) . Bridge formation requires calcium-dependent homologous self-association of AF core structures, and calcium-independent

heterologous association of AFs with membrane-associated ARs (Jumblatt et al. 1980), probably via their associated glycans (Misevic and Burger 1990a; 1993). Estimates from AF binding studies suggest that each sponge cell may be associated with up to 28,000 AFs in vivo (Jumblatt et al. 1980).

The core AF is an extracellular proteoglycan (Henkart et al. 1973). Intriguingly, electron and atomic force microscopy of purified AFs from different demosponge species reveals interspecies differences in AF structure. AFs from *Halichondria panicea* (Jarchow et al. 2000), *Halichondria bowerbankii* (Humphreys et al. 1977), *Haliclona oculata* (Humphreys et al. 1977), *Suberites domuncula* (Müller et al. 1978a), *Suberites* (formerly *Ficulina*) *ficus* (Jarchow et al. 2000) and *Terpios zeteki* (Humphreys et al. 1977) are linear, and very similar in appearance to other classical proteoglycans (Fernàndez-Busquets and Burger 2003). Conversely, AFs from *Clathria* (formerly *Microciona*) *prolifera* (CpAFs) (Humphreys et al. 1975; 1977; Jarchow et al. 2000), *Clathria parthena* (Henkart et al. 1973), *Geodia cydonium* (Müller and Zahn 1973), and *Oscarella tuberculata* (Humbert-David and Garrone 1993) display a sunburst-like morphology with a circularised backbone that is otherwise very similar to the linear form. The circular, sunburst-like proteoglycan form appears to be unique to sponges (Fernàndez-Busquets and Burger 2003) and therefore is either a convergent trait, the result of secondary loss in a number of species whose ancestor possessed both forms, or it represents the ancestral AF form that was subsequently linearised in several demosponge clades.

The circularised AF from *C. prolifera* remains the best studied AF to date, and is comprised of a twenty-subunit central ring and twenty radiating arms, with each ring subunit binding one arm (Jarchow et al. 2000). Each ring subunit is encoded by MAFp3 and is coupled to one or two g200 glycan molecules, while each arm is encoded by MAFp4 and binds about 50 g6 glycans (Jarchow et al. 2000).

2.2.2 Core AF and AF-related sequences

Messenger RNA (mRNA), genomic DNA (gDNA) and protein sequences from *C. prolifera* MAFp3 and MAFp4 have been elucidated (Fernàndez-Busquets et al. 1996; Fernàndez-Busquets and Burger 1997; Fernàndez-Busquets et al. 1998). Both sequences have been shown to be highly polymorphic (Fernàndez-Busquets and Burger 1997; Fernàndez-Busquets et al. 1998), exhibiting small- (i.e. nucleotide-level) and large-scale (i.e. intronic, exonic and length variants) differences between

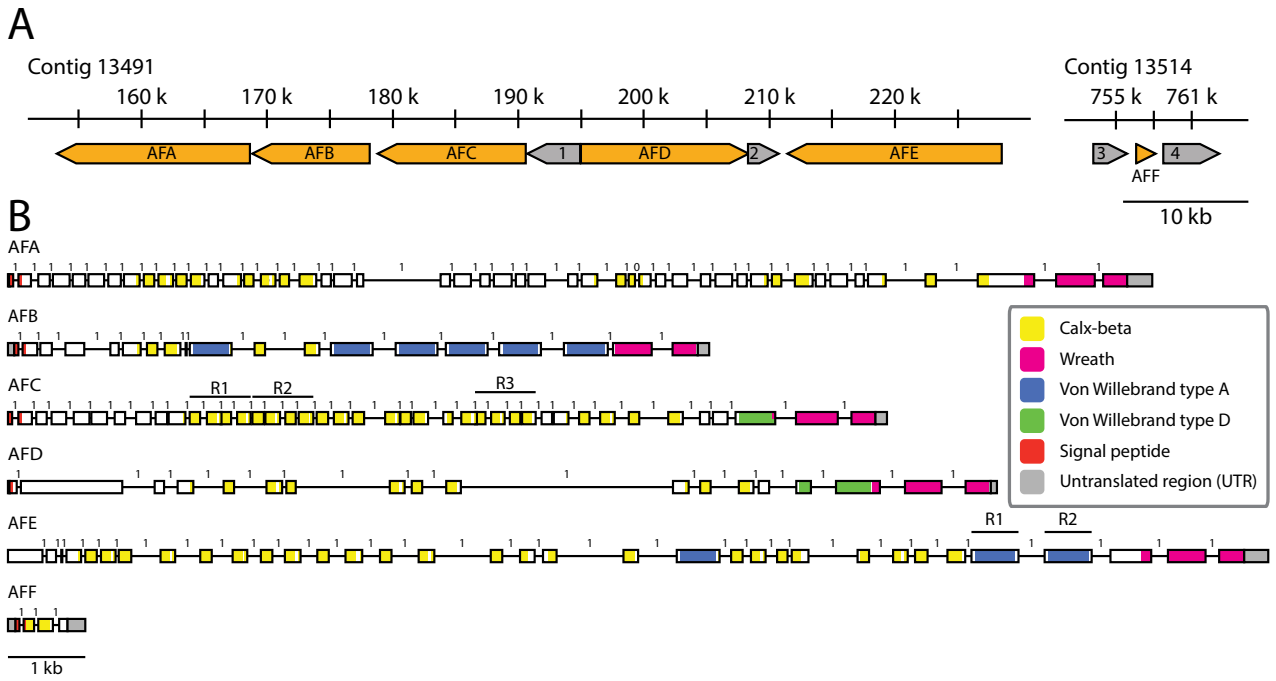


Figure 2.1 Genomic organisation and domain organisation of the *A. queenslandica* aggregation factor genes

Six *aggregation factor (AF)* genes are encoded in the *A. queenslandica* genome. (A) Five *AqAFs* (*AqAFA* - *AqAFE*) are clustered in an ~80 kb region on Contig 13491. Two non-AqAF genes are nested within the cluster: *autophagy-related protein 13-like* (1) and *sn1-specific diacylglycerol lipase beta-like* (2). The sixth *AqAF*, *AqAFF*, sits separately in the genome on Contig 13514, and is flanked by *zinc finger MYND domain-containing protein 10-like* (3) and *similar to centrosomal protein KIAA1731* (4). Non-AqAFs were identified based on the best BLASTp or BLASTx hit obtained from NCBI. *AqAFs* are shown in orange and non-*AqAFs* in grey. Chromosomal gene orientation is indicated by arrowheads representing the 3' end of each gene. (B) The gene model prediction for each *AqAF* gene is shown, with boxes representing exons. Each gene is oriented 5' to 3'. Genomic DNA regions encoding protein domains (Calx-beta, Von Willebrand type A (VWA), Von Willebrand type D (VWD) and Wreath domains) are coloured accordingly. Numbers above introns indicate the phase of each intron. *AqAFC* R1 – R3 and *AqAFE* R1 – R2: location of three (*AqAFC*) or two (*AqAFE*) repeated sequences encoded within the genomic DNA of each gene. The *AqAFE* repeats are independent of those present in *AqAFC*. Exons and introns are drawn to scale.

and within individuals (Fernàndez-Busquets and Burger 1997). Multiple isoforms have been identified from within single *C. prolifera* individuals (Fernàndez-Busquets and Burger 1997). Transcript analyses suggest that *MAFp3* and *MAFp4* are transcribed together as a single contiguous mRNA (Fernàndez-Busquets and Burger 1997; Fernàndez-Busquets et al. 1998). However, chemical dissociation of the core CpAF produces intact ring structures that lack attached arm subunits, suggesting that mature MAFp3 and MAFp4 peptides are independent (Jarchow et al. 2000). This apparent independence implies the presence of a post-translational peptide processing event in CpAF assembly (Fernàndez-Busquets and Burger 1997; Jarchow et al. 2000). *MAFp4* isoforms encode between three and fifteen Calx-beta

domains (Gauthier 2009), while *MAFp3* does not encode any known domain types (Fernàndez-Busquets and Burger 1997; Fernàndez-Busquets et al. 1998; Gauthier 2009). While *C. prolifera* AF-induced aggregation is primarily thought to be mediated by AF-associated glycan subunits, an *in vitro* study demonstrated that recombinant MAFp3 can induce cellular reaggregation in the absence of additional complex components (Jarchow et al. 2000). This suggests that the proteinaceous AF backbone may play a greater role in AF functionality beyond simply acting as a passive scaffold to support functional carbohydrate moieties.

An AF-related sequence, *GEOCY_AF*, was identified in a *G. cydonium* complementary DNA (cDNA) library based on sequence similarity to the *C. prolifera* core AF sequences (Müller et al. 1999b). This sequence encodes two Sushi domains and a region equivalent to the *C. prolifera* MAFp3 sequence. A second AF-associated sequence was later identified from a *G. cydonium* cDNA library, this time using antibodies raised against a fraction of enriched AF isolate (Schütze et al. 2001). This sequence, with the confusingly similar name *AF_GEOCY*, bears little resemblance to the core AF protein in *C. prolifera* or *G. cydonium*, instead being equipped with a single BAR domain and appearing to be a BIN1 homologue (Schütze et al. 2001).

The sponge *S. domuncula* also possesses an AF-related sequence, *SdSLIP* (Wiens et al. 2005). This sequence encodes one Calx-beta domain and also shares significant sequence similarity with *C. prolifera MAFp3*. SdSLIP was originally identified based on its sequence similarity to the putative *G. cydonium* aggregation factor core protein, *GEOCY_AF,* (Müller et al. 1999b). Curiously, however, SdSLIP does not appear to be a classical AF, acting instead as a binding partner of the bacterial endotoxin lipopolysaccharide (LPS) (Wiens et al. 2005). This suggests an unexplored relationship between the sponge allorecognition and bacterial defence systems. Regardless of the cellular function/s of SdSLIP, *S. domuncula* does possess an aggregation factor-mediated adhesion system; however, in this species the central AF is linear and has a higher protein content than that seen in *C. prolifera* (over 80% protein, compared with about 50% protein in *C. prolifera* (Henkart et al. 1973; Müller et al. 1978a).

The genome sequence of the demosponge *Amphimedon queenslandica* (Srivastava et al. 2010) encodes six putative aggregation factor (AF) genes, named *AqAFA* through to *AqAFF* (Figure 2.1a)

(Gauthier 2009). These sequences were identified by sequence similarity matches to the *CpAF* isoforms and to *SdSLIP* (Gauthier 2009). Membrane topology predictions from translated peptide sequences indicate that all AqAF proteins are secreted, except perhaps for AqAFE which is predicted to occur extracellularly yet lacks a signal peptide (Figure 2.1b) (Gauthier 2009). Members of three protein domain families are predicted to be present within the AqAF protein coding sequences (Figure 2.1b). As in MAFp4 and SdSLIP, Calx-beta domains are present in all AqAF proteins, in varying numbers (from one in AqAFF to twelve in AqAFC) (Gauthier 2009). However, unlike other AF or AF-related sequences, the AqAFs also contain Von Willebrand domains, with Von Willebrand type A (VWA) domains present in AqAFB and AqAFE, and Von Willebrand type D (VWD) domains in AqAFC and AqAFD (Gauthier 2009).

AqAFA to *AqAFE* are situated in an 80 kilobase pair kb (kb) gene cluster on a single chromosome (i.e. scaffold), oriented head-to-tail (except *AqAFD* which is inverted; Figure 2.1a) (Gauthier 2009). Two non-AF genes are also nested within this cluster, *autophagy-related protein 13-like* (Aqu1.225773/Aqu2.2.38626_001) and *sn1-specific diacylglycerol lipase beta-like* (Aqu1.225776/Aqu2.2.38628_001) (Gauthier 2009). *AqAFF* sits apart from the main cluster on a separate scaffold (Figure 2.1a) and is flanked by *zinc finger MYND domain-containing protein 10-like* (Aqu1.228576/Aqu2.2.42295_001) and *similar to centrosomal protein KIAA1731* (Aqu1.228578/Aqu2.2.42297_001) (Gauthier 2009).

Although some progress has been made towards better understanding the protein components of the AF core, most studies of the AF complex have been biochemical in nature. These studies have, in particular, focussed on the role of AF-associated glycan moieties in mediating adhesion specificity (reviewed by Fernàndez-Busquets and Burger 2003). However, MAFp3 sequence polymorphism is known to be correlated with tissue graft acceptance/rejection has been demonstrated (Fernàndez-Busquets and Burger 1997; Fernàndez-Busquets et al. 1998), and recombinant MAFp3 has been shown to induce cellular reaggregation *in vitro* in the absence of other AF components (Jarchow et al. 2000). These findings both suggest a role for the AF protein backbones in AF complex activity, beyond acting as a simple scaffold for functionally important carbohydrate residues.

Table 2.1 General properties of *A. queenslandica AF* genes

Gene	Accession numbers	gDNA size	cDNA size	Exon number	Median intron length	Domain architecture	Signal peptide?	GC content (gDNA)	Intergenic distances (down | upstream)
AqAFA	Aqu1.225771 Aqu2.1.38623_001	15.44 kb	9.09 kb	48	61 bp	SP (signal peptide) – 7 x Calx-beta – 1x Wreath	Yes	34%	Overlap | 120 bp
AqAFB	Aqu1.225772 Aqu2.1.38624_001	9.46 kb	5.96 kb	19	134.5 bp	SP – 2x Calx-beta – 6x VWA – 1x Wreath	Yes	38%	120 bp | 543 bp
AqAFC	hom.g29438.t1 Aqu2.1.38625_001	11.83 kb	7.85 kb	41	50.5 bp	SP – 12x Calx-beta – 1x VWD – 1x Wreath	Yes	34%	543 bp | 110 bp
AqAFD	1457081+2 Aqu2.1.38626_001	13.34 kb	5.16 kb	18	112.5 bp	SP – 5x Calx-beta – 1x VWD – 1x Wreath	Yes	32%	96 bp | 29 bp
AqAFE	hom.g29441.t1 Aqu2.1.38629_001	17.03 kb	8.42 kb	34	221 bp	9x Calx – 3x VWA – 1x Wreath	No	35%	704 bp | 1489 bp
AqAFF	Aqu1.228577 Aqu2.1.42296_001	1.04 kb	0.51 kb	4	58 bp	SP – 1x Calx-beta	Yes	36%	40 bp | 10 bp

gDNA = genomic DNA, cDNA = complementary DNA, kb = kilobase pairs, bp = base pairs, SP = signal peptide

In the present study, I first searched for likely AF candidate sequences within the genomes and transcriptomes of thirteen sponge species, including a newly-sequenced full transcriptome from *C. prolifera*. This analysis resulted in the identification of over 150 AF-like sequences from sponge species distributed across the Porifera. For the second part of this chapter, I investigated the relationship between genomic architecture and secondary protein structure in the six putative AF genes encoded in the genome of *A. queenslandica*. I report a remarkable conservation of genomic structure in these genes, including the tight restriction of protein domains to precise exon modules, and an intron phase distribution that differs significantly from that seen in the genome as a whole. This genomic constraint is juxtaposed with a high level of sequence divergence amongst domain sequences within and between individual AF genes.

2.3 Methods

2.3.1 A note on nomenclature

The aggregation factor (AF) complex is comprised of various protein and carbohydrate components, the identities and functions of which are not yet fully understood. In addition, some components and functionalities of the complex appear to differ between sponge species, and therefore may not be broadly applicable to the phylum Porifera as a whole. For the purposes of this chapter, any general reference to AFs refers to the core structure or encoding genes. References to other components of the system or to the complex as a whole will be explicitly stated.

The first part of this chapter describes the identification of a suite of sequences with features similar to known AF and AF-related sequences. Members of this list, either collectively or individually, are referred to as 'AF-like'. AF-like sequences deemed to represent probable AFs are referred to as 'AF candidates' or 'putative AFs'.

2.3.2 *A. queenslandica* AF sequence information

General information about the *A. queenslandica* AF (AqAF) protein and genomic DNA (gDNA) sequences is given in Table 2.1. A full list of AqAF accession numbers for different databases is given in Appendix 2.1 for cross-referencing purposes. Most analyses of *AqAF* sequence features described here are based on unpublished, in-house gene models (version Aqu2.1; S. Fernandez Valverde, B.

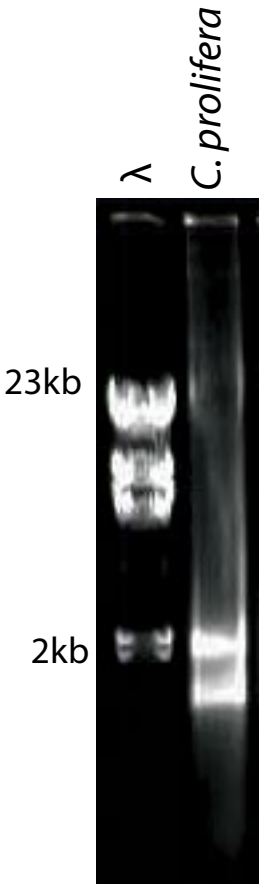


Figure 2.2 Qualitative analysis of *C. prolifera* RNA and DNA quality
A sample containing *C. prolifera* RNA and DNA was resolved on a 0.5x TBE 1% agarose gel by electrophoresis. Sample size was determined based on the lambda DNA size marker (Invitrogen).

Degnan and S. Degnan, unpublished data). The intron phase analysis in Calx-beta domain-containing genes and genome-wide, however, used the published Aqu1 gene models (Srivastava et al. 2010) available on the *A. queenslandica* Ensembl Metazoa genome browser (Kersey et al. 2014), as genome-wide manual analysis with the newer gene models was not practical. Intron phase calculations from the *AqAF* genes were based on the Aqu2.1 dataset; differences between phase values for *AqAF* models in the Aqu1 and Aqu2.1 sets are relatively minor and unlikely to impact on the interpretation of the results in this study. Sequence alignments from *A. queenslandica* Calx-beta domain-containing genes are also based on Aqu1 sequences.

2.3.3 Generation of the Wreath domain HMM model

A multiple sequence alignment previously generated by M. Gauthier (Gauthier 2009), showing the MAFp3 region from *C. prolifera* and homologous regions from AqAFC (*A. queenslandica*) and SdSLIP (*S. domuncula)*, was used to generate a profile hidden Markov model (HMM; Appendix 2.2), which I have termed the Wreath domain. The model was generated with the hmmbuild tool and verified using hmmsearch, using default parameters. Both of these tools are available in the HMMER 3.0 software package (Eddy 1998). Domain hits were counted if hmmsearch returned an expect (e)-value equal or less than 10^{-4} for the region in question. Tests of the new model resulted in the identification of Wreath domains in AqAFA through AqAFE, and in novel sequences from other sponge species. Wreath domains were not detected in AqAFF, non-AF predicted proteins from *A. queenslandica*, or in any non-sponge species. These results were confirmed by BLASTp searches, with the MAFp3 region used as a search query to probe the same datasets used for HMM analyses.

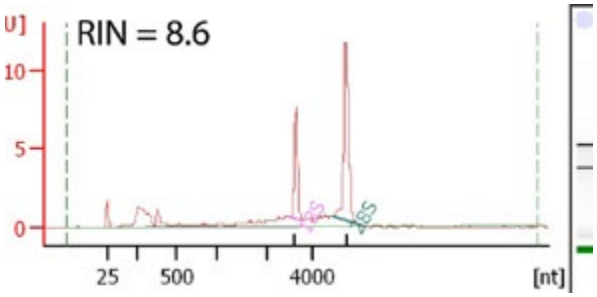


Figure 2.3 Quantitative analysis of *C. prolifera* RNA quality
A Bioanalyser 2100 trace for DNase-treated *C. prolifera* RNA, prior to transcriptome sequencing. RIN – RNA integrity number.

2.3.4 Calx-beta, VWA and VWD phylogenetic domain distribution

HMM models for Calx-beta (Pfam PF03160), VWA (Pfam PF00092) and VWD (Pfam PF00094) domains were used to probe the translated gene models of multiple species with a wide taxonomic distribution (Appendix 2.3) using hmmsearch as per Chapter 2.3.3.

2.3.5 Calx-beta domain multiple sequence alignments

The peptide sequences of all Calx-beta domain-containing genes from *A. queenslandica* and *Nematostella vectensis* were downloaded from Ensembl Metazoa (Kersey et al. 2014) using the BioMart tool (Kinsella et al. 2011). The HMM Search function of DoMosaics (Moore et al. 2014) was used to identify conserved domain types, using the HMMER 3.0 hmmsearch and hmmplan binary files (Eddy 1998) and all Pfam-A domain profiles (current version as of 31.04.14) (Finn et al. 2006), and run with default parameters. Resulting Calx-beta domain sequences were exported. For each gene encoding four or more Calx-beta domains ($n_{Aq} = 8$, $n_{Nv} = 10$), multiple sequence alignments were generated from all Calx-beta domain sequences, by running 100 iterations of the MUSCLE software (Edgar 2004) built into Geneious Pro 5.0.2 with default parameters. Minor alignment alterations were performed manually in Geneious to remove mostly-gapped positions. Sequence logos were generated in WebLogo 3.4 (Crooks et al. 2004), with custom colours used to distinguish polar, non-polar, acidic, and basic amino acids.

2.3.6 Sequencing data used for AF identification

a. Ephydatia muelleri and Sycon ciliatum

Ephydatia muelleri translated mRNA sequences (T-PEP) were downloaded from Compagen (http://www.compagen.org) (Hemmrich and Bosch 2008). Translated peptide sequences from the then-unpublished *S. ciliatum* genome (Fortunato et al. 2015) were provided by M. Adamska and M. Adamski (personal communication).

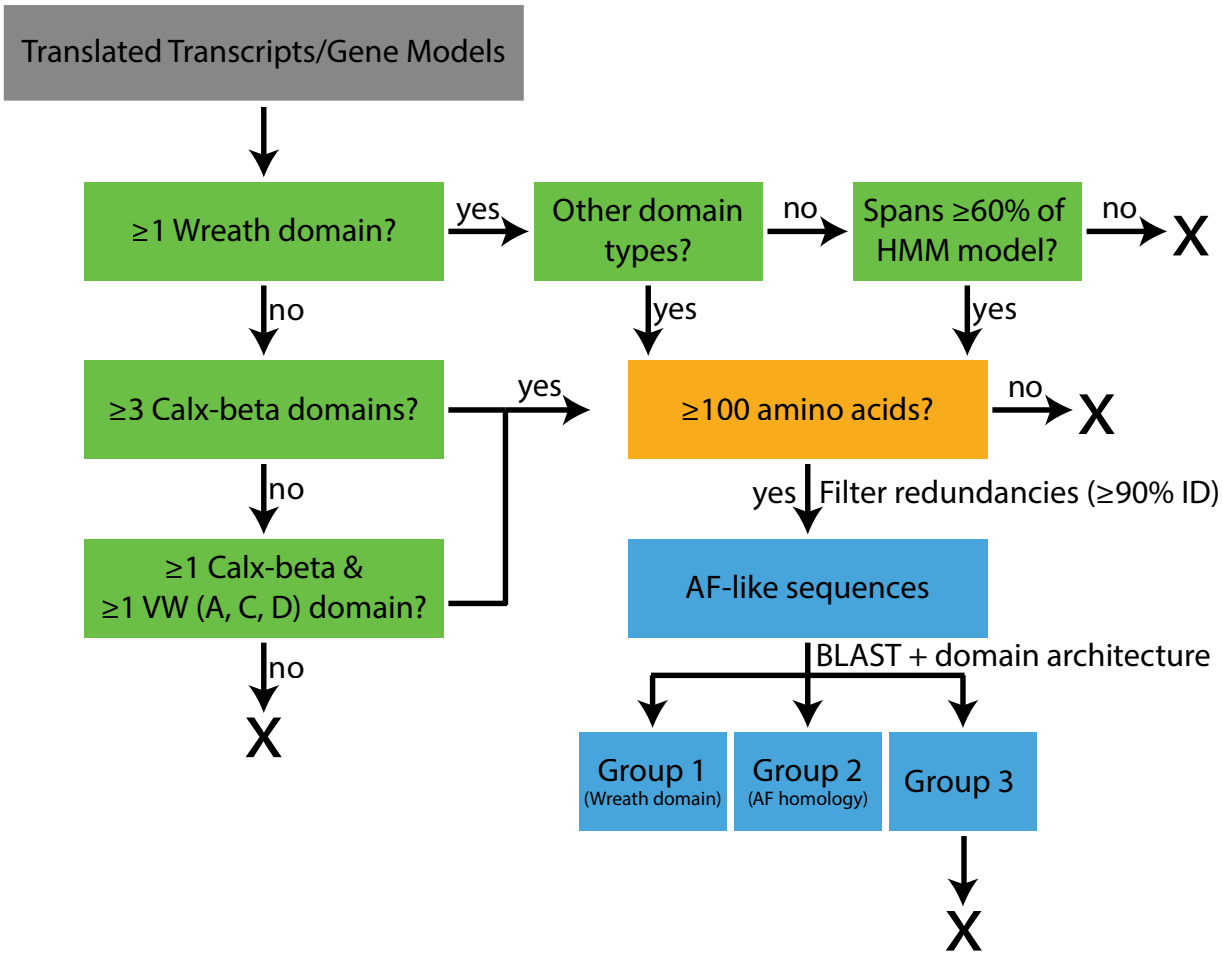


Figure 2.4 Methodology for AF candidate sequence identification
Flowchart depicting the filtering process to isolate AF-like and candidate AF sequences from whole-genome or -transcriptome datasets. Sequences possessing Wreath, Calx-beta, VWA, or VWD domains were identified by searching sequence datasets with HMM profiles. Sequences were eliminated (X) if they encoded only a Wreath domain and this domain did not cover at least 60% of the HMM model. Short or redundant sequences were also removed. The resulting list was divided into three groups, based on domain architecture and sequence similarity. Group 1 sequences possess a Wreath domain, with or without other domain types. Group 2 sequences have a top BLAST hit to a known AF sequence from *A. queenslandica*, *C. prolifera* or *S. domuncula*, but do not possess a Wreath domain. Group 3 sequences represent all other sequences identified, and were not considered AF candidates for the purposes of this analysis.

b. C. prolifera

A high-quality *C. prolifera* sample (Figure 2.2) provided by X. Fernandez-Busquets was treated with Deoxyribonuclease I (Amplification Grade; Invitrogen) according to manufacturer's directions, in order to remove contaminating genomic DNA. Sample quality was checked using an Agilent Bioanalyser 2100 (Figure 2.3). The sample was submitted to Macrogen Ltd. (Seoul, Korea) for transcriptome sequencing with a 100 base pair (bp), paired-end, stranded Illumina HiSeq 2000 protocol.

Transcriptome preparation and *de novo* transcript assembly was performed by S. Fernandez Valverde. Briefly, overall sequencing quality was determined using FastQC v0.10.1 (http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc), run with default parameters. Raw sequencing reads were quality filtered using Trimmomatic v0.20 (Bolger et al. 2014). The first 7 bp of each read were cropped, and reads were trimmed if the average quality within a window of 4 bp was below 15. Unpaired or short (<60 bp) reads were discarded. Remaining reads were assembled *de novo* using Trinity v2013-08-14 (Grabherr et al. 2011) using default parameters except for a lower transcript size of 200 bp. The longest open reading frame (ORF) between stop codons was determined for each assembled transcript using the program getorf from the EMBOSS 6.5.7 software package (Rice et al. 2000).

c. Oscarella carmela

The *O. carmela* whole genome assembly dataset (http://www.compagen.org) (Nichols et al. 2012) was submitted to Augustus 2.6.1 (Stanke et al. 2006), in order to generate new gene models for this species. Augustus was run with the *A. queenslandica* training set, with settings singlestrand=true, alternatives-from-evidence=true and uniqueGeneId=true.

d. Other species

Aphrocallistes vastus, Chondrilla nucula, Corticium candelabrum, Crella elegans, Ircinia fasciculata, Petrosia ficiformis, Spongilla lacustrus, Pseudospongosorites suberitoides and *Sycon coactum* (Riesgo et al. 2012) nucleotide datasets were converted to predicted ORFs as described for *C. prolifera*. For *C. elegans*, sequences from all three available developmental stages were pooled prior to analysis.

2.3.7 Identification of AF-like sponge sequences

Sequences from the translated transcriptomes and genomes from each species listed in Section 2.3.6 were filtered to generate a list of AF-like sequences (Figure 2.4). Sequences equipped with Calx-beta, VWA, VWD or Wreath domains (maximum e-value 10^{-4}) were identified using HMM search methods as described in Section 2.3.4. Sequences were considered AF-like if they were greater than 100 amino acids (aa) in length and possessed either (a) three or more Calx-beta domains, (b) one or more Wreath domain (which, for sequences not predicted to encode additional domain types or

sequence features such as transmembrane domains or signal peptides, had to span 60% or more of the Wreath HMM model), or (c) one or more Calx-beta domain plus one or more VWA or VWD domain. To remove redundancies, sequences within each species were clustered into groups sharing at least 90% sequence identity, using the default parameters of the cd-hit tool (Li and Godzik 2006), available via the CD-HIT Suite server (Huang et al. 2010). Only the representative sequence from each cluster (as determined by cd-hit; equivalent to the longest sequence) was passed through for further analysis.

AF-like sequences were further sorted to identify putative AF candidates. Overall domain architecture for each sequence was determined using DoMosaics (as per Section 2.3.5), and signal peptides and transmembrane domains were predicted using Phobius (Käll et al. 2004). The Personal BLAST Navigator (PLAN) tool (He et al. 2007) was used to perform a batch BLASTP search for the top hit (maximum e-value 10^{-4}) in the NCBI 20121015 NR database. Sequences were assigned to one of three groups based on the domain and BLAST (basic local alignment search tool) results (see Figure 2.4 for group assignment criteria).

2.3.8 Calculation of intron phase distribution frequencies

Genome-wide intron phase frequencies were determined for all protein-coding genes in each of *A. queenslandica, Helobdella robusta, Lottia gigantea, N. vectensis* and *Trichoplax adhaerens*. Intron phase values were retrieved from the Ensembl Metazoa (Kersey et al. 2014) genome browsers for each species, using the BioMart data mining tool (Kinsella et al. 2011). BioMart automatically assigns an intron phase value to each exon (including the first exon of a gene), based on the phase of the previous intron. As the first exon of a gene is by definition never preceded by an intron, the phase value incorrectly associated with the first exon of every gene was deleted. Any negative values (again, a quirk of the BioMart output) were also deleted. Within each species, all remaining intron phase values were summed and used to calculate genome-wide frequencies of each intron phase, the standard deviations of the mean, and significance values, following the methods described by Fedorov et al. (1992; 1998). For each species, I determined whether the observed frequency of each intron phase (P_{obs}) was significantly different from a random distribution (P_{rand}) of 0.33 per phase; values were considered statistically significant if $|P_{rand} - P_{obs}| > 3\sigma$ (Fedorov et al. 1998).

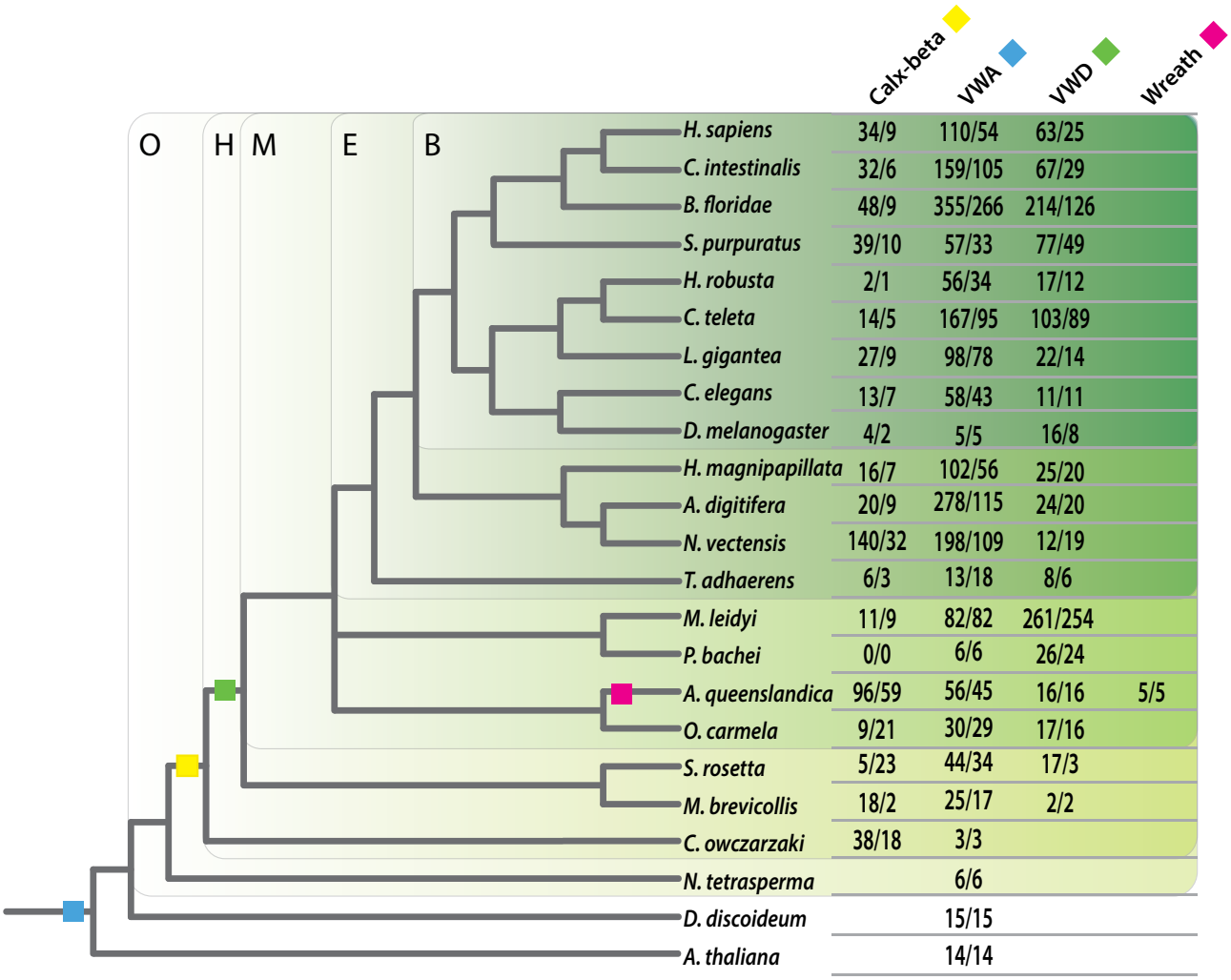


Figure 2.5 Phylogenetic distribution of Calx-beta, VWA, VWD and Wreath domains

The table (right) gives Calx-beta, Von VWA, VWD and Wreath domain and domain-encoding gene counts for a selection of eukaryote model species (for the full data table, see Appendix 3.3). Counts are written in the form 'domain count/gene count'. Putative evolutionary origins of each domain type are mapped to the phylogenetic tree as coloured squares (*left*); colours are given above each domain name (*right*). Green boxes separate the tree into the main phylogenetic groupings: Bilateria (B), Eumetazoa (E), Metazoa (M), Holozoa (H) and Opisthokonta (O).

Genome-wide phase data are not readily available for species without an Ensembl Metazoa genome browser. To allow statistical comparisons for such species in later analyses, I created an additional dataset ("Reference Set") comprising the combined counts of phase 0, 1 and 2 introns from the five species analysed above. This dataset was analysed as above to determine whether the frequencies observed in this reference set (P_{ref}) were significantly different from a random phase distribution ($|P_{rand} - P_{ref}| > 3\sigma$). To test the representativeness of the Reference Set, I also compared the observed

intron phase frequencies in each contributing species to the overall Reference values ($|P_{ref} - P_{obs}| > 3\sigma$); no significant difference was found for any species (data not shown), suggesting that this dataset is sufficiently representative to apply to other basal metazoan species.

The above analyses were repeated for datasets of Calx-beta domain-containing genes from each of eleven species. For *A. queenslandica*, *H. robusta, L. gigantea, N. vectensis* and *T. adhaerens*, phase data was again gathered from each species' Ensembl Metazoa genome browser, with the BioMart search filtered to include only those genes annotated with one or more Pfam Calx-beta domains (Pfam:PF03160). For *Branchiostoma floridae, Capitella teleta, Hydra magnipapillata* and *Monosiga brevicollis*, Calx-beta domain-containing genes were isolated via HMM-based searches for Calx-beta domains as described in Section 2.3.4. Phase values for each intron in these genes were determined manually. Phase distributions were again analysed as above, comparing observed frequencies to those seen in the Reference Set ($|P_{ref} - P_{calx}| > 3\sigma$) A final dataset, comprising only the six AqAF genes (based on the Aqu2.1 gene models), was also analysed. I tested whether the frequencies in the AqAF dataset (P_{AF}) differed significantly from the *A. queenslandica* Calx-beta-encoding gene set ($|P_{calx} - P_{AF}| > 3\sigma$).

2.4 Results

2.4.1 The *A. queenslandica* AFs encode a novel protein domain

The *C. prolifera* MAFp3 protein plays a key functional role in AF structure and self-adhesion by forming the central ring of the core AF sunburst structure (Jarchow et al. 2000). The ring structure of circular AFs is equivalent to the rod-like backbone of linear AFs (Henkart et al. 1973) . Regions exhibiting MAFp3 sequence similarity are also present in SdSLIP from *S. domuncula* (Wiens et al. 2005; Gauthier 2009) and in all AqAFs except AqAFF (this work; Gauthier (2009)). Protein domains can be defined as protein structural units that form an independent fold within a protein, and mediate a particular protein function (Richardson 1981). Considering the demonstrated functional importance, structural independence and multi-species distribution of this region, I propose that MAFp3 and homologous sequences be considered representatives of a novel protein domain, becoming the fourth domain type of the AqAFs. I suggest the name 'Wreath domain' due to its role in *C. prolifera* AF central ring formation (Jarchow et al. 2000). A multiple sequence alignment of the Wreath region from MAFp3, SdSLIP and AqAFC (Gauthier 2009) was used to generate an HMM for this new putative

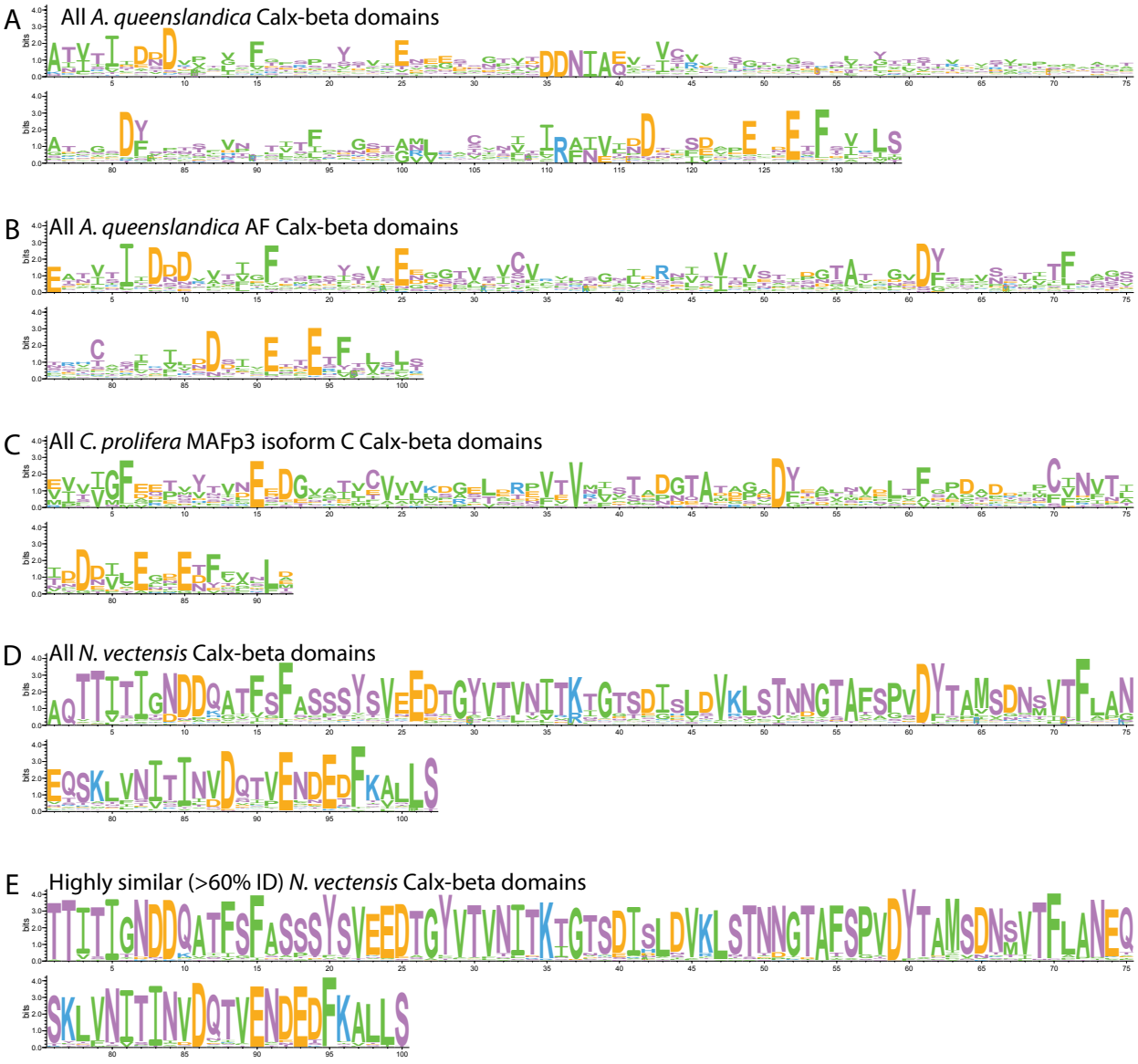


Figure 2.6 Sequence homology within selected Calx-beta domain-containing proteins

Sequence logos of all Calx-beta domains from (A) all *A. queenslandica* proteins possessing four or more Calx-beta domains, (B) all *A. queenslandica* AFs, (C) the *C. prolifera* AF MAFp3 isoform C, (D) all *N. vectensis* proteins possessing four or more Calx-beta domains and (E) all *N. vectensis* proteins possessing four or more domains that show an average amino acid sequence identity of 60% or greater between domains. Individual logos for all such proteins containing four or more Calx-beta domains are shown in Appendix 2.4. Nonpolar amino acids – green, polar – purple, acidic – orange, basic – blue.

domain (Appendix 2.2). HMM searches with this model identified a single Wreath domain in all AqAF sequences except AqAFF. The Wreath domain was not identified in any non-AF *A. queenslandica* genes, or in any analysed non-sponge species (Figure 2.5, Appendix 2.4).

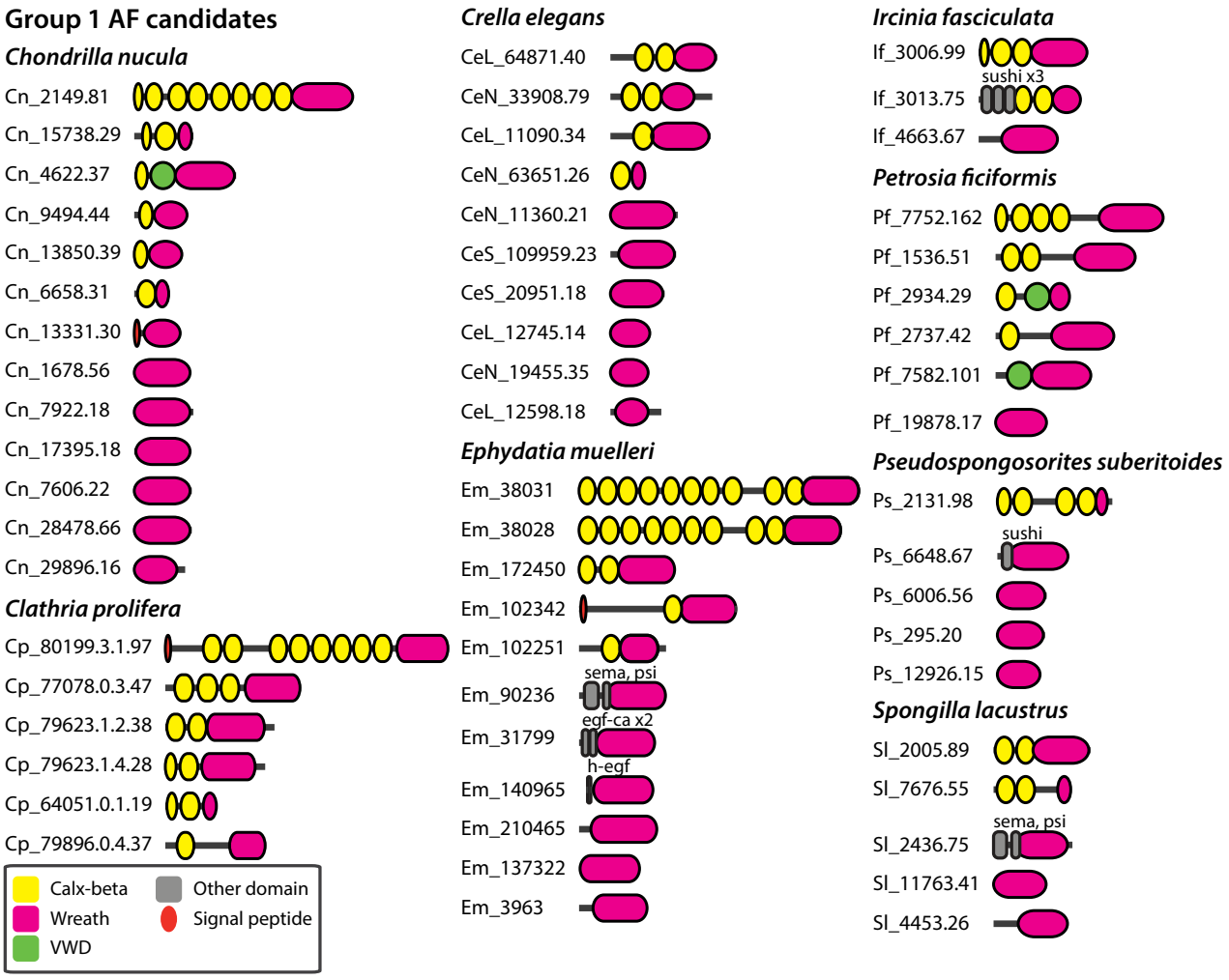


Figure 2.7 Domain architecture of Group 1 AF candidates
The domain architectures for all AF-like sequences equipped with a Wreath domain. Domains and other sequence features are represented as coloured shapes. Domain types not present in known AF or AF-related sequences from *A. queenslandica, C. prolifera* or *S. domuncula* are depicted in grey and named above each domain. Sequence names describe the species, accession number of the original sequence, and the number of the longest translated ORF for that sequence as the last digit (e.g. Cp_80199.3.1.97 represents ORF 97 from sequence 80199.3.1, in *C. prolifera*). All sequences and features are drawn to scale.

2.4.2 Phylogenetic distribution of domain types present in AqAFs

To better understand the evolution of the AqAF domain building blocks, I surveyed the translated genomes of a phylogenetically widely-distributed set of species for genes encoding Calx-beta, VWA, VWD and Wreath domains (Figure 2.5, Appendix 2.4). Calx-beta domains are present in all holozoan species analysed, but not in any fungi, amoebozoa, protist, plant or archaea species. Calx-beta domains were, however, identified in a number of bacterial species. All but two of these species were isolated from marine environments (Schlesner et al. 2004; Sohn et al. 2004; Schäfer et al. 2005; Oh et al. 2010;

Figure 2.8 Domain architecture of Group 2 AF candidates
The domain architectures of all AF-like sequences exhibiting highest sequence similarity to a known AF or AF-related sequence from *A. queenslandica, C. prolifera* or *S. domuncula* (and lacking a Wreath domain) are depicted. Domain types not present in known AF or AF-related sequences are depicted in grey and named above each domain. Sequence names describe the species, accession number of the original sequence, and the number of the longest translated ORF for that sequence as the last digit (E.g. Av_39733.0.1.93) represents ORF 93 from sequence 39733.0.1, in *Aphrocallistes vastus*). All sequences and features are drawn to scale.

2011). *Oscillochloris trichoides* was isolated from a warm hydrogen sulphide spring (Keppen et al. 1993), while *Pedobacter saltans* is a soil bacterium (Steyn et al. 1998). *A. queenslandica* encodes a large number of Calx-beta domains (n = 96), the second highest number from any species tested behind *N. vectensis* (n = 140). The large number of Calx-beta domains in both of these species appear to be the result of separate lineage-specific expansions; Calx-beta domain counts in other analysed species from the same phyla are comparatively low (*Acropora digitifera*, [n = 9] and *Hydra magnipapillata* [n = 7], and *O. carmela* [n = 9], for cnidarians and sponges respectively). VWA domains are evolutionarily

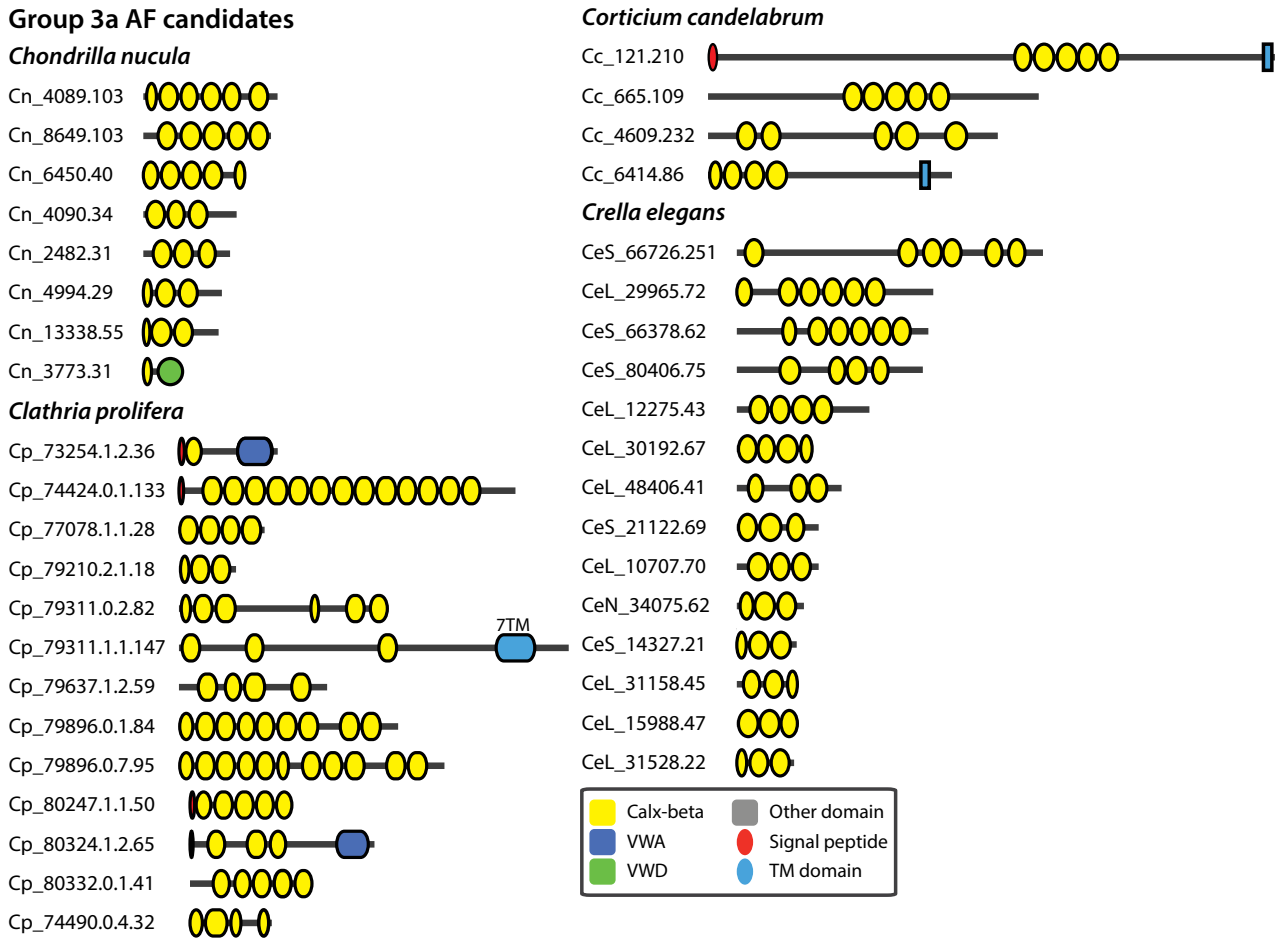


Figure 2.9 Domain architecture of Group 3 AF candidates

(Part 1 of 2)
The domain architectures of AF-like sequences not fulfilling criteria for Groups 1 or 2 are shown. Group 3a sequences are similar to known AF sequences but lack identifying features of likely AFs. Group 3b sequences possess other domain types marking them as likely members of other gene families. Domain types not present in known AF or AF-related sequences are depicted in grey and named above each domain. Sequence names describe the species, accession number of the original sequence, and the number of the longest translated ORF for that sequence as the last digit (E.g. Cp_73254.1.2.36 represents ORF 36 from sequence 73254.1.2, in *C. prolifera*). All sequences and features are drawn to scale.

ancient, being identified in all species tested with the exceptions of the yeast species *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe,* the archaea *Haladaptatus pauchihalophilus* and the bacterium *Pedobacter saltans* (Appendix 2.4). In contrast, VWD domains are comparatively younger, being identified in metazoans and choanoflagellates, but not the fellow holozoan *Capsaspora owczarzaki*. Intriguingly, the VWD domains were also found in the excavate amoeba *Naegleria gruberi*. Wreath domains were not identified in any non-sponge species tested.

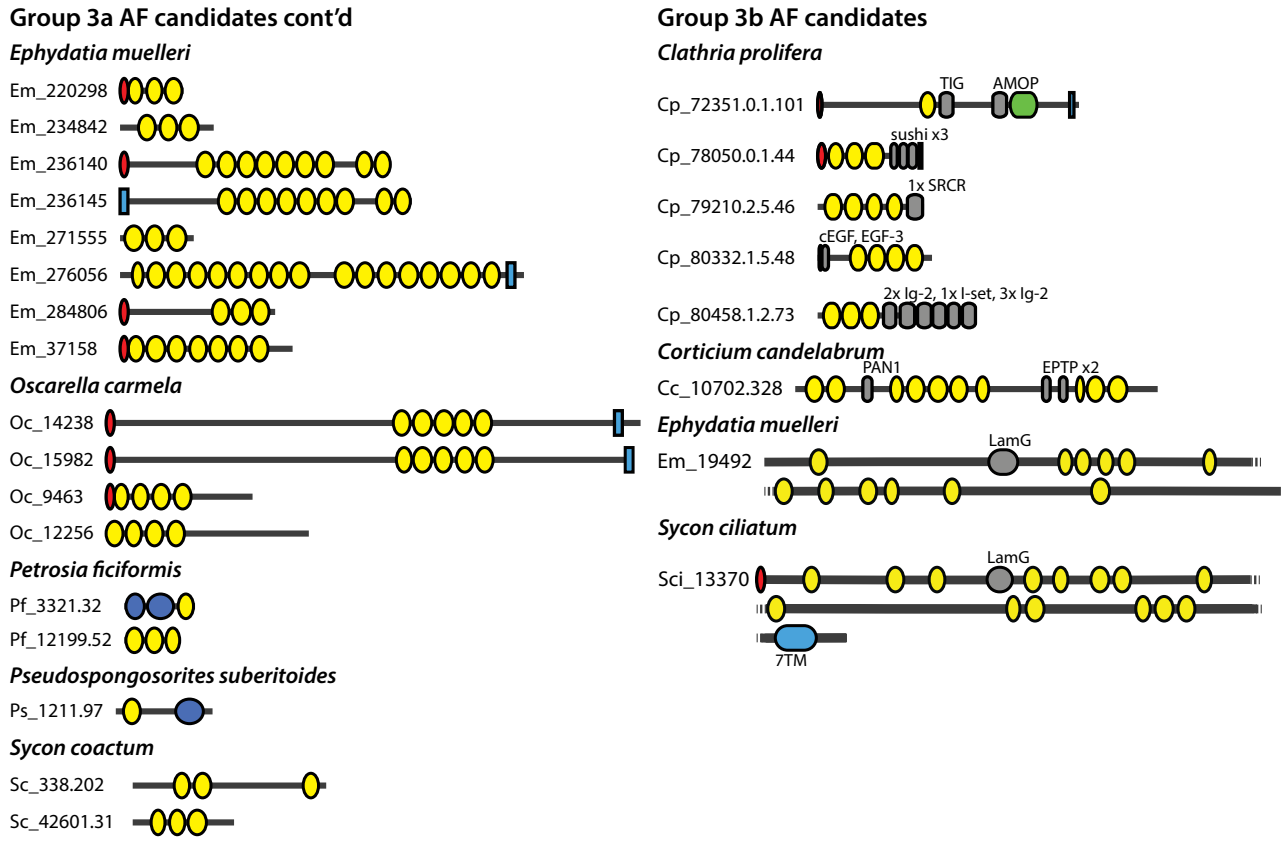


Figure 2.9 Domain architecture of Group 3 AF candidates

(Part 2 of 2)

2.4.3 AqAF domain sequence alignments

A. queenslandica Calx-beta domains exhibit a low level of sequence identity to other Calx-beta domains within the same gene (average 29% identity; Figure 2.6, Appendix 2.5). Multiple sequence alignments from genes encoding four or more Calx-beta domains show that only a few key residues are conserved between domains (Figure 2.6). A large proportion of these conserved residues are the amino acids aspartic acid (D) and glutamic acid (E), and correspond to those residues identified by Hilge et al. (2006) as key Ca^{2+}-binding residues. Low sequence identities are also observed between the Calx-beta domains of the AqF subset of this gene list (average 26% identity), or of the *C. prolifera* sequence MAFp3 isoform C (Figure 2.6). Although the *A. queenslandica* and *N. vectensis* genomes both encode unusually high numbers of Calx-beta domains relative to other analysed species (Section 2.4.2), the *N. vectensis* Calx-beta domains are not as highly diversified as in *A. queenslandica* (Figure 2.6). Only three *N. vectensis* Calx-beta domain-containing genes encode Calx-beta domains exhibiting

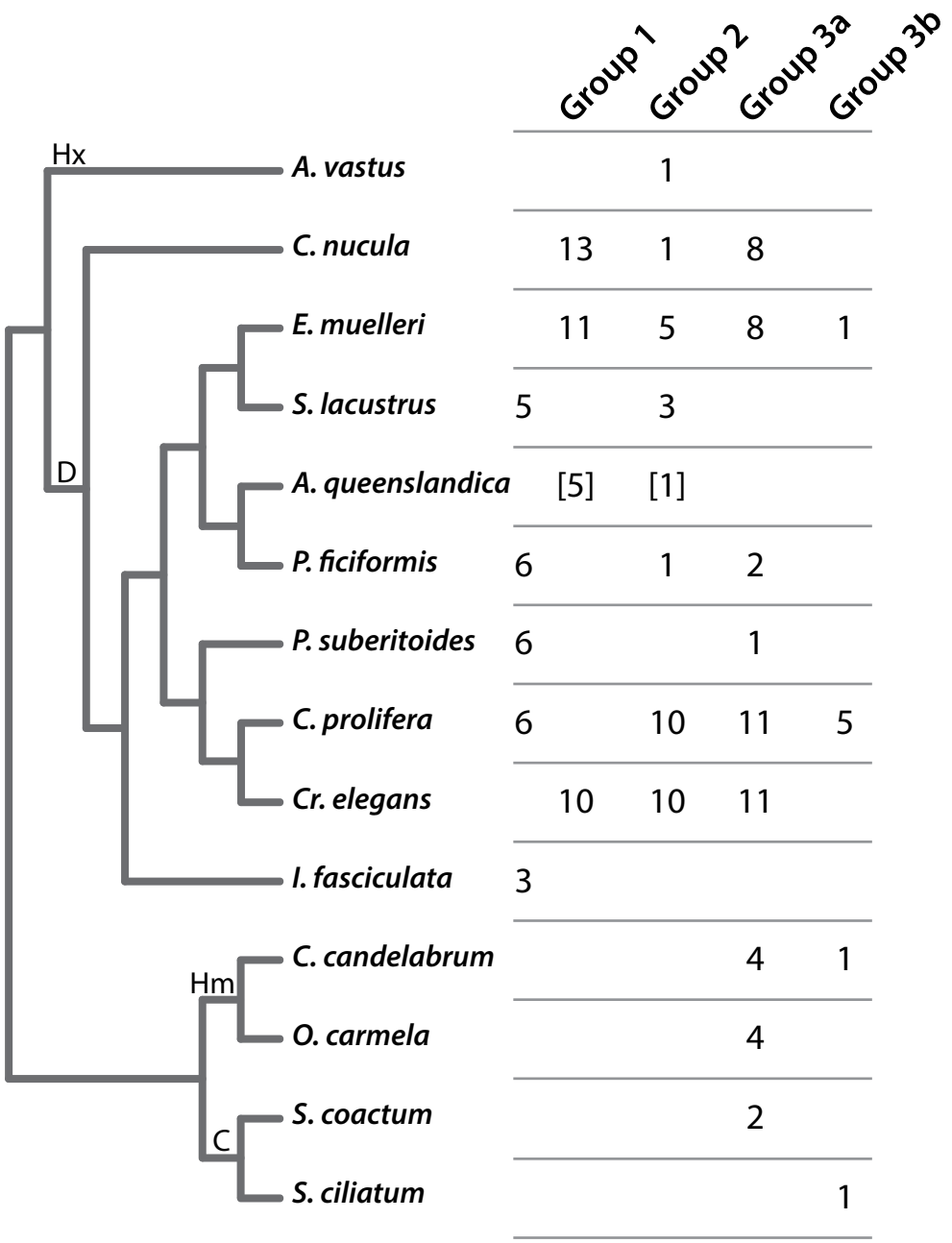


Figure 2.10 Phylogenetic distribution of Group 1 and 2 AF candidates, and Group 3 AF-like sequences

The phylogenetic relationships between analysed sponge species is depicted on the left (Thacker et al. 2013). The table gives the number of sequences per species assigned to Groups 1 (i.e. possessing a Wreath domain), 2 (i.e. having homology to known AFs or AF-related sequences from *A. queenslandica, C. prolifera* or *S. domuncula*), 3a (i.e. sequences equipped with Calx-beta, VWA or VWD domains only, with no sequence homology to known AFs) and 3b (i.e. sequences that appear to be members of other gene families). The counts given for *A. queenslandica* refer to the *AqAF* genes encoded in the genome for this species. Letters refer to sponge classes – Calcarea (C), Demospongia (D), Homoscleromorpha (Hm), Hexactinellida (Hx).

low sequence identity to one another (average 64% identity). The domains in the remaining seven analysed *N. vectensis* genes share a high level of sequence identity (average 81%) both between and within genes. These seven genes are architecturally diverse, ranging in Calx-beta domain count from four to forty domains (data not shown). No analysed *N. vectensis* genes bear significant sequence similarity to any analysed *A. queenslandica* genes (data not shown)

2.4.4 Search criteria for AF candidate identification

Recent advances in sequencing technology have led to the availability of genome or transcriptome data from a large number of sponge species distributed across the phylum. I searched for candidate AF sequences in the genomes or transcriptomes of thirteen sponge species (*Aphrocallistes vastus, Chondrilla nucula, C. prolifera, Corticium candelabrum, Crella elegans, Ephydatia muelleri, Ircinia fasciculata, O. carmela* (genome), *Petrosia ficiformis, Pseudospongosorites suberitoides, Spongilla lacustrus, Sycon ciliatum* (genome) and *Sycon coactum*). I used known features of the *A. queenslandica* and *C. prolifera* AFs, plus the *S. domuncula* AF-related protein SdSLIP, to develop a sequence filtering workflow, based on the domain architecture and sequence similarity of each analysed sequence (Figure 2.4). Sequences were considered for further study if they possessed (a) three or more Calx-beta domains, (b) a Wreath domain or (c) VWA or VWD domain/s coupled to one or more Calx-beta or Wreath domains (Figure 2.4). Sushi domains, as seen in the candidate core *G. cydonium* AF, *GEOCY_AF*, were not included as search criteria as this form has only been observed in one species and has not been well characterised.

The presence or absence of signal peptides or transmembrane domains, overall protein domain architecture and the best BLAST hit were determined for each AF-like sequence (n = 155; Appendix 2.6). Sequences were divided into three groups based on the latter two pieces of information. Group 1 sequences (n = 59; Figure 2.7) are all equipped with a Wreath domain, regardless of the overall domain architecture of the encoded protein. In *C. prolifera*, the Wreath domain encodes the AF ring subunit that conveys AF assembly functionality. Homologous regions have not been identified outside the sponge AFs and the AF-related SdSLIP protein from *S. domuncula*, meaning that any gene possessing this domain is likely to be an AF. Group 2 sequences (n = 32; Figure 2.8) do not encode a Wreath domain, but have a top BLAST hit to an AF or AF-related sequence from *A. queenslandica*, *C. prolifera* or *S. domuncula*. Finally, Group 3 sequences (n = 64; Figure 2.9) comprise all remaining AF-like sequences.

Group 3 contains diverse sequences that fulfil the filtration criteria outlined above, but that do not have additional features or properties identifying them as likely AFs (Group 3a; n = 56) or that are equipped with other domain types identifying them as probable members of other protein families (Group 3b; n = 8) (Appendix 2.6). For the purposes of this preliminary study, I considered Group 1 and 2 members to be candidate AF sequences, and Group 3 sequences to be AF-like but probably (though not definitely) not true AFs.

2.4.5 AF candidate sequences from thirteen sponge species

a. Group 1 - Wreath domain-equipped sequences

AF candidates belonging to Group 1 (Figure 2.7) were identified in *C. nucula* (n = 13), *C. prolifera* (n = 6), *C. elegans* (n = 10, all stages combined), *E. muelleri* (n = 11), *I. fasciculata* (n = 3), *P. ficiformis* (n = 6), *P. suberitoides* (n = 5) and *S. lacustrus* (n = 5); that is, all demosponge species (and no others) analysed were found to possess multiple AF candidates equipped with a Wreath domain (Figure 2.7; Figure 2.10).

As in known *A. queenslandica* and *C. prolifera* AF sequences, all examined demosponges encode transcripts encoding Calx-beta and Wreath domains together. Between one and ten Calx-beta domains were found in each of these Calx-beta + Wreath sequences. Three such sequences also encode a signal peptide (*C. nucula* Cn_13331.30, *C. prolifera* Cp_80199.3.1.97 and *E. muelleri* Em_102342), indicating that the 5' end of these sequences is intact and that their encoded protein products are secreted. Similar to the *A. queenslandica* AFs, some *C. nucula* (Cn_4622.37) and *P. ficiformis* (Pf_2934.29 and Pf_7582.101) AF candidates possess VWD domains coupled to their Wreath domains.

Most analysed demosponge species also encode transcripts comprised of a single Wreath domain. All but one of these sequences (*C. nucula* Cn_13331.30) lack signal peptides, so it is currently unknown whether these represent true biological transcripts or truncated sequence fragments. Several Group 1 sequences also exhibit Wreath domains coupled to novel domain types not seen in known AFs. The two closely-related freshwater haploscleromorph species *E. muelleri* (Em_90236) and *S. lacustrus* (Sl_2436.75) both encode a protein equipped with one copy each of Sema (PF01403), PSI (PF01437) and Wreath domains. *E. muelleri* also encodes two proteins (Em_31799 and Em_140965) containing

EGF-related domain types (Calcium-binding EGF domain, PF07645; human growth factor-like EGF domain, PF12661). Finally, Sushi domains, as previously documented in the *G. cydonium* candidate core AF *GEOCY_AF*, are present in one sequence each from *I. fasciculata* (If_3013.75, 3 copies) and *P. suberitoides* (Ps_6648.67, 1 copy).

Three *C. prolifera* Group 1 sequences are highly similar to previously reported MAFp3 isoforms. Cp_79623.1.2.38 exhibits 99% identity to both MAFp3 isoforms B and C, Cp_79623.1.4.28 is 89% identical to MAFp3 isoform D, and Cp_64051.0.1.19 shares 99% identity with MAFp3 isoform E. As the MAFp3 isoforms are similar to one another, the new *C. prolifera* sequences also share high sequence identity with other isoforms and with each other. These new sequences are shorter than those identified in previous studies and probably do not represent full-length sequences. The remaining *C. prolifera* Group 1 sequences, while somewhat similar to characterised MAFp3 isoforms, appear to represent novel sequences.

b. Group 2 - Sequences exhibiting AF sequence homology

Group 2 candidates (Figure 2.8) - that is, sequences lacking a Wreath domain but exhibiting top BLAST hits to a known *A. queenslandica* or *C. prolifera* AF or to the AF-related *S. domuncula* SdSLIP, were identified in *A. vastus* (n = 1), *C. nucula* (n = 1), *C. prolifera* (n = 8), *C. elegans* (n = 13, all stages combined), *E. muelleri* (n = 5), *P. ficiformis* (n = 1) and *S. lacustrus* (n = 3). All sequences possess Calx-beta domains in numbers ranging from 3 to 19.

As in Group 1 sequences, only a small number of Group 2 transcripts are predicted to encode a signal peptide; these are present in two sequences from *C. prolifera* (Cp_68734.0.1.105 and Cp_77978.0.7.38) and one from *E. muelleri* (Em_225017). It is unclear whether the remaining sequences simply lack signal peptides or if the sequences are not complete. Transmembrane domains are predicted within six sequences (*C. prolifera* Cp_79465.0.4.99, *C. elegans* CeL_65310.42, *E. muelleri* Em_225017 and Em_133978, *S. lacustrus* Sl_3459.106 and Sl_13008.32). In all cases, the transmembrane domains are situated towards the C-terminal, relative to the other domains. Several transmembrane domain-equipped sequences encode long stretches of Calx-beta domains (for example, having 6 and 15 Calx-beta domains in the two *E. muelleri* sequences). Proteins with similar organisations are not observed in

the *A. queenslandica* genome. While *A. queenslandica* does encode proteins equipped with Calx-beta and transmembrane domains together, these all appear to display either a small number of Calx-beta domains (1-2 per sequence), or domain architectures known to be well-conserved in non-AF protein family members (data not shown).

A small number of Group 2 sequences (n = 3) are equipped with domain types novel to AF sequences. Two *C. prolifera* sequences (Cp_79465.0.4.99 and Cp_77978.0.7.38) and one *C. elegans* sequence (CeS_76241.66) are equipped with multiple domains belonging to the immunoglobulin superfamily (IgSF; Ig2, PF13895; I-set, PF07679; V-set PF07686), in addition to Calx-beta domains.

c. Group 3 - additional AF-like sequences

Remaining AF-like sequences that did not fulfil criteria for Groups 1 or 2 were partitioned into Subgroups 3a and 3b (Figure 2.9) depending on whether their architecture was generally similar to other AFs (Group 3a) or if the sequence encoded other domain types, suggesting that the sequences are probable members of other protein families (Group 3b).

Group 3a sequences were identified in *C. nucula* (n = 8), *C. prolifera* (n = 13), *C. candelabrum* (n = 4), *C. elegans* (n = 14, all stages combined), *E. muelleri* (n = 8), *O. carmela* (n = 4), *P. ficiformis* (n = 2), *P. suberitoides* (n = 1) and *S. coactum* (n = 2). Group 3a members are mostly comprised of Calx-beta domains, in numbers ranging from one to seventeen domains. A small number of sequences encode VWA (*C. prolifera* Cp_73254.1.2.36 and Cp_80324.1.2.65, *P. ficiformis* Pf_3321.32, and *P. suberitoides* Ps_1211.97) or VWD (*C. nucula* Cn_3773.31) domains. Signal peptides are present in sequences from *C. prolifera* (Cp_73254.1.2.36, Cp_74424.0.1.133, Cp_80247.1.1.50, Cp_80324.1.2.65), *C. candelabrum* (Cc_121.210), *E. muelleri* (Em_220298, Em_236140, Em_284806 and Em_37158) and *O. carmela* (Oc_14238, Oc_15982 and Oc_9463). Transmembrane domains were identified in *C. candelabrum* (Cc_121.210 and Cc_6414.86), *E. muelleri* (Em_236145 and Em_276056) and *O. carmela* (Oc_14238 and Oc_15982). A seven transmembrane receptor is also predicted to be present in a single *C. prolifera* sequence (Cp_79311.1.1.147).

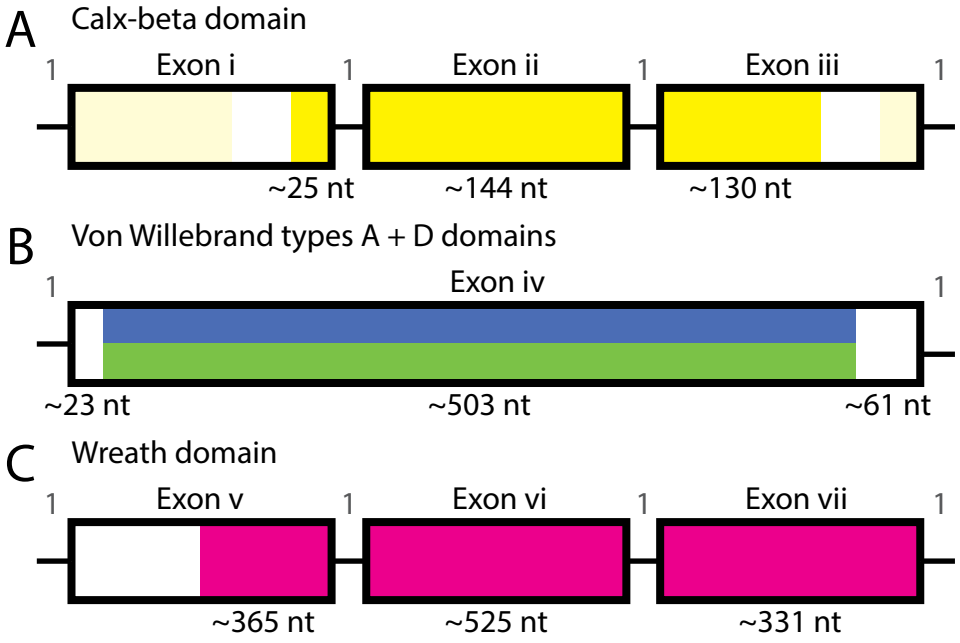


Figure 2.11 Generalised exon organisation of Calx-beta, VWA and VWD, and Wreath domains

The majority of *A. queenslandica* Calx-beta, VWA/VWD and Wreath domains are encoded by exons that are organised in a consistent way between domains within the AqAFs. (A) Most Calx-beta domains are encoded by a three-exon domain module, covering a small portion (average 25 nucleotides; nt) of exon i, the entirety (average 144 nt) of exon ii and about two-thirds of exon iii (average 130 nt). This pattern then repeats, commencing at the end of exon iii. (B) The modular pattern for VWA and VWD domains is similar to one another. Here, a single exon encodes a single domain, with a short spacer region at the start (average 23 nt) and end (average 61 nt) of each exon. (C) The Wreath domains from AqAFA, C, D and E are encoded by the final three exons of each gene. The Wreath domain region covered by exon A is variable in size, spanning 148 to 497 nt in different sequences; the regions covered by exons B (range of 66 nt difference between sequences) and C (range of 21 nt difference between sequences) is more consistent between sequences. Precise values are provided in Appendix 5. Grey '1' refers to the phase of the introns flanking each exon. Exons are not to scale within or between models.

Group 3b sequences were identified in *C. prolifera* (n = 5), *C. candelabrum* (n = 1), *E. muelleri* (n = 1) and *S. ciliatum* (n = 1). These sequences all include domain types novel to known AFs. Signal peptides are present in a small number of sequences (*C. prolifera* Cp_72351.0.1.101 and Cp_78050.0.1.44, *S. ciliatum* Csi_13370). The latter sequence from *S. ciliatum* also encodes a seven transmembrane receptor.

2.4.6 Genomic organisation of *A. queenslandica AFs*

The availability of the complete *A. queenslandica* genome sequence allows for a more in-depth analysis of the AF genes in this species than is possible at present for other sponge species (since probable

Table 2.2 Genome-wide intron phase frequencies of basal holozoan protein-coding genes

GENE SET	PHASE FREQUENCY		
	PHASE 0	PHASE 1	PHASE 2
A. queenslandica (N_g = 563321)	P = 0.46 σ = 0.001* N_i = 62582	P = 0.33 σ = 0.001 N_i = 45260	P = 0.21 σ = 0.001* N_i = 29409
H. robusta (N_g = 23432)	P = 0.45 σ = 0.001* N_i = 52137	P = 0.33 σ = 0.001* N_i = 38091	P = 0.23 σ = 0.001* N_i = 26503
L. gigantea (Ng = 23340)	P = 0.43 σ = 0.001* N_i = 50304	P = 0.35 σ = 0.001* N_i = 40517	P = 0.23 σ = 0.001* N_i = 26144
N. vectensis (N_g = 24773)	P = 0.48 σ = 0.002* N_i = 51029	P = 0.29 σ = 0.001* N_i = 31449	P = 0.23 σ = 0.001* N_i = 24452
T. adhaerens (N_g = 11520)	P = 0.49 σ = 0.002* N_i = 42157	P = 0.27 σ = 0.002* N_i = 23441	P = 0.23 σ = 0.001* N_i = 19846
Reference set	P = 0.46 σ = 0.001* N_i = 258209	P = 0.32 σ = 0.001* N_i = 178758	P = 0.22 σ = 0.001* N_i = 126354

P = phase frequency; σ = standard deviation of the mean; * = statistically significant difference from a random frequency distribution of 0.33 per phase; Ng = total number of genes surveyed; Ni = total number of introns per phase. Reference set values were calculated by adding the intron counts from all species and calculating phase frequency and statistics as per the other samples.

AFs were not identified in *O. carmela* or *S. ciliatum*, the two other sponge species with sequenced genomes). I examined the relationship between AF gene sequences, domain architecture and genomic structure in *A. queenslandica*. Six AF genes are predicted to be present in the *A. queenslandica* genome (Figure 2.1) (Gauthier 2009). *AqAFA* to *AqAFE* each encode a contiguous sequence equivalent to *C. prolifera* MAFp4 + MAFp3, and possess Calx-beta, VWA or VWD, and Wreath domains (except AqAFA which contains neither VWA nor VWD domains, and AqAFF which lacks VWA, VWD and Wreath domains). *AqAFA* to *AqAFE* are large genes (each spanning a genomic region between 9.5 and 17.0 kb in length) with many exons (between 18 and 48 exons per gene; Table 2.1) (Gauthier 2009). In contrast, *AqAFF* is smaller (1.0 kb) and possesses four introns (Table 2.1) (Gauthier 2009). When

Table 2.3 Intron phase frequencies of Calx-beta domain-containing genes from basal holozoan protein-coding genes

GENE SET	PHASE FREQUENCY		
	PHASE 0	PHASE 1	PHASE 2
Reference set	P = 0.46 σ = 0.001 N_i = 258209	P = 0.32 σ = 0.001 N_i = 178758	P = 0.22 σ = 0.001 N_i = 126354
A. queenslandica # Ng = 49	P = 0.15 σ = 0.014^ N_i = 101	P = 0.77 σ = 0.016^ N_i = 510	P = 0.06 σ = 0.010^ N_i = 50
A. queenslandica AFs only Ng = 6	P = 0.006 σ = 0.018^ >
N_i = 1	P = 0.99 σ = 0.018^ >
N_i = 157	P = 0.00 σ = n/a^ >
N_i = 0
A. queenslandica non-*AFs* only Ng = 43	P = 0.19 σ = 0.001^ N_i = 97	P = 0.70 σ = 0.001^ N_i = 353	P = 0.10 σ = 0.001^ N_i = 50
B. floridae Ng = 9	P = 0.43 σ = 0.025 N_i = 167	P = 0.46 σ = 0.025^ N_i = 178	P = 0.11 σ = 0.016^ N_i = 43
C. teleta Ng = 7	P = 0.42 σ = 0.049 N_i = 43	P = 0.37 σ = 0.048 N_i = 38	P = 0.21 σ = 0.040 N_i = 22
H. robusta # Ng = 1	P = 0.70 σ = 0.145 N_i = 7	P = 0.10 σ = 0.095 N_i = 1	P = 0.20 σ = 0.126 N_i = 2
H. magnipapillata Ng = 7	P = 0.41 σ = 0.058 N_i = 30	P = 0.38 σ = 0.057 N_i = 28	P = 0.21 σ = 0.047 N_i = 15
L. gigantea # Ng = 9	P = 0.42 σ = 0.036 N_i = 78	P = 0.43 σ = 0.036 N_i = 80^	P = 0.16 σ = 0.026 N_i = 29
M. brevicollis Ng = 5	P = 0.48 σ = 0.090 N_i = 15	P = 0.39 σ = 0.087 N_i = 12	P = 0.13 σ = 0.060 N_i = 4
N. vectensis # Ng = 32	P = 0.17 σ = 0.018^ N_i = 72	P = 0.44 σ = 0.024^ N_i = 191	P = 0.39 σ = 0.023^ N_i = 169
T. adhaerens # Ng = 3	P = 0.36 σ = 0.091 N_i = 10	P = 0.32 σ = 0.088 N_i = 9	P = 0.32 σ = 0.088 N_i = 9

P = phase frequency; σ = standard deviation of the mean; Ng = total number of genes surveyed; Ni = total number of introns per phase; # = data from the Ensembl Metazoa genome browser; data was collected manually for all other species. ^ = significant difference from average phase distribution of analysed holozoan genomes ("Reference set"); > = significant difference from the *A. queenslandica* full dataset (i.e. AF and non-AF genes). Note that the *L. gigantea* phase 1 frequency was significantly different from the Reference Set value, but not from that seen in the *L. gigantea* genome-wide phase 1 introns.

considered *in toto*, the *AqAFs* have a median intron length of 72 bp (Table 2.1), a value slightly smaller than the genome-wide median (81 bp) (Srivastava et al. 2010). However, when analysed individually, the median intron sizes of *AqAFA*, *AqAFC* and *AqAFF* are shorter than the genome-wide value (61, 51 and 58 bp, respectively), while those in *AqAFB*, *AqAFD* and *AqAFE* are longer (135, 113 and 221 bp, respectively). The AF cluster is tightly packed, with a median intergenic distance of 103 bp (when including flanking and nested genes); this value is smaller than the median genome-wide intergenic region size of 824 bp (Srivastava et al. 2010).

Two sets of highly similar repeats are present in the gDNA encoding *AqAFC* and *AqAFE* (Figure 2.1b). In *AqAFC*, three repeat units span intron 10 to exon 14, intron 14 to exon 18, and intron 26 to exon 30. These repeats cover both intron and exon sequences and share about 85% pairwise sequence identity to one another. Two repeats are present in *AqAFE*, in exons 30 and 31 (96% pairwise identity). These repeats do not cover any intronic sequences and do not bear any sequence similarity to the *AqAFC* repeats (data not shown). It is currently unknown whether these repeats are real or represent genome sequencing artefacts.

2.4.7 Modular exon structure of protein domains

To investigate the relationship between AqAF domain architecture and genomic structure, the positions of all AqAF Calx-beta, VWA, VWD, and Wreath domains were mapped back to the underlying genomic DNA (gDNA) sequences (Figure 2.1b). Most AqAF Calx-beta domains from AqAFA to AqAFE are encoded by a module spanning three exons. When averaged across all sequences, these Calx-beta domains span the final 25 bp of the Exon i, the entirety of Exon ii (average 144 bp) and the first 130 bp of Exon iii. This pattern repeats itself, starting in the final 25 bp of Exon iii (Figure 2.11a; Appendix 2.7). VWA and VWD domains (with the exception of AqAFD VWA domain number 1, which spans two exons) all map to single exons (Exon iv), with a short spacer sequence at the beginning (17 - 29 bp) and end (25 - 124 bp) of each exon (Figure 2.11b; Appendix 2.7). Finally, for all AqAFs except AqAFF (in which the domain is absent) and AqAFB (where the domain is encoded by two exons) the Wreath domain is encoded by the final three exons of each gene (Exons v to vii), commencing partway through the antepenultimate exon and running to the end of the sequence. The length of the

first exon encoding the Wreath domain is variable between sequences, while the other two exons are more consistently sized (Figure 2.11c; Appendix 2.7).

2.4.8 Intron phase distribution patterns in *AqAFs* and other Calx-beta domain-encoding sequences

In order to further investigate the genomic structure of the AqAF genes, I determined the intron phase distribution patterns of the AqAFs, and compared them to those patterns observed in the full suite of Calx-beta domain-containing genes of eight invertebrate species (*A. queenslandica, Branchiostoma floridae, Capitella teleta, H. magnipapillata, Helobdella robusta, Lottia gigantea, Nematostella vectensis* and *Trichoplax adhaerens*) and one choanoflagellate species (*Monosiga brevicollis*), as well as the full complement of protein-coding genes from five of these species (*A. queenslandica, H. robusta, L. gigantea, N. vectensis* and *T. adhaerens*).

Genome-wide intron phase frequencies are non-random in all analysed species (Table 2.2). Each species follows an approximate distribution pattern of ~50% phase 0, ~30% phase 1, and ~20% phase 2 introns; these values are similar to those previously reported in various eukaryotes by Csuros et al. (2011). The values determined here are statistically significantly different from an expected random distribution of 33% per phase, in all cases except the phase 1 introns of *A. queenslandica*. As all five analysed species displayed similar phase distribution patterns, phase counts from each species were summed and used to estimate a generalised intron phase distribution pattern for basal metazoan species (Table 2.2). The distributions seen in the individual contributing species did not differ significantly from the generalised value. These reference values allowed statistical comparisons between subsets of genes and the genome as a whole, in species where genome-wide phase data is not readily available. All subsequent comparisons to genome-wide phase values discussed below involved this generalised reference dataset (including those performed in species where genome-wide phase data is available).

The *A. queenslandica* Calx-beta domain-containing genes show an intron phase distribution that is significantly different from the corresponding basal metazoan genome-wide values, with frequencies of 15% phase 0, 77% phase 1 and 6% phase 2; this trend remains when examining only the non-AF Calx-beta domain containing *A. queenslandica* genes (Table 2.3). A more extreme difference is the trend observed in the *A. queenslandica* AF-only dataset; here all introns except one (which is in

phase 0) occur in phase 1 (n = 157; Figure 2.1b; Table 2.3). These values are not only statistically significantly different from the genome-wide reference dataset, but also from the *A. queenslandica* Calx-beta domain-containing gene set. This result also differs from the *C. prolifera* AFs, which are equipped with phase 0 introns only (Fernàndez-Busquets and Burger 1999).

The strong bias towards phase 1 introns observed in the *A. queenslandica* Calx-beta domain-containing subset is not maintained within the Calx-beta domain-containing genes of other analysed species. The analysed datasets of only three other species, *B. floridae, L. gigantea* and *N. vectensis*, also exhibit phase 1 frequency distributions that are significantly different from the genome-wide reference set (note for *L. gigantea,* the phase 1 frequency is statistically different from the invertebrate reference dataset, but not from the *L. gigantea*-specific dataset; the other two phases are significantly different, however) and in all three cases the frequency of phase 1 introns is roughly equal to that of another phase (i.e. ~40% phases 0 and 1 in *L. gigantea and B. floridae*; ~40% phases 1 and 2 in *N. vectensis*. Significant spikes of enrichment for any single phase are not observed elsewhere.

The low numbers of Calx-beta domain containing genes, and introns contained therein, from *H. robusta*, *M. brevicollis* and *T. adhaerens* genomes impede the collection of meaningful statistics about phase distribution frequencies or patterns from these species; it is clear from these low numbers, however, that these species deploy Calx-beta domains in a way that is very different to *A. queenslandica*.

2.5 Discussion

2.5.1 Candidate aggregation factors are present in demosponge and hexactinellid sponges

The AF complex is a multimeric proteoglycan assembly that facilitates cellular recognition and adhesion between sponge cells (Popescu and Misevic 1997). In *C. prolifera*, the core AF is encoded by MAFp3 and MAFp4 (Fernàndez-Busquets et al. 1996; Fernàndez-Busquets and Burger 1997; Fernàndez-Busquets et al. 1998) which appear to be encoded by a single transcript (Fernàndez-Busquets and Burger 1997; Fernàndez-Busquets et al. 1998) and later cleaved to produce independent protein subunits (Jarchow et al. 2000). Related sequences are present in *A. queenslandica* (Gauthier 2009; Srivastava et al. 2010) and *S. domuncula* (SdSLIP) (Wiens et al. 2005). For the first part of this research, I sought to catalogue the candidate AFs that exist across the Porifera, with the goal of determining

the evolutionary origin point of these genes. To this end, I surveyed the genomes or transcriptomes of thirteen sponge species (plus the known *A. queenslandica* AFs) to find relevant sequences. Based on known sequences, I defined a candidate AF sequence as one equipped with a Wreath domain (Group 1) or multiple Calx-beta domains plus top BLAST matches to known AFs (Group 2). Using these criteria, I conclude that AFs are a demosponge + hexactinellid-specific innovation (Figures 2.7-2.8, Figure 2.10, Appendix 2.6).

2.5.2 Group 1 AF sequences are present in all analysed demosponge species

Group 1 AF candidates are those AF-like sequences equipped with a Wreath domain, regardless of their additional domain content or sequence properties (Figure 2.7). The Wreath domain is a defining motif shared by sequence homologues of the *C. prolifera* MAFp3 protein sequence, which in this species plays a functional role in AF assembly (Jarchow et al. 2000). The Wreath domain has not been identified to date outside known or probable AFs, with the exception of the *S. domuncula* protein SdSLIP, which possesses a Wreath domain but also has LPS-binding functionality (Wiens et al. 2005). Despite the unusual nature of SdSLIP, possession of a Wreath domain is currently the best indication that an unknown sequence represents a putative AF. Group 1 sequences are present in all demosponge species tested, in numbers ranging from three to thirteen transcripts per species.

In *C. prolifera* , all identified Group 1 sequences are exclusively comprised of Calx-beta and Wreath domains, suggesting that, as in *A. queenslandica*, members of the AF suite in this species are fairly uniform in terms of domain architecture. Three identified *C. prolifera* sequences (Cp_79623.1.2.38, Cp_79623.1.4.28 and Cp_64051.0.1.19) show high sequence identity to one another and to MAFp3 (with best matches to isoforms B/C, D and E respectively). However, all three sequences are much shorter than their corresponding known MAFp3 isoforms and therefore probably represent fragmented sequence assemblies. Other *C. prolifera* Group 1 sequences exhibit lower sequence identity to known sequences in this species, and may therefore represent novel members of the CpAF gene family. Beyond *C. prolifera*, sequences comprised solely of Calx-beta and Wreath domains represent at least one-third of Group 1 sequences in all other examined species. As in the Group 1 *C. prolifera* sequences, the majority of these sequences are short to moderate in length, relative to known *A. queenslandica* and *C. prolifera* AFs. Just four instances of long transcripts with a large number of Calx-beta domains (n

≥ 8) were observed (*C. nucula* Cn_2149.81, *C. prolifera* Cp_80199.3.1.97 and *E. muelleri* Em_38031 and Em_38028). The absence of long AF sequences elsewhere may indicate that, on average, most AFs are truly shorter than those present in *A. queenslandica* or *C. prolifera*. However, especially given the general lack of signal peptides in these short sequences, it is probable that long sequences were simply not captured during RNA sequencing, or that sequencing reads were not joined into long transcripts during *de novo* assembly.

An *A. queenslandica* AF-like domain composition of Calx-beta, VWD, and Wreath domains is present in three sequences (*C. nucula* Cn_4622.37, *P. ficiformis* Pf_2934.29 and Pf_7582.101). The sparse distribution of sequences equipped with VWD and Wreath domains together means that reconstruction of the evolutionary origin of this domain coupling is not currently possible (Figure 2.10). The VWD domain may have incorporated into the AFs in the demosponge ancestor, or may be the product of several independent domain shuffling events in the different sponge lineages; more sequencing data from a wider range of sponge species is required before a meaningful conclusion can be drawn. In *A. queenslandica*, AqAFB and AqAFE are equipped with VWA domains; however, no Group 1 or 2 candidate AF sequences are predicted to contain this domain type. VWA domains were identified in four Group 3a sequences; however for the purposes of this study these sequences are not considered to be likely AFs. It therefore may be the case that the inclusion of a VWA domain in the *A. queenslandica* AFs is an *Amphimedon*-specific innovation; this is currently unclear without more comparative transcriptomic and genomic data.

All examined demosponge species except *C. prolifera* encode at least one sequence possessing a Wreath domain only (Figure 2.7). These sequences could represent truncated sequences, but it also remains possible that these are true independent transcripts. Indeed, one Wreath domain-only sequence includes a signal peptide (*C. nucula* Cn_13331.30), suggesting that the 5' end of this sequence is complete. In *C. prolifera*, the Wreath domain and arm subunit regions appear to be transcribed as a single contiguous mRNA (Fernàndez-Busquets and Burger 1997; Fernàndez-Busquets et al. 1998), but the two subunits are independent in their mature protein forms (Jarchow et al. 2000). If the free Wreath domain sequences observed in the present study are real, they may be the result of a post-transcriptional processing event, captured by RNA sequencing, that separated the Wreath and arm

regions. It may also be the case that the Wreath region is expressed independently. However, longer transcripts encoding both Wreath and Calx-beta domains were identified in all species in which free Wreath domain sequences were also present, indicating that transcription of free Wreath domains is not obligatory.

Perhaps the most intriguing Group 1 sequences are those coupling Wreath domains to novel domain types not observed in other well-characterised AFs. Such sequences were observed in *E. muelleri* (Em_90236, Em_31799 and Em_140965)*, I. fasciculata* (If_3013.75)*, P. suberitoides* (Ps_6648.67) and *S. lacustrus* (Sl_2436.75). In these proteins, domain types such as Sema, PSI, Sushi and EGF-related domains are seen coupled to Wreath domains, suggesting that the Wreath domain may be involved in other functions beyond AF bridge formation. This possibility is supported by the LPS-binding role that the AF-related SdSLIP plays in *S. domuncula*. Contrary to the novel Group 1 sequences, however, SdSLIP is comprised of a single copy of each of a Calx-beta and Wreath domain, rather than possessing novel domain types. It is probable, but unconfirmed, that the Wreath domains in these novel proteins would still facilitate backbone formation, either ring-shaped or linear, although the role of these hypothetical structures is unknown.

EGF-related (reviewed by Campbell and Bork 1993) and Sushi (reviewed by Day et al. 1989) domains are promiscuous domains (Basu et al. 2008) that often mediate protein-protein interactions in a range of molecules, including those with cell adhesion or immune functions. Although a precise role for these domain types in candidate AFs identified here or in *G. cydonium GEOCY_AF* is currently unknown, their inclusion is not wholly surprising due to their wide distribution in proteins from other self-nonself recognition and immune systems (data not shown). Sema and PSI domains, as seen in the present study in the freshwater sponges *E. muelleri* and *S. lacustrus*, are perhaps best known for their role in semaphorin-mediated axon guidance (Kolodkin et al. 1993), but have also been implicated in cell adhesion and migration processes (reviewed by Casazza et al. 2007). Although the Sema and PSI domains have been observed together in representative taxa from choanoflagellates (data not shown), sponges and ctenophores (Ryan et al. 2013), no instances of a Sema-PSI-Wreath domain combination has been observed in the *A. queenslandica* genome (data not shown). The function of this novel domain combination in *E. muelleri* and *S. lacustrus* candidate AFs is mysterious. However, it is possible that

the Wreath domain allows these molecules to form circular or linear backbones, and that the Sema-PSI region mediates cell-cell or cell-extracellular matrix tethering (Casazza et al. 2007).

2.5.3 Group 2 sequences are present in demosponges and hexactinellids

Group 2 sequences, those not encoding Wreath domains but which are top BLAST hits to other known AFs or AF-related sequences from *A. queenslandica, C. prolifera* or *S. domuncula*, are present in *A. vastus* (the sole hexactinellid species analysed in this study) and all analysed demosponges except *I. fasciculata* and *P. suberitoides* (Figure 2.8; Figure 2.10). The majority of Group 2 sequences contain Calx-beta domains only, with some sequences also containing signal peptides and/or transmembrane domains. A small number of *C. prolifera* (Cp_79465.0.4.99 and Cp_77978.0.7.38) and *C. elegans* (CeS_76241.66) sequences also encode domains belonging to the immunoglobulin superfamily. Besides these sequences, however, overall the homogenous nature of the Group 2 sequences is striking, possibly indicating that non-Wreath domain-equipped AFs do not tend to include additional novel domain types.

2.5.4 Group 3 sequences

Group 3 sequences are AF-like but do not contain additional sequence properties identifying them as candidate AFs (Figure 2.9). Group 3a sequences most likely represent sequences equipped with Calx-beta domains that play non-AF functions, while Group 3b sequences appear to be members of other protein families. Group 3 sequences are reported here but were considered unlikely to be true AFs.

2.5.5 Phylogenetic distribution of sponge AFs

Group 1 and 2 members represent potential novel AF sequences and were identified in demosponges and hexactinellid species. Group 1 or 2 sequences were not identified in any analysed homoscleromorph (*O. carmela* or *C. candelabrum*) or calcareous (*S. coactum* or *S. ciliatum*) sponges, despite the availability of full genome sequences for *O. carmela* and *S. ciliatum*. These four species are present in a clade separate from the demosponge + hexactinellid lineage (Figure 2.10) (Thacker et al. 2013). Therefore, the present dataset suggests that AFs, at least in the form best known from *C. prolifera*, are a demosponge + hexactinellid-specific innovation. While failure to detect AF-like sequences in large datasets, particularly those generated from transcriptome libraries, is not definitive evidence of their absence from the sponge species sequenced, it is striking that the datasets analysed

for *O. carmela* and *S. ciliatum* represent full genome sequences, in which successful sequencing is not context- or expression-dependent, increasing the likelihood of sequence detection.

Self-nonself recognition and cell reaggregation phenomena have been studied in representative species of the calcareous sponges. Calcareous sponges are capable of discrimination between self and nonself at the tissue level (Amano 1990). However, tests of the cellular reaggregation capacity of calcareous sponges have suggested that these sponges undergo primary aggregation only, that is, aggregation that is not facilitated by a soluble aggregation factor (Müller 1982). It therefore appears that the AFs are absent in at least some species of calcareous sponge (Müller 1982); the results reported in the present study appear to support this conclusion. To the best of my knowledge, the presence of AFs and cellular reaggregation functionality has not been investigated in homoscleromorph sponges to date. Humbert-David and Garrone (1993) reported the presence in *O. tuberculata* of a circular molecule closely resembling the circular core AF structure; however, it is unknown whether this represents a true AF, and if so, whether the underlying protein sequence is similar to that in *C. prolifera* and other characterised species.

2.5.6 Limitations of AF candidate identification

I report the identification of 155 AF-like sequences from thirteen sponge species; 91 of these sequences met additional criteria to be considered candidate aggregation factors. The methods used to identify these sequences are suitable for preliminary analysis of candidate sequences for hypothesis generation and further study. However, further research is required to verify the nature of these sequences.

First, the majority of the datasets analysed here are the result of *de novo* assembly of short sequencing reads, derived from mRNA transcripts. Therefore, sequences that are unexpressed or lowly expressed in the biological context sampled may not be captured. The quality of these datasets is also reliant on read assembly - sequence truncations, splits and incorrect isoform assignment are common phenomena in datasets such as these, and may lead to sequences either failing to meet the filtration criteria used here, being present in a truncated form, or being represented multiple times as several partial sequences belonging to a longer transcript. These assembly issues may particularly impact the AFs, as these sequences are expected to be highly allelic with multiple forms present both between and

within individuals (Fernàndez-Busquets and Burger 1997; Fernàndez-Busquets et al. 1998); *de novo* assembly tools are therefore likely to struggle in these regions. These issues are compounded by the fact that only one of the datasets used here (the *C. prolifera* transcriptome) was generated in-house, meaning that data quality in the other datasets is harder to assess and improve.

A key assumption made for this analysis was that the AF domain architecture in other species is similar to that in *A. queenslandica* and *C. prolifera*, and therefore that filtering sequences based on domain architecture is appropriate. However, considering the similarities that exist between the *A. queenslandica* and *C. prolifera* AFs, despite these species not being particularly closely related relative to the rest of the demosponges, it seems unlikely that other demosponge species, particularly other haplosclerids or poecilosclerids (i.e. the orders to which *A. queenslandica* and *C. prolifera*, respectively, belong) would develop an entirely different secondary structure for their AFs that could still support a proteoglycan structure similar to those known to be present in various demosponge species (Henkart et al. 1973; Müller and Zahn 1973; Humphreys et al. 1975; 1977; Jarchow et al. 2000). Any novel sequences equipped with domain types not seen in *A. queenslandica* and *C. prolifera* AFs would still be detected here, unless these sequences had abandoned Wreath domains or long stretches of Calx-beta domains entirely. This major evolutionary revision of AF structure seems unlikely, given the level of conservation between *A. queenslandica* and *C. prolifera* AFs.

AF-like sequences were considered to be candidate AFs if they exhibited top sequence similarity to known *A. queenslandica* or *C. prolifera* AFs, or to the AF-related *S. domuncula* sequence, SdSLIP. Due to the sequence variability expected between AF sequences, it is possible that some AFs were falsely assigned to Group 3 due to poor BLAST matches. Any sequences assigned to Group 3b are unlikely to represent AFs, as all sequences in this group were similar to genes in other non-AF gene families. The most probable false negative sequences would be those lacking any BLAST annotations. However this effect is likely to be minor, as only five such sequences were present in the current dataset, and four of these were equipped with Wreath domains (and therefore designated as Group 1 sequences) (Appendix 2.6).

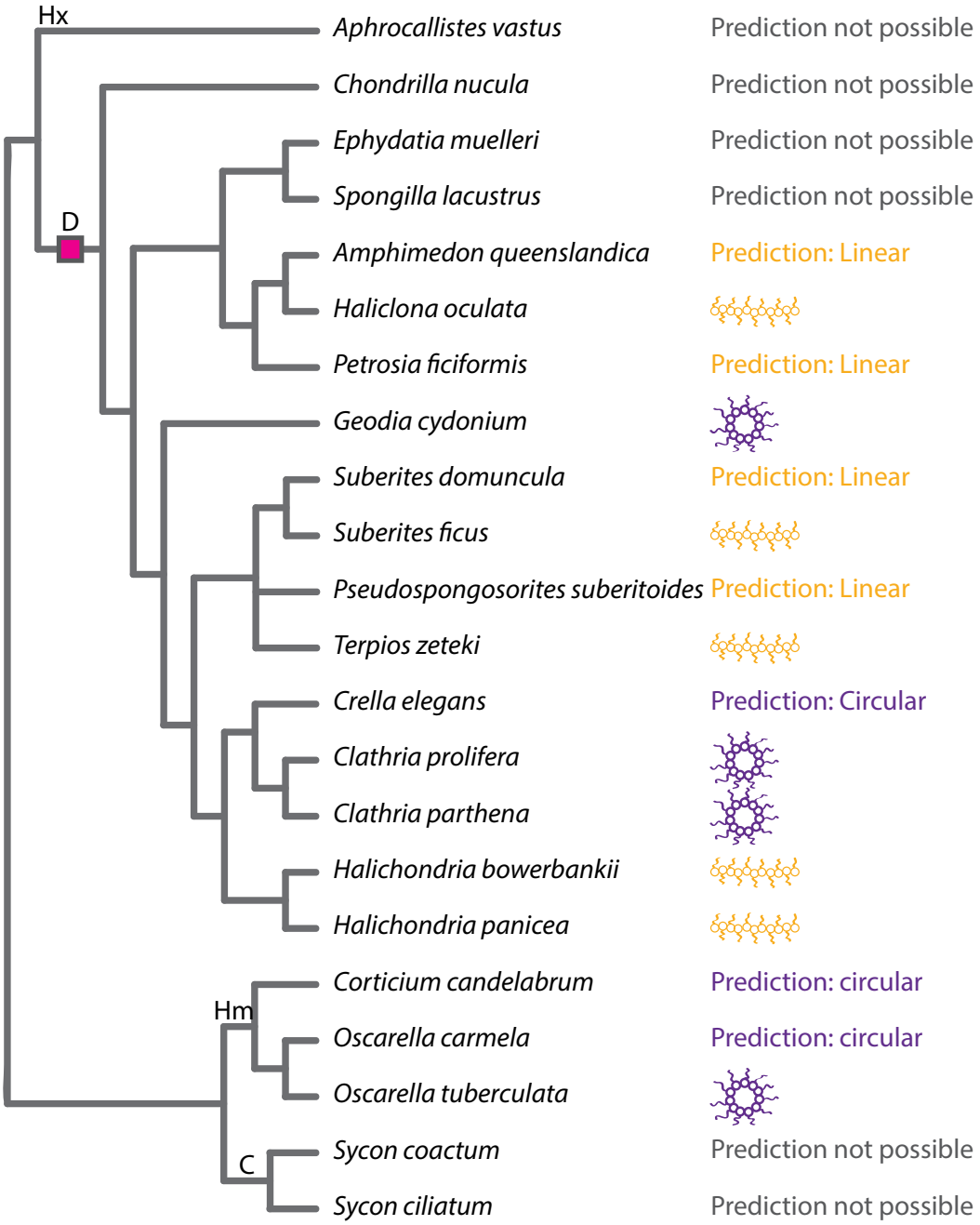


Figure 2.12 Known and predicted sponge AF core morphologies
The phylogenetic relationships between twenty-two sponge species (those for which AF structures are available, plus those used for the AF transcript identification portion of this study) are depicted in the tree (*left*). The macromolecular structure of the core AF has previously reported for nine sponge species, and is depicted on the right. AFs from *C. parthena, C. prolifera, G. cydonium* and *O. tuberculata* are circular (purple structures), with a central ring (equivalent to MAFp3 in *C. prolifera*) and radiating arms (*C. prolifera* MAFp4). Arms and ring subunits appear in a 1:1 stoichiometry. AFs in *Halichondria bowerbankii, H. panicea, Haliclona oculata, S. ficus* and *Terpios zeteki* are similar in overall structure to the circular form, but with a linear backbone (orange structures). Predictions can be made regarding AF core structure in some additional species, based on the forms present in closely-related species (*right*). The pink box indicates the evolutionary origin of the Wreath domain. Letters refer to sponge classes – Calcarea (C), Demospongia (D), Homoscleromorpha (Hm), Hexactinellida (Hx)

It is also possible that some Group 2 sequences received a top BLAST hit to known AF sequences simply because both were equipped with a large number of Calx-beta domains, with repeated instances of conserved residues building to a high degree of non-evolutionarily significant sequence similarity. It should be noted, however, that Group 3 sequences also contain many sequences encoding large numbers of Calx-beta domains (up to 19 in one sequence). The possession of a large number of Calx-beta domains thus does not automatically lead to a positive best BLAST match between two sequences, and therefore to erroneous assignment to Group 2. If this were the case, it would be expected that sequences encoding long stretches of Calx-beta domains would not be present in Group 3.

The *C. prolifera* AF sequences were originally partially determined by short peptide sequencing of purified AFs with known functional involvement in cellular reaggregation (Fernàndez-Busquets et al. 1996). The *A. queenslandica* AFs and novel AF candidates identified in the present study exhibit sequence homology and similar sequence properties to the *C. prolifera* AFs. However, no functional studies have been performed on their encoded proteins or purified AF complexes to date. Therefore, it currently remains unknown whether these sequences actually play any role in sponge cell adhesion and self-nonself recognition. The list of sequences identified here is therefore intended to serve as a preliminary set of hypotheses about the presence and properties of aggregation factors across the poriferans; these hypotheses can later be tested experimentally. An important piece of future research will be to purify known functional AFs from various species, and correlate their structural and functional properties with the sequence properties encoding the AF protein backbone, in order to better understand the interplay between AF sequence, structure and function.

2.5.7 Macromolecular structure of the AFs

The circular sunburst-like AF form is a proteoglycan structure that appears so far to be unique to sponges (Fernàndez-Busquets and Burger 2003). However, while the AFs of *G. cydonium* and *C. prolifera* are the best studied to date, they appear to be unusual in the larger context of the demosponges; it currently appears that linear AFs have a broader distribution throughout the demosponge lineage than the circular form seen in these two species (Figure 2.12). However, the sole probable AF structure isolated outside the demosponges, from the homoscleromorph *O. tuberculata*, is also circular. It is currently unknown precisely how the protein backbones of the AFs contribute to AF structure (and,

subsequently, function), as the only two species with both sequence and structural information available (*C. prolifera* and *G. cydonium*) show the circular form (Müller and Zahn 1973; Humphreys et al. 1975; 1977; Müller et al. 1978b). Indeed, homoscleromorphs appear to lack AF sequences altogether, and yet appear to possess AF or AF-like structures. Analysis of the phylogenetic distributions of the linear and circular AF forms allows the inference of the AF structures in the species which were studied here; it should be acknowledged that these are very tentative predictions that do not take the place of biochemical analyses. Circular AFs are found in *C. prolifera* (Humphreys et al. 1975; 1977) and *C. parthena* (Henkart et al. 1973). These species are all representative poecilosclerids, of which *C. elegans* is also a member. It can therefore be inferred that *C. elegans* AFs may also be circular. The demosponges of the family Suberitidae, *S. domuncula* (Müller et al. 1978a), *S. ficus* (Jarchow et al. 2000) and *Terpios zeteki* (Humphreys et al. 1977), each contain linear AFs; it is therefore probable that *P. suberitoides* AFs are also linear. The *Haliclona oculata* AFs are linear (Humphreys et al. 1977), perhaps indicating that AFs in the closely-related species, *A. queenslandica* and *P. ficiformis*, are also linear. This designation is less clear as there is only one representative structure available for this clade, however if this indeed is the case, it would provide compelling evidence that the structural form of the AFs is not tightly linked with the protein backbone sequence or that small changes are responsible for differences in AF form, since the *A. queenslandica* and *C. prolifera* AFs are quite similar at a domain architecture level (though not particularly at the sequence level), but would display different structural forms of the AFs. Finally, the homoscleromorph *O. tuberculata* possesses a circular AF (or AF-like) structure (Humbert-David and Garrone 1993), which suggests that such a form would also be found in the other studied homoscleromorph species. It is not currently possible to predict the structures of other examined species from this study, without the availability of structural information from a wider range of species. Ideally, future studies would be performed so as to produce both sequencing and structural information from each species, and particularly to examine instances of both circular and linear AFs. This may help to elucidate which, if any, sequence features of the AFs correlate with a circular or linear structure.

2.5.8 Protein domains associated with AF-like sequences

a. Calx-beta domains

The most prevalent feature of the *A. queenslandica* and *C. prolifera* AFs are the Calx-beta domains. The *A. queenslandica* genome contains a high number of Calx-beta domains (n = 96) and Calx-beta domain-containing genes (n = 59) compared with other representative basal metazoan species (Figure 2.5; Appendix 2.4). 36% of the *A. queenslandica* Calx-beta domains (n = 35) are included within the six AqAF proteins (Figure 2.1b). The homoscleromorph sponge *O. carmela* encodes fewer Calx-beta domains within its genome (n = 21), suggesting that the recurrence of this domain in *A. queenslandica* represents a lineage-specific expansion. It is not currently possible to determine when this radiation occurred, without genome sequences from a wider range of sponge species. However, as AFs (at least those similar to the ones analysed here) appear to be demosponge + hexactinellid-specific (Section 2.5.5), it is not unreasonable to predict that high Calx-beta domain numbers are limited to these taxa. A similar spike in Calx-beta domain numbers is observed in *Nematostella vectensis,* but not in other analysed cnidarians, *A. digitifera* and *H. magnipapillata*; however, this radiation is presumably evolutionarily independent to that observed in *A. queenslandica*.

Calx-beta domains appear to have originated in the holozoan common ancestor (Figure 2.5), as they were not found in any other examined eukaryotes. Intriguingly, Calx-beta domains were also discovered in high numbers in a range of mostly-marine bacteria species (Appendix 2.4). The history of these bacterial domains is unclear. While the presence of these domains in both holozoans and some bacteria may represent convergent evolution, it is also possible that the domains were transferred to bacteria via lateral gene transfer. It is unlikely that these domains share a direct common ancestor, due to the unparsimonious requirements for mass loss events in all intervening lineages.

Calx-beta domains were first reported by Schwarz and Benzer (Schwarz and Benzer 1997) in the *Drosophila melanogaster* Na^{+}-Ca^{2+} exchanger protein Calx. Calx-beta domains are composed of β-strands that come together to form a β-sandwich conformation (Schwarz and Benzer 1997; Hilge and Aelen 2006). Many Calx-beta domains contain high-affinity calcium binding sites (Matsuoka et al. 1997), which bind up to four calcium ions (Nicoll et al. 2006). Aggregation factor complex stabilisation is calcium dependent (Jumblatt et al. 1980). Calcium binding tests have determined that the *C. parthena*

AF possesses over 1000 Ca^{2+} binding sites, plus an additional large population of weaker Ca^{2+} binding sites that rely on a higher Ca^{2+} concentration for binding activity (Cauldwell et al. 1973). It has been proposed that the former population stabilises AF complex formation, while the latter allows AF-cell binding (Cauldwell et al. 1973).

b. VWA and VWD domains

The *A. queenslandica* AFs, with the exceptions of AqAFA and AqAFF, possess VWA or VWD domains. Various AF candidates distributed across the poriferans possess VWD domains, but the inclusion of VWA domains in candidate AFs has not been observed outside *A. queenslandica* (Appendix 2.6). The evolutionary origin of a VWD-equipped AF is unclear, as the distribution of these sequences as observed in the present study is polyphyletic. VWD-equipped AFs are present in a small number of species distributed across the demosponges; VWD domains therefore may have incorporated in the demosponge ancestor (or earlier) and subsequently been lost in several lineages. Alternatively, the incorporation of VWD domains into AFs may have occurred in several distinct lineages.

VWA and VWD domains have vastly different evolutionary origins. VWD domains are present in the Metazoa, as well as in *M. brevicollis* and *N. gruberi.* In contrast, VWA domains are an ancient domain family, being found in all examined taxa, with the exceptions of *Saccharomyces cerevisiae* and *Pedobacter saltans*. These findings support the results of Whittaker and Hynes (Whittaker and Hynes 2002), who previously demonstrated the wide phylogenetic distribution of VWA domains, and expand upon their work by examining the genomes of a wider range of species than were available in 2002. The profile HMM models of VWA and VWD domains available on Pfam show little sequence similarity between the two domain types.

The role that the VWA or VWD domains play in the AFs is mysterious. However, VWA domains have been proposed to functionally participate in protein adhesion and aggregation in proteins such as integrins (Whittaker and Hynes 2002). The VWA MIDAS (metal ion-dependent adhesion site) motif has been implicated in divalent cation-dependent (usually Mg^{2+}, but also Ca^{2+}) ligand binding (Cantí et al. 2005); MIDAS motifs are present within each VWA domain in the AqAFs (data not shown). As AF functionality is Ca^{2+} and Mg^{2+} dependent (Galtsoff 1925; Humphreys et al. 1960), it is possible

that the incorporation of VWA domains into the AqAFs aids cation-mediated aggregation in some way. The VWD domains lack a MIDAS motif, and the role of these domains in the AqAFs remains unclear.

c. Wreath domains

The MAFp3 region in *C. prolifera* is responsible for the formation of the central ring of the AF structure in this species, and subsequently for homologous self-interactions between individual AF structures (Jarchow et al. 2000). It is expected, but not experimentally verified, that the equivalent structure in linear AFs is encoded by a homologous sequence. Regions exhibiting MAFp3 sequence homology are present in *A. queenslandica*, *S. domuncula* and all demosponge sequences examined here, although a functional role for these homologous regions is yet to be empirically verified. In light of the key functional role of this region in *C. prolifera*, its independent structure and multi-species distribution, I propose that this region represents novel protein domain, the Wreath domain.

The Wreath domain appears to be a demosponge-specific evolutionary novelty. The majority of sequences that include a Wreath domain display domain organisations very similar to the *C. prolifera* AF sequences. However, in a limited number of newly-identified sequences presented here, the Wreath domain is coupled to domain types unknown from previously identified AFs. It is unclear whether these sequences represent AFs, AF complex-associated proteins, or unique proteins that have co-opted Wreath domain functionality for novel purposes. Assuming that such novel sequences assemble as in the AFs, it is unclear whether they would form circular or linear structures.

2.5.9 The *A. queenslandica* AFs exhibit a low level of sequence similarity

The Calx-beta domains in *A. queenslandica* AqAF and non-AqAF proteins, and in *C. prolifera* MAFp3 isoform C exhibit a low level of sequence similarity within and between proteins (Figure 2.6); only a small number of residues are conserved between domains. Notably, almost all residues identified by Hilge and Alean (2006) as being involved in Ca^{2+} binding are conserved in most AqAF Calx-beta domains (data not shown). This suggests that while a large amount of mutation has occurred in the Calx-beta domains of the AqAFs and other *A. queenslandica* genes, these domains can still function so long as the key functional residues are preserved. In contrast to the *A. queenslandica* sequences, most of the *N. vectensis* Calx-beta domain-encoding genes are highly similar both within and between

themselves, with only four of ten genes displaying a lower level of sequence identity between domains from the same gene. This demonstrates that Calx-beta domains in marine invertebrate species are not obligatorily diverse.

2.5.10 The *A. queenslandica* AFs are highly structurally constrained

The *A. queenslandica* AFs exhibit a high degree of structural constraint at the genomic level. The Calx-beta and Von Willebrand domains conform to precise boundaries within their encoding exon, throughout the AqAFs. For the VWA and VWD domains, these modules comprise a single exon, with a short spacer of non-domain sequence at either end. The simple genomic architecture of these domains likely facilitated their duplication and spread through the AFs from their presumptive ancestral form (Patthy 1996). The AqAF Calx-beta domains show a more complex modular structure, with domains in all genes except AqAFF conforming to a repeated three-exon organisation.

A. queenslandica AF and non-AF genes encoding Calx-beta domains exhibit remarkably consistent genomic architectures, despite their large numbers and presumably different functions. Overall, the introns of all *A. queenslandica* Calx-beta domain-containing genes show a highly significant over-representation of phase 1 introns, with 77% of introns in genes possessing Calx-beta domains (and almost 100% for the AqAFs) being in phase 1, compared with 33% genome-wide. This bias towards phase 1 introns was not observed in Calx-beta domain-equipped genes from any other analysed holozoan species. It is of particular interest to note that intron phase in CpAFs is biased towards phase 0 (Fernàndez-Busquets and Burger 1999). Therefore, intron phase bias may be a characteristic trait of the AFs, but with the precise nature of this bias being species-specific. It is notable that while the phases of the AF introns differ between these species, the phase bias in each is such that all exons in these genes are symmetrical – that is, all exons are flanked by introns in the same phase. Symmetrical exons are a requirement for exon rearrangement processes such as domain shuffling or alternative splicing, to prevent disruption to the transcriptional reading frame of the resulting mRNA (Patthy 1987; Fedorov et al. 1998). Therefore, such rearrangement processes could be occurring in the AFs, either to diversify the AF genes between species, or to allow the generation of diversified AF transcripts between individuals of the same species. This phenomenon could be investigated further with additional transcriptome or genome sequencing data.

2.5.11 The genetic dissimilarity and structural constraint of the *A. queenslandica* AFs may contribute to AF diversity

The analysis of the genetic properties of the *A. queenslandica* AFs reveals an apparent paradox - that is, the low sequence conservation within and between genes, in sharp contrast with the marked structural conservation constraining these diverse sequences. It is possible that this phenomenon may be explained by the AFs' potential role in allorecognition and the need to generate between-individual diversity. Any molecular system involved in individual-specific processes requires an underlying level of polymorphism, in order to generate individual-specific labels that can be recognised by particular individuals. Generation of this required variation could be potentially achieved by one of (or several) possible mechanisms; examples include high allelic variance, alternative splicing, somatic recombination, or variation in associated non-protein molecules such as glycans.

2.6 Conclusion

The *AFs* are putative sponge allorecognition genes; six such genes are present in the *A. queenslandica* genome. The ability to discriminate between self and nonself does not manifest in the sponge until two weeks post metamorphosis (Gauthier and Degnan 2008). In Chapter 3, I investigate the gene expression profiles of the six *AqAF* genes across sponge life history, and conclude that the AqAFs play a novel developmental role, possibly in tandem with their putative adult allorecognition functionality.

CHAPTER 3 - DEVELOPMENTAL EXPRESSION OF THE *AMPHIMEDON QUEENSLANDICA* AGGREGATION FACTOR GENES

3.1 Abstract

Amphimedon queenslandica individuals do not acquire immunological competence until two weeks post-metamorphosis (wpm). As adult sponge allorecognition is putatively mediated by the aggregation factor (AF) complex, I hypothesised that the onset of allorecognition competency is triggered by the initiation of AF expression at 2 wpm. Using a genome-wide gene expression dataset, I traced the expression of the AF genes across development from the early cleavage-stage embryo to the fully mature adult sponge. This revealed that the AF genes are very highly expressed at all developmental stages, but exhibit a particularly large spike in expression at metamorphosis. I identified a suite of 122 other *A. queenslandica* genes with expression profiles that were highly correlated with those of one or more AF genes. This list of genes is statistically enriched for those with functions involved with developmental cell signalling roles. This study represents the first analysis of AF gene expression and potential functions across development. The expression of the AF genes in the absence of immunological competence in the developing sponge suggests that the AFs may play an important cell adhesion and/or signalling role in development, possibly operating in tandem with some of the developmental genes with which the AFs share an expression pattern.

3.2 Introduction

3.2.1 Normal development in *Amphimedon queenslandica*

The demosponge *Amphimedon queenslandica* has become a model species for investigation of the evolution and development of basal metazoans (Degnan et al. 2008a). *A. queenslandica* has a biphasic pelagobenthic lifecycle consisting of a hermaphroditic benthic adult stage, fertilisation via spermcast spawning, the internal brooding of embryos, and the release of pelagic larvae (Leys and Degnan 2001).

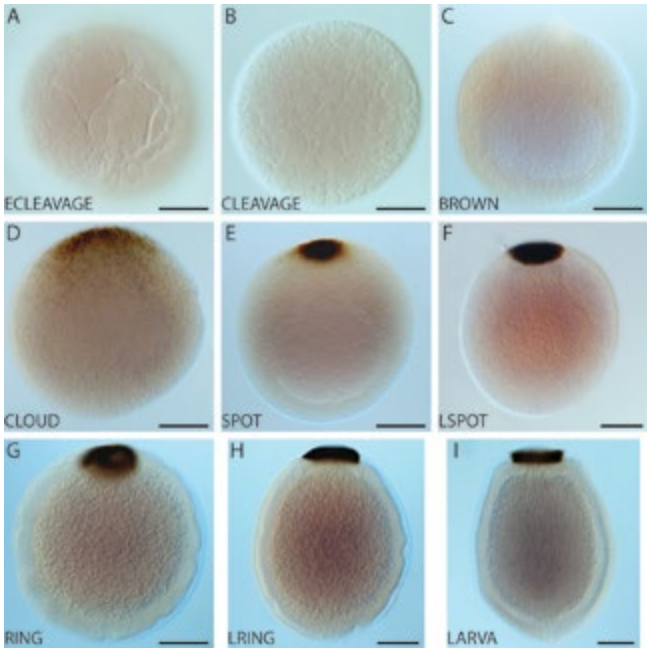


Figure 3.1 Embryonic development of *Amphimedon queenslandica*
Whole mount light micrographs of fixed embryos and larva. Posterior is to the top in panels C to I. Scale bar: 100 µm. Image by G. Richards (2010).

The progression of embryogenesis has been extensively characterised for *A. queenslandica* (see for example Leys and Degnan 2001; 2002; Degnan et al. 2005; Adamska et al. 2007; 2010; Nakanishi et al. 2014). *A. queenslandica* adults possess numerous brood chambers, each containing 20 - 150 asynchronously developing embryos (Leys and Degnan 2001). Multiple fathers contribute genetic material to the different embryos within a single brood chamber (K. Maritz, A. Calcino, and S. Degnan, unpublished data). Embryos can be staged based on the location of pigment cells over time. These cells are initially spread across the surface of the embryo (Figure 3.1c), but later migrate towards the embryo posterior pole (Figure 3.1d), then form a spot (Figure 3.1e-f) and finally a ring (Figure 3.1g-h) (Richards 2010). In larvae the mature pigment ring (Figure 3.1i) enables negatively phototactic swimming behaviour prior to settlement (Figure 3.2) (Leys:2001vy; but see Degnan and Degnan 2010).

A. queenslandica larvae are developmentally competent to initiate settlement and metamorphosis after about 4 – 6 hours in the water column (Figure 3.2), and high settlement capacity is retained until at least 32 hours post emergence (hpe) (Degnan and Degnan 2010). Settlement and metamorphosis can be induced by exposure to environmental settlement cues such as crustose coralline algae (CCA) (Degnan and Degnan 2010) or the articulate coralline alga *Amphiroa* sp. (S. Degnan and B. Degnan, personal communication). While larvae are capable of settling in the absence of an inductive cue, the overall rate of settlement within a cohort is much lower under these conditions (Degnan and Degnan 2010). Newly-emerged *A. queenslandica* larvae are not immediately able to respond to environmental settlement cues, and indeed it appears that early exposure to inductive substrates such as CCA leads to a delay in the onset

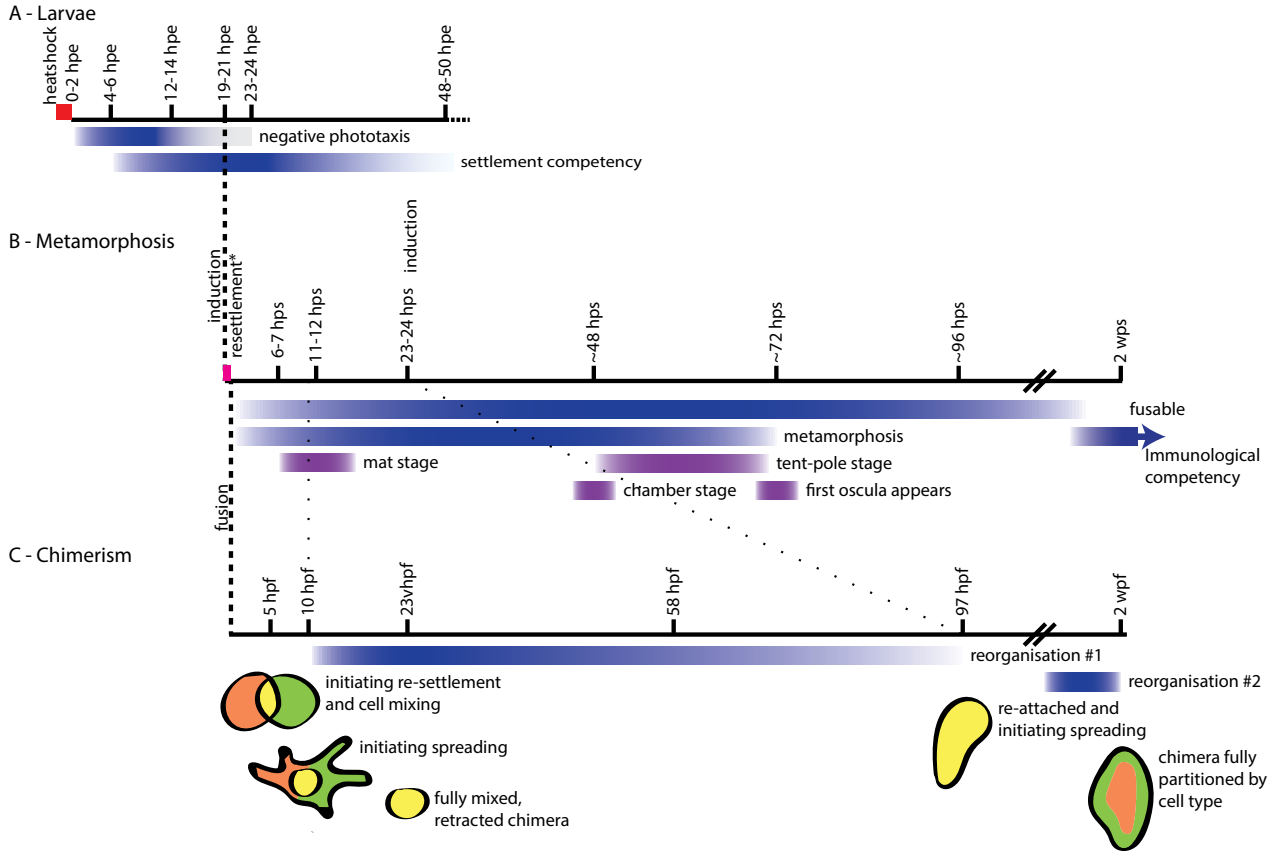


Figure 3.2 Normal and chimeric development of *A. queenslandica* larvae and juveniles

General times and phenomena characterising development of *A. queenslandica* individuals from the newly-emerged free-swimming larva to the fully-metamorphosed juvenile. (A) Larval release occurs naturally on a daily cycle; the number of released larvae can be enhanced by a two hour heatshock of a few degrees above ambient temperature (red box). Larvae collected at the end of this two hour period are therefore considered to be 0 – 2 hours post emergence (hpe). Larvae are developmentally competent to settle from about 4 – 6 hpe to about 48 hpe, although cohort-level settlement rate declines from about 32 hpe. For the CEL-Seq experiment analysed in this chapter, larvae were induced to settle at 19 – 21 hpe (dashed line). (B) For the CEL-Seq experiment, competent larvae were exposed to an inductive cue for 1 hour (pink box). After this time, unsettled larvae were discarded. Settled postlarvae were either kept on algae, or resettled on glass coverslips for observation after 48 hps. (C) Newly settled postlarvae (as shown in B) were resettled in contact with other conspecific postlarvae in experiments described by Gauthier and Degnan (2008). Cartoons depict the major developmental changes that occur in the chimera, in terms of morphology and cellular mixing. Here, cells from the two fused individuals are red and green, respectively, while yellow represents a mixed population of cells from each individual. Dotted lines show the point that chimeric development diverges from normal juvenile development; the ~97 hours post fusion (hpf) is morphologically similar from the normal ~24 hpe juvenile. Although dashed lines represent precise times at which induction and fusion were performed, these could occur at other times and the subsequent timelines would remain as shown. Approximate time ranges of behavioural and physiological processes are shown as blue shaded bars; morphological stages are shown in purple shaded bars.

of competency to undergo settlement and metamorphosis (Degnan and Degnan 2010).

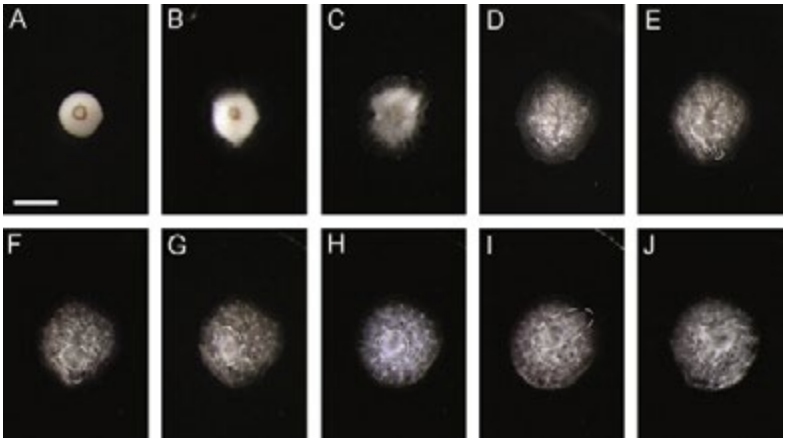


Figure 3.3 Morphological characteristics of postlarvae
Example of a postlarva (A) 0.5 hours post metamorphosis (hpm), (B) 6 hpm, (C) 24 hpm, (D) 48 hpm, (E) 72 hpm, (F) 96 hpm, (G) 120 hpm, (H) 144 hpm, (I) 168 hpm, and (J) 216 hpm. (A – C) The pigment ring is still apparent; (E – J) the osculum appears approximately 3 days post metamorphosis (dpm) and is indicated with a dotted circle in panels (G) and (J). Scale bar: 250 mm. Image by M. Gauthier and B. Degnan (2008).

Larval settlement proceeds by the initiation of substrate contact, rotation on the larval anterior pole, and the flattening and spreading of the anterior larval hemisphere across the substrate (Figure 3.3a) (Leys and Degnan 2002). By about 6 - 7 hours post settlement (hps), the postlarva has entered the mat stage (Figures 3.1, 3.3b), and has begun to spread across the substrate, and evidence of a large degree of apoptosis is present in the epithelium towards the edge of the spreading juvenile (Nakanishi et al. 2014). The sponge aquiferous system becomes apparent from about 48 hps (chamber stage; Figures 3.1, 3.3d), when a system of choanocyte-lined canals is first observed (Nakanishi et al. 2014). The following ~24 hours marks the tent-pole formation stage, where vertical spicule clusters raise the outer exopinacocyte layer into a 'tent-like' appearance (Figure 3.1) (Nakanishi et al. 2014). The appearance of the first osculum at about 72 hps marks the end of metamorphosis and the ability of the sponge to begin filter feeding (Figures 3.1, 3.3e-j). Individuals are considered to be juveniles, rather than postlarvae, from this point forward.

3.2.2 Allogeneic perturbations to normal *A. queenslandica* development

The ability of sponge larvae and postlarvae to fuse and form chimeric sponges has been noted both observationally and experimentally (Wilson 1907; Van de Vyver 1975; Uriz 1982; Simpson 1984; Ilan and Loya 1990; Maldonado 1998; McGhee 2006; Gauthier and Degnan 2008). The cellular fate of postlarval and juvenile chimeras has been tracked for three weeks in *A. queenslandica* (Figure 3.2) (Gauthier and Degnan 2008). In this species, sponges are capable of fusion at any point from the

completion of settlement until two weeks post metamorphosis (wpm). While spontaneous larval fusion is a relatively common phenomenon in this species and others (see cited references above), fused *A. queenslandica* larvae have not been observed to settle and metamorphose (Gauthier and Degnan 2008). For newly-fused postlarvae, cells from the two individual sponges intermingle, and postlarval development appears to proceed as normal until about 10 hours post fusion (hpf; Figure 3.4a). At this point, the chimeras undergo a ~12 hour process of partial or total reorganisation, whereby the chimera retracts or forms a ball, and may detach entirely from the substrate (Figure 3.4b-j). Over the next 2-3 days, the chimera then reattaches and recommences metamorphosis. The cells of the two contributing individuals remain intermingled throughout this process. For chimeras formed from newly-settled postlarvae, morphology at 97 hpf is similar to a normal postlarval individual at 24 hps (Gauthier and Degnan 2008).

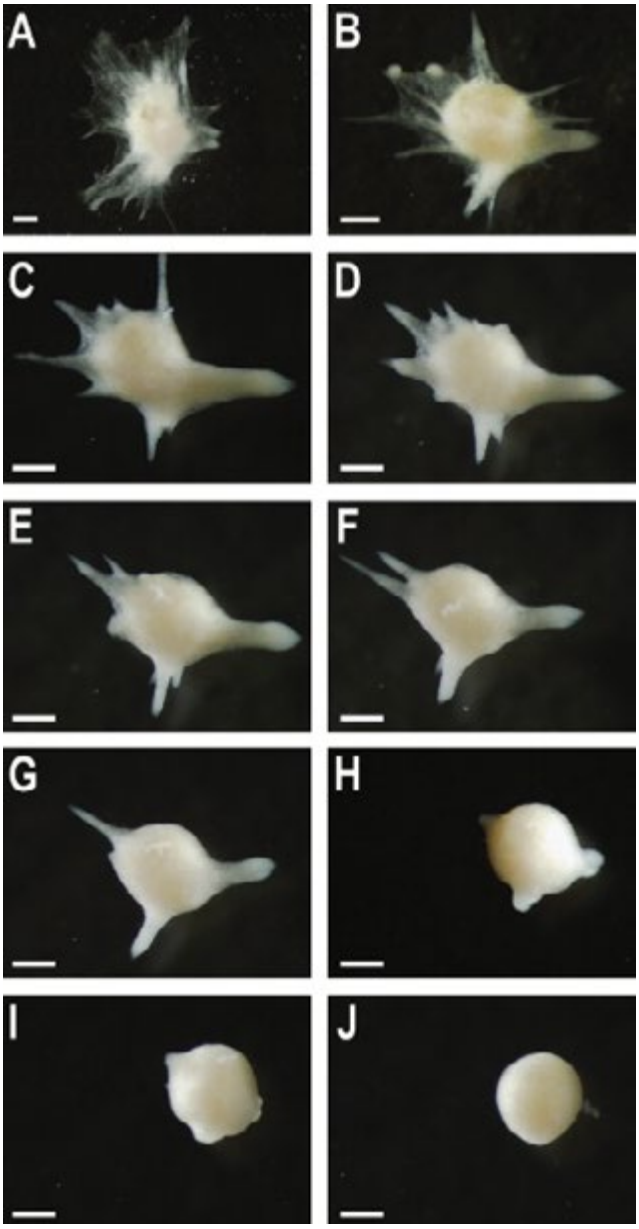


Figure 3.4 Postlarval chimerism
Time course documenting the resorption steps to ball formation in a chimeric postlarva resulting from the fusion of two individuals. The series of images are from a single chimera. Chimeras (A) 10 hours post fusion (hpf), (B) 15.5 hpf, (C) 17 hpg, (D) 18 hpf, (E) 19 hpf, (F) 20.5 hpf, (G) 21 hpf, (H) 21.5 hpf, (I) 22 hpf and (J) 23 hpf. Scale bar: 100 µm. Image by M. Gauthier and B. Degnan (2008).

Two week old *A. queenslandica* chimeras and juveniles undergo shifts in their allorecognition capabilities. At two weeks post fusion (wpf), chimeric juveniles undergo a cell sorting process, whereby the cells of one individual form the choanocytes, and those of the other individual form the pinacocytes and mesohyl (Gauthier and Degnan 2008). At 2 wpm, normal unfused juveniles lose the ability to fuse with other

individuals. The genetic and/or molecular basis of this transition is unknown. However, as chimeras display simultaneous sorting of cells by cell type and by individual, it is possible that this process is governed by a single molecule with both cell type- and individual-level specificity.

3.2.3 Aggregation factors and *A. queenslandica* development

Aggregation factors (AFs) are implicated in the adult sponge response to self-nonself challenges. AFs mediate the species-specific reaggregation of dissociated adult sponge cells *in vitro.* Cell-free experimental systems, demonstrating that this species-specificity remains when testing bead-coupled xenogeneic AF molecules, reiterates that the species-specific nature of this process resides within the AF complex (Müller et al. 1974; Jumblatt et al. 1980; Misevic and Finne 1987; Popescu and Misevic 1997; Misevic 1999; Jarchow and Burger 1998; Jarchow et al. 2000; Bucior et al. 2004). The AFs are also associated with the adult sponge tissue graft response, with *Clathria prolifera* individuals exhibiting upregulated AF expression in response to both allo- and autografts (Fernàndez-Busquets et al. 1998). The AF complex has also been shown to be recruited to the allograft interface in this species (Fernàndez-Busquets et al. 1998).

Little is known about the potential gene expression levels and functions, if any, of the AFs during sponge development. Given the role of the AFs in adult allorecognition, I hypothesised that initiation of allorecognition competency is triggered by the onset of *AqAF* expression around 2 wpm. I therefore sought to determine when *A. queenslandica AF* (*AqAF*) gene expression initiates in *A. queenslandica,* and whether the onset of this expression correlates with the activation of sponge allorecognition capabilities in the juvenile. In this chapter, I examine the quantitative expression of the AqAFs across normal sponge development, from the early cleavage-stage embryo through to the fully mature adult. The use of a large genome-wide sequencing dataset spanning 82 developmental time points has allowed me to finely trace the expression profiles of the *AqAFs*, and of a suite of other genes whose expression profiles are highly correlated to that of the *AqAFs*.

Table 3.1 Developmental stages of interest

State	Stage	No. Samples	Comment
Embryo	Cleavage	7	
Embryo	Brown	7	
Embryo	Cloud	7	
Embryo	Spot	5	
Embryo	Late Spot	8	Includes 2 samples identified as spot (morphologically) but grouped transcriptomically with late spot stage individuals
Embryo	Ring	7	
Embryo	Late Ring	6	
Larvae	Pre-competent larvae	5	Includes 1 sample each from 0-2, 2-4, 3-5, 4-6, 5-7 hpe
Larvae	Competent larvae	4	Includes 1 sample each from 6-8, 8-10, 9-11, 10-12 hpe
Larvae	Late larvae	2	Includes 1 sample each from 24-26 and 48-50 hpe
Juvenile	1 hps	3	
Juvenile	6-7 hps	3	
Juvenile	11-12 hps	3	
Juvenile	23-24 hps	3	
Juvenile	Tent Pole + Chamber	6	Includes 3 samples each of juveniles identified morphologically as tent-pole and chamber stage, but transcriptomically grouped together
Juvenile	Oscula	3	A single osculum is present
Adult	Adult	3	

3.3 Methods

3.3.1 Generation of a genome-wide expression quantification dataset using CEL-Seq

Analysis of ontogenetic expression levels of AqAFs and other genes was performed using a genome-wide expression dataset from 82 time points across *A. queenslandica* development (S. Fernandez Valverde, N. Nakanishi, K. Roper, B. Degnan, S. Degnan, unpublished data). Briefly, developmental

tissue from single sponge individuals was collected and processed by N. Nakanishi and K. Roper. Total RNA samples were extracted and used as input for CEL-Seq (Cell Expression by Linear amplification and Sequencing) as described by Hashimshony et al. (2012). Assembly of the sequencing reads, normalisation and quantification of genome-wide expression values, and developmental staging of samples was performed by S. Fernandez Valverde. Here, sequencing reads were mapped to the *A. queenslandica* genome, and expression of the 3' end of each *A. queenslandica* gene model (version Aqu2.1) was quantified according to the CEL-Seq protocol (Hashimshony et al. 2012). Precise ordering of the 82 samples was resolved using the BLIND clustering method (Anavy et al. 2014), which uses increasing transcriptional entropy of samples, rather than morphology, as a measure of developmental progression (Anavy et al. 2014). Larval samples were not re-ordered, as the collection of these stages was based on precisely-known maternal release times.

For the present analysis, the reordered set of 82 time points was grouped into 17 ontogenetic stages spanning the embryonic (n = 7), larval (n = 3), juvenile (n = 6) and adult (n = 1) stages of sponge development and metamorphosis (Table 3.1). Some of these 17 stages included individuals of mixed ages (such as the pre-competent larval time point, which included individuals ranging from 0 - 7 hpe) or of mixed morphological state (for example, samples that were morphologically identified as spot-stage embryos were included in both the spot and late spot groups, based on the results of the BLIND reordering process).

The normalised count values of the six *AqAF* genes across 82 time points were extracted from the genome-wide list; the average expression value for each of the 17 developmental stages was used for some analyses, as specified.

3.3.2 Statistical analysis of ontogenetic *AqAF* expression

Pairwise statistical differences in expression between each of the 17 developmental stages were calculated for of each *AqAF* gene. To do so, the normalised count values for each of the 82 CEL-Seq samples for each AqAF gene were used as input for one-way ANOVA and Tukey's HSD (honest significant difference) tests in R (http://www.R-project.org/) within the RStudio environment (http://www.rstudio.org). Circos plots (Krzywinski et al. 2009) were generated using the online version of the

Circos tableviewer tool (http://mkweb.bcgsc.ca/tableviewer), and show those pairs of developmental stages that exhibit statistically significant expression differences to one another. Data values used as input for the Circos plots were set to reflect the p-value generated by the Tukey's HSD test results (Appendix 3.1), such that the lower the p-value, the greater the width of the ribbons, within (but not between) a Circos plot. Integers to designate ribbon width were set at 500 ($p \leq 0.0001$), 50 ($p \leq 0.001$), 5 ($p \leq 0.01$) and 1 ($p \leq 0.05$) to reflect the differences in scale between the p-values.

3.3.3 Identification of genes exhibiting *AqAF*-like ontogenetic gene expression profiles

Mean expression values for every *A. queenslandica* gene were calculated, for each of the seventeen broad stages of sponge development. Developmental-wide expression values were summed for each gene. As the *AqAFs* are highly expressed, only those genes above the 75th percentile of total expression were selected for this correlation analysis (note that the bottom 50th percentile of genes exhibited expression levels of zero).

A correlation matrix comparing the developmental expression trends of all *A. queenslandica* genes was generated using R within the RStudio environment. An F-statistic was calculated in order to generate a p-value for each correlation value. After rearrangement of the resulting data table, genes whose expression pattern was significantly ($p \leq 0.05$) highly correlated ($cor \geq 0.95$) with any of the AqAFs were identified (n = 122 unique genes, not including correlated AqAFs). The commands used to perform this analysis are provided in Appendix 3.2.

All identified genes were analysed using the HMM Search function of DoMosaics (Moore et al. 2014) to identify conserved domain types, using the HMMER 3.0 hmmsearch and hmmplan binary files (Eddy 1998) and all Pfam-A domain profiles (version as per 31.04.14) (Finn et al. 2006), and run with default parameters. Signal peptides and transmembrane domains were predicted using Phobius (Käll et al. 2004).

3.3.4 Expression-based clustering

An unscaled heatmap showing the expression levels of each identified correlated gene was generated using the R function heatmap.2 within the gplots package (http://www.cran.r-project.org/web/packages/gplots/index.html).

3.3.5 Gene ontology enrichment analysis

The list of genes exhibiting correlations in developmental gene expression profiles to the AqAFs was analysed to identify significantly enriched gene ontology (GO) terms. Genome-wide GO annotation was performed for all *A. queenslandica* gene models (version Aqu2.1) by S. Fernandez Valverde using Blast2GO version 2.8 (Conesa and Götz 2008). The list of 122 co-expressed genes (plus all six AqAFs) was used as input for two-sided GO enrichment analysis using the Fisher's Exact Test tool (FDR [false discovery rate] p-value cutoff ≤ 0.05) included in Blast2GO Basic version 3.0.5, in order to identify over- or under-represented GO terms. The full enrichment list was restricted to include only the most specific GO terms therein; this restricted list was used for all further analyses. Biological Process and Molecular Function GO terms were examined further; the Cellular Component GO term was not deemed to be of interest at this time.

Examination of the Fisher's Exact Test results indicated that the GO term list was saturated with terms associated with a single pair of sequences, both identified as TGF-β receptor type-1 genes (Aqu2.1.41568_001 and Aqu2.1.41569_001). In order to better analyse the remaining genes and their associated GO terms, the Fisher's Exact Test was re-run in Blast2GO, omitting Aqu2.1.41568_001 and Aqu2.1.41569_001.

The annotation results from the BLAST (basic local alignment search tool) phase of the Blast2GO analysis were manually examined to identify potential mis-annotations or -attributions.

3.3.6 GO term clustering

Statistically enriched GO terms were clustered based on semantic similarity (SimRel measure) using the software REVIGO (Supek et al. 2011). Similar GO terms with a redundancy of >0.7 were

collapsed. The coordinates of the resulting semantic space scatterplot were exported and used to graph the GO term clusters in GraphPad Prism version 6.0 for Mac (http://www.graphpad.com).

All Venn diagrams were generated using the online tool Venny (http://bioinfogp.cnb.csic.es/tools/venny/index.html).

3.4 Results

3.4.1 Quantitative analysis of *A. queenslandica AF* expression across development

AqAF expression levels were tracked across development from the cleavage-stage embryo to the adult sponge, using an in-house genome-wide expression dataset comprising embryonic (n = 7), larval (n = 3), juvenile (n = 6) and adult (n = 1) developmental stages (82 sub-stages/time points; Table 3.1). Each *AqAF* gene is expressed in all developmental stages examined (Figure 3.5). The six *AqAF* genes are expressed at very high levels across sponge development, relative to other *A. queenslandica* genes, with expression values above the 75th percentile of genome-wide expression levels for most stages (Figure 3.6). For 62% of total developmental time point expression observations, gene expression levels are in the top 90th percentile relative to the rest of the genome at the relevant developmental stages (Figure 3.6).

The expression of each AqAF gene shows similar, but not identical, profiles across development. *AqAFA*, -*C*, and -*D* are all statistically highly correlated with *AqAFE*, but not with each other, in terms of overall expression profiles across development (discussed further in section 3.4.2). More specifically, a slow steady increase in *AqAFB* and *AqAFD* expression occurs in the embryo stage (Figures 3.5, 3.7), with statistically significant differences in expression observed in pairwise comparisons between some early- and later-stage embryos (Figure 3.8, Appendix 3.1). Besides these changes, *AqAF* expression is relatively stable across embryonic and larval stages (Figures 3.5, 3.7, 3.8; Appendix 3.1). However, a statistically significant increase in expression occurs at the transition between the ring and late ring embryo for *AqAFF*, and between the late ring embryo and post-competent larvae for *AqAFA* and *AqAFC* (Figure 3.8, Appendix 3.1).

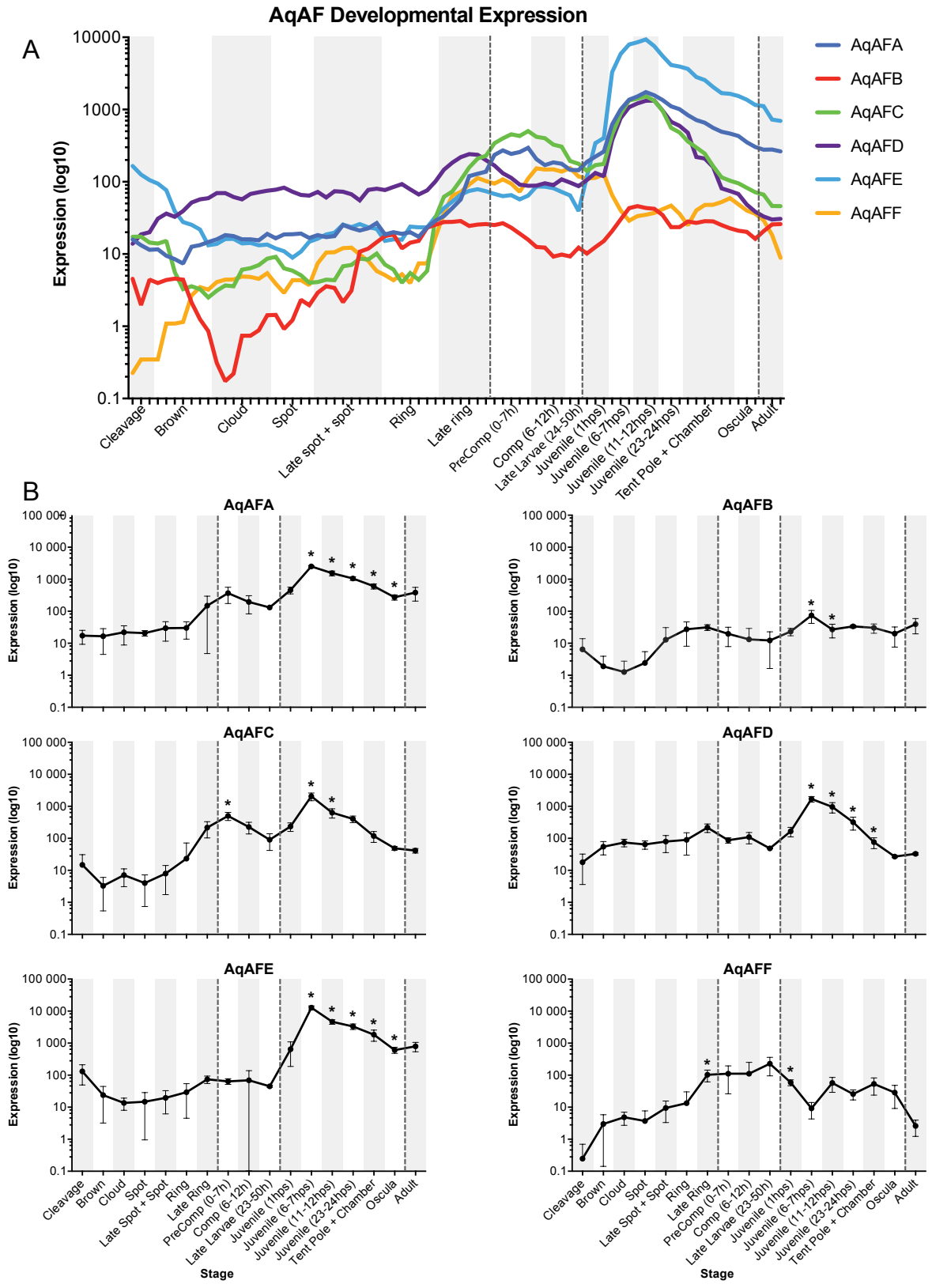


Figure 3.5 Developmental expression of *AqAF* genes
(legend over page)

←

Figure 3.5 Developmental expression of *AqAF* genes

(See previous page)

(A) *AqAF* log10 normalised gene expression levels across 82 time points. Labels and alternating grey and white bands denote seventeen developmental stages; curves represent a moving average (period = 5) of expression values over time. The 82 time points were ordered using the BLIND clustering method. (B) Average gene expression levels of each AqAF gene per broad developmental stage. Error bars depict the standard deviation of expression within each stage. Asterisks indicate those developmental stages where gene expression levels are statistically significantly different from those in the previous stage ($p \leq 0.05$). In all graphs, successive developmental stages are alternatively shaded grey and white; dashed lines mark transitions between embryo, larval, juvenile and adult stages. As plot (A) uses a moving average while those in (B) are averaged within a stage, the precise timing of peak and nadir points may differ between parts (A) and (B).

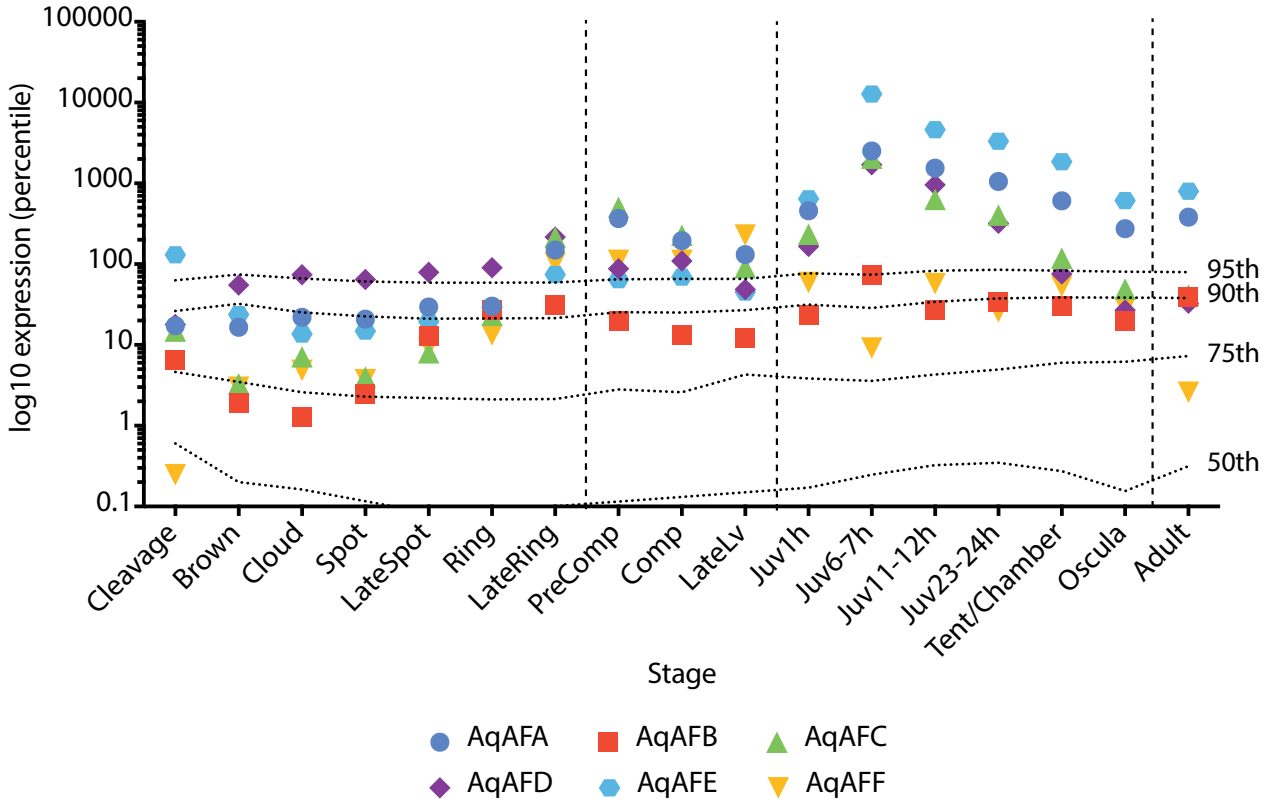


***Figure 3.6 A. queenslandica* AF expression relative to genome-wide percentiles**

Coloured data points represent the log10 normalised counts of *AqAF* gene expression in each developmental stage. Dashed lines show the genome-wide percentiles (50th – 95th) of transcript abundance in each developmental stage. Lines showing the 5th, 10th and 25th percentiles are not visible as these represent transcript counts of 0 across all stages.

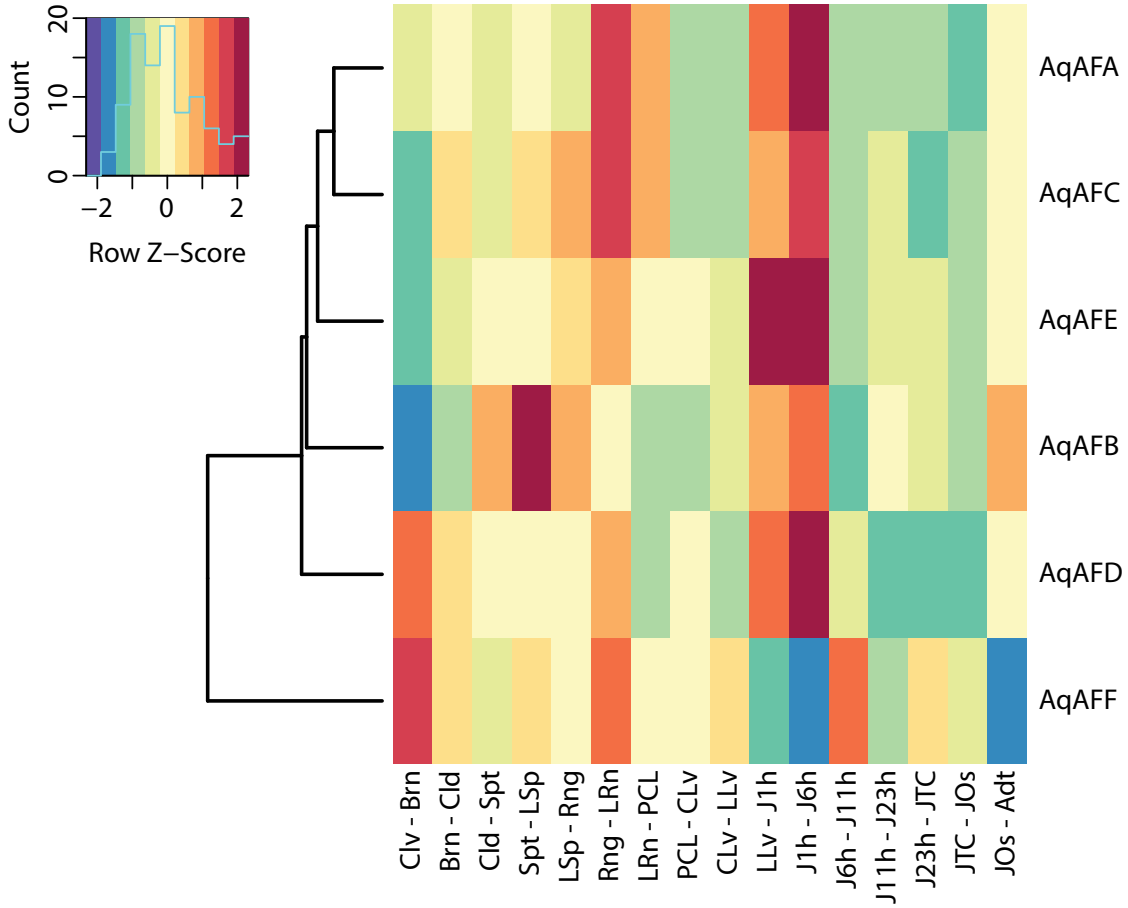


Figure 3.7 Patterns of *AqAF* gene expression changes across development
Heatmap depicting log2 expression fold changes to the *AqAFs* across sponge development. Values are scaled within each row, and rows are clustered based on expression similarity. Yellow boxes represent no change in expression, blue indicates downregulation and red indicates upregulation. Column names represent the transitions between successive developmental stages; stage abbreviations: Clv - cleavage, Brn - brown, Cld - cloud, Spt - spot, LSp - late spot, Rng - ring, LRn - late ring, PCL - pre-competent larvae, CLv - competent larvae, LLv - late larvae, J1h - 1 hps juvenile, J6h - 6-7 hps juvenile, J11h - 11-12 hps juvenile, J23h - 23-24 hps juvenile, JTC - tent or chamberstage juvenile, JOs - one-oscula juvenile, Adt - adult.

Figure 3.8 Statistically significant differences in *AqAF* expression across *A. queenslandica* development
(See next page)
The outer segments of each Circos plot depict the 17 broad developmental stages of sponge development, from embryonic cleavage (red) to adults (purple). An arrow marks the earliest developmental stage, and developmental stages progress clockwise. Stages exhibiting statistically significant differences in gene expression ($p \leq 0.05$), as determined by a Tukey's HSD test, are connected by coloured ribbons. Increasing connector widths within, but not between, plots represent decreasing p-values ($p \leq 0.0001$, 0.001, 0.01 and 0.05). Stage abbreviations: Clv - cleavage, Brn - brown, Cld - cloud, Spt - spot, LSp - late spot, Rng - ring, LRn - late ring, PCL - pre-competent larvae, CLv - competent larvae, LLv - late larvae, J1h - 1 hps juvenile, J6h - 6-7 hps juvenile, J11h - 11-12 hps juvenile, J23h - 23-24 hps juvenile, JTC - tent or chamber-stage juvenile, JOs - one-oscula juvenile, Adt - adult.

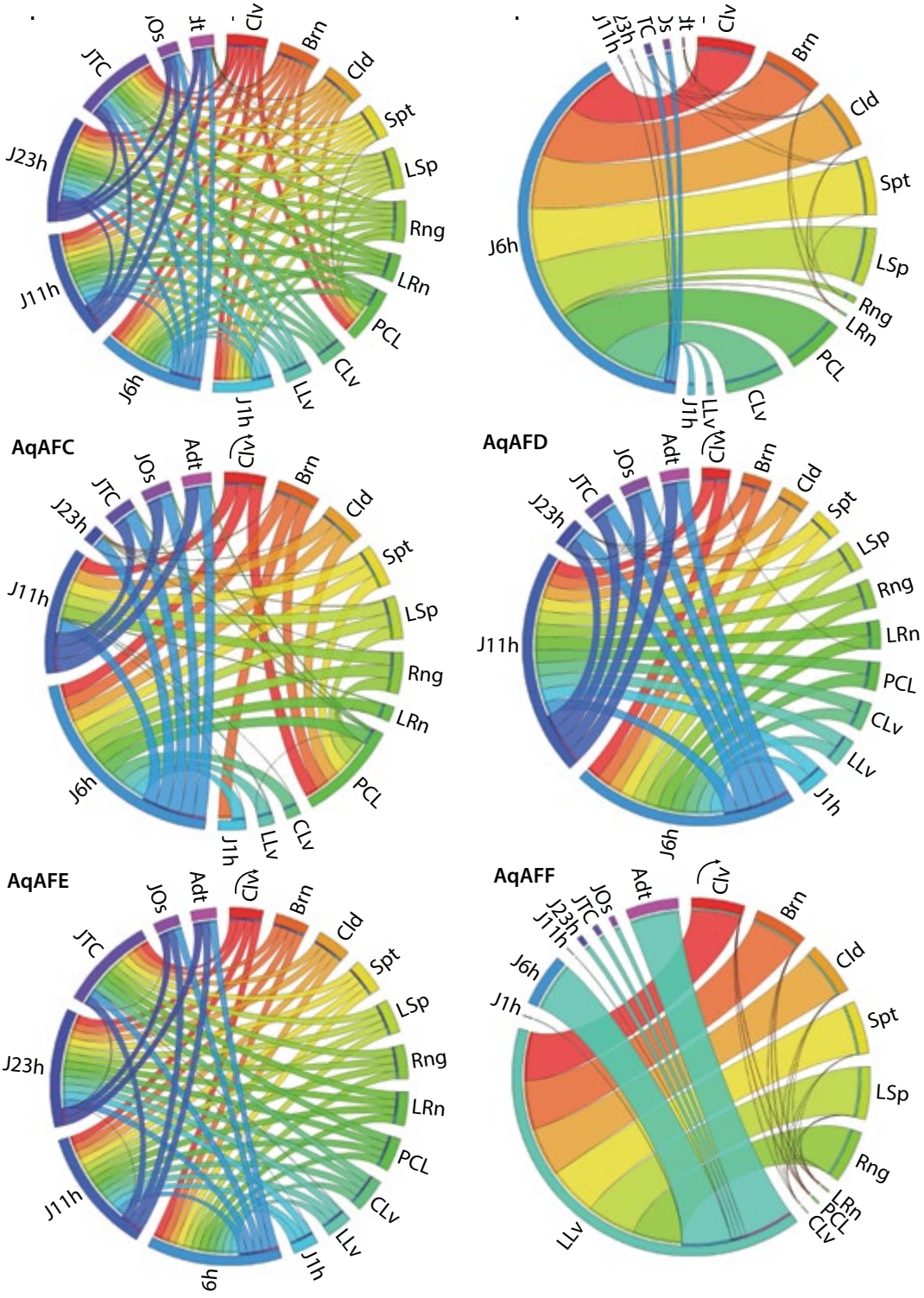


Figure 3.8 Statistically significant differences in *AqAF* expression across *A. queenslandica* development
(legend on previous page)

AqAFA to *AqAFE* are strongly upregulated in the early postlarva, with each gene exhibiting a ~1.5 - 4.5 fold increase (unscaled log2 fold change value) in expression between 0 - 1 and 6 - 7 hps (Figures 3.5, 3.7). The *AqAF* genes are amongst the most highly expressed genes at 6 - 7 hps, with *AqAFA, C, D* and *E* expression levels in the 99th percentile of genome-wide expression (Figure 3.6). For *AqAFA* to *AqAFE*, gene expression during the 6 - 7 hpe period is significantly different from that observed at all other developmental stages, except for the *AqAFB* 6 hps vs. adult pairwise comparison (Figure 3.8, Appendix 3.1). Later postlarval development sees a steady decline in *AqAFA* to *AqAFE* expression (Figure 3.5), although expression levels remain high relative to the rest of the genome (Figure 3.6). Expression appears to plateau between the single-osculum juvenile and the adult; however, as sequencing data are not available for older juveniles, further fluctuations in expression between these two chronologically distant stages cannot be ruled out.

The *AqAFF* expression profile differs from the other *AqAF* genes (Figure 3.2). In the embryo, *AqAFF* and *AqAFC* expression profiles are similar (Figure 3.2), and larval expression of *AqAFF* does not follow a markedly different trend from the other *AqAFs*. However, *AqAFF* expression does not increase in the early postlarval stage, and in fact exhibits a drop (though not statistically significant) in expression between 0 - 1 and 6 - 7 hps, when all other genes exhibit a large expression increase.

3.4.2 Identification of potentially co-expressed genes

As the *AqAFs* are clearly expressed prior to the onset of allorecognition capabilities, I sought to identify other genes that display similar developmental expression profiles to the *AqAFs*, in order to better understand the potential role/s of the AqAFs during sponge development.

I performed a genome-wide correlation analysis to identify relatively highly-expressed genes that exhibit a statistically significant correlation with the *AqAFs* in development-wide expression values. The commands used to perform this analysis take the overall expression trend of each gene across development, and use this information to perform pairwise comparisons between all *A. queenslandica* genes. Genes with expression trends correlated with those of *AqAFA* (n = 74), *AqAFC* (n = 26), *AqAFD* (n = 48) and *AqAFE* (n = 62) were identified; *AqAFB* and *AqAFF* expressions were not found to be correlated with any surveyed genes (Figure 3.9). As stated in section 3.4.1, *AqAFA*, *C,* and *D* are

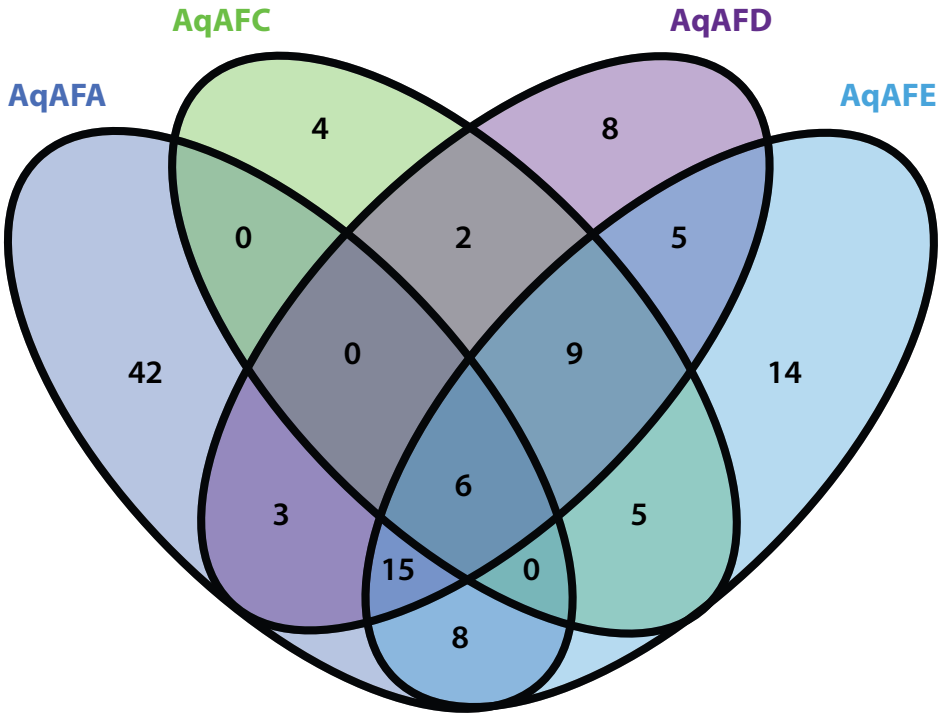


Figure 3.9 Potential coexpression of *AqAFs* and other genes
Venn diagram summarising the suite of genes exhibiting similar expression patterns to one or more of *AqAFA, AqAFC, AqAFD* and *AqAFE*. No genes exhibited a statistically significant correlation with *AqAFB* or *AqAFF*, so these genes are not included here.

significantly highly correlated with expression of *AqAFE*, but not correlated with each other. In total, expression of 122 unique, non-*AqAF* genes was significantly correlated with the expression of one or more *AqAF* gene (Figure 3.9, Appendix 3.3). These genes are henceforth referred to as being 'co-expressed' with the *AqAFs*; this should not be taken to imply co-regulation or shared function between and within the *AqAFs* and other genes of interest. Six genes correlated with all four of *AqAFA, C, D* and *E* (Figure 3.9, Appendix 3.3). Each *AqAF* is also correlated with a subset of genes not shared with any other AqAFs (AqAFA: n = 42, AqAFC: n = 4, AqAFD: n = 8, AqAFE: n = 14; Figure 3.9). The co-expressed show similar, but not identical, expression profiles across development to each other and to the *AqAFs* (Figure 3.10).

3.4.3 Analysis of statistically enriched GO terms

To investigate the putative functions of those genes co-expressed with the *AqAFs*, I performed two GO enrichment analyses to identify those GO terms that were over-represented in this gene list,

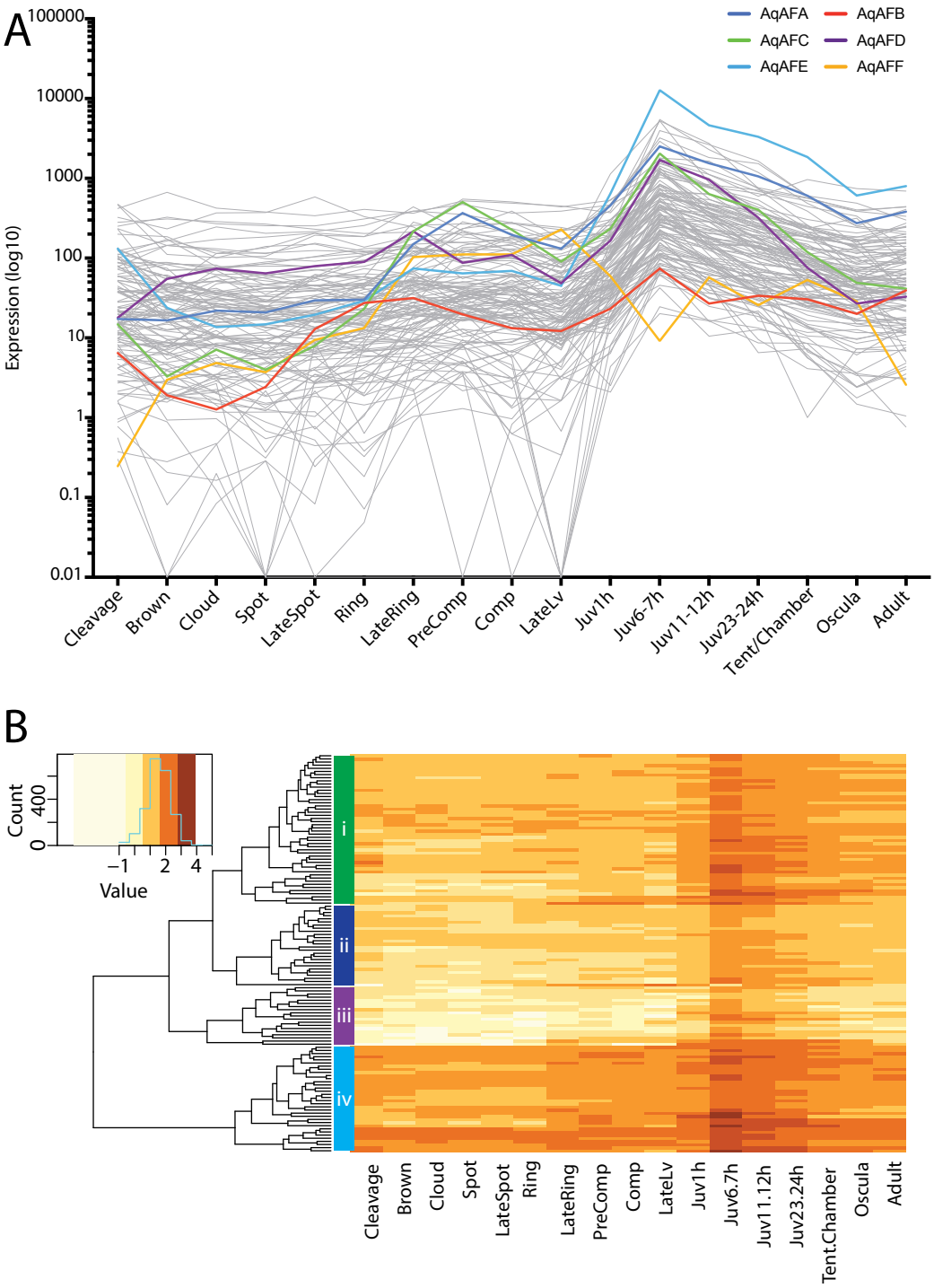


Figure 3.10 Expression of *A. queenslandica AFs* and other genes with correlated expression values

122 genes were identified that exhibited highly similar trends in expression pattern across development to the *AqAFs*. (A) The average log10 expression of each gene across development is shown in grey, while the *AqAFs* are coloured. (B) Heatmap depicting unscaled log10 expression levels of all genes across development. Coloured boxes (i - iv) highlight the four major clusters of genes, based on the dendrogram to the left. Cluster i contains *AqAFC*, ii contains *AqAFB* and *AqAFF* (which are not statistically correlated with any genes), and iv contains *AqAFA*, *AqAFD* and *AqAFE*. Members of each gene cluster are listed in Appendix 3.3

relative to the rest of the genome. For the first analysis, I analysed the GO terms associated with all genes identified as exhibiting correlated expression with the *AqAFs.* Of the 36 total enriched GO terms (Appendices 3.4, 3.5), 14 terms were associated with a single pair of sequences (Aqu2.1.41568_001 and Aqu2.1.41569_001), which were identified as *TGF-β receptor 1* genes (Appendix 3.3). Three of these enriched GO terms were TGF-β ligand- or receptor-binding related, two were associated with SMAD functionality, and eight represented developmental terms not apparently relevant to sponge biology (e.g. neuron fate commitment, palate development, etc.; Appendices 3.4, 3.5). To better analyse the enriched GO terms associated with other co-expressed genes, the GO enrichment analysis was repeated with the TGF-β receptor 1 genes omitted. This analysis produced a smaller list of highly related GO terms (Figure 3.11), and indicated that cell signalling genes are abundant amongst the set of genes co-expressed with the AqAFs.

3.4.4 Identity assignment to genes of interest

Genes were manually categorised based on their sequence homologues, domain architecture, and GO terms. Selected categorised genes are listed in Table 3.2, full details are provided in Appendix 3.3. This categorisation further emphasises that signal pathway, extracellular matrix and protein regulation molecules are co-expressed with the *AqAFs*.

3.5 Discussion

A. queenslandica allogeneic competency develops approximately two weeks after the commencement of settlement and metamorphosis. At this point, individual juveniles lose the ability to fuse with conspecifics, and chimeras undergo a cell partitioning process whereby each individual within a chimera contributes to the formation of different cell types (Gauthier and Degnan 2008). The molecular basis of juvenile allorecognition at 2 wpm, and of the transition to allogeneic competency, remains unexplored. In adults, the sponge-specific proteoglycan AF complex is involved, at least in part, in different types of self-nonself recognition behaviour. AFs play a direct functional role in the species-specific reaggregation of dissociated sponge cells (Moscona 1968; Humphreys 1970; Henkart et al. 1973; Müller and Zahn 1973). The *C. prolifera* AFs also respond to conspecific allorecognition challenge; expression of MAFp3 and MAFp4 is upregulated in response to auto- and allogeneic tissue contact (Fernàndez-Busquets et al. 1998), and AF molecules localise to the point of contact between

allogeneic tissue grafts (Fernàndez-Busquets et al. 1998; 2002). The precise mechanism/s of AF action in sponge tissue grafts has not been well characterised. However, studies of reaggregating sponge cells in the demosponge *Geodia cydonium* have revealed a capacity for AF-mediated cell signalling through control of protein kinase C, Ras and calcium activity (reviewed by Müller et al. 1990), suggesting these processes may also regulate the response to tissue contact.

In this chapter, I sought to characterise the expression profiles of the six *AqAF* genes across *A. queenslandica* development. In light of the proposed allorecognition role for the AFs, I hypothesised that activation of the sponge allorecognition system at 2 wpm may be triggered by the initiation of *AqAF* expression. This led to the prediction that *AqAF* expression would not be observed in the early stages of sponge development. The findings that the six *AqAF* genes are expressed at very high levels at all stages of *A. queenslandica* development, and that the expression profiles of these genes are correlated with those of a suite of cell signalling and other developmentally important genes, do not appear to support the hypothesis for a role for the *AqAFs* in triggering allogeneic competency. However, the lack of expression data for >3 dps juveniles, and the absence of functional studies, does not allow full refutation of the idea that the *AqAFs* drive the activation of allorecognition capabilities at 2 wpm, further to potential separate role/s for these genes in early development.

3.5.1 Possible explanations for *AqAF* developmental expression

Two alternative hypotheses are raised by the findings presented in this chapter. First, it may be the case that sponge allorecognition does indeed exist in some previously-undocumented form in early sponge development, and that the *AqAFs* are involved in this allorecognition process in some

Figure 3.11 Treemaps of other enriched GO terms *(See over page)*

Each section represents the statistically enriched GO terms (for Biological Process and Molecular Function) associated with the list of genes potentially coexpressed with the AqAFs, with the two TGF-β type 1 genes (Aqu2.1.41568_001 and Aqu2.1.41569_001) removed prior to performing the enrichment analysis. Each coloured box represents an enriched GO term, and the list of accession numbers for genes associated with each GO term are listed. Venn diagrams show the number of shared genes between adjacent GO terms. The bottom left box for the biological process section contains two GO terms identified as redundant by Blast2Go. Those genes annotated with the "regulation of phosphate metabolic process" GO term are highlighted by an asterisk. The two GO terms for this box are shown separately for all relevant Venn diagrams, with "regulation of phosphate metabolic process" in green and "regulation of GTPase" activity in blue.

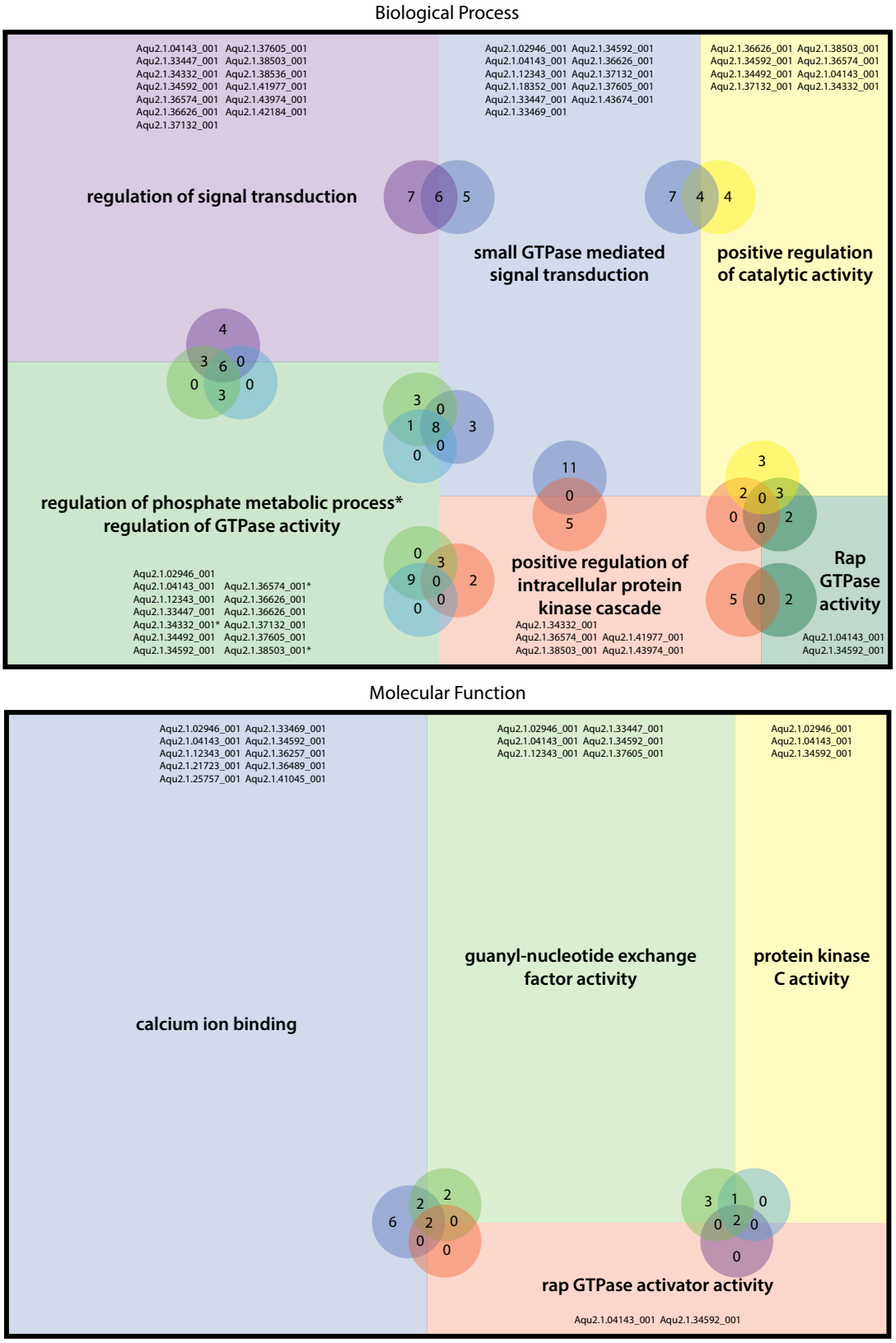


Figure 3.11 Treemaps of other enriched GO terms
(legend on previous page)

Table 3.2 Selected genes of interest co-expressed with the *AqAFs*

(Part 1 of 2)

Extracellular matrix molecules	
paxillin-like isoform x2	Aqu2.1.36574_001
calcium and integrin-binding protein 1	Aqu2.1.41045_001
hyaluronan mediated motility receptor (RHAMM)	Aqu2.1.14715_001
talin-1-like isoform x1	Aqu2.1.38632_001
talin-2 isoform x1	Aqu2.1.12470_001
Collagen alpha-2 chain	Aqu2.1.32089_001
G-protein coupled receptors (GPCRs)	
5-HT7 receptor	Aqu2.1.22312_001
gamma-aminobutyric acid (GABA) type b receptor subunit 2-like	Aqu2.1.39154_001
low quality protein: probable g-protein coupled receptor 112	Aqu2.1.36489_001
GTPase activating proteins (GAPs)	
ARF-GAP1	Aqu2.1.37132_001
ARF-GAP2	Aqu2.1.36626_001
rho gtpase-activating protein 6 isoform x4	Aqu2.1.34492_001
Protein kinases and related proteins	
serine/threonine-protein kinase TAO1-like	Aqu2.1.34332_001
MAP2K2	Aqu2.1.38503_001
protein tyrosine kinase	Aqu2.1.41977_001
tyrosine-protein kinase 223-like	Aqu2.1.43528_001
mob kinase activator 1a isoform x1	Aqu2.1.42093_001
Ras superfamily small GTPases	
ras guanyl-releasing protein 3	Aqu2.1.02946_001
ras guanyl-releasing protein 3	Aqu2.1.04143_001
ras-related protein rab-3b	Aqu2.1.18352_001
ras guanine nucleotide exchange factor	Aqu2.1.33447_001
ras guanyl-releasing protein 3	Aqu2.1.34592_001
rho-related gtp-binding protein	Aqu2.1.43674_001
ef-hand calcium-binding domain-containing protein 4b	Aqu2.1.33469_001
TGF-B signaling pathway	
tgf-beta receptor type-1	Aqu2.1.41568_001
tgf-beta receptor type-1	Aqu2.1.41569_001
Transcription factors	
tcf lef transcription factor	Aqu2.1.43974_001
t-box transcription factor tbx5-a-like	Aqu2.1.27488_001

Table 3.2 Selected genes of interest co-exprssed with the *AqAFs*
(Part 2 of 2)

Ubiquitination	
Kelch-like protein 20 (KLHL20)	Aqu2.1.20837_001
ubiquitin carboxyl-terminal hydrolase 33 isoform x2	Aqu2.1.36843_001
nedd8-conjugating enzyme ubc12	Aqu2.1.23767_001
e3 ubiquitin-protein ligase hecw2-like	Aqu2.1.27066_001
protein fem-1 homolog c	Aqu2.1.43650_001
f-box only protein 7	Aqu2.1.38681_001
Wnt signalling pathway	
nephrocystin-3	Aqu2.1.32091_001
frizzled-B	Aqu2.1.39914_001
nucleoredoxin	Aqu2.1.39833_001
tcf lef transcription factor [see also – transcription factors]	Aqu2.1.43974_001

way. Second, the *AqAFs* may instead facilitate the normal developmental and morphogenic processes occurring across sponge development.

a. Hypothesis: Sponge allorecognition may be active earlier than previously reported

Given the involvement of the AFs in adult sponge allorecognition, I originally proposed that the onset of *AqAF* expression triggers, and would therefore correlate with, the initiation of allogeneic competency. This was not found to be the case, as the *AqAFs* are active from the earliest developmental stage surveyed (the cleavage-stage embryo). However, since *AqAF* expression is correlated with adult allorecognition functionality, the inference could be drawn that the expression of the *AqAFs* in early sponge development indicates the existence of allorecognition functionality earlier than previously reported. No evidence for functional allorecognition during development has been reported from *A. queenslandica,* with larvae, postlarvae and >2 wpm juveniles capable of fusion with conspecific individuals (Gauthier and Degnan 2008). Thus, in light of the lack of evidence for allorecognition phenomena in the early sponge, this hypothesis seems improbable.

b. Hypothesis: The AqAFs may play a novel role in sponge development

The *AqAFs* are dynamically - but consistently very highly - expressed at all stages of sponge development. *AqAF* expression correlates with the expression of a suite of genes which function in

cell signalling, morphogenesis and other key developmental roles. Therefore I suggest that the *AqAFs* represent a suite of sponge-specific developmental or morphogenesis molecules, with complementary allorecognition roles arising later in sponge development. The hypothetical developmental function/s of the *AqAFs* may operate in tandem with some of the genes exhibiting correlated expression profiles to the *AqAFs*. Of the 122 correlated sequences identified here, a subset of notable genes is described in further detail below. Possible relationships between the genes of interest and the *AqAFs* are discussed.

3.5.2 Notable genes of interest that are co-expressed with the AqAFs

a. The Wnt pathway

Three members of the Wnt/β-catenin (canonical) signalling pathway - Frizzled (FrzB), Axin and TCF - are encoded by genes co-expressed with the *AqAFs*. The metazoan Wnt/β-catenin signalling pathway regulates numerous developmental processes, which in bilaterians include the establishment of axial and segment polarity, limb formation and organ development (reviewed by Cadigan and Nusse 1997). The *A. queenslandica* genome encodes all key elements of the Wnt/β-catenin signalling pathway (Adamska et al. 2007; Richards 2010). Analysis of the spatial expression patterns of these molecules in embryogenesis has suggested a role for the Wnt/β-catenin specification of the sponge anterior-posterior axis, and of the two tissue layers that form during gastrulation (Adamska et al. 2007; 2010). The Wnt/β-catenin pathway also appears to play a role in formation or remodelling of the sponge aquiferous system, because chemical deregulation (i.e. global activation) of the pathway in the homoscleromorph sponge *Oscarella lobularis* has been shown to trigger the ectopic formation of ostia in adults (Lapébie et al. 2008).

Figure 3.12 Molecular associations of co-expressed genes →

See over page

(A-B) Bilaterian Wnt/β-catenin (A) and TGF-β (B) signaling pathways in bilaterians. Known functional interactions are depicted. Phosphorylation events are depicted with circle-ending arrows, ubiquitination events are represented with triangle-ending arrows. Pathway adapted from (2010). (C) A simplified focal adhesion complex. Structures adapted from (Hammerschmidt and Wedlich 2008). (D) A hypothetical representation of the putative AF-RHAMM interaction that occurs via the HA-like molecule incorporated into the AF complex. For genes that are co-expressed with the *AqAFs* across sponge development, the encoded molecules are shown in pink. Molecules encoded in the *A. queenslandica* genome are shown in dark grey, molecules that are absent are in light grey. AF molecules are shown in orange. The solid curved line in each diagram represents the cell surface, and the area below the curve represents the cytoplasm. The dashed lines represent the cell nucleus.

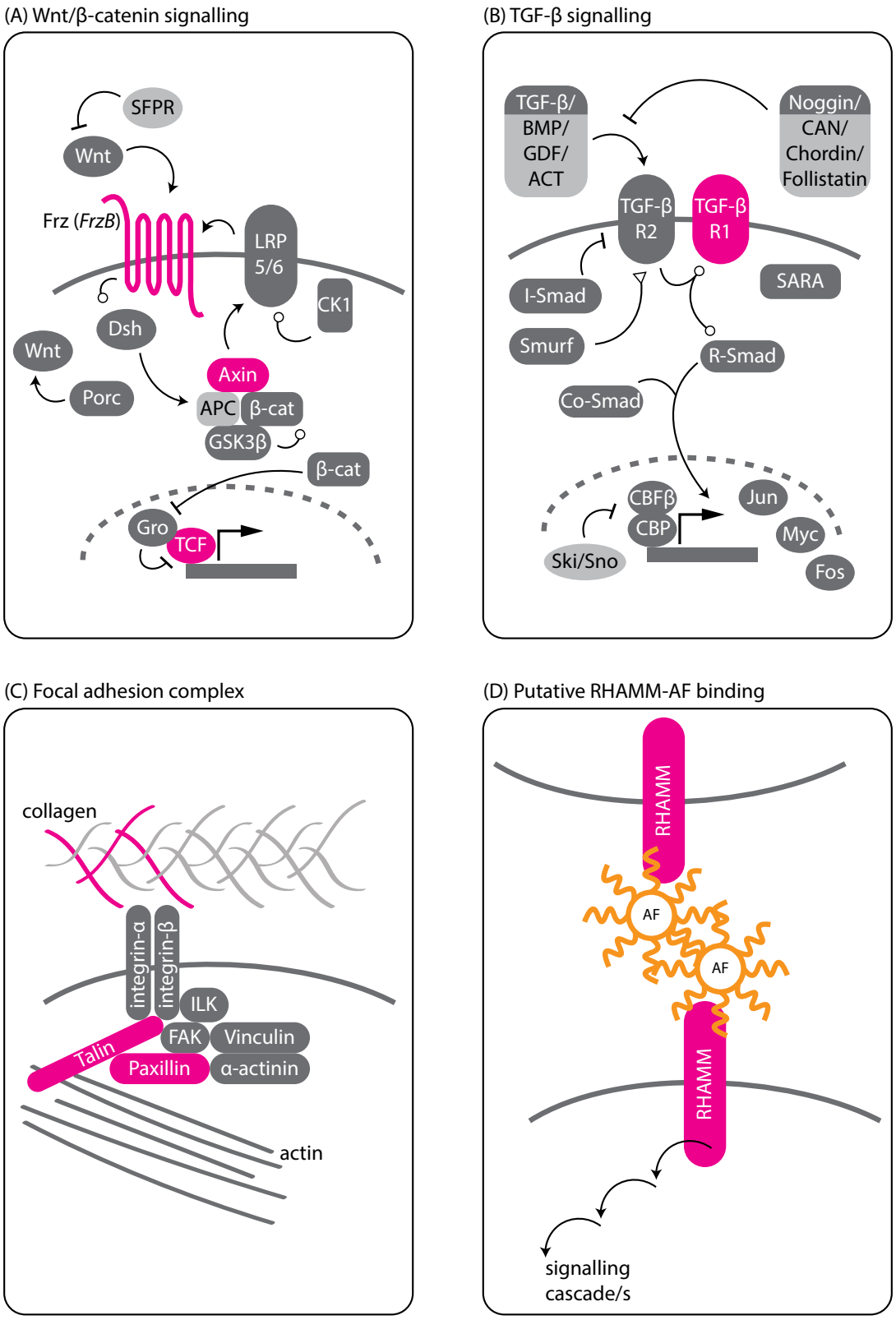


Figure 3.12 Molecular associations of co-expressed genes
(Legend on previous page)

The bilaterian Wnt/β-catenin pathway (Figure 3.12a) is activated by the binding of Wnt to form a receptor complex with the receptors Frizzled and LRP5/6. The *A. queenslandica* genome encodes two Frizzled receptors, Frizzled A and B (Adamska et al. 2010), of which only Frizzled B is co-expressed with the *AqAFs*. Wnt binding triggers a signalling cascade which ultimately leads to the release of β-catenin from an inhibitory complex that includes the scaffolding protein Axin. Free β-catenin translocates to the nucleus. Here, the transcription factor TCF/LEF (encoded in *A. queenslandica* by a single gene, *TCF*) is repressed by Groucho; β-catenin displaces Groucho, allowing TCF/LEF to drive the transcription of Wnt pathway target genes (for a more detailed review of the Wnt/β-catenin pathway see Saito-Diaz et al. 2013).

Two non-canonical Wnt pathways (the planar cell polarity [PCP] (McEwen and Peifer 2000) and Wnt-Ca^{2+} (Miller et al. 1999) pathways) operate alongside the canonical Wnt/β-catenin pathway in bilaterians. Like Wnt/β-catenin signalling, both non-canonical pathways are activated by the binding of a Wnt ligand to a Frizzled receptor, and involve the Dishevelled signalling molecule. However, the downstream molecules in each pathway differ from each other and from those in the canonical pathway. Key members of each non-canonical pathway are absent from the *A. queenslandica* genome, implying that these pathways do not function in sponges (Adamska et al. 2010). However, based on the expression patterns of certain Wnt pathway components in embryonic development, the possibility exists that an ancestral or derived non-canonical Wnt pathway may indeed operate in the sponge (Adamska et al. 2010). In particular, FrzB is a possible candidate receptor for this hypothetical non-canonical Wnt pathway (Adamska et al. 2010).

The co-expression of the *TCF* transcription factor gene with the *AqAFs* is particularly intriguing, in light of earlier reports suggesting an involvement of TCF, and therefore perhaps of the Wnt/β-catenin pathway, in allorecognition. TCF expression in the demosponge *Suberites domuncula* is upregulated following allogeneic, but not autogeneic, contact in both tissue grafts and dissociated cell reaggregation experiments (Müller et al. 2002). Application of the human immunosuppressant drug FK506 inhibits the rejection response in both types of allogeneic challenge experiment (Müller et al. 2001; 2002), and prevents TCF upregulation in reaggregating allogeneic cells (the effect on TCF in treated tissue grafts was not tested) (Müller et al. 2002). These results may indicate that the Wnt/β-catenin pathway

is activated downstream of the sponge allorecognition response, and suggest a functional relationship between the AFs as "frontline" allorecognition molecules, and the downstream signalling pathway. Developmental co-expression of the *AqAFs* and *TCF*, even in the absence of allogeneic competency and challenge, may indicate that the putative functional link between these systems endures from a cooperative relationship that emerges early in sponge development.

b. The TGF-β signalling pathway

Two *TGF*-β *receptor type 1* genes are co-expressed with the *AqAFs* across development. The TGF-β pathway is a metazoan innovation (Huminiecki et al. 2009; Richards 2010), and is a key player in developmental cell signalling processes. The bilaterian role of the TGF-β signalling pathway has been well-studied, and is associated with a range of processes including specification of the embryonic axes and germ layers, organogenesis, formation of the Spemann organiser, epithelial-to-mesenchymal transition, and wound repair (reviewed by Wu and Hill 2009). In *A. queenslandica,* the TGF-β signalling pathway is thought to work in cooperation with Wnt/β-catenin signalling to specify axial polarity during embryogenesis (Adamska et al. 2007).

TGF-β signalling is initiated by ligand-receptor binding (Figure 3.12b). Two broad sub-families of pathway ligands exist - the TGF-βs and the BMPs (Shi and Massagué 2003). The *A. queenslandica* genome encodes eight TGF-β ligands, but no BMPs (Srivastava et al. 2010). TGF-β receptors are serine-threonine kinases. Five receptor genes are present in the *A. queenslandica* genome - three of type 1 and two of type 2 (Srivastava et al. 2010; Conaco et al. 2012). Ligand-receptor binding triggers the phosphorylation of SMAD proteins. These subsequently translocate to the nucleus, and by interact with various cofactors or transcription factors, regulate the expression of TGF-β signalling target genes. The TGF-β signalling pathway is reviewed in depth by Massagué (1998).

The TGF-β pathway has not been implicated in sponge allorecognition processes to date. It is therefore not currently possible to infer whether or not TGF-β signalling has a functional relationship to the AFs in normal or immunologically challenged sponges, which could explain the co-expression of the *AqAFs* and two TGF-β receptor type 1 genes across development. Further research is required to explore this finding further.

c. The focal adhesion complex

Several focal adhesion complex genes are co-expressed with the *AqAFs* across development. For example, components of the integrin-linked focal adhesion complex - *Paxillin* and two *Talin* genes (*talin1*- and *2-like*) - were identified in the present study. Focal adhesions represent sites where the extracellular matrix is linked to the actin cytoskeleton via integrin receptors and the intracellular focal adhesion complex; these regions are important for regulating the interplay between locomotion and substrate adhesion (Figure 3.12c). Paxillin is a scaffold protein component of the cytoplasmic focal adhesion complex. Paxillin is an important regulator of Rho GTPases (Price et al. 1998), and is also associated with ARF-GAPs (ADP-ribosylation factor GTPase activating proteins), which regulate ARF GTPases (Turner et al. 2001). Several Rho GTPase and ARF-GAP genes are co-expressed with the *AqAFs* across development. However, it is unknown whether these are functionally equivalent with those that interact with Paxillin in bilaterians. Talin allows the physical linkage of integrins to actin, but is also important for inside-out integrin activation (Nayal et al. 2004). This activation is enhanced by an association between talin and the membrane phospholipid PIP_2 (phosphatidylinositol bisphosphate), which accumulates, for example, following cellular binding to fibronectin (McNamee et al. 1993). It is interesting to note that PIP_2 synthesis has also been shown to be triggered by AF-induced reaggregation of dissociated *G. cydonium* cells (Müller et al. 1987), and that the breakdown products of PIP_2 (inositol triphosphate [IP_3] and diaclyglycerol [DAG]) act as second messengers that control cellular calcium and active protein kinase C levels. These play a necessary role in the downstream response to AF binding (Müller et al. 1987; 1990). One of the two talin genes identified here, talin 1-like (Aqu2.1.38632_001) is situated close to *AqAFE* in the genome. These two large genes are separated by a 5.8 kb region that encodes acyl coA desaturase and sphingosine-1-phosphate phosphatase (data not shown).

It appears that the AFs interact with integrin receptors during cellular reaggregation, and possibly trigger downstream integrin-mediated signalling pathways. The *C. prolifera* MAFp3 sequence encodes an RGD binding sequence (Fernàndez-Busquets and Burger 2003), as does the Wreath domain-encoding portion of *AqAFD* (data not shown), suggesting that these proteins can physically interact with integrin. Addition of an RGD peptide, which binds β-integrin, to dissociated *S. domuncula* cells blocks AF-mediated reaggregation, and appears to mimic the downstream signalling effects of AF-cell binding

(Wimmer et al. 1999b). Finally, autograft fusion in *G. cydonium* results in the upregulation of integrin transcription (Wimmer et al. 1999a). The three lines of evidence discussed here – the potential role of integrin signalling win allorecognition, the developmental co-expression of *talin1-* and *2-like* and *paxillin* with the *AqAFs*, and the possible genetic linkage of talin1-like with the main *AqAF* locus – suggest that the *AqAFs* may be associated with focal adhesion functionality during sponge development.

d. The hyaluronan-mediated motility receptor

Hyaluronan (HA) is a large extracellular matrix glycosaminoglycan that also localises intracellularly (Evanko and Wight 1999). HA, and HA-receptor binding, promotes a diverse set of biophysical and biochemical states during development, normal cell physiology, and in cancers, by mediating cellular behaviours such as proliferation, location, cytoskeletal organisation and signal transduction (Toole 2001; Turley et al. 2002; Vigetti et al. 2014). A key HA-binding receptor, RHAMM (hyaluronan-mediated motility receptor) (Turley 1982), localises both to the cell surface (Crainie et al. 1999) and intracellularly (Entwistle et al. 1996; Assmann et al. 1999; Lynn et al. 2001). Different cellular processes are regulated by differentially-localised RHAMM receptors. For example, intracellular RHAMM receptors are associated with cytoskeletal assembly processes, while cell surface RHAMM mediates kinase and other signalling pathways, the dis/assembly of focal adhesions, cell motility, and other processes (Turley et al. 2002).

The *Clathria prolifera* MAFp3 cDNA encodes a HA-binding motif (Fernàndez-Busquets et al. 1996); binding studies with biotinylated HA have revealed that this motif is indeed functional (Kuhns et al. 1998). HA-binding motifs are also predicted within the coding sequences of *AqAFA* to *AqAFE* (Appendix 3.6), although these have not been functionally tested. Atomic force microscopy of the *C. prolifera* AF core has identified the presence of an HA-like molecule that appears to join the MAFp3-encoded arm subunits to the MAFp4-encoded central ring (Jarchow et al. 2000). The *C. prolifera* AF can thus be tethered to RHAMM *in vitro*, via the incorporated HA molecule (Kuhns et al. 1999). Treatment of purified *C. prolifera* AF with hyaluronidase (HAase) blocks this binding ability (Kuhns et al. 1999) and causes the disassembly of the sunburst-like AF structure (Jarchow et al. 2000).

The RHAMM-HA^{AF} binding ability raises a potential mechanism by which the AFs could mediate cell signalling and motility in the sponge (Figure 3.12d) (Kuhns et al. 1998). Since the AFs occur extracellularly, it is likely that this would occur via cell surface RHAMM signalling. It may be the case that RHAMM-HA^{AF} binding, and subsequent signalling, is involved in AF-mediated allorecognition processes. If this were the case, RHAMM-HA^{AF} binding would require individual-level specificity, to prevent heterologous AF-receptor binding. To the best of my knowledge, this has not been reported in sponges or other systems. However, as HA may take on different conformations with variable receptor specificity (Day and Sheehan 2001), this is not completely implausible.

Regardless of any immunological role of RHAMM-HA^{AF} binding, the co-expression of the *AqAFs* with the *RHAMM* gene across development lends further support to the hypothesis that AF-HA-RHAMM interactions moderate non-allogeneic biological processes. As stated above, it appears most likely that this interaction would occur via the cell surface RHAMM pathway; therefore this putative relationship may drive the cellular motility and reorganisation processes that occur during sponge development.

e. Regulation of phosphorylation and ubiquitination

The *AqAFs* are co-expressed with a number of genes putatively involved in de/ubiquitination signals and the transfer of phosphate molecules. The latter category includes genes identified as GTPase activating proteins (GAPs), small GTPases, and protein kinases. The control of ubiquitination and phosphorylation is an important mechanism in developmental and homeostatic cell signalling, as these signals are used in most signalling pathways, with functions including the activate and deactivate protein functionality, and specification of binding strength between molecular targets.

f. Scavenger receptor cysteine-rich domain-containing proteins

A candidate aggregation receptor (AR) gene has been cloned from *G. cydonium* (Blumbach et al. 1998). The longest gene product for this sequence encodes fourteen SRCR (scavenger receptor cysteine-rich) domains, six Sushi domains, and a transmembrane domain. Alternatively spliced isoforms, one encoding twelve SRCR domains and a transmembrane domain, the other ten SRCR domains only, have also been identified (Pancer et al. 1997). No *A. queenslandica* AR has been identified using functional

or genomic studies, although a large number of SRCR domain-encoding genes are encoded within the *A. queenslandica* genome (B. Yuen, personal communication). Two such genes were found to be co-expressed with the *AqAFs* across development; one encodes a signal peptide, five SRCR domains and a transmembrane domain, the other encodes three SRCR domains and a Protein Tyrosine Kinase domain (Appendix 3.3). The former sequence resembles the mid-length splice variant of the *G. cydonium* AR. However, no functional information is available to assign either sequence as a candidate AR.

3.5.3 Proposed experiments

In this chapter I have demonstrated that the *AqAFs* exhibit a shared expression pattern with 122 other *A. queenslandica* genes. However, the potential functional relationship between the *AqAFs* and the genes identified here has not yet been explored. I therefore propose a series of experiments to investigate this question further.

a. Spatiotemporal expression of AqAFs and other genes

The *AqAFs* are very highly expressed across sponge development. However, the spatial localisation of expression of these genes across development remains unexplored. Therefore, *in situ* hybridisation of the six *AqAF* genes in embryos, larvae, postlarvae and adults would provide valuable insight into the putative developmental role/s of the *AqAFs*. It is currently unknown whether the six *AqAF* genes participate in complementary (i.e. expressed and functioning together) or distinct (i.e. expressed and/or functioning separately) processes; therefore, it is of particular interest to determine whether the six *AqAF* genes are expressed in the same or different body regions, and cell types, to one another.

Investigation of the spatial expression patterns of other genes co-expressed with the *AqAFs* would provide valuable information in two ways. First, expression tracking of these genes would allow the detection of correlations in expression patterns with the *AqAFs* across development. The list of genes of interest could then be partitioned according to whether or not their spatiotemporal expression patterns correlate with one or more *AqAF* genes. Second, while the spatiotemporal expression patterns of components of the Wnt/β-catenin and TGF-β pathways have been elucidated in developing embryo and, to a lesser extent, in free-swimming larvae (Adamska et al. 2007; 2010), their expression in the postlarval and juvenile sponge remains unexplored. As the list of co-expressed sequences contains a

number of developmentally important genes, this analysis would be of general interest to the research community, as it would provide a greater understanding of sponge developmental processes. As it is not practical to perform this analysis for all 122 co-expressed genes, a set of likely interesting gene candidates would have to be selected.

b. Does RHAMM bind the AFs in vivo and does this influence self-nonself recognition?

The AFs are believed to bind RHAMM via their incorporated HA-like molecules, suggesting that RHAMM could play a role in the allorecognition response in adult sponges. To test this hypothesis, I propose a two-part experiment. First, it is important to confirm that the RHAMM receptor and the AFs actually interact *in vivo* in *A. queenslandica*. This could be tested using a chemical cross-linking approach to detect protein-protein interactions, and could allow detection of interactions between the AFs and RHAMM and other receptors (Tang and Bruce 2009). The next step would be to test whether this putative interaction between the AFs and RHAMM is involved in allorecognition. Previous studies have used an antibody directed towards RHAMM to block RHAMM-HA^{AF} binding (Kuhns et al. 1997). It may, therefore, be valuable to test whether this antibody could affect the *A. queenslandica* graft response. Sponge auto- and allografts could be incubated with anti-RHAMM, and the effect on graft responses subsequently monitored. If a phenotypic effect on grafting was observed, analysis of the associated change in gene expression could reveal the effect this process had on the molecular graft response.

3.5.4 An evolving paradigm of AqAF developmental involvement?

The AFs are proposed to play an allorecognition role in adult sponges. This self-nonself recognition mechanism occurs via the homotypic AF-AF interaction, to allow adhesion of like cells. To a certain extent, this adhesion is a passive mechanism reliant on the adhesive properties of the AF molecules; this explains the species-specific nature of AF binding even in cell-free systems (Moscona 1968; Humphreys 1970; Henkart et al. 1973; Müller and Zahn 1973). However, it also appears that AF binding triggers an intracellular signalling cascade that regulates the cellular AF-binding response. For example, AF-AR binding in dissociated *G. cydonium* cells triggers an increase in intracellular calcium levels and an activation of protein kinase C; together these molecules stimulate intracellular signalling which

triggers processes including cell proliferation, protein phosphorylation, and increased transcription, translation and DNA replication (reviewed by Müller et al. 1990).

Immature *A. queenslandica* individuals do not acquire immunological competence until 2 wpm, but expression of the *AqAFs* is very high prior to this time. This finding, plus the co-expression of the *AqAFs* with a suite of developmentally important genes, raises the possibility that the AqAFs also play an important role during sponge development and metamorphosis. Functional testing is still required to confirm this suggestion, and to investigate the specific putative function/s of the AFs. It is likely that the putative AF developmental role is mechanistically similar to AF function in dissociated adult cells; namely, that the AFs mediate cell-cell interactions, and trigger intracellular signalling in response to this binding, to promote cell proliferation, migration, cell-matrix interactions, and/or cytoskeletal remodelling, in a developmental context.

The AF complex represents an elaborate conglomerate of multiple protein and glycan subunits. The *A. queenslandica* genome encodes six *AqAF* genes, five of which are equipped with the necessary elements to form the head and arm subunits of the core AF complex. It is currently unknown whether the different AF genes work separately, or cooperate within a single cell type, developmental stage or even AF complex. In their role as allodeterminants, the AFs are predicted to carry a high level of polymorphism to facilitate the individual specificity required of a molecule that recognises and discriminates between conspecifics Chapter 1.1.3). It appears that AF diversity is carried not only within the proteinaceous subunits, but also within the attached glycan residues (Fernàndez-Busquets and Burger 1997). Sponges may also be able to regulate AF activity via the differential glycosylation of the AF complex (Fernàndez-Busquets et al. 2002). Adding to this complexity is the suggestion that the AFs are capable of binding multiple types of receptor to mediate different cell processes. Such implicated receptors to date include the *G. cydonium* AR (Blumbach et al. 1998), integrins (Wimmer et al. 1999b), and RHAMM (Kuhns et al. 1999).

AFs therefore appear to possess at least five layers of variability: the potential to utilise different genes, gene sequence polymorphisms, glycan polymorphisms, differential glycosylation levels, and different AF-receptor binding combinations. Altering one or several of these aspects may allow the

modular regulation of cell-cell interaction and signalling functionality, and the fine-tuning of biological variables such as binding kinetics, specificity, and spatiotemporal functionality.

Some *AqAF* genes/variants may confer cell type, rather than, or in addition to, individual-level specificity. This could be useful in facilitating cell migration and patterning processes during development. The observation that the 2 wpm juvenile partitions cell types in an individual-specific manner (Gauthier and Degnan 2008), suggesting that the same molecule (either the AFs or others) confers both cell type- and individual-level specificity, may lend support to this idea.

In Chapters 2 and 3, I explored the normal genomic features of the *AqAFs*, and the expression profiles of these genes across *A. queenslandica* development. In Chapter 4, I investigate the potential contributions of two mechanisms – alternative splicing and nucleotide polymorphism – by which the *AqAFs* could be diversified, such as to generate the level of between-individual variability expected of a putative allorecognition molecule. I first took a PCR (polymerase chain reaction)-based approach to search for transcriptional length variants in a single *AqAF* gene (AqAFC), before embarking on a large-scale transcriptome survey for *AqAF* alternative splicing across *A. queenslandica* development. In the second half of this study, I investigated the amount and type of sequence polymorphisms present transcriptome-wide and in the *AqAF* genes across four adult individuals. I show that the *AqAFs* undergo intron retention to produce novel truncated AF forms, and that these genes display a high level of sequence polymorphism between individuals.

CHAPTER 4 - POLYMORPHISM IN THE *AMPHIMEDON QUEENSLANDICA* AGGREGATION FACTOR GENES

4.1 Abstract

Precise allorecognition reactions rely on the existence of an underlying polymorphic molecular system to facilitate nonself rejection. Such polymorphism could exist on the level of the genome, nucleotide, transcript, protein and/or molecular complex; the use of multiple mechanisms and differential regulation thereof potentially allows for more precise control over this diversification. Sponge aggregation factors (AFs) are putative allorecognition molecules and have been previously demonstrated to be highly polymorphic in the demosponge *Clathria prolifera*. However, as the full AF gene complement in this species is not fully resolved, it is not currently possible to determine the full extent of variation amongst the AFs between individuals and relative to the underlying genome sequences. Therefore, I sought to catalogue and characterise the level of AF nucleotide diversity and alternatives splicing in the model demosponge *A. queenslandica.* AF transcripts in this species exhibit multiple intron retention events, suggesting a role for the nonsense mediated decay pathway in AF activity regulation. A subset of intron retention events also introduce signal peptides to the resulting predicted protein sequence, suggesting the existence of a novel set of truncated AF proteins that may compete with full-length AFs for target substrate binding sites. The *A. queenslandica* AFs are also highly polymorphic at a nucleotide level, showing an over-representation of non-synonymous variants relative to the transcriptome as a whole. Therefore, the *A. queenslandica* AFs may use alternative splicing and nucleotide-level sequence variants - with or without the contribution of other mechanisms - to generate the between-individual variability required of putative allorecognition molecules.

4.2 Introduction

Self-nonself recognition reactions, regardless of the level of observation (i.e. species, individual, cell type etc.), occur as a three phase process involving detection of a neighbouring entity, recognition of this entity as self or nonself, and a discriminatory action that excludes self or nonself as appropriate

(Chapter 1.1.2). Three possible classes of recognition exist - self recognition (which occurs, for example, in the plant self-sterility system) (Nasrallah 2005), nonself recognition (seen, for instance, during fungal mating) (Hall et al. 2010) and self and nonself recognition (for example, in T cell-mediated immunity) (Boehmer and Kisielow 1990; Wu et al. 2009). In self recognition, only cells or molecules possessing 'labels' marking themselves as self are accepted (Burnet 1971; Coombe and Ey 1984; Boehm 2006). This is the simplest of the three possible recognition mechanisms, and therefore probably the most ancient.

Regardless of the exact mechanism used, the primary requirement for recognition is a capacity for highly precise decision making, to prevent costly errors during downstream discrimination (Tsutsui 2004). Precision requires an underlying genetic or molecular system that is sufficiently diverse to produce unique labels for each self unit (Hildemann 1979; Grosberg 1988; Tsutsui 2004). A key challenge for allorecognition reactions is the need to facilitate rejection between incompatible individuals that nonetheless share the same basic genome and thus a roughly identical complement of allorecognition molecules. Generation of between-individual gene product diversity may be controlled at the level of the genome (e.g. somatic recombination), nucleotide (e.g. sequence polymorphisms), transcript (e.g. alternative splicing or RNA editing), protein (e.g. post-translational modification or protein complex assembly), and/or molecular complex (e.g. the addition of non-protein moieties such as glycans).

In addition to their well-characterised role in species-specific cell reaggregation (Wilson 1907; Humphreys 1963; Moscona 1968; Humphreys 1970; Curtis and Van de Vyver 1971; Henkart et al. 1973; Müller and Zahn 1973), aggregation factors (AFs) have been implicated in the sponge allorecognition response (Fernàndez-Busquets and Burger 1999). For example, expression of *MAFp3* and *MAFp4*, which encode elements of the demosponge *Clathria prolifera* AF core structure appear to be upregulated in self and nonself tissue grafts (Fernàndez-Busquets et al. 1998), and MAFp3 protein has been shown to accumulate at the allograft interface (Fernàndez-Busquets et al. 1998). The AFs are therefore expected to possess the required properties of allorecognition molecules (Chapter 1.1.3-4), including undergoing diversification to allow between-individual recognition and downstream discrimination. Biochemical studies have revealed a role for AF complex glycan subunits in generating between-individual variability (Fernàndez-Busquets and Burger 1997), but a high degree of polymorphism has also been observed

at the nucleotide level within the *C. prolifera* AFs (Fernàndez-Busquets and Burger 1997; Fernàndez-Busquets et al. 1998). As the genomic complement of *AF* genes for this species has not yet been resolved, the extent of this polymorphism and the contribution of other possible mechanisms remains unknown. General mechanisms for creating transcript- and nucleotide-level differences between individuals are discussed briefly below – see also a review by Ghosh et al. (2011) – with particular attention payed to the known or hypothetical roles of these mechanisms in the sponge *AFs*.

a. Alternative splicing

Alternative splicing allows a single gene to produce multiple protein isoforms, which may be expressed together and/or in a context-specific manner. Allorecognition or other immune genes from a variety of invertebrate species undergo diversification by alternative splicing. For example, the ascidian *Botryllus schlosseri* histocompatibility-associated genes *fester* and *uncle fester* are both alternatively spliced (Nyholm et al. 2006; McKitrick et al. 2011), as is *alr1* from the colonial hydroid *Hydractinia symbiolongicarpus* allorecognition system (Rosa et al. 2010). Perhaps the most dramatic example of alternative splicing in the invertebrate immune system is the *Dscam* gene (Schmucker et al. 2000), which in the mosquito *Anopheles gambiae* undergoes alternative splicing with the potential to generate 32,000 unique transcripts, and has been implicated in the pathogen response mechanism of this species (Dong et al. 2006).

Alternative splicing may take several forms, including the use of alternative intron donor or acceptor sequences, intron retention, transcript initiation or termination within a canonical intron, exon skipping, or use of an alternative terminal exon (Figure 4.1). The most prevalent form of alternative splicing in non-eumetazoan eukaryotes - including

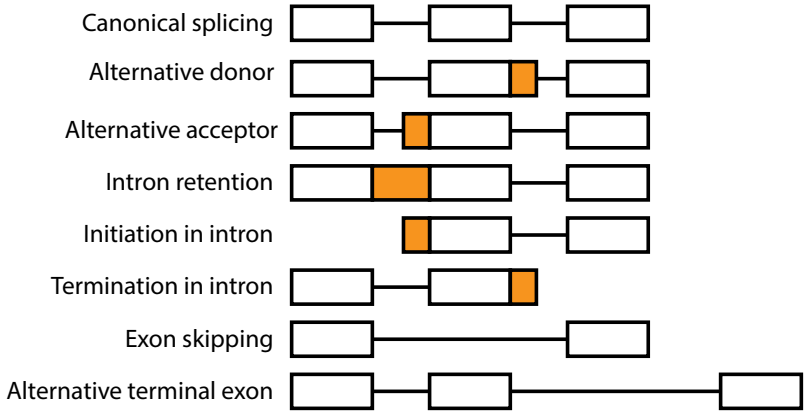


Figure 4.1 Types of alternative splicing
Each diagram defines an alternative splicing category, relative to a hypothetical canonical intron-exon organisation (top).

protists (McGuire et al. 2008), fungi (McGuire et al. 2008), plants (Kim et al. 2006; Wang and Brendel 2006; McGuire et al. 2008), choanoflagellates (Westbrook 2011), and sponges (S. Fernandez Valverde and B. Degnan, manuscript in preparation) - is intron retention, while exon skipping is least common in these taxa. Conversely, eumetazoans generally follow the reverse trend - exon skipping is the most frequently-observed splice change, while intron retention is the least common (Kim et al. 2006; Sugnet et al. 2004; McGuire et al. 2008). Therefore, it appears that a fundamental change in splicing regulation occurred at the metazoan-eumetazoan boundary.

The modular nature of the *Amphimedon queenslandica AF* (*AqAF*) genes may be a sign that alternative splicing acts to diversify these sequences. The *AqAFs* show a highly significant over-representation of symmetrical exons (i.e. in which exons are flanked by introns in the same phase), with all but one *AqAF* intron being in Phase 1 (Chapter 2.4.8). Exons in the *C. prolifera* AFs are also symmetrical, although all introns in these genes are exclusively in Phase 0 (Fernàndez-Busquets and Burger 1999). Alternative splicing, like exon shuffling, relies on symmetrical exons to maintain the transcriptional reading frame of the resulting transcript, so this feature of the *AqAFs* makes them plausible candidates for this means of diversification. The organisation of *AqAF* protein domains into uni- or multi-exon modules (Chapter 2.4.7) could potentially provide a further source of variation, if between-domain rearrangements, and the production of chimeric domain sequences, were to occur.

b. Nucleotide polymorphisms

The requirement for high levels of diversity within self-nonself recognition systems can favour the accumulation of many rare alleles within a population, effectively limiting compatibility to true instances of self or close kinship rather than random matches due to chance (Tsutsui 2004). Nucleotide-level variants are a common source of diversity amongst characterised self-nonself recognition molecules. For example, within the *B. schlosseri FuHC* histocompatibility locus, *BHF, fester, Hsp40-L, mFuHC* and *sFuHC* all exhibit high levels of nucleotide diversity between individuals (De Tomaso et al. 2005; Nyholm et al. 2006; Nydam et al. 2013a; 2013b; Voskoboynik et al. 2013). *H. symbiolongicarpus,* in which fusion rates of less than 5% have been recorded, also possesses a similarly highly polymorphic system. This species' two allodeterminant genes, *alr1* and *alr2*, both encode transmembrane proteins with hypervariable extracellular regions that are equipped with repeated domains, and possess codons

found to be under positive selection (Nicotra et al. 2009; Rosa et al. 2010). The rich allelic nature of these genes facilitates the aforementioned low rates of colony fusion - for example, Gloria-Soria et al. (2012) identified 198 unique allelic variants of *alr2* in a single study population. Reports of positive selection acting on genes from other self-nonself recognition systems - including the *Sp185/333* suite from the sea urchin *Strongylocentrotus purpuratus* (Terwilliger et al. 2006), the parasite defense gene *FREP3* from the freshwater snail *Biomphalaria glabrata* (Zhang et al. 2001), and various panaeidin antimicrobial peptide genes from the penaeid shrimp (Padhi et al. 2007) - further highlight the importance of nucleotide-level sequence variation for self-nonself recognition diversity.

The core components of the *Clathria prolifera* AF protein complex, MAFp3 and MAFp4, are coded by a contiguous mRNA transcript but appear to be cleaved post-transcriptionally to generate independent MAFp3 and MAFp4 protein subunits (Fernàndez-Busquets and Burger 1997; Fernàndez-Busquets et al. 1998; Jarchow et al. 2000). The sequences encoding MAFp3 and MAFp4 exhibit a high degree of nucleotide-level polymorphism. Sequence polymorphism in this species correlates with self-nonself decision making, with a 99.5% correlation between sponge tissue graft behaviour (fusion or rejection) and *MAFp3* restriction fragment length polymorphism (RFLP) profiles (identity or dissimilarity) between individuals (Fernàndez-Busquets and Burger 1997). MAFp4 also displays similar RFLP disparity between individuals. These findings suggest that the AFs possess the level of variability expected of molecules involved in the recognition stage of allorecognition.

c. RNA editing

While the presence of genomically-encoded sequence polymorphisms within a population is the more common and better understood example of nucleotide-level differences between individuals, an intriguing alternative exists in the form of RNA editing. RNA editing occurs post-transcriptionally, where a sequence is altered via nucleotide insertion, deletion or modification (Simpson 1996; Gott and Emeson 2000). RNA editing has been shown to play a role in *S. purpuratus* innate immunity, with post-transcriptional nucleotide changes adding an additional layer of complexity to the *Sp185/333* system (Buckley et al. 2008).

One of the most prevalent forms of RNA editing involves the deamination of adenosine residues in double-stranded RNA substrates into inosines (A-to-I editing) (Bass and Weintraub 1988; Wagner et al. 1989). A-to-I editing is mediated by ADAR (adenosine deaminase acting in RNA) editing molecules; editing of other nucleotide substrates is performed by other molecules. ADAR editing can modify and regulate gene product output, for example via codon modification (as inosines are interpreted as guanosines by the cell) or influencing splice site and small RNA functionality (Nishikura 2010). ADAR family members exist in bilaterians and cnidarians (Jin et al. 2009; Keegan et al. 2011), but while ADARs were recently identified in the genome of the ctenophore *Pleurobrachia bachei* (Moroz et al. 2014), previous studies have not identified this class of molecules in *A. queenslandica* or other sponges (Keegan et al. 2011). In Chapter 5, I revisit this issue and report that ADAR protein family members are indeed present in the earliest branching metazoan lineages, including numerous sponge species. RNA editing is therefore a potential fourth mechanism by which the *A. queenslandica* AFs could acquire between-individual diversity.

In this chapter, I investigate the potential contributions of alternative splicing and nucleotide polymorphisms to *AqAF* diversification. First, I present the results of a survey of *AqAF* transcripts from individuals spanning the *A. queenslandica* lifecycle (precompetent larvae, competent larvae, juveniles and adults) with the goal of determining whether the *AqAFs* undergo alternative splicing in a normal, immunologically unchallenged context. I show that the *AqAFs* undergo multiple intron retention events across the six *AqAF* genes and across developmental time, and that a subset of these intron retention events is predicted to promote the transcription of novel short protein isoforms derived from the C-terminal end of the full sequence. Second, I examine the *AqAFs* at a nucleotide level, and document the amount and nature of the sequence polymorphisms that exist in four adult *A. queenslandica* individuals on both a transcriptome-wide and *AqAF*-specific scale. I show that putative sequence polymorphisms exist in five of the six *AqAF* genes in all individuals, and that changes predicted to cause non-synonymous amino acid changes are over-abundant relative to the frequency of these observed in the transcriptome as a whole. I conclude that the *AqAFs* display a degree of diversification that may facilitate between-individual self-nonself recognition in sponges.

4.3 Methods

4.3.1 Transcriptome-based analysis of alternative splicing

Four in-house transcriptome datasets were previously generated using a polyA-selection, 100 base pair, paired-end, stranded Illumina HiSeq 2000 protocol, and multiplexed with four libraries on a single lane of an Illumina flow cell. RNA for these datasets was derived from multiple precompetent larvae, competent larvae and juvenile *A. queenslandica* individuals and a single adult individual. Transcripts were assembled *de novo* and compared to the Aqu2.0 *A. queenslandica* gene models using a standard PASA (program to assemble spliced alignments) (Haas 2003) pipeline to identify and classify putative alternatively spliced transcripts. Tissue sample collection and RNA extraction was performed by A. Calcino, and read preparation, assembly and PASA annotation (including alternative splicing detection) was performed by S. Fernandez Valverde. Only the alternative acceptor, alternative donor, alternative exon, skipped exon, retained intron, initiation within an intron, or termination within an intron categories of splicing events were considered for downstream analyses. The nucleotide sequences of all unfiltered putatively alternatively spliced *AqAF* transcripts were extracted and manually compared to the Aqu2.1 genomic DNA (gDNA) and messenger RNA (mRNA) sequences using CodonCode Aligner version 3.7.1.1. Transcripts confirmed to alter *AqAF* structure were selected for further analysis. Sequence truncations were not inherently of interest unless these transcripts also contained a splicing event of interest.

4.3.2 PCR-based analysis of alternative splicing

a. Preparation of larval genetic material for the polymerase chain reaction

Thirty *A. queenslandica* larvae from multiple mothers were collected as described by Leys et al. (2008), allowed to develop for 10 hours post emergence (hpe), and preserved in RNA Later (Ambion) for later use. All centrifugations during RNA extraction were performed at 14,680 revolutions per minute (rpm). Preserved larvae were transferred to 200 µL Tri Reagent (Sigma) and ground to release RNA. An additional 50 µL Tri Reagent was added, and samples were left at room temperature for 5 minutes before centrifugation for 10 minutes at 4°C. The supernatant as collected, vigorously mixed with 25 µL bromochloropropane (BCP), left at room temperature for 15 minutes, and then centrifuged for 15 minutes at 4°C. The resulting top aqueous layer was combined with 62.5 µL each of isopropanol and high-salt precipitation solution (0.8 M sodium citrate, 1.2 M NaCl). After a 10 minute incubation

at room temperature, the sample was centrifuged for a further 10 minutes at 4°C. The supernatant was discarded and a standard 70% ethanol wash was performed on the pellet. Pellets were eluted in DNase and RNase-free distilled water (Gibco, Life Technologies). RNA was quantified using a NanoDrop spectrophotometer (Thermo Scientific) and run on a 1% TBE (Tris-Borate-Acetate) agarose gel to check sample quality.

b. Primer design

Oligonucleotide primer pairs were designed to amplify the middle portions of *AqAFC*. Primers were designed using Primer3 version 2.0.0 (Koressaar and Remm 2007) and Vector NTI Advance 10 (Invitrogen), and were supplied by Sigma-Aldrich. Three primer sets were designed to sit within exons, while the fourth forward primer (F39) was designed to sit within intron 18 (Figure 4.2). Full primer details are given in Table 4.1.

c. Polymerase chain reaction and product purification

AqAFC cDNA fragments were amplified by the polymerase chain reaction (PCR) using one of three reaction mixtures listed in Appendix 4.1. Reactions were run on a PCR thermocycler following the cycling conditions listed in Appendix 4.2. PCR products were visualised on a 1% TAE (Tris-Acetate-EDTA) agarose gel, before bands of interest were excised and DNA was extracted as described by Boyle and Lew (1995).

d. Cloning and sequencing

Purified PCR products were cloned by ligation into the pGEM-T easy vector using the pGEM-T-easy kit (Promega) following the manufacturer's directions. Competent XL1-Blue *Escherichia coli* cells were transformed by heat-shock (1 minute at 42°C) and grown overnight on LB-ampicillin (100 µg/mL ampicillin) agar plates that had been streaked with 0.75 mg each of X-gal and IPTG. Positive colonies were verified using PCR and prepared for sequencing using the Big Dye Terminator 3.1 Cycle Sequencing kit (Applied Biosystems) according to directions supplied by the Australian Genome Research Facility (AGRF). Sanger sequencing was performed by AGRF. Sequences were trimmed and aligned to the Aqu2.1 *AqAFC* gene in CodonCode Aligner version 3.7.1.1.

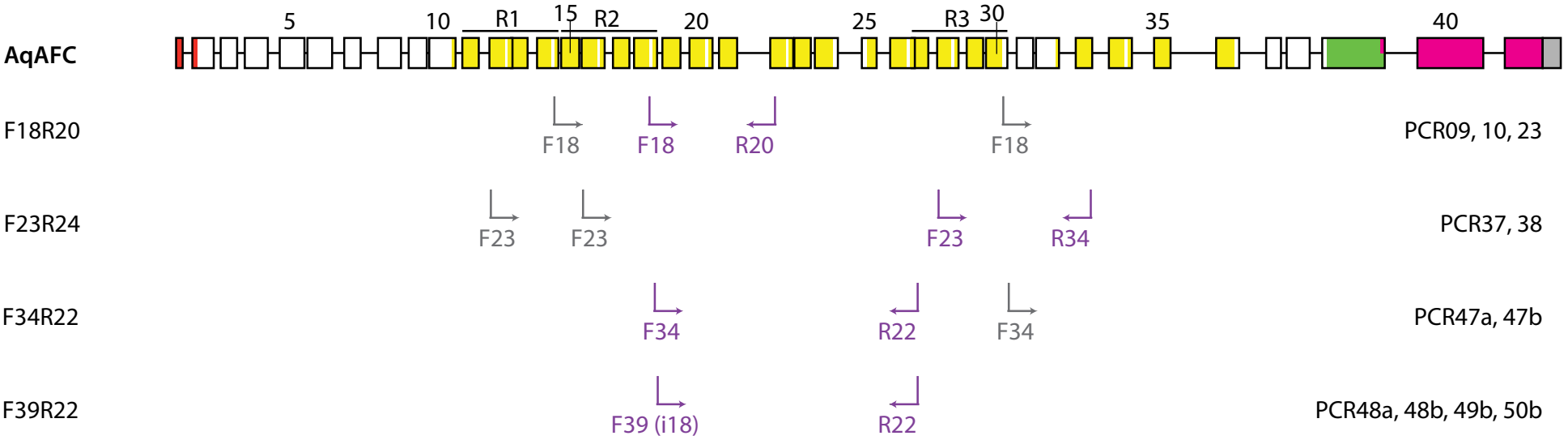


Figure 4.2 *AqAFC* primer locations
Binding sites for primers used in the *AqAFC* alternative splicing study. Primers binding unique regions could not be designed for some regions due to the presence of three highly similar repeat regions (R1 - R3) within AqAFC. Therefore, all possible binding sites are shown. Actual binding sites based on sequencing results are shown in purple. Naming codes for resulting alternatively spliced PCR products (as used in Figure 4.4) are listed to the right.

Table 4.1 Primer details for *AqAFC*

Primer Pair	Forward Primer Sequence	Fwd Tm (°C)	Reverse Primer Sequence	Rev Tm (°C)	Set Tm (°C)	mRNA product Length (bp)	Extension time
F18R20	TAGCTCGGATCAATTTGTTGA	62.1	TGAGTCATGCTGTCAGAAACG	64.1	61	1290	90 sec
F23R24	GGAGTTGATTATAATTTGCCCAGT	62.8	TGACAACAACAGCATCAGCA	64.3	62	991	90 sec
F34R22	TGCTGACAGTGCTACATCAA	60.7	TGACTGGGCTAGATCCTTCTTC	63.4	59	1411	100 sec
F39R22	ACCATTAGCAACTTGTTGTTCC	61.7	TGACTGGGCTAGATCCTTCTTC	63.4	59	1224	100 sec

4.3.3 Whole-transcriptome sequencing data for probabilistic variant detection

Whole-transcriptome sequencing data were prepared from four adult *A. queenslandica* individuals (sponges A to D). Sponges A and B were also used as control samples for the graft transcriptome analysis presented in Chapter 6; preparation of these samples is described in detail in Chapter 6.3.1-6.3.5. Sponge C was also used for the alternative splicing analysis described above (Chapter 4.3.1). Sponge D was prepared following a polyA-selection, 100 base pair, paired-end, unstranded Illumina HiSeq 2000 protocol and was run across an entire lane of an Illumina flow cell.

4.3.4 Probabilistic variant detection

The four adult individuals were examined to identify putative sequence variants. Trimmed sequencing reads were mapped to the Aqu2.1 gene model-annotated *A. queenslandica* genome in CLC Genomics Workbench version 6.5.1 using a similarity fraction value of 0.8 and default parameters for all other settings. The probabilistic variant detection tool was used to identify sequence variants, based on a diploid prediction model and using default parameters. Variants were annotated with exon numbers and their predicted effects on splice sites or encoded amino acids. Poorly-supported variant calls were filtered using CLC Genomics Workbench's filter marginal variants tool, run with default parameters. Variants mapping to the six *AqAF* genes were extracted for further analysis.

All statistical comparisons between conditions were performed using the paired T-test tool in GraphPad Prism version 6.0 for Mac (http://www.graphpad.com), using the percentages of each observation in each sponge dataset.

4.3.5 Haplotype reconstruction

Manual contig walks were performed to identify linked *AqAF* variants, and to ultimately reconstruct the two full-length *AqAF* alleles from each analysed individual (Figure 4.3). Read mapping results for the *AqAF* regions were manually examined. For every pair of adjacent predicted variants, the encoding reads were scanned to determine whether one or more reads existed that encoded both variants; this would indicate that they were linked on a single allele. This was performed for each subsequent set of variants, until a link could not be formed between a pair. Each contiguous region of linked variants (encoding the fragments of two alleles) was referred to as a 'haplotype block'.

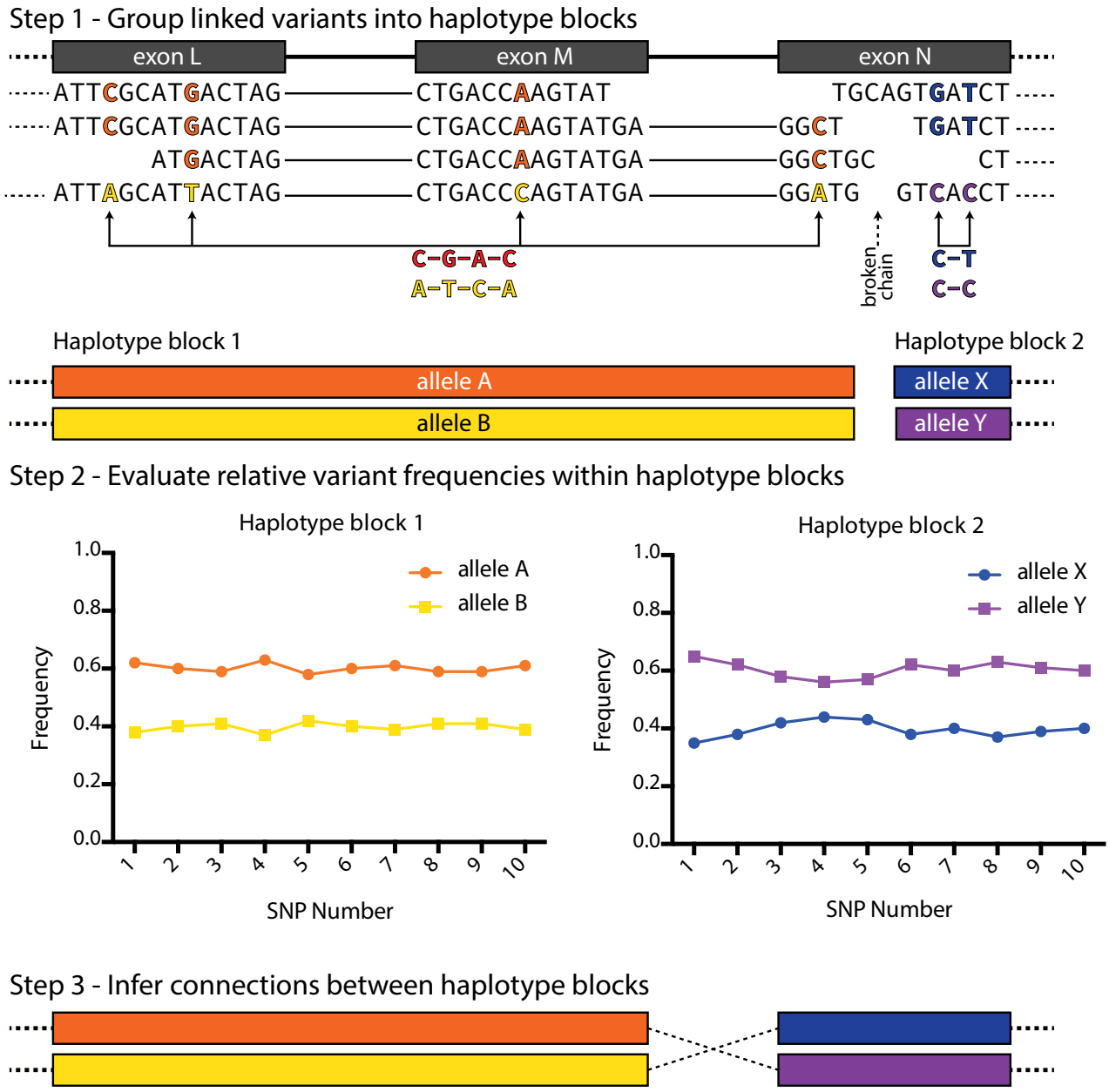


Figure 4.3 Allele reconstruction methods
Schematic showing a hypothetical worked example of full-length allele sequence reconstruction from individual variant site information. Briefly, all adjacent variants with sequence evidence for linkage were strung together into haplotype blocks, with two allele fragments per chain (Step 1). The frequencies of all variants within an allele fragment were averaged (Step 2), and allele fragments exhibiting similar average frequencies between adjacent haplotype chains were inferred to be linked (Step 3).

The probabilistic variant analysis output lists the proportion of reads encoding each nucleotide option per variant site. These values were averaged across each of the two allele fragments per haplotype block; in most cases one allele tended to be detected at a higher frequency than the other. This information

Table 4.2 Observed numbers of alternatively spliced *A. queenslandica AF* transcripts

		Pre-Competent larvae	Competent Larvae (RNA-Seq)	Competent Larvae (PCR)	Juvenile	Adult
AqAFA	IR	1	2		1	1
	SiI	1				
	EiI					1
AqAFB	IR	1				
	SiI					1
	EiI					4
AqAFC	IR	3	1	2**, 6		
	SiI	4	1	3	1	
	EiI			1	1	
AqAFD	IR	1	1			
	SiI	3	1		1	1
	EiI					
AqAFE	IR		2		2	
	SiI				2*	
	EiI		1		1	
AqAFF	IR		1**, 3		1	
	SiI					
	EiI					

PCL = pre-competent larvae, CL = competent larvae, Juv = juvenile, Ad = adult
IR = intron retention; SiI = starts in intron; EiI = ends in intron
* Unknown sequence; ** multiple events per transcript

was used to infer which allele fragments from neighbouring haplotype blocks were part of the same full length sequence. It was assumed that linked alleles from adjacent blocks should exhibit similar expression abundances to one another. Therefore, for each pair of adjacent haplotype blocks, those alleles exhibiting the higher expression would be linked, as would those with the lower expression level. In this way, the inferred alleles along the length of each AqAF were reconstructed.

4.4 Results

The AFs' putative allorecognition role (Fernàndez-Busquets and Burger 1999) means that these molecules are expected to exhibit a high degree of diversity between individuals. I therefore investigated the potential contributions of two possible mechanisms of generating diversity in these genes - alternative splicing and sequence polymorphism.

4.4.1 Alternative splicing of *AqAFs* across *A. queenslandica* development

The *AqAFs* are large genes comprised of many exons, all but one of which are flanked by introns in phase 1 (Chapter 2.4.8). Such symmetrical exons are often associated with alternative splicing or exon shuffling processes to prevent disruption of the transcriptional reading frame of the resulting mRNA (Patthy 1987; Fedorov et al. 1998). I therefore investigated whether the over-representation of symmetrical *AqAF* exons is a sign that these genes undergo alternative splicing as a means of generating the sequence variability expected of allorecognition molecules. To do so, I examined the *AqAF* transcripts present in a whole-transcriptome alternative splicing dataset generated from *de novo* assembled precompetent larval, competent larval, juvenile and adult transcripts. I also used PCR to amplify and sequence a portion of the *AqAFC* competent larval cDNA, as transcript assembly for this gene is complicated by the presence of three highly similar repeat regions therein (Chapter 2.4.6).

Transcripts encoding putative alternatively spliced *AqAF* variants (i.e. conflicts between expected and observed exon boundaries) were identified from one or more developmental stage for each of the six genes. A total of 56 variant *AqAF* transcripts were identified (including 11 *AqAFC* PCR products), each exhibiting either intron retention (53%), transcript initiation within an intron (32%; including two transcripts where the first exon was preceded by an unknown sequence) or transcript termination within an intron (15%; including one transcript with unknown sequence) events (Table 4.2). Note that as many assembled transcripts in this dataset are not complete, some intron initiation or termination events may actually represent instances of intron retention. No alternative exon usage was identified for any developmental stage for any *AqAF* gene. Changes to 75% of variant transcripts are predicted to introduce one or more premature termination codons, and 14% of variant transcripts are predicted to encode signal peptides (with or without an upstream termination codon), that may allow transcription of novel protein isoforms (Figure 4.4; Table 4.3). Eleven percent of transcripts lack both stop codons

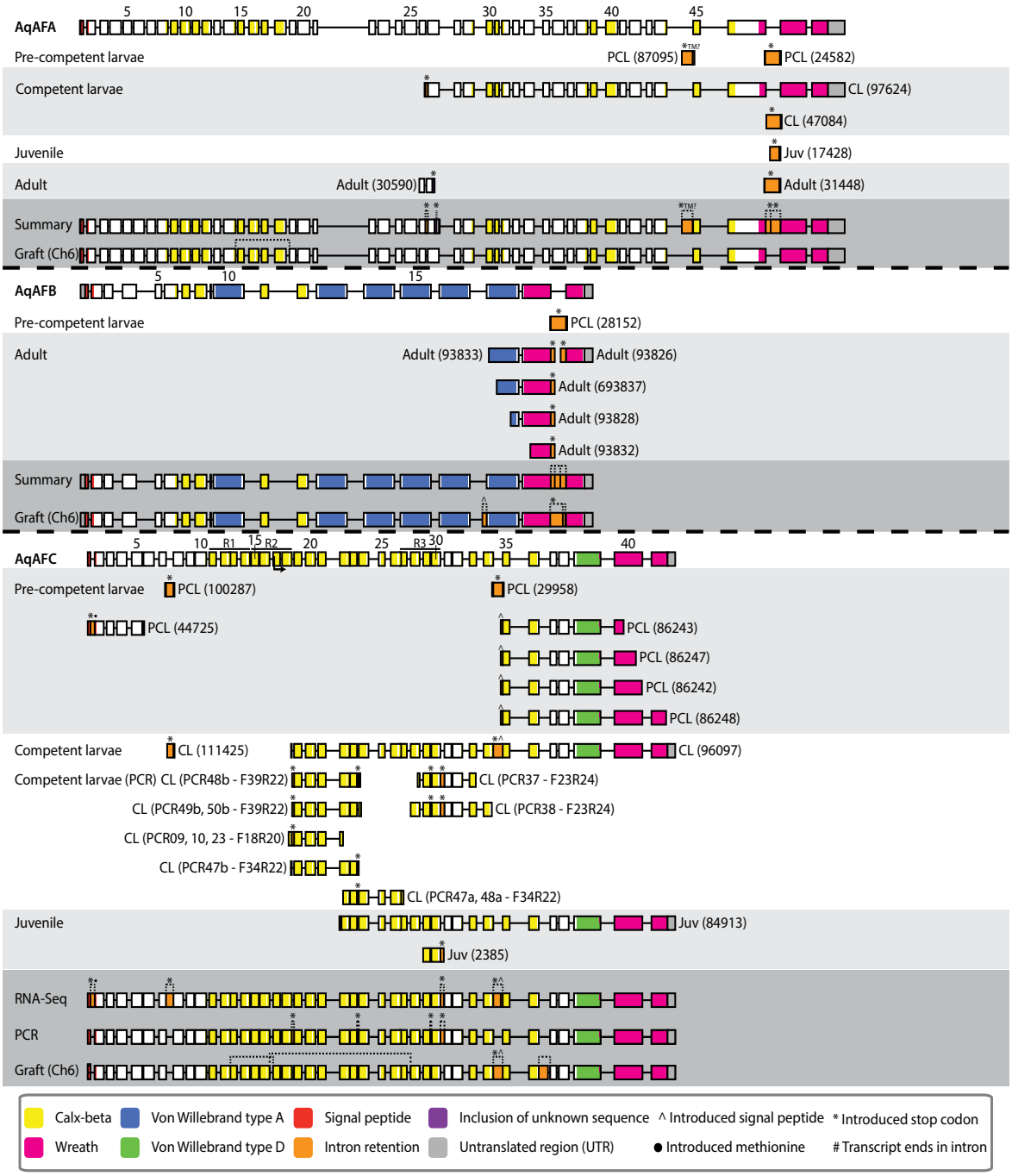


Figure 4.4 Alternatively spliced *AqAF* transcripts in sponge development *(Part 1 of 2)*

For each *AqAF* gene, the Aqu2.1 gene model prediction (top line) and putative alternatively spliced transcripts from each developmental stage are shown. Boxes represent exons (every fifth exon is numbered) and the connecting lines represent introns; regions encoding protein domains are coloured accordingly. Orange boxes represent intron inclusion events, while purple boxes represent inclusions of unknown sequence. Regions where domain type predictions overlap are depicted by overlapping colours. Exons and introns are drawn to scale. Symbols above each model represent predicted effects on the encoded proteins (see key). Two summaries are given for each gene (bottom lines), in which all observed changes from this experiment ('Summary') and the adult tissue graft experiment discussed later in Chapter 6 ('Graft (Ch6)') are annotated on the full-length gene models. For *AqAFC*, two summaries from this experiment are given - one from the RNA-Seq analysis and another from the PCR analysis. No graft summary is provided for *AqAFF* as no alternatively spliced transcripts were identified for this gene in Chapter 6.

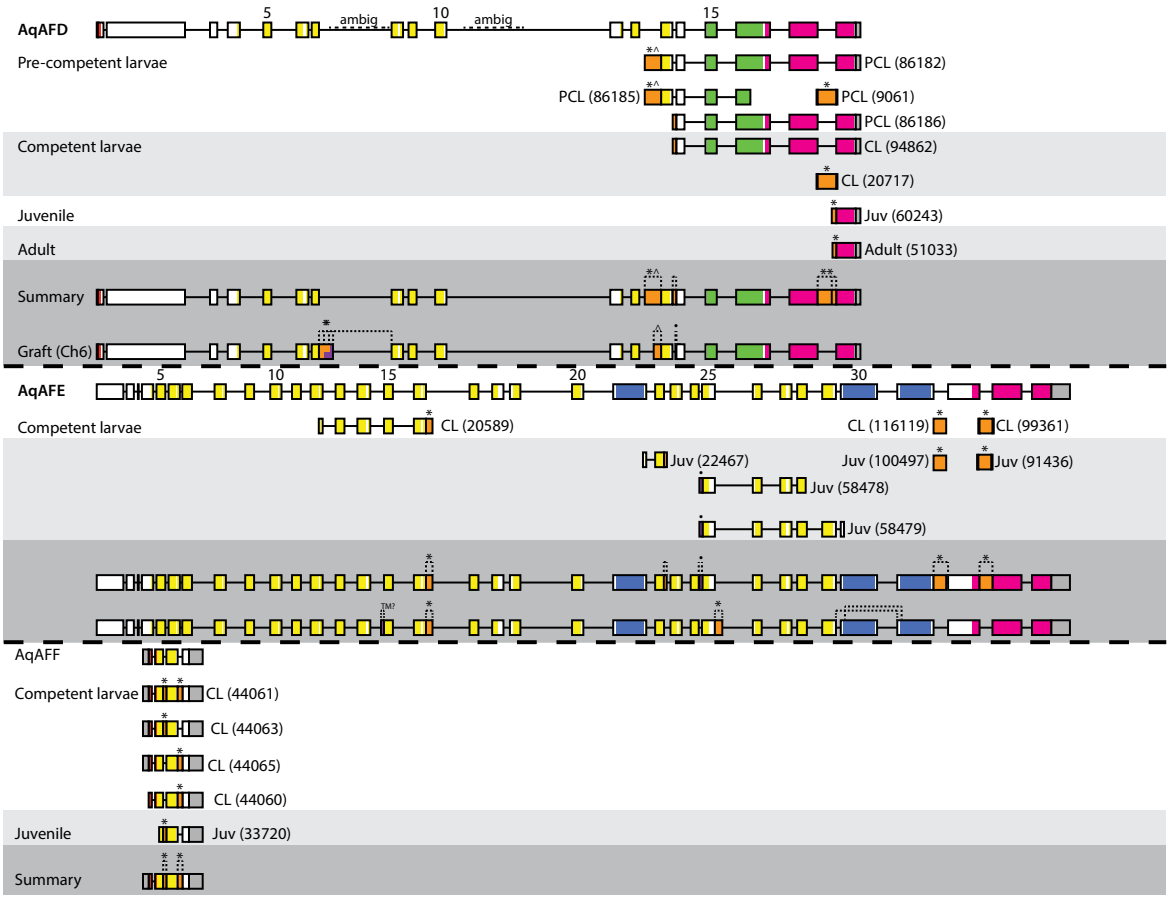


Figure 4.4 - Alternatively spliced AqAF transcripts in sponge development *(Part 2 of 2)*

and signal peptides, and maintain the normal transcriptional reading frame along their length. The domain and intron-exon architectures of all putatively spliced transcripts are shown in Figure 4.4, and the protein-level changes that these events are predicted to cause are discussed further in Table 4.3.

4.4.2 Detection of transcriptome-wide nucleotide variants

Whole-transcriptome sequencing data from four adult *A. queenslandica* individuals (Sponges A to D) were surveyed to identify putative sequence polymorphisms within both the transcriptomes as a whole and, the *AqAF* genes more specifically. Between ~197,000 (Sponge B) and ~398,000 (Sponge D) total potential variant sites were detected in each dataset; this disparity is a direct consequence of the differences in sequencing depth between individuals (Table 4.4, expanded in Appendix 4.3). The number of variants per 1000 sequencing reads decreases with increasing library size (Table 4.4, Appendix 4.3), presumably because above a certain sequencing depth threshold, increasing read counts does not

Table 4.3 Predicted effects of *A. queenslandica AF* alternative splicing on encoded proteins

(Part 1 of 2)

AqAFA			
Position	Stage/s	Type of change	Predicted translational effect
Intron 26	CL	Intron retention	Introduces a stop codon. Reading frame resumes downstream.
Exon 27	A	Novel sequence	Encodes 1 aa before introducing stop codon (Transcript encodes first half of Exon 27 before introducing unknown sequence)
Intron 44	PCL	Starts in intron	Introduces a stop codon. After 11 aa, introduces a transmembrane domain (Unknown if TM represents true TM or misclassified SP)
Intron 46	PCL, CL, A	Intron retention	Introduces a stop codon (Very short transcript)
Intron 46	J	Starts in intron	Introduces a stop codon (Very short transcript)
AqAFB			
Position	Stage/s	Type of change	Predicted translational effect
Intron 18	PCL	Intron retention	Introduces a stop codon
Intron 18	A (4)	Ends in intron	Introduces a stop codon
Intron 18	A	Starts in intron	Introduces a stop codon
AqAFC			
Position	Stage/s	Type of change	Predicted translational effect
Intron 1	PCL	Intron retention	Introduces a stop codon. Reading frame resumes downstream, including methionine. No signal peptide (SP) predicted.
Intron 7	PCL	Intron retention	Introduces a stop codon
Intron 7	CL	Starts in intron	Introduces a stop codon
Intron 18	CL-PCR	Intron retention	Introduces a stop codon
Intron 18	CL-PCR (2)	Starts in intron	Introduces a stop codon
Intron 23	CL-PCR (2)	Intron retention	Introduces a stop codon
Intron 23	CL-PCR	Ends in intron	Introduces a stop codon
Intron 29	CL-PCR (2)	Intron retention	Introduces a stop codon
Intron 30	CL-PCR (2)	Intron retention	Introduces a stop codon
Intron 30	J	Ends in intron	Introduces a stop codon
Intron 34	PCL	Intron retention	Introduces a stop codon
Intron 34	CL	Intron retention	Introduces a stop codon. Predicted SP
Intron 34	PCL (4)	Starts in intron	Predicted SP

Table 4.3 Predicted effects of *A. queenslandica AF* alternative splicing on encoded proteins
(Part 2 of 2)

AQAFD			
POSITION	STAGE/S	TYPE OF CHANGE	PREDICTED TRANSLATIONAL EFFECT
Intron 12	PCL (2)	Starts in intron	Introduces a stop codon. Predicted SP
Intron 13	PCL, CL	Starts in intron	Maintains reading frame
Intron 17	PCL, CL	Intron retention	Introduces a stop codon
Intron 17	J, A	Starts in intron	Introduces a stop codon
AQAFE			
POSITION	STAGE/S	TYPE OF CHANGE	PREDICTED TRANSLATIONAL EFFECT
Intron 16	CL	Ends in intron	Introduces a stop codon
Intron 22	J	Ends in intron	Maintains reading frame
Exon 24	J (2)	Novel sequence	Introduces a methionine. No SP predicted. (Transcript encodes exon 24 before introducing unknown sequence)
Intron 31	CL, J	Intron retention	Introduces a stop codon
Intron 32	CL, J	Intron retention	Introduces a stop codon
AQAFF			
POSITION	STAGE/S	TYPE OF CHANGE	PREDICTED TRANSLATIONAL EFFECT
Introns 2 and 3	CL	Intron retention	Introduces a stop codon
Intron 2	CL, J	Intron retention	Introduces a stop codon
Intron 3	CL (2)	Intron retention	Introduces a stop codon

significantly increase genomic coverage and, therefore, the number of detected polymorphisms. The sequencing error rate is expected to be 0.1% based on the Illumina HiSeq 2000 specifications in the year 2012 (Glenn 2011), although this value is likely to be an underestimate of the actual error rate (Wall et al. 2014). Filtering of low-frequency nucleotide differences was performed prior to analysis, reducing the expected number of false positive nucleotide variants.

An average of 25.6% (± 0.8%) total detected variants was predicted to be non-synonymous, i.e. causing an amino acid change (Table 4.4, Appendix 4.3). A much smaller group of variants (3.8% ± 0.6%) was predicted to alter canonical intron splice sites (Table 4.4, Appendix 4.3). Analysis of a 'consensus' dataset, comprising only those variant sites present in all four sponge individuals, produced very similar results to analysis of the four complete datasets; here, 25.1% and 3.5% of variants were predicted to

Table 4.4 General nucleotide variant information (abridged)

	Transcriptome-wide		AqAF genes	
	Consensus	Average (Raw)	Consensus	Average (Raw)
Basic variant statistics				
Mapped reads	-	-	-	-
Total variants	34,156	300,977.8	49	407.3
Variants / 1000 reads	-	-	-	-
Predicted false positives (0.1%)	34.2*	301.0	0.0	0.4
Variant type (percentage of total variants)				
Insertion	1.4%	2.4%	0.0%	1.4%
Deletion	2.0%	3.3%	0.0%	1.6%
MNV (Multi-nucleotide variants)	3.8%	4.9%	8.2%	7.0%
SNV (Single-nucleotide variants)	92.6%	89.0%	91.8%	90.0%
Replacement	0.1%	0.4%	0.0%	0.1%
Single nucleotide transitions vs transversions (percentage of total SNPs)				
Total CDS SNVs	31,633	266,697.8	45	366.0
Transitions	75.5%	73.2%	84.4%	74.5%
Transversions	24.5%	26.8%	15.6%	25.5%
Individual SNVs (percentage of total SNPs)				
Total CDS SNVs	31,633	266,697.8	45	366.0
A → G - transition	20.7%	17.9%	26.7%	20.3%
A → C - transversion	2.8%	2.9%	2.2%	3.1%
A → T - transversion	4.1%	4.8%	2.2%	3.5%
G → A - transition	17.5%	18.8%	24.4%	21.3%
G → C - transversion	2.6%	2.7%	8.9%	3.9%
G → T - transversion	2.6%	3.0%	0.0%	2.7%
C → A - transversion	2.6%	3.0%	2.2%	2.9%
C → G - transversion	2.7%	2.7%	0.0%	2.6%
C → T - transition	17.1%	18.7%	15.6%	17.4%
T → A - transversion	4.1%	4.7%	0.0%	4.6%
T → G - transversion	2.9%	2.9%	0.0%	2.2%
T → C - transition	20.2%	17.9%	17.8%	15.6%
Predicted effects (percentage of total variants)				
Total variants	34,156	300,977.8	49	407.3
Amino acid change	25.1%	25.6%	34.7%	40.1%
Non-conservative change	-	-	-	18.9%
Splice change	3.5%	3.8%	0.0%	3.8%

The full version of this table is available in Appendix 4.3

alter amino acids or splice sites, respectively (Table 4.4, Appendix 4.3). When the distribution of total variants was broken down by the form each change took, single nucleotide changes (SNPs) were the most commonly detected variant type (89.0% ± 1.7%); other variant types - multi-nucleotide changes (4.9% ± 0.6%), deletions (3.3% ± 0.6%), insertions (2.4% ± 0.4%), and replacements (0.4% ± 0.1%) - were relatively rarer (Table 4.4, Appendix 4.3). Looking specifically at only those SNPs located within coding regions, transitions (purine-purine, G <—> A, or pyrimidine-pyrimidine, C<—>T) were, as expected, most common (Table 4.4, Appendix 4.3). Transitions were 5.9 times more likely to occur than transversions, after accounting for the larger number of possible transversion events. However, comparisons between individual transition or transversion classes revealed statistical differences between the frequencies of most types of changes (Table 4.5).

4.4.3 Nucleotide variants within the *AqAF* locus

A total of 967 unique variant sites, relative to the reference genome, were identified within the *AqAFA* to *AqAFE* across the four studied individuals; 49 sites were identified within all four sponge samples (Table 4.4, Appendix 4.3). No variant sites were detected within *AqAFF*. When accounting for the presence of canonical nucleotides at a variant site, 99% of sites within the full *AqAF* dataset were biallelic (i.e. only two nucleotide types identified across all reads from all individuals for a given position), with only five positions exhibiting three nucleotide types

Table 4.5 Significant differences between transcriptome-wide SNP distribution categories

	G ↑ A	A ↑ G	T ↑ C	C ↑ T	C ↑ A	T ↑ A	C ↑ G	T ↑ G	A ↑ C	G ↑ C	A ↑ T	G ↑ T
A → G		**	*	*								
G → A				**								
C → T				*								
T → C												
A → C						****		**		***	****	**
A → T							***	***	****	****	*	****
G → C								**		**	***	
G → T									**	****	***	**
C → A										**	***	
C → G											****	**
T → A												****
T → G												

* p ≤ 0.05; ** p ≤ 0.01; *** p ≤ 0.001; **** p ≤ 0.0001

across all individuals (data not shown). The *AqAF* sequence variants were statistically enriched ($p \leq 0.001$) for non-synonymous changes (40.1% ± 1.1%) compared to the transcriptome-wide average (25.6% ± 0.77%) (Table 4.6). No change in the frequency of variants predicted to alter intron splice sites was observed between the *AqAFs* and the whole-transcriptome datasets (Table 4.6). The *AqAF* variants exhibit fewer instances of insertion (1% fewer; $p \leq 0.05$) and deletion (1.7% fewer; $p \leq 0.05$) and more of multi-nucleotide variants (2.1% more; $p \leq 0.01$) relative to what is observed transcriptome-wide (Table 4.6). The *AqAFs* show a slight but statistically significant ($p \leq 0.05$ for each) reduction in the proportion of coding region G-to-A (18.1% ± 0.04%) and A-to-G (14.8% ± 1.2%) transitions relative to the transcriptome as a whole (17.36% ± 0.1% and 19.06% ± 0.2%, respectively) (Figure 4.5). However, the frequencies of other individual substitutions, and of the average frequency of transitions and transversions overall, remained constant (Figure 4.5, Table 4.6). Comparisons between individual transition and transversion classes within the *AqAF* variants revealed that A-to-G transitions were statistically less common than G-to-A ($p \leq 0.01$), C-to-T ($p \leq 0.05$) and T-to-C ($p \leq 0.001$) changes. A-to-T transversions also occurred at a statistically higher rate ($p \leq 0.05$) than G-to-C changes (Table 4.7). However, contrary to transcriptome-wide observations, all other pairwise comparisons between either transitions or transversions were not significantly different from one another (Table 4.7).

Table 4.6 Significant differences between genome-wide and AF-specific variant categories

	Genome	AqAF	Significance
Variant type			
Insertion	2.4%	1.4%	$p \leq 0.05$
Deletion	3.3%	1.6%	$p \leq 0.05$
MNV	4.9%	7.0%	$p \leq 0.01$
SNP	89.0%	90.0%	-
Replacement	0.4%	0.1%	-
Transitions vs. transversions			
Transition	74.0%	74.8%	-
Transversion	25.9%	25.2%	-
Individual SNPs			
A → G*	17.4%	14.8%	$p \leq 0.05$
G → A*	19.1%	18.1%	$p \leq 0.05$
C → T*	19.1%	21.7%	-
T → C*	18.5%	20.2%	-
A → C	2.9%	2.7%	-
A → T	4.5%	4.5%	-
G → C	2.8%	2.6%	-
G → T	3.1%	3.4%	-
C → A	2.8%	3.0%	-
C → G	2.5%	2.7%	-
T → A	4.6%	3.4%	-
T → G	2.7%	3.0%	-
Predicted effects			
Amino acid change	25.6%	40.1%	$p \leq 0.001$
Splice change	3.8%	3.8%	-

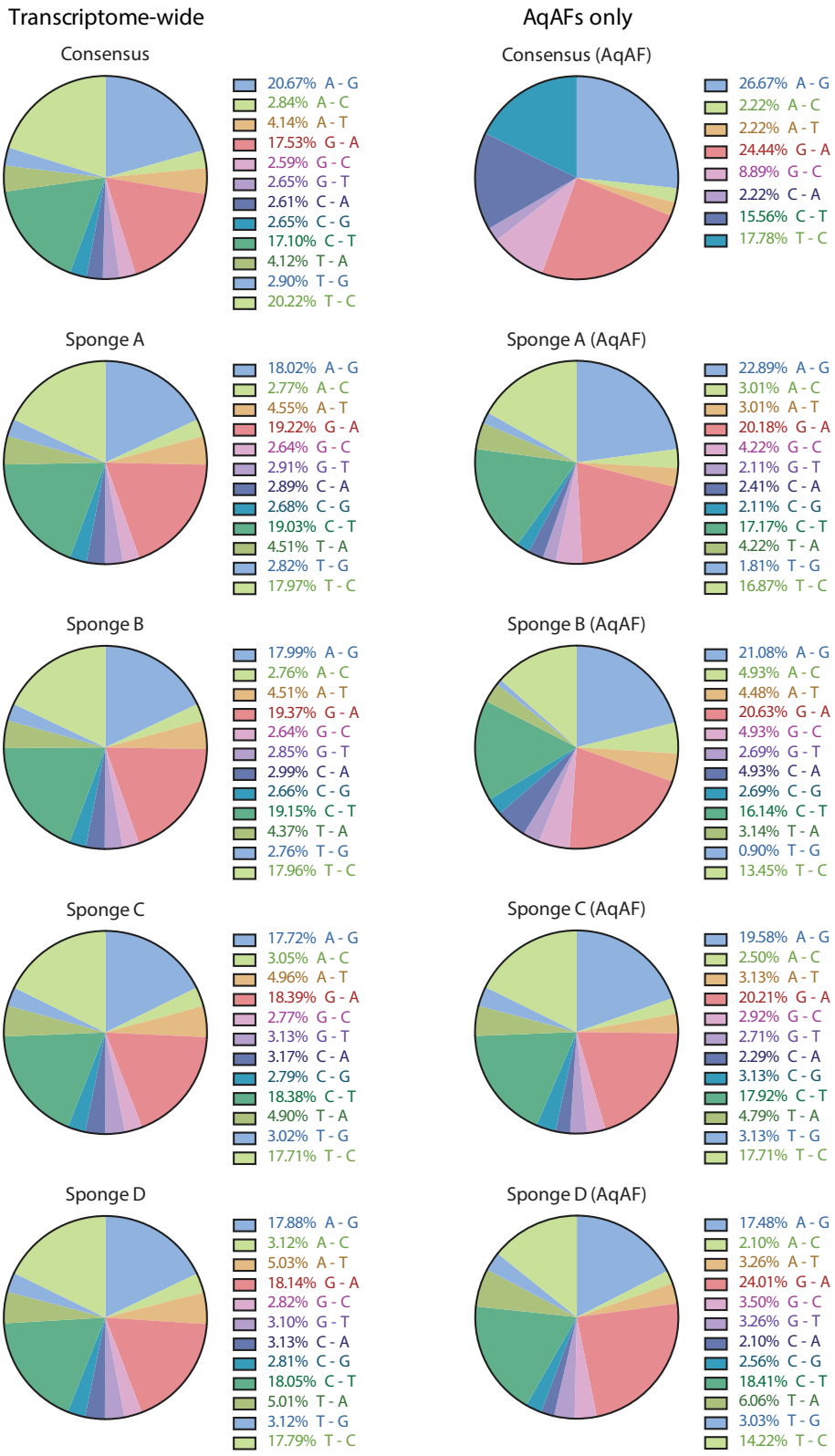


Figure 4.5 SNP substitution frequencies
Each pie chart shows the distribution of the different SNP substitute categories per sponge sample, both transcriptome-wide and in the *AqAF* genes only. The results of a consensus dataset, where only those variants present in all four sponge individuals are shown. This dataset contains only those single-nucleotide changes that were localised within a coding region of an *A. queenslandica* gene, and are given relative to the correct orientation of each gene on the chromosome.

4.4.4 *AqAF* haplotype reconstruction

Table 4.7 Significant differences between AF-specific SNP distribution categories

	G ↑ A	A ↑ G	T ↑ C	C ↑ T	C ↑ A	T ↑ A	C ↑ G	T ↑ G	A ↑ C	G ↑ C	G ↑ T	T ↑ A	G ↑ T
A → G		**	*	***									
G → A													
C → T													
T → C													
A → C													
A → T							*						
G → C													
G → T													
C → A													
C → G													
T → A													
T → G													

* p ≤ 0.05; ** p ≤ 0.01; *** p ≤ 0.001; **** p ≤ 0.0001

To investigate each putative *AqAF* nucleotide variant in context, I sought to reconstruct the full-length alleles of each *AqAF* gene from four adult sponge individuals. This was achieved through manual examination of the identified variants within their mapped sequencing reads, which were visualised on an annotated *A. queenslandica* genome browser. Each neighbouring pair of predicted variants was examined to pinpoint instances where adjacent variants were encoded by a single sequencing read and, therefore, by the same allele (Figure 4.3, *step 1*). By walking along the assembled *AqAF* locus, paired chains of variants could be identified (referred to as haplotype blocks), with each member of the pair representing a fragment of one of the two alleles from the diploid *A. queenslandica* genome (Figure 4.3, *step 1*).

Each allele fragment within a haplotype block represents a reconstructed piece of a full-length allele. While two alleles for a single gene are not necessarily expected to exhibit identical quantitative expression levels to one another, the expression of each allele should, in theory, remain constant across its length. The average frequency of all nucleotide variants was calculated per allele fragment per haplotype block (Figure 4.3, *step 2*). Allele fragments of neighbouring haplotype blocks were inferred to be linked if their average expression frequency values were similar (Figure 4.3, *step 3*). For two of the four sponges (sponges A and B), two full-length alleles per individual were successfully reconstructed for all *AqAF* genes. Despite each variable position only encoding one or two different nucleotides across

all sponges, different variant combinations were used to produce four unique alleles per gene from two individuals. Intriguingly, sponge C appeared to possess at least four alleles, despite each variant site again displaying a maximum of two possible nucleotide options (data not shown). Reconstruction of the alleles for this sponge was therefore not pursued further. The first exon of *AqAFA* in sponge D appears to also encode four alleles; this region could therefore not be easily reconstructed. However, as only two alleles were detected across the rest of *AqAFA* and the other *AqAFs*, these alleles were successfully reconstructed and included in further analyses. As above, alleles from sponge D were unique within and between sponges.

a. AqAFA

The three examined sponge individuals - sponges A, B and D - exhibited similar numbers of *AqAFA* nucleotide polymorphisms (i.e. SNPs, insertions/deletions etc.) to one another, with an average of 15.6 variant sites per 1000 base pairs (bp) of coding sequence (Table 4.8). Synonymous, conservative and non-conservative changes were distributed across the length of the sequence in all six reconstructed alleles (Figure 4.6a; Appendix 4.4). However, sponge D exhibited more variants in the first 20 exons of *AqAFA* than did the other two sponges, which in turn possessed a greater number of variants in the following 20 exons than did sponge D (Figure 4.6a). One exon 15 variant in the two sponge D alleles is predicted to cause a frameshift during protein translation (Figure 4.6a). The retention of intron 46, as identified in the alternative splicing experiment (Figure 4.4), is supported by the identification of nucleotide variants in exon 46 and intron 46 that are predicted to alter intron splice sites (Figure 4.6a). However, other predicted intron splice site nucleotide changes or intron retention events are not mutually supported by one another (Figure 4.6a).

b. AqAFB

Unlike for *AqAFA*, sponges A and B exhibit a much lower number of total variant sites (average 4.9 sites per 1000 bp coding sequence) than sponge D (17.2 sites per 1000 bp) (Table 4.8). In sponges A and B, variants are localised solely between exon 14 and intron 18, with relatively high variant frequencies in exons 17 and 18. Sponge D exhibits more variants in the end region of *AqAFB*, and also variants in exons 3 and 7 (Figure 4.6b). A single frame shift variant is predicted within exon 16 of sponge D allele 1. The two observed intron retention events for this gene, one of which is predicted

Table 4.8 Total and scaled variants per *A. queenslandica AF* gene

	Total Variants						Variants per 1000 bp coding sequence					
	Consensus	Average (Raw)	Sponge A	Sponge B	Sponge C	Sponge D	Consensus	Average (Raw)	Sponge A	Sponge B	Sponge C	Sponge D
AqAFA	14	156	144	142	160	178	1.5	17.2	15.8	15.6	17.6	19.6
AqAFB	10	74	41	31	111	113	1.7	12.4	6.9	5.2	18.7	19.0
AqAFC	8	42.8	22	16	51	82	1.0	5.6	2.9	2.1	6.6	10.7
AqAFD	14	48	31	32	57	72	2.7	9.3	6.0	6.2	11.0	13.9
AqAFE	3	84	124	20	151	41	0.4	9.9	14.6	2.4	17.8	4.8
AqAFF	0	0	0	0	0	0	0	0	0	0	0	0

to introduce a novel signal peptide to the translated protein, are both supported by nucleotide variants, within intron 16 and exon 17, and intron 18, respectively (Figure 4.6b).

c. AqAFC

AqAFC exhibits an average of 2.3 variant sites per 1000 bp for sponges A and B, in contrast to 10.0 variant sites per 1000 bp for sponge D (Table 4.8). Sponge A is homozygous for a majority of variants along its length. The majority of sponge A and B variants are synonymous nucleotide substitutions, which are largely restricted to the end region of this gene. Variants are distributed more evenly across the length of *AqAFC* in sponge D; however, while allele 1 displays a mix of synonymous, conservative and non-conservative changes, the majority of allele 2 changes are synonymous or conservative. Two of the eight predicted intron retention events are supported by nucleotide variants in exon 8 and intron 35, respectively. Other predicted splice site nucleotide variants and alternatively spliced transcripts were not mutually supportive in this instance (Figure 4.6a).

d. AqAFD

An average of 6.1 variant sites per 1000 bp was observed for *AqAFD* in Sponges A and B; variants in these sponges are restricted to exons 12 to 18. These changes are mostly synonymous nucleotide substitutions, with a smaller number of conservative and non-conservative variants detected across this region. Sponge D possesses 13.4 variants per 1000 bp, which are located between exons 2 and 19.

Two of the four putatively retained introns are supported by nucleotide variants in sponge D that are predicted to alter intron splice sites, in exon 6 and intron 12 respectively (Figure 4.6d).

e. AqAFE

Sponges B and D encode an average of 3.5 variant sites per 1000 bp of *AqAFE*. Sponge A, in contrast, encodes a much larger number of variant sites (13.4 sites per 1000 bp) (Table 4.8). For sponge B, only three *AqAFE* exons (exons 1, 18 and 23) contain nucleotide variants; the rest of the observed variants all fall within canonical introns. Sponge D possesses a cluster of variants between exons 16 to 19, plus extras in exon 11, intron 1 and intron 30. Sponge A variants are distributed across the length of the gene. One of the seven putative intron retention events for this gene is supported by the presence of a nucleotide variant in intron 25 from sponge A (Figure 4.6e).

4.5 Discussion

Aggregation factors have been implicated in allorecognition (Fernàndez-Busquets and Burger 1999) and are therefore predicted to display high levels of between-individual variability consistent with this role (Chapter 1.1.3). Such variability could exist on a genomic, nucleotide, transcript, protein and/or molecular complex level. Multiple diversification methods could be used in combination, and differential regulation of these processes could allow fine-tuned control of diversity between individuals or in a context-dependent manner. In this chapter, I sought to catalogue and characterise the contributions of two potential sources of *AqAF* diversity - alternative splicing and nucleotide variants - across development and between individuals, respectively.

Figure 4.6 Distribution of allelic variants across *AF* gene models →
(Begins over page)
Depicts the proportions of synonymous, non-synonymous conservative, non-synonymous non-conservative, and intronic variants detected for each exon or intron per allele per *AqAF* gene. All values are scaled per 100 bp of intron/exon sequence. The gene model at the top represents the Aqu2.1 gene model for each *AqAF* gene, annotated with all observed alternative splicing events (orange boxes) and their predicted effects on the encoded proteins (see key). Instances where alternative splicing events are supported by the predicted nucleotide variants are marked with an arrow. As no *AqAFF* variants were detected in this study, this gene is not shown here.

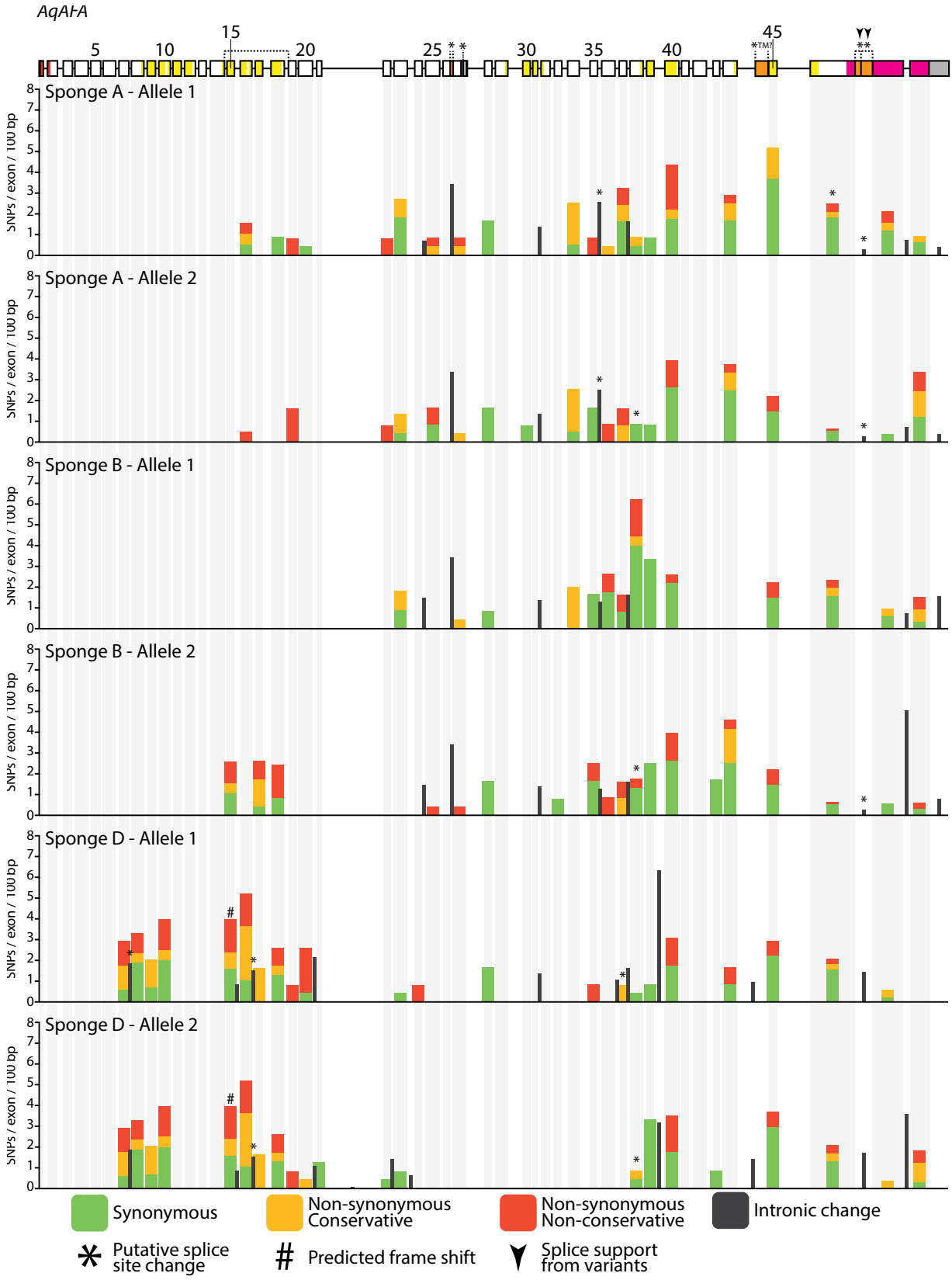


Figure 4.6 Total and scaled variants per *A. queenslandica AF* gene
(Part 1 of 5)

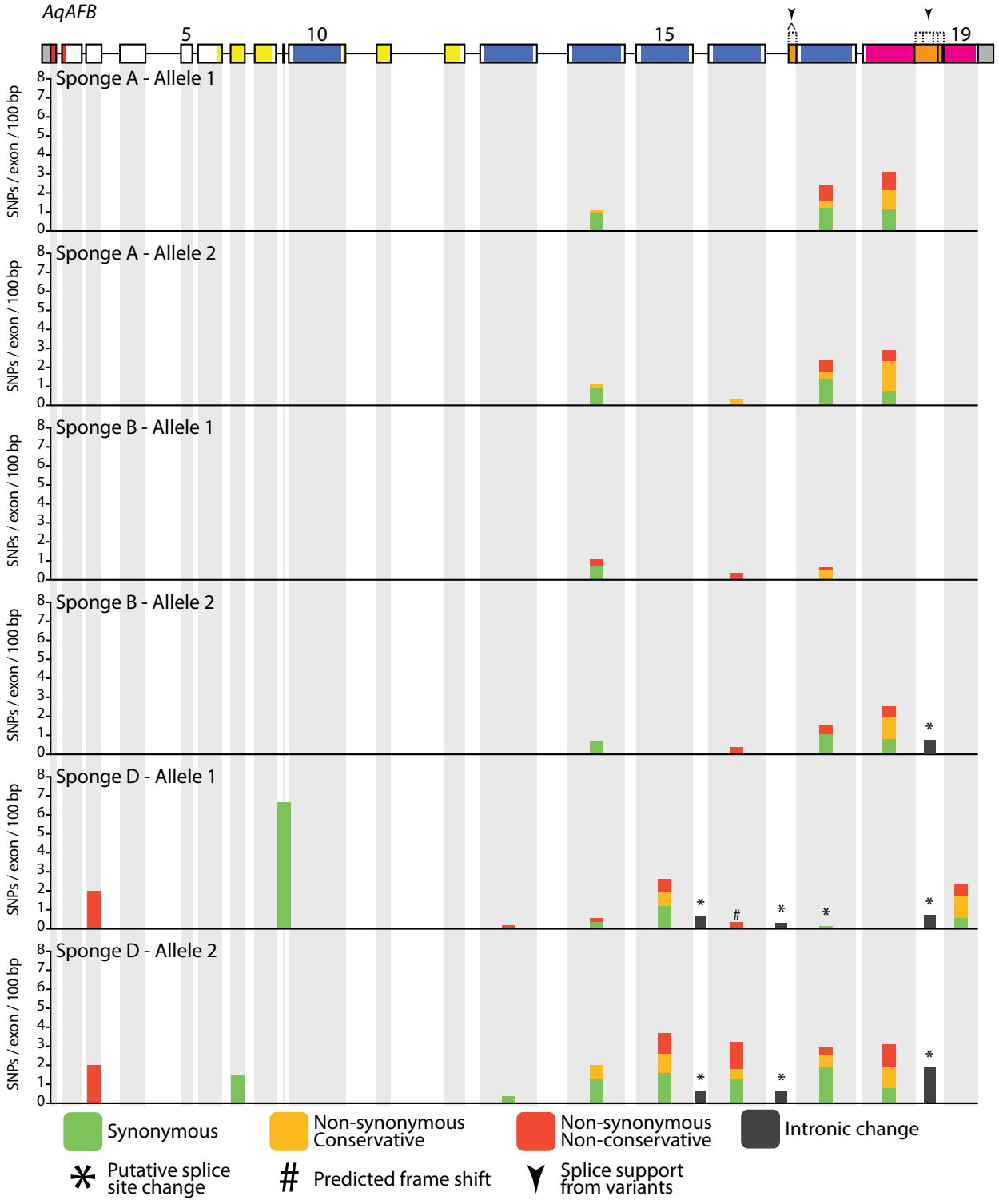


Figure 4.6 - Total and scaled variants per *A. queenslandica AF* gene
(Part 2 of 5)

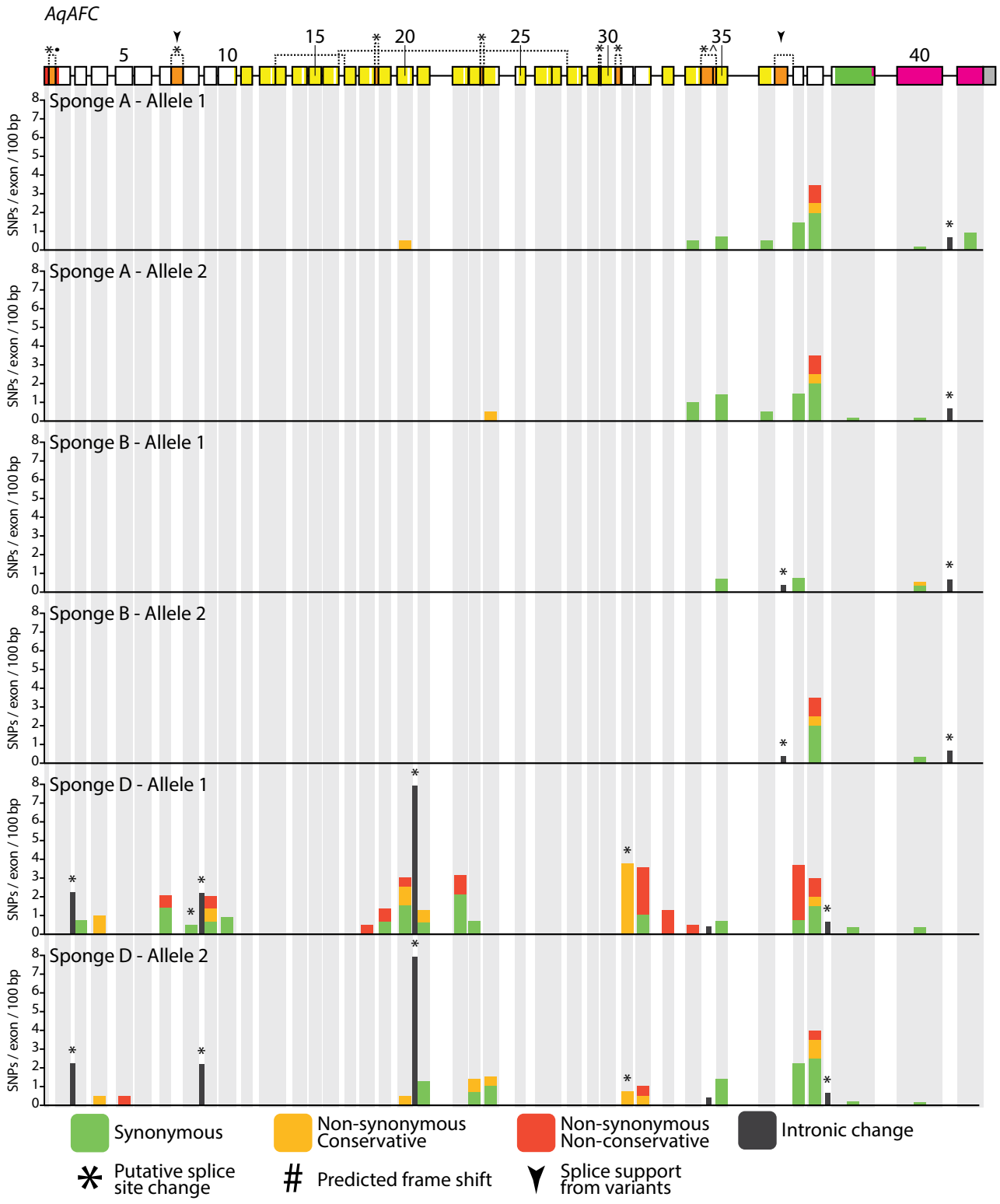


Figure 4.6 - Total and scaled variants per *A. queenslandica AF* gene
(Part 3 of 5)

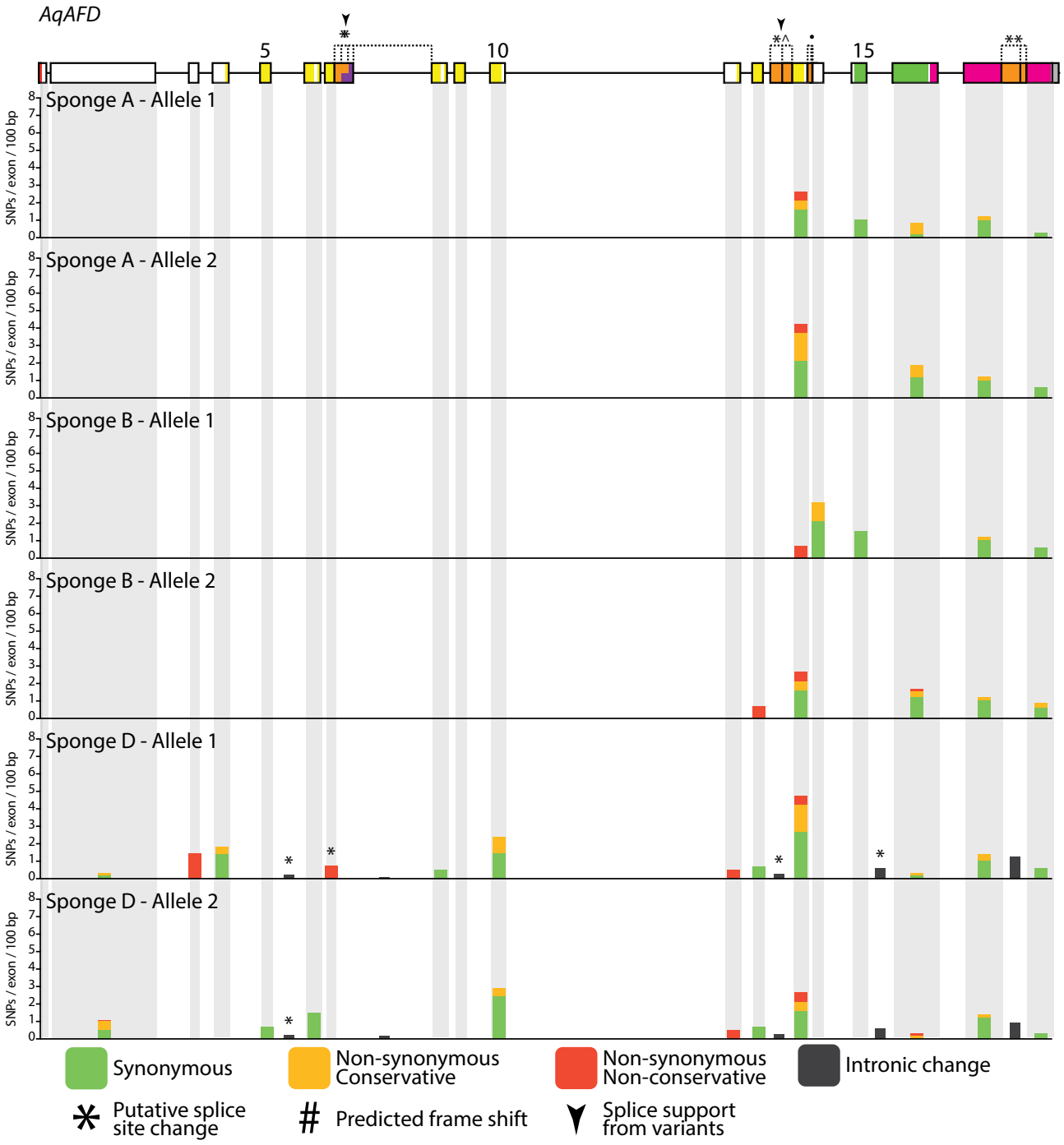


Figure 4.6 - Total and scaled variants per *A. queenslandica AF* gene
(Part 4 of 5)

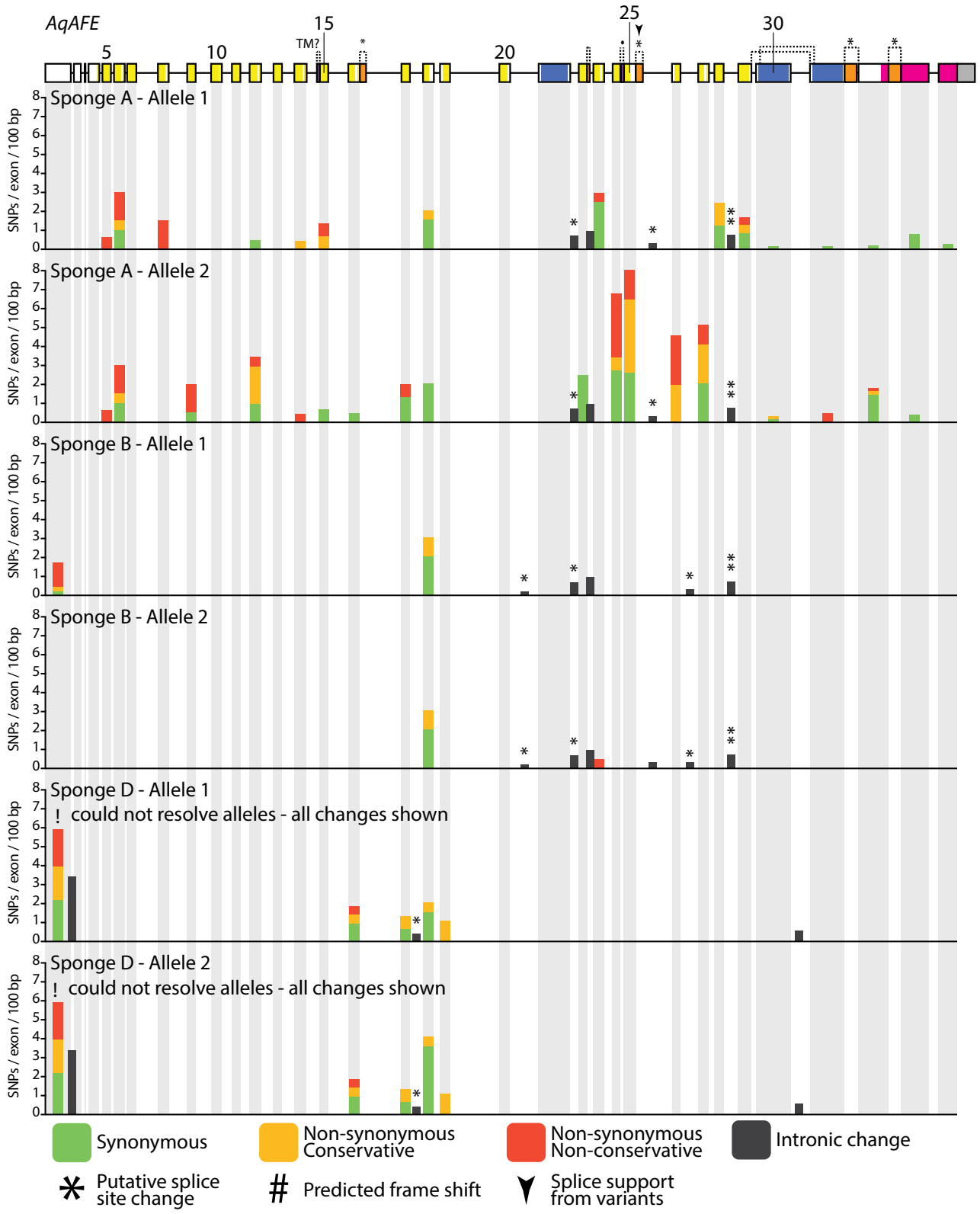


Figure 4.6 - Total and scaled variants per *A. queenslandica AF* gene
(Part 5 of 5)

4.5.1 The *A. queenslandica AFs* do not undergo exon rearrangement

The *AqAFs* are architecturally constrained at the genomic level, as they are built entirely from symmetrical exons (with each flanked by Phase 1 introns; Chapter 2.4.8) that encode uni- or multi-exon domain modules (Chapter 2.4.7). I hypothesised that alternative splicing is used to rearrange *AqAF* exons post-transcriptionally, to produce novel exon combinations within or between domains, or shortened protein isoforms due to exon skipping. However, I did not find evidence to support this hypothesis, as the six *AqAF* genes did not show any indications of exon skipping or rearrangement. I therefore conclude that exon rearrangement via either alternative splicing or pre-transcriptional genomic processes (as the method of variant detection used here cannot distinguish between pre- and post-transcriptional changes) is not a widespread mechanism of *AqAF* diversification. However, one cannot exclude the possibility that alternative splicing of the *AqAFs* occurs somewhere in the *A. queenslandica* lifecycle or under specific environmental conditions that have not been surveyed here.

The modular structure and intron phase bias of the *AqAFs* may instead reflect the evolutionary history of the *AFs*. On a sequence level, the *AFs* of *A. queenslandica* and *C. prolifera* are the best studied to date. These *AF* sequences share a moderate degree of sequence similarity and similar domain architectures between species (Gauthier 2009) and both are constructed from symmetrical exons (Fernàndez-Busquets and Burger 1999). However, while all *AqAF* exons are flanked by Phase 1 introns (Chapter 2.4.8), all elucidated *C. prolifera* introns are in phase 0 (Fernàndez-Busquets and Burger 1999). In addition, while the *AqAF* introns have a median length of 72 bp (Chapter 2.4.6), *C. prolifera MAFp4* introns are much larger, ranging from 300 - 650 bp (Fernàndez-Busquets and Burger 1997). Average exon sizes of the *AFs* from these two species are, however, similar (Fernàndez-Busquets and Burger 1997). Structural information from other sponge species has also revealed that large differences in AF protein complex structure exist between species, with the circular AF ring structure apparently limited to the demosponge orders Poecilosclerida and Astrophorida; other examined species have either been shown or predicted to be linear in form (Figure 2.12). The AF genes have therefore undergone a high degree of reorganisation since the divergence of sponges from their common ancestor. Extensive exon shuffling has most likely occurred in the *A. queenslandica* and/or *C. prolifera* lineages after divergence from their common ancestor. Changes to intron size and phase have occurred; chance shifts in the phase of some *AF* introns in one or both lineages were probably perpetuated throughout the

genes by continued selection for inter-compatible symmetrical exons. The *AFs* are fast-evolving genes, meaning that genomic information from a wider distribution of demosponge species is required to better elucidate the similarities and differences between the *AF* gene complements and their structural and sequence properties from different species, and to understand the evolutionary processes that shaped the divergence of this gene family.

4.5.2 Retention of the *A. queenslandica AF* introns may allow AF regulation via nonsense mediated decay

A total of 56 alternatively spliced *AqAF* transcripts were detected from precompetent larvae, competent larvae, juvenile and adult sponges. All alternative splicing events involved the full or partial inclusion of canonical intron sequences, i.e. intron retention (53%), transcript initiation in an intron (32%) or transcript termination in an intron (15%) events. The biological distinction between these different classifications cannot be fully resolved at this time, due to the fragmented nature of some transcripts within these datasets. Forty-seven percent of *AqAF* alternatively spliced transcripts begun or ended within an intron and as such were classified as intron initiation or termination events. While some or all of these may represent the true transcript start or end positions, the numbers are likely to be overestimated given that the intron initiation and termination categories make up just 1% of transcriptome-wide alternative splicing events (S. Fernandez Valverde and B. Degnan, unpublished data). Therefore, a number of these events in the *AqAFs* are most likely incompletely assembled and therefore misclassified instances of intron retention. For this reason all *AqAF* alternative splicing events are discussed with the assumption that they represent intron retention, whether fully or partially. The bias within the AqAFs towards full or partial intron retention, rather than alternative exon usage, is consistent with the observation that 76% of alternative splicing events across the *A. queenslandica* transcriptome are predicted to cause full or partial inclusion of intron sequences (i.e. intron retention, alternative intron acceptor/donor, intron initiation/termination) in the resulting transcripts (S. Fernandez Valverde and B. Degnan, manuscript in preparation). While it remains possible that some observed intron retention instances are derived from pre-mRNA transcripts captured by RNA-Seq before intron splicing, the lack of transcripts predicted to encode multiple intron retention events suggests that the impact of these events is minor. However, intron retention events of interest should ideally be verified by PCR before future analysis continues.

The majority (75%) of *AqAF* intron inclusion events are predicted to introduce a premature termination codon (PTC) to the resulting protein product. These transcripts are therefore potential targets of the nonsense mediated decay (NMD) pathway. NMD is a mRNA surveillance mechanism by which the cell can detect and degrade erroneously spliced transcripts containing PTCs (Losson and Lacroute 1979). However, NMD has also emerged as a regulatory mechanism by which an organism can regulate transcript abundance and subsequent activity in a spatiotemporal manner (reviewed by Ge and Porse 2013). Therefore, this could potentially represent a further means by which the sponge can regulate AF activity and allorecognition; it is unlikely that the extensive intron retention observed across the *AqAFs*, with particular retention events observed in multiple transcriptomes, is purely due to mis-splicing. Alternatively, if the PTC-containing AqAF transcripts were protected from NMD in some way (as occurs, for instance, in the RNA editing molecule ADAR1) (Lykke-Andersen et al. 2007), this would suggest an alternative, unknown role for these transcripts.

4.5.3 The *A. queenslandica AFs* may encode novel truncated protein isoforms

A subset of *AqAFC* and *AqAFD* intron inclusion events (14% of transcripts) are predicted to introduce signal peptides to their resulting transcripts. All such transcripts from *AqAFC* and *AqAFD* are predicted to encode all (*AqAFC*) or part (*AqAFD*) of a Calx-beta domain, one Von Willebrand type D domain, and a Wreath domain (Figure 4.4). Similar short transcripts, predicted to encode a signal peptide and Wreath domain, with or without a Calx-beta domain, have been detected from the sponge species *Chondrilla nucula* and *Ephydatia muelleri* (Chapter 2.4.5). These transcripts, as in *A. queenslandica*, may represent isoforms resulting from alternative splicing of a longer gene sequence.

The inclusion of novel signal peptides in particular alternatively spliced *AqAF* transcripts provides good evidence that these observed intron retention events are both real and functional, as predicted signal peptide sequences are unlikely to be encoded by an intron by chance. The roles of the putative resulting novel proteins are unknown. In *C. prolifera*, the ring and arm subunits (MAFp3 and MAFp4, respectively) appear to be encoded by a single contiguous mRNA before being cleaved post-translationally to produce independent peptides (Fernàndez-Busquets and Burger 1997; Jarchow et al. 2000). One explanation for the novel signal peptides in *AqAFC* and *AqAFD* could therefore be that production of the independent ring subunit here occurs pre-translationally in some cases. This process, however, does not appear to be obligatory, as longer *AqAFC* transcripts that lacked novel signal

peptides were also predicted. Therefore, the shortened AqAF proteins may instead play some other regulatory role, such as competition with full-length AqAF proteins for binding targets. All transcripts possessing a putative signal peptide sequence sit close to the start of the assembled transcript. It is unknown whether this observation is biologically meaningful, for example if the sequences possess a novel transcription initiation site or if post-transcriptional RNA cleavage occurred prior to sequence capture by RNA-Seq. This could be tested using RACE-PCR (rapid amplification of cDNA ends - polymerase chain reaction) to determine the full-length transcript variant sequences and to determine whether the putative novel transcript start sites are real or artifactual. Searches for predicted signal peptides in other intron sequences could be performed in order to predict other possible intron retention events; these predictions could be tested using PCR.

4.5.4 *AqAF* alternative splicing does not appear to be age-specific

The majority of observed *AqAF* intron retention events were found in just one or a few of the four examined developmental stages. However, no clear patterns of developmental regulation of *AqAF* alternative splicing could be discerned from the present analysis. It should be acknowledged that a lack of transcripts exhibiting an intron retention event for a particular developmental stage does not constitute conclusive evidence that this event does not occur. Transcriptome sequencing and assembly instead allows the broad surveying of alternative splicing events within a particular locus. These results can in future be used to design more targeted analyses to confirm the developmental distributions of alternative splicing events of interest, for instance by taking a focussed PCR and sequencing approach. Here, primer pairs flanking putative intron retention events of interest, ideally flanking (1) multiple candidate intron retention events to reduce labour and experimental costs and (2) an intron not expected to be alternatively spliced, to detect possible instances of gDNA contamination. PCRs should be performed for each primer pair using complementary DNA (cDNA) derived from multiple individuals at different developmental stages. A total of 25 introns across the six *AqAF* genes (including introns flagged in Chapter 6) were found to exhibit intron retention, several of which are situated close together, so this could be performed relatively easily.

4.5.5 The *AqAFs* show an overabundance of non-synonymous changes

Regions of the *AqAF* genes may be under positive selection. Variants detected within the *AqAF* genes showed a statistically significant enrichment in non-synonymous nucleotide changes (average 40%) relative to the transcriptome as a whole (average 26%). No accompanying shift in the frequencies of transitions or transversions, or in specific nucleotide substitutions (except for a small but significant decrease in A —> G and G—> A transitions) was observed. The frequency of non-synonymous changes is not evenly distributed across the six haplotypes of the six *AqAF* genes. Several haplotypes from *AqAFA* (n = 2), *AqAFB* (n = 4), *AqAFC* (n = 1) and *AqAFE* (n = 4) possessed a greater number of non-synonymous changes than synonymous changes; the remaining haplotypes showed more synonymous than non-synonymous changes. Therefore, positive selection may be acting on at least some of the *AqAF* gene regions. External verification is required to support this claim, for example by again taking a PCR amplification and sequencing approach. Here, primer pairs targeting apparent variation hotspots of interest would be used to amplify genomic DNA sequences from multiple sponge individuals, followed by sequencing of the resulting PCR products. Multiple replicates from each individual should be performed to minimise the effects of PCR or sequencing errors. Statistical analyses of the ratios of synonymous and non-synonymous polymorphisms between individuals could then be performed. Although this method would not directly allow the distinction of separate alleles, analysis of the Sanger sequencing trace profiles of each sequence would reveal heterozygous positions per individual. The effects of each detected variant on the encoded amino acid can then be determined.

4.5.6 Nucleotide variant study limitations

It is important that the results of the nucleotide-level variant detection study presented above be interpreted in light of a number of caveats. First, the Illumina HiSeq 2000 sequencing platform has an inherent error rate of 0.1% errors per base per read (Glenn 2011). This error rate is likely an underestimate due to other inherent biases as discussed for example by Wall et al. (2014). While the software used to detect variants includes a filtering step to remove low frequency variants, it is likely that some false positive hits remain in this dataset. Other false positive or negative hits may be introduced if the reference genome sequence contains errors. Particular variants of interest within the *AqAFs* or elsewhere should therefore be verified using other methods such as PCR and sequencing.

Conversely, it is possible that the current analysis underestimates the level of diversity present in the *AqAF* locus and elsewhere. The variant analysis was performed on sequencing reads mapped to the *A. queenslandica* genome using standard mapping parameters, including a minimum similarity fraction per read of 0.8; reads not meeting the mapping parameters were discarded prior to variant detection analysis. Therefore, the possibility remains that particularly divergent reads may have been discarded during the mapping process. While this is a desirable feature of the mapping algorithm in most circumstances, it may hide the true level of diversity within variant loci. This could be explored by re-mapping the reads using less strict mapping parameters and repeating the variant detection analysis.

Finally, the haplotype reconstruction analysis was performed while making a key assumption that should be acknowledged. Where possible, linked variants were grouped to form haplotype blocks, each comprised of two allele fragments; the average variant frequency was calculated for each allele. Alleles of neighbouring haplotype blocks were linked by inferring that joined alleles should be expressed at roughly the same frequency as one another (Figure 4.3). However, if this inference was invalid at a particular region (for instance, if low sequencing coverage in a particular region skewed the average allele frequencies), neighbouring alleles could be erroneously joined, meaning that the resulting allele sequence would be incorrect. Therefore, the allele reconstructions, while informative, should be taken as a guide only and considered with caution.

4.5.7 Conclusion

Allorecognition genes are predicted to display between-individual differences that reflect the need to reject nonself individuals within a population. In sponges, the AFs are predicted to fulfil this role, and in *C. prolifera* the AFs have been shown to be allelic, with sequence differences between individuals correlated with differential graft responses (Fernàndez-Busquets and Burger 1997). I have shown that the *AqAFs* undergo alternative splicing in the form of full or partial intron retention, and that a number of these retention events are predicted to encode signal peptide sequences that may allow the *AqAFs* to produce novel shortened protein isoforms. At a nucleotide level, I detected a suite of apparent sequence polymorphisms within the *AqAFs*. In particular, I determined that the proportion of nucleotide changes predicted to encode amino acid changes is significantly greater than that observed

across the whole transcriptome, suggesting that the *AqAFs* may be under positive selection to help generate the between-individual gene product diversity predicted of these molecules.

As discussed in Chapter 4.2c, RNA editing is a second possible mechanism by which the *AqAFs* and other genes could become diversified at the nucleotide level. In Chapter 5, I investigate a major class of RNA editing molecules, the ADARs, which had previously been reported absent from sponges. I show that these molecules are indeed present in *A. queenslandica* and other sponge species, suggesting that RNA editing is mechanistically possible in sponges. I speculate on the significance of these findings for the evolution of metazoan RNA editing.

Chapter 5 - The origin of the ADAR gene family and animal RNA editing

5.1 Abstract

ADAR (adenosine deaminase acting on RNA) proteins convert adenosine into inosine in double-stranded RNAs and have been shown to increase gene product diversity in a number of bilaterians, particularly mammals and flies. This enzyme family appears to have evolved from an ADAT (adenosine deaminase acting on tRNA) ancestor, via the addition of a double-stranded RNA binding domain. The modern vertebrate ADAR family is comprised of ADAD, ADAR2 and ADAR1, each of which has a conserved domain architecture. To reconstruct the origin of this protein family, I identified and categorised ADAR family members encoded in the genomes and/or transcriptomes of early-branching metazoan and closely related non-metazoan taxa, including thirteen sponge and ten ctenophore species. I demonstrate that the ADAR protein family is a metazoan innovation, with the three ADAR subtypes being present in representatives of the earliest phyletic lineages of animals – sponges and ctenophores – but not in other closely related choanoflagellate and filasterean holozoans. *ADAR1* is missing from all ctenophore genomes and transcriptomes surveyed. Depending on the relationship of sponges and ctenophores to the rest of the Metazoa, this is consistent with either *ADAR1* being lost in ctenophores, as it has been in multiple metazoan lineages, or being an innovation that evolved after ctenophores diverged from the rest of the animal kingdom. The presence of Z-DNA binding domains in some sponge ADARs indicates an ancestral ADAR included this domain and it has been lost in multiple animal lineages. The ADAR family appears to be a metazoan innovation, with all family members in place in the earliest phyletic branches of the crown Metazoa. The presence of ADARs in sponges and ctenophores is consistent with A-to-I editing being a post-transcriptional regulatory mechanism that was used by the last common ancestor to all living animals and subsequently has been preserved in most modern lineages.

5.2 Introduction

RNA editing is a process of post-transcriptional RNA modification characterised by the insertion, deletion or modification of nucleotides (Simpson 1996; Gott and Emeson 2000). One of the most prevalent forms of RNA editing is mediated by the ADAR (adenosine deaminase acting on RNA) class of editing molecules, that work both selectively and non-selectively to deaminate adenosine residues into inosines (A-to-I editing) in double-stranded RNA (dsRNA) substrates (Bass and Weintraub 1988; Wagner et al. 1989). This editing can modify and regulate gene product output, for example via codon modification (as inosines are interpreted as guanosines by the cell), and influence splice site and small RNA functionality (Nishikura 2010).

ADARs and A-to-I editing have been shown or proposed to play a role in diverse biological processes, the extent of which are not yet fully understood. Perhaps the best-studied role of ADARs is their involvement in editing neuronal receptor and ion channel components in taxa such as flies, squid and vertebrates (Jantsch and Öhman 2008). ADARs have also been implicated in regulatory pathway roles, with suggested functions for A-to-I editing in RNAi antagonists (Scadden and Smith 2001), in pro- or antiviral mechanisms (Samuel 2011), and in the silencing of transposons and related sequences (Athanasiadis et al. 2004). Gene-level regulation may also occur through editing-induced sequestration of transcripts within organelles (Ng et al. 2013) or modification of splice sites (Rueter et al. 1999; Solomon et al. 2013). The primordial functionalities of the earliest ADAR systems are currently unknown.

ADATs (adenosine deaminase acting on tRNA) are critical proteins found in all eukaryotes. ADAT1 is equipped with a single adenosine deaminase (AD) domain, and is responsible for deamination of an adenosine in the tRNA wobble position into inosine (Gerber 1998), and does not play a role in RNA editing. ADARs appear to have originated via the incorporation of a double-stranded RNA binding (dsRB) domain-encoding region into the *ADAT1* coding sequence (Gerber 1998). Duplication of this ancestral *ADAR* gene, and subsequent coding sequence and domain architecture diversification, has led to the generation of the ADAR family.

ADAR family members exist in bilaterians and cnidarians (Jin et al. 2009; Keegan et al. 2011), and were recently identified in the genome of the ctenophore *Pleurobrachia bachei* (Moroz et al. 2014). They have not been found in the placozoan *Trichoplax adhaerens,* or in several non-metazoan eukaryotes, including choanoflagellates, fungi or plants, although these surveys have been limited in scope (Jin et al. 2009; Keegan et al. 2011). In this chapter, I identify and categorise ADAR protein family members present in the earliest branching metazoan lineages, including thirteen sponge and ten ctenophore species. I thus conclude that the full, or nearly full, repertoire of ADAR protein family members existed in the last common ancestor to all contemporary animals.

5.3 Methods

5.3.1 Sources of sequence data

We searched for ADAR candidates in the genomes of *Acropora digitifera* (Shinzato et al. 2011), *A. queenslandica* (Srivastava et al. 2010), *Aplysia californica* (Broad Institute 2009), *Arabidopsis thaliana* (The Arabidopsis Genome Initiative 2000), *Branchiostoma floridae* (Putnam et al. 2008), *Caenorhabditis elegans* (The C. elegans Sequencing Consortium 1998), *Capitella teleta* (Simakov et al. 2013), *Capsaspora owczarzaki* (Suga et al. 2013),*Ciona intestinalis* (Dehal et al. 2002), *Dictyostelium discoideum* (Eichinger et al. 2005), *Drosophila melanogaster* (Adams et al. 2000), *Helobdella robusta* (Simakov et al. 2013), *Hydra magnipapillata* (Chapman et al. 2010), *Lottia gigantea* (Simakov et al. 2013), *Mnemiopsis leidyi* (Ryan et al. 2013), *Monosiga brevicollis* (King et al. 2008), *Nematostella vectensis* (Putnam et al. 2007), *Neurospora tetrasperma* (Ellison et al. 2011), *Oscarella carmela* (http://www.compagen.org) (Nichols et al. 2012), *Pleurobrachia bachei* (Moroz et al. 2014), *Salpingoeca rosetta* (Fairclough et al. 2013), *Strongylocentrotus purpuratus* (Sea Urchin Genome Sequencing Consortium 2006), *Sycon ciliatum* (Fortunato et al. 2015) (details of analysed sequences available in Additional file 1 of (details of analysed sequences available in Additional File 1 of Grice and Degnan 2015b) *and Trichoplax adhaerens* (Srivastava et al. 2008). Transcriptome data was analysed from sponge species *Aphrocallistes vastus, Chondrilla nucula, Corticium candelabrum, Ircinia fasciculate, Petrosia ficiformis, Pseudospongosorites suberitoides, Spongilla lacustrus* and *Sycon coactum* (Riesgo et al. 2012), *Crella elegans* (non-reproductive tissue sample) (Pérez-Porro et al. 2013), *Ephydatia muelleri* (http://www.compagen.org), and *Clathria prolifera* (unpublished dataset, S. Fernandez Valverde and B. Degnan; details of analysed transcripts are provided in Additional File 1 of Grice and Degnan (2015b).

5.3.2 Identification of ADAR candidates from available draft genomes

HMMER 3.0 (Eddy 1998) was used to probe the unfiltered and filtered translated gene models from the genomes of each analysed species for AD domains (Pfam:PF02137) with a maximum Expect (e-) value of 0.001. As confirmation, the *H. sapiens* ADAR1 protein sequence (Ensembl: ENST00000368474) was used as a query for reiterative PSI-BLAST (position-specific iterative basic local alignment search tool) searches against the NCBI refseq protein database for each species in turn (Altschul et al. 1997), and also for BLAST (basic local alignment search tool) searches in the genome browsers for each species. Domain architecture of the hits identified by each method was determined using Pfam (Punta et al. 2011), and sequences containing ADAR-associated domains (AD, dsRB (Pfam: PF00035) and ZB (Pfam: PF02295) domains) were selected. To be counted, each domain had a maximum e-value of 0.001, however a small number of putative domains with higher e-values were manually compared to the Pfam seed domain sequences; those deemed to be of sufficient similarity were included in subsequent analyses. Where identical, or very similar, sequences were identified using different search methods, the hit from the translated gene model dataset was used. Accession numbers and sequence sources are listed in Appendix 5.1.

5.3.3 Preparation of translated sequences from sponge and ctenophore transcriptomes

Gene models for *Oscarella carmela* were predicted by submitting the whole genome assembly (http://www.compagen.org) (Hemmrich and Bosch 2008; Nichols et al. 2012) to the Augustus v2.6.1 program (Stanke et al. 2006). Augustus was run using the *A. queenslandica* training set, with settings singlestrand=true, alternatives-from-evidence=true and uniqueGeneId=true; all other settings were run as default. Predicted amino acid sequences were extracted from the resulting file. Translated peptide sequences for *Ephydatia muelleri* were downloaded from Compagen (http://www.compagen.org) (Hemmrich and Bosch 2008). For remaining transcriptome datasets, the longest open reading frame between stop codons was determined for each sequence, using the program getorf available in the EMBOSS v6.5.7 software package (Rice et al. 2000).

5.3.4 Identification of ADAR candidates from available sponge and ctenophore transcriptomes

Open reading frames were interrogated via hmmsearch and the domain architectures of resulting sequences were verified using Pfam, as for the genomic sequences above.

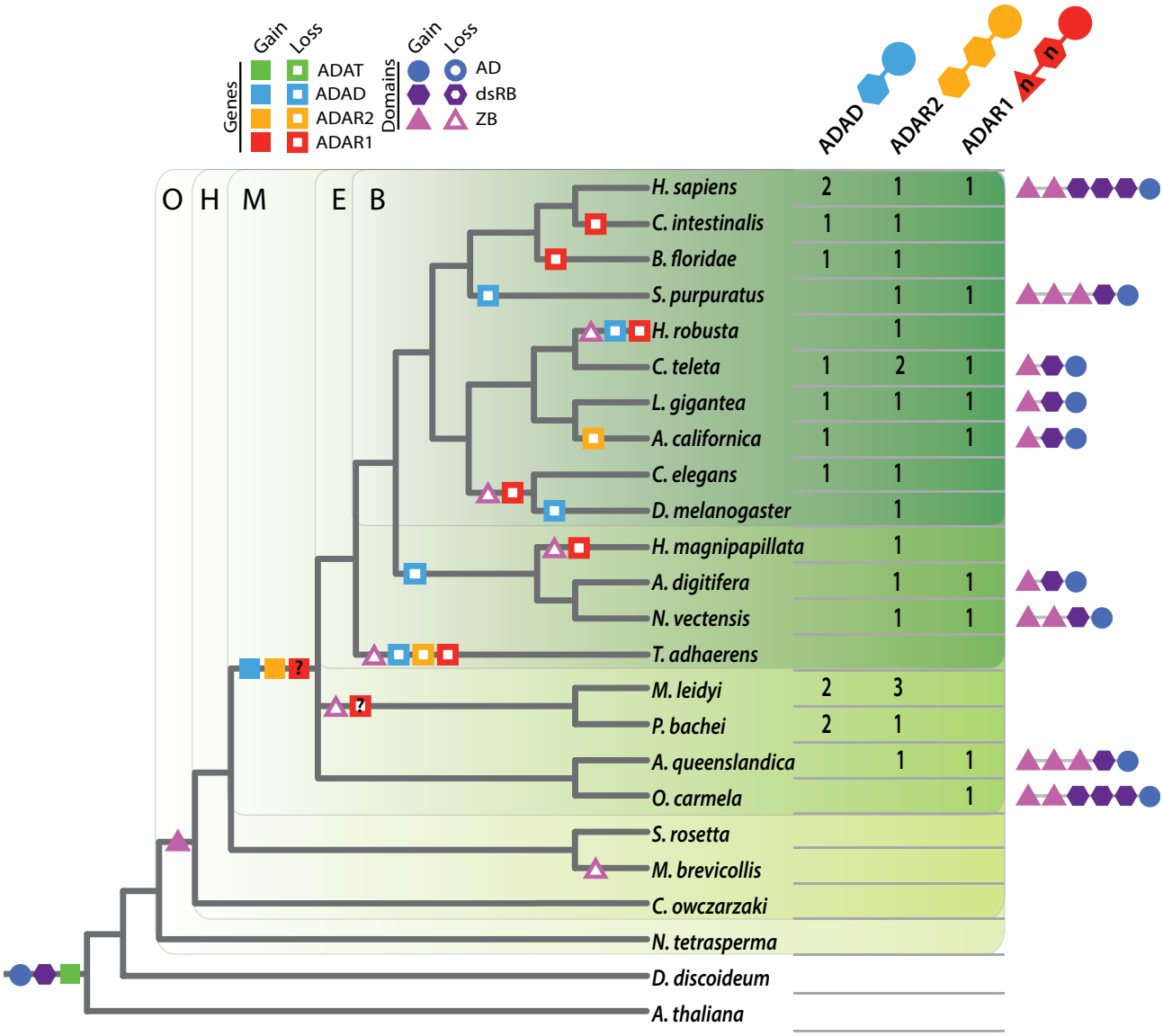


Figure 5.1 Reconstruction of ADAR gene and domain evolution
The table (right) lists the number of ADAR family members identified in each species. ADARs are classified based on their domain architecture, as shown by the 'ball-and-stick' protein models above each ADAR name. The Z-DNA/RNA binding (ZB) and double-stranded RNA binding (dsRB) domains of the ADAR1 model are marked with an 'n' to indicate that multiple copies of these domains may be present in different species. The domain architectures of all ADAR1-like proteins are depicted on the far right. The ADAR gene counts were used to reconstruct ADAT/ADAR evolution, as mapped to the phylogenetic tree as coloured squares (left). Searches for adenosine deaminase (AD), dsRB and ZB domains were performed to determine the phylogenetic positions of whole-genome domain origin and loss events, regardless of ADAT/ADAR complement; these events are also mapped to the tree as coloured shapes. Green boxes separate the tree into the main phylogenetic groupings: Bilateria (B), Eumetazoa (E), Metazoa (M), Holozoa (H) and Opisthokonta (O). For clarity, I present the sponge and ctenophore lineages on equal footing, and depict all three ADARs as present in the metazoan stem. The loss and gain of the ADAR1-like gene is marked with a question mark to illustrate the uncertainty in reconstructing these evolutionary events, which are elaborated upon further in Figure 5.3 and Figure 5.4.

Sequence redundancies were observed in the transcriptomes of a number of species. To counter this, I partitioned sequences into groups sharing over 90% sequence identity, using the default parameters of the tool cd-hit (Li and Godzik 2006), available via the CD-HIT Suite server (Huang et al. 2010). I assigned the representative sequence from each cluster, as determined by cd-hit, to its relevant ADAR category. ADAR family member counts were mapped to a sponge-ctenophore phylogenetic tree (Thacker et al. 2013; Moroz et al. 2014). Accession numbers of selected candidates are listed in Appendix 5.1.

5.3.5 Phylogenetic tree generation

AD domain sequences from non-bilaterian ADAD-, ADAR1- and ADAR2-like sequences were used to generate a multiple sequence alignment, generated with 100 iterations of the built-in MUSCLE algorithm (Edgar 2004) in Geneious Pro 5.0.2 (http://www.geneious.com) (Kearse et al. 2012). The *A. queenslandica* ADAT sequence Aqu1.212905 was also included as an outgroup. The alignment was manually refined in Geneious Pro, and submitted to the Gblocks webserver with the least stringent settings to further trim poorly-aligned regions (Castresana 2000; Talavera et al. 2007). The ProtTest 2.4 webserver (Abascal et al. 2005) was used to analyse the AD domain alignment and determine the best model selection method to use in generating phylogenetic trees, based on the AIC criterion. The best model was found to be LG+G. A maximum likelihood tree with 1000 bootstrap replicates was generated using the PhyML 3.0 webserver (Guindon et al. 2010), with the SPR method of tree improvement and five random starting trees. The resulting tree was visualized in FigTree 1.4 (http://tree.bio.ed.ac.uk/software/figtree/) and aesthetic modifications were made during manuscript preparation.

5.4 Results and Discussion

5.4.1 ADARs are present in the earliest branching metazoan lineages

I identified ADARs in a number of key opisthokont and eukaryote taxa for which a draft genome is available. HMM and BLAST-based search methods were used to identify AD domain-encoding genes, and domain architecture predictions were employed to narrow this list to likely ADAR candidates (Appendix 5.1). ADAR sequences can be partitioned into three categories based on their overall domain architecture (Figure 5.1): ADAD-like (one dsRB domain and one AD domain); ADAR2-like (two dsRB and one AD domain); and ADAR1-like (any number of Z-DNA/RNA binding (ZB; z-alpha) and dsRB domains and one AD domain). These categories are based on *Homo sapiens* gene names

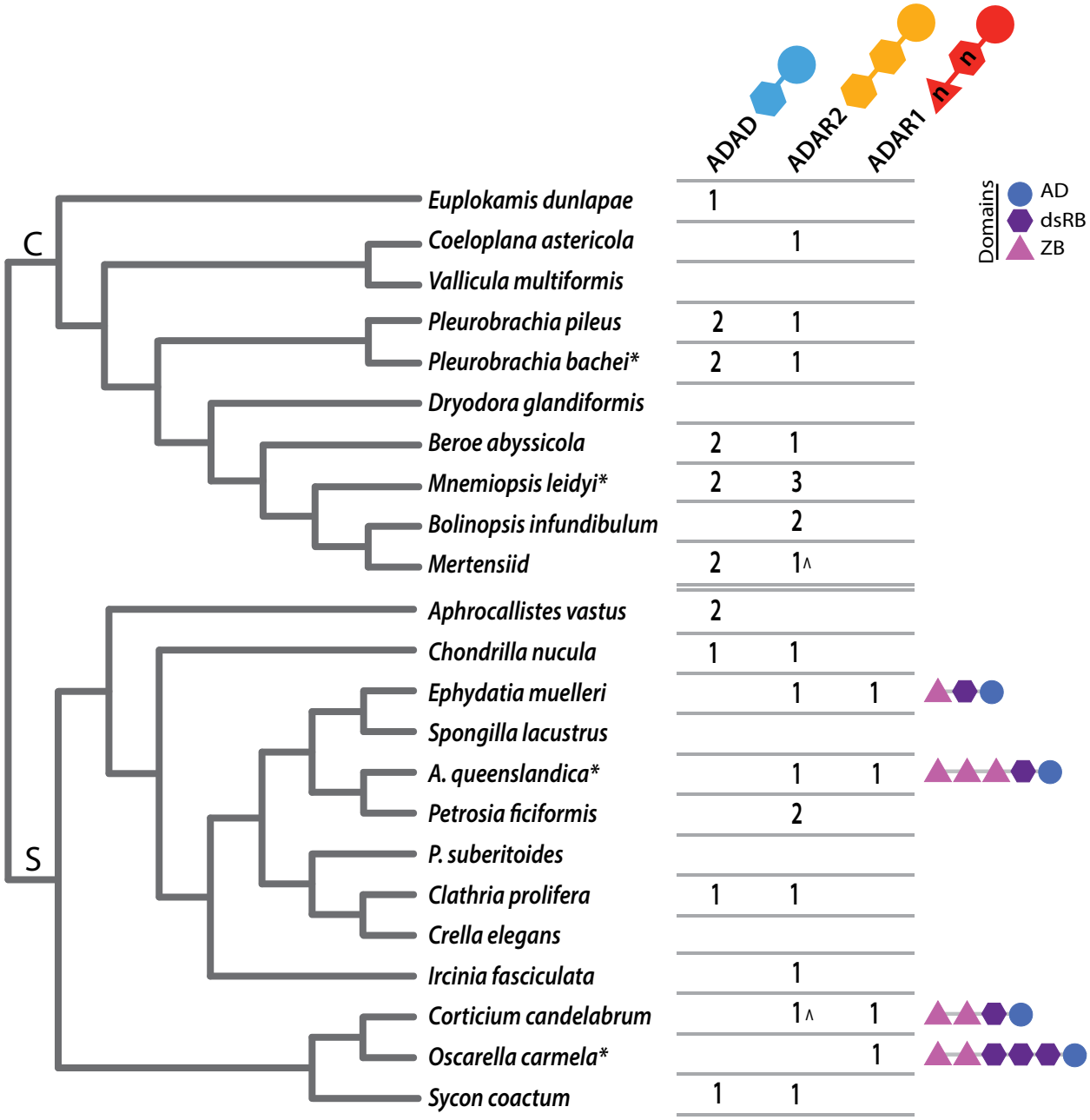


Figure 5.2 ADAR family member distribution in sponges and ctenophores
As in Figure 5.1, the number of candidate ADAR family members identified in each sponge and ctenophore genome (indicated by an asterisk) or transcriptome is shown. The domain architectures of ADAR1-like sequences are given on the far right. The phylogenetic relationships within the ctenophore (C, top) and sponge (S, bottom) lineages are depicted to the left. ADAR2 sequences indicated by a ^ are predicted to encode three dsRB domains. *A. queenslandica* and *Pseudospongosorites suberitoides* are abbreviated to conserve space.

and domain architectures. The *H. sapiens* ADAD sequence, while related to ADAR1 and ADAR2, is not implicated in RNA editing. ADAT-like sequences were identified in all species analysed (data not shown). I did not find evidence for ADAR3-like sequences in invertebrates, which possess an ADAR2-like architecture with an additional arginine-rich R-domain (Melcher et al. 1996).

I identified novel candidate ADAR genes in the genomic sequences of representative species of two of the earliest-branching animal lineages – sponges (*A. queenslandica* and *Oscarella carmela*) and ctenophores (*Mnemiopsis leidyi*); our methodology also isolated the ADAR candidates recently reported from the ctenophore *Pleurobrachia bachei* (Moroz et al. 2014). I identified one each of an *ADAR1*- and *ADAR2-like* gene in *A. queenslandica*, a single *ADAR1-like* gene in *O. carmela*, and two *ADAD*- and three *ADAR2-like M. leidyi* genes (Figure 5.1). Of the previously identified *P. bachei* ADAR candidates (Moroz et al. 2014), I categorised two sequences as *ADAD-like* and one as *ADAR2-like*, based on our domain architecture criteria (a comparison with candidates identified by Moroz et al. (Moroz et al. 2014) is provided in Appendix 5.1). Analysis of the *Sycon ciliatum* genome reveals that this calcarean sponge possesses *ADAD*-, *ADAR2*- and *ADAR1-like* genes (Appendix 5.1). The presence of multiple ADAR types in sponges, ctenophores and other invertebrates is consistent with the metazoan last common ancestor being already equipped with a suite of ADARs comparable to the repertoire that exists in humans and other modern bilaterians, and that ADAR gene and domain loss occurred independently in multiple metazoan lineages (Figure 5.1).

5.4.2 ADARs in the metazoan last common ancestor

Sponges and ctenophores are of significant evolutionary interest because they are considered the two earliest-branching metazoan lineages. However, questions remain as to whether sponge or ctenophores are the sister group to the rest of the Metazoa (Ryan et al. 2013). Although both taxa have multiple ADAR family members, all four examined species, *A. queenslandica, O. carmela, M. leidyi* and *P. bachei*, differ in their complement of ADAR genes. To facilitate a reconstruction of the evolution of the ADAR family, I searched for candidate ADAR sequences within the transcriptomes of an additional eleven sponge and eight ctenophore species (Figure 5.2; Appendix 5.1). Across the analysed sponge species, I identified candidate transcripts belonging to all three ADAR categories, *ADAD-, ADAR2*- and *ADAR1-like*. In no instance did a single species possess transcripts belonging

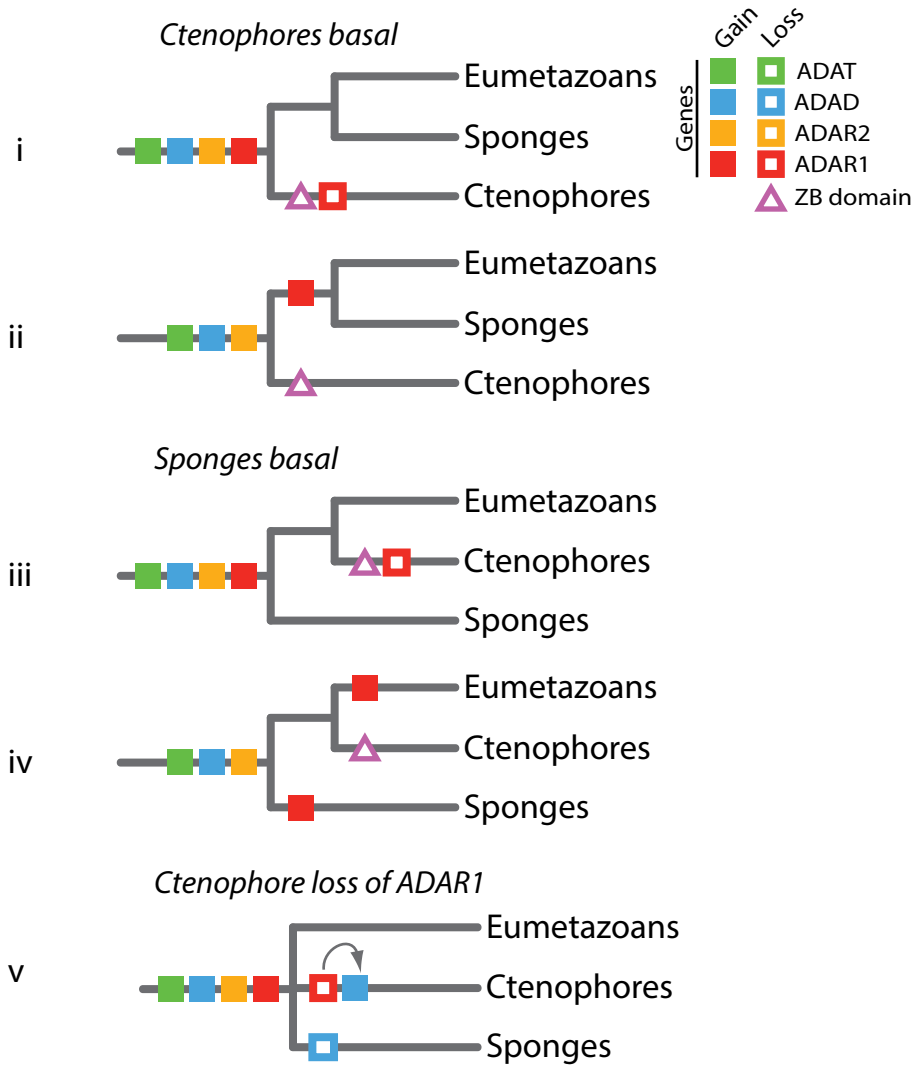


Figure 5.3 Possible scenarios for ADAR evolution in the metazoan ancestor
Five different scenarios of gene gain and loss events could explain the ADAR family distribution observed in sponges, ctenophores and eumetazoans, depending on whether sponges or ctenophores are the earliest-branching metazoan lineage. Filled and blank shapes represent gene (coloured squares) or ZB domain (triangles) gain and loss events, respectively. In panel v, the arrow represents the possible conversion of an ADAR1-like sequence to an ADAD-like architecture via domain loss.

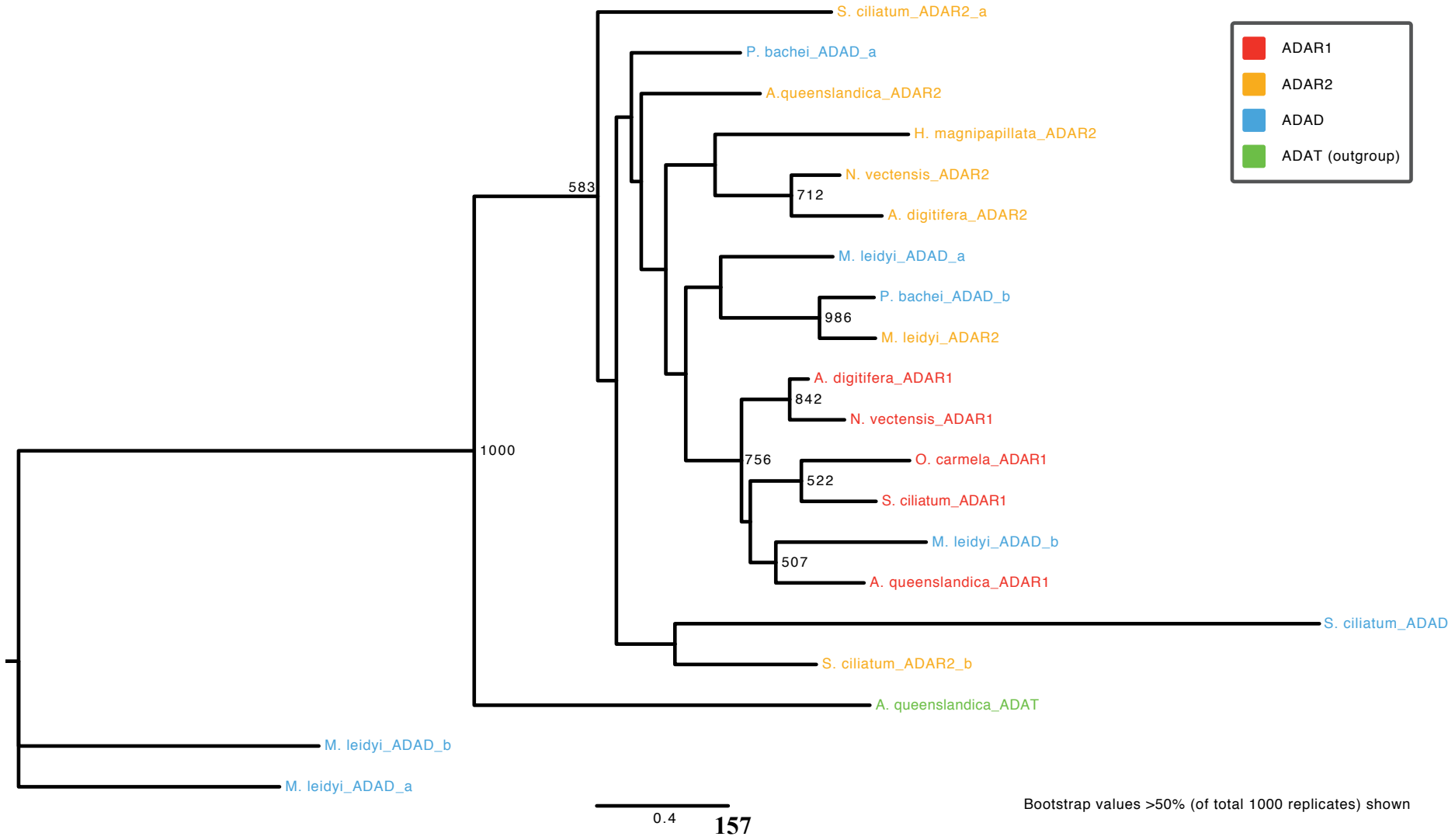
to all three ADAR types (Figure 5.2); *ADAD*-, *ADAR2*- and *ADAR1-like* genes however are present in the *S. ciliatum* genome (Appendix 5.1). In ctenophores, no *ADAR1-like* transcripts were identified in any species; only *ADAD*- and *ADAR2-like* transcripts were identified, either together or separately. It should be noted, as these searches were performed on transcriptome data, that the failure to identify ADAR family members in particular species is not necessarily indicative that these sequences are

absent from the genome; the overall lineage-specific trends do however allow insight into the taxonomic distribution of this protein family.

Until the relative phyletic positions of sponges and ctenophores are fully resolved, multiple reconstructions of ADAR evolution are obtained depending if sponges or ctenophores are the earlier-branching phylum. ADAD-, ADAR2- and ADAR1-like proteins are all present in the sponge lineage, but ADAR1-like proteins, and indeed ZB domains entirely (data not shown), are absent in ctenophores. From this I conclude that ADAT-, ADAD- and ADAR2-like sequences were all present in the metazoan ancestor. ADAR1-like proteins were either present and subsequently lost in the ctenophore lineage, or gained later. If ctenophores branch first, the *ADAR1-like* gene was either lost in this taxon, along with the ZB domain (Figure 5.3, panel i) or gained in the sponge + eumetazoan clade after diverging from ctenophores (Figure 5.3, panel ii). Alternatively, if sponges are the most basal metazoans, the *ADAR1-like* gene was either lost in ctenophores (Figure 5.3, panel iii) or gained independently in both the sponge and eumetazoan groups (Figure 5.3, panel iv). Scenario iv appears to be less likely, as it would require *ADAR1-like* genes to evolve twice. A phylogenetic analysis of the ADAR family-associated AD domains from all analysed non-bilaterian genomes provided poor resolution regarding the evolutionary relationships between ADAD-, ADAR2 and ADAR1-like sequences (Figure 5.4). However, as in earlier phylogenetic analyses of eumetazoan AD domains (Keegan et al. 2011), the AD domains from non-bilaterian ADAR1-like sequences were found to form a cluster with reasonable bootstrap support, suggesting that the ADAR1-like gene has undergone little diversification across evolutionary history. Interestingly, the AD domain of an *M. leidyi* ADAD-like gene is also present in this ADAR1-like AD domain cluster (Figure 5.4). This raises the possibility of a fifth evolutionary scenario of ADAR evolution (Figure 5.3, panel v) where the metazoan ancestor encoded all three ADAR family members, and that domain loss events converted a ctenophore ADAR1-like protein into a protein with ADAD-like architecture leaving ctenophores with two genes classifiable as ADAD-like. However, due to the poor bootstrap support for this tree overall, and as no *P. bachei* domain sequences are present in this cluster (Figure 5.4), it is currently unclear whether this result is evolutionarily significant.

Figure 5.4 Phylogenetic analysis of adenosine deaminase domains

Phylogenetic tree showing the relationship between AD domains from ADAD-, ADAR1- and ADAR2-like proteins. The tree was run with 1000 bootstrap replicates; bootstrap values greater than 500 are shown. While several branch points are not well supported, the ADAR1-like AD domains (and an additional ADAD-like sequence from *Mnemiopsis leidyi*) form a bootstrap-supported cluster. The *A. queenslandica* ADAT gene AD domain is included as an outgroup but was not explicitly designated as such for tree generation.



5.4.3 Domain architecture of the ADAR1-like genes

ADAR1-like genes were identified in a diverse set of metazoans, and are present in a variety of domain conformations (Figures 5.1-2, far right). Human and other vertebrate *ADAR1* genes encode two ZB, three dsRB, and one AD domain, while the sea urchin *Strongylocentrotus purpuratus* genome encodes a protein equipped with three ZB, one dsRB and one AD domain. The *Nematostella vectensis* ADAR1 protein possesses two ZB (one of which is divergent), one dsRB and one AD domain. All ADAR1-like proteins identified in the other studied non-deuterostome eumetazoan taxa encode one copy each of the ZB, dsRB and AD domains. Interestingly, a diversity of domain architectures are encoded amongst the *ADAR1-like* genes and transcripts of sponges. In *A. queenslandica,* the *ADAR1-like* gene encodes three ZB, one dsRB and one AD domain, identical to the architecture of the *S. purpuratus* ADAR1, while the *O. carmela* gene encodes the vertebrate-like domain complement of two ZB, three dsRB and one AD domain (Figures 5.1-2); the *S. ciliatum* genome encodes an ADAR1-like protein with two ZB, one dsRB and one AD domain (Appendix 5.1). I also identified *ADAR1-like* transcripts from *Ephydatia muelleri* and *Corticium candelabrum*. These sequences both possess one dsRB and one AD domain, and the *E. muelleri* sequence contains one ZB domain while the *C. candelabrum* sequence has two (Figure 5.2).

The diversity of ADAR1-like architectures present in modern sponges complicates the resolution of the ancestral ADAR1-like form. However, a combination of one ZB, one dsRB and one AD domain remains the most parsimonious ancestral conformation; this form is seen within the sponge lineage (*E. muelleri*) and in all analysed non-deuterostome eumetazoan species except *N. vectensis*. ADAR1-like domain diversification has occurred in the sponge lineage, perhaps indicative of molecular tinkering allowing the testing and retaining in various species of different ADAR1-like domain architecture combinations. It is currently unknown whether similar levels of interspecies diversity exist in other phyla or classes.

5.4.4 Origin of the metazoan ADAR protein family

ADAT genes are present throughout eukaryotes and are responsible for the deamination of adenosine into inosine for tRNA functionality (Gerber 1998). Although AD and dsRB domains evolved prior to eukaryotic cladogenesis (Figure 5.1), the first evidence of these domains coming together to

form an ancestral ADAR exists in the lineage leading to the crown Metazoa. This is likely to have occurred when a duplicated *ADAT* gene was coupled to a gene or part of a gene encoding one – or possibly more – dsRB domains, via domain shuffling. It appears most plausible that the first ADAR had one copy each of a dsRB and AD domain and thus was ADAD-like. This new gene then duplicated and incorporated a second dsRB domain, forming an *ADAR2-like* gene. The formation of the *ADAR1-like* gene involved the incorporation of one or more ZB domains into either an *ADAD-* or *ADAR2-like* gene. It is not clear which of these two family members was the original acceptor for the ZB domain, however, the combination of a single ZB and dsRB domain together in a number of species (Figures 5.1-2, far right) suggests the former is more likely. The ADAR suite was thus in place early in metazoan history. Minor alterations, namely gene loss and duplication events, have occurred in some animal lineages (Figures 5.1-2), but dramatic expansion and diversification events do not characterise the evolutionary history of the ADAR family.

5.4.5 Conclusions

The ancestral role of the ADARs is currently unknown. Indeed, the biochemical functionality of basal metazoan ADAR protein family members in A-to-I editing remains to be tested experimentally. The existence of a diversified gene family in the earliest branching lineages of animals, but not in their close unicellular holozoan and fungal relatives, is consistent with this gene family being an animal-specific innovation. The evolution of metazoan multicellularity and complexity was accompanied by a wide range of genomic innovations (Srivastava et al. 2010). The origin and expansion of the ADAR gene family occurred prior to the diversification of crown metazoans, as is the case for microRNAs and piwiRNAs, and many transcription factor and signalling pathway families(Grimson et al. 2008; Degnan et al. 2009; Richards and Degnan 2009). The maintenance of the ADAR gene family in most modern phyla suggests that RNA editing was and remains an essential part of the genomic zootype and metazoan regulatory toolkit.

The existence of the ADAR gene family in the *A. queenslandica* genome implies that RNA editing occurs in this species as a means of post-transcriptional regulation or diversification. Functional studies are required to confirm that A-to-I editing occurs in sponges as has been characterised in bilaterians. The gene targets of this putative editing are currently unknown, however it remains possible that RNA

editing acts to diversify the *A. queenslandica* aggregation factor (*AqAF)* genes, further to the alternative splicing and genomically-encoded nucleotide polymorphism reported in Chapter 4. Preliminary studies into this question are currently ongoing in the Degnan lab, but suggest that RNA editing may indeed act upon certain regions of the *AqAFs* in some individuals (K. Roper, personal communication).

In Chapters 2 to 4, I examined the sequence properties and activity of the *AqAFs* across sponge development and in a normal, unchallenged biological context. For the final element of this research, I investigate the qualitative (i.e. splicing) and quantitative (i.e. expression profiles) responses of the *AqAF* genes to auto- and allogeneic challenge. I performed auto- and allografts between pairs of sponges, and followed the grafted samples for up to three days, before generating transcriptomes for each time point within the experiment using RNA sequencing. In Chapter 6, I examine the *AqAF* splice variants and quantitative expression profiles across the graft transcriptomes, and identify other non-AF genes that differentially respond to graft challenge.

CHAPTER 6 - TRANSCRIPTOMIC PROFILING OF THE ALLORECOGNITION RESPONSE TO GRAFTING IN THE DEMOSPONGE *AMPHIMEDON QUEENSLANDICA*

6.1 Abstract

Sponge grafting experiments simulate the effects of tissue contact between self or nonself individuals in the field. Previous graft studies in the demosponge *Clathria prolifera* showed that the aggregation factor (AF) genes in this species are upregulated in response to self and nonself tissue contact, and that AF proteins accumulate at the site of nonself contact. I took a transcriptomic approach to investigate AF activity in self and nonself grafts in *Amphimedon queenslandica*. I performed a series of auto- and allografts, and observed and sampled these over a period of three days, before generating fourteen transcriptome datasets spanning the auto- and allograft response. The AF genes are highly but stably expressed across the auto- and allograft time courses. A number of putatively alternatively spliced AF transcripts were expressed in grafted tissue, including some that encoded novel signal peptides. On a genome-wide scale, the nonself graft response appears to involve a broad downregulation of normal biological processes, rather than the mounting of an intense defensive response.

6.2 Introduction

Coral reefs are densely-populated ecosystems that display remarkable levels of biodiversity. In such a crowded environment, space can become a limiting resource, and sessile invertebrates in particular often face intense competition for habitat and growth space. For example, one study determined that 42% of microhabitats (i.e. gastropod shells) for the colonial hydrozoan *Hydractinia echinata* must be shared between two or more colonies (Yund et al. 1987). Similar population crowding has been observed at Woods Hole, Massachusetts, where multi-individual clumps of the sponge *Clathria prolifera* were identified at relatively high (20%) frequencies within the population (Fernàndez-Busquets and Burger 1997). Crowding in reef ecosystems means the chance of direct contact between conspecific individuals or members of different species is high. Conspecific tissue fusion can at times be beneficial, for example by allowing an individual to re-fuse with its own tissue following fragmentation or growth around an

object, or through increased survivorship and subsequent reproductive output associated with increased size (Bonner 1966; 1988; 2000). However, there is often a cost associated with conspecific fusion, since individuals within a chimera are at risk of parasitism whereby the stem cells of one fusion partner gain disproportionate access to the germ line and monopolise reproductive output (Buss 1982). For this reason, tissue fusion is generally limited to genetically-identical individuals or close kin (Grosberg 1988). The decision to fuse with or reject a potential partner is mediated by the allorecognition (or self-nonself recognition) system.

6.2.1 Sponge immune challenges

The sponge has been a useful model organism for the study of cell adhesion and self-nonself recognition systems for almost 150 years, with adult tissue grafting experiments first described in 1869 (Vaillant). Sponge grafts aim to experimentally emulate the effects of natural sponge-sponge contact, as may occur between two regions of a single sponge individual due to wound repair or growth around a jagged substrate (self), or between different individuals due to overgrowth (nonself). Grafting is performed by apposing two pieces of sponge tissue, either from different parts of the same sponge (autograft) or from two different sponges of the same (allograft) or different (xenograft) species (Moscona 1968; Hildemann et al. 1979; Jokiel et al. 1982; Neigel and Avise 1985; Ilan and Loya 1990; McGhee 2006; Gauthier and Degnan 2008). These experiments have demonstrated that sponges are capable of distinguishing between self and nonself. Adult tissue fusion is limited almost exclusively to autografts, although fusion between different sponge individuals has been observed in rare cases at rates inversely proportional to the physical distances between sponge graft partner habitats (Jokiel et al. 1982; Neigel and Avise 1985; McGhee 2006). This trend can be explained broadly by the general decrease in genetic similarity between individuals with increasing distance (Jokiel et al. 1982; Fernàndez-Busquets and Burger 1999). It appears in at least some cases, however, that compatible sponges represent clonally-reproduced derivatives of a single genetic individual (Jokiel et al. 1982).

Typical self grafts that undergo fusion are characterised by the breakdown of the pinacoderm layers separating the two pieces of tissue, with the interface between the graft donors becoming invisible over time (Ilan and Loya 1990; Fernàndez-Busquets and Burger 1997; Fernàndez-Busquets et al. 2002). Responses to allografts, however, vary extensively even within a single sponge genera (Van

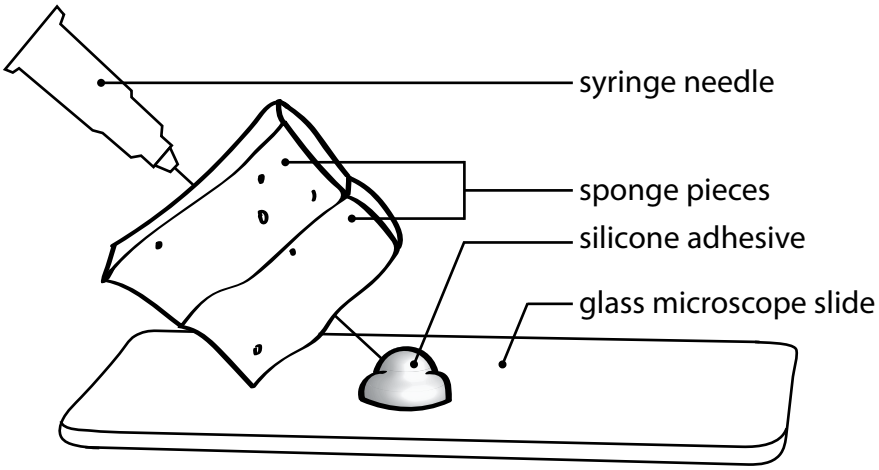


Figure 6.1 Sponge graft setup
Sponge pieces of approximately 1.5 x 3 cm were placed with cut surfaces touching, and were held together with a fresh syringe needle. The needle was stuck in a mound of dried silicon glue on a labeled glass microscope slide to hold the graft underwater, with the plastic attachment of the needle positioned such as to prevent the two tissue pieces from separating.

de Vyver and Barbieux 1983). Reactions can be fast, such as in *Clathria prolifera*, which responds to allografting in two to six hours (Humphreys 1994; Fernàndez-Busquets and Burger 1997), or slow, as in *Callyspongia diffusa*, which can take up to a week to react (Hildemann et al. 1979; 1980; Bigger et al. 1981; Yin and Humphreys 1996; Fernàndez-Busquets and Burger 1997; 1999). Processes that characterise graft rejection may include tissue necrosis of one or both graft partners (Hildemann et al. 1979; 1980; Bigger et al. 1981; Fernàndez-Busquets and Burger 1997; 1999), collagen deposition to form a physicochemical barrier between the apposing tissue (Van de Vyver 1975; Kaye and Ortiz 1981; Buscema and Van de Vyver 1983; Van de Vyver and Barbieux 1983; Humphreys and Reinherz 1994; Humphreys 1994; Fernàndez-Busquets and Burger 1997), cellular migration to the point of contact (Curtis et al. 1982; Van de Vyver and Barbieux 1983; Humphreys 1994; Humphreys and Reinherz 1994; Fernàndez-Busquets and Burger 1999; Fernàndez-Busquets et al. 2002), and phagocytic or cytotoxic reactions (Hildemann et al. 1980; Bigger et al. 1981; Van de Vyver and Barbieux 1983; Yin and Humphreys 1996). Qualitative and quantitative responses to tissue grafts are replicable and predictable (Hildemann et al. 1980; Hildemann and Linthicum 1981; Fernàndez-Busquets and Burger 1997), between both first-party (sponge A:B replicates) and third-party (where A:B fusion predicts identical A:C and B:C reactions) grafts (Bigger et al. 1981; Kaye and Ortiz 1981; Neigel and Avise 1985). This specificity and repeatability indicates that recognition responses are governed by an

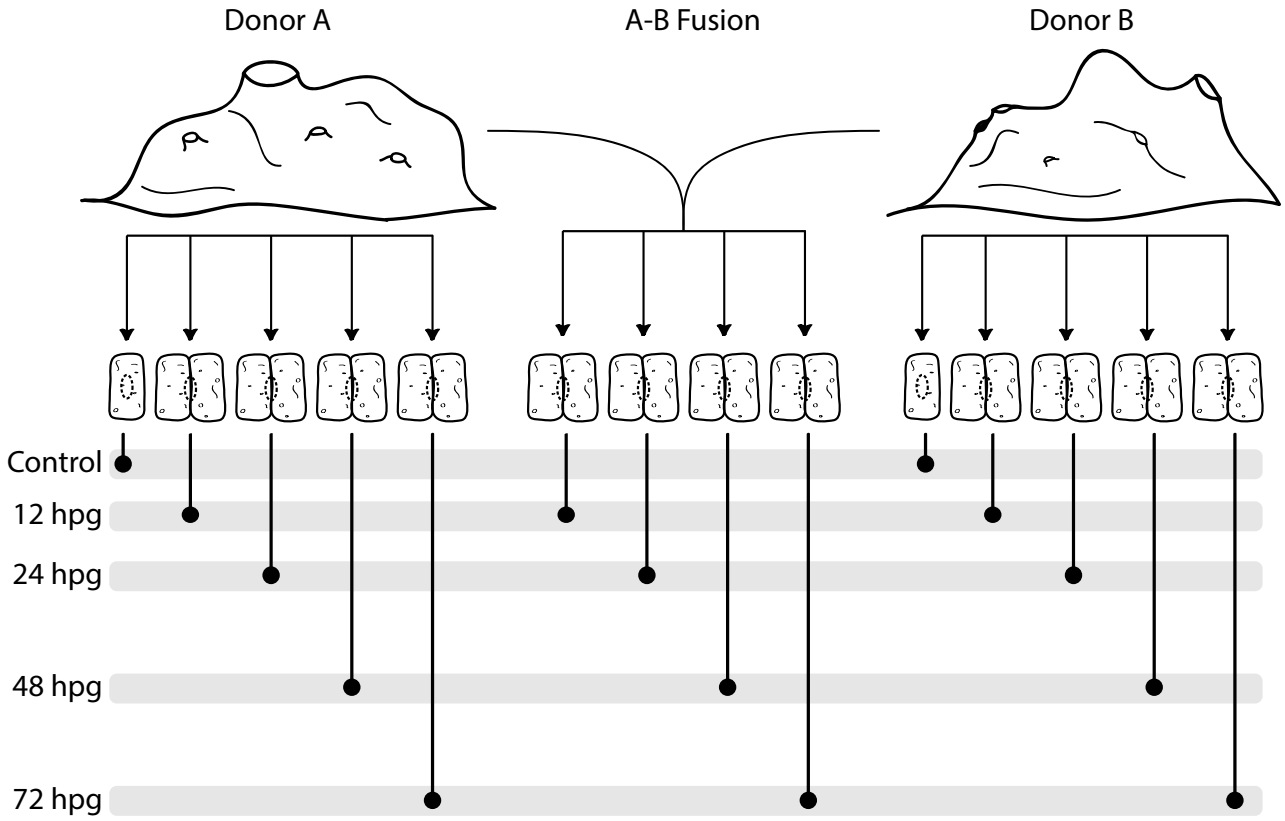


Figure 6.2 Graft sampling regime
Two sponge individuals were used to perform autografts and allografts. For autografts, two pieces of either sponge Donor A or Donor B were apposed and kept in running sea water until sampling at 12, 24, 48 or 72 hours post grafting (hpg; black dots). Allografts were performed by bringing two pieces of tissue, one from each sponge donor, into contact and sampling across the same time course as for the autografts. Samples of pre-grafted tissue were also taken from Donors A and B prior to preparing the grafts (Control). At the time of sampling, a small slice of tissue was taken at the points indicated (dashed circle). Each graft was discarded after sampling.

underlying polymorphic genetic system, rather than by environmental or random effects (Fernàndez-Busquets and Burger 1999).

6.2.2 Aggregation factors in sponge tissue grafts

In addition to their well-characterised role in mediating species-specific cellular reaggregation (Chapter 1.3.2), and their potential developmental function/s (Chapter 3), sponge aggregation factors (AFs) are putatively involved in the individual-specific response to tissue grafting. The AF core protein-coding genes *MAFp3* and *MAFp4* appear to be upregulated in *C. prolifera* auto- and allografts compared with normal tissue (Fernàndez-Busquets et al. 1998), and MAFp3 protein accumulates at the site of allogeneic contact (Fernàndez-Busquets et al. 1998). Additionally, the AF genes in *C. prolifera*

(CpAFs) (Fernàndez-Busquets and Burger 1997) and *A. queenslandica* (AqAFs; Chapter 4.4.3) are highly polymorphic within and between individuals, indicating that the AFs fulfil this requirement of a self-nonself recognition system. Comparisons of *C. prolifera* graft response and CpAF polymorphism, as measured by the restriction fragment length polymorphism (RFLP) profiles of each individual, have also revealed a ~100% correlation between RFLP profile similarity/dissimilarity and fusion/rejection outcomes (Fernàndez-Busquets and Burger 1997). These findings all demonstrate a correlation between the AFs and alloimmune challenge.

Table 6.1 Graft nomenclature

	SELF		NONSELF
	AA	BB	AB
Donor	DA	DB	n/a
12 hpg	T12AA	T12BB	T12AB
24 hpg	T24AA	T24BB	T24AB
48 hpg	T48AA	T48BB	T48AB
72 hpg	T72AA	T72BB	T72AB

6.2.3 Introduction to the study

In previous chapters, I characterised the six AF genes (*AqAFA – AqAFE*) and transcripts from *A. queenslandica* and other sponge species (Chapter 2), and examined *AqAF* developmental gene expression (Chapter 3), nucleotide polymorphism, and alternative splicing (Chapter 4) under normal, unchallenged conditions. For the final portion of this thesis, I sought to place the activity of these genes in context, by studying *AqAF* expression levels and alternative splicing across the physiological response to self and nonself graft challenge in *A. queenslandica*. I performed a series of autograft and allograft time course experiments, and generated the first multi-transcriptome dataset from one such experiment to follow sponge auto- and allograft response over time. I examined the expression profiles of the *AqAFs* across the graft response using both this dataset and qPCR (real-time quantitative polymerase chain reaction). Alternatively spliced *AqAF* transcript variants were also identified and characterised from the graft trancriptomes. Finally, the transcriptome dataset was surveyed to identify the broader functional changes that occur in response to nonself grafting challenge in *A. queenslandica*.

6.3 Methods

6.3.1 Tissue grafting of adult sponges

Four grafting experiments were performed in total. For each experiment, two large adult *A. queenslandica* specimens that were not growing in the immediate vicinity of one another were collected from Shark Bay, Heron Island (Great Barrier Reef, Australia) (Leys et al. 2008). The sponges were transported to the laboratory at Heron Island Research Station and were maintained outdoors, but shaded, in tanks of constantly-flowing unfiltered sea water that was pumped off the reef flat. Three graft time courses, two self and one nonself, were produced from the two sponge donor specimens within each experiment.

To prepare the grafts, each sponge was removed from its rocky substrate, and a small sample of donor sponge tissue was taken and placed in RNA Later (Ambion) to serve as the control (0 hours post grafting, hpg) time point for each sponge. Each sponge was cut into twelve pieces of about equal size (approximately 3 x 1.5 cm). Autografts and allografts were prepared by apposing two pieces of tissue from the same (autograft) or different (allograft) individual, with their internal cut surfaces touching. Each graft was skewered together with a fresh syringe needle, which was stuck into a mound of dried silicon glue on a labelled glass slide (Figure 6.1). To minimise sample handling, a separate graft was examined and sampled at each time point. Therefore, each experiment comprised twelve graft samples - four self grafts from each of the two sponges, and four nonself grafts (Figure 6.2). The grafts were kept in a tank with flow-through sea water, and exposed to ambient, shaded light, until they were due to be sampled.

Auto- and allografts were sampled at 12, 24, 48 and 72 hours post grafting (hpg; Figure 6.2). At each time point, one graft from each of the one nonself and two self time courses was retrieved and taken to the lab for observation and tissue sampling. Each graft was removed from its attached slide and needle and briefly examined to assess tissue health and fusion state. Graft pairs were separated where this could be done gently and without excessive force. Small slices of tissue were taken from the graft interface, taking care to take approximately equal amounts of tissue from each side. The samples were then placed in RNA Later. Grafts were discarded after sampling.

Table 6.2 Transcriptome sequencing

LIBRARY	TOTAL BASES	READ COUNT	TRIMMED READ COUNT	GC (%)	Q20 (%)	Q30 (%)
Donor A	2,809,914,132	27,820,932	26,228,938	42.1	96.3	91.2
Donor B	2,663,155,678	26,367,878	25,007,728	41.9	96.6	91.7
T12 AA	2,435,185,144	24,110,744	22,538,612	41.4	95.9	90.6
T12 BB	2,494,153,186	24,694,586	23,222,990	41.9	96.2	91.1
T12 AB	2,682,231,144	26,556,744	24,878,304	40.1	96.3	91.2
T24 AA	2,229,386,534	22,073,134	20,895,012	41.5	96.5	91.5
T24 BB	2,249,828,934	22,275,534	20,912,980	41.1	96.2	91
T24 AB (A)	2,109,872,022	20,889,822	19,686,010	42.1	96.3	91
T24 AB (B)	2,006,084,624	19,862,224	18,581,488	42	95.9	90.5
T24 AB (C)	1,762,637,456	17,451,856	16,136,524	41.7	95.4	89.5
T48 AA	2,475,671,196	24,511,596	23,032,818	41.5	96.2	91.1
T48 BB	2,356,682,894	23,333,494	21,971,450	40.9	96.4	91.3
T48 AB	2,041,949,926	20,217,326	19,135,330	43.1	96.4	91.2
T72 AA	2,277,699,076	22,551,476	21,139,356	41.1	96.1	90.9
T72 BB	2,421,596,402	23,976,202	22,378,750	42.1	95.8	90.3
T72 AB	2,477,912,992	24,533,792	23,158,336	42	96.4	91.4

6.3.2 Graft sample nomenclature

For one of the four graft experiments, tissue samples were prepared for whole-transcriptome sequencing and qPCR (Chapter 6.3.3). A system of nomenclature was developed to identify each graft time point within the experiment (Table 6.1). The two control samples are referred to as 'Donor A' and 'Donor B' respectively. Self grafts are designated as 'AA' or 'BB' depending on their sponge of origin, and nonself grafts as 'AB'. Each sample was given a name based on the time of sampling (Donor, T12 to T72) and sponge of origin (A or AA, B or BB, AB). Therefore, T12AA refers to the self graft derived from the Donor A sponge that was examined at 12 hpg, and so forth.

6.3.3 RNA extraction from graft tissue

RNA from the selected graft time course was extracted and prepared for whole-transcriptome sequencing and qPCR. Separate extractions were performed for the two applications. For all extractions, a total of 200 mg tissue per extraction was used (100 mg tissue from each side of the graft interface, where applicable). All centrifugations were performed at 14,680 rpm (revolutions per minute). Tissue was added to 800 µL Tri Reagent (Sigma), heated to 55oC for 30 minutes, and briefly ground to maximise RNA release. An additional 200 µL Tri Reagent was added, and samples were left at room temperature for 5 minutes before centrifugation for 10 minutes at 4oC. The supernatant was collected, vigorously mixed with 100 µL bromochloropropane (BCP), left at room temperature for 15 minutes, and then centrifuged for 15 minutes at 4oC. The resulting top aqueous layer was combined with 250 µL each of isopropanol and high-salt precipitation solution (0.8 M sodium citrate, 1.2 M NaCl). After a 10 minute incubation at room temperature, the sample was centrifuged for an additional 10 minutes at 4oC. The supernatant was discarded and a standard 70% ethanol wash was performed on the pellet. Each pellet was eluted in DNase and RNase-free distilled water (Gibco, Life Technologies). RNA quantity and quality was checked using a Qubit 2.0 (Invitrogen by Life Sciences) and Bioanalyser 2100 (Agilent).

6.3.4 Transcriptome sequencing

RNA samples were submitted to Macrogen Ltd. (Seoul, Korea) for RNA-Sequencing (RNA-Seq) high throughput sequencing following a polyA-selection, 100 base pair, paired-end, unstranded Illumina HiSeq 2000 protocol. Samples were multiplexed with eighteen libraries run on a single lane of the Illumina flow cell. For the

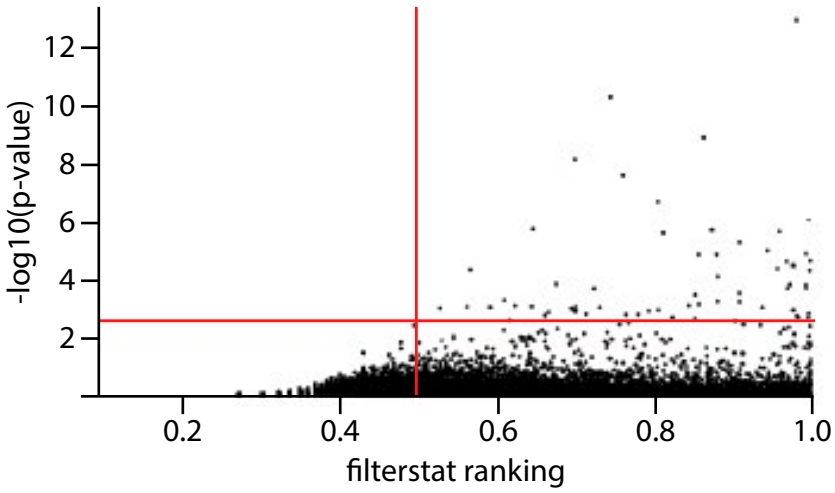


Figure 6.3 Analysis of independent filtering criteria
The scatterplot shows all *A. queenslandica* genes in rank order of expression across the graft time course (x-axis, scaled 0 to 1), against the negative log p-values for each gene (y-axis). The red lines indicate that the 50% of genes with the lowest read counts (vertical) do not achieve an unadjusted p-value less than 0,003 (horizontal; ~2.5 on the –log10 scale), and can therefore be eliminated without negatively affecting downstream analysis.

T24AB time point, 3 different RNA extractions were performed using the original tissue sample. All three RNA samples were sequenced due to initial concerns about RNA quality and quantity. These samples were named T24AB_A, B and C. General sequencing statistics are provided in Table 6.2.

Table 6.3 Quartile distributions of genewise read counts

QUANTILE	READ COUNT
25%	0
50%	53
75%	1,367
100%	901,606
Mean	3,139

6.3.5 Transcriptome preparation and analysis

Sequenced transcriptome libraries were evaluated to determine overall sequencing quality, using FastQC 0.10.0 (non-interactive mode, run with Java 1.6.0_22; http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc). In light of these results, Trimmomatic 0.22 (Bolger et al. 2014) was used to trim poor-quality bases or entire reads, using a headcrop length of 13 base pairs (bp), a sliding window size of 4 bp and average quality of 15, and a minimum read length of 36 bp. All other settings were run with their default values. Trimmed read counts are provided in Table 6.2. The quality of remaining paired reads was again verified using FastQC prior to further analysis. All three T24AB samples were deemed to be of sufficient quality for this experiment; sequence reads from three all samples were pooled in further analyses unless otherwise stated.

6.3.6 Read mapping and counting

Gene-level read counts were determined by mapping trimmed sequencing reads to the Aqu2.1-model annotated *A. queenslandica* genome using the CLC Genomics Workbench 6.5.1 RNA-Seq tool (CLCbio) with default parameters. An artificial nonself 'donor' sample was also generated by combining the Donor A and B reads in a single analysis. Two count matrices were generated, with columns corresponding to different samples and rows to the Aqu2.1 gene models; the first table showed RPKM (reads per kilobase of transcript per million mapped reads) values for principal component analysis (PCA; Chapter 6.3.7) and the second showed total gene-wise read counts for differential expression analysis (Chapter 6.3.8-9).

Table 6.4 Details of qPCR primer pairs

Gene	Forward Primer Sequence	Forward Tm (°C)	Reverse Primer Sequence	Reverse Tm (°C)	mRNA product length (bp)
*AqECH**	GGTGAACGTATTGGTGAGTTC	60.9	GTTTCTCAAGGAAGGCAGTC	60.5	172
*AqGADP**	GCACCTTCTGCTGATGCT	61.8	ACGACCATCACGCCATTT	64.1	147
*AqHPRT**	CAGACGATGAAAACAAGACTG	60	TAGTAATGAGCAGGGACACAG	59.4	127
*AqILF2**	GCACTGAAAAGGAGGAAAGA	60.9	TGTACCAAAACCTTGAACACGA	64.1	191
*AqNFkB**	TCTCTTACAGCAAACAATCCTC	60.6	CTTACCACAGAGAGATTCATTGAC	61.3	156
*AqSDHA**	CGGGGAGTGGTAGCTATGAA	63.8	TGAAACTGTACAAACTCCATGTCT	61.4	194
AqAFA	GTCTGTGGCACTGGGTCTA	61.4	CAGGCTCTGCTCCAGTAAC	60.2	157
AqAFB	CTCACTCCACCTCCAGAAG	60	GGGAAGAGAGAGTGGAAGG	60.4	160
AqAFC	GTGGCAGCTAGCGATACAG	61.2	CCGTCTCTCCTTCTGAGAC	59.1	100
AqAFD	GATGGTACCCTTCGTCCTG	62	CTGACCAGCCTGAGTCCTA	60.6	116
AqAFE	CAGGAGAGAGTGTGCTGTC	58.6	CAGAGGTCAGAGAGGAGGT	58.5	156

Sequences and melting temperatures (Tm) for each individual primer are given, as well as the expected product length. A 58°C primer annealing temperature and 30 s extension time was set for all PCR reactions using these primers. *Denotes candidate reference gene

6.3.7 Principal component analysis

RPKM values were used as input for PCA using BLIND (Anavy et al. 2014). BLIND was run with default parameters, to examine the 0.9th quantile of dynamically expressed genes as selected by the program, with sample order determined using a measure of sample entropy, and results scaled using the percentage of scaled variance.

Table 6.5 qPCR primer amplification efficiency

	EFFICIENCY	ERROR	SLOPE
Optimal	1.8 - 2.2	<0.2	-3.1 to -3.58
AqAFA	2.082	0.023	-3.139
AqAFB	1.848	0.013	-3.751
AqAFC	1.729	0.031	-4.204
AqAFD	1.995	0.027	-3.333
AqAFE	1.876	0.006	-3.658
AqECH	1.731	0.020	-4.197
AqGAPD	2.016	0.008	-3.283
AqHPRT	1.823	0.012	-3.834
AqILF2	1.856	0.051	-3.723
AqNFKB	1.951	0.025	-3.446
AqSDHA	1.913	0.027	-3.548
AqYWHAZ1	1.904	0.007	-3.575

6.3.8 Assessment of filter statistics for independent filtering

Multiple testing correction in differential gene expression (DGE) analysis is important in order to reduce the number of false positives in the resulting dataset, however, such corrections can also reduce the detection power of the analysis (Dudoit et al. 2003). Detection power can be improved by reducing the number of tests required in an analysis (Bourgon et al. 2010), for example by filtering lowly-expressed genes that would be unlikely to be flagged as significantly differentially expressed if they were included. Deletion of these statistically uninformative genes, and therefore reduction of the number of required statistical tests, can potentially allow detection of a greater number of statistically significant expression changes than if the dataset was not filtered (Bourgon et al. 2010).

The Bioconductor packages genefilter (v1.46.1; http://bioconductor.org/packages/release/bioc/html/genefilter.html) and DESeq (v1.16.0) (Anders and Huber 2010) can be used together to determine an optimal filtering threshold, where genes with count values in the bottom n-th percentile (when gene-wise counts are summed across all samples) can be removed from the dataset without losing genes

that would be flagged as significantly differentially expressed. For the present study, this analysis was performed as per the genefilter vignette 'Diagnostics for independent filtering' (http://bioconductor.org/packages/release/bioc/html/genefilter.html).

Table 6.6 qPCR thermocycling conditions

STAGE	TEMPERATURE - TIME
Denaturation (1x)	95°C - 10 min
Cycling (50x)	95°C - 5 s 58°C - 10 s 72°C - 45 s
Melt (1x)	97°C - 10 s 55°C - 30 s 95°C - na
Cool (1x)	50°C - 30 s

A script describing the full analytical methods to perform the independent filtering analysis is provided in Appendix 6.1. Briefly, the total gene read count matrix generated in Chapter 6.3.6 was imported into R (http://www.R-project.org/), and DESeq 1.16.0 (Anders and Huber 2010) was used to generate a countDataSet object. Samples were grouped according to time and graft state (for example the group 'T12-self' contained samples T12AA and T12BB, while 'T12-nonself' contained sample T12AB). Each sample was also annotated with

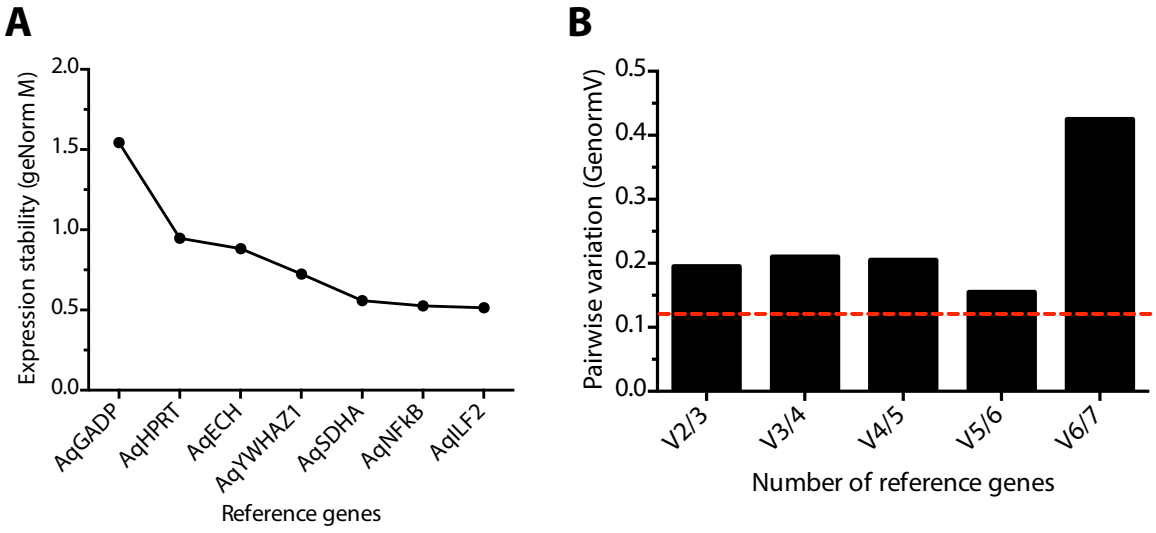


Figure 6.4 geNorm analysis of candidate qPCR reference genes
(A) Expression stabilities (geNorm M value) of the seven candidate housekeeping genes across the graft time course. (B) geNorm calculations of the optimal number of candidate housekeeping genes to use as standards for qPCR analysis. Stable genes would optimally exhibit a pairwise variation value (geNorm V value) below 0.15 (red line). The five most stable genes (*AqILF2, AqNFkB, AqSDHA, AqYWHAZ1* and *AqECH*) were recommended for use as reference genes, as this combination was closest to the optimal level. *AqECH* was, however, omitted from further analyses due to contamination of the no-template control sample.

its sponge of origin, namely sponge A, B or AB. The generalised linear modelling (GLM) stage of the analysis was performed as per the genefilter vignette, with graft state taken as 'condition', while sponge of origin was taken as 'type'.

Table 6.7 Trinity *de novo* assembly statistics

SAMPLE	TOTAL TRANSCRIPTS	TOTAL COMPONENTS	N50
Donor A	56937	26969	1672
Donor B	54496	25307	1892
T12 AA	58703	26949	1723
T12 BB	70044	29181	2021
T12 AB	60192	28719	1884
T24 AA	54366	26569	1761
T24 BB	55661	25930	1635
T24 AB	88060	34575	2404
T48 AA	52108	25740	1701
T48 BB	55634	25291	1637
T48 AB	58599	27140	1767
T72 AA	54389	26850	1723
T72 BB	61724	29823	1692
T72 AB	57323	26417	1894

The results of this analysis demonstrated that the bottom 50% of genes, as ranked by total genewise counts across samples, could be removed from the analysis without eliminating any genes likely to be designated as differentially expressed (for an unadjusted p-value of 0.003, as per the genefilter vignette) (Figure 6.3). This corresponded to removal of any genes with a total genewise count ≤53 (Table 6.3). Use of this 50% filtered dataset for DGE analysis resulted in the identification of a greater or equal number of differentially expressed genes than were identified in identical test analyses in which less-filtered datasets (e.g. removal of genes in bottom 40% of expression, with counts < 1 cpm [counts per million], total rowsum < 10 etc.) were used (data not shown).

6.3.9 Differential gene expression analysis

DGE analysis was performed using EdgeR version 3.6.8 (Robinson and Smyth 2007a; 2007b; Robinson et al. 2010; McCarthy et al. 2012). A script describing the full analytical methods to perform this analysis is available in Appendix 6.1. Briefly, to help compensate for the lack of replication available for this experiment, a reduced experimental model was generated in which the within-time course samples were grouped together, and the common dispersion across all genes was calculated using this model (common dispersion = 0.1606744). The analysis was then re-run with the full explanatory

model, where samples were grouped by treatment (AA, BB or AB) and time (0 hpg to 72 hpg). The common dispersion value determined above was also used for this analysis. Genes exhibiting statistically significant ($p \leq 0.01$) changes of four-fold or greater (log2) expression were identified using EdgeR's GLM functionality (Appendix 6.2).

6.3.10 qPCR

Fresh RNA was extracted as described in Chapter 6.3.3, taking tissue from the same graft time course used for transcriptome sequencing. One milligram of RNA was treated to remove genomic DNA (gDNA) contamination with DNase I (Invitrogen), according to manufacturer's directions. This RNA was reverse transcribed using the SuperScript III (ssIII) reverse transcriptase (RT) system (Invitrogen) according to manufacturer's directions, but using 1.5 µL 50 uM oligoDT (Promega), 1.5 µL 10 mM dNTPs, 3uL 5x first strand buffer, 0.75 µL 0.1 M DTT, 0.375 µL RNAsin (Promega), 0.375 µL ssIII and 7.5 µL RNA. Reverse transcriptase-free (no-RT) controls, in which the ssIII was replaced with an equal volume of DNase and RNase-free water (Gibco, Invitrogen) were also prepared for each sample in order to check for gDNA contamination. Sample PCRs (polymerase chain reaction) were run to confirm the absence of gDNA contamination in the no-RT controls (data not shown).

Primer pairs for use in qPCR were designed to amplify short (100 - 160 bp) fragments of *AqAFA* to *AqAFE* (Table 6.4). *AqAFF* was not tested due to its small size and lack of similarity to the other *AqAFs*. Primers were designed using Primer3 2.0.0 (Koressaar and Remm 2007) and Vector NTI Advance 10 (Invitrogen), and were supplied by Sigma-Aldrich. In-house primers for candidate reference genes *Enoyl CoA hydratase (AqECH)*, *Glyceraldehyde-3-phosphate dehydrogenase (AqGAPD), Hypoxanthine phosphoribosyltransferase (AqHPRT), Interleukin enhancer binding factor 2 (AqILF2), Nuclear factor kappaB (AqNF-κB), Succinate dehydrogenase complex subunit A (AqSDHA),* and *Tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein, zeta polypeptide (AqYWHAZ1)* were also selected for use (Table 6.4).

qPCR was performed using a Roche Lightcycler 480 Real-Time PCR System (Roche Applied Science). Standard curves were generated for each primer pair, using cDNA diluted 1:10, 1:20, 1:50, 1:100 and 1:500. qPCR error and efficiency estimates, and standard curve slopes, are provided in Table

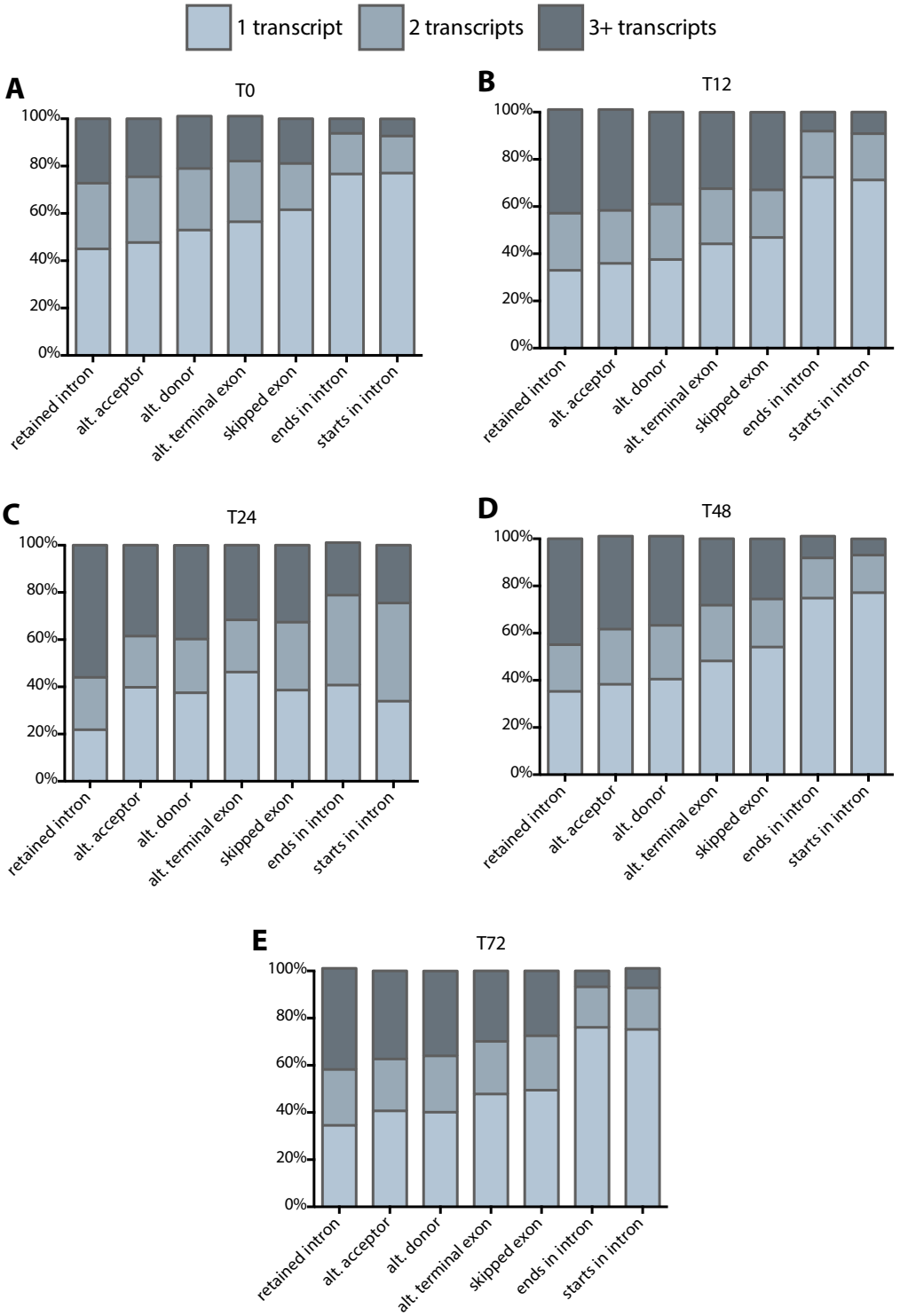


Figure 6.5 Filtering criteria for transcriptome-wide alternative splicing events

(A-E) Transcriptional support for each type of alternative splicing event. Each bar represents the percentage of splice events in the unfiltered dataset supported by one, two, or three or more transcripts within each time point (self and nonself combined). Splicing events supported by fewer than three transcripts were filtered prior to further analysis to reduce noise.

Table 6.8 Self and nonself graft response scoring

	12 HPG			24 HPG			48 HPG			T72		
	×	~	✓	×	~	✓	×	~	✓	×	~	✓
Self	4	2	2	6	1	1	8			8		
Nonself	4			1	3		1		3	4		

✓ = fusion, × = rejection, ~ = ambiguous/partial fusion

6.5. qPCR was performed using 3 µL of 1:50 diluted cDNA in a 15 µL reaction mixture of 7.5 µL SYBR green mastermix (Roche Applied Sciences), 0.75 µL bovine serum albumin (BSA) and 0.5 µL each of 5 uM (*AqAFA - AqAFE, AqECH, AqGAPD, AqHPRT, AqILF2, AqNFkB*) or 10 uM (*AqSDHA, AqYWHAZ1*) forward and reverse primers. For each gene, cDNA samples were run in triplicate, as were a no-template control (in which DNase and RNAse-free water was used in place of a cDNA template) and a calibrator cDNA sample (derived from 35 assorted grafted and ungrafted sponges) that was used in all qPCR runs to account for inter-run variation. The qPCR thermoprofile used for all runs is provided in Table 6.6; an annealing temperature of 58°C was used for all primers.

Reference gene stability was assessed using the geNorm (Vandesompele et al. 2002) algorithm within qbase+ 2.6.1 (Biogazelle), which determined that the combination of *AqECH, AqILF2, AqNFkB, AqSDHA and AqYWHAZ1* was optimal for downstream expression normalisation (Figure 6.4). However, *AqECH* was omitted as contamination was detected in the no-template control samples. Calibrated normalised relative quantities (CNRQ) of samples for all genes were calculated using qbase+ 2.6.1. No statistically significant differences between samples were detected by one-way analyses of variance (ANOVAs) performed within qbase+.

6.3.11 Detection of putative alternatively spliced transcripts

Trinity (release 2012-06-08) (Grabherr et al. 2011) was used for *de novo* transcriptome assembly, in conjunction with Bowtie 0.12.7 (Langmead et al. 2009), Java 1.6.0_22 and Samtools 0.1.18 (Li et al. 2009). Jellyfish k-mer counting was assigned 20 gigabytes (GB) memory, and a glue factor of 0.1 was used for the Trinity analysis; all other parameters were run as default. Assembly quality was assessed using the TrinityStats tool included in the 2013-02-25 Trinity release (Grabherr et al. 2011); assembly statistics are provided in Table 6.7.

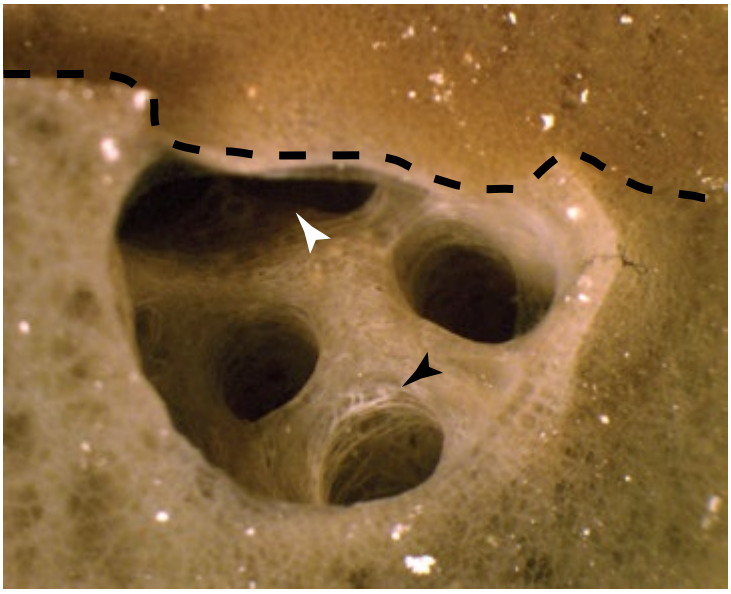


Figure 6.6 Tissue remodeling of an osculum following self grafting
An internal osculum bisected during graft preparation and placed at the autograft interface triggered the adjacent self tissue to remodel to form a continuous chamber inside the new tissue by 72 hours post grafting (hpg). Black arrow – white tissue can be seen at the cut surface of the osculum that was not in contact with self tissue, and signs of tissue healing are apparent by 72 hpg. White arrow – indicates where the chamber continued into the other half of the sponge autograft. Visual inspection of the chamber revealed that it continued deep inside the tissue.

Transcripts for all samples were compared to the Aqu2.0 gene models using PASA (program to assemble spliced alignments; release 2012-06-25) (Haas 2003) to identify and classify putative alternatively spliced transcripts genome-wide and, more specifically, amongst the *AqAF* genes. PASA annotation was performed by S. Fernandez Valverde, using the transcripts generated above. All transcript datasets for each time point were analysed together. A standard PASA pipeline was followed, using a minimum percentage of isoform coverage value of 40, and a stringent alignment overlap setting of 30.

For transcriptome-wide alternative splicing statistics, the data output was filtered to reduce the impact of spuriously-supported splice changes, by removing all transcripts in each time point which were supported by fewer than three transcripts (Figure 6.5). Support could come from the same (i.e. if multiple transcripts were present in the one individual) and/or different (i.e. AA, BB and/or AB) individuals. Only the alternative acceptor, alternative donor, alternative exon, skipped exon, retained intron, starting in intron or ending in intron categories of splicing events were considered for downstream analyses.

The nucleotide sequences of all unfiltered putative alternatively spliced transcripts mapping to the *AqAF* genes were extracted and were manually compared to the Aqu2.1 gDNA and messenger RNA (mRNA) sequences using CodonCode Aligner version 3.7.1.1. Transcripts confirmed to alter AqAF structure were selected for further analysis. Sequence truncations were not considered unless these transcripts also contained a splicing event of interest.

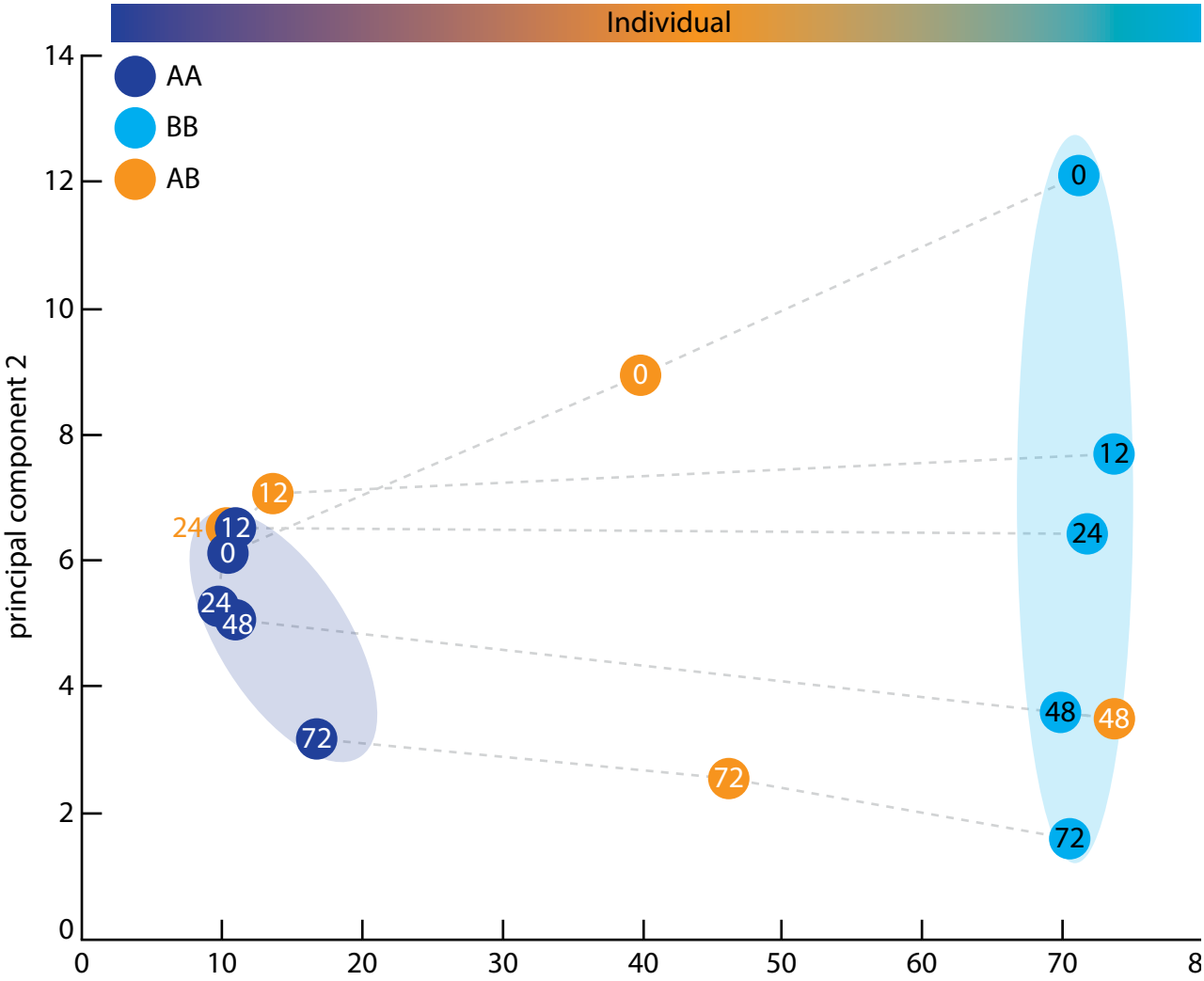


Figure 6.7 Principal component analysis of dynamically expressed genes
Each circle represents a transcriptome within the graft time course. Dots are coloured by donor sponge (A/AA – dark blue, B/BB – light blue, AB – orange) and numbered by time point (0 – donor sample, 12 – 12 hours post grafting (hpg), 24 – 24 hpg, 48 – 48 hpg, 72 – 72 hpg). Shaded rings group the A/AA and B/BB samples, respectively. Dashed lines link samples representing the same time point from different time courses. Shaded bars at the top and right sides of the graph summarise the results of the analysis, showing the biological variables that best explain the sample separations observed across each axis. Sample separation is based on the top 0.9th quantile of dynamically expressed genes, as determined by BLIND.

6.3.12 Venn diagrams

All Venn diagrams were generated using the online tool Venny (http://bioinfogp.cnb.csic.es/tools/venny/index.html).

6.3.13 Heatmaps

Unscaled heat maps showing log2 fold changes between genes of interest were generated using the R function heatmap.2 within the gplots package (http://www.cran.r-project.org/web/packages/gplots/index.html) using default clustering parameters.

6.3.14 Gene ontology

Gene ontology annotation of the *A. queenslandica* Aqu2.1 gene models was performed by S. Fernandez Valverde using Blast2GO version 2.8 (Conesa and Götz 2008). Annotations were manually reformatted for downstream analysis by W. Hatleberg. The Cytoscape software (Shannon et al. 2003) plugin, BiNGO (Maere et al. 2005), was run with default parameters to identify Biological Process and Molecular Function gene ontology (GO) terms that were statistically significantly over-enriched in the gene lists of interest, relative to the rest of the *A. queenslandica* genome. Enriched GO terms were clustered based on semantic similarity (SimRel measure) using the software REVIGO (Supek et al. 2011). Similar GO terms with a redundancy of >0.7 were collapsed. Gene counts per enriched GO term were used to determine treemap layouts.

6.4 Results

Sponge grafting experiments were first performed in 1869 (Vaillant) and have been well-described in the literature since this time. However, advances in DNA and RNA sequencing technologies mean that the sponge graft response can now be studied on a transcriptome-wide scale. I therefore performed a classical self and nonself grafting experiment between *A. queenslandica* individuals, and analysed the quantitative and qualitative changes in expression that occurred across the graft time course.

6.4.1 Physiological responses to sponge tissue grafting

Four graft experiments were performed, with each experiment using tissue from two sponge individuals to generate one nonself and two self time courses. Grafts were observed at 12, 24, 48 and 72 hpg to determine the physiological response to self or nonself contact. Tissue samples were collected at each time point.

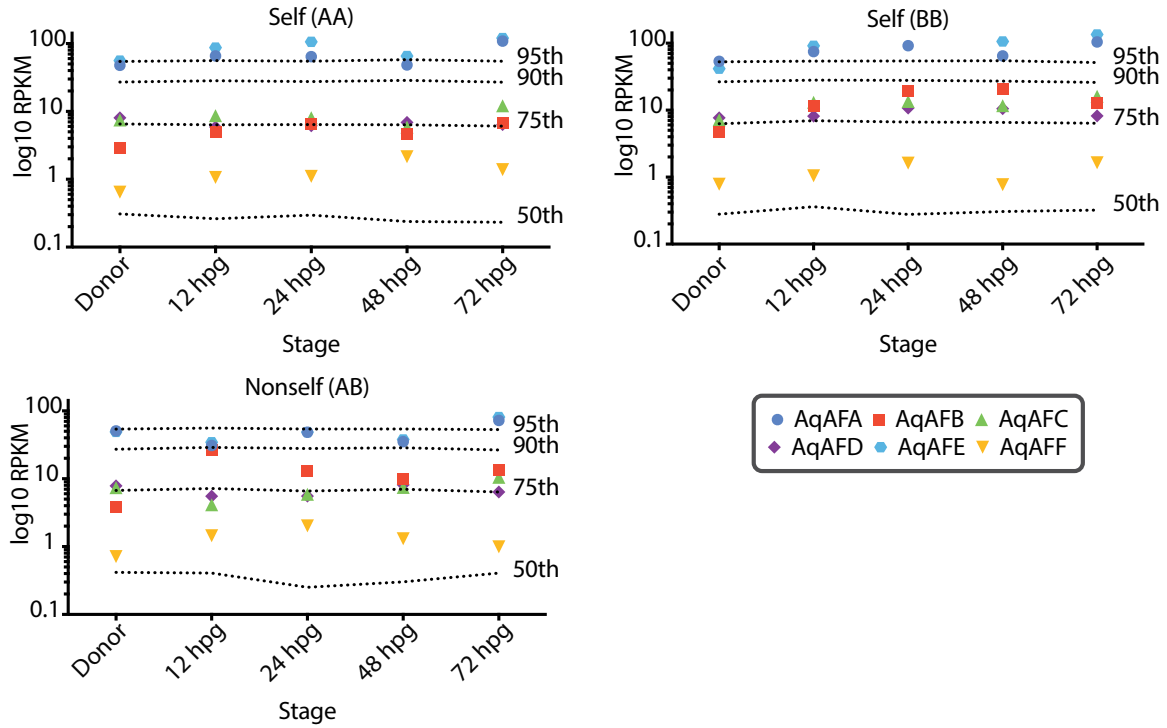


Figure 6.8 *A. queenslandica AF* expression levels in graft transcriptomes, relative to transcriptome-wide percentiles

Coloured data points represent the log10 normalised counts (measured in RPKM - reads per kilobase per million mapped reads) of *AqAF* gene expression in each graft time point, across the three time courses. Dashed lines show the transcriptome-wide percentiles (50th – 95th) of transcript abundance in each graft stage (complete, unfiltered datasets). Lines showing the 5th, 10th and 25th percentiles are not visible as these represent transcript counts of 0 across all stages.

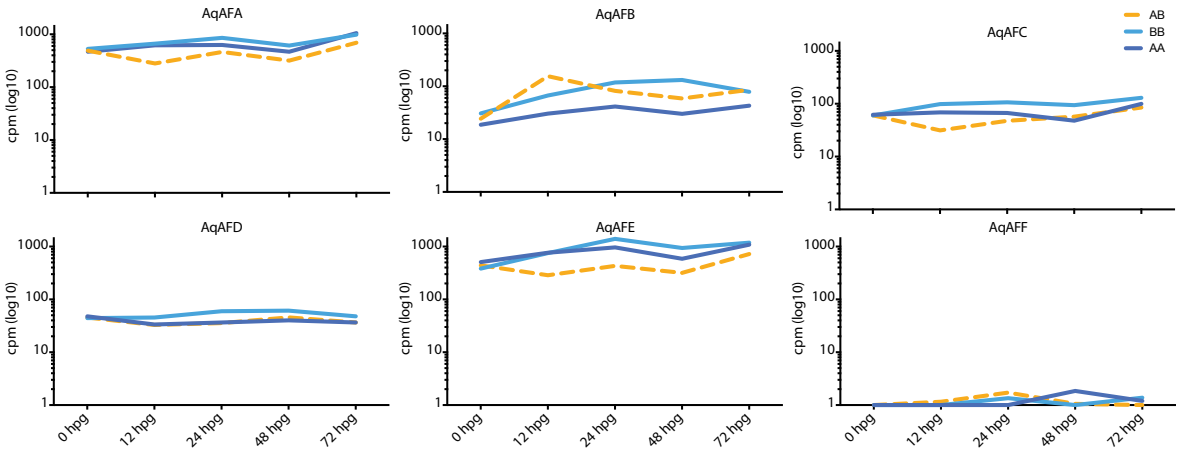


Figure 6.9 *A. queenslandica* AF gene expression response to tissue grafting in transcriptome data

For each *A. queenslandica AF* gene, each datapoint represents the expression level (measured in read counts per million sequencing reads, cpm) of the gene at a particular time point (0 to 72 hours post grafting, hpg) within a self (AA, dark blue; BB, light blue) or nonself (AB, orange) graft time course.

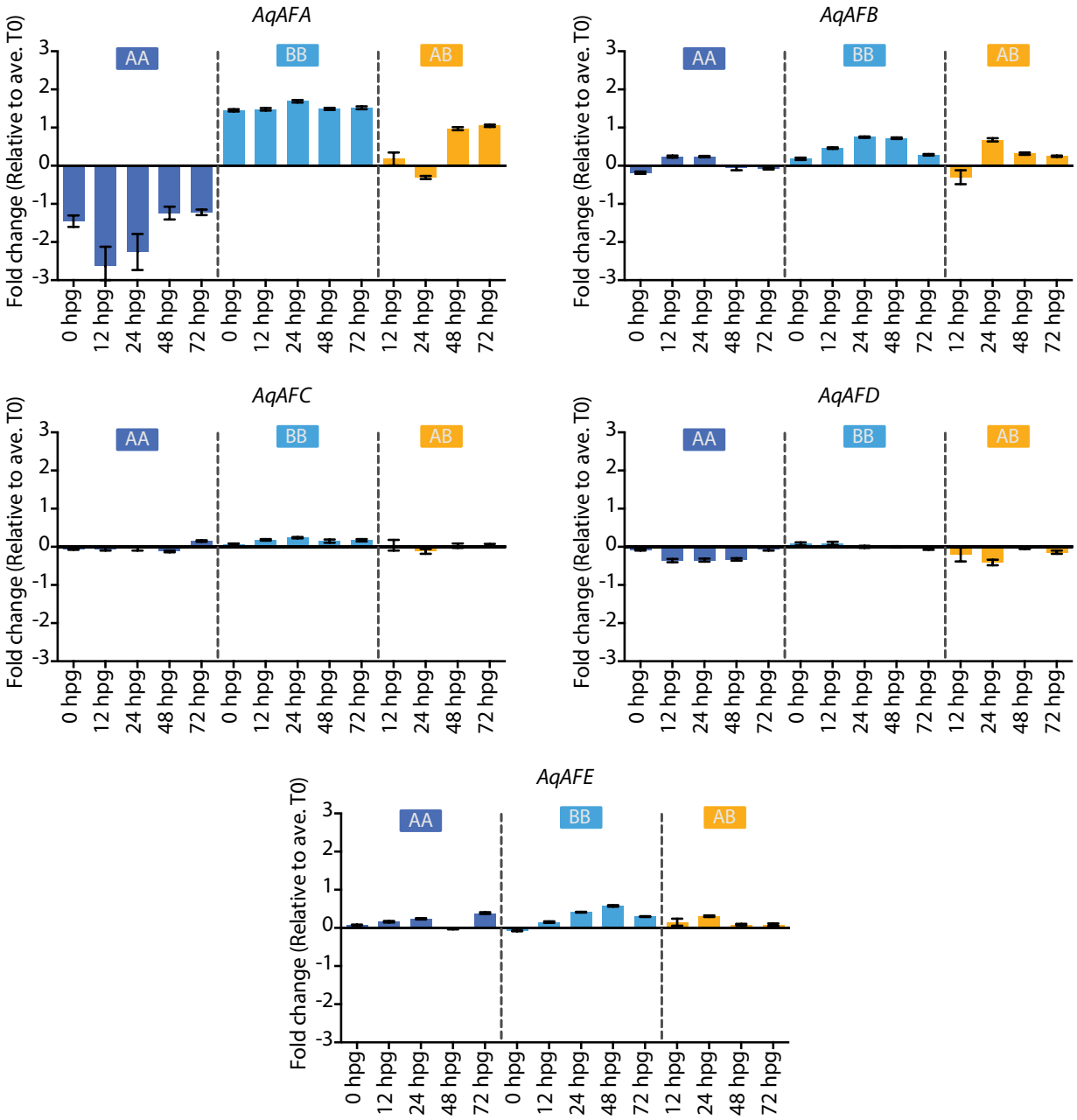


Figure 6.10 *A. queenslandica AF* gene expression response to tissue grafting in qPCR data

For each *A. queenslandica AF* gene, each datapoint represents difference in expression levels (fold change; the CNRQ value produced by qbase+) between the gene at a particular time point within a self or nonself (AA, dark blue; BB, light blue; AB, orange) graft (0 to 72 hours post grafting, hpg), and the mean expression of that gene in the two ungrafted donor sponges (i.e. 0 hpg AA and BB).

a. Autografts

For six of the eight autograft time courses, early signs of tissue fusion were first observed at 12 hpg. The seventh sponge initiated fusion by 24 hpg, and the eighth by 48 hpg (Table 6.8). Bonds between tissue samples grew progressively stronger as the experiment progressed, with all self samples unambiguously fused by 48 hpg. In general, by 72 hpg the two tissue pieces could not be separated with reasonable force, and the line dividing the tissues was difficult to see. Signs of tissue remodelling were also observed by 72 hpg. For example, in one sample, a bisected osculum originally sat on one side of the point of fusion, and by 72 hpg the internal tissue from both sides of the graft appeared to have remodelled to develop a new chamber (Figure 6.6).

b. Allografts

Twelve hours after grafting, all four allograft samples remained unfused. However, at 24 and 48 hpg, several of the samples exhibited signs of partial fusion (Table 6.8). Here, weak fibrous connections were present between apposed tissue slices, although these bonds were easily broken with a light amount of force. By 72 hpg, no fusion between grafted tissue slices was ever observed. Both tissue partners within the grafts appeared healthy, although the cut surfaces at times appeared fibrous and whitened.

6.4.2 Transcriptome sequencing and statistics

One of the four graft experiments, comprised of one nonself and two self time courses sampled at 0, 12, 24, 48 and 72 hpg, was selected for whole-transcriptome sequencing and subsequent analysis. A tissue sample from the interface of each graft was taken at each time point, and RNA was extracted and prepared for Illumina high-throughput sequencing. Final sequencing datasets each contained between 17.5 (T24AB_C) and 27.8 (Donor A) million reads (Table 6.2). The average GC count per library was 42.3%, which was slightly higher than the genomic average across all *A. queenslandica* genes (38.1% as calculated using the *A. queenslandica* genome data available through BioMart) (Kinsella et al. 2011); Srivastava:2010ie. Sequencing reads were trimmed for quality, resulting in the loss of approximately 6% of reads per sample, and shortening of the remaining reads (Table 6.2).

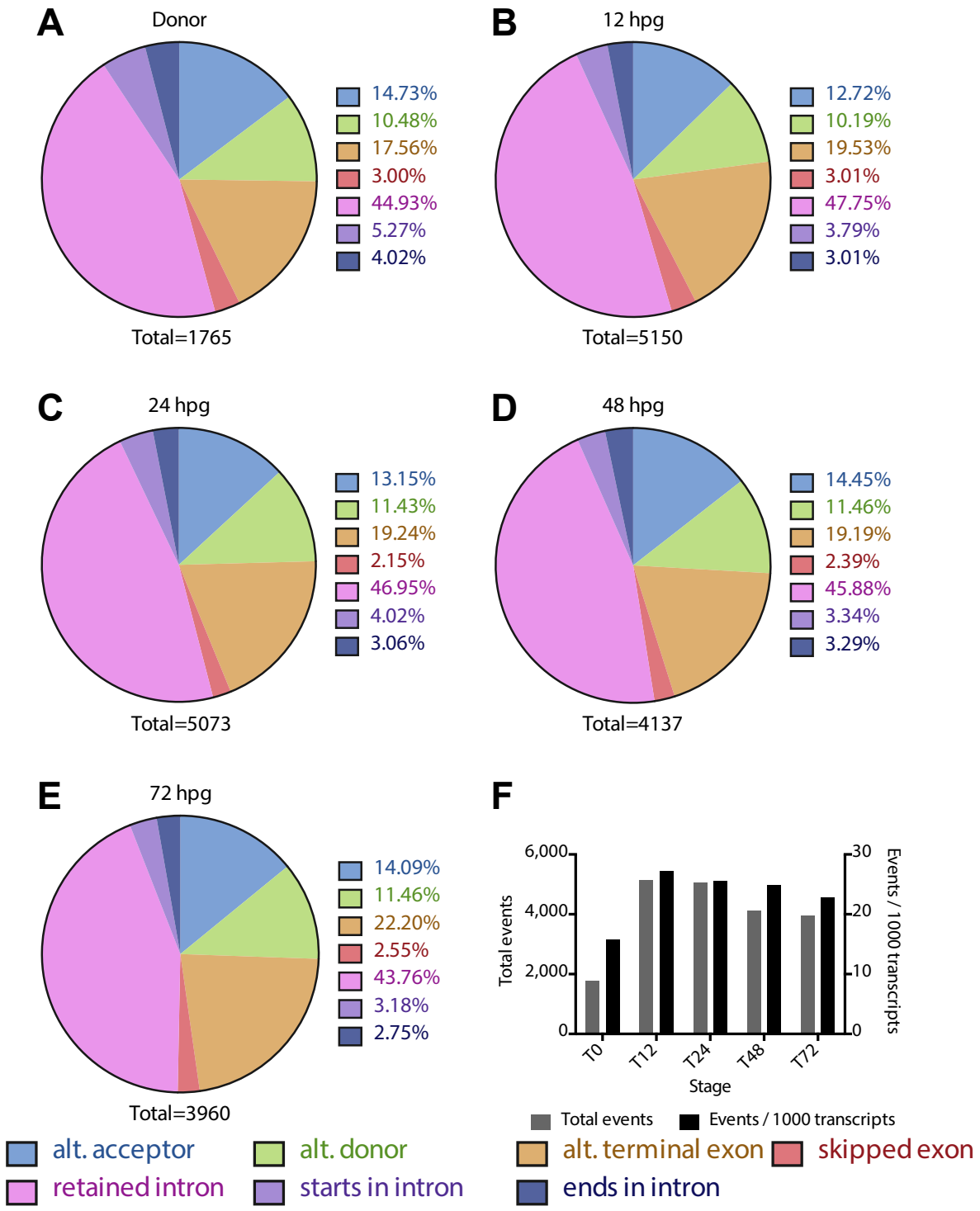


Figure 6.11 Alternative splicing event distribution and frequency across the graft time course

(A – E) Each pie chart represents the transcriptome-wide proportion of each of the possible alternative splicing events of interest. The number below each pie chart represets the total number of alternative splicing events. (F) Numbers of alternative splicing events per stage in total (grey bars, left axis) and scaled per 1000 transcripts analysed (black bars, right axis).

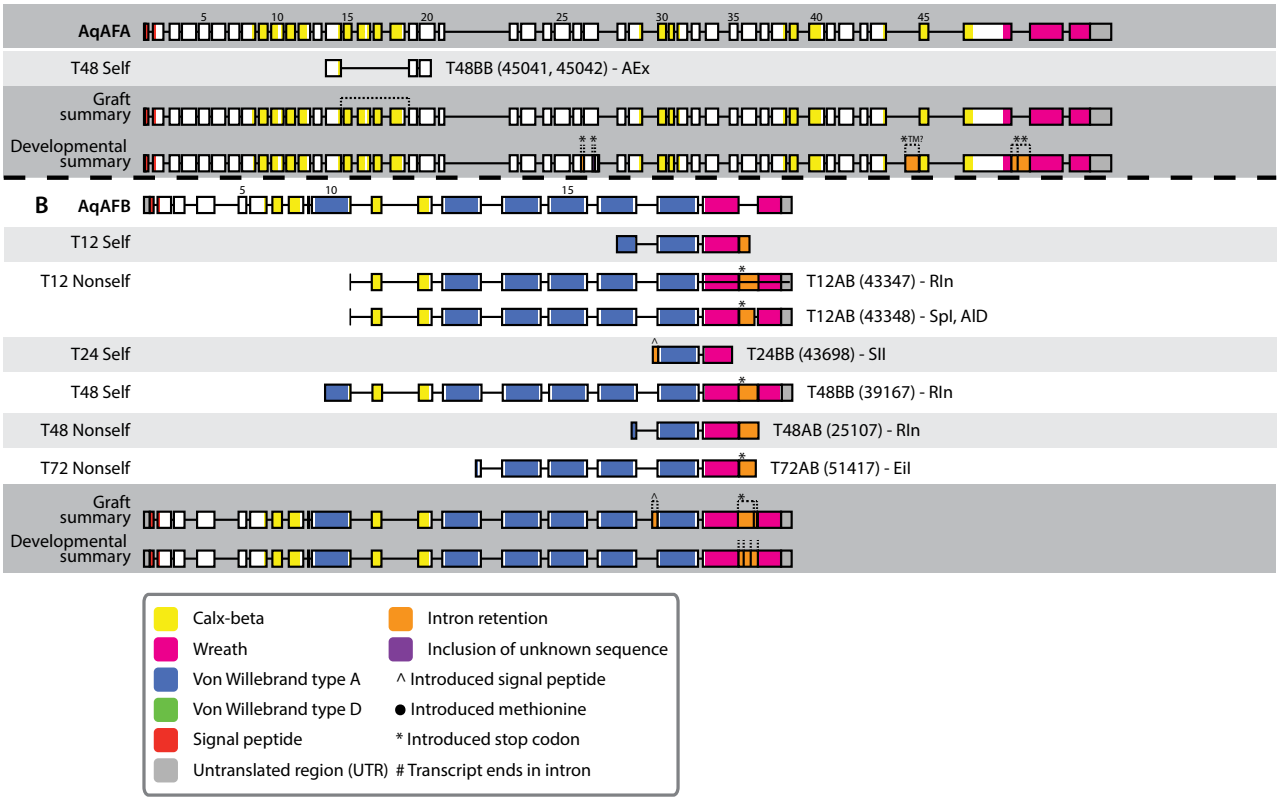


Figure 6.12 Putatively alternatively spliced *A. queenslandica* AF transcripts

(Part 1 of 2)
For *AqAFA* to *AqAFF*, the Aqu2.1 gene model prediction (top line) and putative alternatively spliced spliced transcripts from each graft time point are shown. Boxes represent exons (every fifth exon is numbered) and the connecting lines represent introns; regions encoding protein domains are coloured accordingly. Orange boxes represent intron inclusion events, while purple boxes represent inclusions of unknown sequence. Regions where domain type predictions overlap are depicted by overlapping colours. Exons and introns are drawn to scale. Symbols above each model represent predicted effects on the encoded proteins (see key). Two summaries are given for each gene (bottom lines), in which all observed changes from this experiment ('Summary') and the developmental experiment discussed in Chapter 4 ('Developmental summary'). No data is provided for *AqAFF* as no alternatively spliced transcripts were identified for this gene in the present study.

6.4.3 Principal component analysis

Genetic identity, rather than immune state, appears to be the primary factor promoting gene expression differences between samples, when considering the most dynamically-expressed genes across all samples. In the PCA results (Figure 6.7), the AA and BB autogeneic graft samples formed two separate clusters along the first principal component. The autogeneic samples then showed a chronological separation of samples by hours post grafting along the second principal component. Although both the AA and BB time courses displayed this trend, between-sample variation was greater in the BB time course, with samples spread out across the second principal component, while the AA

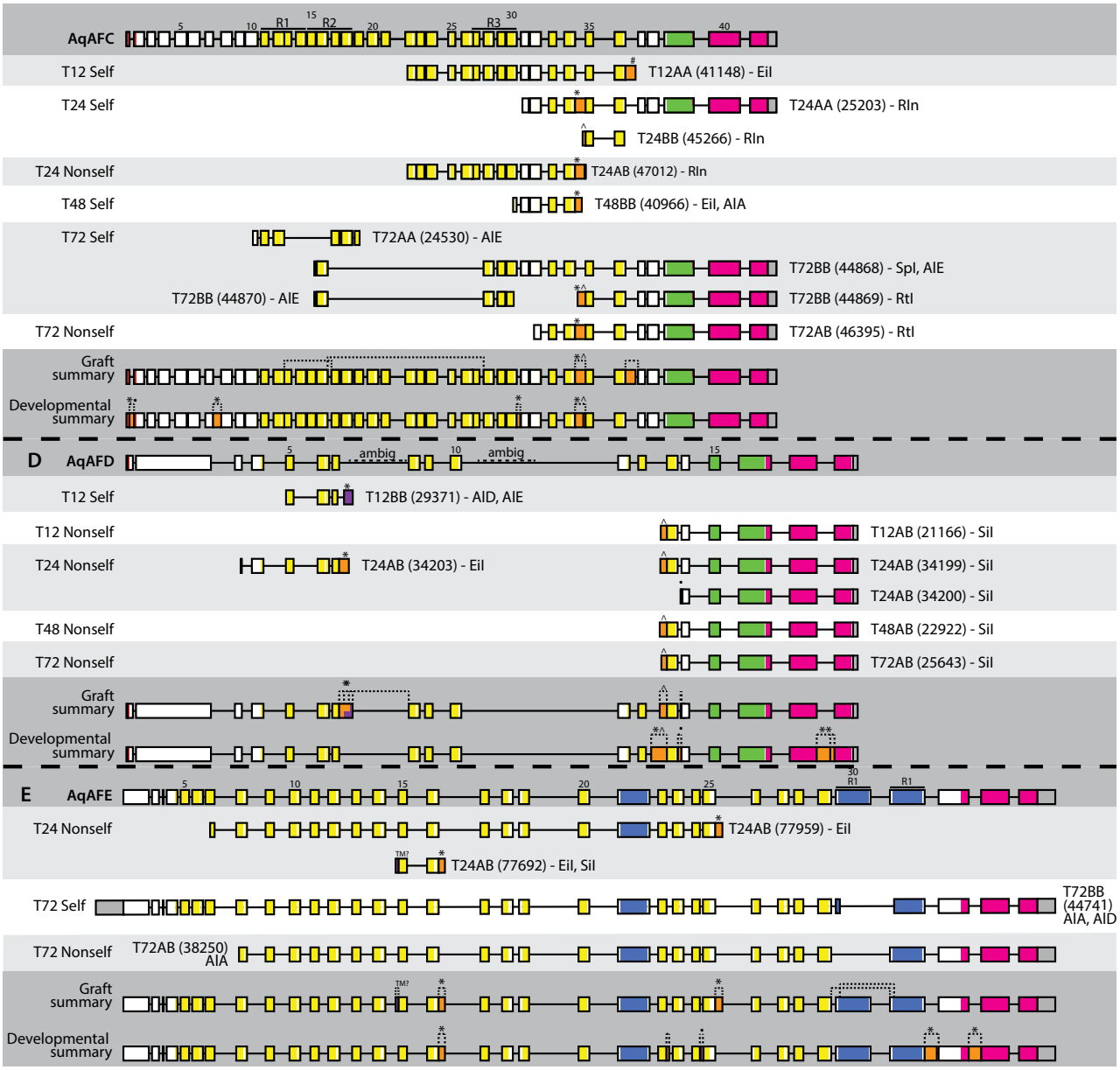


Figure 6.12 Putatively alternatively spliced *A. queenslandica* AF transcripts

(Part 2 of 2)

Table 6.9 Putatively alternatively spliced *A. queenslandica AF* transcripts

(Part 1 of 2)

AqAFA			
Position	Sample/s	Type of change	Predicted translational effect
Exons 15 – 18	T48BB	Exon skipping	Loss of two Calx-beta domains
AqAFB			
Position	Sample/s	Type of change	Predicted translational effect
Intron 16	T24BB	Starts in intron	Introduces methionine. Signal peptide (SP) support weak but present
Intron 18	T12AA, T12AB (2), T48BB, T48AB, T72AB	Intron retention / ends in intron	Encodes 9 amino acids (aa) before introducing stop codon. Premature truncation of Wreath domain.
AqAFC			
Position	Sample/s	Type of change	Predicted translational effect
Exons 13 - 16	T72AA	Exon skipping	Loss of two Calx-beta domains (Overlaps with repetitive exons (Chapter 2.4.6) – unclear if sequence variants or misassembly)
Exons 17 - 26	T72BB (2)	Exon skipping	Loss of five Calx-beta domains (Overlaps with repetitive exons (Chapter 2.4.6) – unclear if sequence variants or misassembly)
Intron 34	T24AA, T24AB, T48BB, T72AB	Intron retention / ends in intron	Encodes 15 aa before introducing stop codon. Reading frame re-established, including methionine.
Intron 34	T24BB, T72BB	Starts in intron	T72BB encodes 15 aa before introducing stop codon. Predicted SP in both sequences (and canonical gDNA intron)
Intron 36	T12AA	Ends in intron	Reading frame maintained until end of assembled transcript

Table 6.9 Putatively alternatively spliced *A. queenslandica AF* transcripts

(Part 2 of 2)

AqAFD			
Position	Sample/s	Type of change	Predicted translational effect
After exon 7	T12BB	Unknown sequence incorporated after exon 7	Sequence encodes 5 aa before introducing a stop codon (Source unknown, but portion of canonical intron 7 ambiguous – extra sequence may belong to this region)
Intron 7	T24AB	Stops in intron	Encodes 49 aa before introducing a stop codon
Intron 12	T12AB, T24AB, T48AB, T72AB	Starts in intron	Early introduction of methionine; signal peptide predicted (Present in all nonself samples only)
Intron 13	T24AB	Starts in intron	No disruption to translational reading frame; no methionine/signal peptide
AqAFE			
Position	Sample/s	Type of change	Predicted translational effect
Before exon 15	T24AB (same transcript as below)	Inclusion of unknown sequence	Putative transmembrane domain
Intron 16	T24AB (same transcript as above)	Ends in intron	Translational reading frame maintained until two stop codons at end of transcript
Intron 25	T24AB	Ends in intron	Encodes two amino acids before introducing stop codon
Exon 30	T72AB	Exon skipping	Removal of VWA domain (Overlaps with repetitive exons (Chapter 2.4.6) – unclear if sequence variants or misassembly)
Exon 30 – 31	T72BB	Exon skipping	Splice two VWA domains together; one fewer VWA domain in total (Overlaps with repetitive exons (Chapter 2.4.6) – unclear if sequence variants or misassembly.)

samples formed a much tighter cluster (Figure 6.7). The AB allogeneic samples did not cluster along either principal component; instead, individual samples tended to group with similarly-staged samples from either AA or BB time courses (Figure 6.7). The Donor AB sample fell between the AA and BB samples along the first principal component, which was expected because Donor AB is an artificial sample formed by merging the sequencing reads from Donors A and B. T12AB and T24AB sat within the tight AA cluster, while T48AB fell close to T48BB. T72AB was aligned with the Donor AB sample along the first principal component, and with T72AA and T72BB along the second. Therefore, at each time point, samples from the three time courses tended to fall within the same general region along the first principal component, with time points arranged along the axis of the second principal component in general chronological order (Figure 6.7).

6.4.4 AqAF expression in tissue grafts

The *AqAFs* were consistently highly expressed at all points within the auto- and allograft time courses, relative to the transcriptome as a whole (i.e. before independent filtering by expression

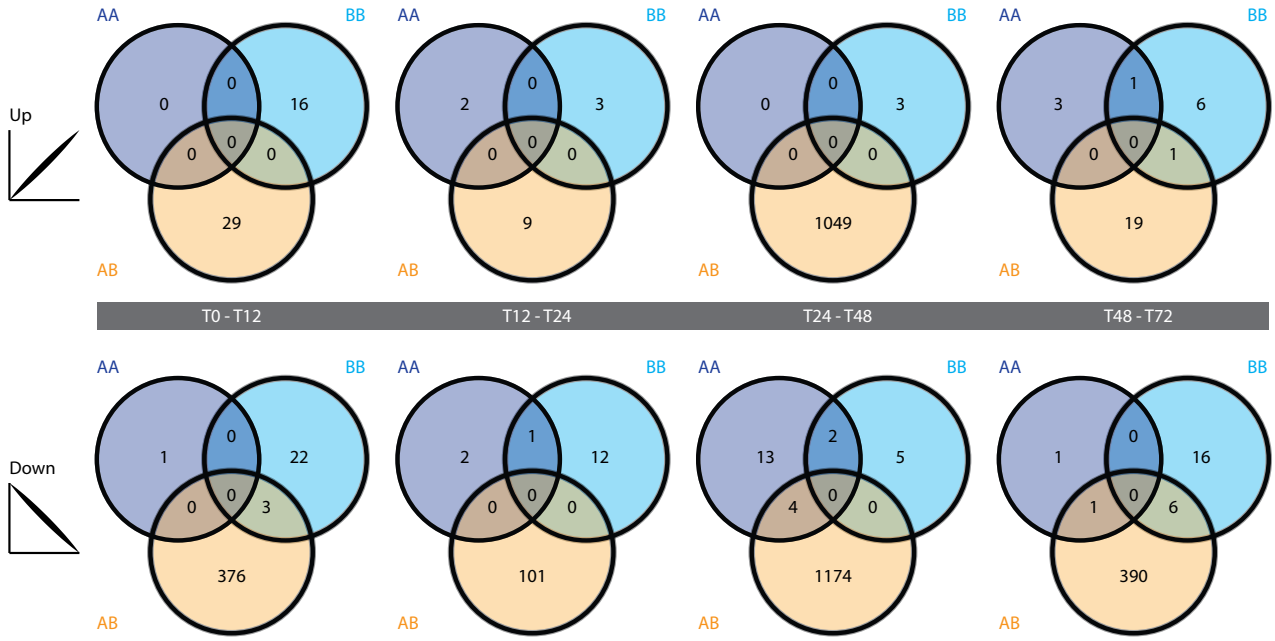


Figure 6.13 Differentially expressed gene counts
Each Venn diagram shows the number of differentially expressed genes that are up- (top) or downregulated (bottom) with an observed fold change of 4-fold or greater between pairs of successive time points, in the AA (dark blue), BB (light blue) and/or AB (orange) time courses.

level; Figure 6.8). Fold changes between successive stages were less than 2 in all instances, except for *AqAFB* between 0 and 12 hpg in the AB time course. None of the six *AqAF* genes were found to be significantly differentially expressed between any adjacent time points in the graft transcriptomes (Figure 6.9), however this should be re-tested in future with greater replication.

qPCR on RNA derived from the same graft time course as the transcriptome dataset did not reveal any statistically significant differences in *AqAF* expression between individual or grouped time points. However, the divergence between the two donor sponges, and the lack of biological replication, means that the occurrence of biologically meaningful changes cannot be ruled out. When examining the expression of *AqAFA* despite the absence of statistical support, the AA and BB autograft time courses overall showed large differences in expression to one another, although very little change occurred across each of the time courses (Figure 6.10). The T12AB and T24AB samples were intermediate between the two self extremes, while the T48AB and T72AB samples were more similar to the expression levels in the BB time course for these stages. *AqAFB* to *AqAFE* did not show large fold changes in expression relative to the average ungrafted control state (Figure 6.10).

6.4.5 Alternative splicing in the graft time course

a. Transcriptome-wide changes

De novo assembled transcripts were generated for each graft transcriptome. Assembly N50 values ranged from 1635 bp (donor A) to 2404 bp (T24 AB) (Table 6.7); these values are higher, on average, than those reported from other recently-published sponge transcriptomes (Riesgo et al. 2012; 2014). The PASA assembly pipeline was used to compare the newly assembled graft transcripts to the *A. queenslandica* Aqu2.0 gene models, in order to identify potential instances of alternative splicing. PASA designates differences between transcripts and the gene models as belonging to one of seven categories of interest: alternative use of intron donor or acceptor sites, intron retention, the start or end of a transcript exon within a canonical intron, alternative terminal exons or exon skipping (Figure 4.1). As seen in the developmental transcriptome datasets discussed in Chapter 4, intron retention was the most commonly observed alternative splicing category in the control tissue, comprising 45% of total alternative splicing observations (Figure 6.11). Exon skipping (3%), transcript termination inside an intron (5%) and transcript initiation inside an intron (4%) were the least commonly-observed categories

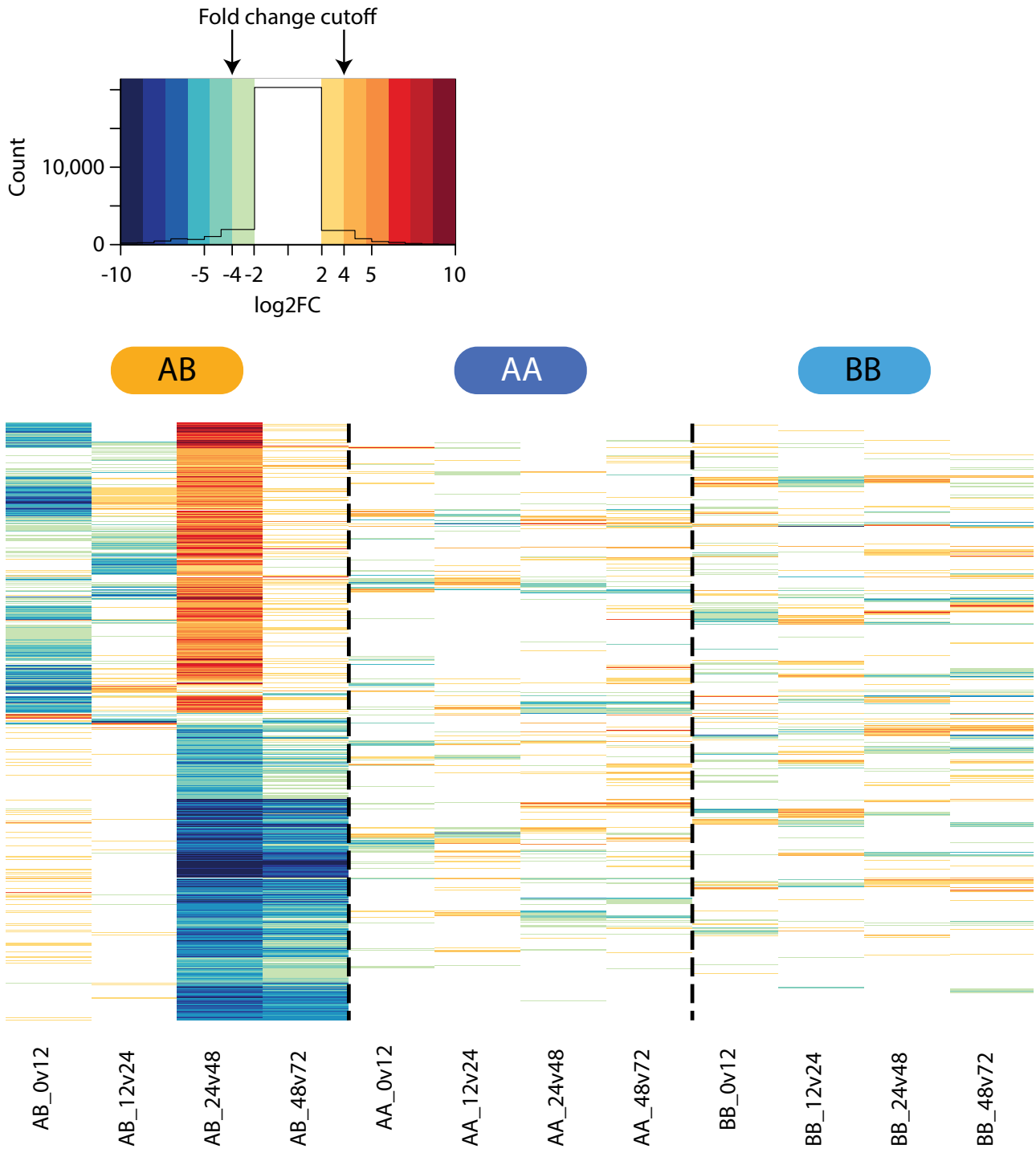


Figure 6.14 Differential gene expression in the nonself graft timecourse
Heatmap showing the log2 fold changes in expression across the self and nonself graft timecourses. Only genes found to be statistically differentially expressed, and exhibiting a 4+-fold expression change, in one or more pairs of nonself timepoints are shown.

(Figure 6.11). Alternative terminal exons, intron acceptors and intron donors were identified in 18%, 15% and 10% of cases, respectively (Figure 6.11). The graft response does not appear to promote wide-scale changes in alternative splicing, as the relative proportions of each splicing category remained stable across time (Figure 6.11a-e). Similarly, although the numbers of total splicing events varied between samples, the number of changes in the grafted samples is approximately proportional to the total number of transcripts analysed at each time point (Figure 6.11f). A slight increase in events per 1000 transcripts was observed in the grafted samples relative to the donors. However, this change may be explained by the lower number of samples contributing to the donor time point, which would in turn reduce the number of identified events with three or more instances of transcript support.

b. AqAF-specific changes

The unfiltered list of alternative splicing events that localised to the *AqAF* genes was manually examined to identify and characterise the transcriptional changes occurring relative to the Aqu2.1 gene models. Alternatively spliced transcripts were found for *AqAFA* to *AqAFE* (Appendix 6.3). The majority of observed changes were intron retention events, or transcripts ending or beginning within an intron. No alternatively spliced *AqAF* transcripts were identified in either of the two donor samples. The domain and intron-exon architectures of the alternatively spliced transcripts are shown in Figure 6.12, and the putative protein-level changes that these splicing events would cause are discussed in Table 6.9.

6.4.6 Differential gene expression

RNA-Seq reads from all graft samples were mapped back to the *A. queenslandica* genome to determine the read counts per Aqu2.1 gene model. These counts were then used to identify genes exhibiting statistically significant fold changes between successive pairs of time points. The two self time courses, AA and BB, were analysed separately in light of the finding that between-individual differences were the primary source of variance between samples (Figure 6.7). For this reason, and the general lack of replication available, I chose a strict fold change selection threshold - four-fold or greater (log2) changes in expression between successive pairs of time points – to avoid spurious results.

All tested comparisons in the two self time courses exhibited low numbers of statistically significant differentially expressed genes at the filtering threshold used, with very little overlap between genes identified for the AA and BB time courses. Greater numbers of differentially expressed genes were identified in the four nonself comparisons (Figure 6.13).

The highest number of differentially expressed genes was identified in the 24 to 48 hpg category, where over 1500 genes were both up- and down-regulated at 48 hpg relative to 24 hpg. This is therefore the most prominent time period on a heatmap displaying the log2 fold change in expression of all genes that were differentially expressed in one or more nonself graft comparisons (Figure 6.14). When the differentially expressed genes are considered as two groups, based on whether they are up- or down-regulated between 24 and 48 hpg, it can be seen that within a group, genes tended to behave similarly to one another across the graft time course (Figure 6.14). For those upregulated between 24 and 48 hpg, genes tended to be downregulated between 0 and 12 hpg or 12 and 24 hpg, relative to ungrafted expression levels. A small subset of the genes downregulated between 0 and 12 hpg were upregulated slightly between 12 and 24 hpg. The genes in this broad group increased in expression between 24 and 48 hpg, before either increasing further or exhibiting an expression plateau between 48 and 72 hpg (Figure 6.14). For those genes that were downregulated between 24 and 48 hpg, expression either remained constant or increased slightly between 0 and 12 hpg. Most genes remained stable between 24 and 48 hpg, before decreasing in expression between 24 and 48 hpg, and again between 48 and 72 hpg (Figure 6.14). When examining all differentially expressed genes in the self time courses, most genes exhibited no or small changes in expression (Figure 6.14), as expected based on the DGE counts presented in Figure 6.13. No clear trends were observed when examining the expression of these genes in the two self time courses (Figure 6.14).

6.4.7 Gene ontology analysis

To explore the sponges' putative functional response to grafting, each list of differentially expressed genes (Appendix 6.4) within the nonself time course was analysed to identify GO terms which were statistically significantly enriched amongst the genes of interest, relative to the genome as a whole. Treemaps showing these results are presented in Appendix 6.5. In particular, these results reveal that chronological progression of the sponge graft response is associated with the downregulation

of genes involved in key biological processes such as cell signalling, transcription and translation and molecular transport.

6.5 Discussion

AFs are putative allorecognition molecules which are implicated in auto- and allograft responses in the demosponge *Clathria prolifera* (Fernàndez-Busquets and Burger 1997; Fernàndez-Busquets et al. 1998; Fernàndez-Busquets and Burger 1999). In the present chapter, I examined the *A. queenslandica* physiological graft response, before generating fourteen whole-transcriptome sequencing datasets spanning the duration of the physiological self fusion and nonself rejection processes in this species. I traced the quantitative expression profiles of the *AqAF* genes across the graft time course using both this transcriptome dataset and qPCR, and catalogued the set of alternatively spliced transcripts generated from the *AqAFs* in grafted tissue. Finally, I performed a preliminary analysis of the global changes in gene expression that occur across the graft time course. This represents the first longitudinal, high-throughput sequencing approach applied to understanding the molecular allorecognition response in sponges.

6.5.1 Physiological self and nonself graft responses in *A. queenslandica*

Of the eight examined self graft time courses, fusion was observed for all samples by 48 hpg. Observed variability in the onset time of initial fusion likely represents inter-individual variation, but possibly also inconsistencies in contact surfaces between grafts and/or failure to observe weak bonds between tissue pieces, which may have broken while unpinning the grafts. By 72 hpg, graft interfaces were difficult to discern, and the tissue pieces could not be separated without force. In all four nonself graft time courses, rejection had occurred by 72 hpg. Tissue in the rejected grafts remained alive and healthy, with no signs of necrosis obvious to the naked eye. Three of the four nonself graft time courses exhibited signs of transitory fusion between 12 and 48 hpg. Here, weak bonds appeared to join the two pieces of tissue, and light force was required to separate the two slices after removal of the pin holding them together. The bonds between tissue pieces may not represent true early fusion, but rather, for example, fibrous material produced during graft rejection that randomly interlaced due to proximity of the two tissue pieces. However, it may be that a degree of tissue fusion is required early in the rejection process, to allow cellular infiltration of the graft interface, direct cell-cell contact

between cells of the opposing individuals, and subsequent immune rejection. Such a phenomenon has been reported elsewhere, for instance with the discovery of tissue bridges spanning the nonself graft interface in other sponge species (see for instance Hildemann et al. 1980; Bigger et al. 1981; Buscema and Van de Vyver 1984; Fernàndez-Busquets et al. 2002; Fernàndez-Busquets and Burger 2003). Blocking the graft interface with an artificial membrane, permeable to diffusible factors but not cells, has also been shown to inhibit the rejection response (Bigger et al. 1981), further suggesting that direct cell-cell interactions are critical for sponge allorecognition. Transitory fusion has also been observed in the allorecognition response of the colonial hydroid *Hydractinia symbiolongicarpus*. Fusion-rejection reactions in this species are largely under the control of two tightly-linked, highly polymorphic genes, *alr1* and *alr2* (Rosa et al. 2010). Fusibility assays have determined that two contacting colonies require at least one shared allele at both *arl1* and *alr2* for recognition as self and subsequent successful fusion, while nonself identification and rejection occurs if the colonies do not share any alleles at either gene (Cadavid et al. 2004). However, if two colonies, most likely recombinants, share at least one allele at only one of the two genes, a process called transitory fusion occurs, whereby colonies fuse for a number of days before commencing a normal rejection response (Cadavid et al. 2004). A similar process may be occurring in *A. queenslandica*, though given that the majority of nonself grafts in this experiment exhibited signs of transitory fusion, it seems unlikely that this hypothetical response is limited to genetically-similar individuals in this species. Microscopic analysis of cellular activity at the nonself graft interface is required to understand the nature of this apparent transitory fusion.

6.5.2 Transcriptome and qPCR data do not reveal dynamic expression of *AqAF* genes in grafted tissue

Statistically significant differences in *AqAF* expression were not observed within the one nonself or two self graft time courses, using either whole-transcriptome or qPCR analysis. It should be noted, however, that negative statistical results for the qPCR analysis could be due, in part, to lack of biological replication of this experiment, given the variation detected between individual sponges (Figure 6.7). The lack of dynamic *AqAF* expression in response to grafting may suggest that the AqAFs are not involved in the self or nonself graft responses in *A. queenslandica*, that the *AqAFs* are indeed dynamically expressed but were not detected for technical or analytical reasons, or that the *AqAF* genes are ubiquitously expressed regardless of alloimmune state. The *AqAFs* are very highly expressed relative to the rest

of the genome in ungrafted (Figure 3.6, Figure 6.8) and grafted (Figure 6.8) tissue, which may lend support to this latter hypothesis.

The lack of *AqAF* expression response to grafting is surprising, as previous studies have reported that *MAFp3* and *MAFp4* expression increases in both auto- and allografted tissue (Fernàndez-Busquets et al. 1998). There are several possible explanations for this difference. First, the AFs may not be involved in *A. queenslandica* allorecognition. However, if the *CpAFs* are indeed dynamically expressed in *C. prolifera,* this explanation seems unlikely as this would require large evolutionary shifts in the molecules deployed in the allorecognition response to have occurred within a single class of sponge. Second, regulation of the *AqAF* response may occur downstream of transcription. As the *AqAF* genes are very highly expressed at all developmental stages (Figure 3.6) and grafted tissue samples (Figure 6.8), it may be that the *AqAFs* are ubiquitously transcribed, but that differential control of translation, AF complex assembly, glycosylation, or extracellular molecule deployment is responsible for *AqAF* regulation. While this would again suggest that the *A. queenslandica* and *C. prolifera* allorecognition systems are quite different, it is here not implausible that changes to gene regulation might occur since these species diverged from their common ancestor. Finally, it may be the case that the original reports of *CpAF* activity in *C. prolifera* grafts (Fernàndez-Busquets et al. 1998) do not accurately reflect AF expression patterns. While *MAFp3* and *MAFp4* expression was shown to increase in grafted tissue (Fernàndez-Busquets et al. 1998), the same study showed considerable variation in *MAFp3/MAFp4* expression in various ungrafted conditions - for example, between different cut or whole individuals, samples taken from the same individual 24 hours apart, and between ungrafted tissue slices across time and/or different individuals. As this data is entirely qualitative, however, the apparent fluctuations in expression are difficult to interpret, and separating the individual-, allogeneic-, isogeneic- and daily cycle-specific effects on MAFp3/MAFp4 expression in this species is complicated. This could be examined in a fully replicated quantitative study of the dynamics of AF gene expression in different individuals and species, whereby the expression levels of genes were compared at different points in the day-night cycle and the tissue healing process, as well as in self and nonself grafts.

6.5.3 The *A. queenslandica* genome does not undergo wide-scale alternative splicing changes across most of the graft time course

The relative distributions of the different types of alternative splicing events observed transcriptome-wide in control ungrafted tissue were similar to those observed in analyses performed on other adult whole transcriptome datasets (discussed in Chapter 4). The largest difference between these analyses was the observation of a higher level of transcript initiation (5%) and termination (4%) within introns in the present study (compared with 0.5% each in previous analyses; S. Fernandez Valverde and B. Degnan, manuscript in preparation). This finding is likely due to residual noise from incompletely assembled transcripts exhibiting intron retention events that were not removed despite preliminary transcript filtering (Figure 6.5). However, further filtering was not undertaken as the relative proportion of these splicing categories is small, and remained consistent in all five transcriptome groups analysed, rather than exhibiting sample-specific bias.

Overall alternative splicing frequencies (Figure 6.11f) remained constant across the tissue grafting response, as did the relative distributions of different types of splicing events (Figure 6.11a-e). This finding does not, however, mean that individual genes did not exhibit different splice patterns in response to different alloimmune states. Intron retention was the most commonly observed alternative splicing category, while exon skipping and the introduction of novel initiation or termination sites within introns were the least commonly observed splicing events. This is consistent with observations from other *A. queenslandica* datasets (S. Fernandez Valverde and B. Degnan, manuscript in preparation) and in other non-eumetazoan eukaryotes (Kim et al. 2006; Wang and Brendel 2006; McGuire et al. 2008; Westbrook 2011).

6.5.4 The *AqAFs* exhibit intron retention and possible exon skipping events in a non-allorecognition-specific manner

Alternatively spliced transcripts were identified from *AqAFA* to *AqAFE*. Alternatively spliced *AqAF* transcripts were present at all grafted time points; none, however, were identified from either of the two ungrafted control samples. This finding was unexpected because alternatively spliced transcripts from *AqAFA* to *AqAFD* were previously identified from another *A. queenslandica* adult

transcriptome (Figure 4.4). However, failure to detect transcripts in datasets of this nature does not constitute biological proof of absence.

As is the case transcriptome-wide, full or partial intron retention events were the most commonly observed changes to *AqAF* transcript structure in the grafted sponge. A number of the observed events were technically classified by PASA as instances of transcript initiation or termination within an intron, however the high frequency of assembly truncation within these datasets means it cannot currently be determined whether the majority of these transcripts represent true initiation/termination events or truncated intron retention events that were mis-classified due to assembly artefacts. Contrary to observations discussed in Chapter 4, a small number of exon skipping events were detected, however all but one transcript (from *AqAFA*) were localised to repeated exons (Chapter 2.4.6) and therefore likely to represent assembly artefacts rather than biologically-meaningful splice variants. However, PCR-based sequencing data is required to draw conclusions either way regarding this point.

The majority of retained introns introduce premature termination codons (PTCs) into the transcripts. The presence of these PTCs suggests that these transcripts are possible targets of the nonsense mediated decay (NMD) pathway, which can detect and degrade erroneously spliced transcripts (Losson and Lacroute 1979) but may also serve as a regulatory mechanism to control transcript abundance and gene product activity (reviewed by Ge and Porse 2013). As several particular intron retention events were observed in multiple independent datasets analysed both here and in Chapter 4, it seems likely that these transcripts are biologically significant. It is unknown whether these transcripts are indeed targeted by the NMD pathway or if they are protected and subsequently encode functional RNAs or proteins. The introduced PTCs from *AqAFC* and *AqAFE* would, in the latter case, terminate the encoded protein approximately two-thirds of the way along its length, while those identified from *AqAFB* would result in a termination event partway through the encoded Wreath domain. Truncation of the proteins upstream of the Wreath domain may serve a regulatory function for *AqAF* activity, for example by controlling the amount of protein available to form the AF core structure. Further functional studies are required to confirm the existence of these intron retention events, and to determine whether the subsequent introduction of a PTC results in a functional but truncated protein.

A number of retained introns from *AqAFB, AqAFC* and *AqAFD* introduce putative signal peptide-encoding regions to the sequences, usually either at the start of an assembled transcript or within a retained intron, following an upstream stop codon introduced by the same intron. These introductions are of particular interest given the relative improbability of an intron encoding a putative signal peptide by chance, suggesting that these trancripts are indeed biologically significant. In *AqAFB*, the putative signal peptide occurs immediately upstream of the sequence encoding the Wreath domain, while in *AqAFC* and *AqAFD* the signal peptide occurs a few exons upstream of this domain, so that a Von Willebrand type D and a full or partial Calx-beta domain would also be included in the predicted resulting protein. Short signal peptide- and Wreath domain-encoding transcripts have been identified in other sponge species (Figure 2.7), which may also represent alternatively spliced transcripts similar to those identified here. This suggests that sponges may regulate expression of various AF structures, which may operate in different biological contexts. For example, the short Wreath domain-equipped proteins may form a core AF backbone structure (either linear or circular depending on the species) that could serve as an inhibitory molecule to competitively block some downstream AF-mediated pathway or response. As also seen in Chapter 4, all retained introns encoding novel signal peptide were situated close to the start of the encoding transcripts. This could be an assembly artefact, however it is possible that novel transcription initiation sites exist for these genes.

Alternative splicing of allorecognition molecules appears to be a common strategy to generate diversity or suites of molecules with context- or tissue-dependent roles. For example, *fester* and *uncle fester*, of the *Botryllus schlosseri* histocompatibility system, are both alternatively spliced (Nyholm et al. 2006; McKitrick et al. 2011), as is the *Hydractinia symbiolongicarpus* allorecognition gene *alr1* (Rosa et al. 2010). Notably, these and other immune-related genes (Ghosh et al. 2011) predominantly employ exon skipping to generate alternate isoforms. If real, the intron retention events detected here and previously (Figure 4.4) therefore appear unique amongst other characterised invertebrate allorecognition molecules.

6.5.5 Graft transcriptome samples exhibit greater between-individual than between-time point variance

Principal component analysis of the most dynamically expressed genes across the three graft time courses revealed greater divergence between sponge individuals than between immune states (Figure 6.7). This between-individual variance was not revealed until after sequencing was complete; it is unknown whether this degree of variance is representative of the *A. queenslandica* population as a whole, or if one or both of the sponges used for this analysis was unusually divergent. Regardless, this between-individual difference proved to be a limitation for quantitative analysis of the graft response, as the two self time courses could not be analysed as simple replicates of one another; doing so resulted in the detection of very few differentially expressed genes in all comparisons tested (data not shown). To account for this interindividual variation, I designed a reduced experimental model in which samples within a time course were treated as replicates of one another in order to calculate a global common dispersion value. This common dispersion value was then applied to the full design model, in which each time point was considered separately. This common dispersion value therefore encompasses the self and nonself graft-induced biological variation; it is therefore unsurprising that low counts of differentially expressed genes were identified within the two self graft comparisons. Future studies could repeat the graft experiments and subsequent transcriptome preparation performed here. Improved biological replication would allow a more robust analysis of the changes occurring across the graft time course, as well as the relative contributions of individual diversity and time post grafting on expression dynamics.

6.5.6 Differential gene expression analysis

Relatively low numbers of genes were found to be differentially expressed between successive autograft time points. The allograft time course exhibited more dynamic expression across time, particularly between the 24 and 48 hpg time points (Figure 6.13). Around these times in the allograft time course, transition occurs from a transitory fusion state (which was observed between 12 to 48 hpg, though this timing varied between individuals) to a rejection state. It is therefore possible that the large changes in nonself gene expression at this time are functionally related to this transition. However, further data is required to explore this point.

Sixty-five percent of differentially expressed genes identified within the nonself time course were found to be downregulated within their relevant pair of time points. The downregulated genes were statistically enriched for GO terms associated with key biological processes such as cell signalling, transcription and translation, protein and molecular transport and other metabolic processes (Appendix 6.5). This may indicate that a key response to nonself grafting is the shutdown of regular biological processes, rather than a shift to defensive gene expression. As small tissue slices were taken directly from the graft interface, it is unknown whether this hypothetical shutdown is localised to the point of contact, or extends deeper into the grafted tissue. Cell-type specific infiltration of the graft interface could also impact the transcriptional landscape in the immediate vicinity of the graft interface. Normal cell signalling appears to be downregulated in response to nonself grafting; for instance, genes with functions associated with signalling pathways such as ubiquitin transferase, or GTP or metal ion binding activity were downregulated at 12 hpg relative to the control state (Appendix 6.5), while genes with more generalised cell signalling roles were downregulated at both 48 and 72 hpg relative to the previous time points (Appendix 6.5). However, a suite of other cell signalling genes were also upregulated at 48 hpg, perhaps indicating a shift to rejection signalling processes, or that previously-downregulated cell signalling genes were being reactivated at this time.

A transcriptional shutdown in response to graft rejection has been reported in microarray analyses of gene expression in the botryllid ascidian *Botryllus schlosseri*. In this species, rejection reactions are asymmetric, where one graft partner develops morphological signs of rejection (the 'rejected' individual), while the other partner does not (the 'rejecting' individual) (Oren et al. 2010). Rejected individuals within a graft showed limited gene upregulation relative to the naive state, but extensive downregulation of genes involved in protein biosynthesis, cell structure and motility, and immune function; rejecting individuals showed limited changes relative to the naive state (Oren et al. 2010). Here it was hypothesised that the rejected individual undergoes a period of tissue self-destruction, in order to facilitate physical tissue separation from the rejecting individual, and to inhibit interference of this separation process by the immune or tissue healing systems (Oren et al. 2010). It is possible that a similar tissue avoidance strategy is in place in *A. queenslandica*. Additionally, although no obvious physiological signs of a 'rejected/rejecting' hierarchy have been noted within *A. queenslandica*,

characterisation of the morphological graft response has not been extensive to date, meaning that such hierarchy may operate in a molecular or physiological manner in some or all instances of graft rejection.

6.5.7 Conclusion

The *A. queenslandica* allorecognition decision appears to occur over a period of three days after grafting. Self grafts initiated fusion between 12 to 48 hpg, and the graft interface had nearly completely disappeared after 72 hpg. The outward nonself graft response does not appear to be aggressive (e.g. involving chemical attack of one graft partner), with both tissue partners remaining alive and healthy for the duration of the graft response. Allografts may undergo a period of transitory fusion between 12 and 48 hpg, where weak bonds formed between the tissue slices, possibly to allow direct cell-cell contact between the rejecting tissues. A preliminary analysis of the global transcriptional changes occurring during this time suggests that the allograft response is characterised by the shutdown of normal biological processes, rather than the initiation of a defensive response. Grafted tissue also does not appear to use alternative splicing on any large scale during the allorecognition response. Contrary to prior reports from the demosponge *C. prolifera* (Fernàndez-Busquets et al. 1998), expression of the the *A. queenslandica* AFs was not found to change during the auto- or allograft responses. It remains unknown whether this indicates that the AqAFs are not involved in allorecognition, or if they are involved but ubiquitously highly expressed. However, alternatively spliced AqAF isoforms, some of which were equipped with novel signal peptides, were identified within graft tissue, although no clear correlation between these isoforms and graft state was observed.

Chapter 7 - General Discussion

7.1 Overview

Sponges are representatives of one of the oldest extant metazoan lineages and are an informative model phylum for studying the transition to a multicellular state. Allorecognition - discrimination between self and conspecific nonself upon physical contact - is a key requirement for successful multicellularity (Buss 1987). Aggregation factors (AFs) are sponge-specific proteoglycans that drive selective reaggregation of dissociated sponge cells, and are also the proposed molecular determinants of sponge allorecognition (Chapter 1.4). The glycan components of the AFs are important mediators of AF-AF and AF-cell interactions (Misevic and Finne 1987; Misevic and Burger 1990a; 1990b; 1993). However, the head subunit region of the protein backbone may also aid cell aggregation (Jarchow et al. 2000), suggesting that this backbone is functionally important beyond serving as a passive scaffold for its attached glycan moieties. For this thesis, I sought to further explore the properties of this protein backbone, an avenue of inquiry that is increasingly feasible with the advent of accessible genome and RNA sequencing technologies.

The AF proteins appear to be a hexactinellid + demosponge-specific innovation (based on a definition that a candidate AF should possess a Wreath domain, or multiple Calx-beta domains plus top sequence similarity to other known AFs), that in *Amphimedon queenslandica* are comprised of Calx-beta, Von Willebrand and Wreath domains. Some *A. queenslandica AF* (*AqAF*) transcripts are diversified by intron retention and appear to generate novel shortened AF isoforms, and analysis of nucleotide polymorphism between individuals indicates the *AqAF* sequences vary between individuals and may be under positive selection. The *AF* genes were not found to demonstrate variable transcript levels in response to self or nonself tissue grafting, possibly being regulated upstream by the nonsense mediated decay (NMD) pathway and/or other mechanisms. In contrast, the AFs were very highly expressed across sponge development, suggesting the existence of a novel developmental role for these genes.

7.2 Evolution of poriferan aggregation factors

AFs have been studied to varying degrees in several model sponge species, particularly the demosponges *Clathria prolifera* and *Geodia cydonium* (Chapter 1.3). To the best of my knowledge, no serious attempts have been made to catalogue the AFs (either the proteoglycan complexes or underlying sequences) that exist across the Porifera. The first goal of my thesis, therefore, was to perform the first systematic survey of AFs in multiple sponge genomes and transcriptomes, in order to infer the evolutionary origin point of the AFs.

7.2.1 What is an AF?

The precise identification of candidate AF sequences across the phylum Porifera is reliant on the use of accurate criteria for sorting AF from non-AF sequences. In their 1996 study, Fernàndez-Busquets et al. isolated the native *C. prolifera* AF proteoglycan complex and performed N-terminal sequencing to determine a short portion of AF amino acid sequence. Degenerate primers were designed to target matching nucleotide sequences in a complementary DNA (cDNA) library (Fernàndez-Busquets et al. 1996). Therefore, a direct relationship exists between the known functional AF complex and the derived DNA/protein sequences for this species. The *C. prolifera* sequences and that of the related *Suberites domuncula* protein, SLIP, were used for similarity searches against the *A. queenslandica* genome, resulting in the identification of six candidate *AF* genes from this species (Gauthier 2009).

The *A. queenslandica*, *C. prolifera* and *S. domuncula* AF and AF-like predicted proteins do not possess large stretches of highly similar sequence between genes or species; instead only certain, structurally important (Hilge and Aelen 2006) residues are maintained (Figure 2.6). All sequences, however, show similar domain architectures; all contain one or multiple Calx-beta domains, and most (Gauthier 2009) contain the region that was originally identified as MAFp3 (Fernàndez-Busquets et al. 1996) but which I have classified as a probable novel protein domain, the Wreath domain (Chapter 2.4.1). In addition, the *AqAFs* incorporate up to six Von Willebrand type A or D domains (Gauthier 2009). Because of the architectural consistency between species, and for reasons of practicality (as manual inspection of divergent sequence similarity results for large datasets is both subjective and slow), I decided to use domain architecture as the main criterion for selection of candidate AFs, with sequence similarity as a secondary requirement. Sushi domains, as seen in the candidate core *G.*

cydonium AF, *GEOCY_AF*, were not included as search criteria as this form has only been observed in one species and has not been well characterised.

Wreath domains in *C. prolifera* form the circular backbone of the sunburst-like AF core (Jarchow et al. 2000). As circular proteoglycans have not been observed outside the sponges or playing non-AF sponge roles (reviewed by Fernàndez-Busquets and Burger 2003), the presence of a Wreath domain is currently the best indicator of a likely AF. However, it is unknown at present whether non-AF genes might possess Wreath domains, or if AF genes in some species (for instance, those exhibiting a linear AF form) may lack this domain type. It is also possible that some true AF sequences were truncated during *de novo* transcript assembly, resulting in a transcript without a Wreath domain. All sequences equipped with multiple Calx-beta domains and displaying top sequence similarity to a known AF were included as Group 2 candidate sequences, to allow for the latter two possibilities. Ideally, future studies would focus on trying to find direct links between the AF proteoglycan complexes of different species and their underlying protein sequences (as per Fernàndez-Busquets et al. 1996), to better establish whether the AF candidate filtering criteria used in the present work are indeed valid and representative.

7.2.2 Where did the AFs evolve?

A model candidate AF sequence should contain either a Wreath domain (Group 1), or multiple Calx-beta domains plus top sequence similarity to a known AF sequence (Group 2). Given this definition, the AFs appear to be a demosponge + hexactinellid-specific innovation. Wreath domains, and sequences equipped therewith, were identified only in demosponges (Figure 2.10). The sole available hexactinellid *Aphrocallistes vastus* possessed a single Group 2 sequence and no Group 1 sequences. When the genomes or transcriptomes of additional hexactinellid species are sequenced - an almost inevitable eventuality given the increasing uptake of sequencing technology - searches should be performed to help confirm the absence of the Wreath domain, and to verify whether likely AF candidates are present in this species. This will help resolve the evolutionary origin of the sponge AFs.

No likely AFs were identified in the examined calcareous or homoscleromorph sponges, despite the availability of complete genome sequences for two of the four species (*Oscarella carmela* and *Sycon ciliatum*) (Fortunato et al. 2015; Nichols et al. 2012). This is intriguing in light of the presence

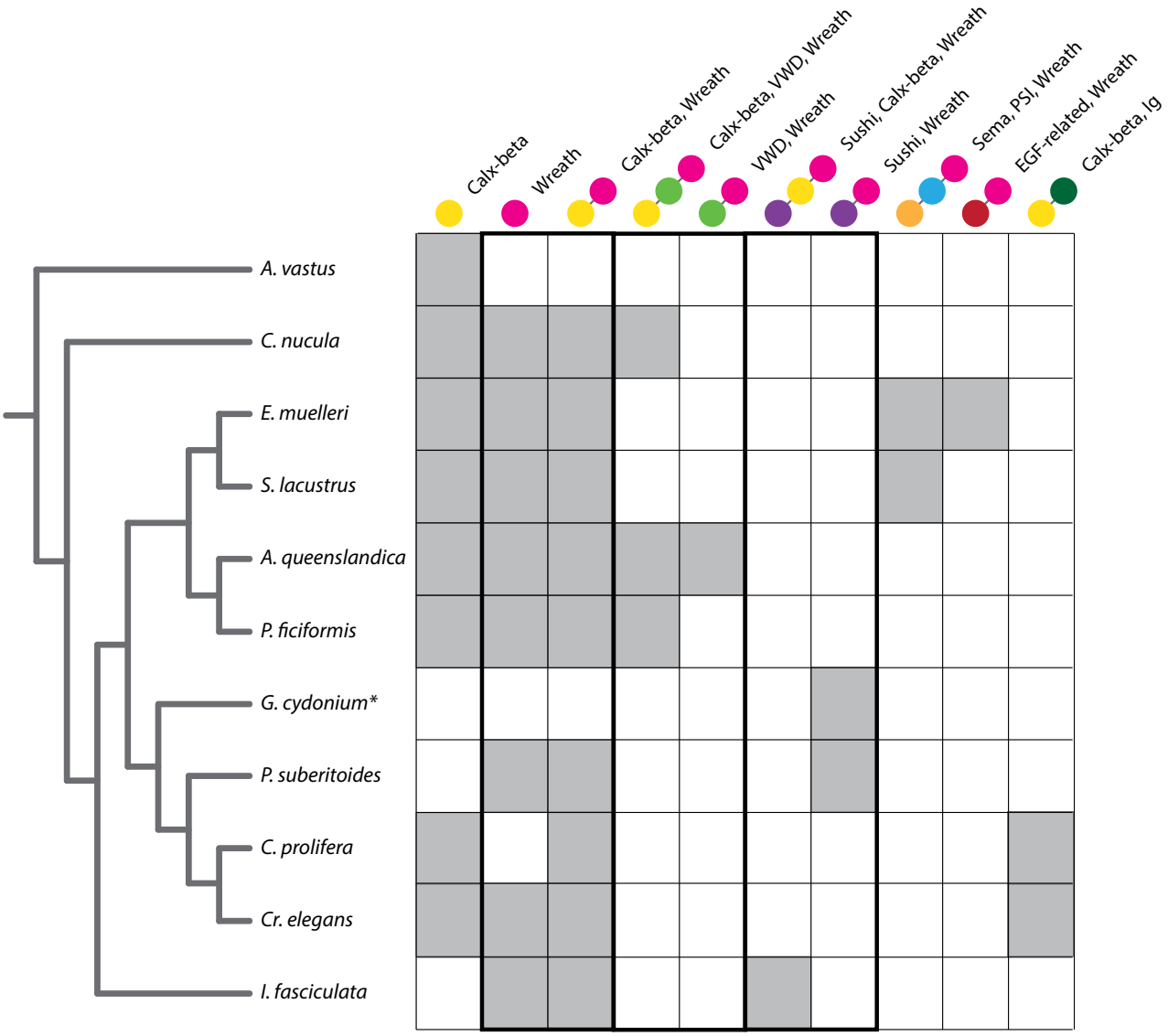


Figure 7.1 Phylogenetic distribution of AF candidate domain architectures
The domain architectures of all identified Group 1 and 2 AF candidates (Chapter 2) for each analysed species are shown. Domain architecture combinations are shown at the top of the table, with each domain type shown only once per model; recurring domain types within single sequences were grouped together. The phylogenetic relationships between all sponge species, as determined by Thacker et al. (Thacker et al. 2013), is shown at the left. Grey boxes indicate the presence of one or more proteins encoding each model type. * includes only the *G. cydonium* protein described by Müller et al. (Müller et al. 1999a).

of a circular AF-like structure in the homoscleromorph *Oscarella tuberculata* (Humbert-David and Garrone 1993), and the ability of the calcareous sponge *Leucandra abratsbo* to discriminate between self and nonself in graft experiments (Amano 1990). Calcareous sponges, however, cannot undergo AF-mediated secondary cellular reaggregation (Müller 1982). These findings suggest that either the AFs are not actually key players in sponge allorecognition (despite the other evidence supporting this

hypothesis; Chapter 1.4), or that the mechanisms of allorecognition response differ between calcareous sponges and demosponges.

7.2.3 What do the AF proteins look like?

The Group 1 and 2 candidate AFs are composed primarily of Wreath and/or Calx-beta domains. While sequences encoding this basic architecture are present in all examined demosponge and hexactinellid species, some also contained sequences with additional domain types. The AFs of three species (*A. queenslandica*, *Chondrilla nucula* and *Petrosia ficiformis*) are equipped with Von Willebrand domains, and two species (*Pseudospongosorites suberitoides* and *Ircinia fasciculata*) incorporate Sushi domains in different combinations. *G. cydonium* candidate AF sequence also encodes one Wreath and two Sushi domains (Müller et al. 1999b). The distributions of both the Von Willebrand and Sushi domain-equipped sequence types are such as to suggest that similar domain architectures were encoded within the ancestral demosponge genome and were lost in multiple subsequent lineages, or that these forms arose through convergent evolution (Figure 7.1). The two studied representatives of the haplosclerid suborder Spongillina (*Ephydatia muelleri* and *Spongilla lacustrus*) both encode sequences equipped with PSI, Sema and Wreath domains, suggesting that this domain combination is an evolutionary novelty limited to this group (Figure 7.1). Similarly, the two studied sponges from the order Poecilosclerida (*C. prolifera* and *Crella elegans*) both include different combinations of immunoglobulin superfamily domains in some Group 2 sequences; this could again either represent diversification from a poecilosclerid ancestral form, or convergent evolution (Figure 7.1). Finally, several transcripts from *E. muelleri* incorporate a variety of EGF-related domains along with Wreath domains, however this form was not observed outside this species (Figure 7.1).

The identification of these new domain combinations in candidate sequences further emphasises the importance of a clear definition of an AF. AF sequences are currently best defined by the presence of a Wreath domain, however an optimal definition would rely not only the sequence features, but also the functions of characterised AFs. In Chapter 2 I proposed that the Wreath domain facilitates AF circular or linear backbone formation (depending on species, though it is possible that the Wreath domain might facilitate both forms in a single species), regardless of its associated domains and functions. Functional studies are required to determine whether the hypothetical backbones that are

formed by Wreath domain-equipped sequences with other novel domain types contribute to AF-like cell adhesion, or play other unknown roles.

7.3 Diversification of the *A. queenslandica* AFs

The demosponge *A. queenslandica* is presently the only sponge species that is equipped with AFs and also has an available sequenced genome. I performed the first in-depth characterisation of the genomic and transcriptomic properties of the *AqAF* genes; the majority of this thesis details the outcomes of this research. To fulfil the second and third goals of this project, I investigated the normal genomic features of the *AqAFs*, and the potential ways in which these sequences might be diversified to generate molecules with sufficient variability to mediate self-nonself recognition between conspecifics.

7.3.1 Genomic architecture and splicing of the *A. queenslandica* AFs

The five main *AqAFs* (*AqAFA* - *AqAFE*) are large genes that are mostly comprised of many short introns and exons (Table 2.1). Ninety-nine percent of the *AqAF* exons are symmetrical (i.e. flanked by introns in the same phase), meaning that exon rearrangement of the resulting transcript could occur without disruption to the translational reading frame. Alternative splicing is a commonly observed form of immune and allorecognition molecule diversification (reviewed by Ghosh et al. 2011), occurring for example in the allorecognition molecules of the ascidian *Botryllus schlosseri* (Nyholm et al. 2006; McKitrick et al. 2011) and the colonial hydroid *Hydractinia symbiolongicarpus* (Rosa et al. 2010). However, searches for alternative exon usage across seventeen *A. queenslandica* transcriptomes and with the polymerase chain reaction failed to reveal convincing evidence that this is a widespread mechanism of AF diversification. Instead, I identified instances of intron retention across all six *AqAFs*. The majority of these retention events introduce premature termination codons. NMD may act upon these unviable transcripts, either to remove erroneously spliced transcripts, or to regulate *AqAF* transcript abundance (reviewed by Ge and Porse 2013). The observation of retention of the same introns in multiple trancsriptomes may suggest that the latter process is used as a control mechanism for the AFs. A subset of *AqAFB, AqAFC* and *AqAFD* retention events, however, also introduced predicted novel signal peptides, that preceded the final Calx-beta, Von Willebrand and Wreath domain (or Von Willebrand and Wreath domain only, in the case of *AqAFB*) of each gene. Similar short transcripts encoding predicted signal peptides were observed in *C. nucula* and *E. muelleri*. Without knowing the

genomic sequences encoding these transcripts, it is unknown whether these are the result of alternative splicing or represent fully transcribed genes; given that much longer AF candidates were identified from both species, the former option seems likely.

7.3.2 Sequence variation in the *AqAFs*

I detected a high degree of nucleotide variability between the *AqAFs* of three adult *A. queenslandica* individuals. In total, the *AqAFs* of each individual displayed an average of ~400 variant nucleotide sites, which together result in the existence of six unique alleles per gene across the three individuals. The *AqAFs* also show a significant increase in non-synonymous nucleotide changes relative to the frequency observed across the genome as a whole. This suggests that positive selection may act upon the *AqAFs* to drive sequence diversification of these putative allorecognition molecules. This, however, remains to be statistically tested. Nucleotide polymorphisms in the present study were detected within short sequencing reads produced by high-throughput RNA sequencing with both alleles per individual mixed; analysis of positive selection would be better surveyed within discrete alleles generated by cloning and direct sequencing. Downstream statistical analysis could be performed, for example, as per Nicotra et al. (2009).

The observation of sequence polymorphism across the *AqAFs* supports the findings of Fernàndez-Busquets et al., who demonstrated the existence of variability in both the *C. prolifera* AFs and their associated glycans (Fernàndez-Busquets and Burger 1997). Polymorphism has also been observed in the well-characterised self-nonself recognition systems of other invertebrates, such as *B. schlosseri* (De Tomaso et al. 2005; Nyholm et al. 2006; Nydam and De Tomaso 2012; Nydam et al. 2012; 2013a; 2013b; Voskoboynik et al. 2013), *H. symbiolongicarpus* (Nicotra et al. 2009; Rosa et al. 2010) and the sea urchin *Strongylocentrotus purpuratus* (Nair 2005). Such polymorphic allorecognition molecules are often also found to be under positive selection, as seen in the *B. sclosseri* genes *fester* (Nydam and De Tomaso 2012), *Hsp40-L* (Nydam et al. 2013a), *mFuHC*, and *sFuHC* (Nydam et al. 2012); *alr1* (Rosa et al. 2010) and *alr2* (Nicotra et al. 2009) from *H. symbiolongicarpus,* and *Sp185/333* from *S. purpuratus* (Nair 2005).

I have demonstrated the existence of the ADAR (adenosine deaminase acting on RNA) class of RNA editing molecules in the earliest phyletic branches of the crown Metazoa, by surveying the genomes and transcriptomes of thirteen sponge and ten ctenophore species. This finding supports that of Moroz et al. (2014), who identified ADAR sequences in the genome of the ctenophore *Pleurobrachia bachei*. Together, these results suggest that this post-transcriptional regulatory mechanism was in place in the last common ancestor to the metazoans, and has been preserved in *A. queenslandica* and other lineages. It is therefore mechanistically possible that the *AqAFs* are diversified by RNA editing, as occurs for example in the *Sp185/333* transcripts of *S. purpuratus* (Buckley et al. 2008). While functional investigation of this hypothesis is beyond the scope of this thesis, preliminary studies elsewhere in the Degnan lab suggest that extra-genomic *AqAF* nucleotide variability does exist in some individuals (K. Roper, personal communication).

7.4 Expression of the *A. queenslandica AF* genes

After analysing the genomically-encoded *AqAF* genes and the potential ways in which these genes are diversified between individuals, the final goal of this thesis was to investigate the activity of the *AqAFs in vivo*. To do so, I analysed the changtes in *AqAF* gene expression across sponge life history in a normal, non-immunologically challenged context, before surveying for potential changes to this expression pattern in adult sponges upon tissue contact with another conspecific individual.

7.4.1 A putative developmental role for the *A. queenslandica* AFs

The developmental expression profile of the *AqAFs* is significantly correlated with 122 other *A. queenslandica* genes, most of which have cell signalling related functions. *AqAF* expression is also very high relative to the rest of the genome, particularly shortly after the commencement of metamorphosis, and occurs prior to the onset of sponge immunocompetency. As many of these other genes play core roles in development and basic sponge biology, I propose that the AFs work together with these molecules to play an important developmental function, in addition to their putative allorecognition role in the mature sponge. A joint role in self-nonself recognition and development for the *AqAFs* would not be surprising, as similar dual functions have been observed elsewhere. For instance, in the ascidian *Boltenia villosa*, a number of innate immune-associated genes are upregulated during metamorphosis (Roberts et al. 2007). Similarly, *A. queenslandica* Toll pathway components, *AmqMyD88, AmqIgTIR1,*

AmqlgTIR2, and *AmqTollip* are developmentally expressed, but also respond transcriptomically upon exposure to microbial signals in the form of lipopolysaccharide (LPS) endotoxin and the marine bacterium *Vibrio harveyi* (Gauthier 2009).

While I have demonstrated a statistically significant correlation in expression between the *AqAFs* and the other identified genes, and discussed previously-described relationships between the *AqAFs* and the other gene systems, I have not attempted to demonstrate a mechanistic connection between the genes to show that they are co-regulated. Such information would, however, be informative. This could be tested by the application of various drugs that affect the binding partners, or of antibodies that block the *AqAFs* in development, and observation of the resulting phenotypes.

7.4.2 *A. queenslandica AF* expression does not change in response to tissue grafting

The *AqAF* genes are very highly expressed in adult sponges, however in the present study, *AqAF* gene expression did not appear to be affected by self or nonself tissue grafting. This result is contrary to the findings of Fernàndez-Busquets et al. (Fernàndez-Busquets et al. 1998), who reported that the *C. prolifera* AF genes *MAFp3* and *MAFp4* appear to be upregulated in both auto- and allografts. The considerable genetic variability observed between *A. queenslandica* individuals in this study may indicate that a larger sample size is required to detect statistically significant expression changes. However, if the *AqAFs* are indeed stably expressed in grafted tissue, this may indicate that AF activity is controlled above the level of transcription. For instance, if selective intron retention occurs within the *AqAF* genes, NMD may serve as a post-transcriptional regulatory mechanism to control the rate of AqAF production. Alternatively, the *AqAFs* may be ubiquitously expressed, and later selectively glycosylated when enhanced aggregative activity is required. It is also possible that AF activity is modulated downstream, for instance by controlling the expression or activity of the aggregation receptor/s (AR) or of associated signalling molecules.

7.5 Synthesis of findings

Early studies of the cell reaggregation process were complicated by the great diversity of responses to cell-cell contact, including complete intermingling of xenospecific cells, and complete or partial sorting by species, individual or cell type (reviewed Fernàndez-Busquets and Burger 1999). The

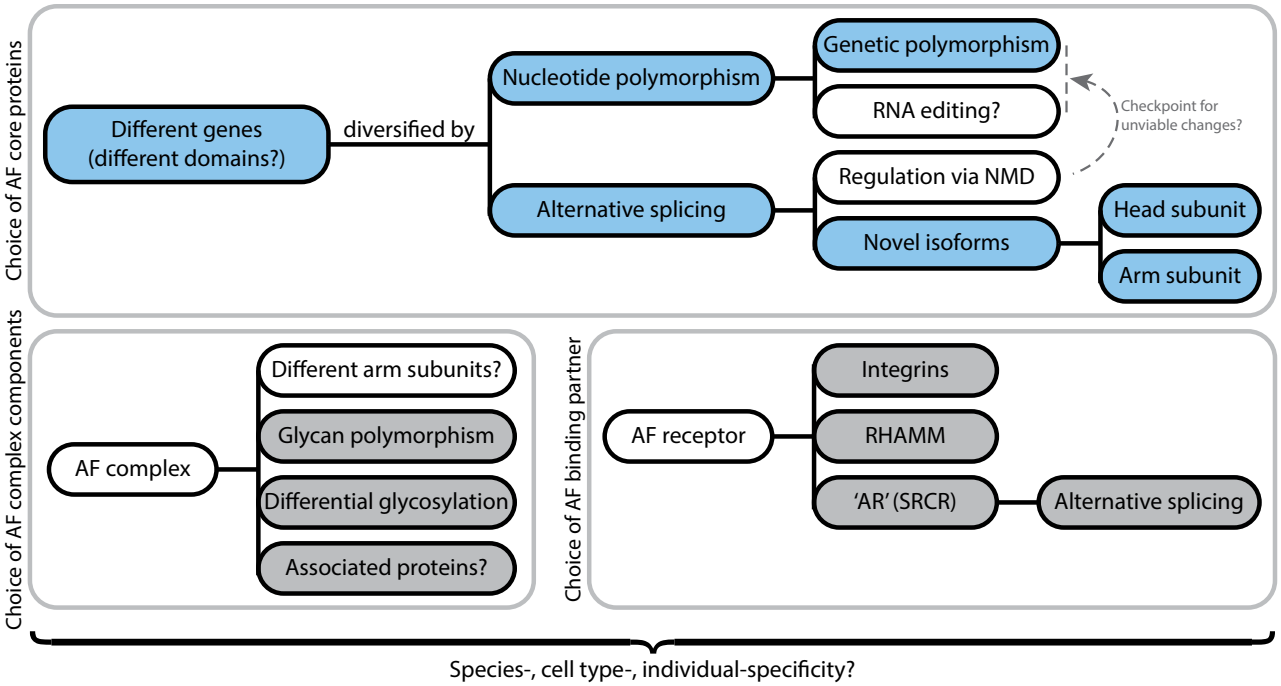


Figure 7.2 Proposed mechanism of AF versatility
See text for details.

differences in response appear to differ depending on experimental setup, phylogenetic relationships between species, and the rates at which reaggregation occurs (Fernàndez-Busquets and Burger 1999). By synthesising previous findings from the literature with the conclusions of the present work, I speculate that the AFs are versatile molecules that facilitate a variety of processes in different biological contexts, including the ability to mediate cell-cell interactions in species-, cell type-, and individual-specific manners within a single individual. This synthesis is outlined below and in Figure 7.2; it should, of course, be noted that this model is speculative and requires further research to support some connections drawn between lines of evidence.

7.5.1 Choice of AF core proteins

The *A. queenslandica* genome encodes six *AqAF* genes, each of which encodes similar, but distinct, domain conformations. It is currently unknown how *A. queenslandica* deploys the different *AqAF* genes, and whether they work together or independently in different contexts (or a combination of both). The genes do show similar expression profiles across development, suggesting a degree of cooperation between the genes. However, the existence of multiple similar *AqAF* genes may suggest that this sponge, and others, can 'choose' which gene/s to deploy in a context-dependent manner, in

order to enhance AF versatility. In species displaying a broader range of domain types coupled to Wreath domains - such as *E. muelleri,* where Wreath domains are coupled to Calx-beta, EGF-related or Sema and PSI domains - the outcomes of this choice may be more pronounced. The different arm subunits of the AF core are likely to have different binding partners or specificities (discussed further below), allowing AFs or AF-like structures to perform different functions.

For a given *AF* gene, further diversity appears to be introduced in the form of genomically-encoded polymorphisms. As discussed in Chapter 4.4.3, the *A. queenslandica* AFs display a higher proportion of non-synonymous nucleotide polymorphisms than is seen across the genome as a whole, suggesting that the *AqAFs* may be under positive selection. Such nucleotide changes have the potential to alter the secondary structure of the resulting protein, as well as adding or removing particular functional motifs such as glycosylation and protein binding sites. Further post-transcriptional nucleotide changes may also occur, if RNA editing does indeed operate upon the *AqAF* sequences. To allow greater freedom for experimentation with sequence diversification, NMD may act as a check point to detect non-viable transcripts containing premature termination codons caused by errant nucleotide polymorphisms.

Alternative splicing has the potential to further modify particular AF sequences. I have shown that intron retention occurs across the *AqAFs*, with two possible outcomes. First, certain intron retention events in *AqAFB, AqAFC* and *AqAFD* are predicted to introduce a novel signal peptide to the resulting sequence. In all cases but one, the retained intron sits at the start of the *de novo* assembled transcript, suggesting the existence of alternative transcriptional start sites for these transcripts. A single *AqAFC* transcript from competent larvae is predicted to introduce a signal peptide inside a longer transcript; it is currently unknown whether this is real or an asselbly artifact. All introduced signal peptides sit towards the end of the transcripts, and in these cases, shortened sequences encoding the Wreath domain region (sometimes with an attached Von Willebrand and/or Calx-beta domain) are predicted to be produced. Alternatively, in most instances intron retention introduces a premature termination codon to a sequence, upstream of the Wreath domain. If such transcripts are successfully translated, the resulting protein would be a partial or full arm subunit region. Alternatively, truncated transcripts may be targeted by the NMD pathway, possibly as a way to regulate expression of the *AqAFs* in lieu of changes to transcriptional abundance.

7.5.2 Choice of AF complex components

AF-mediated cell adhesion relies on a complex association of molecules, that include the core AF protein, associated glycan subunits and proteins, and a membrane-bound aggregation receptor (reviewed by Fernàndez-Busquets and Burger 2003). Differential assembly of the AF complex may represent a way by which the sponge could modulate AF behaviour and activity. The *C. prolifera* AF protein core is comprised of head and arm subunits. These are transcribed from a contiguous piece of RNA but later cleaved to produce independent subunits for the mature complex (Fernàndez-Busquets and Burger 1997; Jarchow et al. 2000), that are held together by glycan-glycan or glycan-protein interactions (Jarchow et al. 2000). As discussed above, some *AqAF* transcripts are truncated within the arm subunit. I speculate that these truncated forms may be incorporated into the AF complex, allowing the sponge to 'mix and match' different AF arm subunits to further alter AF diversity where appropriate.

Previous studies have examined the glycans associated with the AF complex, and found these moieties are also highly polymorphic between individuals (Fernàndez-Busquets and Burger 1997), demonstrating another way by which the AF complex might be diversified. Similarly, differential glycosylation of the AFs may be employed to regulate the adhesiveness of the AF complex. As AF binding relies on the polyvalent adhesiveness of many glycans acting in tandem (Garcia-Manyes et al. 2006), changing the glycosylation state of the AFs is likely to be an important regulator of AF activity *in vivo,* as has been proposed to be the case in the *C. prolifera* graft response (Fernàndez-Busquets et al. 2002).

Finally, a variety of additional proteins are associated with the AFs, such as the BIN1 protein from *G. cydonium* (Schütze et al. 2001) and p68 and p210 from *M. prolifera* (Varner et al. 1988; Varner 1995; 1996). The p210 protein may exhibit polymorphism at the protein level (Fernàndez-Busquets and Burger 1997), the polymorphic or splice state of the other proteins is unknown. However, if these protease are indeed diversified, it may introduce an additional layer of complexity to the AFs.

7.5.3 Choice of AF binding partner

A putative AR has been identified in *G. cydonium* (Blumbach et al. 1998). The longest form of this sequence encodes fourteen SRCR (scavenger receptor cysteine-rich) and six Sushi domains, plus

a transmembrane domain. However, the encoding gene appears to be alternatively spliced to generate shorter isoforms without some Sushi and/or the transmembrane domains (Blumbach et al. 1998). It has not been tested to date whether this gene is polymorphic between individuals. While no AR has been functionally identified in *A. queenslandica,* this species does encode a large number of SRCR domain-equipped proteins, some of which are expressed together with the *A. queenslandica* AFs across development (Appendix 3.3) and others which also possess Sushi and transmembrane domains (B. Yuen, personal communication). However, the AF complex may interact with multiple receptor types, allowing cell-cell interactions to trigger a range of downstream responses in a context-dependent manner. For example, the AFs may bind integrins (this work; Wimmer et al. 1999b; Fernàndez-Busquets and Burger 2003), suggesting that AF binding can promote downstream integrin signalling. The AFs from *A. queenslandica* and *C. prolifera* also contain a hyaluronan-binding motif (Fernàndez-Busquets et al. 1996; Kuhns et al. 1998), which may allow them to bind the hyaluronan receptor RHAMM (hyaluronan-mediated motility receptor) and trigger downstream signalling (Turley et al. 2002). The AFs may also have other binding partners; these could potentially be identified by pull-down assays using tethered AF proteoglycans as bait.

Therefore, I posit that the AFs are highly versatile molecules, that function by tethering neighbouring cells together by forming molecular bridges between them. However, the various AF core and complex properties discussed above allow this basic functionality to be applied to different contexts as required across the sponge lifecycle or in response to external stimuli.

7.6 Recommendations for future study

This study represents the first broad-ranging genomic analysis of candidate AF genes to date, and provides insights into the mechanisms by which the AFs might be diversified and regulated in the sponge. The AFs are complex molecules that I propose to be multi-purpose molecules that mediate cell-cell adhesions in individual-, cell type- and species-specific manners. The findings presented in the present work provide a foundation upon which to base and direct subsequent research; some such experiments are discussed throughout this document. Overall, I contend that the most informative avenue of future inquiry would be attempts to functionally confirm several of the findings discussed here. Also important is the resolution of a definition of an AF, that should be based on both sequence

features and experimental tests of cell aggregation abilities. Investigation of novel candidate AF forms, equipped with domain architectures that differ from the basic form seen in *C. prolifera* and *A. queenslandica*, will allow further resolution of this question of AF definition.

References

Abascal F, Zardoya R, Posada D. 2005. ProtTest: selection of best-fit models of protein evolution. *Bioinformatics* **21**: 2104–2105.

Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, Amanatides PG, Scherer SE, Li PW, Hoskins RA, Galle RF, et al. 2000. The genome sequence of *Drosophila melanogaster*. *Science* **287**: 2185–2195.

Adamska M, Degnan SM, Green KM, Adamski M, Craigie A, Larroux C, Degnan BM. 2007. Wnt and TGF-β expression in the sponge *Amphimedon queenslandica* and the origin of metazoan embryonic patterning. *PLoS ONE* **2**: e1031.

Adamska M, Larroux C, Adamski M, Green K, Lovas E, Koop D, Richards GS, Zwafink C, Degnan BM. 2010. Structure and expression of conserved Wnt pathway components in the demosponge *Amphimedon queenslandica*. *Evol Dev* **12**: 494–518.

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**: 3389–3402.

Amano S. 1990. Self and non-self recognition in a calcareous sponge, *Leucandra abratsbo*. *Biol Bull* **179**: 272–278.

Anavy L, Levin M, Khair S, Nakanishi N, Fernandez-Valverde SL, Degnan BM, Yanai I. 2014. BLIND ordering of large-scale transcriptomic developmental timecourses. *Development* **141**: 1161–1166.

Anders S, Huber W. 2010. Differential expression analysis for sequence count data. *Genome Biol* **11**: R106.

Assmann V, Jenkinson D, Marshall JF, Hart IR. 1999. The intracellular hyaluronan receptor RHAMM/IHABP interacts with microtubules and actin filaments. *J Cell Sci* **112**: 3943–3954.

Athanasiadis A, Rich A, Maas S. 2004. Widespread A-to-I RNA editing of Alu-containing mRNAs in the human transcriptome. *PLoS Biol* **2**: e391.

Bass BL, Weintraub H. 1988. An unwinding activity that covalently modifies its double-stranded RNA substrate. *Cell* **55**: 1089–1098.

Basu MK, Carmel L, Rogozin IB, Koonin EV. 2008. Evolution of protein domain promiscuity in eukaryotes. *Genome Res* **18**: 449–461.

Baumann CT, Lim CS, Hager GL. 1999. Intracellular localization and trafficking of steroid receptors. *Cell Biochem Biophys* **31**: 119–127.

Benabentos R, Hirose S, Sucgang R, Curk T, Katoh M, Ostrowski EA, Strassmann JE, Queller DC, Zupan B, Shaulsky G, et al. 2009. Polymorphic members of the *lag* gene family mediate kin discrimination in *Dictyostelium*. *Curr Biol* **19**: 567–572.

Bigger CH, Hildemann WH, Jokiel PL, Johnston IS. 1981. Afferent sensitization and efferent cytotoxicity in allogeneic tissue responses of the marine sponge *Callyspongia diffusa*. *Transplantation* **31**: 461–464.

Blumbach B, Pancer Z, Diehl-Seifert B, Steffen R, Munkner J, Müller I, Müller WEG. 1998. The putative sponge aggregation receptor. *J Cell Sci* **111**: 2635–2644.

Blumenthal T. 1998. Gene clusters and polycistronic transcription in eukaryotes. *BioEssays* **20**: 480–487.

Bodmer W. 1972. Evolutionary significance of the HL-A system. *Nature* **237**: 139–145.

Boehm T. 2006. Quality control in self/nonself discrimination. *Cell* **125**: 845–858.

Boehmer von H, Kisielow P. 1990. Self-nonself discrimination by T cells. *Science* **248**: 1369–1373.

Bogaerts A, Beets I, Schoofs L, Verleyen P. 2010. Antimicrobial peptides in *Caenorhabditis elegans*. *ISJ* **7**: 45–52.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**: 2114–2120.

Bonner JT. 2000. *First signals: The evolution of multicellular development*. Princeton University Press.

Bonner JT. 1966. *Size and cycle*. Princeton University Press, Princeton.

Bonner JT. 1988. *The evolution of complexity*. Princeton University Press, Princeton.

Bonner JT, Slifkin MK. 1949. A study of the control of differentiation: the proportions of stalk and spore cells in the slime mold *Dictyostelium discoideum*. *Am J Bot* **36**: 727–734.

Bork P, Holm L, Sander C. 1994. The immunoglobulin fold. *J Mol Biol* **242**: 309–320.

Bourgon R, Gentleman R, Huber W. 2010. Independent filtering increases detection power for high-throughput experiments. *Proc Natl Acad Sci USA* **107**: 9546–9551.

Boyle JS, Lew AM. 1995. An inexpensive alternative to glassmilk for DNA purification. *Trends Genet* **11**: 8.

References

Broad Institute. 2009. Aplysia Genome Project ed. Broad Institute. http://www.broadinstitute.org/science/projects/mammals-models/vertebrates-invertebrates/aplysia/aplysia-genome-sequencing-project (Accessed August 7, 2014).

Bucior I, Scheuring S, Engel A, Burger MM. 2004. Carbohydrate-carbohydrate interaction provides adhesion force and specificity for cellular recognition. *J Cell Biol* **165**: 529–537.

Buckley KM, Terwilliger DP, Smith LC. 2008. Sequence variations in *185/333* messages from the purple sea urchin suggest posttranscriptional modifications to increase immune diversity. *J Immunol* **181**: 8585–8594.

Burnet FM. 1971. Self recognition in colonial marine forms and flowering plants in relation to evolution of immunity. *Nature* **232**: 230–235.

Buscema M, Van de Vyver G. 1984. Cellular aspects of alloimmune reactions in sponges of the genus *Axinella* I. *Axinella verrucosa* and *Axinella damicornis*. *J Exp Zool* **229**: 7–17.

Buscema M, Van de Vyver G. 1983. Variability of allograft rejection processes in *Axinella verrucosa*. *Dev Comp Immunol* **7**: 613–616.

Buss LW. 1982. Somatic cell parasitism and the evolution of somatic tissue compatibility. *Proc Natl Acad Sci USA* **79**: 5337–5341.

Buss LW. 1987. *The Evolution of Individuality*. Princeton University Press, Princeton.

Cadavid LF, Powell AE, Nicotra ML, Moreno M, Buss LW. 2004. An invertebrate histocompatibility complex. *Genetics* **167**: 357–365.

Cadigan KM, Nusse R. 1997. Wnt signaling: a common theme in animal development. *Genes Dev* **11**: 3286–3305.

Campbell ID, Bork P. 1993. Epidermal growth factor-like modules. *Curr Opin Struct Biol* **3**: 385–392.

Cantí C, Nieto-Rostro M, Foucault I, Heblich F, Wratten J, Richards MW, Hendrich J, Douglas L, Page KM, Davies A, et al. 2005. The metal-ion-dependent adhesion site in the Von Willebrand factor-A domain of α2δ subunits is key to trafficking voltage-gated Ca^{2+} channels. *Proc Natl Acad Sci USA* **102**: 11230–11235.

Casazza A, Fazzari P, Tamagnone L. 2007. Semaphorin signals in cell adhesion and cell migration: functional role and molecular mechanisms. *Adv Exp Med Biol* **600**: 90–108.

Castresana J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol* **17**: 540–552.

Cauldwell CB, Henkart P, Humphreys T. 1973. Physical properties of sponge aggregation factor. A unique proteoglycan complex. *Biochemistry* **12**: 3051–3055.

Chapman JA, Kirkness EF, Simakov O, Hampson SE, Mitros T, Weinmaier T, Rattei T, Balasubramanian PG, Borman J, Busam D, et al. 2010. The dynamic genome of *Hydra*. *Nature* **464**: 592–596.

Conaco C, Neveu P, Zhou H, Arcila ML, Degnan SM, Degnan BM, Kosik KS. 2012. Transcriptome profiling of the demosponge *Amphimedon queenslandica* reveals genome-wide events that accompany major life cycle transitions. *BMC Genomics* **13**: 209.

Conesa A, Götz S. 2008. Blast2GO: A comprehensive suite for functional analysis in plant genomics. *Int J Plant Genomics* **2008**: 619832.

Conrad J, Zahn RK, Kurelec B, Uhlenbruck G, Müller WEG. 1981. Aggregation of sponge cells: Immunological characterization of the species-specific *Geodia* aggregation factor. *J Supramol Struct Cell Biochem* **17**: 1–9.

Coombe D, Ey P. 1984. Self/non-self recognition in invertebrates. *Q Rev Biol* **59**: 231–255.

Crainie M, Belch AR, Mant MJ, Pilarski LM. 1999. Overexpression of the receptor for hyaluronan-mediated motility (RHAMM) characterizes the malignant clone in multiple myeloma: identification of three distinct RHAMM variants. *Blood* **93**: 1684–1696.

Crooks GE, Hon G, Chandonia J-M, Brenner SE. 2004. WebLogo: a sequence logo generator. *Genome Res* **14**: 1188–1190.

Csuros M, Rogozin IB, Koonin EV. 2011. A detailed history of intron-rich eukaryotic ancestors inferred from a global survey of 100 complete genomes. *PLoS Comput Biol* **7**: e1002150.

Curtis AS. 1979. Histocompatibility systems, recognition and cell positioning. *Dev Comp Immunol* **3**: 379–387.

Curtis AS, Van de Vyver G. 1971. The control of cell adhesion in a morphogenetic system. *J Embryol Exp Morphol* **26**: 295–312.

Curtis ASG, Kerr J, Knowlton N. 1982. Graft rejection in sponges - Genetic structure of accepting and rejecting populations. *Transplantation* **33**: 127–133.

Day AJ, Campbell RD, Reid K. 1989. The mosaic nature of the complement proteins. In *Progress in Immunology - Proceedings of the 7th International Congress Immunology Berlin 1989* (eds. F. Melchers, E.D. Albert, H. Von Boehmer, M.P. Dierich, L. Du Pasquier, K. Eichmann, D. Gemsa, O. Götze, J.R. Kalden, S. Kaufmann, et al.), pp. 209–212, Springer Berlin Heidelberg.

References

Day AJ, Sheehan JK. 2001. Hyaluronan: polysaccharide chaos to protein organisation. *Curr Opin Struct Biol* **11**: 617–622.

De Tomaso AW, Nyholm SV, Palmeri KJ, Ishizuka KJ, Ludington WB, Mitchel K, Weissman IL. 2005. Isolation and characterization of a protochordate histocompatibility locus. *Nature* **438**: 454–459.

Degnan B, Leys S, Larroux C. 2005. Sponge development and antiquity of animal pattern formation. *Integr Comp Biol* **45**: 335–341.

Degnan BM, Adamska M, Craigie A, Degnan SM, Fahey B, Gauthier M, Hooper JNA, Larroux C, Leys SP, Lovas E, et al. 2008a. The demosponge *Amphimedon queenslandica*: Reconstructing the ancestral metazoan genome and deciphering the origin of animal multicellularity. *Cold Spring Harb Protoc* **2008**: doi:10.1101–pdb.emo108.

Degnan BM, Vervoort M, Larroux C. 2009. Early evolution of metazoan transcription factors. *Curr Opin Genet Dev* **19**: 591–599.

Degnan SM, Craigie A, Degnan BM. 2008b. Genotyping Individual Amphimedon Embryos, Larvae, and Adults. *Cold Spring Harb Protoc* **2008**: pdb.prot5098–pdb.prot5098.

Degnan SM, Degnan BM. 2010. The initiation of metamorphosis as an ancient polyphenic trait and its role in metazoan life-cycle evolution. *Phil Trans R Soc B* **365**: 641–651.

Dehal P, Satou Y, Campbell RK, Chapman J, Degnan B, De Tomaso A, Davidson B, Di Gregorio A, Gelpke M, Goodstein DM, et al. 2002. The draft genome of *Ciona intestinalis*: insights into chordate and vertebrate origins. *Science* **298**: 2157–2167.

Destoumieux D, Bulet P, Loew D, Van Dorsselaer A, Rodriguez J, Bachère E. 1997. Penaeidins, a new family of antimicrobial peptides isolated from the shrimp *Penaeus vannamei* (Decapoda). *J Biol Chem* **272**: 28398–28406.

Dishaw LJ, Giacomelli S, Melillo D, Zucchetti I, Haire RN, Natale L, Russo NA, De Santis R, Litman GW, Pinto MR. 2011. A role for variable region-containing chitin-binding proteins (VCBPs) in host gut-bacteria interactions. *Proc Natl Acad Sci USA* **108**: 16747–16752.

Dong Y, Taylor HE, Dimopoulos G. 2006. AgDscam, a hypervariable immunoglobulin domain-containing receptor of the *Anopheles gambiae* innate immune system. *PLoS Biol* **4**: e229.

Dudoit S, Shaffer JP, Boldrick JC. 2003. Multiple hypothesis testing in microarray experiments. *Stat Sci* **18**: 71–103.

Dunham P, Anderson C, Rich AM, Weissmann G. 1983. Stimulus-response coupling in sponge cell aggregation: Evidence for calcium as an intracellular messenger. *Proc Natl Acad Sci USA* **80**: 4756–4760.

Eberhard M. 1975. The evolution of social behavior by kin selection. *Q Rev Biol* **50**: 1–33.

Eddy SR. 1998. Profile hidden Markov models. *Bioinformatics* **14**: 755–763.

Edelman GM. 1987. CAMs and Igs - Cell adhesion and the evolutionary origins of immunity. *Immunol Rev* **100**: 11–45.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32**: 1792–1797.

Edwards YJ, Perkins SJ. 1995. The protein fold of the von Willebrand factor type A domain is predicted to be similar to the open twisted β-sheet flanked by α-helices found in human ras-p21. *FEBS Lett* **358**: 283–286.

Eichinger L, Pachebat JA, Glöckner G, Rajandream MA, Sucgang R, Berriman M, Song J, Olsen R, Szafranski K, Xu Q, et al. 2005. The genome of the social amoeba *Dictyostelium discoideum*. *Nature* **435**: 43–57.

Ellison CE, Stajich JE, Jacobson DJ, Natvig DO, Lapidus A, Foster B, Aerts A, Riley R, Lindquist EA, Grigoriev IV, et al. 2011. Massive changes in genome architecture accompany the transition to self-fertility in the filamentous fungus *Neurospora tetrasperma*. *Genetics* **189**: 55–69.

Entwistle J, Hall CL, Turley EA. 1996. HA receptors: regulators of signalling to the cytoskeleton. *J Cell Biochem* **61**: 569–577.

Erwin DH, Laflamme M, Tweedt SM, Sperling EA, Pisani D, Peterson KJ. 2011. The Cambrian conundrum: early divergence and later ecological success in the early history of animals. *Science* **334**: 1091–1097.

Evanko SP, Wight TN. 1999. Intracellular localization of hyaluronan in proliferating cells. *J Histochem Cytochem* **47**: 1331–1341.

Fairclough SR, Chen Z, Kramer E, Zeng Q, Young S, Robertson HM, Begovic E, Richter DJ, Russ C, Westbrook MJ, et al. 2013. Premetazoan genome evolution and the regulation of cell differentiation in the choanoflagellate *Salpingoeca rosetta*. *Genome Biol* **14**: R15.

Fay P. 1992. Oxygen relations of nitrogen fixation in cyanobacteria. *Microbiol Rev* **56**: 340–373.

Fedorov A, Fedorova L, Starshenko V, Filatov V, Grigor'ev E. 1998. Influence of exon duplication on intron and exon phase distribution. *J Mol Evol* **46**: 263–271.

Fedorov A, Suboch G, Bujakov M, Fedorova L. 1992. Analysis of nonuniformity in intron phase distribution. *Nucleic Acids Res* **20**: 2553–2557.

References

Fernàndez-Busquets X, Burger MM. 1999. Cell adhesion and histocompatibility in sponges. *Microsc Res Tech* **44**: 204–218.

Fernàndez-Busquets X, Burger MM. 2003. Circular proteoglycans from sponges: first members of the spongican family. *Cell Mol Life Sci* **60**: 88–112.

Fernàndez-Busquets X, Burger MM. 1997. The main protein of the aggregation factor responsible for species-specific cell adhesion in the marine sponge *Microciona prolifera* is highly polymorphic. *J Biol Chem* **272**: 27839–27847.

Fernàndez-Busquets X, Gerosa D, Hess D, Burger MM. 1998. Accumulation in marine sponge grafts of the mRNA encoding the main proteins of the cell adhesion system. *J Biol Chem* **273**: 29545–29553.

Fernàndez-Busquets X, Kammerer RA, Burger MM. 1996. A 35-kDa protein is the basic unit of the core from the 2 x 10^4-kDa aggregation factor responsible for species-specific cell adhesion in the marine sponge *Microciona prolifera*. *J Biol Chem* **271**: 23558–23565.

Fernàndez-Busquets X, Kuhns WJ, Simpson TL, Ho M, Gerosa D, Grob M, Burger MM. 2002. Cell adhesion-related proteins as specific markers of sponge cell types involved in allogeneic recognition. *Dev Comp Immunol* **26**: 313–323.

Finn RD, Mistry J, Schuster-Böckler B, Griffiths-Jones S, Hollich V, Lassmann T, Moxon S, Marshall M, Khanna A, Durbin R, et al. 2006. Pfam: clans, web tools and services. *Nucleic Acids Res* **34**: D247–51.

Fortunato SAV, Adamski M, Ramos OM, Leininger S, Liu J, Ferrier DEK, Adamska M. 2015. Calcisponges have a ParaHox gene and dynamic expression of dispersed NK homeobox genes. *Nature* **514**: 620–623.

Fugmann SD, Messier C, Novack LA, Cameron RA, Rast JP. 2006. An ancient evolutionary origin of the *Rag1/2* gene locus. *Proc Natl Acad Sci USA* **103**: 3728–3733.

Galindo BE, Vacquier VD, Swanson WJ. 2003. Positive selection in the egg receptor for abalone sperm lysin. *Proc Natl Acad Sci USA* **100**: 4639–4643.

Galtsoff PS. 1925. Regeneration after dissociation (an experimental study on sponges) I. Behavior of dissociated cells of *Microciona prolifera* under normal and altered conditions. *J Exp Zool* **42**: 183–221.

Garcia-Manyes S, Bucior I, Ros R, Anselmetti D, Sanz F, Burger M, Fernàndez-Busquets X. 2006. Proteoglycan mechanics studied by single-molecule force spectroscopy of allotypic cell adhesion glycans. *J Biol Chem* **281**: 5992–5999.

Gauthier M, Degnan BM. 2008. Partitioning of genetically distinct cell populations in chimeric juveniles of the sponge *Amphimedon queenslandica*. *Dev Comp Immunol* **32**: 1270–1280.

Gauthier MEA. 2009. Developing a sense of self. University of Queensland, Brisbane.

Gavrilets S. 2010. Rapid transition towards the division of labor via evolution of developmental plasticity. *PLoS Comput Biol* **6**: e1000805.

Ge Y, Porse BT. 2013. The functional consequences of intron retention: Alternative splicing coupled to NMD as a regulator of gene expression. *BioEssays* **36**: 236–243.

Gerber A. 1998. Tad1p, a yeast tRNA-specific adenosine deaminase, is related to the mammalian pre-mRNA editing enzymes ADAR1 and ADAR2. *EMBO J* **17**: 4780–4789.

Geuze HJ, Slot JW, Strous GJAM, Peppard J, vonFigura K, Hasilik A, Schwartz AL. 1984. Intracellular receptor sorting during endocytosis: Comparative immunoelectron microscopy of multiple receptors in rat liver. *Cell* **37**: 195–204.

Ghosh J, Lun CM, Majeske AJ, Sacchi S, Schrankel CS, Smith LC. 2011. Invertebrate immune diversity. *Dev Comp Immunol* **35**: 959–974.

Glenn TC. 2011. Field guide to next-generation DNA sequencers. *Mol Ecol Resour* **11**: 759–769.

Gloria-Soria A, Moreno MA, Yund PO, Lakkis FG, Dellaporta SL, Buss LW. 2012. Evolutionary genetics of the hydroid allodeterminant *alr2*. *Mol Biol Evol* **29**: 3921–3932.

Goldsby HJ, Dornhaus A, Kerr B, Ofria C. 2012. Task-switching costs promote the evolution of division of labor and shifts in individuality. *Proc Natl Acad Sci USA* **109**: 13686–13691.

Gott JM, Emeson RB. 2000. Functions and mechanisms of RNA editing. *Annu Rev Genet* **34**: 499–531.

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, et al. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol* **29**: 644–652.

Graham GJ. 1995. Tandem genes and clustered genes. *J Theor Biol* **175**: 71–87.

Grice LF, Degnan BM. 2015a. How to Build an Allorecognition System: A Guide for Prospective Multicellular Organisms. In *Evolutionary Transitions to Multicellular Life* (eds. I. Ruiz-Trillo and A.M. Nedelcu), Springer, Dordrecht Heidelberg New York London.

Grice LF, Degnan BM. 2015b. The origin of the ADAR gene family and animal RNA editing. *BMC Evol Biol* **15**: 4.

Grimson A, Srivastava M, Fahey B, Woodcroft BJ, Chiang HR, King N, Degnan BM, Rokhsar DS, Bartel DP. 2008. Early origins and evolution of microRNAs and Piwi-interacting RNAs in animals. *Nature* **455**: 1193–1197.

Grosberg RK. 1988. The evolution of allorecognition specificity in clonal invertebrates. *Q Rev Biol* **63**: 377–412.

Grosberg RK, Strathmann RR. 2007. The evolution of multicellularity: A minor major transition? *Annu Rev Ecol Evol Syst* **38**: 621–654.

Guindon S, Dufayard J-F, Lefort V, Anisimova M, Hordijk W, Gascuel O. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol* **59**: 307–321.

Haas BJ. 2003. Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res* **31**: 5654–5666.

Halbleib JM, Nelson WJ. 2006. Cadherins in development: cell adhesion, sorting, and tissue morphogenesis. *Genes Dev* **20**: 3199–3214.

Hall C, Welch J, Kowbel DJ, Glass NL. 2010. Evolution and diversity of a fungal self/nonself recognition locus. *PLoS ONE* **5**: e14055.

Hammerschmidt M, Wedlich D. 2008. Regulated adhesion as a driving force of gastrulation movements. *Development* **135**: 3625–3641.

Harada Y. 2013. Allorecognition between compound ascidian colonies. *Zool Sci* **30**: 694–698.

Harpaz Y, Chothia C. 1994. Many of the immunoglobulin superfamily domains in cell adhesion molecules and surface receptors belong to a new structural set which is close to that containing variable domains. *J Mol Biol* **238**: 528–539.

Hashimshony T, Wagner F, Sher N, Yanai I. 2012. CEL-Seq: single-cell RNA-Seq by multiplexed linear amplification. *Cell Rep* **2**: 666–673.

He J, Dai X, Zhao X. 2007. PLAN: a web platform for automating high-throughput BLAST searches and for managing and mining results. *BMC Bioinformatics* **8**: 53.

Hellgren O, Ekblom R. 2010. Evolution of a cluster of innate immune genes (*β-defensins*) along the ancestral lines of chicken and zebra finch. *Immunome Res* **6**: 3.

Hemmrich G, Bosch TCG. 2008. Compagen, a comparative genomics platform for early branching metazoan animals, reveals early origins of genes regulating stem-cell differentiation. *BioEssays* **30**: 1010–1018.

Henkart P, Humphreys S, Humphreys T. 1973. Physical properties of sponge aggregation factor. *Biochemistry* **12**: 3045–3050.

Hibino T, Loza-Coll M, Messier C, Majeske AJ, Cohen AH, Terwilliger DP, Buckley KM, Brockton V, Nair SV, Berney K, et al. 2006. The immune gene repertoire encoded in the purple sea urchin genome. *Dev Biol* **300**: 349–365.

Hildemann WH. 1979. Immunocompetence and allogeneic polymorphism among invertebrates. *Transplantation* **27**: 1–3.

Hildemann WH, Bigger CH, Johnston IS, Jokiel PL. 1980. Characteristics of transplantation immunity in the sponge, *Callyspongia diffusa*. *Transplantation* **30**: 362–367.

Hildemann WH, Johnson IS, Jokiel PL. 1979. Immunocompetence in the lowest metazoan phylum - Transplantation immunity in sponges. *Science* **204**: 420–422.

Hildemann WH, Linthicum DS. 1981. Transplantation immunity in the Palaun sponge, *Xestospongia exigua*. *Transplantation* **32**: 77–80.

Hilge M, Aelen J. 2006. Ca^{2+} regulation in the Na^{+}/Ca^{2+} exchanger involves two markedly different Ca^{2+} sensors. *Mol Cell* **22**: 15–25.

Hill E, Broadbent ID, Chothia C, Pettitt J. 2001. Cadherin superfamily proteins in *Caenorhabditis elegans* and *Drosophila melanogaster*. *J Mol Biol* **305**: 1011–1024.

Hohenester E, Tisi D, Talts JF, Timpl R. 1999. The crystal structure of a laminin G-like module reveals the molecular basis of alpha-dystroglycan binding to laminins, perlecan, and agrin. *Mol Cell* **4**: 783–792.

Hooper JNA, van Soest RWM. 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'. *Zootaza* **1314**: 31–39.

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

Humbert-David N, Garrone R. 1993. A six-armed, tenascin-like protein extracted from the Porifera *Oscarella tuberculata* (Homosclerophorida). *Eur J Biochem* **216**: 255–260.

References

Huminiecki L, Goldovsky L, Freilich S, Moustakas A, Ouzounis C, Heldin C-H. 2009. Emergence, development and diversification of the TGF-β signalling pathway within the animal kingdom. *BMC Evol Biol* **9**: 28.

Humphreys S, Humphreys T, Sano J. 1977. Organization and polysaccharides of sponge aggregation factor. *J Supramol Struct* **7**: 339–351.

Humphreys T. 1970. Biochemical analysis of sponge cell aggregation. *Symp Zool Soc Lond* **25**: 325–334.

Humphreys T. 1963. Chemical dissolution and *in vitro* reconstruction of sponge cell adhesions I. Isolation and functional demonstration of the components involved. *Dev Biol* **8**: 27–47.

Humphreys T. 1994. Rapid allogeneic recognition in the marine sponge *Microciona prolifera:* Implications for evolution of immune recognition. *Ann N Y Acad Sci* **712**: 342–345.

Humphreys T, Humphreys S, Moscona AA. 1960. A procedure for obtaining completely dissociated sponge cells. *Biol Bull* **119**: 294–294.

Humphreys T, Reinherz EL. 1994. Invertebrate immune recognition, natural immunity and the evolution of positive selection. *Immunol Today* **15**: 316–320.

Humphreys T, Yonemoto W, Humphreys S, Anderson D. 1975. Purification of *Microciona prolifera* aggregation factor. *Biol Bull* **149**: 430.

Hurst LD, Lercher MJ. 2005. Unusual linkage patterns of ligands and their cognate receptors indicate a novel reason for non-random gene order in the human genome. *BMC Evol Biol* **5**.

Ikegami T, Okada T, Hashimoto M, Seino S, Watanabe T, Shirakawa M. 2000. Solution structure of the chitin-binding domain of *Bacillus circulans* WL-12 chitinase A1. *J Biol Chem* **275**: 13654–13661.

Ilan M, Loya Y. 1990. Ontogenetic variation in sponge histocompatibility responses. *Biol Bull* **179**: 279–286.

Ispolatov I, Ackermann M, Doebeli M. 2012. Division of labour and the evolution of multicellularity. *Phil Trans R Soc B* **279**: 1768–1776.

Jantsch MF, Öhman M. 2008. RNA editing by adenosine deaminases that act on RNA (ADARs). In *RNA Editing* (ed. H.U. Göringer), pp. 51–84, Springer Verlag, Berlin.

Jarchow J, Burger MM. 1998. Species-specific association of the cell-aggregation molecule mediates recognition in marine sponges. *Cell Adhes Commun* **6**: 405–414.

Jarchow J, Fritz J, Anselmetti D, Calabro A, Hascall VC, Gerosa D, Burger MM, Fernàndez-Busquets X. 2000. Supramolecular structure of a new family of circular proteoglycans mediating cell adhesion in sponges. *J Struct Biol* **132**: 95–105.

Jensen JD, Wong A, Aquadro CF. 2007. Approaches for identifying targets of positive selection. *Trends Genet* **23**: 568–577.

Jin Y, Zhang W, Li Q. 2009. Origins and evolution of ADAR-mediated RNA editing. *IUBMB Life* **61**: 572–578.

Jokiel PL, Hildemann WH, Bigger CH. 1982. Frequency of intercolony graft acceptance or rejection as a measure of population structure in the sponge *Callyspongia diffusa*. *Mar Biol* **71**: 135–139.

Jumblatt JE, Schlup V, Burger MM. 1980. Cell-cell recognition: specific binding of *Microciona* sponge aggregation factor to homotypic cells and the role of calcium ions. *Biochemistry* **19**: 1038–1042.

Kastan MB, Bartek J. 2004. Cell-cycle checkpoints and cancer. *Nature* **432**: 316–323.

Kaye H, Ortiz T. 1981. Strain specificity in a tropical marine sponge. *Mar Biol* **63**: 165–173.

Käll L, Krogh A, Sonnhammer ELL. 2004. A combined transmembrane topology and signal peptide prediction method. *J Mol Biol* **338**: 1027–1036.

Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, Buxton S, Cooper A, Markowitz S, Duran C, et al. 2012. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* **28**: 1647–1649.

Keegan LP, McGurk L, Palavicini JP, Brindle J, Paro S, Li X, Rosenthal JJC, O'Connell MA. 2011. Functional conservation in human and *Drosophila* of metazoan ADAR2 involved in RNA editing: Loss of ADAR1 in insects. *Nucleic Acids Res* **39**: 7249–7262.

Keppen OI, Baulina OI, Lysenko AM, Kondrateva EN. 1993. A new green bacterium belonging to the Chloroflexaceae family. *Mikrobiologiia* **62**: 267–275.

Kersey PJ, Allen JE, Christensen M, Davis P, Falin LJ, Grabmueller C, Hughes DST, Humphrey J, Kerhornou A, Khobova J, et al. 2014. Ensembl Genomes 2013: scaling up access to genome-wide data. *Nucleic Acids Res* **42**: D546–52.

Kidd T, Brose K, Mitchell KJ, Fetter RD, Tessier-Lavigne M, Goodman CS, Tear G. 1998. Roundabout controls axon crossing of the CNS midline and defines a novel subfamily of evolutionarily conserved guidance receptors. *Cell* **92**: 205–215.

Kim E, Magen A, Ast G. 2006. Different levels of alternative splicing among eukaryotes. *Nucleic Acids Res* **35**: 125–131.

Kimura M. 1968. Evolutionary rate at the molecular level. *Nature* **217**: 624–626.

Kimura M. 1977. Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution. *Nature* **267**: 275–276.

King N. 2004. The unicellular ancestry of animal development. *Dev Cell* **7**: 313–325.

King N, Westbrook MJ, Young SL, Kuo A, Abedin M, Chapman J, Fairclough S, Hellsten U, Isogai Y, Letunic I, et al. 2008. The genome of the choanoflagellate *Monosiga brevicollis* and the origin of metazoans. *Nature* **451**: 783–788.

Kinsella RJ, Kahari A, Haider S, Zamora J, Proctor G, Spudich G, Almeida-King J, Staines D, Derwent P, Kerhornou A, et al. 2011. Ensembl BioMarts: a hub for data retrieval across taxonomic space. *Database* **2011**: bar030.

Kirk DL. 2005. A twelve-step program for evolving multicellularity and a division of labor. *BioEssays* **27**: 299–310.

Kolodkin AL, Matthes DJ, Goodman CS. 1993. The semaphorin genes encode a family of transmembrane and secreted growth cone guidance molecules. *Cell* **75**: 1389–1399.

Koressaar T, Remm M. 2007. Enhancements and modifications of primer design program Primer3. *Bioinformatics* **23**: 1289–1291.

Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA. 2009. Circos: An information aesthetic for comparative genomics. *Genome Res* **19**: 1639–1645.

Kuhns WJ, Bramson S, Simpson TL, Burkart W, Jumblatt J, Burger M. 1980. Fluorescent antibody localization of *Microciona prolifera* aggregation factor and its baseplate component. *Eur J Cell Biol* **23**: 73–79.

Kuhns WJ, Burger MM, Turley E. 1999. Hyaluronic acid: A component of the aggregation factor secreted by the marine sponge, *Microciona prolifera*. *Biol Bull* **197**: 277–279.

Kuhns WJ, Fernàndez-Busquets X, Burger MM, Ho M, Turley E. 1998. Hyaluronic acid-receptor binding demonstrated by synthetic adhesive proteoglycan peptide constructs and by cell receptors on the marine sponge *Microciona prolifera*. *Biol Bull* **195**: 216–218.

Kuhns WJ, Ho M, Burger MM, Turley E. 1997. Binding of hyaluronic acid to cellular receptors of the marine sponge *Microciona prolifera*. *Biol Bull* **193**: 243–244.

Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10**: R25.

Lapébie P, Gazave E, Ereskovsky A, Derelle R, Bézac C, Renard E, Houliston E, Borchiellini C. 2008. WNT/beta-catenin signalling and epithelial patterning in the homoscleromorph sponge *Oscarella*. *PLoS ONE* **4**: e5823–e5823.

Leahy DJ, Hendrickson WA, Aukhil I, Erickson HP. 1992. Structure of a fibronectin type III domain from tenascin phased by MAD analysis of the selenomethionyl protein. *Science* **258**: 987–991.

Leckband D, Prakasam A. 2006. Mechanism and dynamics of cadherin adhesion. *Annu Rev Biomed Eng* **8**: 259–287.

Leshko-Lindsay LA, Corces VG. 1997. The role of selectins in *Drosophila* eye and bristle development. *Development* **124**: 169–180.

Leys SP, Degnan BM. 2001. Cytological basis of photoresponsive behavior in a sponge larva. *Biol Bull* **201**: 323–338.

Leys SP, Degnan BM. 2002. Embryogenesis and metamorphosis in a haplosclerid demosponge: gastrulation and transdifferentiation of larval ciliated cells to choanocytes. *Invertebr Biol* **121**: 171–189.

Leys SP, Larroux C, Gauthier M, Adamska M, Fahey B, Richards GS, Degnan SM, Degnan BM. 2008. Isolation of *Amphimedon* developmental material. *Cold Spring Harb Protoc* **2008**: pdb.prot5095.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**: 2078–2079.

Li W, Godzik A. 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**: 1658–1659.

Losson R, Lacroute F. 1979. Interference of nonsense mutations with eukaryotic messenger RNA stability. *Proc Natl Acad Sci USA* **76**: 5134–5137.

Lykke-Andersen S, Piñol-Roma S, Kjems J. 2007. Alternative splicing of the ADAR1 transcript in a region that functions either as a 5'-UTR or an ORF. *RNA* **13**: 1732–1744.

Lynn BD, Turley EA, Nagy JI. 2001. Subcellular distribution, calmodulin interaction, and mitochondrial association of the hyaluronan-binding protein RHAMM in rat brain. *J Neurosci Res* **65**: 6–16.

Lyon JD, Vacquier VD. 1999. Interspecies chimeric sperm lysins identify regions mediating species-specific recognition of the abalone egg vitelline envelope. *Dev Biol* **214**: 151–159.

References

Maere S, Heymans K, Kuiper M. 2005. BiNGO: a Cytoscape plugin to assess overrepresentation of Gene Ontology categories in Biological Networks. *Bioinformatics* **21**: 3448–3449.

Maldonado M. 1998. Do chimeric sponges have improved chances of survival? *Mar Ecol Prog Ser* **164**: 301–306.

Massagué J. 1998. TGF-beta signal transduction. *Annu Rev Biochem* **67**: 753–791.

Matsunaga T, Mori N. 1987. The origin of the immune system: The possibility that immunoglobulin superfamily molecules and cell adhesion molecules of chicken and slime mold are all related. *Scand J Immunol* **25**: 485–495.

Matsuoka S, Nicoll D, He Z, Philipson KD. 1997. Regulation of the cardiac Na^{+}-Ca^{2+} exchanger by the endogenous XIP region. *J Gen Physiol* **109**: 273–286.

McCarthy DJ, Chen Y, Smyth GK. 2012. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Res* **40**: 4288–4297.

McEwen DG, Peifer M. 2000. Wnt signaling: Moving in a new direction. *Curr Biol* **10**: R562–R564.

McGhee K. 2006. The importance of life-history stage and individual variation in the allorecognition system of a marine sponge. *J Exp Mar Biol Ecol* **333**: 241–250.

McGuire AM, Pearson MD, Neafsey DE, Galagan JE. 2008. Cross-kingdom patterns of alternative splicing and splice recognition. *Genome Biol* **9**: R50.

McKitrick TR, De Tomaso AW. 2010. Molecular mechanisms of allorecognition in a basal chordate. *Sem Immunol* **22**: 34–38.

McKitrick TR, Muscat CC, Pierce JD, Bhattacharya D, De Tomaso AW. 2011. Allorecognition in a basal chordate consists of independent activating and inhibitory pathways. *Immunity* **34**: 616–626.

McNamee HP, Ingber DE, Schwartz MA. 1993. Adhesion to fibronectin stimulates inositol lipid synthesis and enhances PDGF-induced inositol lipid breakdown. *J Cell Biol* **121**: 673–678.

McNeill H. 2000. Sticking together and sorting things out: adhesion as a force in development. *Nat Rev Genet* **1**: 100–108.

Melcher T, Maas S, Herb A, Sprengel R, Higuchi M, Seeburg PH. 1996. RED2, a brain-specific member of the RNA-specific adenosine deaminase family. *J Biol Chem* **271**: 31795–31798.

Metz EC, Robles-Sikisaka R, Vacquier VD. 1998. Nonsynonymous substitution in abalone sperm fertilization genes exceeds substitution in introns and mitochondrial DNA. *Proc Natl Acad Sci USA* **95**: 10676–10681.

Meylan E, Tschopp J, Karin M. 2006. Intracellular pattern recognition receptors in the host response. *Nature* **442**: 39–44.

Micali CO, Smith ML. 2006. A nonself recognition gene complex in *Neurospora crassa*. *Genetics* **173**: 1991–2004.

Michod RE. 2007. Evolution of individuality during the transition from unicellular to multicellular life. *Proc Natl Acad Sci USA* **104**: 8613–8618.

Middha S, Wang X. 2008. Evolution and potential function of fibrinogen-like domains across twelve *Drosophila* species. *BMC Genomics* **9**.

Miller CA, Buckley KM, Easley RL, Smith LC. 2010. An *Sp185/333* gene cluster from the purple sea urchin and putative microsatellite-mediated gene diversification. *BMC Genomics* **11**: 575.

Miller JR, Hocking AM, Brown JD, Moon RT. 1999. Mechanism and function of signal transduction by the Wnt/β-catenin and Wnt/Ca^{2+} pathways. *Oncogene* **18**: 7860–7872.

Misevic G, Finne J. 1987. Involvement of carbohydrates as multiple low affinity interaction sites in the self-association of the aggregation factor from the marine sponge *Microciona prolifera*. *J Biol Chem* **262**: 5870–5877.

Misevic GN. 1999. Molecular self-recognition and adhesion via proteoglycan to proteoglycan interactions as a pathway to multicellularity: Atomic force microscopy and color coded bead measurements in sponges. *Microsc Res Tech* **44**: 304–309.

Misevic GN, Burger MM. 1993. Carbohydrate-carbohydrate interactions of a novel acidic glycan can mediate sponge cell adhesion. *J Biol Chem* **268**: 4922–4929.

Misevic GN, Burger MM. 1990a. Involvement of a highly polyvalent glycan in the cell-binding of the aggregation factor from the marine sponge *Microciona prolifera*. *J Cell Biochem* **43**: 307–314.

Misevic GN, Burger MM. 1990b. The species-specific cell-binding site of the aggregation factor from the sponge *Microciona prolifera* is a highly repetitive novel glycan containing glucuronic acid, fucose, and mannose. *J Biol Chem* **265**: 20577–20584.

Moore AD, Held A, Terrapon N, Weiner J, Bornberg-Bauer E. 2014. DoMosaics: software for domain arrangement visualization and domain-centric analysis of proteins. *Bioinformatics* **30**: 282–283.

Moroz LL, Kocot KM, Citarella MR, Dosung S, Norekian TP, Povolotskaya IS, Grigorenko AP, Dailey C, Berezikov E, Buckley KM, et al. 2014. The ctenophore genome and the evolutionary origins of neural systems. *Nature* **510**: 109–114.

References

Moscona AA. 1968. Cell aggregation: Properties of specific cell-ligands and their role in the formation of multicellular systems. *Dev Biol* **18**: 250–277.

Moscona AA. 1963. Studies on cell aggregation: Demonstration of materials with selective cell-binding activity. *Proc Natl Acad Sci USA* **49**: 742–747.

Müller WEG. 1982. Cell membranes in sponges eds. G.H. Bourne and J.F. Danielli. *Int Rev Cytol* **77**: 129–181.

Müller WEG, Blumbach B, Müller IM. 1999a. Evolution of the innate and adaptive immune systems - Relationships between potential immune molecules in the lowest metazoan phylum (Porifera) and those in vertebrates. *Transplantation* **68**: 1215–1227.

Müller WEG, Gamulin V, Rinkevich B, Spreitzer I, Weinblum D, Schröder HC. 1994. Ubiquitin and ubiquitination in cells from the marine sponge *Geodia cydonium*. *Biol Chem Hoppe-Seyler* **375**: 53–60.

Müller WEG, Koziol C, Müller IM, Wiens M. 1999b. Towards an understanding of the molecular basis of immune responses in sponges: the marine demosponge *Geodia cydonium* as a model. *Microsc Res Tech* **44**: 219–236.

Müller WEG, Krasko A, Skorokhod A, Bünz C, Grebenjuk VA, Steffen R, Batel R, Schröder HC. 2002. Histocompatibility reaction in tissue and cells of the marine sponge S*uberites domuncula* in vitro and in vivo: central role of the allograft inflammatory factor 1. *Immunogenetics* **54**: 48–58.

Müller WEG, Müller I, Pondeljak V, Kurelec B, Zahn R. 1978a. Species-specific aggregation factor in sponges: Isolation, purification and characterization of the aggregation factor from *Suberites domuncula*. *Differentiation* **10**: 45–53.

Müller WEG, Müller I, Zahn R. 1976a. Species-specific aggregation factor in sponges V. Influence on programmed syntheses. *Biochim Biophys Acta* **418**: 217–225.

Müller WEG, Müller I, Zahn R, Kurelec B. 1976b. Species-specific aggregation factor in sponges VI. Aggregation receptor from the cell surface. *J Cell Sci* **21**: 227–241.

Müller WEG, Müller I, Zahn RK. 1974. Two different aggregation principles in reaggregation process of dissociated sponge cells (*Geodia cydonium*). *Experientia* **30**: 899–902.

Müller WEG, Rottmann M, Diehl-Seifert B, Kurelec B, Uhlenbruck G, Schröder HC. 1987. Role of the aggregation factor in the regulation of phosphoinositide metabolism in sponges. Possible consequences on calcium efflux and on mitogenesis. *J Biol Chem* **262**: 9850–9858.

Müller WEG, Steffen R, Lorenz B, Batel R, Kruse M, Krasko A, Müller IM, Schröder HC. 2001. Suppression of allograft rejection in the sponge *Suberites domuncula* by FK506 and expression of genes encoding FK506-binding proteins in allografts. *J Exp Biol* **204**: 2197–2207.

Müller WEG, Ugarković D, Gamulin V, Weiler BE, Schröder HC. 1990. Intracellular signal transduction pathways in sponges. *Electron Microsc Rev* **3**: 97–114.

Müller WEG, Zahn R, Kurelec B, Uhlenbruck G, Vaith P, Müller I. 1978b. Aggregation of sponge cells, XVIII. Glycosyltransferase associated with the aggregation factor. *Hoppe Seylers Z Physiol Chem* **359**: 529–537.

Müller WEG, Zahn RK. 1973. Purification and characterization of a species-specific aggregation factor in sponges. *Exp Cell Res* **80**: 95–104.

Nair SV. 2005. Macroarray analysis of coelomocyte gene expression in response to LPS in the sea urchin. Identification of unexpected immune diversity in an invertebrate. *Physiol Genomics* **22**: 33–47.

Nakanishi N, Sogabe S, Degnan BM. 2014. Evolutionary origin of gastrulation: insights from sponge development. *BMC Biol* **12**: 1–9.

Narasimha M, Brown NH. 2006. Integrins and associated proteins in Drosophila development. In Integrins and Development (ed. E. Danen), Landes Bioscience.

Nasrallah J. 2005. Recognition and rejection of self in plant self-incompatibility: comparisons to animal histocompatibility. *Trends Immunol* **26**: 412–418.

Nayal A, Webb DJ, Horwitz AF. 2004. Talin: an emerging focal point of adhesion dynamics. *Curr Opin Cell Biol* **16**: 94–98.

Nei M, Rooney AP. 2005. Concerted and birth-and-death evolution of multigene families. *Annu Rev Genet* **39**: 121–152.

Neigel JE, Avise JC. 1983. Histocompatibility bioassays of population structure in marine sponges: Clonal structure in *Verongia longissima* and *Lotrochota birotulata*. *J Hered* **74**: 134–140.

Neigel JE, Avise JC. 1985. The precision of histocompatibility response in clonal recognition in tropical marine sponges. *Evolution* **39**: 724–732.

Neigel JE, Schmahl GP. 1984. Phenotypic variation within histocompatibility-defined clones of marine sponges. *Science* **224**: 413–415.

Ng SK, Weissbach R, Ronson GE, Scadden ADJ. 2013. Proteins that contain a functional Z-DNA-binding domain localize to cytoplasmic stress granules. *Nucleic Acids Res* **41**: 9786–9799.

Ng YK, Wu W, Zhang L. 2009. Positive correlation between gene coexpression and positional clustering in the zebrafish genome. *BMC Genomics* **10**.

Nichols SA, Roberts BW, Richter DJ, Fairclough SR, King N. 2012. Origin of metazoan cadherin diversity and the antiquity of the classical cadherin/β-catenin complex. *Proc Natl Acad Sci USA* **109**: 13046–13051.

Nicoll DA, Sawaya MR, Kwon S, Cascio D, Philipson KD, Abramson J. 2006. The crystal structure of the primary Ca^{2+} sensor of the Na^{+}/Ca^{2+} exchanger reveals a novel Ca^{2+} binding motif. *J Biol Chem* **281**: 21577–21581.

Nicotra ML, Powell AE, Rosengarten RD, Moreno M, Grimwood J, Lakkis FG, Dellaporta SL, Buss LW. 2009. A hypervariable invertebrate allodeterminant. *Curr Biol* **19**: 583–589.

Nishikura K. 2010. Functions and regulation of RNA editing by ADAR deaminases. *Annu Rev Biochem* **79**: 321–349.

Norman DG, Barlow PN, Baron M, Day AJ, Sim RB, Campbell ID. 1991. Three-dimensional structure of a complement control protein module in solution. *J Mol Biol* **219**: 717–725.

Nydam ML, De Tomaso AW. 2012. The *fester* locus in *Botryllus schlosseri* experiences selection. *BMC Evol Biol* **12**: 249.

Nydam ML, Hoang TA, Shanley KM, De Tomaso AW. 2013a. Molecular evolution of a polymorphic HSP40-like protein encoded in the histocompatibility locus of an invertebrate chordate. *Dev Comp Immunol* **41**: 128–136.

Nydam ML, Netuschil N, Sanders E, Langenbacher A, Lewis DD, Taketa DA, Marimuthu A, Gracey AY, De Tomaso AW. 2013b. The candidate histocompatibility locus of a basal chordate encodes two highly polymorphic proteins. *PLoS ONE* **8**: e65980.

Nydam ML, Taylor AA, De Tomaso AW. 2012. Evidence for selection on a histocompatibility locus. *Evolution* **67**: 487–500.

Nyholm SV, Passegue E, Ludington WB, Voskoboynik A, Mitchel K, Weissman IL, De Tomaso AW. 2006. *fester*, a candidate allorecognition receptor from a primitive chordate. *Immunity* **25**: 163–173.

Oh H-M, Kang I, Vergin KL, Lee K, Giovannoni SJ, Cho J-C. 2011. Genome sequence of *Oceanicaulis* sp. strain HTCC2633, isolated from the Western Sargasso Sea. *J Bacteriol* **193**: 317–318.

Oh H-M, Kang I, Yang S-J, Jang Y, Vergin KL, Giovannoni SJ, Cho J-C. 2010. Complete genome sequence of strain HTCC2170, a novel member of the genus *Maribacter* in the family *Flavobacteriaceae*. *J Bacteriol* **193**: 303–304.

Oka H, Watanabe H. 1957. Colony specificity in compound ascidians as tested by fusion experiments. *Proc Japan Acad* **33**: 657–659.

Oren M, Paz G, Douek J, Rosner A, Amar K-O, Rinkevich B. 2013. Marine invertebrates cross phyla comparisons reveal highly conserved immune machinery. *Immunobiology* **218**: 484–495.

Oren M, Paz G, Douek J, Rosner A, Fishelson Z, Goulet TL, Henckel K, Rinkevich B. 2010. 'Rejected' vs. "rejecting" transcriptomes in allogeneic challenged colonial urochordates. *Mol Immunol* **47**: 2083–2093.

Padhi A, Verghese B, Otta SK, Varghese B, Ramu K. 2007. Adaptive evolution after duplication of penaeidin antimicrobial peptides. *Fish Shellfish Immunol* **23**: 553–566.

Pancer Z, Amemiya CT, Ehrhardt GRA, Ceitlin J, Gartland GL, Cooper MD. 2004. Somatic diversification of variable lymphocyte receptors in the agnathan sea lamprey. *Nature* **430**: 174–180.

Pancer Z, Munkner J, Müller I, Müller WEG. 1997. A novel member of an ancient superfamily: Sponge (*Geodia cydonium*, Porifera) putative protein that features scavenger receptor cysteine-rich repeats. *Gene* **193**: 211–218.

Patthy L. 1996. Exon shuffling and other ways of module exchange. **15**: 301–10; discussion 311–2.

Patthy L. 1987. Intron-dependent evolution: preferred types of exons and introns. *FEBS Lett* **214**: 1–7.

Pál C, Hurst LD. 2003. Evidence for co-evolution of gene order and recombination rate. *Nat Genet* **33**: 392–395.

Pérez-Porro AR, Navarro-Gómez D, Uriz MJ, 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. *Mol Ecol Resour* **13**: 494–509.

Pfeifer K, Frank W, Schröder HC, Gamulin V, Rinkevich B, Batel R, Müller IM, Müller WEG. 1993. Cloning of the polyubiquitin cDNA from the marine sponge *Geodia cydonium* and its preferential expression during reaggregation of cells. *J Cell Sci* **106**: 545–554.

Popescu O, Misevic GN. 1997. Self-recognition by proteoglycans. *Nature* **386**: 231–232.

Price LS, Leng J, Schwartz MA, Bokoch GM. 1998. Activation of Rac and Cdc42 by integrins mediates cell spreading. *Mol Biol Cell* **9**: 1863–1871.

Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, Boursnell C, Pang N, Forslund K, Ceric G, Clements J, et al. 2011. The Pfam protein families database. *Nucleic Acids Res* **40**: D290–D301.

Putnam NH, Butts T, Ferrier DEK, Furlong RF, Hellsten U, Kawashima T, Robinson-Rechavi M, Shoguchi E, Terry A, Yu J-K, et al. 2008. The amphioxus genome and the evolution of the chordate karyotype. *Nature* **453**: 1064–1071.

Putnam NH, Srivastava M, Hellsten U, Dirks B, Chapman J, Salamov A, Terry A, Shapiro H, Lindquist E, Kapitonov VV, et al. 2007. Sea anemone genome reveals ancestral eumetazoan gene repertoire and genomic organization. *Science* **317**: 86–94.

Ratcliff WC, Denison RF, Borrello M, Travisano M. 2012. Experimental evolution of multicellularity. *Proc Natl Acad Sci USA* **109**: 1595–1600.

Rice P, Longden I, Bleasby A. 2000. EMBOSS: The European Molecular Biology Open Software Suite. *Trends Genet* **16**: 276–277.

Richards GS. 2010. The origins of cell communication in the animal kingdom: Notch signalling during embryogenesis and metamorphosis of the demosponge. University of Queensland.

Richards GS, Degnan BM. 2009. The dawn of developmental signaling in the Metazoa. *Cold Spring Harb Symp Quant Biol* **74**: 81–90.

Richardson JS. 1981. The anatomy and taxonomy of protein structure. *Adv Protein Chem* **34**: 167–339.

Riesgo A, Andrade SC, Sharma PP, Novo M, Pérez-Porro AR, Vahtera V, González VL, Kawauchi GY, Giribet G. 2012. Comparative description of ten transcriptomes of newly sequenced invertebrates and efficiency estimation of genomic sampling in non-model taxa. *Front Zool* **9**: 33.

Riesgo A, Farrar N, Windsor PJ, Giribet G, Leys SP. 2014. The analysis of eight transcriptomes from all poriferan classes reveals surprising genetic complexity in sponges. *Mol Biol Evol* **31**: 1102–1120.

Rinkevich B, Porat R, Goren M. 1995. Allorecognition elements on a urochordate histocompatibility locus indicate unprecedented extensive polymorphism. *Proc R Soc Lond B* **259**: 319–324.

Rinkevich B, Weissman IL. 1987. A long-term study on fused subclones in the ascidian *Botryllus schlosseri*: the resorption phenomenon (Protochordata: Tunicata). *J Zool* **213**: 717–733.

Rinkevich B, Weissman IL, De Tomaso AW. 1998. Transplantation of Fu/HC-incompatible zooids in *Botryllus schlosseri* results in chimerism. *Biol Bull* **195**: 98–106.

Roberts B, Davidson B, MacMaster G, Lockhart V, Ma E, Wallace SS, Swalla BJ. 2007. A complement response may activate metamorphosis in the ascidian *Boltenia villosa*. *Dev Genes Evol* **217**: 449–458.

Robinson MD, McCarthy DJ, Smyth GK. 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**: 139–140.

Robinson MD, Smyth GK. 2007a. Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics* **23**: 2881–2887.

Robinson MD, Smyth GK. 2007b. Small-sample estimation of negative binomial dispersion, with applications to SAGE data. *Biostatistics* **9**: 321–332.

Rosa SFP, Powell AE, Rosengarten RD, Nicotra ML, Moreno MA, Grimwood J, Lakkis FG, Dellaporta SL, Buss LW. 2010. *Hydractinia* allodeterminant *alr1* resides in an immunoglobulin superfamily-like gene complex. *Curr Biol* **20**: 1122–1127.

Rossetti V, Schirrmeister BE, Bernasconi MV, Bagheri HC. 2010. The evolutionary path to terminal differentiation and division of labor in cyanobacteria. *J Theor Biol* **262**: 23–34.

Rothenberg B. 1978. The self recognition concept: An active function for the molecules of the major histocompatibility complex based on complementary interaction of protein and carbohydrate. *Dev Comp Immunol* **2**: 23–37.

Rottmann M, Schröder HC, Gramzow M. 1987. Specific phosphorylation of proteins in pore complex-laminae from the sponge *Geodia cydonium* by the homologous aggregation factor and phorbol ester. Role of protein kinase C in the phosphorylation of DNA topoisomerase II. *EMBO J* **6**: 3939–3944.

Rueter SM, Dawson TR, Emeson RB. 1999. Regulation of alternative splicing by RNA editing. *Nature* **399**: 75–80.

Ryan JF, Pang K, Schnitzler CE, Nguyen A-D, Moreland RT, Simmons DK, Koch BJ, Francis WR, Havlak P, NISC Comparative Sequencing Program, et al. 2013. The genome of the ctenophore *Mnemiopsis leidyi* and its implications for cell type evolution. *Science* **342**: 1242592.

Saito-Diaz K, Chen TW, Wang X, Thorne CA, Wallace HA, Page-McCaw A, Lee E. 2013. The way Wnt works: Components and mechanism. *Growth Factors* **31**: 1–31.

Samuel CE. 2011. Adenosine deaminases acting on RNA (ADARs) are both antiviral and proviral dependent on the virus. *Virology* **411**: 180–193.

Scadden ADJ, Smith CW. 2001. RNAi is antagonized by A→I hyper-editing. *EMBO Rep* **2**: 1107–1111.

Schäfer H, McDonald IR, Nightingale PD, Murrell JC. 2005. Evidence for the presence of a CmuA methyltransferase pathway in novel marine methyl halide-oxidizing bacteria. *Environ Microbiol* **7**: 839–852.

Schlesner H, Rensmann C, Tindall BJ, Gade D, Rabus R, Pfeiffer S, Hirsch P. 2004. Taxonomic heterogeneity within the *Planctomycetales* as derived by DNA-DNA hybridization, description of *Rhodopirellula baltica* gen. nov., sp. nov., transfer of *Pirellula marina* to the genus *Blastopirellula* gen. nov. as *Blastopirellula marina comb*. nov. and emended description of the genus *Pirellula*. *Int J Syst Evol Microbiol* **54**: 1567–1580.

Schmucker D, Clemens JC, Shu H, Worby CA, Xiao J, Muda M, Dixon JE, Zipursky SL. 2000. *Drosophila* Dscam is an axon guidance receptor exhibiting extraordinary molecular diversity. *Cell* **101**: 671–684.

Schröder HC, Kuchino Y, Gramzow M, Kurelec B, Friese U, Uhlenbruck G, Müller WEG. 1988. Induction of *ras* gene expression by homologous aggregation factor in cells from the sponge *Geodia cydonium*. *J Biol Chem* **263**: 16334–16340.

Schütze J, Krasko A, Diehl-Seifert B, Müller WEG. 2001. Cloning and expression of the putative aggregation factor from the marine sponge *Geodia cydonium*. *J Cell Sci* **114**: 3189–3198.

Schwarz EM, Benzer S. 1997. *Calx*, a Na-Ca exchanger gene of *Drosophila melanogaster*. *Proc Natl Acad Sci USA* **94**: 10249–10254.

Sea Urchin Genome Sequencing Consortium. 2006. The genome of the sea urchin *Strongylocentrotus purpuratus*. *Science* **341**: 941–952.

Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. 2003. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* **13**: 2498–2504.

Shapiro L, Kwong PD, Fannon AM, Colman DR, Hendrickson WA. 1995. Considerations on the folding topology and evolutionary origin of cadherin domains. *Proc Natl Acad Sci USA* **92**: 6793–6797.

Shi Y, Massagué J. 2003. Mechanisms of TGF-β signaling from cell membrane to the nucleus. *Cell* **113**: 685–700.

Shinzato C, Shoguchi E, Kawashima T, Hamada M, Hisata K, Tanaka M, Fujie M, Fujiwara M, Koyanagi R, Ikuta T, et al. 2011. Using the *Acropora digitifera* genome to understand coral responses to environmental change. *Nature* **476**: 320–323.

Simakov O, Marletaz F, Cho S-J, Edsinger-Gonzales E, Havlak P, Hellsten U, Kuo D-H, Larsson T, Lv J, Arendt D, et al. 2013. Insights into bilaterian evolution from three spiralian genomes. *Nature* **493**: 526–531.

Simpson JH, Kidd T, Bland KS, Goodman CS. 2000. Short-range and long-range guidance by Slit and its Robo receptors. Robo and Robo2 play distinct roles in midline guidance. *Neuron* **28**: 753–766.

Simpson L. 1996. RNA editing. *Annu Rev Neurosci* **19**: 27–52.

Simpson TL. 1984. Gamete, embryo, larval development. In *The Cell Biology of Sponges*, pp. 341–413, Springer Verlag, New York.

Smith L, Hildemann WH. 1986. Allograft rejection, autograft fusion and inflammatory responses to injury in *Callyspongia diffusa* (Porifera; Demospongia). *Phil Trans R Soc B* **226**: 445–464.

Smith VJ, Fernandes JMO, Kemp GD, Hauton C. 2008. Crustins: Enigmatic WAP domain-containing antibacterial proteins from crustaceans. *Dev Comp Immunol* **32**: 758–772.

Sohn JH, Lee J-H, Yi H, Chun J, Bae KS, Ahn T-Y, Kim S-J. 2004. *Kordia algicida* gen. nov., sp. nov., an algicidal bacterium isolated from red tide. *Int J Syst Evol Microbiol* **54**: 675–680.

Solomon O, Oren S, Safran M, Deshet-Unger N, Akiva P, Jacob-Hirsch J, Cesarkas K, Kabesa R, Amariglio N, Unger R, et al. 2013. Global regulation of alternative splicing by adenosine deaminase acting on RNA (ADAR). *RNA* **19**: 591–604.

Spellman PT, Rubin GM. 2002. Evidence for large domains of similarly expressed genes in the *Drosophila* genome. *J Biol* **1**.

Spiegel M. 1954. The role of specific surface antigens in cell adhesion. Part 1. The reaggregation of sponge cells. *Biol Bull* **107**: 130–148.

Spieth J, Brooke G, Kuersten S, Lea K, Blumenthal T. 1993. Operons in *C. elegans*: polycistronic mRNA precursors are processed by trans-splicing of *SL2* to downstream coding regions. *Cell* **73**: 521–532.

Springer TA. 1997. Folding of the N-terminal, ligand-binding region of integrin alpha-subunits into a beta-propeller domain. *Proc Natl Acad Sci USA* **94**: 65–72.

Srivastava M, Begovic E, Chapman J, Putnam NH, Hellsten U, Kawashima T, Kuo A, Mitros T, Salamov A, Carpenter ML, et al. 2008. The *Trichoplax* genome and the nature of placozoans. *Nature* **454**: 955–960.

Srivastava M, Simakov O, Chapman J, Fahey B, Gauthier MEA, Mitros T, Richards GS, Conaco C, Dacre M, Hellsten U, et al. 2010. The *Amphimedon queenslandica* genome and the evolution of animal complexity. *Nature* **466**: 720–726.

Stanke M, Tzvetkova A, Morgenstern B. 2006. AUGUSTUS at EGASP: Using EST, protein and genomic alignments for improved gene prediction in the human genome. *Genome Biol* **7**: S11.

Steyn PL, Segers P, Vancanneyt M, Sandra P, Kersters K, Joubert JJ. 1998. Classification of heparinolytic bacteria into a new genus, *Pedobacter*, comprising four species: *Pedobacter heparinus* comb. nov., *Pedobacter piscium* comb. nov., *Pedobacter africanus* sp. nov. and *Pedobacter saltans* sp. nov. proposal of the family *Sphingobacteriaceae* fam. nov. *Int J Syst Bacteriol* **48**: 165–177.

Stoner DS, Rinkevich B, Weissman IL. 1999. Heritable germ and somatic cell lineage competitions in chimeric colonial protochordates. *Proc Natl Acad Sci USA* **96**: 9148–9153.

Stoner DS, Weissman IL. 1996. Somatic and germ cell parasitism in a colonial ascidian: possible role for a highly polymorphic allorecognition system. *Proc Natl Acad Sci USA* **93**: 15254–15259.

Strassmann JE, Queller DC. 2011. Evolution of cooperation and control of cheating in a social microbe. *Proc Natl Acad Sci USA* **108**: 10855–10862.

Suga H, Chen Z, de Mendoza A, Sebé-Pedrós A, Brown MW, Kramer E, Carr M, Kerner P, Vervoort M, Sánchez-Pons N, et al. 2013. The *Capsaspora* genome reveals a complex unicellular prehistory of animals. *Nat Comms* **4**: 2325.

Sugnet CW, Kent WJ, Ares M, Haussler D. 2004. Transcriptome and genome conservation of alternative splicing events in humans and mice. *Pac Symp Biocomput* 66–77.

Supek F, Bošnjak M, Škunca N, Šmuc T. 2011. REVIGO summarizes and visualizes long lists of gene ontology terms. *PLoS ONE* **6**: e21800.

Talavera D, Hospital A, Orozco M, delaCruz X. 2007. A procedure for identifying homologous alternative splicing events. *BMC Bioinformatics* **8**: 260.

Tang X, Bruce JE. 2009. Chemical cross-linking for protein-protein interaction studies. *Methods Mol Biol* **492**: 283–293.

Terwilliger DP, Buckley KM, Mehta D, Moorjani PG, Smith LC. 2006. Unexpected diversity displayed in cDNAs expressed by the immune cells of the purple sea urchin, *Strongylocentrotus purpuratus*. *Physiol Genomics* **26**: 134–144.

Thacker RW, Hill AL, Hill MS, Redmond NE, Collins AG, Morrow CC, Spicer L, Carmack CA, Zappe ME, Pohlmann D, et al. 2013. Nearly complete 28S rRNA gene sequences confirm new hypotheses of sponge evolution. *Integr Comp Biol* **53**: 373–387.

The Arabidopsis Genome Initiative. 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408**: 796–815.

The C. elegans Sequencing Consortium. 1998. Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science* **282**: 2012–2018.

Toole BP. 2001. Hyaluronan in morphogenesis. *Semin Cell Dev Biol* **12**: 79–87.

Tsutsui N. 2004. Scents of self: The expression component of self/non-self recognition systems. *Ann Zool Fennici* **41**: 713–724.

Turley EA. 1982. Purification of a hyaluronate-binding protein fraction that modifies cell social behavior. *Biochem Biophys Res Commun* **108**: 1016–1024.

Turley EA, Noble PW, Bourguignon LYW. 2002. Signaling properties of hyaluronan receptors. *J Biol Chem* **277**: 4589–4592.

Turner CE, West KA, Brown MC. 2001. Paxillin-ARF GAP signaling and the cytoskeleton. *Curr Opin Cell Biol* **13**: 593–599.

Uriz MJ. 1982. Reproducción en Hymellíacídoll sallguíllea (Grant, 1926): Biología de la larva y primeros estadios postlarvarios. *Invest Pesq* **46**: 29–39.

Vaillant ML. 1869. Note on the vitality of a sponge of the family Corticatæ (*Tethya lyncurium*, Lamarck). *Ann Mag Nat Hist* **3**: 172–172.

Van de Vyver G. 1975. Phenomena of cellular recognition in sponges. *Curr Top Dev Biol* **10**: 123–140.

Van de Vyver G, Barbieux B. 1983. Cellular aspects of allograft rejection in marine sponges of the genus *Polymastia*. *J Exp Zool* **227**: 1–7.

Vandesompele J, De Preter K, Pattyn F, Poppe B, Van Roy N, De Paepe A, Speleman F. 2002. Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. *Genome Biol* **3**: research0034.

Varner JA. 1995. Cell adhesion in sponges: potentiation by a cell surface 68 kDa proteoglycan-binding protein. *J Cell Sci* **108**: 3119–3126.

Varner JA. 1996. Isolation of a sponge-derived extracellular matrix adhesion protein. *J Biol Chem* **271**: 16119–16125.

Varner JA, Burger MM, Kaufman JF. 1988. Two cell surface proteins bind the sponge *Microciona prolifera* aggregation factor. *J Biol Chem* **263**: 8498–8508.

Vigetti D, Karousou E, Viola M, Deleonibus S, De Luca G, Passi A. 2014. Hyaluronan: Biosynthesis and signaling. *Biochim Biophys Acta* **1840**: 2452–2459.

Voskoboynik A, Newman AM, Corey DM, Sahoo D, Pushkarev D, Neff NF, Passarelli B, Koh W, Ishizuka KJ, Palmeri KJ, et al. 2013. Identification of a colonial chordate histocompatibility gene. *Science* **341**: 384–387.

Wagner RW, Smith JE, Cooperman BS, Nishikura K. 1989. A double-stranded RNA unwinding activity introduces structural alterations by means of adenosine to inosine conversions in mammalian cells and *Xenopus* eggs. *Proc Natl Acad Sci USA* **86**: 2647–2651.

Wall JD, Tang LF, Zerbe B, Kvale MN, Kwok P-Y, Schaefer C, Risch N. 2014. Estimating genotype error rates from high-coverage next-generation sequence data. *Genome Res* **24**: 1734–1739.

Wang B-B, Brendel V. 2006. Genomewide comparative analysis of alternative splicing in plants. *Proc Natl Acad Sci USA* **103**: 7175–7180.

Weinbaum G, Burger MM. 1973. Two component system for surface guided reassociation of animal cells. *Nature* **244**: 510–512.

Werner T, Liu G, Kang D, Ekengren S, Steiner H, Hultmark D. 2000. A family of peptidoglycan recognition proteins in the fruit fly *Drosophila melanogaster*. *Proc Natl Acad Sci USA* **97**: 13772–13777.

Westbrook MW. 2011. Introns and alternative splicing in choanoflagellates. University of California, Berkeley.

Whittaker C, Hynes R. 2002. Distribution and evolution of von Willebrand/integrin a domains. *Mol Biol Cell* **13**: 3369–3387.

Wiens M, Korzhev M, Krasko A, Thakur NL, Perovic-Ottstadt S, Breter HJ, Ushijima H, Diehl-Seifert B, Müller IM, Müller WEG. 2005. Innate immune defense of the sponge *Suberites domuncula* against bacteria involves a MyD88-dependent signaling pathway. *J Biol Chem* **280**: 27949–27959.

Wilson H. 1907. On some phenomena of coalescence and regeneration in sponges. *J Exp Zool* **5**: 245–258.

Wimmer W, Blumbach B, Diehl-Seifert B, Koziol C, Batel R, Steffen R, Müller IM, Müller WEG. 1999a. Increased expression of integrin and receptor tyrosine kinase genes during autograft fusion in the sponge *Geodia cydonium*. *Cell Adhes Commun* **7**: 111–124.

Wimmer W, Perovic S, Kruse M, Schröder HC, Krasko A, Batel R, Müller WEG. 1999b. Origin of the integrin-mediated signal transduction. Functional studies with cell cultures from the sponge *Suberites domuncula*. *Eur J Biochem* **260**: 156–165.

Wouters MA, Rigoutsos I, Chu CK, Feng LL, Sparrow DB, Dunwoodie SL. 2005. Evolution of distinct EGF domains with specific functions. *Prot Sci* **14**: 1091–1103.

Wright CF, Christodoulou J, Dobson CM, Clarke J. 2004. The importance of loop length in the folding of an immunoglobulin domain. *Protein Eng Des Sel* **17**: 443–453.

Wu MY, Hill CS. 2009. TGF-β superfamily signaling in embryonic development and homeostasis. *Dev Cell* **16**: 329–343.

Wu YL, Zheng ZY, Jiang YH, Chess L, Jiang H. 2009. The specificity of T cell regulation that enables self-nonself discrimination in the periphery. *Proc Natl Acad Sci USA* **106**: 534–539.

Wulff JL. 1986. Variation in clone structure of fragmenting coral reef sponges. *Biol J Linn Soc* **27**: 311–330.

Yang Z, Swanson WJ, Vacquier VD. 2000. Maximum-likelihood analysis of molecular adaptation in abalone sperm lysin reveals variable selective pressures among lineages and sites. *Mol Biol Evol* **17**: 1446–1455.

Yin CQ, Humphreys T. 1996. Acute cytotoxic allogeneic histoincompatibility reactions involving gray cells in the marine sponge, *Callyspongia diffusa*. *Biol Bull* **191**: 159–167.

Yund PO, Cunningham CW, Buss LW. 1987. Recruitment and postrecruitment interactions in a colonial hydroid. *Ecology* **68**: 971–982.

Zelensky AN, Gready JE. 2005. The C-type lectin-like domain superfamily. *FEBS Journal* **272**: 6179–6217.

Zhang S-M, Adema CM, Kepler TB, Loker ES. 2004. Diversification of Ig superfamily genes in an invertebrate. *Science* **305**: 251–254.

Zhang S-M, Léonard PM, Adema CM, Loker ES. 2001. Parasite-responsive IgSF members in the snail *Biomphalaria glabrata*: Characterization of novel genes with tandemly arranged IgSF domains and a fibrinogen domain. *Immunogenetics* **53**: 684–694.

Zhang X, Smith TF. 1998. Yeast “operons”. *Microb Comp Genomics* **3**: 133–140.

APPENDICES

A note on additional files

A number of appendices described throughout this thesis are impractically large for inclusion in a book-style manuscript. Therefore, these files are available to download via Cloudstor+, a cloud storage and sharing web service run by AARNet (Australian Academic and Research Network). These appendices are referenced in-text, and their titles and descriptions are listed in this Appendices section, in the order in which they would normally occur. These online-only files are highlighted with an asterisk here and in the List of Appendices.

Download information for these files is as follows:

Short URL: `http://bit.ly/1akHXys` *(NB: this will redirect to the full URL given below)*

Long URL: `https://cloudstor.aarnet.edu.au/plus/public.php?service=files&t=9e32112bb74faeafd1f2bc25aba61678`

Password: `amphimedon`

Link expiry date: None

These files are also available upon request to the author (Laura Grice) at `lfgrice@gmail.com` or `laura.grice@uqconnect.edu.au` (as of 02.04.15).

Appendix 2.1 Accession numbers for *A. queenslandica AFs* in popular sequence databases

Gene	JGI	Ensembl Metazoa	NCBI	Aqu1	Aqu2.0	Aqu2.1
AqAFA	Aqu1.225771	Aqu1.225771	XP_003384474.1	Aqu1.225771	Aqu2.34606_001	Aqu2.1.38623_001
AqAFB	Aqu1.225772	Aqu1.225772	XP_003384475.1	Aqu1.225772	Aqu2.34607_001	Aqu2.1.38624_001
AqAFC	hom.g29438.51	Aqu1.225773	XP_003384476.1	Aqu1.225773	Aqu2.34608_001	Aqu2.1.38625_001
AqAFD	Aqu0: 1457081 + 1457082	N/A	XP_003384477.1	N/A	Aqu2.34610_001	Aqu2.1.38627_001
AqAFE	hom.g29441.t1	Aqu1.225777	XP_003384479.1	Aqu1.225777	Aqu2.34612_001	Aqu2.1.38629_001
AqAFF	Aqu1.228577	Aqu1.228577	XP_003387347.1	Aqu1.228577	Aqu2.37939_001	Aqu2.1.42296_001

This table provides the AqAF accession numbers from different sequencing databases. JGI = http://genome.jgi-psf.org/ (no longer available online); Ensembl Metazoa = metazoa.ensembl.org/Amphimedon_queenslandica; NCBI = NCBI peptide database, http://www.ncbi.nlm.nih.gov/protein; Aqu1, Aqu2.0, Aqu2.1 = local in-house genome browser.

Appendix 2.2 Hidden Markov model (HMM) for the sponge Wreath domain*

* Available online via CloudStor+ (http://bit.ly/1akHXys; pw = amphimedon)

Appendix 2.3 Online sources of genome and transcriptome datasets used for this study*

URLs at which genome and transcriptome datasets were downloaded. Data accurate as of 21.03.12 except where otherwise stated. For genome datasets, downloaded files were the translated amino acid sequences of each gene model. For transcriptome datasets, downloaded files were nucleotide sequences of expressed transcripts; these sequences were translated as described in Chapter 2.4.

* Available online via CloudStor+ (http://bit.ly/1akHXys; pw = amphimedon)

Appendix 2.4 Calx-beta, VWA, VWD and Wreath domain and gene counts*

Counts of the total number of Calx-beta, VWA, VWD, and Wreath domains, and genes encoding these domains, present in the genomes of a phylogenetically diverse species. A subset of the data presented in this file is shown in Figure 2.5.

* Available online via CloudStor+ (http://bit.ly/1akHXys; pw = amphimedon)

Appendix 2.5 Sequence homology within Calx-beta domain-containing proteins*

Sequence logos of all Calx-beta domains from A. queenslandica and N. vectensis proteins possessing four or more Calx-beta domains. The A. queenslandica sequences include three of the six AF genes (AqAFA, AqAFC, AqAFE) and other non-AF genes (Aqu1 codes). Nonpolar amino acids – green, polar amino acids – purple, acidic amino acids – orange, basic amino acids – blue.

* Available online via CloudStor+ (http://bit.ly/1akHXys; pw = amphimedon)

Appendix 2.6 Details of all AF-like sequences from thirteen sponge species

Group	Species	Sequence	Size (aa)	Wreath domain?	Top AF Homology?	SP?	TM domain?	Domain Organisation
2	*A. vastus*	Av_39733.0.1.93	1073		Yes (AqAFD)			4x Calx-beta
1	*C. nucula*	Cn_13331.30	222	Yes	Yes (AqAFB)		Yes (12 - 32 aa)	Wreath (53% coverage)
1	*C. nucula*	Cn_13850.39	257	Yes	Yes (MAFp3D)			1x Calx-beta, 1x Wreath
1	*C. nucula*	Cn_15738.29	321	Yes	Yes (MAFp3core)			2x Calx-beta, 1x Wreath
1	*C. nucula*	Cn_1678.56	306	Yes	Yes (AqAFE)			1x Wreath
1	*C. nucula*	Cn_17395.18	303	Yes	Yes (AqAFD)			Wreath
1	*C. nucula*	Cn_2149.81	1193	Yes				8x Calx-beta, 1x Wreath
1	*C. nucula*	Cn_28478.66	311	Yes	Yes (SdSLIP)			1x Wreath
1	*C. nucula*	Cn_29896.16	276	Yes				Wreath
1	*C. nucula*	Cn_4622.37	554	Yes				1x Calx-beta, 1x VWD, 1x Wreath
1	*C. nucula*	Cn_6658.31	192	Yes				1x Calx-beta, 1x Wreath
1	*C. nucula*	Cn_7606.22	317	Yes				1x Wreath
1	*C. nucula*	Cn_7922.18	322	Yes				1x Wreath
1	*C. nucula*	Cn_9494.44	294	Yes	Yes (MAFp3D)			Wreath, Calx-beta
2	*C. nucula*	Cn_2401.38	540		Yes (MAFp3D)			3x Calx-beta
3a	*C. nucula*	Cn_13338.55	412					3x Calx-beta
3a	*C. nucula*	Cn_2482.31	473					3x Calx-beta
3a	*C. nucula*	Cn_3773.31	212					1x Calx-beta, 1x VW
3a	*C. nucula*	Cn_4089.103	735					6x Calx-beta
3a	*C. nucula*	Cn_4090.34	513					3x Calx-beta
3a	*C. nucula*	Cn_4994.29	430					3x Calx-beta
3a	*C. nucula*	Cn_6450.40	558					5x Calx-beta

Group	Species	Sequence	Size (aa)	Wreath domain?	Top AF Homology?	SP?	TM domain?	Domain Organisation
3a	*C. nucula*	Cn_8649.103	695					5x Calx-beta
1	*C. prolifera*	Cp_64051.0.1.19	282	Yes	Yes (MAFp3E)			2x Calx-beta, 1x Wreath
1	*C. prolifera*	Cp_77078.0.3.47	746	Yes	Yes (SdSLIP)			3x Calx-beta, 1x Wreath
1	*C. prolifera*	Cp_79623.1.2.38	598	Yes	Yes (MAFp3B)			2x Calx-beta, 1x Wreath
1	*C. prolifera*	Cp_79623.1.4.28	548	Yes	Yes (MAFp3B)			2x Calx-beta, 1x Wreath
1	*C. prolifera*	Cp_79896.0.4.37	549	Yes	Yes (MAFp3D)			1x Calx-beta, 1x Wreath
1	*C. prolifera*	Cp_80199.3.1.97	1553	Yes	Yes (MAFp3D)	Yes (1 - 32 aa)		8x Calx-beta, 1x Wreath
2	*C. prolifera*	Cp_68734.0.1.105	1619		Yes (MAFp3D)	Yes (1 - 29aa)		6x Calx-beta
2	*C. prolifera*	Cp_74490.0.3.92	1351		Yes (MAFp3C)			9x Calx-beta
2	*C. prolifera*	Cp_75360.0.2.42	715		Yes (AqAFC)			5x Calx-beta
2	*C. prolifera*	Cp_77978.0.7.38	607		Yes (AqAFC)	Yes (1 - 33 aa)		2x IG, 3x Calx-beta
2	*C. prolifera*	Cp_79465.0.4.99	1496		Yes (AqAFC)	Yes (1 - 19 aa)	Yes (1206 - 1229 aa)	2x IG, 4x Calx-beta
2	*C. prolifera*	Cp_80038.1.6.182	2697		Yes (MAFp3C)			19x Calx-beta
2	*C. prolifera*	Cp_80038.1.7.61	959		Yes (MAFp3C)			7x Calx-beta
2	*C. prolifera*	Cp_80410.1.6.27	420		Yes (MAFp3D)			3x Calx-beta
3a	*C. prolifera*	Cp_73254.1.2.36	542			Yes (1 - 31 aa)		Calx-beta, VWA

Group	Species	Sequence	Size (aa)	Wreath domain?	Top AF Homology?	SP?	TM domain?	Domain Organisation
3a	*C. prolifera*	Cp_74424.0.1.133	1853			Yes (1 - 29 aa)		13x Calx-beta
3a	*C. prolifera*	Cp_77078.1.1.28	471					4x Calx-beta
3a	*C. prolifera*	Cp_79210.2.1.18	316					3x Calx-beta
3a	*C. prolifera*	Cp_79311.0.2.82	1152					6x Calx-beta
3a	*C. prolifera*	Cp_79311.1.1.147	2145				Yes, 7TM (1752 - 1995 aa)	3x Calx-beta
3a	*C. prolifera*	Cp_79637.1.2.59	815					4x Calx-beta
3a	*C. prolifera*	Cp_79896.0.1.84	1207					9x Calx-beta
3a	*C. prolifera*	Cp_79896.0.7.95	1462					11x Calx-beta
3a	*C. prolifera*	Cp_80247.1.1.50	563			Yes (1 - 27 aa)		5x Calx-beta
3a	*C. prolifera*	Cp_80324.1.2.65	1038			Yes (1 - 29 aa)		3x Calx-beta, 1x VWA
3a	*C. prolifera*	Cp_74490.0.4.32*	450					4x Calx-beta
3a	*C. prolifera*	Cp_80332.0.1.41*	676					5x Calx-beta
3b	*C. prolifera*	Cp_72351.0.1.101	1533			Yes (1 - 29 aa)	Yes (1476 - 1501 aa)	Calx-beta, TIG, AMOP, VWD
3b	*C. prolifera*	Cp_78050.0.1.44	613			Yes (1 - 49 aa)	Yes (593 - 612 aa)	3x Calx-beta, 3x SUSHI
3b	*C. prolifera*	Cp_79210.2.5.46	624					4x Calx-beta, 1x SRCR

Group	Species	Sequence	Size (aa)	Wreath domain?	Top AF Homology?	SP?	TM domain?	Domain Organisation
3b	*C. prolifera*	Cp_80332.1.5.48	670					2x EGF, 4x Calx-beta
3b	*C. prolifera*	Cp_80458.1.2.73	973					3x Calx-beta, 6x IG
3a	*C. candelabrum*	Cc_121.210	3151			Yes (1 - 19 aa)	Yes (3056 - 3077 aa)	5x Calx-beta
3a	*C. candelabrum*	Cc_4609.232	1581					5x Calx-beta
3a	*C. candelabrum*	Cc_6414.86	1330				Yes (1166 - 1191 aa)	4x Calx-beta
3a	*C. candelabrum*	Cc_665.109	1805					5x Calx-beta
3b	*C. candelabrum*	Cc_10702.328	2111					10x Calx-beta, 2x EPTP, 1x PAN
1	*C. elegans (L)*	CeL_11090.34	542	Yes	Yes (AqAFA)			1x Calx-beta, 1x Wreath
1	*C. elegans (L)*	CeL_12598.18	280	Yes				1x Wreath
1	*C. elegans (L)*	CeL_12745.14	214	Yes	Yes (AqAFE)			Wreath
1	*C. elegans (L)*	CeL_64871.40	581	Yes	Yes (SdSLIP)			2x Calx-beta, 1x Wreath
2	*C. elegans (L)*	CeL_10397.66	490		Yes (MAFp3D)			4x Calx-beta
2	*C. elegans (L)*	CeL_64595.71	509		Yes (MAFp3C)			4x Calx-beta
2	*C. elegans (L)*	CeL_65310.42	650		Yes (MAFp3C)		Yes (331 - 359 aa)	3x Calx-beta
3a	*C. elegans (L)*	CeL_10706.70	449					3x Calx-beta
3a	*C. elegans (L)*	CeL_29965.72	1071					6x Calx-beta
3a	*C. elegans (L)*	CeL_31158.45	332					3x Calx-beta
3a	*C. elegans (L)*	CeL_31528.22	309					3x Calx-beta
3a	*C. elegans (L)*	CeL_48406.41	569					3x Calx-beta
3a	*C. elegans (L)*	CeL_12275.43*	724					4x Calx-beta
3a	*C. elegans (L)*	CeL_15988.47*	335					3x Calx-beta

Group	Species	Sequence	Size (aa)	Wreath domain?	Top AF Homology?	SP?	TM domain?	Domain Organisation
3a	*C. elegans (L)*	CeL_30192.67*	408					4x Calx-beta
1	*C. elegans (NR)*	CeN_11360.21	363	Yes	Yes (SdSLIP)			Wreath
1	*C. elegans (NR)*	CeN_19455.35	205	Yes	Yes (MAFp3D)			Wreath
1	*C. elegans (NR)*	CeN_33908.79	555					2x Calx-beta, 1x Wreath
1	*C. elegans (NR)*	CeN_63651.26	187	Yes	Yes (MAFp3core)			Calx-beta, Wreath
2	*C. elegans (NR)*	CeN_47673.58	416		Yes (MAFp3C)			4x Calx-beta
2	*C. elegans (NR)*	CeN_9832.47	324		Yes (MAFp3C)			3x Calx-beta
3a	*C. elegans (NR)*	CeN_34075.62	371					3x Calx-beta
2	*C. elegans (S)*	CeS_14267.87	667		Yes (MAFp3C)			5x Calx-beta
2	*C. elegans (S)*	CeS_17790.48	709		Yes (MAFp3E)			4x Calx-beta
2	*C. elegans (S)*	CeS_69081.29	444		Yes (AqAFC)			4x Calx-beta
2	*C. elegans (S)*	CeS_70119.59	884		Yes (AqAFE)			7x Calx-beta
2	*C. elegans (S)*	CeS_74113.36	531		Yes (MAFp3C)			5x Calx-beta
2	*C. elegans (S)*	CeS_74209.19	252		Yes (MAFp3C)			3x Calx-beta
2	*C. elegans (S)*	CeS_76241.66	997		Yes (MAFp3D)			5x Calx-beta, 3x IG
2	*C. elegans (S)*	CeS_80842.81	542		Yes (MAFp3C)			5x Calx-beta
3a	*C. elegans (S)*	CeS_14327.21	322					3x Calx-beta
3a	*C. elegans (S)*	CeS_21122.69	445					3x Calx-beta
3a	*C. elegans (S)*	CeS_66378.62	1047					6x Calx-beta
3a	*C. elegans (S)*	CeS_66726.251	1670					6x Calx-beta
3a	*C. elegans (S)*	CeS_80406.75	994					4x Calx-beta
1	*C. elegans (S)*	CeS_109959.23	354	Yes				Wreath
1	*C. elegans (S)*	CeS_20951.18	296	Yes	Yes (MAFp3E)			Wreath
1	*E. muelleri*	Em_102251	473	Yes	Yes (MAFp3core)			1x Calx-beta, 1x Wreath

Group	Species	Sequence	Size (aa)	Wreath domain?	Top AF Homology?	SP?	TM domain?	Domain Organisation
1	*E. muelleri*	Em_102342	861	Yes	Yes (AqAFD)	Yes (1 - 23 aa)		1x Calx-beta, 1x Wreath
1	*E. muelleri*	Em_137322	330	Yes	Yes (AqAFA)			1x Wreath
1	*E. muelleri*	Em_140965	366	Yes	Yes (AqAFC)			1x hEGF, 1x Wreath
1	*E. muelleri*	Em_172450	522	Yes	Yes (AqAFA)			2x Calx-beta, 1x Wreath
1	*E. muelleri*	Em_210465	425	Yes	Yes (AqAFD)			1x Wreath
1	*E. muelleri*	Em_31799	409	Yes				2x EGF-CA, 1x Wreath
1	*E. muelleri*	Em_38028	1424	Yes	Yes (SdSLIP)			9x Calx-beta, 1x Wreath
1	*E. muelleri*	Em_38031	1526	Yes	Yes (SdSLIP)			10x Calx-beta, 1x Wreath
1	*E. muelleri*	Em_3963	371	Yes	Yes (AqAFE)			1x Wreath
1	*E. muelleri*	Em_90236	466	Yes	Yes			1x sema, 1x PSI, 1x Wreath
2	*E. muelleri*	Em_133978	982		Yes (MAFp3C)		Yes (898 - 924 aa)	6x Calx-beta
2	*E. muelleri*	Em_187482	566		Yes (MAFp3C)			3x Calx-beta
2	*E. muelleri*	Em_187484	338		Yes (MAFp3C)			3x Calx-beta
2	*E. muelleri*	Em_225017	2354		Yes (MAFp3C)	Yes (1 - 23aa)	Yes (2252 - 2276 aa)	15x Calx-beta
2	*E. muelleri*	Em_57511	375		Yes (MAFp3C)			3x Calx-beta
3a	*E. muelleri*	Em_220298	359					3x Calx-beta
3a	*E. muelleri*	Em_234842	547					3x Calx-beta
3a	*E. muelleri*	Em_236140	1577					9x Calx-beta
3a	*E. muelleri*	Em_236145	1696					9x Calx-beta
3a	*E. muelleri*	Em_271555	432					3x Calx-beta

Group	Species	Sequence	Size (aa)	Wreath domain?	Top AF Homology?	SP?	TM domain?	Domain Organisation
3a	*E. muelleri*	Em_276056	2353				Yes (2252 - 2277 aa)	17x Calx-beta
3a	*E. muelleri*	Em_284806	903					3x Calx-beta
3a	*E. muelleri*	Em_37158	991					7X Calx-beta
3b	*E. muelleri*	Em_19492	5803					12x Calx-beta, 1x LamG
1	*I. fasciculata*	If_3006.99	593	Yes	Yes (MAFp3D)			3x Calx-beta, 1x Wreath
1	*I. fasciculata*	If_3013.75	551	Yes				3x Sushi 2x Calx-beta 1x Wreath
1	*I. fasciculata*	If_4663.67	426	Yes	Yes (AqAFD)			Wreath
3a	*O. carmela*	Oc_14238 (Scaffold 11397)	3112				Yes (2956 - 2982 aa)	5x Calx-beta
3a	*O. carmela*	Oc_15982 (Scaffold 13981)	3071				Yes (3018 - 3042 aa)	5x Calx-beta
3a	*O. carmela*	Oc_9463 (Scaffold 6160)	834					4x Calx-beta
3a	*O. carmela*	Oc_12256 (Scaffold 8996)	1179					4x Calx-beta
1	*P. ficiformis*	Pf_1536.51	764	Yes	Yes (AqAFD)			2x Calx-beta, 1x Wreath
1	*P. ficiformis*	Pf_19878.17	278	Yes	Yes (AqAFA)			1x Wreath, internal ITI-HC-C?
1	*P. ficiformis*	Pf_2737.42	639	Yes	Yes (AqAFE)			1x Calx-beta
1	*P. ficiformis*	Pf_2934.29	400	Yes	Yes (AqAFE)			1x Calx-beta, 1x VW, 1x Wreath
1	*P. ficiformis*	Pf_7582.101	517	Yes	Yes (AqAFE)			1x VW, 1x Wreath
1	*P. ficiformis*	Pf_7752.162	913	Yes	Yes (AqAFC)			4x Calx-beta, 1x Wreath
2	*P. ficiformis*	Pf_9904.21	335		Yes (MAFp3C)			3x Calx-beta

Group	Species	Sequence	Size (aa)	Wreath domain?	Top AF Homology?	SP?	TM domain?	Domain Organisation
3a	*P. ficiformis*	Pf_12199.52	323					3x Calx-beta
3a	*P. ficiformis*	Pf_3321.32	410					2x VW, 1x Calx-beta
1	*P. suberitoides*	Ps_12926.15	232	Yes	Yes (MAFp3core)			1x Wreath
1	*P. suberitoides*	Ps_2131.98	622	Yes	Yes (MAFp3D)			4x Calx-beta, 1x Wreath
1	*P. suberitoides*	Ps_295.20	256	Yes	Yes (SdSLIP)			Wreath
1	*P. suberitoides*	Ps_6006.56	266	Yes	Yes (SdSLIP)			Wreath
1	*P. suberitoides*	Ps_6648.67	387	Yes	Yes (AqAFC)			1x Sushi, 1x Wreath
3a	*P. suberitoides*	Ps_1211.97	564					1x Calx-beta, 1x VW
1	*S. lacustrus*	Sl_11763.41	287	Yes	Yes (MAFp3E)			1x Wreath
1	*S. lacustrus*	Sl_2005.89	525	Yes	Yes (SdSLIP)			2x Calx, 1x Wreath
1	*S. lacustrus*	Sl_2436.75	429	Yes				1x sema, 1x PSI, 1x Wreath
1	*S. lacustrus*	Sl_4453.26	405	Yes	Yes (AqAFD)			Wreath
1	*S. lacustrus*	Sl_7676.55	417	Yes	Yes (MAFp3E)			2x Calx-beta, 1x Wreath
2	*S. lacustrus*	Sl_13008.32	517		Yes (MAFp3C)		Yes (426 - 450 aa)	3x Calx-beta
2	*S. lacustrus*	Sl_3459.106	614		Yes (AqAFC)		Yes (597 - 613 aa)	4x Calx-beta
2	*S. lacustrus*	Sl_4654.49	671		Yes (MAFp3C)			5x Calx-beta
3b	*S. ciliatum*	Sci_13370				Yes (1 - 38aa)	Yes, 7TM (5821 - 6075 aa)	13x low e, 5x e-4, 3x e-3 Calx-beta, LamG
3a	*S. coactum*	Sc_338.202	1126					3x Calx-beta
3a	*S. coactum*	Sc_42601.31	588					3x Calx-beta

SP - signal peptide, TM = transmembrane domain

Appendix 2.7 A. queenslandica AF exonic domain sizes

Gene	Domain	Domain #	Exon Span	# Exons	Exon 1	Exon 2	Exon 3
AqAFA	Calx-beta	1	8 to 10	3	9 aa / 28 nt	48 aa / 147 nt	42 aa / 128 aa
AqAFA	Calx-beta	2	10 to 12	3	22 nt	138 nt	125 nt
AqAFA	Calx-beta	3	14 to 16	3	25 nt	126 nt	113 nt
AqAFA	Calx-beta	4	16 to 18	3	25 nt	123 nt	161 nt
AqAFA	Calx-beta	5	29 to 32	4	25 nt	123 nt + 80	57 nt
AqAFA	Calx-beta	6	38 to 40	3	25 nt	120 nt	158 nt
AqAFA	Calx-beta	7	44 to 46	3	25 nt	135 nt	125 nt
AqAFA	Wreath	1	46 to 48	Wreath	148nt exon 46 (617 spacer)	516 (all)	326 (all)
AqAFB	Calx-beta	1	6 to 8	3	22 nt	138 nt	131 nt
AqAFB	Calx-beta	2	10 to 12	3	10 nt ex 10	141 nt	128 nt
AqAFB	VWA	1	10	spacer 23 start, 40 end	501 nt		
AqAFB	VWA	2	13	spacer 23 start, 40 end	501 nt		
AqAFB	VWA	3	14	spacer 23 start, 46 end	486		
AqAFB	VWA	4	15	spacer 26 start, 34 end	513		

Gene	Domain	Domain #	Exon Span	# Exons	Exon 1	Exon 2	Exon 3
AqAFB	VWA	5	16	spacer 29 start, 82 nt end	444		
AqAFB	VWA	6	17	spacer 17 start, 52 end	516		
AqAFB	Wreath	1	18 to 19	Wreath	(8nt spacer start) 508nt exon 18	18nt spacer end, 326nt content	
AqAFD	Calx-beta	1	10 to 12	3	25nt exon 10	144nt	128
AqAFD	Calx-beta	2	12 to 14	3	25nt exon 12	138 nt	130
AqAFD	Calx-beta	3	14 to 16	3	28 nt exon 14	144	128
AqAFD	Calx-beta	4	16 to 18	3	25nt exon 16	138	131 nt
AqAFD	Calx-beta	5	18 to 20	3	31 nt exon 18	147	131 nt
AqAFD	Calx-beta	6	20 to 22	3	25 nt exon 20	156	131
AqAFD	Calx-beta	7	22 to 24	3	25 nt exon 22	141	131 nt
AqAFD	Calx-beta	8	25 to 26	2	115 exon 25	134	
AqAFD	Calx-beta	9	26 to 28	3	25nt exon 26	144 nt	128 nt
AqAFD	Calx-beta	10	28 to 30	3	25nt exon 28	138 nt	131 nt
AqAFD	Calx-beta	11	32 to 34	3	25nt exon 32	153	128 nt
AqAFD	Calx-beta	12	34 to 36	3	25 nt exon 34	141	131

Gene	Domain	Domain #	Exon Span	# Exons	Exon 1	Exon 2	Exon 3
AqAFD	VWD	1	39	VW	23 nt start, 124 end	390	
AqAFD	Wreath	1	39 to 41	Wreath	362 nt spacer, 175 exon 39	567 (all)	323 (all)
AqAFD	Calx-beta	1	4 to 6	3	22 nt exon 4	147	125 nt
AqAFD	Calx-beta	2	6 to 8	3	25 nt exon 6	138	128
AqAFD	Calx-beta	3	8 to 10	3	28 nt exon 8	144	125
AqAFD	Calx-beta	4	11 to 13	3	25 exon 11	141	125
AqAFD	VWD	1	16	Absent in Pfam. High E value (0.142), but CDD score OK (e-3)	Start exon 15 (28nt spacer), 170 bp exon 15	Exon 16: 446nt, then 145 spacer at end	
AqAFD	Wreath	1	16 to 18	Wreath	497 spacer, 94nt exon 16	501	329nt, 12nt spacer end
AqAFE	Calx-beta	1	23 to 25	3	28nt exon 23	147	143
AqAFE	Calx-beta	1	4 to 6	3	19ant exon 4	159	128
AqAFE	Calx-beta	2	26 to 27	2	118 exon 26	131 exon 27	
AqAFE	Calx-beta	2	6 to 8	3	29mt exon 6	165	131
AqAFE	Calx-beta	3	27 to 29	3	28nt exon 27	162 exon	122
AqAFE	Calx-beta	3	8 to 10	3	28nt exon 8	150	122

Gene	Domain	Domain #	Exon Span	# Exons	Exon 1	Exon 2	Exon 3
AqAFE	Calx-beta	4	10 to 12	3	28nt exon 10	153	128nt
AqAFE	Calx-beta	5	12 to 14	3	25nt exon 12	147	143
AqAFE	Calx-beta	6	14 to 16	3	25 nt exon 14	147	128 nt
AqAFE	Calx-beta	7	16 to 18	3	25nt exon 16	150	107
AqAFE	Calx-beta	8	19 to 20	2	136nt exon 19	128 nt exon 20	
AqAFE	Calx-beta	9	21 to 23	3	13nt exon 21	162	131
AqAFE	VWA	1	21	VW	20nt at start, 25 at end	534 content	
AqAFE	VWA	2	30	VW	20 at start, 43 at end	567	
AqAFE	VWA	3	31	VW	20 at start, 37nt at end	573	
AqAFE	Wreath	1	32 to 34	Wreath	452 spacer, 97 exon 32	516	344
AqAFF	Calx-beta	1	2 to 3	2	139nt exon 2	131 nt exon 3	

Appendix 3.1 Results of Tukey's HSD analysis for developmental AqAF expression*

The statistical analysis was run in concert with a one-way ANOVA in R. Ticks represent instances where expression levels are significantly different between pairs, crosses represent non-significant expression differences. P-value ranges are indicated for significant differences in expression between stages (**** = $p \leq 0.0001$, *** = $p \leq 0.001$, ** = $p \leq 0.01$, * = $p < 0.05$, not significant = > 0.05). The analysis was performed comparing different (a) genes and (b) timepoints. PC = Pre-competent larvae (0 – 7 hours post emergence, hpe), C = Competent larvae (6 – 12 hpe), late larvae (23 – 50 hpe).

* Available online via CloudStor+ (http://bit.ly/1akHXys; pw = amphimedon)

Appendix 3.2 Commands for identification of correlated gene expression

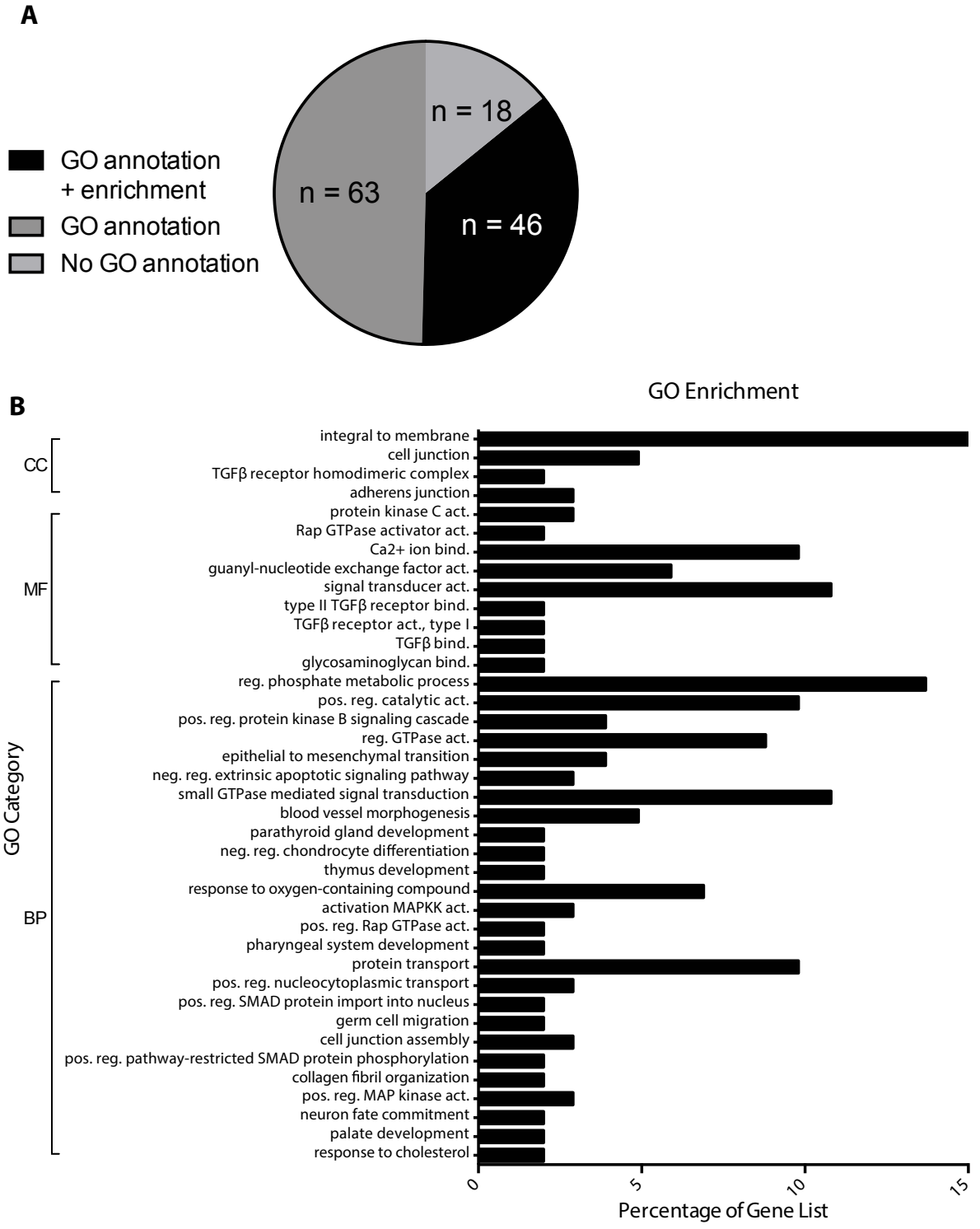
* Available online via CloudStor+ (http://bit.ly/1akHXys; pw = amphimedon)

Appendix 3.3 Genes exhibiting expression correlation to the *AqAFs*

* Available online via CloudStor+ (http://bit.ly/1akHXys; pw = amphimedon)

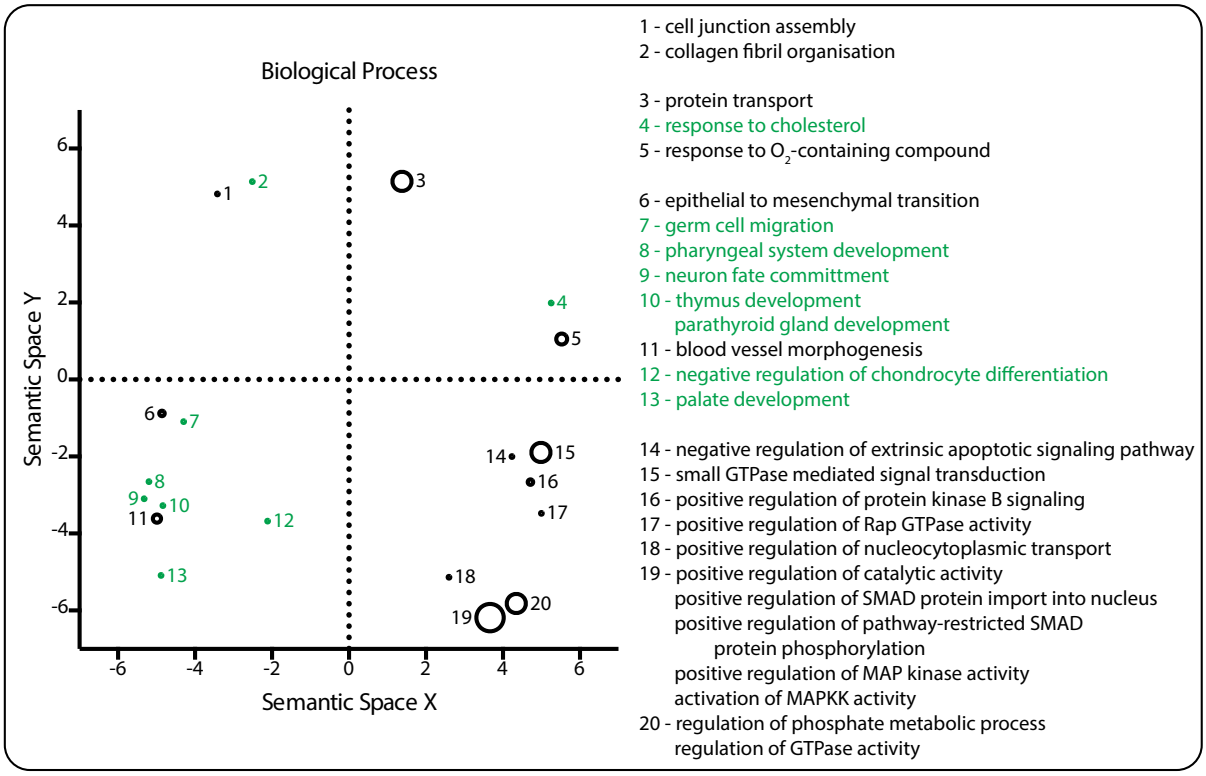
Appendix 3.4 - Statistically enriched Gene Ontology terms from genes with expression pattern correlations with the *AqAFs*

Expression of the AqAFs correlates with that of 122 A. queenslandica genes. (A) Gene annotation and enrichment status of correlated genes. (B) Percentage of correlated genes annotated with enriched Gene Ontology terms.



Appendix 3.5 Distribution of enriched GO terms within semantic space

Each graph represents the statistically enriched GO terms (for Biological Process and Molecular Function) associated with the list of genes potentially coexpressed with the AqAFs. Enriched GO terms are clustered based on the SimRel measure of semantic similarity and plotted on arbitrary X and Y axes. Circle size is proportional to the number of gene sequences annotated with each GO term. Points labelled in green represent those GO terms associated only with the two TGF-β receptor type 1 genes (Aqu2.1.41568_001 and Aqu2.1.41569_001).

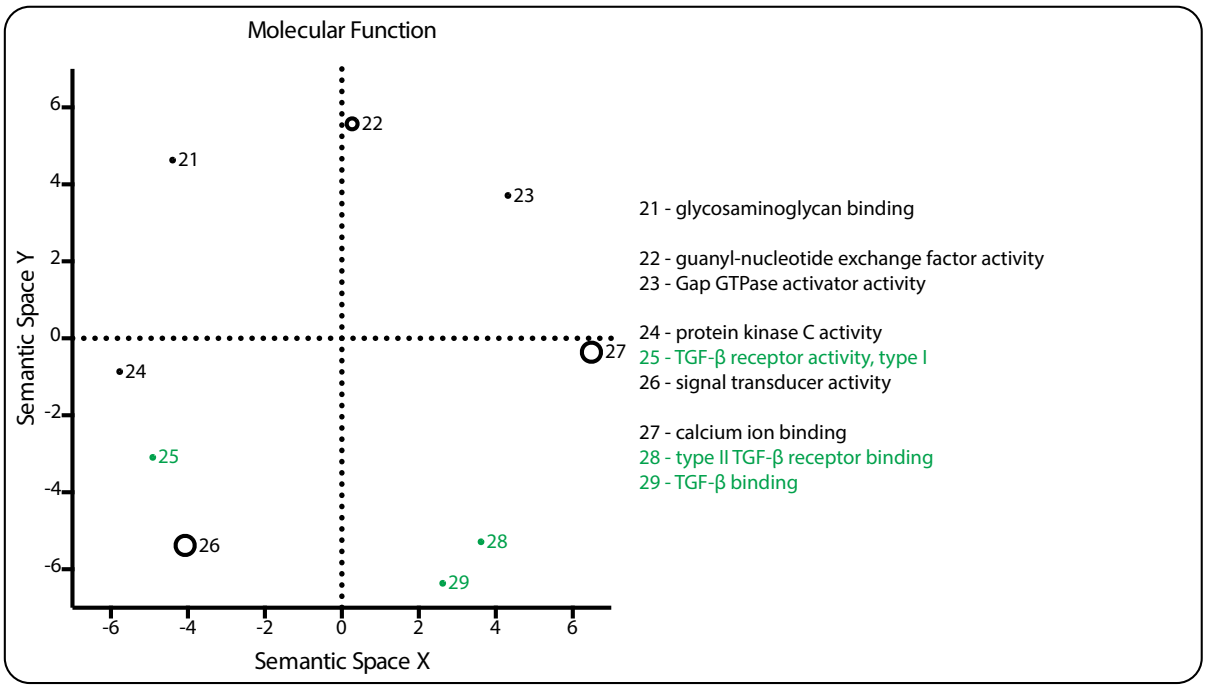


1 - cell junction assembly
2 - collagen fibril organisation

3 - protein transport
4 - response to cholesterol
5 - response to O_2-containing compound

6 - epithelial to mesenchymal transition
7 - germ cell migration
8 - pharyngeal system development
9 - neuron fate committment
10 - thymus development
parathyroid gland development
11 - blood vessel morphogenesis
12 - negative regulation of chondrocyte differentiation
13 - palate development

14 - negative regulation of extrinsic apoptotic signaling pathway
15 - small GTPase mediated signal transduction
16 - positive regulation of protein kinase B signaling
17 - positive regulation of Rap GTPase activity
18 - positive regulation of nucleocytoplasmic transport
19 - positive regulation of catalytic activity
positive regulation of SMAD protein import into nucleus
positive regulation of pathway-restricted SMAD protein phosphorylation
positive regulation of MAP kinase activity
activation of MAPKK activity
20 - regulation of phosphate metabolic process
regulation of GTPase activity



21 - glycosaminoglycan binding

22 - guanyl-nucleotide exchange factor activity
23 - Gap GTPase activator activity

24 - protein kinase C activity
25 - TGF-β receptor activity, type I
26 - signal transducer activity

27 - calcium ion binding
28 - type II TGF-β receptor binding
29 - TGF-β binding

Appendix 3.6 Predicted hyaluronan binding motifs in the *A. queenslandica AFs*

GENE	COORDINATE	SEQUENCE	PATTERN
AqAFA	654 - 662	KhyllrlkK	[RK]-x(7)-[RK]
AqAFA	2577 - 2585	RcelrsstR	[RK]-x(7)-[RK]
AqAFA	2585 - 2593	RrlttfrdR	[RK]-x(7)-[RK]
AqAFA	2824 - 2832	RiriravnK	[RK]-x(7)-[RK]
AqAFA	2949 - 2957	RidvkprnK	[RK]-x(7)-[RK]
AqAFA	3020 - 3028	RyghfesnR	[RK]-x(7)-[RK]
AqAFA	310 - 317	RvrldplK	[RK]-x(6)-[RK]
AqAFA	505 - 512	RsdystrR	[RK]-x(6)-[RK]
AqAFA	654 - 661	KhyllrlK	[RK]-x(6)-[RK]
AqAFA	2586 - 2593	RlttfrdR	[RK]-x(6)-[RK]
AqAFA	2631 - 2638	RlgvrlgR	[RK]-x(6)-[RK]
AqAFA	2817 - 2824	RdfhgvdR	[RK]-x(6)-[RK]
AqAFA	2953 - 2960	KprnkpqR	[RK]-x(6)-[RK]
AqAFB	244 - 251	KtirvhvK	[RK]-x(6)-[RK]
AqAFB	737 - 744	KvtrpstR	[RK]-x(6)-[RK]
AqAFB	1753 - 1760	RdlhlinK	[RK]-x(6)-[RK]
AqAFB	1805 - 1812	KrngvhvR	[RK]-x(6)-[RK]
AqAFB	1837 - 1844	KsvlkekK	[RK]-x(6)-[RK]
AqAFB	486 - 494	RfvadvakK	[RK]-x(7)-[RK]
AqAFB	844 - 852	RiareellK	[RK]-x(7)-[RK]
AqAFB	847 - 855	ReellkngR	[RK]-x(7)-[RK]
AqAFB	852 - 860	KngresvpR	[RK]-x(7)-[RK]
AqAFB	1091 - 1099	KeiatsekK	[RK]-x(7)-[RK]
AqAFB	1222 - 1230	RnefringR	[RK]-x(7)-[RK]
AqAFB	1226 - 1234	RingrsgaR	[RK]-x(7)-[RK]
AqAFB	1569 - 1577	KpeliqriR	[RK]-x(7)-[RK]
AqAFB	1815 - 1823	KrnvflsvK	[RK]-x(7)-[RK]
AqAFB	1837 - 1845	KsvlkekkR	[RK]-x(7)-[RK]
AqAFB	1879 - 1887	RhngdielR	[RK]-x(7)-[RK]

Gene	Coordinate	Sequence	Pattern
AqAFC	2388 - 2396	RysdrvriK	[RK]-x(7)-[RK]
AqAFC	697 - 704	KrftgvlR	[RK]-x(6)-[RK]
AqAFC	918 - 925	KrftgvlR	[RK]-x(6)-[RK]
AqAFC	1585 - 1592	KrftgvlR	[RK]-x(6)-[RK]
AqAFC	2387 - 2394	RrysdrvR	[RK]-x(6)-[RK]
AqAFD	1205 - 1212	KlpnerkK	[RK]-x(6)-[RK]
AqAFD	1210 - 1217	RkkdvriR	[RK]-x(6)-[RK]
AqAFD	1674 - 1681	RswdrsfR	[RK]-x(6)-[RK]
AqAFD	1025 - 1033	RnstriniR	[RK]-x(7)-[RK]
AqAFE	30 - 37	KghlvddR	[RK]-x(6)-[RK]
AqAFE	1417 - 1424	RnistrgR	[RK]-x(6)-[RK]
AqAFE	1630 - 1637	KrqltfpK	[RK]-x(6)-[RK]
AqAFE	2032 - 2039	RgtfthhK	[RK]-x(6)-[RK]
AqAFE	2052 - 2059	RgregasK	[RK]-x(6)-[RK]
AqAFE	2242 - 2249	RgtfthhK	[RK]-x(6)-[RK]
AqAFE	2262 - 2269	RgregasK	[RK]-x(6)-[RK]
AqAFE	2543 - 2550	RrecaviK	[RK]-x(6)-[RK]
AqAFE	37 - 45	RsnddrstK	[RK]-x(7)-[RK]
AqAFE	1306 - 1314	RfstesrtR	[RK]-x(7)-[RK]
AqAFE	1924 - 1932	RltirsseR	[RK]-x(7)-[RK]
AqAFE	2046 - 2054	RqqfndrgR	[RK]-x(7)-[RK]
AqAFE	2256 - 2264	RqqfndrgR	[RK]-x(7)-[RK]
AqAFE	2666 - 2674	RqrmatrvR	[RK]-x(7)-[RK]
AqAFE	2822 - 2830	RyekfdssR	[RK]-x(7)-[RK]

Appendix 4.1 PCR reaction mixtures

INGREDIENT	CONCENTRATION	VOLUME			
		F18R22	F23R24	F34R22	F39R22
Buffer (Promega)	10 x	-	2.5 µL	-	-
Buffer (Thermopol	10 x	2.5 µL	-	2.5 µL	2.5 µL
MgCl2	25 mM	2.5 µL	1 µL	2.5 µL	2.5 µL
dNTP	10 mM	1 µL	0.5 µL	0.5 µL	0.5 µL
Primer (fwd)	10 mM	2 µL	2.5 µL	0.5 µL	0.5 µL
Primer (rev)	10 mM	2 µL	2.5 µL	0.5 µL	0.5 µL
Taq (in-house)	1 U/µL	-	0.25 µL	-	-
Taq (NEB)	1 U/µL	0.125 µL	-	0.125 µL	0.125 µL
BSA	20 x	-	-	2.5 µL	2.5 µL
cDNA	-	2 µL	2 µL	1 µL	1 µL
H2O	-	12.875	13.75 µL	14.875 µL	14.875 µL

Appendix 4.2 Thermocycler conditions for PCR

STAGE	TEMPERATURE - TIME		
	F18-R20	F23-R24	F34R22 R39R22
Denaturation (1x)	95°C - 5 min	94°C - 2 min	95°C - 5 min
Cycling (50x)	95°C - 30 s	94°C - 30 s	95°C - 30 s
	61°C - 30 s	62°C - 30 s	59°C - 30 s
	68°C - 90 s	72°C - 90 s	68°C - 100 s
Final Extension	68°C - 5 min	72°C - 5 min	68°C - 5 min

Appendix 4.3 General nucleotide variant information (full table)*

This is the full version of the table shown in Table 4.4, with information from the four individual sponges included.

* Available online via CloudStor+ (http://bit.ly/1akHXys; pw = amphimedon)

Appendix 4.4 Raw variant counts per *A. queenslandica* gene per allele per sponge

		SPONGE A		SPONGE B		SPONGE C	
		A1	A2	A1	A2	A1	A2
AqAFA	Synonymous	50	36	46	39	45	52
	Conservative	22	14	14	9	20	31
	Non-Conservative	19	18	14	20	30	28
	Intron	9	8	11	16	33	34
AqAFB	Synonymous	18	17	4	14	15	42
	Conservative	8	13	3	6	10	23
	Non-Conservative	10	7	5	8	14	25
	Intron	0	0	0	2	4	8
AqAFC	Synonymous	13	13	4	6	28	17
	Conservative	2	2	1	1	12	8
	Non-Conservative	2	2	0	2	21	3
	Intron	1	1	2	2	8	8
AqAFD	Synonymous	12	18	14	17	23	27
	Conservative	6	8	3	5	11	11
	Non-Conservative	1	1	1	3	5	4
	Intron	0	0	0	0	9	9
AqAFE	Synonymous	23	42	5	4	16	20
	Conservative	7	24	3	2	13	13
	Non-Conservative	10	36	6	1	10	10
	Intron	6	6	6	6	5	5
AqAFF	Synonymous	0	0	0	0	0	0
	Conservative	0	0	0	0	0	0
	Non-Conservative	0	0	0	0	0	0
	Intron	0	0	0	0	0	0

A1, A2 = allele 1, 2

Appendix 6.1 Commands for independent filtering and differential gene expression analysis*

* Available online via CloudStor+ (http://bit.ly/1akHXys; pw = amphimedon)

Appendix 6.2 Filtering of candidate differentially expressed genes by fold change

SAMPLE		DIFFERENTIALLY EXPRESSED GENES (P ≤ 0.01)			DIFFERENTIALLY EXPRESSED GENES (P ≤ 0.01, FC ≥ 4)		
		TOTAL	UP	DOWN	TOTAL	UP	DOWN
0 vs 12 hpg	AA	1	0	1	1	0	1
	BB	56	20	36	41	16	25
	AB	583	77	506	408	29	379
	AA vs AB	19	0	19	19	0	19
	BB vs AB	160	5	155	160	5	155
12 vs 24 hpg	AA	5	2	3	5	2	3
	BB	18	5	13	16	3	13
	AB	131	14	117	110	9	101
	AA vs AB	81	17	64	81	17	64
	BB vs AB	46	14	32	46	14	32
24 vs 48 hpg	AA	22	1	21	19	0	19
	BB	10	3	7	10	3	7
	AB	3365	1812	1753	2227	1049	1178
	AA vs AB	1076	306	770	1039	294	745
	BB vs AB	1480	941	539	1395	885	510
48 vs 72 hpg	AA	6	4	2	6	4	2
	BB	34	8	26	30	8	22
	AB	511	37	474	417	20	397
	AA vs AB	244	7	237	244	7	237
	BB vs AB	95	9	86	95	9	86

Appendix 6.3 Counts of alternatively spliced *AF* transcripts in grafted samples

		0HPG	12 HPG		24 HPG		48 HPG		72 HPG	
			S	NS	S	NS	S	NS	S	NS
AFA	IR									
	SiI									
	EiI									
	Esk						1			
AFB	IR			1			1			
	SiI				1					
	EiI		1					1		1
	Esk									
AFC	IR				1	1				1
	SiI				1				1	
	EiI		1				1			
	Esk								3	
AFD	IR									
	SiI			1		2		1		1
	EiI		1*			1				
	Esk									
AFE	IR									
	SiI					1				
	EiI				2					
	Esk								1	1

IR = Intron retention; SiI = starts in intron; EiI = ends in intron; Esk = exon skipping
S = self; NS = nonself; * = unknown sequence
No AqAFF alternatively spliced transcripts were identified; therefore this gene is not shown here

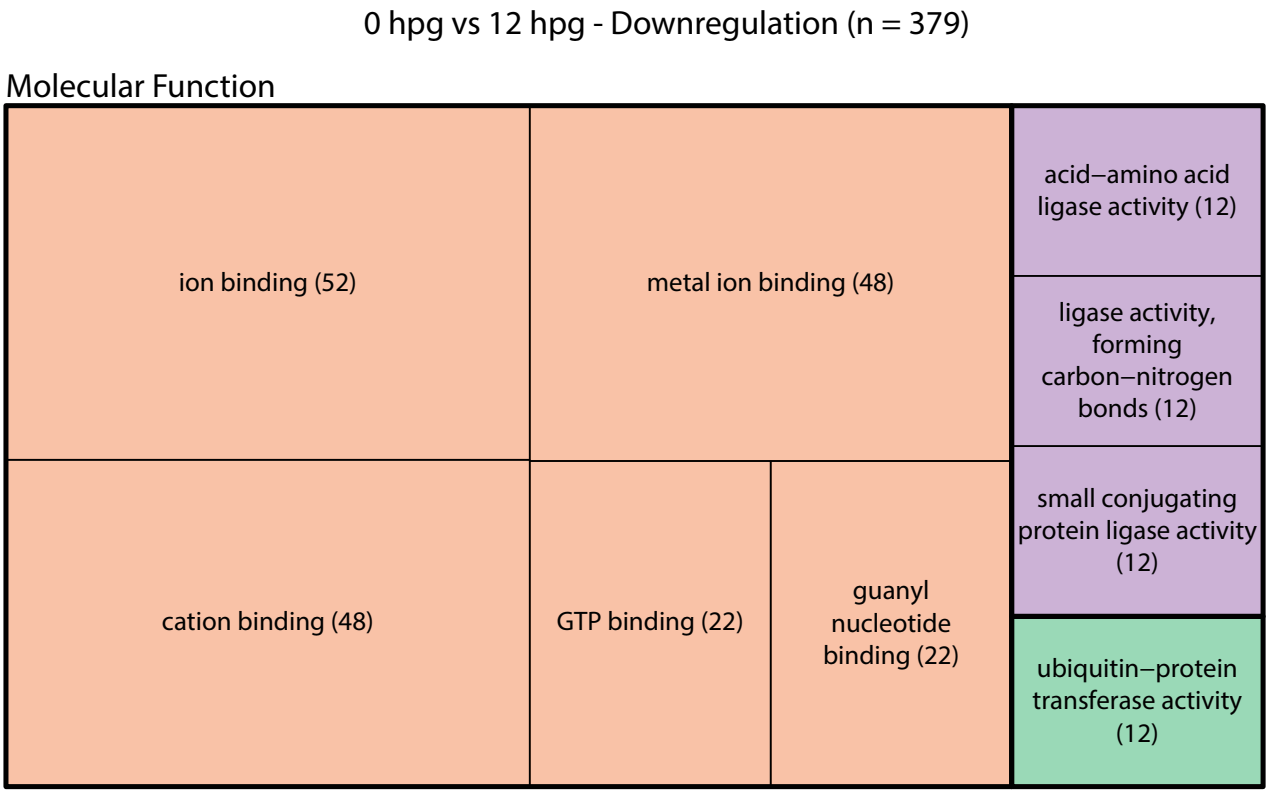
Appendix 6.4 List of 4-fold or higher differentially expressed genes in the graft response*

* Available online via CloudStor+ (http://bit.ly/1akHXys; pw = amphimedon)

Appendix 6.5 Enriched Gene Ontology terms in the nonself time course

(Part 1 of 9)

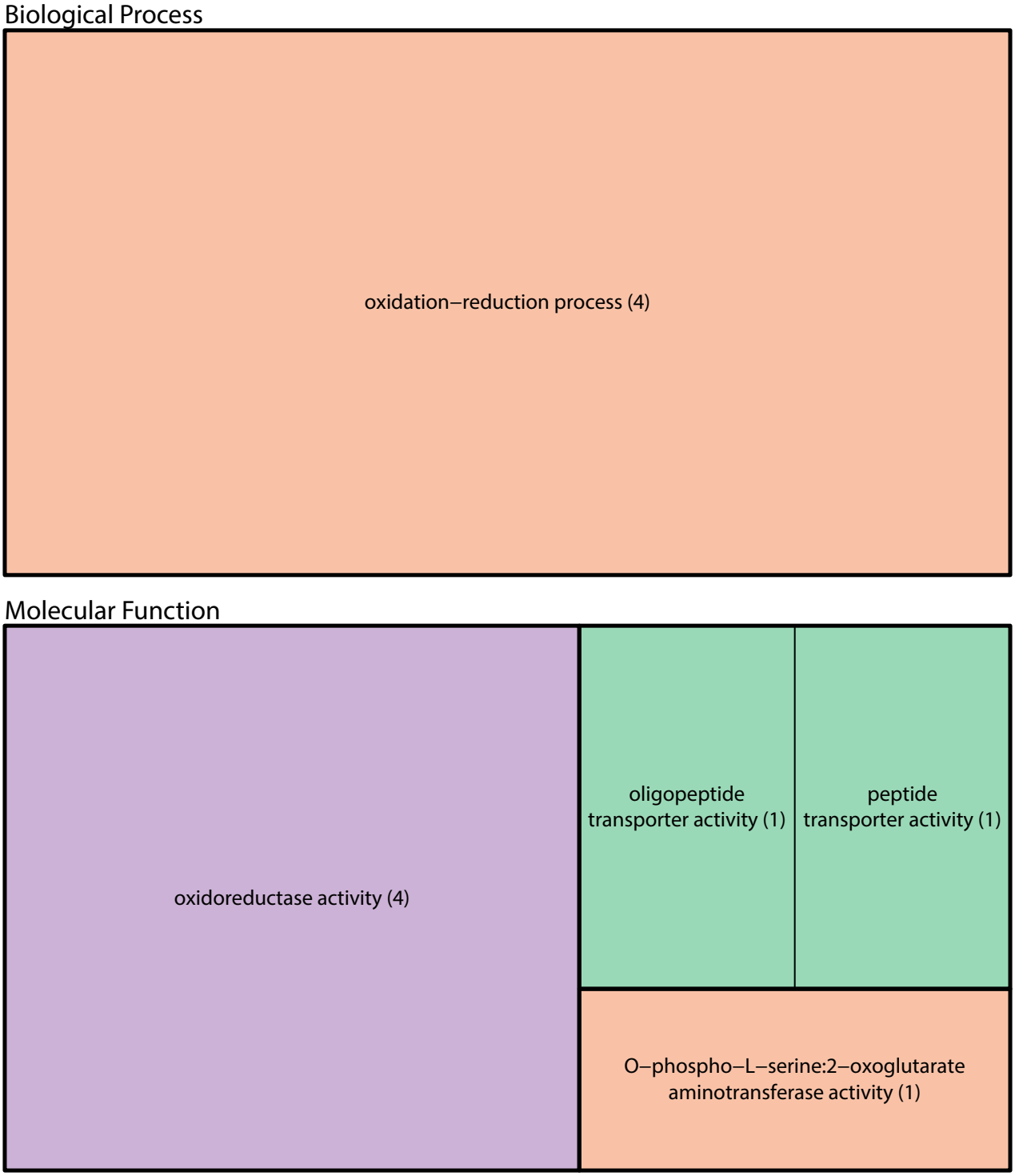
Each treemap represents the statistically enriched Gene Ontology (GO) terms (for Biological Process and Molecular Function terms) associated with the genes which are up- or downregulated at different times in the nonself graft time course. The total number of up- or downregulated genes for each time point is given at the top of each page. Within each treemap, each coloured box represents an enriched GO term associated with the gene list, with box size proportional to the number of genes annotated with that GO term (also shown in brackets). Identically-coloured boxes represent superclusters of loosely related GO terms.



Appendix 6.5 Enriched GO terms for differentially expressed genes in the nonself time course

(Part 2 of 9)

0 hpg vs 12 hpg - Upregulation (n = 29)

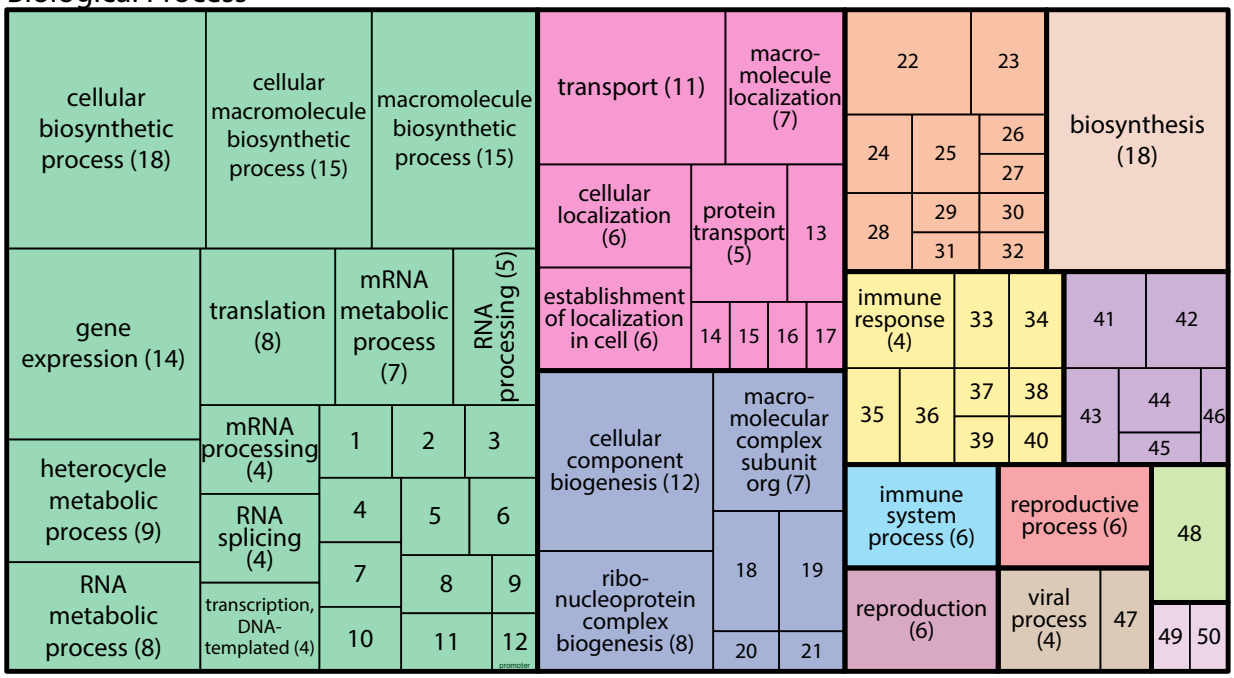


Appendix 6.5 Enriched GO terms for differentially expressed genes in the nonself time course

(Part 3 of 9)

12 hpg vs 24 hpg - Downregulation (n = 101) - 1 of 2

Biological Process

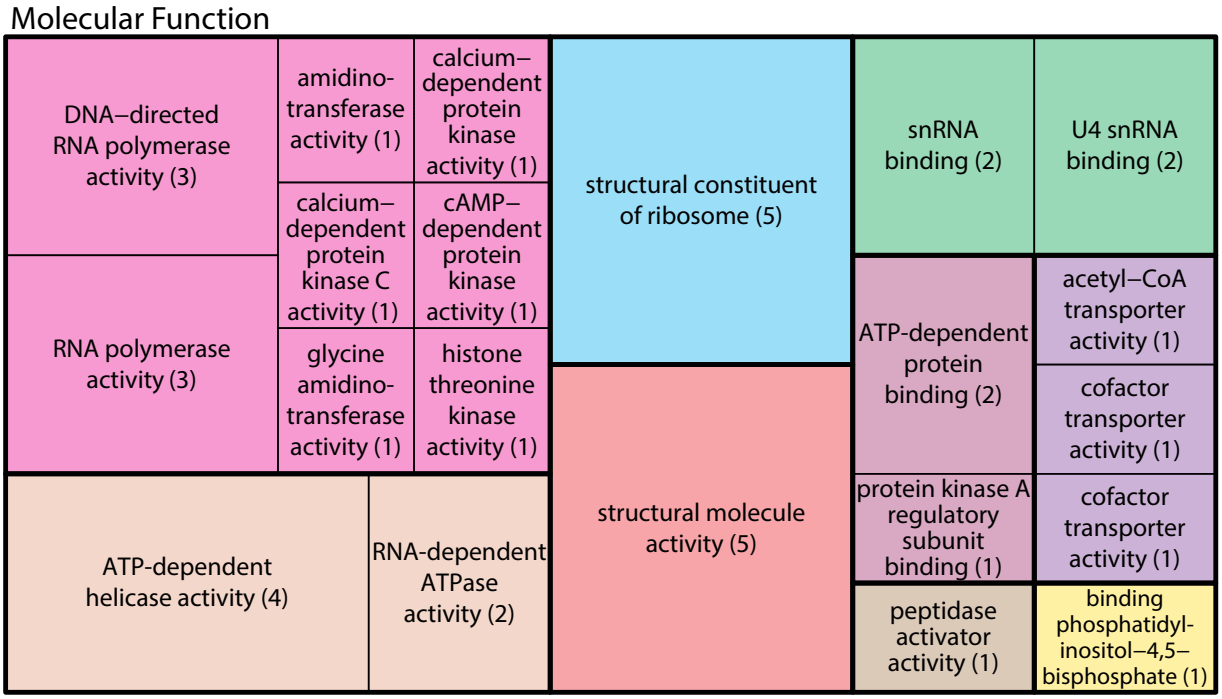


1 - 7–methylguanosine RNA capping (2)
2 - DNA-templated transcription, elongation (2)
3 - DNA-templated transcription, termination (2)
4 - RNA secondary structure unwinding (2)
5 - transcription elongation from RNA pol. III promoter (2)
6 - transcription from RNA pol. III promoter (2)
7 - termination of RNA pol. III transcription (2)
8 - transcription initiation from RNA pol. II promoter (2)
10 - transcription elongation from RNA pol. II promoter (2)
9 - proteasomal ubiquitin-independent protein catabolic process (1)
11 - translational initiation (2)
12 - regulation of transcription from RNA pol. promoter (1)
13 - nucleobase–containing compound transport (3)
14 - acetyl–CoA transport (1)
15 - coenzyme transport (1)
16 - rRNA export from nucleus (1)
17 - rRNA transport (1)
18 - cellular component disassembly (3)
19 - protein complex disassembly (3)
20 - desmosome assembly (1)
21 - histone H3-T6 phosphorylation (1)
22 - regulation of multicellular organismal process (5)
23 - positive regulation of multicellular organismal process (3)
24 - germ cell development (2)
25 - regulation of type I interferon production (2)
26 - activation of protein kinase A activity (1)
27 - negative regulation of glucose import (1)
28 - regulation of cell cycle arrest (2)
29 - positive regulation of behaviour (1)
30 - positive regulation of inflammatory response (1)
31 - positive regulation of cardiac muscle hypertrophy (1)
32 - regulation of macrophage differentiation (1)
33 - response to peptide hormone (2)
34 - transcription-coupled nucleotide excision repair (2)
35 - antigen processing and presentation of exogenous antigen (2)
37 - cellular response to parathyroid hormone stimulus (1)
36 - antigen processing & presentation of exogenous peptide antigen (2)
38 - response to mercury ion (1)
39 - neutrophil chemotaxis (1)
40 - response to parathyroid hormone (1)
41 - nucleobase metabolic process (3)
42 - purine nucleobase metabolic process (3)
43 - ether metabolic process (2)
44 - glycerol ether metabolic process (2)
45 - creatine biosynthetic process (1)
46 - creatine metabolic process (1)
47 - regulation of viral process (2)
48 - multi-organism process (4)
49 - triglyceride catabolism (1)
50 - glycrolipid catabolism (1)

Appendix 6.5 Enriched GO terms for differentially expressed genes in the nonself time course

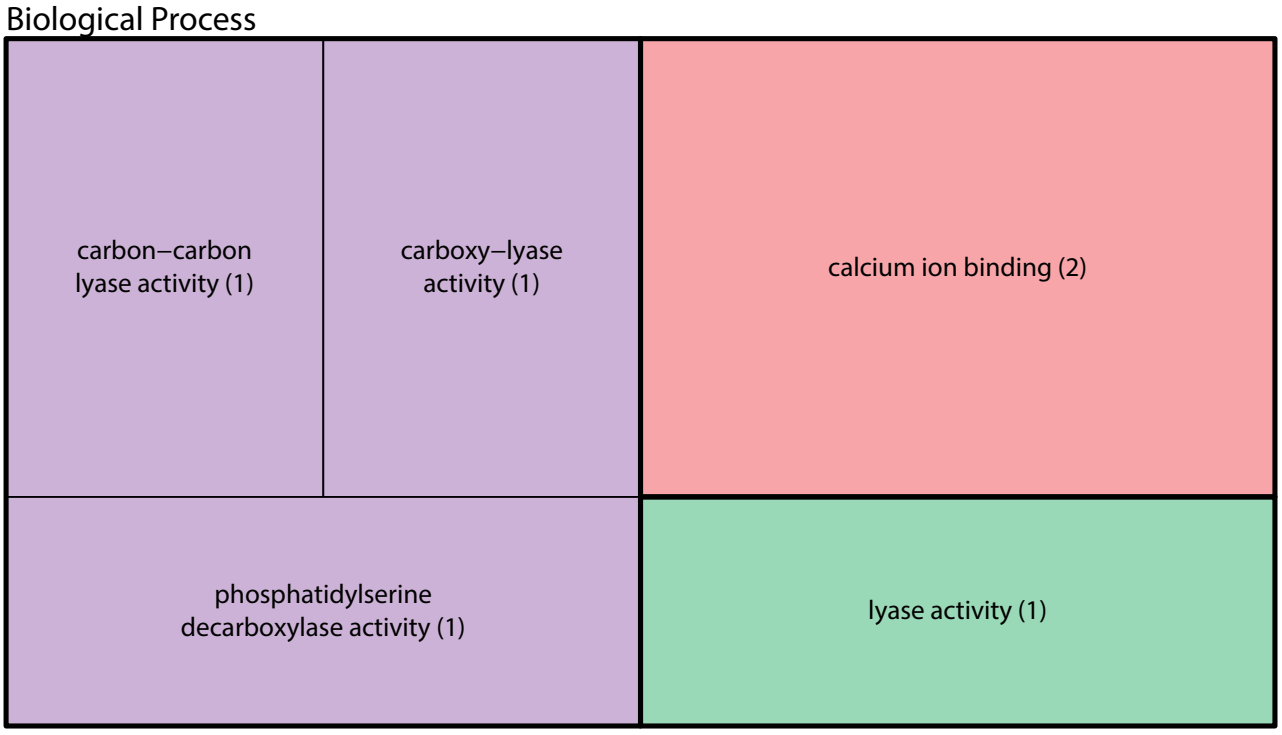
(Part 4 of 9)

12 hpg vs 24 hpg - Downregulation (n = 101) - 2 of 2



Appendix 6.5 Enriched GO terms for differentially expressed genes in the nonself time course
(Part 5 of 9)

12 hpg vs 24 hpg - Upregulation (n = 9)

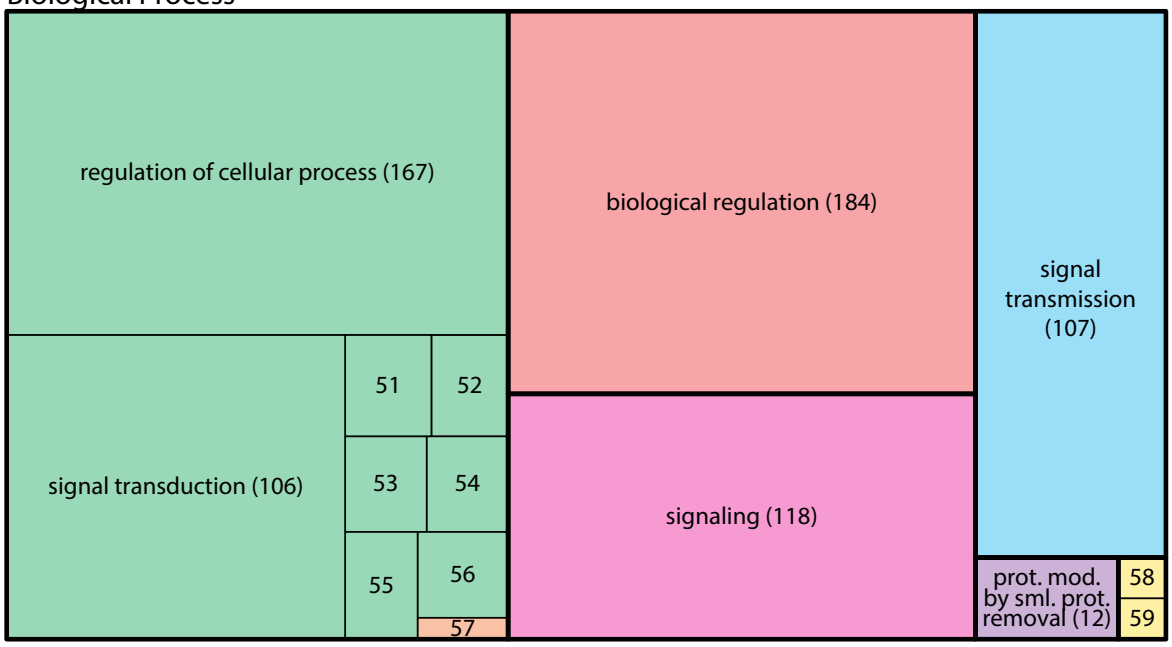


Appendix 6.5 Enriched GO terms for differentially expressed genes in the nonself time course

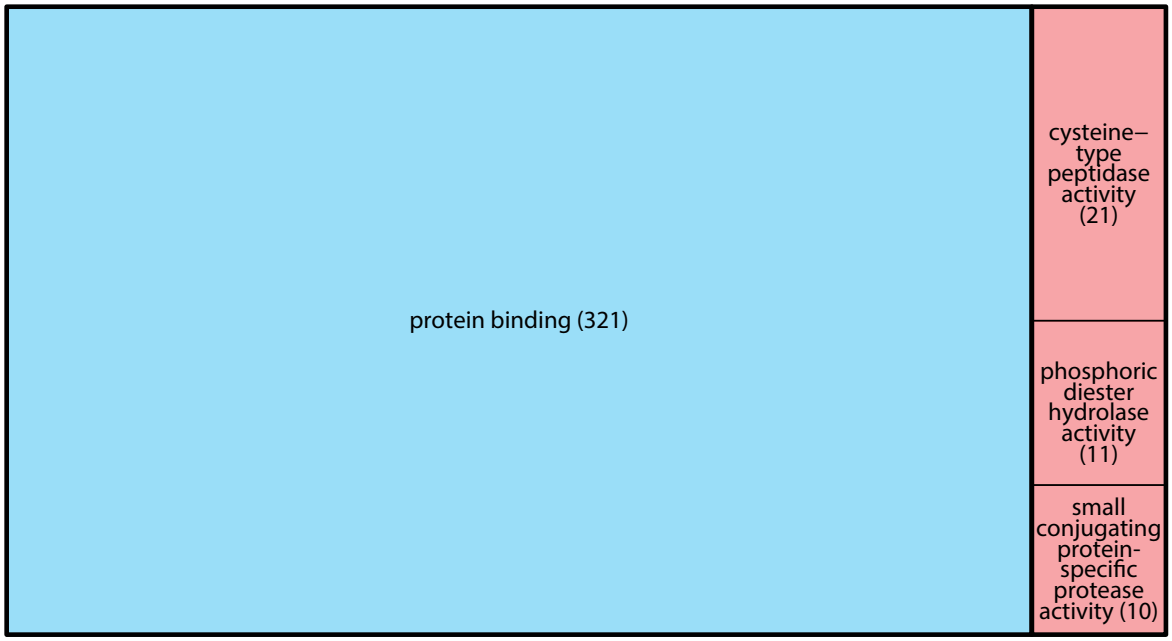
(Part 6 of 9)

24 hpg vs 48 hpg - Downregulation (n = 1178)

Biological Process



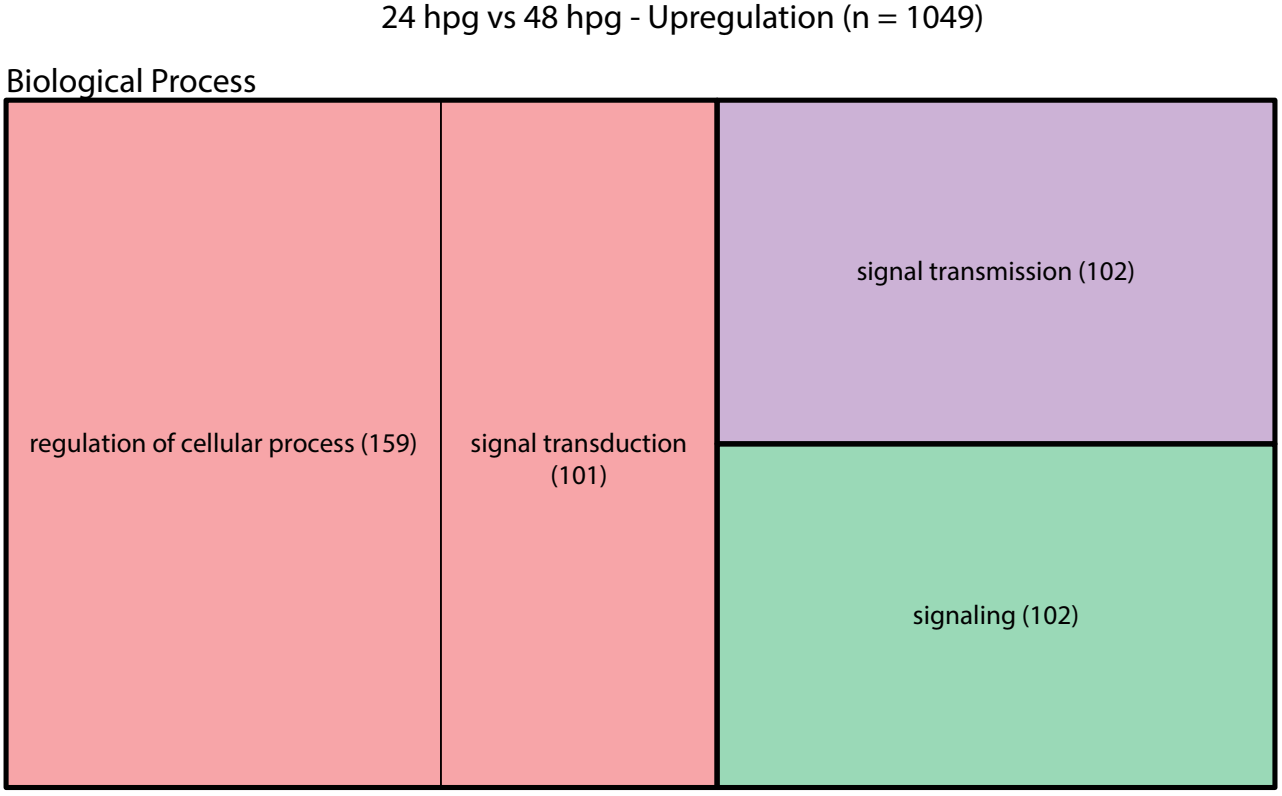
Molecular Function



51 - regulation of cysteine-type endopeptidase activity involved in apoptotic process (9)
52 - base-excision repair (8)
53 - negative regulation of cell cycle process (8)
54 - negative regulation of mitotic cell cycle (8)
55 - positive regulation of cell growth (8)
56 - positive regulation of cell size (8)
57 - synaptonemal complex organisation (2)
58 - oocyte maturation (2)
59 - synaptonemal complex assembly (2)

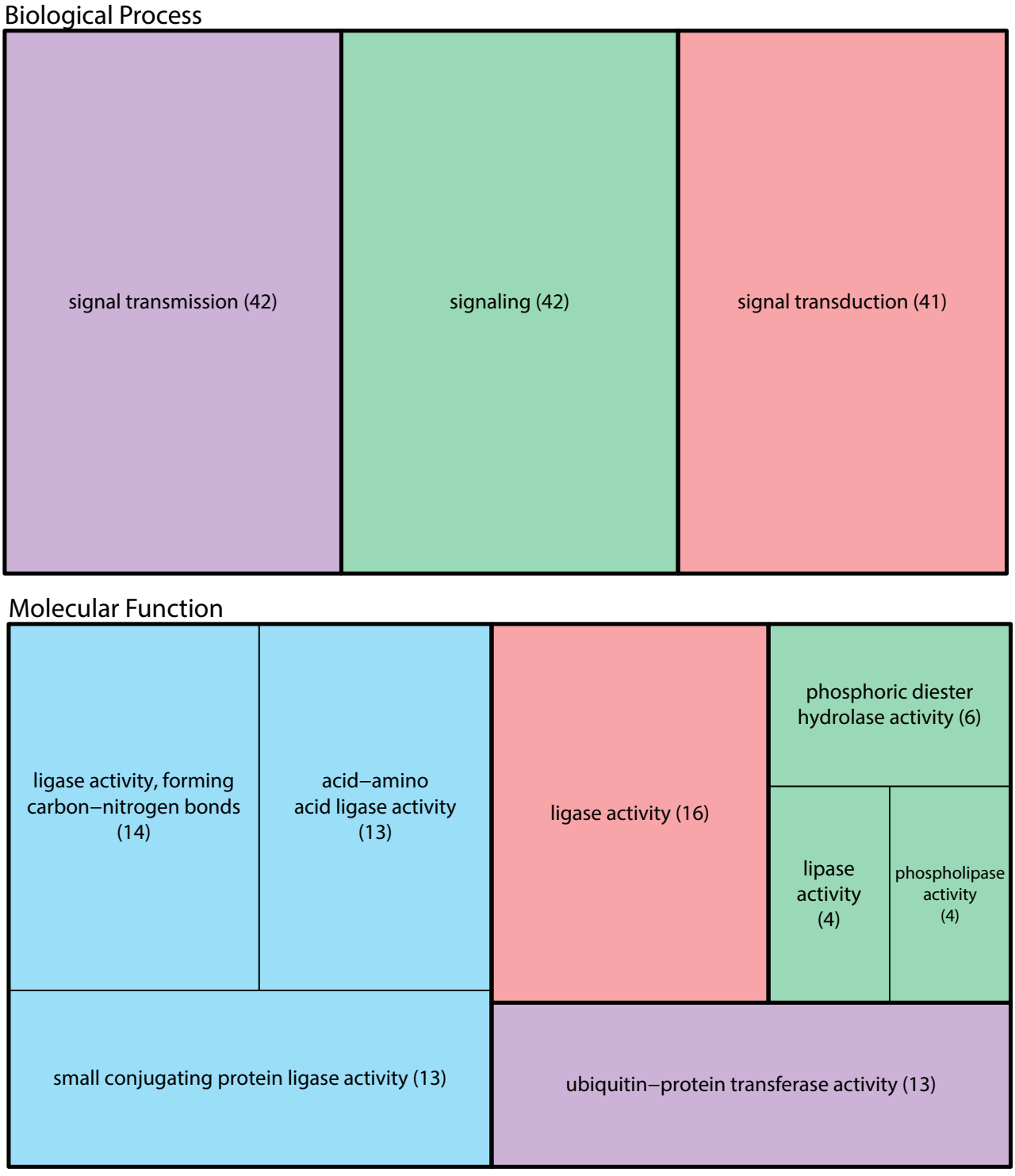
Appendix 6.5 Enriched GO terms for differentially expressed genes in the nonself time course

(Part 7 of 9)



Appendix 6.5 Enriched GO terms for differentially expressed genes in the nonself time course
(Part 8 of 9)

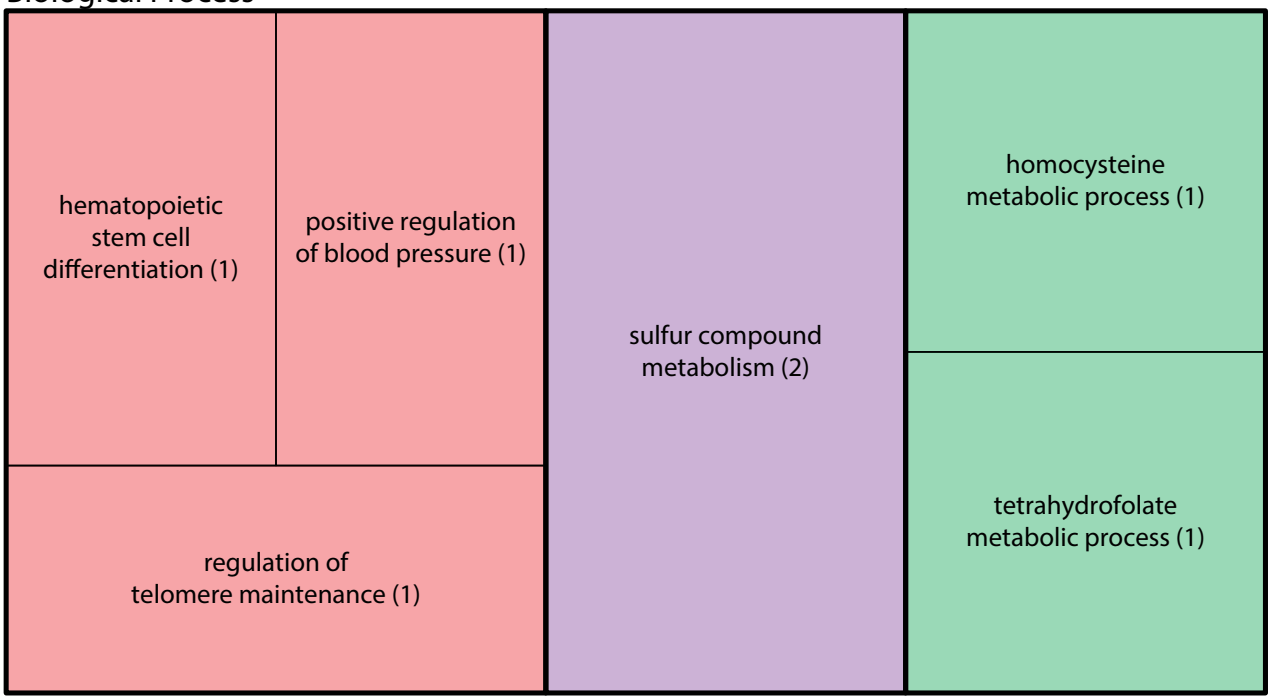
48 hpg 72 hpg - Downregulation (n = 379)



Appendix 6.5 Enriched GO terms for differentially expressed genes in the nonself time course
(Part 9 of 9)

48 hpg vs 72 hpg - Upregulation (n = 20)

Biological Process



Molecular Function