Immunoinformatics: Towards an understanding of species-specific protein evolution using phylogenomics and network theory

Andrew Edward Webb M.Sc. Genetics,

B.Sc. Biotechnology (*cum laude*)

A thesis presented to Dublin City University for the Degree of

Doctor of Philosophy

Supervisor: Dr. Mary J. O'Connell

School of Biotechnology

Dublin City University

January 2015

Declaration

'I hereby certify that this material, which I now submit for assessment on the programme of study leading to the award of Doctor of Philosophy is entirely my own work, that I have exercised reasonable care to ensure that the work is original, and does not to the best of my knowledge breach any law of copyright, and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the text of my work.'

Signed: _____

ID Number: 10114017

Date: _____

Table of Contents

Acknowledgements

First and foremost, I must wholeheartedly give thanks to my supervisor Dr. Mary O'Connell for all her support, enthusiasm, and dedication throughout my time at DCU. I cannot express in words how wonderful working alongside you has been and how truly difficult it was to leave. I know for certain that your mentorship has made me a better researcher, giving me more confidence in my abilities, taught me the benefits of collaboration, and shown me how to operate a truly cooperative and welcoming laboratory environment. And of course, I must thank you for being such a wonderful role model. I know for certain that our paths will cross again and I look forward to those occasions.

Thanks to my fellow BME lab nerdlings, both past and present: Dr. Claire Morgan, Dr. Thomas Walsh, Dr. Mark Lynch, Dr. Noeleen Loughran, Kate Lee, Ann Mc Cartney, Ray Moran, and Dr. Edel Hyland, for being amazing examples, all the support, scientific banter, great times, and of course friendships. I couldn't have asked for better group of lab members. And lastly, I hope our paths cross again and often.

I give sincerely thanks to Prof. James McInerney for his support throughout my time at DCU. Thank you for all your advice and suggestions on my research. And of course, thank you for some truly amazing scientific discussions.

I would also like to thank all my collarborators: Dr. João Pedro de Magalhães (University of Liverpool), Prof. Christine Loscher (DCU), Dr. Scott Edwards (Harvard), and Prof. Heather Ruskin (DCU).

I honestly can't thank all my friends enough for making my life outside the lab so amazing and reminding me there's more to life than research. In particular, I must thank Amy, Lisa, Mark, and Paul for being my Irish family and making my time in Ireland so unforgettable. I couldn't have asked for better friends, which made leaving Ireland so difficult.

Finally, I must acknowledge my family for supporting my dreams – even the crazier ones – throughout my life. Thank you for being such incredible role models and teaching me the importance of hard work and never giving up. Thank you for always being there for me, especially when I failed. And of course, thank you for supporting me in my decision to move half way around the world.

Abbreviations

<Knn>	Average degree of nearest neighbor
ACHE	Acetylcholinesterase (Yt blood group)
ADIPOQ	Adiponectin, C1Q and collagen domain containing
APOA1	Apolipoprotein A-I
AQUA	Automated quality improvement for multiple sequence alignments
ATG9A	Autophagy related 9A
AZIN2	Antizyme inhibitor 2
BCAR1	Breast cancer anti-estrogen resistance 1
BEB	Bayes Empirical Bayes
BF	Complement factor B
BIC	Bayesian Information Criterion
BLAST	Basic Local Alignment Search Tool (DNA and protein sequence similarity searchin
bmeTools	Bioinformatics and Molecular Evolution Tools
C12orf68	Chromosome 12 open reading frame 68
C19orf38	Chromosome 19 open reading frame 38
C1inh	C1-inhibitor
C1Q	Complement component 1, q subcomponent
C1R	Complement component 1, r subcomponent
C1RL	Complement component 1, r subcomponent-like
C1S	Complement component 1, s subcomponent
C2	Complement component 2
C22orf15	Chromosome 22 open reading frame 15
C3	Complement component 3
C3orf49	Chromosome 3 open reading frame 49
C4	Complement component 4
C4BPA	Complement component 4 binding protein, alpha
C5	Complement component 5
C6	Complement component 6
C7	Complement component 7
C8	Complement component 8
C8B	Complement component 8, beta polypeptide
C9	Complement component 9

CARD6	Caspase recruitment domain family, member 6
CCDC181	Coiled-coil domain containing 181
CCDC88A	Coiled-coil domain containing 88A
CD200	CD200 molecule
CD22	CD22 molecule
CD63	CD63 molecule
CDEs	Complete domain events
CDSs	Coding sequence
CFH	Complement factor H
CGEs	Composite gene events
CRHBP	Corticotropin releasing hormone binding protein
CSF2RB	Colony stimulating factor 2 receptor, beta, low-affinity (granulocyte-macrophage)
CTSD	Cathepsin D
d.f.	Degrees of Freedom
DFI	Dynamic flexibility index
DHX9	DEAH (Asp-Glu-Ala-His) box helicase 9
DMP1	Dentin matrix acidic phosphoprotein 1
Dn	Non-synonymous substitutions per non-synonymous site
Ds	Synonymous substitutions per synonymous site
ω	Dn/Ds
DNA	Deoxyribonucleic acid
DUOX1	Dual oxidase 1
ECSIT	ECSIT signalling integrator
EIF2AK2	Eukaryotic translation initiation factor 2-alpha kinase 2
ERBB2IP	Erbb2 interacting protein
ESE	Exonic Splice Enhancer
ESPL1	Extra spindle pole bodies homolog 1 (S. Cerevisiae)
F12	Coagulation factor XII (Hageman factor)
FASTA	Fast All (DNA and protein sequence similarity searching)
FIP1L1	Factor interacting with PAPOLA and CPSF1
GAPT	GRB2-binding adaptor protein, transmembrane
gBGC	GC-biased gene conversion
Gene A	human, chimp, rat, and mouse orthologs of gene A

Gene B	human, chimp, rat, and mouse orthologs of gene B
Gene M	human mosaic of human A gene and human B gene
Gm15440	Predicted gene 15440
GO	Gene Ontology
GPLD1	Glycosylphosphatidylinositol specific phospholipase D1
GRN	Granulin
HGMD	Human Gene Mutation Database
HIV-1	Human immunodeficiency virus
HMMER	Biological sequence analysis using profile hidden Markov models
HSV	Herpes simplex virus
IDEs	Incomplete domain events
IFI30	Interferon, gamma-inducible protein 30
IFIT2	Interferon-induced protein with tetratricopeptide repeats 2
IFNGR1	Interferon gamma receptor 1
IFNGR2	Interferon gamma receptor 2 (interferon gamma transducer 1)
IGF1R	Insulin-like growth factor 1 receptor
IL1RAPL2	Interleukin 1 receptor accessory protein-like 2
IL2RB	Interleukin 2 receptor, beta
IL4R	Interleukin 4 receptor
INPP5D	Inositol polyphosphate-5-phosphatase, 145kda
IQCJ-SCHIP1	IQCJ-SCHIP1 readthrough
IRAK1	Interleukin-1 receptor-associated kinase 1
IRAK4	Interleukin-1 receptor-associated kinase 4
IRF3	Interferon regulatory factor 3
IRF5	Interferon regulatory factor 5
IRF5	Interferon regulatory factor 5
IRF7	Interferon regulatory factor 7
IRF9	Interferon regulatory factor 9
ITGAM	Integrin, alpha M (complement component 3 receptor 3 subunit)
K	Degree
KXD1	Kxdl motif containing 1
LBP	Lipopolysaccharide binding protein

LGALS3	Lectin, galactoside-binding, soluble, 3
LIME1	Lck interacting transmembrane adaptor 1
lnL	Log Likelihood
LRRFIP1	Leucine rich repeat (in FLII) interacting protein 1
LRT	Likelihood-ratio test
LTB4R	Leukotriene B4 receptor
MAC	Membrane attack Complex
MASP1/2	Mannan-binding lectin serine peptidase 1 and 2
MBL	Mannan-binding lectin
MCL	Markov Cluster Algorithm
MCMC	Markov chain Monte Carlo
MD-2	Lymphocyte antigen 96
MetAl	MetAl (not an abbreviation)
MGF	Multi-gene family
ML	Maximum likelihood
MLEC	Malectin
MMP13	Matrix metallopeptidase 13 (collagenase 3)
MN1	Meningioma (disrupted in balanced translocation) 1
MRCA	Most Recent Common Ancestor
MSA	Multiple sequence alignment
MST1R	Macrophage stimulating 1 receptor (c-met-related tyrosine kinase)
MX2	MX dynamin-like gtpase 2
MYD88	Myeloid differentiation primary response 88
N50	50% of assembly is contained in contigs or scaffolds equal to or larger than the value.
NCBI	National Center for Biotechnology Information
NCF1	Neutrophil cytosolic factor 1
NDC80	NDC80 kinetochore complex component
Ne	Effective population size
NEB	Naïve Empirical Bayes
NFκB	Nuclear factor kappa-light-chain-enhancer of activated B cells
NLRP1	NLR family, pyrin domain containing 1
NLRP14	NLR family, pyrin domain containing 14

NLRP5	NLR family, pyrin domain containing 5
NLRP6	NLR family, pyrin domain containing 6
NLRP8	NLR family, pyrin domain containing 8
NLRP9	NLR family, pyrin domain containing 9
noRMD	Normalized Mean Distance
NOS2	Nitric oxide synthase 2, inducible
NUP153	Nucleoporin 153kda
NUP214	Nucleoporin 214kda
OAS2	2'-5'-oligoadenylate synthetase 2, 69/71kda
OLFM4	Olfactomedin 4
Olfr1260	Olfactory receptor 1260
OTUB2	OTU deubiquitinase, ubiquitin aldehyde binding 2
PAMP	Pathogen-associated molecular pattern
PIGV	Phosphatidylinositol glycan anchor biosynthesis, class V
PLCG2	Phospholipase C, gamma 2 (phosphatidylinositol-specific)
PML	Promyelocytic leukemia
PPAN-P2RY11	PPAN-P2RY11 readthrough
PRR	Pattern recognition receptor
PRR5-ARHGAP8	PRR5-ARHGAP8 readthrough
PTK2	Protein tyrosine kinase 2
PTK2B	Protein tyrosine kinase 2 beta
PTPN2	Protein tyrosine phosphatase, non-receptor type 2
QC	Quality control
RASGEF1B	Rasgef domain family, member 1B
REL	V-rel avian reticuloendotheliosis viral oncogene homolog
REST	RE1-silencing transcription factor
RI	Replacement invariable
RNF31	Ring finger protein 31
RSAD2	Radical S-adenosyl methionine domain containing 2
RUNX3	Runt-related transcription factor 3
RV	Replacement variable

SERPING1	Serpin peptidase inhibitor, clade G (C1 inhibitor), member 1
SGO	Single gene orthologs
SI	Silent invariable
SIRPA	Signal-regulatory protein alpha
SIRT1	Sirtuin 1
SLC15A4	Solute carrier family 15 (oligopeptide transporter), member 4
SNAP23	Synaptosomal-associated protein, 23kda
SNP	Single-nucleotide polymorphism
SPP2	Secreted phosphoprotein 2, 24kda
SRRD	SRR1 domain containing
STAT2	Signal transducer and activator of transcription 2, 113kda
SV	Silent variable
TBC1D19	TBC1 domain family, member 19
TCF4	Transcription factor 4
TICAM1	Toll-like receptor adaptor molecule 1
TIRAP	Toll-interleukin 1 receptor (TIR) domain containing adaptor protein
TLR	Toll-Like Receptor
TLR1	Toll-like receptor 1
TLR10	Toll-like receptor 10
TLR2	Toll-like receptor 2
TLR3	Toll-like receptor 3
TLR4	Toll-like receptor 4
TLR5	Toll-like receptor 5
TLR6	Toll-like receptor 6
TLR7	Toll-like receptor 7
TLR8	Toll-like receptor 8
TLR9	Toll-like receptor 9
TMEM119	Transmembrane protein 119
TOLLIP	Toll interacting protein
TRAF3	TNF receptor-associated factor 3
TRAF5	TNF receptor-associated factor 5
TRAF6	TNF receptor-associated factor 6, E3 ubiquitin protein ligase
TRAM	Tir domain-containing adapter molecule 2

TRIF	Toll-like receptor adaptor molecule 1
TRIM25	Tripartite motif containing 25
TRIM5	Tripartite motif containing 5
TRPV2	Transient receptor potential cation channel, subfamily V, member 2
TYK2	Tyrosine kinase 2
TYRO3	TYRO3 protein tyrosine kinase
VDR	Vitamin D (1,25- dihydroxyvitamin D3) receptor
ZGLP1	Zinc finger, GATA-like protein 1
ZNF646	Zinc finger protein 646
θ	Expected heterozygosity
π	Nucleotide diversity (observed)

List of Figures

List of Tables

Abstract

In immunology, the mouse is unquestionably the predominant model organism. However, an increasing number of reports suggest that mouse models do not always mimic human innate immunology. To better understand this discordance at the molecular level, we are investigating two mechanisms of gene evolution: positive selection and gene remodeling by introgression/domain shuffling. We began by creating a bioinformatic pipeline for large-scale evolutionary analyses. We next investigated bowhead genomic data to test our pipeline and to determine if there is lineage specific positive selection in particular whale lineages. Positive selection is a molecular signature of adaptation, and therefore, potential protein functional divergence. Once we had the pipeline troubleshot using the low quality bowhead data we moved on to test our innate immune dataset for lineage specific selective pressures. When possible, we applied population genomics theory to identify potential false-positives and date putative positive selection events in human. The final phase of our analysis uses network (graph) theory to identify genes remodeled by domain shuffling/introgression and to identify species-specific introgressive events. Introgressive events potentially impart novel function and may also alter interactions within a protein network. By identifying genes displaying evidence of positive selection or introgression, we may begin to understand the molecular underpinnings of phenotypic discordance between human and mouse immune systems.

Thesis Aims

(1) Automation of large-scale selective pressure variation – Chapter 2.

We wished to develop a streamlined highly automated method to improve large-scale identification of selective pressure variation and to stream line all processes associated with this type of analysis such as: (i) identification of gene families, (ii) alignment, (iii) phylogenetic reconstruction, (iv) selective pressure analyses, (v) LRTs, and (vi) quality control. We tested this software using novel sequence data from Bowhead Whale.

(2) To determine if positive selection is correlated with phenotypic discordance using the innate immune system of vertebrates as a model – Chapter 3.

We wished to determine if positive selection identified in the innate immune system correlates with known phenotypic discordance thereby providing a putative target to understand the known discordance at the molecular level. And from here we wished to predict sequence changes that may underpin currently unknown discordance cases.

3) To elucidate the role of domain shuffling in the emergence of proteins in vertebrate evolution – Chapter 4.

We wished to qualify and quantify gene remodeling by domain shuffling in vertebrate protein coding gene evolution, and to define the principles that govern how the domains of modular proteins combine to form functional units often in a species-specific manner that could lead to variation in function across species.

Chapter 1: Introduction

1.1 Innate Immunity

1.1.1 General overview of the innate immune system in vertebrates

The immune system of an organism is primarily responsible for defense against and resistance to pathogens. In vertebrates, immunity is divided into two distinct strategies: adaptive immunity and innate immunity. Adaptive immunity is unique to vertebrates and grants immunity from previously encountered pathogens by means of immunological memory [Flajnik and Kasahara, 2010]. A noteworthy aspect of immunological memory is that the survival advantage is confined to an individual [Medzhitov and Janeway, 2000]. Innate immunity evolved in the common ancestor of plants and animals and confers immunity to a wide range of pathogens by recognizing conserved characteristics or pathogen-associated molecular patterns (PAMPs) [Medzhitov and Janeway, 2000]. Innate immunity is responsible for activating multiple inflammatory responses as well as adaptive immunity [Janeway and Medzhitov, 2002]. In contrast with the adaptive immune system, the ability of the innate immune system to detect PAMPs by pattern recognition receptors (PRRs) is heritable.

1.1.2 Major proteins categories involved in innate immunity

Innate immunity involves a number of signaling pathways typically characterized by bespoke interaction networks and their proteins. Deciphering the complexities of these pathways (not to mention pathway-pathway interactions) begins by understanding the basic categories of proteins that are required in innate immune pathways. The protein categories that will be discussed in this section have been limited to those involved in intracellular pathways: PRRs, adaptor proteins, and transcription factors.

The majority of PRRs, including transmembrane and cytosolic PRRs, respond to pathogens by inducing the activation of transcription factors (e.g. NF-κB [nuclear factor kappa-light-chain-enhancer of activated B cells] and IRFs [interferon regulatory factors]) and can activate the adaptive immune response [Mogensen, 2009; Iwasaki and Medzhitov, 2010]. Typically, the response is dependent on the classification of the PRR in question: secreted, transmembrane, or cytosolic [Mogensen, 2009; Iwasaki and Medzhitov, 2010]. Secreted PRRs are responsible for inducing opsonization for phatgocytosis [Iwasaki and Medzhitov, 2010] but are unable to directly activate the adaptive immune response without assistance.

Detection of PAMPs by transmembrane and cytosolic PRRs and subsequent signal transduction requires various adaptor proteins to generate the required immunological response [Jordan *et al.,* 2003]. In innate immunity, a number of adaptor proteins are responsible for encoding binding domains that recognize PAMP-activated PRRs [Jordan *et al.,* 2003]. PRR-bound adaptors (i.e. activated adaptors) are required to facilitate the formation of a protein complex that either generates the required immunological response or binds and activates another adaptor for subsequent rounds of protein complex formation [Pawson, 2007]. It should be noted that many innate immune pathways require multiple rounds of adaptor activation to generate an immune response [Jordan *et al.,* 2003].

Activation of adaptor proteins by transmembrane and cytosolic PRRs eventually forms protein complexes that are responsible for activating a variety of

transcription factors to induce the appropriate immunological responses. The role of activated transcription factors in innate immunity (such as NF-κB and IRFs) is typically to regulate the expression of specific proteins involved in cellular signaling, including the well-documented cytokines [Caamaño and Hunter, 2002; Tamura *et al.,* 2008]. Activation of NF-κB for example leads to regulating the expression of proteins involved in: apoptosis (both inhibitors and activators), immune cell development and function, inflammatory response, and also triggers the adaptive immune system [Caamaño and Hunter, 2002].

1.1.3 Central innate immune proteins and pathways

The innate immune system incorporates a diverse collection of pathways that provide immunity to a range of pathogens (bacteria, parasites, and viruses). For the purpose of this thesis two of the pathways will be reviewed in detail here: the Toll-like Receptor (TLR) signaling pathways and the complement system. These are the two pathways that feature in Chapter 3.

1.1.3.1 The TLR signaling pathways

The TLRs are a well-studied family of hetero/homo-dimeric transmembrane PRRs [Moresco *et al.,* 2011]. It is currently thought that there are eleven TLRs encoded in the human genome, ten of which are expressed (TLRs 1-10) (Table 1.1). The functions of human TLR8 and 10 are still not fully understood, however it has been reported that TLR10 is involved in the innate immune response to influenza infection [Lee *et al.,* 2014] and TLR8 is involved in recognizing RNA of both viral and bacterial origin [Cervantes *et al.,* 2012]. Human TLR11 is not expressed due to pseudogenization [Zhang *et al.,* 2004].

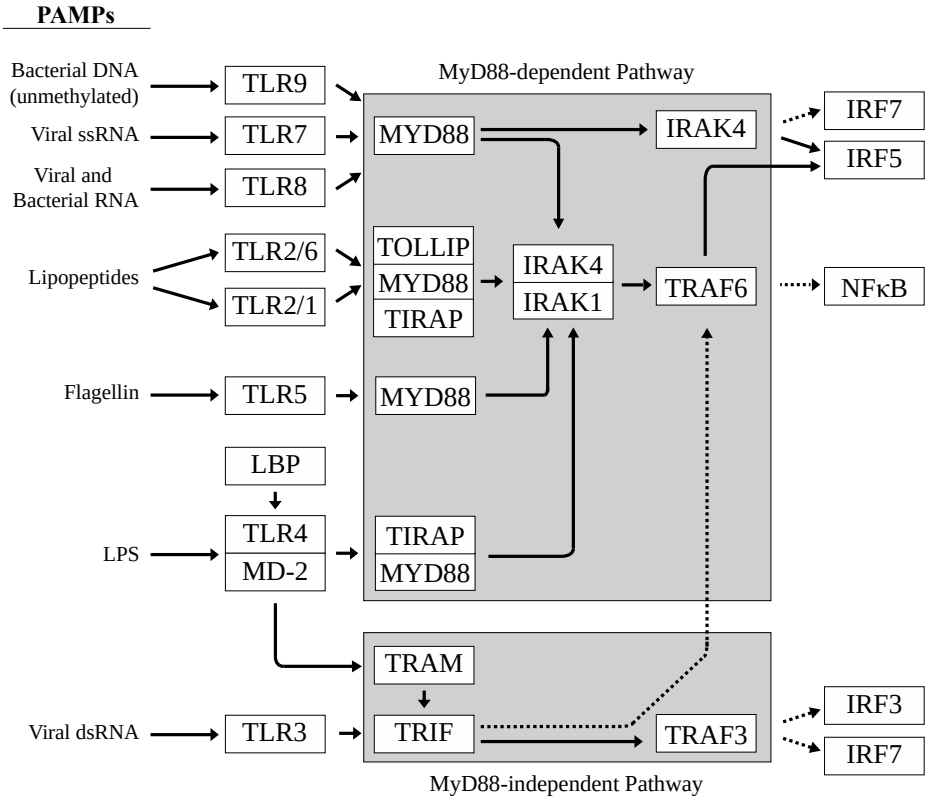
Signal transduction of the TLR signaling pathway requires various adaptor proteins and results in the activation of specific transcription factors to activate the appropriate response (Figure 1.1) [Kanehisa and Goto, 2000]. Of the adaptors involved in the TLR signaling pathway, MyD88 (myeloid differentiation primary response 88) and TRIF (TIR-domain-containing adapter-inducing interferon-β) are of particular importance. MyD88 and TRIF are central adaptors that are required to activate the MyD88-dependent and MyD88-independent pathways, respectively. The MyD88-dependent pathway is responsible for NF-κB activation and inflammatory cytokine release for TLRs 5, 7, 9 and TLR2 (as a heterodimer with TLR1/6). The MyD88-independent pathway (TRIF) is responsible for NF-κB activation and cytokine release for TLR3 alone. Of particular note, TLR4 (which is required to form a complex with MD-2 [Lymphocyte antigen 96] and LBP [Lipopolysaccharide binding protein] for activation) is unique in its ability to activate both pathways with MyD88 responsible for cytokine release and early phase NF-κB activation and TRIF for late phase NF-κB activation [Kawai and Akira, 2007].

Table 1.1: Properties of the functional human TLRs.

Receptor(s)	Dimer Type	Cellular Localization	PAMP	Pathogen(s) Detected	Additional Ligands
TLR2/TLR1	Heterodimeric	Cell Surface	Lipopeptides	Gram-positive bacteria and Fungi	
TLR2/TLR6	Heterodimeric	Cell Surface	Lipopeptides	Gram-positive bacteria and Fungi	
TLR3	Homodimeric	Cell Surface & Intracellular (Endosome)	dsRNA	Viruses	poly I:C
TLR4	Homodimeric	Cell Surface	LPS	Gram-negative bacteria	
TLR5	Homodimeric	Cell Surface	Flagellin	Bacterial flagellum	
TLR7	Homodimeric	Intracellular (Endosome)	ssRNA	Viruses	Resiquimod, Imiquimod, and Loxoribine
TLR8	Homodimeric	Intracellular (Endosome)	ssRNA	Viruses & Bacteria	
TLR9	Homodimeric	Intracellular (Endosome)	Unmethylated DNA	Bacteria	CpG-oligodeoxynucleotides (CpG-ODNs)
TLR10	Homodimeric	Intracellular	Unknown	Virus (influenza)	

The various TLR receptors of humans, the dimer type of the receptor, the cellular localization of each receptor (information from [Kawai and Akira, 2010; Matsumoto *et al.*, 2011; Cervantes *et al.*, 2012; Lee *et al.*, 2014]), the PAMP recognized by each receptor (information from [Moresco *et al.*, 2011; Cervantes *et al.*, 2012]), the pathogen recognized by each receptor (information from [Moresco *et al.*, 2011; Cervantes *et al.*, 2012; Lee *et al.*, 2014]), and any additional molecular ligands of the receptor (reviewed in [Moresco *et al.*, 2011]).

Figure 1.1: Pathway map of the TLR signaling pathway



Simplified pathway map of the TLR signaling pathway (adapted from the KEGG pathway map of the TLR signaling pathway [Kanehisa and Goto, 2000]). Proteins are depicted as white boxes and interactions as arrows; dashed arrows indicate that some interactions are not shown for legibility. The PAMPs detected by each TLR receptor are shown to the left of their respective receptor (obtained from [Moresco *et al.*, 2011; Cervantes *et al.*, 2012]). The pathway is divided into two distinct pathways: the MyD88-dependent pathway and the MyD88-independent pathway. For example, Lipopolysaccharides (LPS) activate TLR4. Activated TLR4 subsequently activates: i) a complex of MyD88 and TIRAP (toll-interleukin 1 receptor domain containing adaptor protein) to activate the MyD88-dependent pathway (resulting in activation of NF-κB and IRF5) and ii) TRIF activates the MyD88-independent pathway (results in activation of IRF3 and IRF7). See table of abbreviations for the name of each protein shown above.

1.1.3.2 Vertebrate TLR repertoires

Drosophila Toll was the first member of the TLRs to be described and it was found to function in embryonic dorsoventral regulation rather than immunity. Later it was discovered that in the presence of the pathogen-recognition adaptor gene, spätzle, a potent antifungal immune response was observed [Lemaitre *et al.,* 1996].

Comparative studies of vertebrate TLRs have found that both the recognition of and response to PAMPs has remained conserved [Roach *et al.,* 2005]. Indeed, evidence of functional conservation has been reported in zebrafish [Purcell *et al.,* 2006], highlighting the ancient and conserved function of TLRs throughout nearly 400 million years of vertebrate evolution [Hedges *et al.,* 2006].

In addition to this functional conservation, vertebrates exhibit numerous instances of clade and/or species-specific TLR family expansion/contraction. For example, Zebrafish encode 17 putative TLR variants, including orthologs for mammalian TLR2-5 and 7-9, the remaining proteins represents divergences of the zebrafish TLR repertoire. For example, zebrafish TLR1 and 18 are the only TLR homologs reported for the mammalian TLR cluster of TLR1, 6, and 10. The zebrafish repertoire also includes multiple duplicated TLRs (4, 5, 8, and 20) and a cluster of fish specific TLRs (19, 20a/b, 21, and 22) [Jault *et al.,* 2004; Meijer *et al.,* 2004]. Similar studies in chicken have revealed the presence of 10 TLRs, with orthologs to mammalian TLR2, 3, 4, 5, and 7 [Alcaide and Edwards, 2011]. Chickens encode two lineage-specific duplications, TLR2A/2B and TLR1A/1B, TLR2A and 2B are homologous to mammalian TLR2, however, TLR1A and 1B

are unique to birds. Of the remaining chicken TLRs, TLR15 is unique to birds and TLR21 is orthologous to zebrafish TLR21. The significance of TLR repertoire divergences is the potential establishment of species/clade-specific innate immunity such as the zebrafish-specific TLR4a/b duplicates that are functionally divergent to mammalian TLR4 [Sullivan *et al.,* 2009].

The Mouse genome encodes thirteen TLRs, twelve of which are functionally expressed: TLR1-9 and 11-13, with TLR1-9 having direct orthologs in human [Roach *et al.,* 2005]. Mouse TLR11-13 do not have any human homologs, rather they share common ancestry with TLR21 of fish and birds. Mouse TLR10, is nonfunctional due to a species-specific retroviral insertion [Hasan *et al.,* 2005].

Phylogenetic reconstruction places several of the TLRs into two clusters: the endocellular TLR cluster (7, 8, and 9) and the heterodimeric TLR cluster (1, 2, 6, and 10) [Roach *et al.,* 2005]. An additional TLR cluster (TLR11) has been documented but lacks a functional human homolog due to the pseudogenization of human TLR11 [Roach *et al.,* 2005; Zhang *et al.,* 2004]. Considering the prevalent nature of divergent TLR repertoires combined with an affinity for functional divergence the TLRs represent an excellent case study for those interested in molecular mechanisms of protein evolution.

1.1.3.3 The complement system

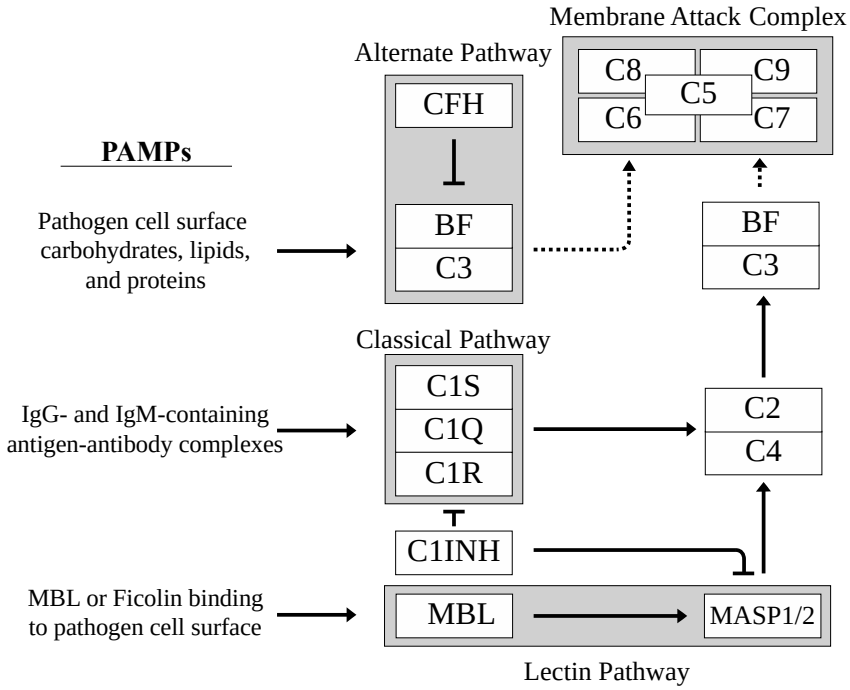
The complement system is a network of proteins involved in host defense and inflammation [Sarma and Ward, 2011]. Activation of the complement system is known to occur through three distinct pathways: the alternative, classical, and

lectin pathways (Figure 1.2) [Kanehisa and Goto, 2000]. The alternative pathway is triggered by C3b (a complex of C3 [complement component 3] and Factor B) binding to carbohydrates, lipids, and proteins found the surface of a variety of pathogens [Sarma and Ward, 2011]. The classical pathway responds to IgG- and IgM-containing antigen-antibody complexes via the C1 complex (C1q [complement component 1, q subcomponent], C1r [complement component 1, r subcomponent], and C1s [complement component 1, s subcomponent]) binding to the Fc portion of IgG and IgM [Sarma and Ward, 2011]. The lectin pathway activates the complement system via mannose-binding lectin (MBL) or Ficolin binding to the surface of pathogens [Sarma and Ward, 2011]. Each pathway results in the formation of the membrane attack complex (MAC) on the cellular surface of the pathogen. The MAC acts as a stable transmembrane pore that leads to lysis of the pathogen [Ehrnthaller *et al.*, 2011]. The complement system is also known to induce opsonization, inflammation, and phagocyte migration, and also to activate the adaptive immune system [Ehrnthaller *et al.*, 2011].

1.1.4 Discordant innate immune responses

On March 13th 2006, six healthy volunteers were administered TGN1412, an immunomodulatory drug developed by TeGenero Immuno Therapeutics for combating autoimmune diseases and leukaemia [Suntharalingam *et al.*, 2006]. Within the next two hours, all six volunteers began to exhibit severe symptoms (e.g. nausea, vomiting, drop in blood pressure, etc.) and eventually multiple organ failure [Suntharalingam *et al.*, 2006]. While all volunteers survived the ordeal, an investigation was launched to determine how the potential lethality of

Figure 1.2: Pathway map of the complement system



Simplified pathway map of the complement system (or cascade) (adapted from the KEGG pathway map of the complement system [Kanehisa and Goto, 2000]). Proteins are depicted as white boxes and interactions as arrows; dashed arrows indicate that some interactions are not shown for legibility. The PAMPs detected by each TLR receptor are shown to the left of their respective receptor (obtained from [Sarma and Ward, 2011]). The pathway is divided into three distinct pathways: the alternative pathway, lectin pathway, and classical pathway. Each pathway results in activation of the membrane attack complex (MAC). For example, the C1 (complement component 1) complex activates the classical pathway in response to IgG- and IgM-containing antigen-antibody complexes. The activated C1 complex then leads to the activation of the MAC to combat the pathogen. See table of abbreviations for the name of each protein shown above.

TGN1412 had gone unnoticed [Attarwala, 2010]. Investigation of TGN1412 found the drug to be superagonist that led to a rapid release of proinflammatory cytokines or a "cytokine storm" in human patients [Suntharalingam *et al.,* 2006; Stebbings *et al.,* 2007]. Animal trials had been completed prior to clinical trial in cynomolgus monkeys (crab-eating macaque) and there had been no evidence of superagonism or cytokine storm. It was concluded that the white blood cells of cynomolgus monkeys were unable to properly mimic TGN1412 response in humans [Stebbings *et al.,* 2007].

TGN1412 presents a worst-case scenario for model organism discordance - a biological response in the model organism that does not mimic human [Mestas and Hughes, 2004]. Traditionally, researchers respond to discordance by selecting more suitable model organisms [Davis, 2008]. While such solutions may be required for expediting research, they ignore the underlying molecular cause. Only by understanding discordance will we truly understand the biology of model organisms.

1.1.4.1 Discordance reported in TRIM5α

TRIM5α represents one of the most prominent and frequently cited instances of model discordance in the innate immune system [Stremlau *et al.,* 2004]. Initially characterized as conferring resistance to human immunodeficiency virus-1 (HIV-1) in rhesus macaque, TRIM5α belongs to the TRIM family of proteins, a group of RING proteins containing the ubiquitin-protein isopeptide ligase RBCC motif [Stremlau *et al.,* 2004; Short and Cox, 2006]. Restriction of HIV-1 by TRIM5α in rhesus macaque is dependent on the SPRY domain, which is capable of

recognizing the HIV-1 capsid protein [Sawyer *et al.,* 2005; Stremlau *et al.*, 2005; Yap *et al.,* 2005; Pertel *et al.,* 2011]. A comparative analysis identified two possible causative polymorphisms for the discordant phenotype between human (susceptible) and macaque (resistant), a single polymorphic residue P332R and a variable region of eight rhesus macaque and six human residues at position 335. Replacing either human variant with the rhesus macaque equivalent sequence conferred resistance to HIV-1 in human [Sawyer *et al.,* 2005; Stremlau *et al.*, 2005; Yap *et al.*, 2005].

1.1.4.2 Discordance reported in the Toll-like Receptors

Since discovery of the first human TLR in 1997 [Medzhitov *et al.,* 1997], the scientific literature has reported a large number of studies in which non-human TLRs produced unexpected phenotypic responses. Concern over these reports has spawned a number of review articles on the subject of unique phenotypes [Rehli, 2002; Mestas and Hughes, 2004; Werling, 2009]. Beyond differences in TLR repertoires, discordant responses have been reported in TLR2, TLR3, TLR4, TLR5, TLR8, and TLR9. The discordant responses of TLR5 and TLR9 are minimal and are mostly attributed to flagellin and CpG-ODNs sequence prefereces respectively [Bauer *et al.,* 2001; Rankin *et al.,* 2001; Pontarollo *et al.,* 2002; Andersen-Nissen *et al.,* 2007; Keestra *et al.,* 2008]. Phenotypic discordance between human and mouse for TLR2 (as a heterodimer) is attributed to the unique ability of mouse TLR2 to respond to tri-lauroylated lipopeptide (Lau_3CSK_4) [Yamamoto *et al.,* 2002; Grabiec *et al.,* 2004]. These are the more mild phenotypic discordances reported, however those involving TLR8, TLR4, and TLR3 are more signifcant.

Human TLR8, but not mouse TLR8, has been reported as capable of conferring NF-κB activation in response to RNA ligands, imidazoquinoline resiquimod (R848), and the immunostimulant derivatives CLO95 (imidazoquinoline resiquimod) and CL075 (thiazoloquinolone) in the absence of polyT-ODN [Jurk *et al.,* 2002; Forsbach *et al.,* 2008; Liu *et al.,* 2010]. Such findings led to the conclusion that TLR8 was non-functional in mouse [Cervantes *et al.,* 2012]. Subsequent sequence comparisons and deletion experiments identified two potentially causative motifs, RQSYA and PGIQ, both of which were missing in rodents [Liu *et al.,* 2010]. Reports suggest that activation of mouse TLR8 by CL075 is possible, but requires the addition of polyT-ODNs to confer activation [Liu *et al.,* 2010].

TLR4 is responsible for recognition of LPS from gram-negative bacteria and was one of the first TLRs to be characterized in vertebrates [Medzhitov *et al.,* 1997; Moresco *et al.,* 2011]. Subsequent research identified that TLR4 is required to associate with the protein MD2 to recognize the hydrophobic domain of LPS (i.e. Lipid A) [Shimazu *et al.,* 1999; Raetz and Whitfield, 2002]. Various LPS molecules have been reported to elicit discordant immune responses, a small number of examples are detailed here for illustrative purposes. Lipid IV_A (a synthetic Lipid A precursor) and LpxL1 LPS of *Neisseria meningitides* are able to induce an innate immune response in mouse but not human TLR4/MD2. The cause of Lipid IV_A discordance was determined to be mutations in both TLR4 and MD2, whereas LpxL1 discordance was dependent on TLR4 mutations alone [Steeghs *et al.,* 2008; Meng *et al.,* 2010]. The LPS molecules of msbB (mutated msbB *E. coli* [Somerville *et al.,* 1996]) and *Porphyromonas gingivalis* are also

documented to elicit species-specific discordance, both LPS molecules antagonize normal LPS based TLR4/MD2 induction in humans whereas they induce an immune response in mouse TLR4/MD2 [Coats *et al.,* 2007]. Human TLR4 but not mouse TLR4 produces an immune response upon recognizing nickel (Ni^{2+}), resulting in human-specific contact hypersensitivity (CHS). This discordant response to nickel was subsequently attributed to the histidine residues H456 and H458 in human TLR4 [Schmidt *et al.,* 2010].

TLR3 is documented to localize to the cellular surface and endosomal compartments and is responsible for recognizing the double-stranded RNA (dsRNA) of viruses [Alexopoulou *et al.,* 2001; Matsumoto *et al.,* 2003]. Recognition of the immunostimulant poly(I:C) has been reported to elicit a discordant response, albeit minor. Mouse macrophages but not human macrophages have been documented to induce TNFα and activate NFκB and IRF-3 in response to poly(I:C) [Fortier *et al.,* 2004; Lundberg *et al.,* 2007]. Beyond poly(I:C) discordance, human TLR3 has been reported to exhibit a restricted anti-viral role in resisting herpes simplex virus encephalitis, whereas mouse TLR3 exhibits broad functionality [Ariffin and Sweet, 2013].

1.1.4.3 Variations reported in inflammation

Inflammation is a critical response triggered by the innate immune system upon infection. The suitability of mouse as a model of inflammation has recently been questioned [Seok *et al.,* 2013]. In the study by Seok *et al.* the gene expression profiles for inflammatory responses due to trauma, burns, and endotoxemia were compared between human and mouse. The results showed little correlation

between the species in terms of gene expression profiles for the same condition [Seok *et al.,* 2013]. They also found no correlation in expression at the pathway level [Seok *et al.,* 2013]. These findings have since been contested. Using the same dataset a strong correlation was found between the inflammation gene expression profiles of human and mouse [Takao and Miyakawa, 2014]. The discrepancy between these reports can be attributed to different standards for data inclusion in the correlation analysis [Takao and Miyakawa, 2014]. At present, the validity of both reports is under question and it has been suggested that an additional study is warranted [Leslie, 2014].

1.1.5 Understanding and predicting model species discordance

The various examples presented in Section 1.1.4 highlight a problem often neglected by biologists; model organisms occasionally are unable to mimic human biology. The remainder of the introduction reviews two distinct evolutionary methodologies with the potential to identify and predict the genetic causes of species discordance. Section 1.2 reviews the application of selective pressure analysis to infer potential functional discordance from protein coding genes under positive selection. Section 1.3 reviews the application of network theory to identify introgressive (non-vertical inheritance) that may attribute to functional discordance.

1.2 Natural Selection and Molecular Evolution

1.2.1 Evolutionary theory

The theory of natural selection was first proposed as the gradual process by which the traits of a population change in frequency depending on their impact on reproductive success [Darwin, 1859]. In the modern era, the study of molecular evolution has sought to understand the process of evolution (including natural selection) from the sequences of organisms and this has shown that saltational events also occur. While the relative roles for genetic drift and selection have been debated in molecular evolutionary biology [Lynch, 2007; Hahn, 2008], the modern synthesis is well developed to describe the processes and patterns we observe in molecular sequence data.

1.2.2 Neutral theory

The neutral theory of evolution postulates that the majority of molecular variations observed within a genome are caused by random genetic drift of neutral alleles rather than natural selection [Kimura, 1968]. Neutral alleles are defined as variation that has no fitness effect on the organism and were believed to be responsible for the vast majority of variation within populations and between species [Kimura, 1968]. Early adoption of the neutral theory was aided by studies demonstrating the highly polymorphic nature of DNA within populations and between species, evidence that could not be explained by adaptive evolution alone [Zuckerkandl and Pauling, 1965; Harris, 1966; Lewontin and Hubby, 1966]. The theory was later expanded upon by the inclusion of nearly neutral mutations, accounting for slightly advantageous or

deleterious mutations that may become fixed within a population due to random genetic drift [Ohta, 1973; Ohta and Gillespie, 1996].

The probability of neutral or nearly neutral mutations becoming fixed within a diploid population with an effective population size N_e due to random genetic drift is shown in the following equation (Equation 1.1).

Equation 1.1: Probability of fixed of neutral mutations within a population.

$$P_x = \frac{1}{2N_e}$$

As the equation states, the probability of neutral or nearly neutral mutations becoming fixed within a population increases with smaller N_e [Kimura, 1968].

1.2.3 Natural Selection

Natural selection was postulated by Darwin to influence the frequency of particular phenotypes within a population depending on their impact on reproductive success [Darwin, 1859]. From a modern molecular standpoint, natural selection may be subdivided into distinct categories: i) positive selection, whereby an advantageous spontaneous mutation increases in frequency within a population and ii) purifying selection, whereby a deleterious mutation decreases in frequency. In the absence of selection, frequency is dependent on neutral evolution and therefore dependent on random genetic drift (and of course *Ne*).

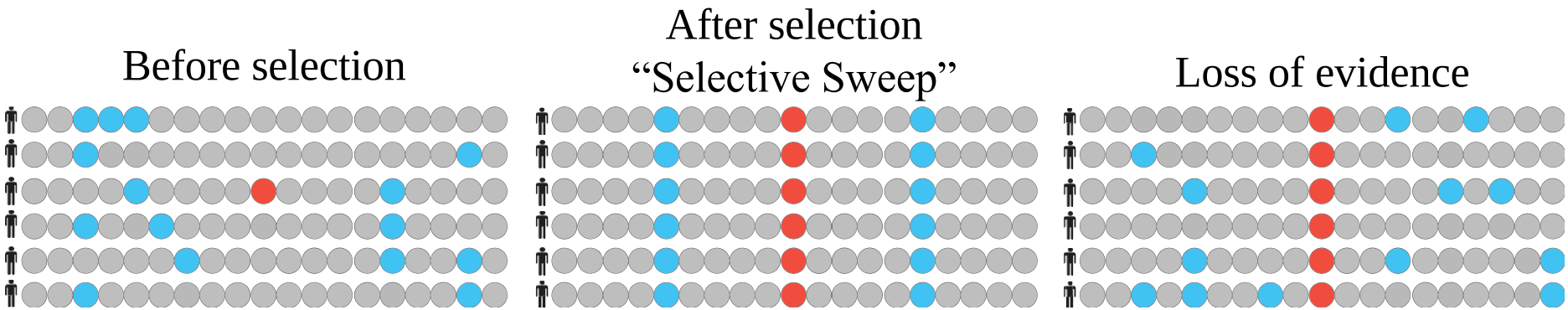
By comparing homologous sequences across populations or species (See Section 1.2.4 for details) it is possible to determine the selective pressures that have acted

upon individual protein coding genes (Section 1.3 the section on how this is done). In general, it is presumed that the majority of the sequence within protein coding genes is evolving under purifying selection due to functional constraints [Hughes, 1999; Peterson *et al.,* 2009]. Nonetheless, studies have identified high levels of positive selection among the protein coding genes of Drosophila [Smith and Eyre-Walker, 2002; Begun *et al.,* 2007], *E. coli* [Charlesworth and Eyre-Walker, 2006], and mammals [Kosiol *et al.,* 2009]. These findings have prompted some to state that the nearly neutral theory may not be an appropriate description of molecular evolution and that an adaptationist regime may be a more appropriate explanation [Hahn, 2008].

1.2.4 Positive selection and functional shift

Population geneticists traditionally define positive selection as a type of natural selection in which a spontaneous mutation that confers an advantage increases in frequency within the population [Sabeti *et al.,* 2006]. In comparison to harmful or neutral spontaneous mutations, positive selection is defined by an increase in the fixation rate of an advantageous allele [Sabeti *et al.,* 2006]. Selection (and possible fixation) of the advantageous allele is documented to confer a hitchhiking effect, whereby neutral, nearly neutral, and deleterious alleles linked to the advantageous allele also increase in frequency [Smith and Haigh, 1974; Chun and Fay, 2011]. Depending on the strength of selection event, hitchhiking may lead to a notable reduction in variation in the region surrounding the advantageous allele termed a selective sweep (Figure 1.3) [Andolfatto, 2001].

Figure 1.3: Schematic for impact of selective sweep at the population level



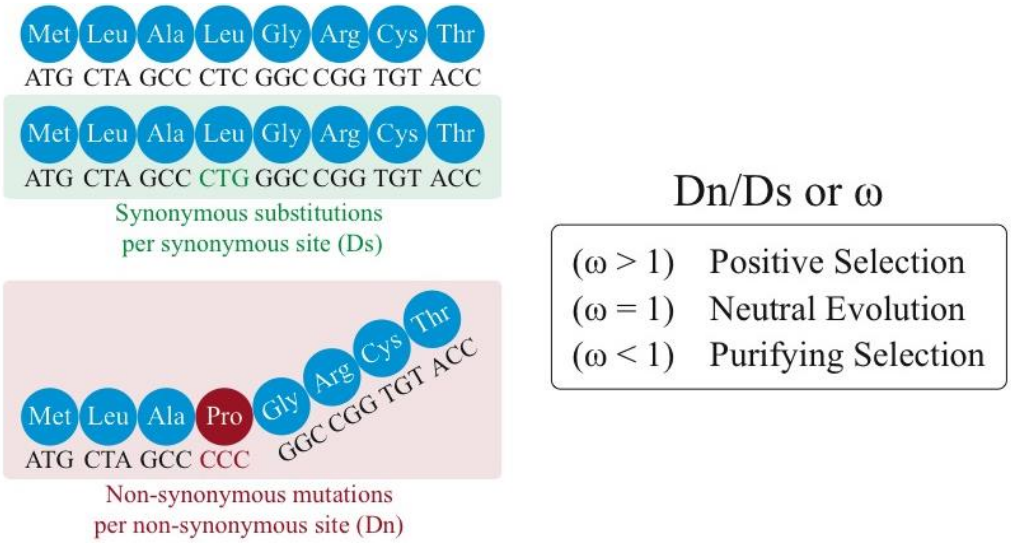
In this simplified scenario, variation in the population of size 6 is shown in blue and advantageous alleles in red. Prior to the selective sweep of an advantageous allele, the population will exhibit a number of neutral alleles (blue) surrounding the locus of the advantageous allele (red) – leftmost panel. After the positive selection and fixation of the advantageous allele in red, a hitchhiking effect will be observed whereby the linked neutral alleles in blue also increase in frequency (i.e. a decrease in variation in the region surrounding the advantageous allele) – central panel. Loss of evidence of the selective sweep will eventually occur due to the occurrence of spontaneous mutations (i.e. slow return of genetic diversity) – rightmost panel.

The various methods developed to determine selective pressure from molecular signatures are described in Section 1.3. Genetic diversity will slowly return to the region due to the occurrence of spontaneous mutations and eventually eliminate the molecular signatures of the selective sweep (this is within approximately 250,000 years for humans) [Sabeti *et al.*, 2006].

From the perspective of species-level comparisons, positive selection is often defined as the molecular signature of species adapting to their environment and for that reason has been hypothesized to be a marker of functional discordance between species [Tennessen, 2008]. The primary method for detecting positive selection between species is calculating the ratio of replacement non-synonymous substitutions per non-synonymous site (Dn) over synonymous substitutions per synonymous site (Ds) [Hurst, 2002]. This ratio of Dn/Ds, termed omega (ω) throughout the thesis has three potential outcomes: i) an $\omega > 1$ is indicative of positive selection, ii) $\omega = 1$ is indicative of neutral evolution, and iii) $\omega < 1$ is indicative of purifying selection (Figure 1.4).

Concerns have been raised over attempts to associate positive selection with functional discordance [Hughes, 2007; Yokoyama, 2008]. It is now widely held that positive selection requires *in-vitro* confirmation by the reconstruction and rational mutagenesis of ancestral proteins [Yokoyama, 2013]. For more details on *in-vitro* confirmation by ancestral reconstruction see Section 1.4.4. More recent studies focusing on the link between positive selection and phenotypic discordance have met with some success.

Figure 1.4: A schematic for how to measure selective pressure variation



The simplified example above shows two outcomes of mutating the fourth codon (CTC or leucine) in a protein. A synonymous substitution (shown in green) mutates the codon CTC to CTG but has no observable protein alteration. A non-synonymous substitution mutates CTC to CCC (shown in red) and results in the substitution of leucine for proline, and this mutation results in an altered conformation (illustrated as a "kink"). The consequence of the observed kink may be separated into three outcomes: neutral (i.e. no fitness effect), deleterious, or advantageous. If neutral, the non-synonymous "kink" (in addition to other neutral non-synonymous substitutions) would be fixed within a species at the same frequency as the synonymous substitutions – as the likelihood of fixation is the same for both the neutral non-synonymous and synonymous mutations – and would result in $\omega = 1$ (i.e. the gene is evolving neutrally) [Hurst, 2002]. If deleterious, the non-synonymous "kink" (in addition to other deleterious non-synonymous substitutions) would be eliminated, resulting in a higher frequency of synonymous substitutions becoming fixed and therefore $\omega < 1$ (i.e. the gene is under purifying selection) [Hurst, 2002]. If advantageous, the non-synonymous

"kink" (in addition to other advantageous non-synonymous substitutions) would be fixed more frequently than synonymous substitutions, thus resulting $\omega > 1$ (i.e. the gene is under positive selection) [Hurst, 2002]. It should be noted that the omega (ω) measurement is not without fault. For example, if a specific region of a protein is under positive selection whereas another region is under purifying selection, ω may incorrectly indicate the protein is evolving neutrally [Hurst, 2002]. In addition, the measurement does not account for the infrequency of transversions in comparison to transitions and therefore results in unrealistic Ds and Dn values [Yang and Bielawski, 2000].

A frequently cited species comparison of TRIM5α, a HIV-1 restriction factor in old world monkeys [Stremlau *et al.,* 2004], used positive selection to identify an 11- to 13-amino acid segment of the SPRY domain responsible for species-specific retroviral restriction [Sawyer *et al.,* 2005]. And a rational mutagenesis approach by Loughran *et al* found that three positively selected residues (N496, Y500, and L504) in the mammal myeloperoxidase protein were responsible for the evolution of this dual functioning enzyme capable of both peroxidation and chlorination activity [Loughran *et al.,* 2012]. Other recent studies have also shown a clear relationship between positive selection and functional discordance [Moury and Simon, 2011; Farhat *et al.,* 2013].

1.2.5 The relationship between orthologs, paralogs, and function

Homologs are genes that share common ancestry and are traditionally classified by their origin, orthologs by speciation events and paralogs by gene duplication [Fitch, 1970]. More recently, discussion centered around the long held assumptions of the ortholog conjecture, i.e. that orthologs are more conserved than paralogs in terms of sequence and function [Dessimoz *et al.,* 2012]. Questions about how orthologs should be defined were raised, for example, should it be by sequence similarly [Gabaldón *et al.,* 2009], domain architecture [Forslund *et al.,* 2011], intron structure [Henricson *et al.,* 2010], protein structure [Peterson *et al.,* 2009], or expression patterns [Huerta-Cepas *et al.,* 2011]. The debate was further complicated by research that found functional annotations more often correlate with paralogs than orthologs, thereby directly challenging the assumptions of the ortholog conjecture [Nehrt *et al.,* 2011]. Subsequent studies found that the reported functional correlation of paralogs was primarily

due to biases introduced by the use of computationally annotated Gene Ontology (GO) terms and they report that the corrected correlations supported the ortholog conjecture [Altenhoff *et al.,* 2012]. The debate on the most appropriate framework to define orthologs continues [Haggerty *et al.,* 2014].

If the ortholog conjecture holds true, paralogs are expected to exhibit greater functional divergence [Dessimoz *et al.,* 2012]. Gene duplication is the biological mechanism that generates paralogs and is hypothesized to provide an opportunity for functional divergence to occur. Gene duplication events, are predicted to frequently result in one of the duplicates becoming pseudogenized by degenerative mutations [Ohno *et al.,* 1970]. However, both duplicates may become preserved by either beneficial mutations that generate a novel function (neofunctionalization) or mutations that necessitate the fixation of both duplicates (subfunctionalization) [Force *et al.,* 1999; Ohno *et al.,* 1970].

In subfunctionalization, duplicates become fixed within a population if partial functional loss necessitates functional complementation of the daughter genes to maintain parental function [Zhang, 2003]. A recent example of subfunctionalization was reported in the H3-H4 histone chaperones ASF1a and ASF1b [Abascal *et al.,* 2013]. The duplication event that produced ASF1a and ASF1b was reported to have occurred in the ancestor of jawed vertebrates. The single ASF1 of *Saccharomyces cerevisiae* is reported to interact with HIRA and CAF-1 chaperones with equal affinity [Abascal *et al.,* 2013]. After the duplication event, ASF1a and ASF1b became divergent in their gene expression profiles and positive selection acted on both paralogs (C-terminal of ASF1b and

N-terminal of both paralogs). Considering these hallmarks of functional divergence, it may not be surprising that ASF1a and ASF1b exhibit preferential interaction affinities, ASF1a with HIRA and ASF1b with CAF-1 [Abascal *et al.*, 2013].

In neofunctionalization, a mutation that confers a novel function may cause both duplicates to become fixed by positive selection or genetic drift [Conant and Wolfe, 2008], such as in the human SRGAP2 genes [Dennis *et al.*, 2012]. SRGAP2 is a neuronal migration gene that is highly conserved in mammal evolution. Multiple "Homo"-specific gene duplication events led to three duplicates of SRGAP2A (SRGAP2B-D) in humans. These duplications had not been previously sequenced or characterized due to a misassembled SRGAP2. Subsequent analysis supported that only SRGAP2A and SRGAP2C were likely functional and that the incomplete duplication that created SRGAP2C had created a novel antagonism mechanism of parental SRGAP2A [Dennis *et al.*, 2012].

1.3 Methods for assessing selective pressure variation

1.3.1 Distance-based methods

One of the earliest approaches for assessing selective pressure variation across sites was developed by Li *et al.*, (1985). The method classified nucleotide positions in coding regions into four categories: non-degenerate sites are classified as non-synonymous (any mutation results in replacement of the amino acid), fourfold degenerate sites are classified as synonymous (all mutations are silent *i.e.* do not change the amino acid), twofold degenerate sites are classified

as synonymous for transitions, and twofold degenerate sites are classified as non-synonymous for transversions. The issue with this method is that the twofold degenerate category overestimates synonymous counts due to the infrequency of transversions in comparison to transitions and later refinements were made to improve the efficiency in this regard [Li, 1993]. This approach to assessing selective pressure has been found to be unsatisfactory as it lacks power to detect positive selection if only a few sites are under selection [Pond and Frost, 2005; Murray, 2011]. Distance-based sliding window approaches were later developed so that selective pressures across coding sequences could be classified along the length of the sequence [Comeron, 1999; Creevey and McInerney, 2003; Fares, 2004; Liang, 2006]. Sliding window based approaches have since been shown to have undesirable characteristics including the estimation of artifactual trends of synonymous and nonsynonymous rate variation and not correcting for multiple testing [Schmid and Yang, 2008].

1.3.2 Phylogeny-based methods

In comparison to the distance-based methods, phylogeny-based methods enable the assessment of selective pressure variation across lineages as well as sites [Creevey and McInerney, 2002; Yang, 2002]. The Creevey-McInerney method uses the G-test to test the hypothesis that sequences are evolving neutrally when the ratio of silent invariable (SI) sites to silent variable (SV) sites is equal to the ratio of replacement invariable (RI) sites to replacement variable (RV) sites [Creevey and McInerney, 2002]. Significant deviations indicate a departure from neutrality. The method implements a rooted phylogeny (that is assumed to be correct) to reconstruct the ancestral sequences at each internal node using a

maximum parsimony approach [Hennig, 1966; Creevey and McInerney, 2002]. The reconstructed phylogeny is then used to identify all substitutions that occurred across the tree and determine whether they resulted in a non-synonymous (replacement) or synonymous (silent) codon change. Significantly high rates – reported by the G-test – of RI are indicative of directional selection whereas significantly high rates of RV are indicative of non-directional selection [Creevey and McInerney, 2002]. The Creevey-McInerney method is effective for detecting selective pressure variation across lineages, but cannot identify sites under lineage-specific positive selection [Creevey and McInerney, 2002].

1.3.2.1 Maximum likelihood methods

Maximum likelihood (ML) methods use models of evolution to determine the likelihood of observing the experimental data given the characteristics of the specified model. A variety of methods have been developed to identify selective pressure variation under a maximum likelihood framework (in addition to methods using a Bayesian framework not specified for brevity) [Massingham and Goldman, 2005; Pond *et al.*, 2005; Yang, 2007]. The sitewise likelihood-ratio (SLR) test was an approach primarily designed to detect evidence of non-neutral evolution and estimate the likelihood of each site under being under either purifying of positive selection [Massingham and Goldman, 2005]. The program HyPhy carries out a variety of likelihood-based analyses, including the assessment of selective pressures acting on sites and lineages combined [Pond *et al.*, 2005]. The codeML program from the PAML software package was designed to identify positive selection acting on specific sites within an alignment and positive selection unique to a specific foreground lineage [Yang,

2007]. The codeML program was selected for our analysis for three reasons: i) the method is highly developed and regularly updated, ii) the robustness of the most recent lineage-site models and their appropriate null models [Yang and dos Reis, 2011], and iii) low false discovery rate of the lineage-site models in the presence of GC content deviations and indels (see Section 1.4 for details) [Fletcher and Yang, 2010; Gharib and Robinson-Rechavi, 2013].

1.3.2.2 CodeML

The codeML program from the PAML package implements a large number of codon substitution models developed to account for various substitution rate characteristics (i.e. Transition/transversion rate bias and codon usage bias) between the various amino acids, and also to estimate the Dn/Ds ratio (ω) [Goldman and Yang, 1994; Yang, 2007]. Codon substitution models have been developed that account for heterogeneous selective pressure at codon sites and that allow for ω to be estimated for a lineage or subset of lineages (referred to in the codeML literature as "foreground lineages") on the phylogeny [Yang *et al.*, 2000; Yang and Nielsen, 2002; Zhang *et al.*, 2005].

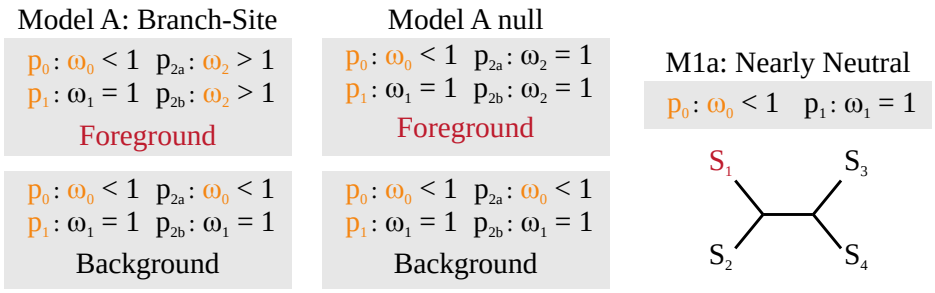
As the primary aims in chapters 2 and 3 of this thesis is the identification of species-specific positive selection, only the relevant codon substitution models will be discussed. The codon-based models of evolution implemented by codeML are nested likelihood models, indicating that models differ in complexity by the addition of free parameters (Figure 1.5). The M1a model is a nearly-neutral model where the ω values are permitted to be below neutrality ($\omega_0 < 1$), and p_0 is the proportion of sites that have that value. The remaining

proportion of sites ($p_1 = 1 - p_0$) is expected to be evolving neutrality ($\omega_1 = 1$). The branch-site (also known as the lineage-specific) model used in the analyses in Chapters 2 and 3 is the updated model A [Zhang *et al.,* 2005]. Previous versions of model A were found to have an unacceptably high rates of false positives (19-54%) when there was a relaxation of purifying selection in the foreground [Zhang, 2004; Zhang *et al.,* 2005]. The modified branch-site model A (hereafter referred to as model A) has four free parameters. Three of these parameters are the same for foreground and background lineages, p_0 = the proportion of sites with the estimated ω below neutrality ($0 < \omega_0 < 1$), and p_1 = the proportion of sites evolving neutrality ($\omega_1 = 1$). The fourth and final free parameter in model A is ω_2 which is the estimated ω in the foreground lineage alone (ω_2 is free to be > 1). The proportion of sites for ω_2 is broken into two site categories: p2a whereby foreground sites are evolving with $\omega_2 > 1$ and background sites are evolving with ω_0 between 0 and 1, and p2b whereby foreground sites are evolving with $\omega_2 > 1$ and background sites are evolving with a $\omega_1 = 1$. The null hypothesis of model A has three free parameters and is referred to throughout as model A null where $\omega_2 = 1$. Model A null allows sites in the background to be evolving neutrally ($\omega_1 = 1$) or under purifying selection ($0 < \omega_0 < 1$), Figure 1.5. The ML implementation of these models could report results from a local minimum on the likelihood plain, for this reason codeML analyses conducted in this thesis used multiple starting omega values (0, 1, 2, 10) as in previous publications to increase the likelihood of finding and reporting estimates from the global minimum [Yang, 1997; Yang *et al.,* 1998; Loughran *et al.,* 2008; Morgan *et al.,* 2010].

The likelihood ratio test (LRT) is used to determine the significance of parameter rich models by a comparison to alterative less parameter rich models [Nielsen and Yang, 1998; Yang *et al.,* 2000]. The LRT is defined as the difference between the log-likelihood (lnL) values of the two models (Δl) multiplied by two, and follows a chi-squared (χ^2) distribution [Nielsen and Yang, 1998]. The degrees of freedom (d.f.) between the two models is then used to determine the χ^2 critical value for significance. If $2\Delta l$ is larger than the χ^2 critical value then the parameter rich model is found to be significant as in Figure 1.5. If a codon substitution model is found to be significant by the LRT, the posterior probability of a specific codon being under positive selection is calculated by applying an Empirical Bayes estimate to ω for each codon. By default codeML uses two Empirical Bayes methods: i) Naïve Empirical Bayes (NEB) [Yang *et al.,* 1998] and ii) Bayes Empirical Bayes (BEB) [Yang *et al.,* 2005]. When possible, the BEB values were used in the analyses conducted in this thesis as they have been reported to be more statistically robust then NEB values particularly for smaller datasets [Anisimova *et al.,* 2002; Yang *et al.,* 2005].

Figure 1.5: Codon based models of substitution used in the analyses

a.



b.

Comparison	Null Model	FP	Alternative Model	FP	d.f.	Critical !2 values
M1a vs. Model A	M1a	2	Model A	4	2	> 5.99 (P = 0.05)
Model A null vs. Model A	Model A null	3	Model A	4	1	> 3.84 (P = 0.05)

(a) The three nested likelihood models required for detecting species-specific positive selection. The parameters for each model are specified within their respective grey boxes, fixed parameters are shown in black whereas free parameters are shown in orange (please note that for branch-site model A ω_2 is a single free parameter). The parameters for branch-site model A and model A null are shown for both the foreground in red (S_1) and background in black (S_2, S_3, and S_4) of the given phylogeny. The free parameters for the nested likelihood models are as follows: Model A with 4 (P_0, P_1, ω_0, and ω_2), Model A null with 3 (P_0, P_1, and ω_0), and M1a with 2 (P_0 and ω_0). The number of free parameters of the nested likelihood models is then used to perform a likelihood ratio test (LRT). (b) The table shows the two comparisons required to determine the significance of Model A. The degrees of freedom (d.f.) of a comparison is determined from the difference in free parameters between the two models. For example, the comparison of M1a (Null) and Model A (alterative) diffeences in 2 free parameters and therefore has a d.f. of 2. As the LRT follows a χ^2 distribution, the d.f. of the models is used to determine the χ^2 critical value for the comparison. The χ^2 critical values given in the table are for a p-value of 0.05.

1.3.2.3 Phylogenetic Reconstruction

Phylogenetic trees describe relationships caused by linear descent and may be constructed from a variety of data types, including (but not limited to): morphological, microRNAs, mitochondrial, and nuclear coding sequences [Wiens, 2004; Pedersen *et al.,* 2006; McCormack *et al.,* 2012; Morgan *et al.,* 2013; Morgan *et al.,* 2014; Yang and Rannala, 2012]. The most common modern methods for phylogeny reconstruction include Bayesian and ML approaches [Felsenstein, 1981; Yang and Rannala, 1997]. Both approaches use the likelihood function and therefore share many statistical properties [Yang and Rannala, 2012]. However, there are major differences between these approaches: (i) ML approaches identify the most probable phylogeny whereas Bayesian approaches search for the most credible trees, and (ii) Bayesian implementations allow a prior hypothesis whereas ML cannot [Huelsenbeck *et al.,* 2002; Yang and Rannala, 2012]. The results of Bayesian inference are also far easier to interpret, with the posterior probability being the support for a given node based on the data and specified model [Yang and Rannala, 2012].

Considering the focus of this thesis is to understand the mechanisms of protein evolution that underpin the formation of vertebrate proteins (namely positive selection and domain shuffling), phylogenetic trees are most suitable for detecting and explaining linear descent such as gene duplication and substitution processes over time such as adaptive evolution (Chapters 2 and 3). However, to capture gene remodeling by domain shuffling a network approach is required (Section 1.5 and Chapter 4).

1.3.4 Population-based methods

1.3.4.1 McDonald–Kreitman test

One of the earliest implementations of population level tests for selective pressure variation was the McDonald–Kreitman test [McDonald and Kreitman, 1991]. The approach uses a simple phylogeny, the advantage of which is to provide the user with the ability to distinguish the amount of variation within a species (polymorphism) from the substitutions between species. The McDonald–Kreitman test works by comparing the ratio of non-synonymous to synonymous polymorphisms within a species to the ratio of non-synonymous to synonymous fixed substitutions between species by means of a contingency table which is used to conduct a G-Test [McDonald and Kreitman, 1991]. Mutations under positive selection will fix within a population more rapidly than by random genetic drift alone [McDonald and Kreitman, 1991; Gillespie, 1998]. Adaptive evolution is therefore observed if the ratio of non-synonymous to synonymous polymorphisms within a population is lower than the ratio of non-synonymous to synonymous variation across species [McDonald and Kreitman, 1991]. It should be noted that the McDonald–Kreitman test requires population data from both species being compared and therefore due to a lack of mouse population data was not employed in this thesis, however other population level approaches are possible with the species we study and these are detailed below.

1.3.4.2 Tajima's D test statistic

Tajima's D test statistic is a population-based approach to determine if a nucleotide sequence in a population is evolving neutrally or evolving under a non-random process [Tajima, 1989]. Tajima's D requires two values to be

calculated from a multiple sequence alignment of the population: the observed nucleotide diversity (π) and the expected heterozygosity (θ). The observed nucleotide diversity is the average number of pairwise nucleotide differences within the population. The expected heterozygosity of a population is calculated using the following equation (Equation 1.2).

Equation 1.2: Expected heterozygosity of a population

$$\theta = \frac{S}{\sum_{i=1}^{n-1} \frac{1}{i}}$$

where S is the number of segregating sites, n is the number of individuals, i is the index of summation, and θ is the expected heterozygosity.

In populations evolving neutrally, the observed and expected values should be equal and therefore return a Tajima's D of approximately zero. If the population however is evolving under a non-random process, Tajima's D may result in a positive value (higher observed to expected *i.e.* possibly balancing selection or population decrease) or it may result in a negative value (lower observed to expected *i.e.* possible selective sweep or population growth) [Simonsen *et al.*, 1995; Tajima, 1989].

Tajima's D has also been applied using sliding window approaches [Rogers *et al.*, 2010]. In these approaches, Tajima's D is calculated at regular intervals from segments of sequence in the genomic region surrounding a nucleotide sequence of interest, as each window is an independent test with no overlap it does not

have the multiple testing flaw of the distance-based sliding window methods [Schmid and Yang, 2008]. In comparison to only calculating Tajima's D for a nucleotide sequence of interest, sliding windows enable the identification of significant departures from neutrality across the sequence [Rogers *et al.,* 2010].

Tajima's D is commonly tested for significance by obtaining a confidence interval for the true value of θ [Simonsen *et al.,* 1995]. Obtaining a confidence interval is achieved by generating a large number of samples under a Wright-Fisher neutral model of genetic variation [Simonsen *et al.,* 1995]. One of the most commonly used approaches to generate a confidence interval is implemented in MS where the user supplies θ, S, or θ and S [Hudson, 2002].

1.3.4.3 Fay and Wu's H test statistic

Fay and Wu's H test statistic is a population-based approach often described as an improvement on Tajima's D [Fay and Wu, 2000]. In comparison to Tajima's D statistic, Fay and Wu's H accounts for the presence of derived alleles (*i.e.* non-ancestral alleles that arose by mutation) by determining the ancestral state of alleles on a phylogenetic tree. Derived alleles are typically expected to be present at lower frequencies than ancestral alleles [Watterson and Guess, 1977]. This expectation becomes violated in the presence of positive selection, resulting in the presence of high-frequency derived alleles [Hamblin *et al.,* 2002]. In populations evolving neutrally, Fay and Wu's H is expected to be approximately zero [Fay and Wu, 2000]. However, if the population is evolving under a non-random process then Fay and Wu's H may result in a positive value (i.e. few high-frequency derived alleles) or negative value (i.e. many high-frequency

derived alleles) [Fay and Wu, 2000]. Like Tajima's D, Fay and Wu's H is a statistical test and therefore resultant values must be tested for significance. This is achieved using a similar approach as described for Tajima's D (Section 1.3.4). Tajima's D and Fay and Wu's test statistics are employed in Chapter 3 using human population genomic data from the 1000 human genomes project [Abecasis *et al.*, 2012].

1.4 Data limitations in analyses of selective pressure variation

Beyond the limitations of specific methods for measuring selective pressure variation, the data itself may have sequencing, assembly and/or alignment errors that can influence the estimates of parameters across sites and lineages [Schneider *et al.*, 2009]. The use of high quality genomes and assemblies is highly recommended for selective pressure analysis (see Section 2.12). The following sections briefly detail some of the major sources of potential error in a selective pressure analysis.

1.4.1 Alignment Error

Alignment error is reported to cause unacceptably high rates of false positives when using model A in codeML [Fletcher and Yang, 2010]. This source of error is not from aligned insertions and deletions but rather from poorly aligned codons [Fletcher and Yang, 2010]. It is advisable therefore to use a variety of alignment methods for a given dataset and independently assess which is the best alignment for the data [Muller *et al.*, 2010]. Programs such as MetAl [Blackburne and Whelan, 2012], AQUA (Automated quality improvement for multiple sequence alignments) [Muller *et al.*, 2010], and NorMD (Normalized

Mean Distance) [Thompson *et al.*, 2001] provide a way of comparing various alignment methods and selecting the most appropriate method for a given multiple sequence alignment (MSA). A combination of the methods MetAl and NorMD was implemented in Chapter 2.

1.4.2 Non-adaptive evolutionary signals mistaken as positive selection

Recombination is the process by which nucleotide sequences exchange genetic information and has been reported to produce new combinations of alleles [Posada and Crandall, 2001]. Recombination has also been documented to alter codon usage [Marais *et al.*, 2001] in addition to affecting the accuracy of phylogenetic reconstruction [Posada and Crandall, 2002]. It has been reported that high levels of recombination may result in an unrealistic LRT analysis and therefore produce molecular signatures indistinguishable from those of positive selection [Anisimova *et al.*, 2003]. Recombination has also been associated with GC-biased gene conversion (gBGC) [Katzman *et al.*, 2011]. gBGC is a neutral process whereby GC content increases due to the DNA mismatch repair machinery favoring G:C pairs at recombination breakpoints [Galtier and Duret, 2007]. gBGC has been reported to associate with false positives in selective pressure analyses, primarily due to the inflation of ω [Ratnakumar *et al.*, 2010]. Studies have used GC and GC3 (wobble base) content to imply evidence of gBGC [Romiguier *et al.*, 2013]. However, a recent report on the branch-site model used in this thesis (modelA) found that deviating GC frequencies have no significant effect on false positives [Gharib and Robinson-Rechavi, 2013].

1.4.3 Purifying Selection acting on silent sites mistaken for positive selection

Exonic splice enhancers (ESEs) are nucleotide sequence motifs that are reported to aid in pre-mRNA splicing and have been reported to be under purifying selection [Cáceres and Hurst, 2013; Parmley *et al.*, 2006; Hurst and Pál, 2001]. ESEs are most likely enriched in regions of exonic sequence that are close to splice sites and they have been proposed as a potential driving force behind the observed reduced rates of synonymous and non-synonymous sites towards the ends of exons [Fairbrother *et al.,* 2002; Parmley *et al.,* 2006; Parmley *et al.,* 2007; Woolfe *et al.,* 2010]. Therefore the presence of ESEs may cause an inflation of ω due to a reduction in Ds rather than an increase in Dn [Parmley *et al.,* 2006]. Recently the net impact of ESEs has been conservatively estimated to result in a 4% reduction in Ds [Cáceres and Hurst, 2013]. While ESEs do present a potential source of false positives, many of the ESE datasets are reported to not reflect the known properties of ESEs (for example: enriched near exon boundaries, associated with weak splice sites, and enriched near longer introns) [Cáceres and Hurst, 2013]. In addition, it is currently unknown how ESEs (or the number of ESEs within a sequence) would alter the false-positive rate of the branch-site model. The presence of ESEs does warrant the exploration of proteins known to have ESEs to determine the effect the reported robustness of the branch-site model (modelA) [Gharib and Robinson-Rechavi, 2013; Yang and dos Reis, 2011] and determine how ESEs should be considered in future analyses.

1.4.4 *In vitro* validation of positive selection

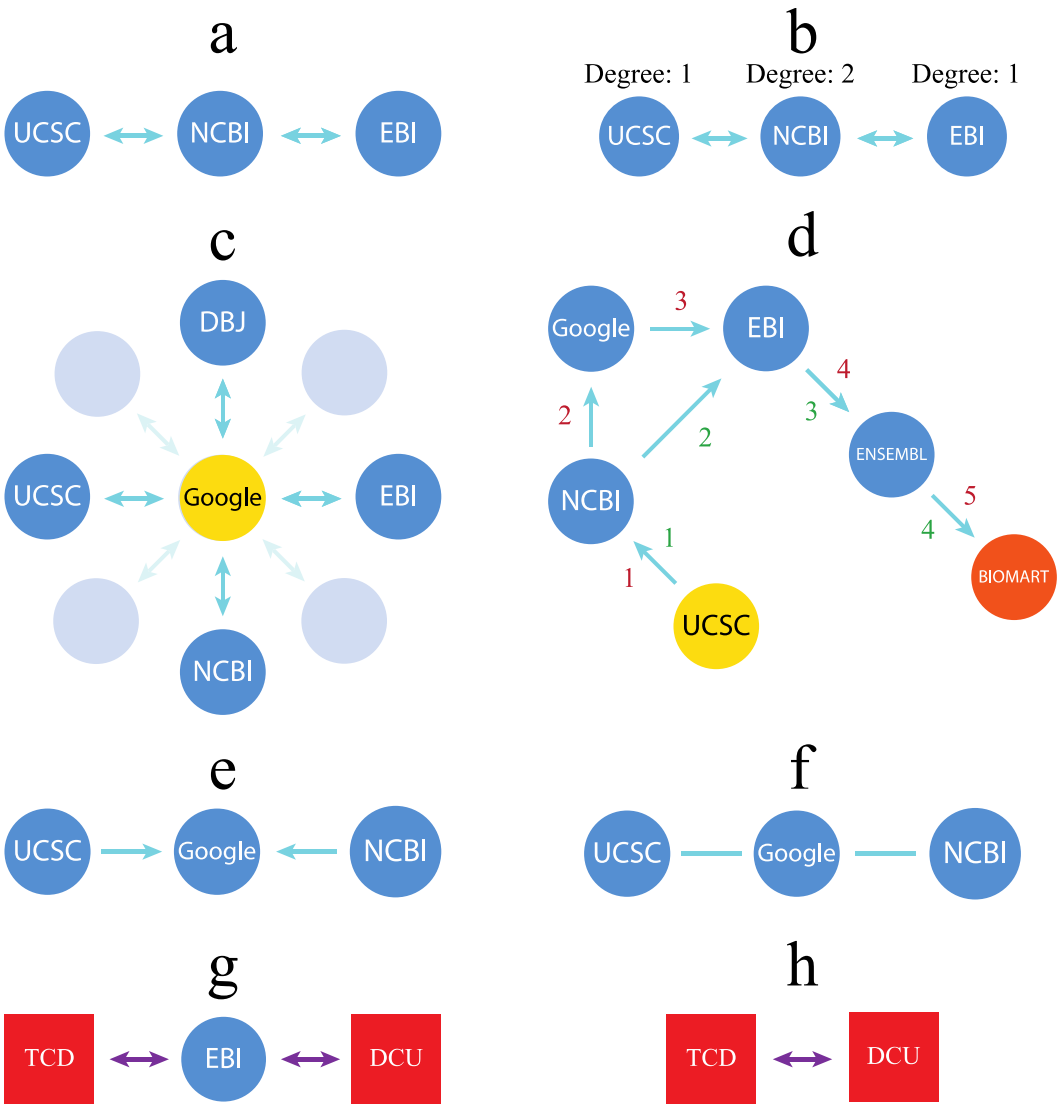
The traditional approach to site-directed mutagenesis mutates specific amino acids within modern day proteins. This simple approach is problematic for *in-vitro* studies of positive selection because it does not account for epistatic interactions with other amino acids in the background, this can lead to erroneous genotype–phenotype correlations [Yokoyama *et al.*, 2012; Yokoyama, 2013]. One possible solution is to first reconstruct the ancestral protein which will provide a more "realistic" background to place the sites of interest [Yokoyama and Radlwimmer, 2001; Shi and Yokoyama, 2003; Bridgham *et al.*, 2006; Harms and Thornton, 2010]. By performing site-directed mutagenesis on these synthesized ancestral proteins it is possible to accurately characterize the relationship between positive selection and function [Yokoyama *et al.*, 2013]. Model A has been previously used to identify positively selected residues in a mammal protein by reconstructing the ancestral protein and then performing rational mutagenesis in that background. It was shown that the positively selected residues elicited a direct functional impact [Loughran *et al.*, 2012]. Such *in vitro* studies were not performed in this thesis as here we are focused on software design for large-scale analyses (Chapter 2) and mechanisms of protein evolution in vertebrates (Chapters 3 and 4).

1.5 Graph theory and molecular evolution

1.5.1 Introduction to graph theory

Graph theory is the subdiscipline of mathematics that studies the nature of graphs, which are mathematical representations of connections between objects (Figure 1.6). Since the initial application of graph theory by computational biologists, the discipline has primarily published research on protein interaction, sequence homology, cell signaling, and genetic association [Katoh and Katoh, 2007; Goh *et al.*, 2007; Bapteste *et al.*, 2012; Franceschini *et al.*, 2012]. One of the most significant discoveries revealed by biological graphs (often referred to as networks) was that despite their complex nature, they shared the common governing principles of scientific and technological graphs (e.g. the Internet and social networks) [Barabasi and Oltvai, 2004]. Of particular importance is that biological graphs are theorized to be scale-free rather than random graphs. In random graphs, the degree of each node does not significantly deviate from the average degree of the graph (Figure 1.7a). In contrast to the random model, the scale-free model is characterized by a power-law degree distribution, which is characterized by a small number of hub nodes that strongly influence the properties of the graph (Figure 1.7b). For example, a protein interaction graph of the TLR signaling pathway displays scale-free characteristics with MyD88 as a central hub node. The existence of hub nodes in scale-free graphs is founded on two concepts, growth and preferential attachment. Growth denotes that graphs grow overtime. Preferential attachment indicates that nodes with higher degree are more likely to gain new connections as they grow [Barabasi and Oltvai, 2004].

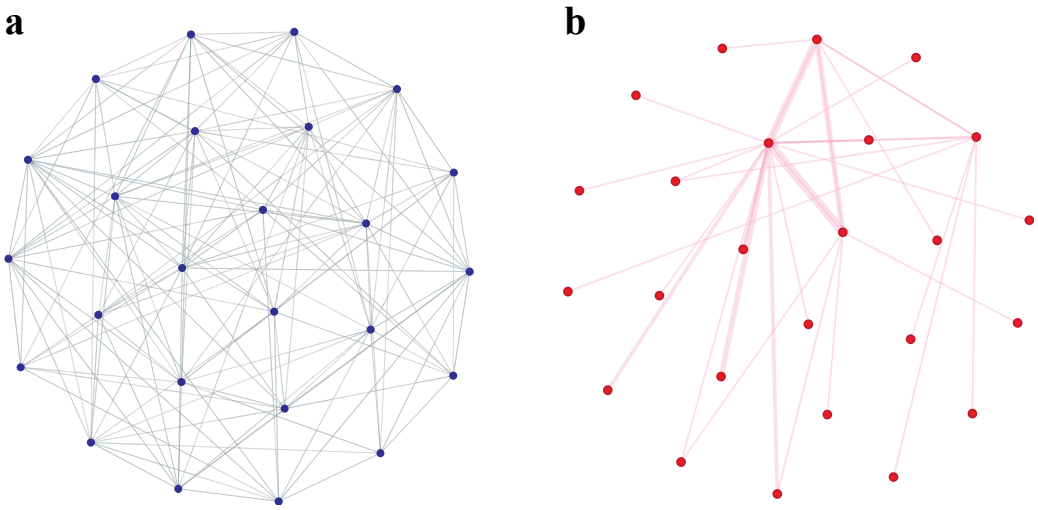
Figure 1.6: Basic graph nomenclature and general types of graphs.



(a) Graphs are composed of objects and the connections between those objects, respectively termed nodes and edges. (b) The degree of a node is the total number of edges it has to other nodes (i.e. nearest neighbors). (c) Nodes may be referred to as hubs if they have a large degree in respect to other nodes (e.g. the Google in yellow). (d) The shortest path between two nodes is the smallest number of edges required to connect the two nodes (e.g. the shortest path from the UCSC genome browser (yellow) to Ensembl BioMart (orange) only requires 4 edges if Google is bypassed. (e) Graphs may be directed and permit edges to have direction associated with them, or (f) undirected if edges have no direction

associated with them. (g) Graphs are bipartite if they incorporate relationships between two independent sets of data (e.g. use of the EBI [European Bioinformatics Institute] database connects DCU and TCD). (h) Relationships between members of the same independent set of a bipartite network may be determined by inferring a unipartite network projection. For example, upon removal of EBI, DCU and TCD are connected in a unipartite graph (unipartite projection of the bipartite graph). This connection is due to both the DCU and TCD nodes being previously associated (i.e. sharing edges) with EBI.

Figure 1.7: The properties of random and scale-free graphs



(a) The Erdös-Rényi model of a random graph is generated by selecting the number of nodes (N) and the probability (p) of a connection and constructs a graph with approximately $\frac{pN(N-1)}{2}$ randomly placed edges [Barabasi and Oltvai, 2004]. The degree distribution (i.e. the distribution of node degree values) of random graphs is expected to follow a Poisson distribution, indicating that most nodes have approximately the same number of edges and do not significantly deviate from the average degree of the graph [Barabasi and Oltvai, 2004]. Therefore the nodes of random graphs are expected to uniform in respect to connectivity. (b) The scale-free model is characterized by a power-law degree distribution, whereby the probability that a node has a connection k is $P(k) \sim k^{-\gamma}$, where γ is a constant typically between 2 and 3 [Barabasi and Oltvai, 2004]. Graphs created from the power-law equation are characterized by a logarithmic decrease in probability of a node existing as the degree of the node increases. Therefore, in scale-free graphs, the majority of nodes exhibit small degree values whereas a small number of nodes exhibit high degree values (i.e. are hub nodes) [Barabasi and Oltvai, 2004]. Because of this, the nodes of scale-free graphs are disproportionate in respect to connectivity.

Duplication events have been proposed as one mechanism that creates preferential attachment in biological networks. As hub proteins naturally exhibit greater connections, they have a higher probability of being connected to a protein that undergoes duplication thereby creating an additional connection [Barabasi and Oltvai, 2004]. Evidence of this concept is illustrated by the duplication events that resulted in TLR7, TLR8, and TLR9 [Leulier and Lemaitre, 2008] each of which shares a connection with MyD88 in a protein interaction graph of the TLR signaling pathway.

1.5.2 Characterizing Graphs

1.5.2.1 Centrality

Centrality is concerned with measuring the influence (or importance) of each node on the structure of the entire graph. While a plethora of measurements have been developed independently, the three most prominent and frequently used measurements of centrality are: degree, closeness, and betweenness [Freeman, 1979].

Degree centrality is the simplest measurement of centrality and is defined as the total number of edges (or adjacencies) for a given node [Freeman, 1979]. The degree of a given node can also be thought of as the initial importance of a given node in the graph. Applying this concept to infection networks, hub nodes unsurprisingly pose the greatest initial risk for spreading an infection [Borgatti, 2005]. The best method for calculating degree centrality of a given node is by using an adjacency matrix of the graph (Figure 1.8), as shown in Equation 1.3:

Equation 1.3: Degree centrality.

$$C_i^{DEG} = \sum_j^N a_{ij}$$

$$a_{ij} := \begin{cases} 1 & if\ i = j \\ 0 & otherwise \end{cases}$$

where i is the given node, j is the remaining nodes, N is the total number of nodes in the graph, and a_{ij} is the adjacency matrix.

The adjacency matrix is necessary to define if an edge exists between the current nodes [Opsahi *et al.*, 2010; Borgatti and Everett 2006].

Closeness centrality is the measurement that states how close a given node is to all other nodes within a graph. The closeness of a given node is the inversed sum of the shortest paths to each of the remaining nodes within the graph [Opsahi *et al.*, 2010; Borgatti and Everett, 2006]. Therefore, nodes with high closeness exhibit a smaller overall distance to the remaining nodes in the graph. In an infection network, nodes exhibiting the highest closeness pose the greatest risk of being positioned for early infection [Borgatti, 2005]. Calculating closeness is achieved using a shortest path adjacent matrix of the graph (Figure 1.8), as shown in Equation 1.4:

Equation 1.4: Closeness centrality.

$$C_i^{CLO} = \frac{N-1}{\sum_j^{N-1} d_{ij}}$$

where i is the given node, j is the remaining nodes, N is the total number of nodes in the graph, and d_{ij} is the shortest path adjacent matrix [Opsahi *et al.*, 2010]. Equation 1.4 is normalized by the total remaining nodes (i.e. $N-1$).

Betweenness centrality is the measurement of centrality that defines the total number of times a given node belongs to the shortest path of two separate nodes [Freeman, 1979]. Nodes with high betweenness are regularly required for connection between nodes that are distantly connected in a graph. For example, shipping canals (e.g. Panama and Suez) typically exhibit high betweenness in global shipping networks as they are frequently visited en route to other ports [Kaluza *et al.*, 2010]. Betweenness centrality is calculated using Equation 1.5:

Equation 1.5: Betweenness centrality.

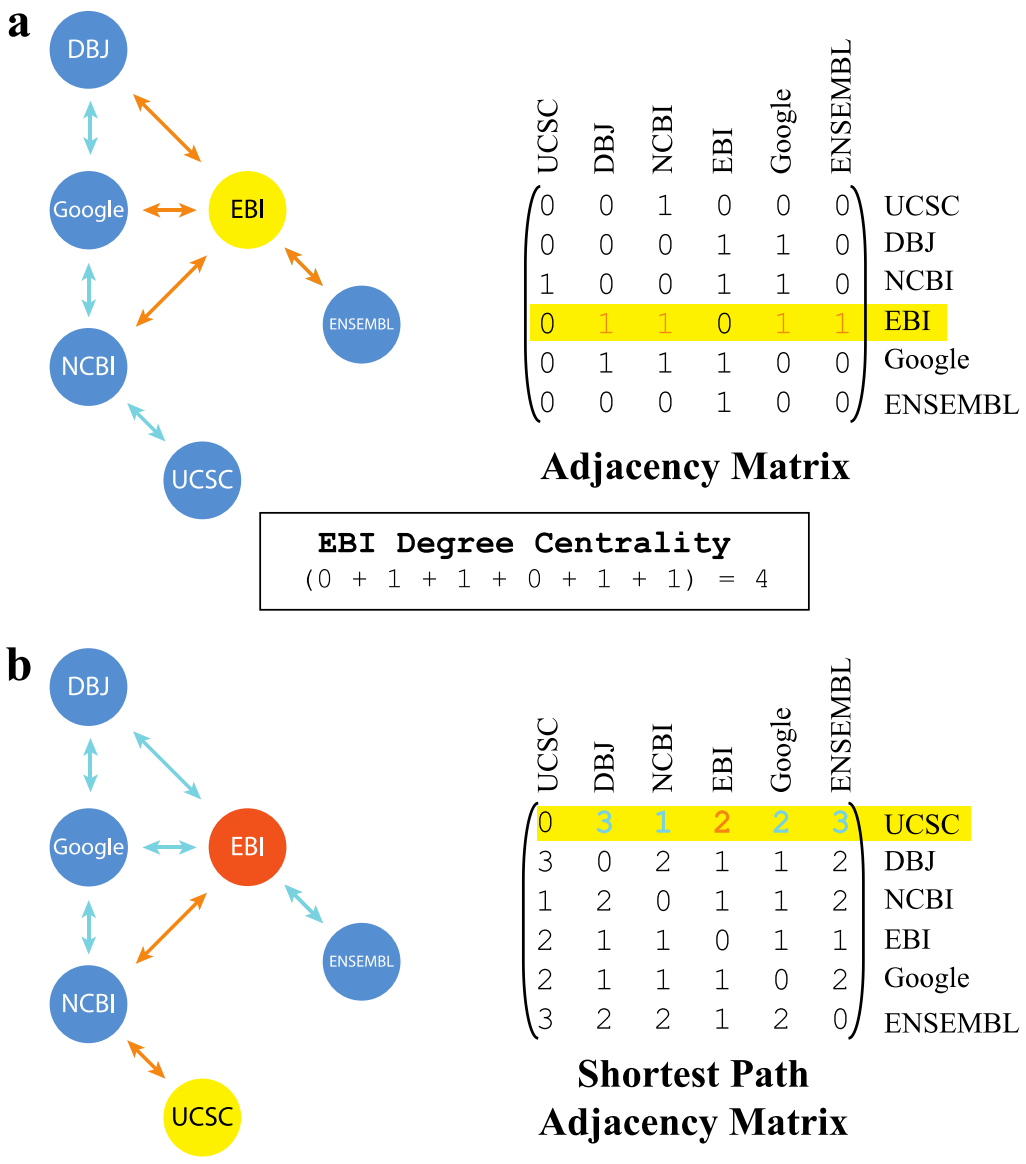
$$C_i^{BET} = \sum_j \sum_k \frac{\sigma(j,k|i)}{\sigma(j,k)}$$

where i is the given node, $\sigma(j,k)$ is the number of shortest paths between the nodes j and k, and $\sigma(j,k|i)$ is the number of those shortest paths that that require i [Borgatti and Everett, 2006]. Betweenness may also be normalized by calculating the maximum number of edges in a graph of N nodes $\left(\frac{2}{(n-1)(n-2)}\right)$, see Equation 1.6 for the normalized betweenness:

Equation 1.6: Normalized betweenness centrality.

$$C_i^{NormBET} = \frac{2\left(\sum_j \sum_k \frac{\sigma(j,k|i)}{\sigma(j,k)}\right)}{(N-1)(N-2)}$$

Figure 1.8: Calculating degree and closeness centrality using adjacency matrices.



(a) Degree centrality can be calculated by summing the number of edges of the node of interest in an adjacency matrix. For example, the degree centrality of the EBI (highlighted in yellow on the matrix) is four due to having edges with the DBJ, Google, NCBI, and ENSEMBL nodes. (b) Closeness centrality can be

calculated by dividing the total number of nodes in the graph minus the given node (i.e. N - 1) by the sum of the shortest paths to all remaining nodes in the graph. This can be accomplished using a shortest path adjacency matrix. For example, the shortest path between UCSC and EBI is two as two edges (shown in orange above) is the fewest number of edges required to connect the nodes. The closeness centrality of UCSC to all other nodes on the graph is 0.45 based on this calculation.

1.5.2.2 Assortativity

Assortativity measures the correlation between degree and node connectivity. Depending on the observed correlation, graphs may be defined as showing assortative, neutral, or disassortative mixing patterns. Assortative mixing is a preference for any given node to attach to other nodes displaying a similar degree (e.g. high-degree nodes attached to high-degree nodes) [Newman, 2002; Newman, 2003]. Disassortative mixing is a preference of attachment to nodes of dissimilar degree (e.g. high-degree nodes with low-degree nodes) [Newman, 2002; Newman, 2003]. Graphs rarely exhibit neutral mixing, where neither a preference for assortative nor disassortative mixing is detected [Newman, 2002; Newman, 2003]. A variety of social, technological, and biological graphs have undergone assortativity measurements. Biological (e.g. protein interaction and metabolic) and technological (e.g. internet and world-wide-web) networks are predominantly disassortative whereas social networks (e.g. co-authorship and actor collaborations) are predominantly assortative [Newman, 2003].

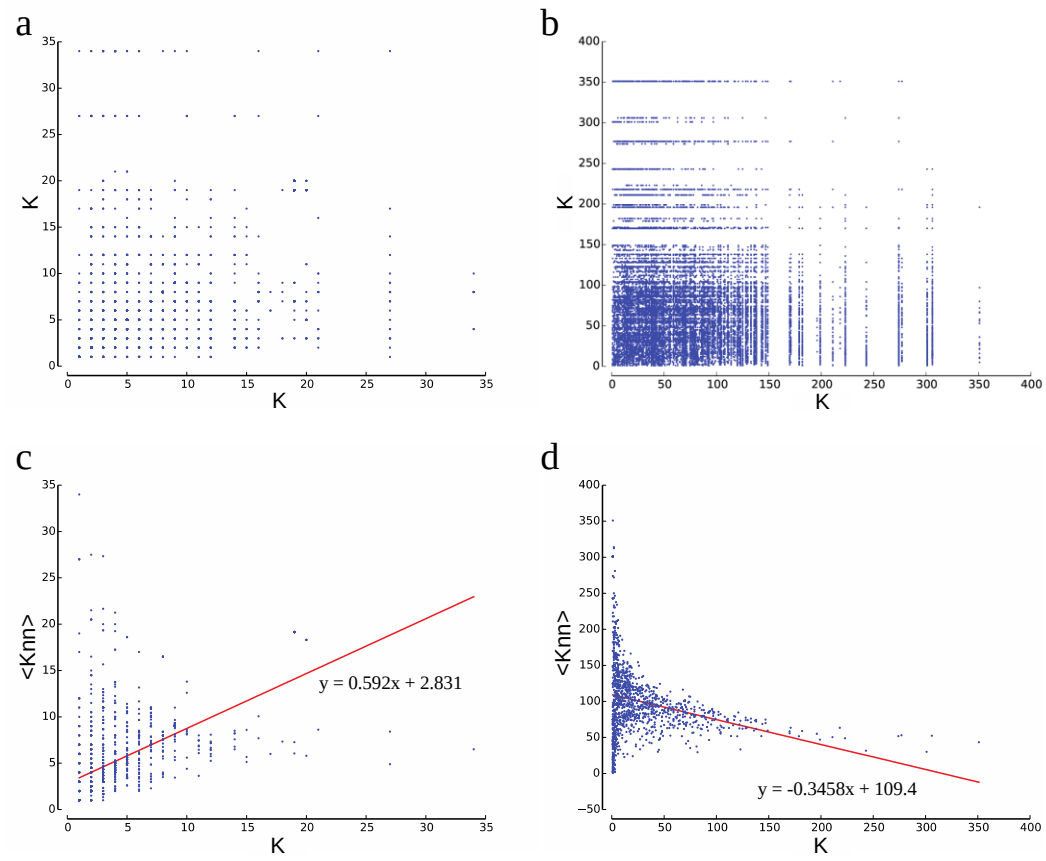
Depending on the robustness of the assortativity analysis, methods of varying complexity have been devised to determine the mixing pattern of a graph. The simplest and often initial method is constructing a chart of each edge in a graph, whereby the axes of the chart are scaled by the degree of the nodes in an edge [Newman, 2003]. It should be noted that edge charts are only ideal for deciphering predominantly assortative or disassortative mixing (Figure 1.9a and b). Another method for investigating degree correlation is plotting average degree of nearest neighbors for a node in respect to the degree of the node in question [Pastor-Satorras and Vespignani, 2001]. The assortativity of the graph is

then estimated from the linear regression of the data - a positive slope indicating assortative, negative indicating disassortative, and a slope of zero indicating neutral (Figure 1.9c and d). Lastly, the assortativity coefficient may be calculated for a graph using a Pearson correlation to determine the linear correlation between degree connectivity [Newman, 2002]. The assortativity coefficient (r) of a graph lies between $-1 \leq r \leq 1$ with positive indicating assortative mixing, negative indicating disassortative, and 0 indicating neutral [Newman, 2002]. To determine the statistical significance of the assortativity coefficient of a graph, a number of randomized graph with the same degree distribution are used to obtain a confidence interval of the assortativity coefficient [Foster *et al.*, 2010].

1.5.2.3 Cliques and Communities

Cliques are defined by graph theory as subgraphs in which every pair of nodes is connected by an edge (*i.e.* fully connected or complete) [Luce and Perry, 1949] (Figure 1.10a). Cliques are often defined by their size k (*i.e.* k-clique), where k is the number of nodes within the clique. Cliques may also be designated as maximal or maximum [Butenko and Wilhelm, 2006]. Maximal cliques are defined as cliques that cannot expand by incorporating neighboring nodes. A maximum clique is the clique of the greatest size in the graph. It should be noted that identifying either the maximum clique or all maximal cliques within a given graph is considered an NP-complete problem, problems in which a given solution may be verified in polynomial time but computing an exact solution cannot be completed in an efficient manner [Karp, 1972; Leeuwen, 1998]. Despite this innate difficultly, algorithms have been developed to approximate the maximum clique or all maximal cliques [Butenko and Wilhelm, 2006].

Figure 1.9: Methods for characterizing graph assortativity

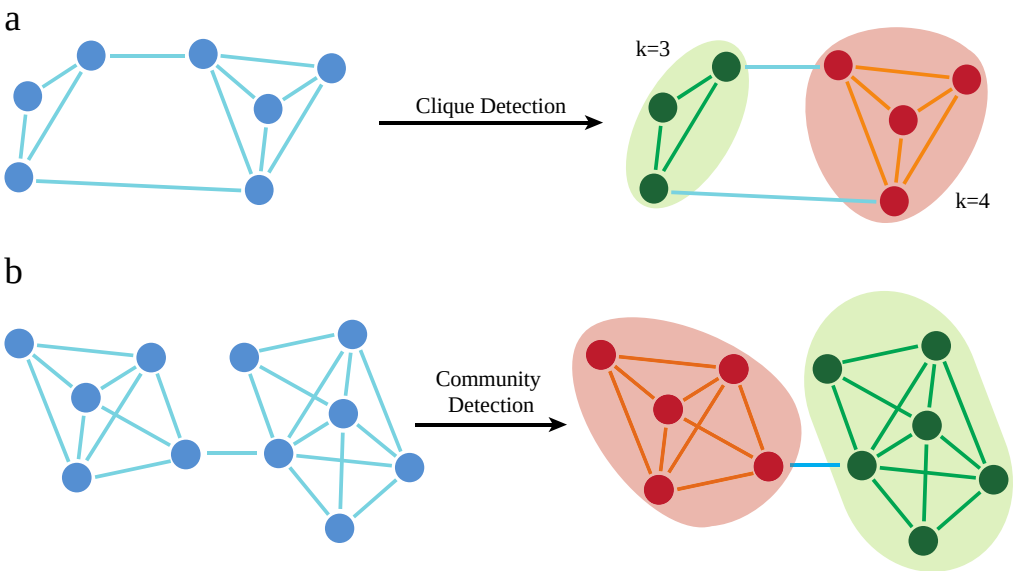


There are two common approaches to plot graph assortativity: (a and b) degree plots and (c and d) neighbor connectivity plots. Two datasets are shown for each plot approach: (a and c) an assortative dataset and (b and d) a disassortative dataset. (a and b) Degree plots are constructed by plotting the degree values (K) of each pair of nodes connected by an edge (one value for each axis). Comparing the mixing pattern generated by (a) the assortative dataset to (b) the disassortative dataset illustrates the difficulty of verifying the assortativity pattern from degree plots. Only highly assortative or disassortative networks will result in patterns that are not ambiguous. (c and d) Neighbor connectivity plots are constructed by plotting the average degree of nearest neighbors (<Knn>) of a given node with the degree (K). In comparison to degree plots, neighbor connectivity plots are able to the accurately identify both weakly assortative and

disassortative graphs. This is achieved by determining the linear regression of <Knn> on the y-axis to and K on the x-axis, with a positive slope indicating assortative mixing and a negative slope indicating disassortative mixing. For example, linear regression was able to correctly identify (c) the assortative dataset with a positive slope and (d) the disassortative dataset with a negative slope.

Communities in graphs are defined as subsets of nodes that exhibit dense internal connections but sparse connections elsewhere [Girvan and Newman, 2002] (Figure 1.10b). Various methods to distinguishing community structure within large graphs have been developed, however, these methods are limited to approximations as identifying the most dense subgraph in a graph is classified as an NP-hard problem [Rivera *et al.,* 2010]. Similar to NP-complete problems, NP-hard problems are unable to be solved in an efficient manner (i.e. polynomial time), however, NP-hard problems occasionally are unable to verify a given solution in polynomial time [Leeuwen, 1998]. Another limitation of many of these community detection methods is user-defined parameters, which if incorrectly specified may give unrealistic results. The NeMo algorithm was selected for this thesis due to the accuracy of the algorithm and the absence of user-defined parameters [Rivera *et al.,* 2010]. NeMo detects communities by calculating a log odds score for observing a certain number of shared neighbors between nodes [Rivera *et al.,* 2010].

Figure 1.10: Visual representation of cliques and communities detection.



(a) A clique is characterized by a subset of nodes in which each pair is connected by an edge. Both cliques shown (k=3 and k=4) are maximal cliques, as they cannot grow larger. The maximum clique of the graph is k=4 (shown in red). (b) A community is characterized by a subset of densely interconnected nodes that are sparsely connected elsewhere.

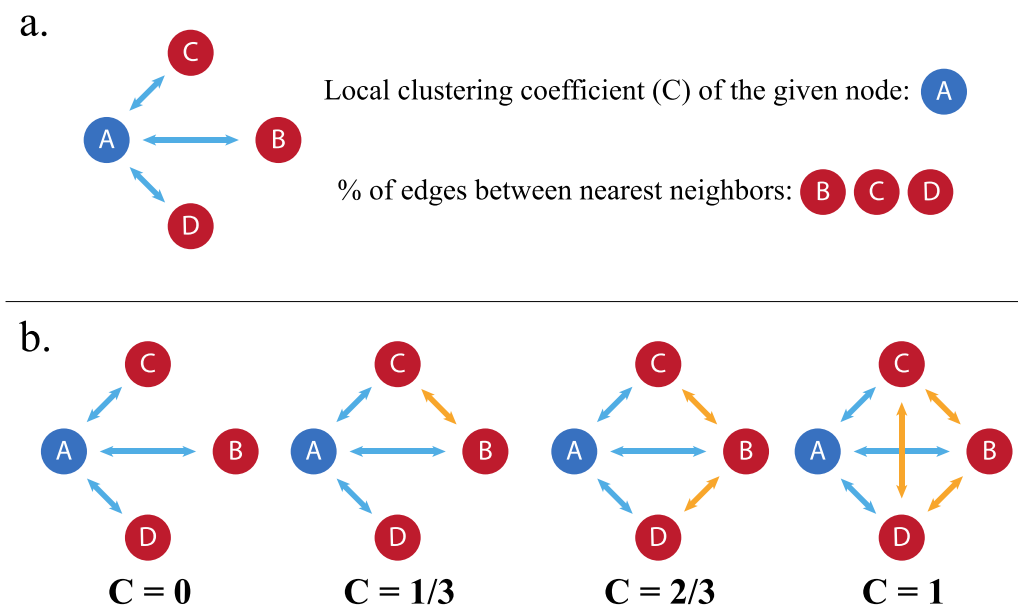
1.5.2.4 Clustering

Two coefficients of graph clustering were employed within this thesis: average clustering and transitivity. Clustering coefficients are typically used to measure the degree to which nodes cluster [Luce and Perry, 1949; Watts and Strogatz, 1998].

Average clustering is the average of the local clustering coefficients of all nodes in a graph [Watts and Strogatz, 1998]. The local clustering coefficient of a given node is defined as the percentage of nearest neighbor (i.e. nodes sharing an edge with the given node) pairs that share an edge (Figure 1.11a). Local clustering coefficients range from 0 to 1; a coefficient of 0 indicates that the nearest neighbors of the given node share no connections whereas a coefficient of 1 indicates the nearest neighbors are completely connected (along with the given node) and are a clique (Figure 1.11b).

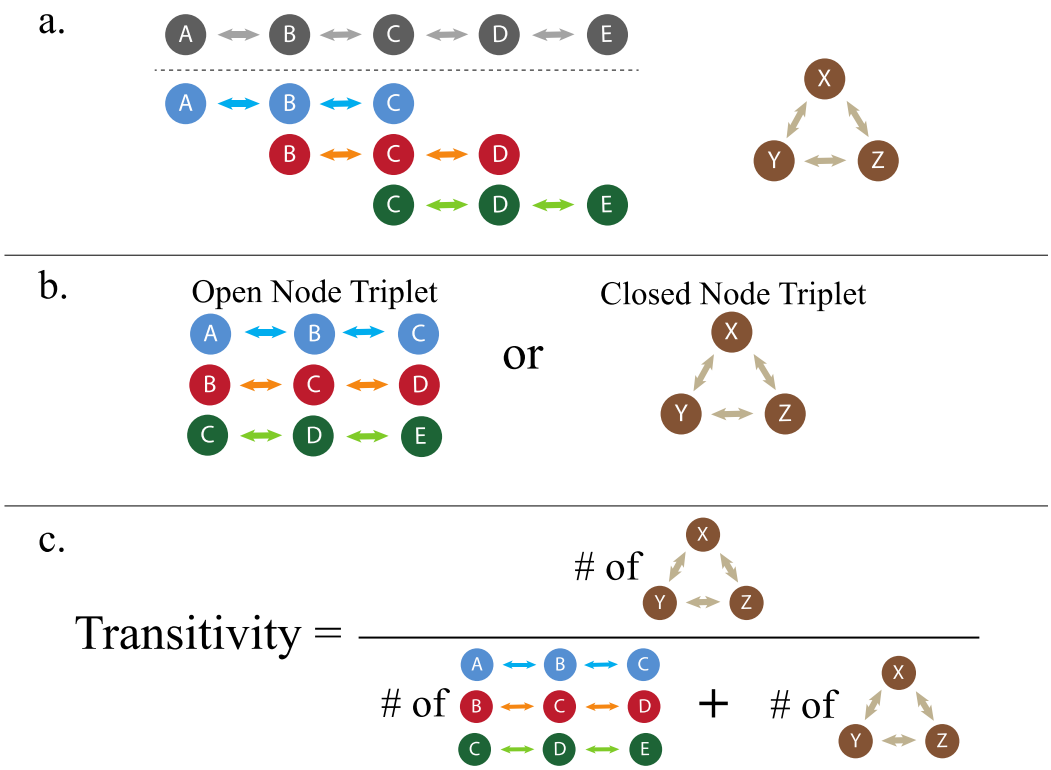
Transitivity is a global clustering coefficient based on the number of node triplets within a graph (Figure 1.12a). Node triplets are either defined as open or closed, an open triplet only possesses two edges and therefore is not fully connected whereas a closed triplet possesses three edges and is fully connected (i.e. a clique) (Figure 1.12b). Transitivity is measured by dividing the closed triplets (often defined as triangles) by the total number of node triplet (open or closed) in the graph (Figure 1.12c). Graph transitivity ranges from 0 to 1; a coefficient of 0 indicates that all triplets are open whereas a coefficient of 1 indicates that all triplets are closed (i.e. cliques).

Figure 1.11: Visual representation of average clustering.



(a) The local clustering coefficient of node A (shown in blue) is the percentage of edges shared by pairs of the nearest neighbors of node A, which are: B, C and D (shown in red). (b) The clustering coefficient (C) of four scenarios: C = 0 if there no edges between nearest neighbors, C = 1/3 if there is a single edge between nearest neighbors, C = 2/3 if there are two edges between nearest neighbors, and C = 1 if the nearest neighbors and the given node (node A) are a clique (i.e. fully connected).

Figure 1.12: Visual representation of transitivity.



(a) The given network has two connected components: 1) nodes A, B, C, D, and E and 2) nodes X, Y, and Z. These components are then broken into triplets (i.e three nodes connected by at least two edges). Component 1 (A, B, C, D, and E) can be broken into three triplets: A-B-C, B-C-D, and C-D-E. Component 2 is only made of a single triplet: X, Y, Z. (b) Triplets are either open and contain only two edges (i.e. not fully connected) or closed – often termed triangles – and contain three edges (i.e. a clique). (c) The transitivity coefficient is calculated by dividing the number of closed triplets (triangles) over the number of all triplets (open and closed).

1.5.3 Graphs and introgressive descent

Of the various biological relationships characterized by graphs, sequence-similarity networks have recently become of particular interest to evolutionary biologists due to their ability to accurately represent the molecular signatures of non-linear or introgressive descent [Bapteste *et al.*, 2013]. Sequence-similarity networks represent sequences as nodes and infer edges from homology data provided by programs such as BLAST, FASTA, or HMMER [Eddy, 1998; Altschul *et al.*, 1990; Lipman and Pearson, 1985]. In a sequence-similarity network the homology connections of both monophyletic orthologs and introgressive descent are characterized by distinct edge patterns (Figure 1.13). Networks therefore enable the evolutionary impact of events such as domain sharing, mosaic genes, plasmids, and phages to be accurately evaluated in addition to the traditional tree-like molecular signatures [Bapteste *et al.*, 2013].

1.5.3.1 Tools for detecting introgressive events in networks

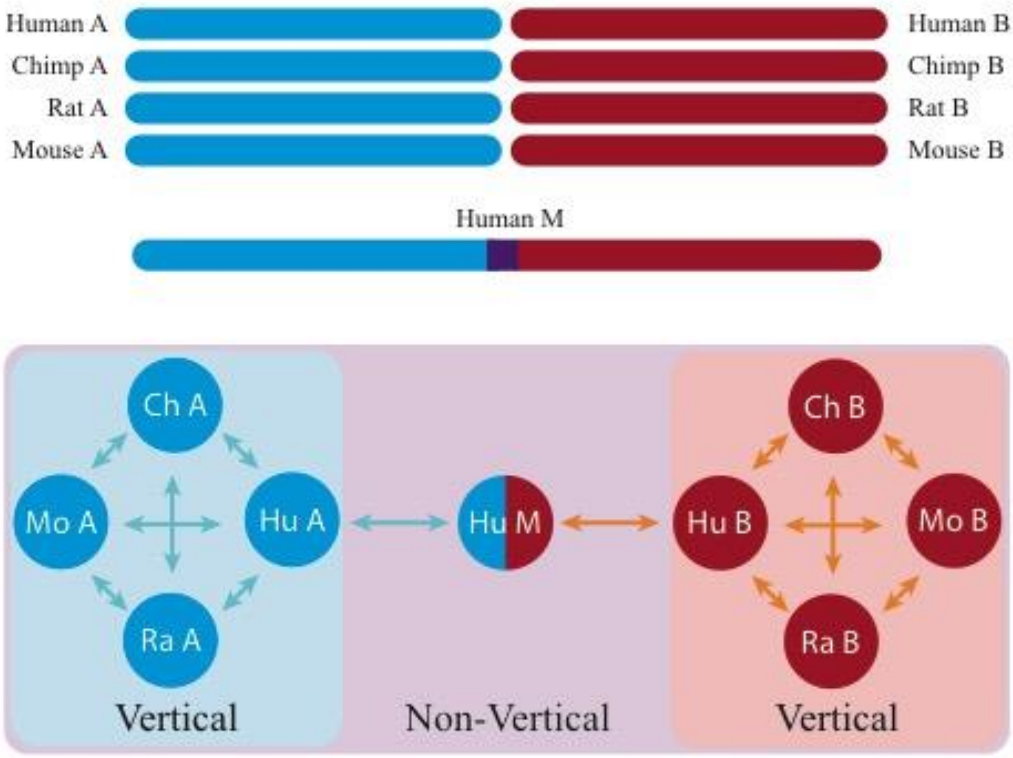
Currently, there are a small number of tools that have been devised to identify introgressive events in large sequence-similarity networks. The recently released program MosaicFinder identifies composite genes – a gene generated by a fusion between two previously separate/distinct genes – by constructing a sequence-similarity graph from BLAST alignment data [Jachiet *et al.*, 2013]. MosaicFinder identifies composite genes by equating them to clique minimal separators [Jachiet *et al.*, 2013]. A clique separator is a node that if removed will cause the graph to separate into connected components, that separator is minimum if no subset of the node also causes separation (i.e. an independent separator that does not require the node in question) [Berry *et al.*, 2010], e.g. Human M in Figure

1.11. MosaicFinder subsequently confirms potential composites using the BLAST output data [Jachiet *et al.,* 2013].

1.5.4 Composite genes and functional discordance.

Composite genes (introgressive events caused by fusion of two or more genes) have been reported to have unique species-specific functional properties [Thomson *et al.,* 2000; Rogers *et al.,* 2010; Molero *et al.,* 2013]. These instances of non-vertical descent have resulted in the alterations to cellular localization (e.g. the human fusion Kua–UEV [Thomson *et al.,* 2000]), distinct regulatory profiles (e.g. the *Drosophila* fusion Quetzalcoatl [Rogers *et al.,* 2010]), and combine of distinct functions (e.g. *Schizosaccharomyces pombe* fusion gene hal3 [Molero *et al.,* 2013]). The ability of composite genes to elicit a variety of species-specific functions demonstrates the importance of characterizing non-vertical events. Therefore the unique ability of graphs to accurately identify non-vertical events provides an additional method to understand functional discordance between species.

Figure 1.13: Basic characteristics of introgressive descent in graphs



Graphs are capable of displaying the unique characteristics of both monophyletic orthologs (vertical decent) and mosaics (non-vertical descent). Monophyletic orthologs typically manifest as cliques as they share coalescent orthologs e.g. Gene A (human, chimp, rat, and mouse orthologs of gene A) in blue & Gene B (human, chimp, rat, and mouse orthologs of gene B) in red. Graphs depict mosaic sequences, human Gene M (human mosaic of human A gene and human B gene), as sparsely connected nodes that typically connect unrelated groups with strong interconnectivity.

Chapter 2: Design and development of the bmeTools package

2.1 Chapter Aim

Analysis of selective pressure heterogeneity requires a large number of steps from ortholog identification through to phylogeny reconstruction and likelihood ratio tests of the codon models applied. Therefore, the primary aim of chapter two was the development of a highly automated bioinformatic pipeline for selective pressure analysis. We called this pipeline the Bioinformatics and Molecular Evolution Tools package or simply bmeTools. The pipeline was designed to minimize potential biases of the user by including software that automates the selection of alignments and/or substitution models based on the sequence data rather than user assumptions. Therefore there were two major aims in the design of this pipeline (1) to simplify the analysis and (2) to reduce potential human error. These goals were achieved by designing a pipeline that included software for all steps from data acquisition to the analysis of the selective pressure results, the major steps involved were: (i) identification of gene families, (ii) alignment, (iii) phylogeny reconstruction, (iv) selective pressure analyses, and (v) Likelihood ratio test calculations to determine the codon based model of best fit for each gene alignment.

The bmeTools package was then applied to the comparative genomic analysis of the newly sequenced Bowhead whale genome to test all functionality of the software.

2.2 Introduction

Since the initial release of the human genome in 2001 [Lander *et al.*, 2001], technological advances in both computing and sequencing have enabled the release of genome assemblies of approximately eighty different vertebrates [Flicek *et al.*, 2014]. In this same period, researchers in the field of molecular evolution have developed techniques capable of evolutionary analyses on a genome-wide scale (for example [Kosiol *et al.*, 2008]). One of the most widely used approaches to estimate the selective pressure variation across homologous protein-coding genes is to calculate the ratio of non-synonymous substitutions per non-synonymous site (Dn) over synonymous substitutions per synonymous site (Ds) (Dn/Ds or ω). An ω value >1 is the classical signature of molecular adaptation and until recently has been theorized to indicate potential functional divergence [Sawyer *et al.*, 2005]. Recently there have been a number of independent studies that have successfully linked positive selection to protein functional divergence in a wide variety of species [Loughran *et al.*, 2012; Moury and Simon, 2011; Levasseur *et al.*, 2006; Sawyer *et al.*, 2005]. Large-scale genomic studies of selective pressure variation across species have the potential therefore to identify e.g. the molecular underpinnings of species-specific traits [Kosiol *et al.*, 2008]. The focus here was to design and implement a pipeline for large-scale selective pressure analysis, thereby positioning us to identifying protein functional shifts between mouse and human that underpin species-specific immune responses [Mestas and Hughes, 2004].

To date, a number of methods and software packages of varying complexity have been released with the purpose of calculating ω [Delport *et al.*, 2010; Yang,

2007; Pond and Frost, 2005]. One of the most highly cited is codeML from the PAML software package [Yang, 2007]. CodeML uses codon-based models of evolution to assess selective pressures in protein coding sequence alignments at specific sites or restricted to sites in a predefined lineage [Yang and dos Reis, 2011]. Operating codeML requires a complex file structure to accurately compute the parameters under multiple nested models, calculate the associated likelihood ratio tests (LRTs), and perform the statistics required to infer putative positive selection. Overcoming these complexities and properly interpreting the results for most evolutionary biologists is achieved by creating in-house software pipelines, which are rarely publically available. Lack of a flexible toolkit for upstream and downstream analyses has proven problematic for many organismal biologists working on next-generation sequence data and is an additional stumbling block for analysis and interpretation of their novel data. Attempts have been made to solve this problem by creating a simplified and automated pipeline for codeML analysis, but most have only focused on the codeML package itself (e.g. BioPerl PAML, Lazarus, etc.) [Hanson-Smith *et al.*, 2010; Stajich *et al.*, 2002; Walsh, 2013]. These attempts have ignored the various steps required prior to and following codeML analysis, which would still present a problem to the non-programming biologist. Enabling codeML analysis for organismal biologists requires a broad functioning pipeline that streamlines and automates the many analyses prior to the codeML stage – such as data collection, homology searching, sequence alignment and phylogenetic reconstruction – as well as automating the various mathematical analyses required to interpret the output from codeML.

Here we present bmeTools, a software package for the automation of codeML and associated upstream and downstream analyses. bmeTools was primarily designed for users unfamiliar with the command-line environment of codeML by eliminating the majority of data manipulation requirements, enabling large-scale analyses, and automatically analyzing codeML output. But bmeTools is equally useful for the more experienced user as it provides a flexible system for a variety of analyses. bmeTools produces results that are easy to interpret and allow simplified assessment and identification of false positive results for inexperienced users. All bmeTools related functions can be found in Appendix 2.

2.3 Aims for selective pressure analysis package

1) Create a simple and comprehensive analysis package to enable researchers with limited bioinformatics experience to conduct large-scale molecular evolutionary analyses such as homology searching, alignment and phylogeny reconstruction as well as codeML analyses.

2) Create a robust and flexible analysis package to enable high-throughput molecular evolutionary analyses for experienced researchers.

2.4 Motivation behind the development of bmeTools

2.4.1 Minimize human error

A major motivation for the bmeTools package was minimizing potential sources of error in selective pressure analyses. Assessing selective pressure variation requires a complex pipeline composed of numerous independent analyses, including: ortholog identification, multiple sequence alignment, phylogenetic reconstruction, and assessment of codon-based models of evolution. The pipeline

requires multiple data manipulation steps to combine the output of each of these different techniques (e.g. parsing BLAST result files to identify homologs and assessment of the suitability of the phylogenetic tree for selective pressure analysis). For researchers with limited bioinformatics knowledge, manual data manipulation is prone to error, is potentially unstandardized, and is difficult to reproduce. bmeTools was designed to be an easily reproducible method to eliminate the need for manual data manipulation by creating functions that automatically complete the majority of data manipulation steps using a standardized approach. In addition, the use of bmeTools should minimize the requirements for inexperienced users to create their own programs, which may be vulnerable to programming errors.

Another potential source of error that motivated the creation of the bmeTools package was the potential for inexperienced researchers to use aberrant pipelines for data analysis. While the procedures within each stage of a selective pressure analysis are independent, there are requirements on the order in which the phases are carried out. bmeTools was designed to mitigate these complications by creating a standardize pipeline of analyses with a specific ordering of phases in the process. In addition, the package encompasses multiple specialized pipelines to accurately assess selective pressure and reduce potential false positives (such as those caused by alignment error [Fletcher and Yang, 2010]).

2.4.2 Increase user productivity

Another motivation in the development of bmeTools was to increase user productivity by automating labour intensive tasks. Increasing productivity in this

respect can be achieved in two ways: i) automation by recursion – used to repeat an analysis on a number of files (e.g. cleaning and translating a directory of genomes), and ii) automation of analysis methods – used to complete tasks that are normally demanding but invariable in execution (e.g. identifying homologs within BLAST output data). Automating these procedures within bmeTools has created an analysis package that is highly scalable (i.e. from a single gene to whole genomes) and that is suitable for the needs of all levels of expertise.

2.5 Rationale behind the development of bmeTools

2.5.1 Selection of python programming language

The primary rationale for using the python programming language was the high productivity of the language [Prechelt, 2000]. The general syntax of Python requires fewer lines than traditional compiler languages (C and C++) while still maintaining a competitive runtime [Prechelt, 2000; Fourment and Gillings, 2008]. Python also incorporates a number of built-in libraries that reduce development time by enabling previously designed functions to be easily incorporated into a program [van der Walt *et al.*, 2011; Sukumaran and Holder, 2010; Hunter, 2007].

'Pythonic' programs, or programs that are minimalistic and highly readable is a major goal [Fourment and Gillings, 2008]. This is beneficial within a PhD environment as any student proficient in python is able to easily understand 'Pythonic' software. High readability also allows for software to be easily maintained and recycled within a laboratory years after the initial development.

Python programs are executable without compilation, and to run python programs, users are only required to install Python on their system.

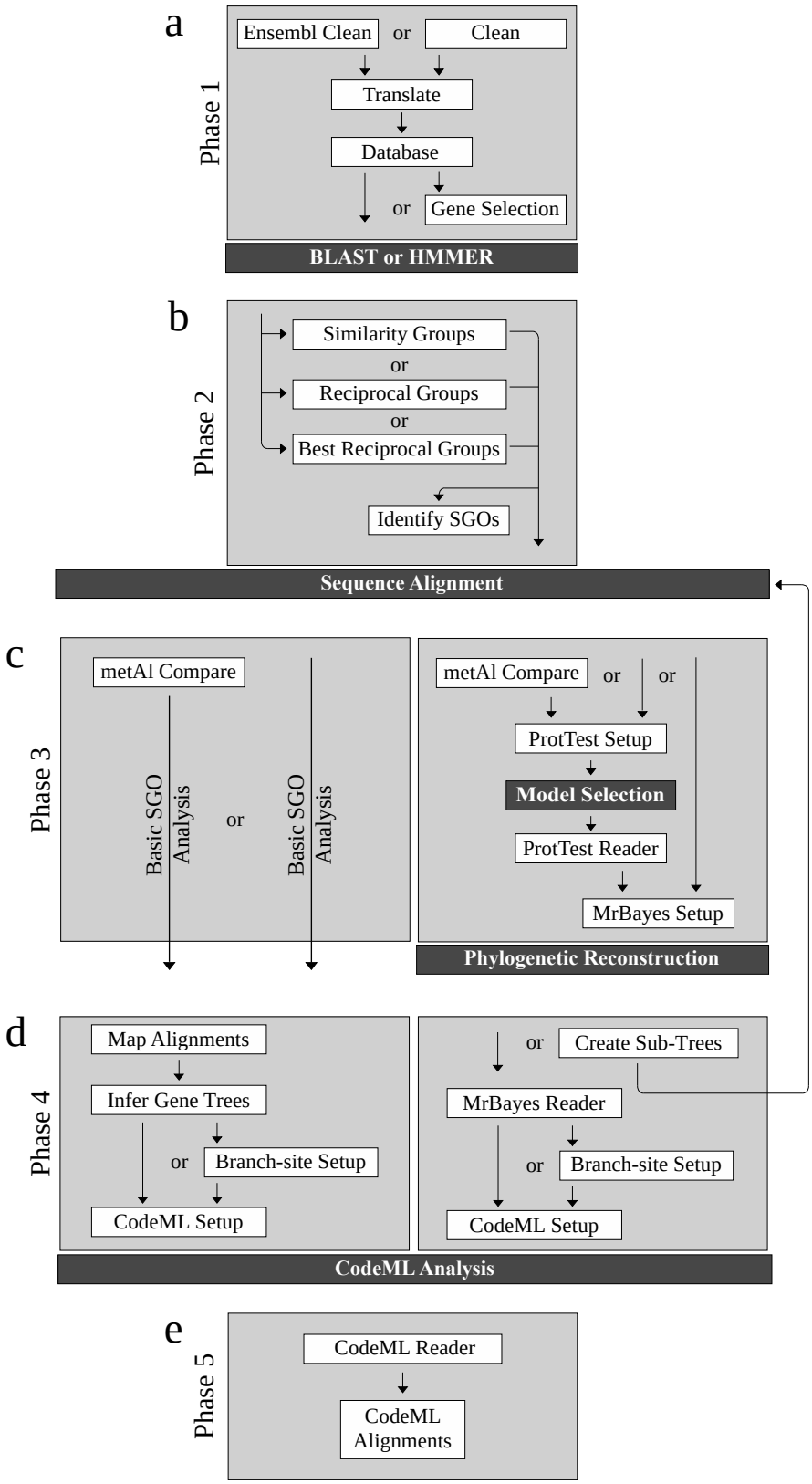
2.5.2 Separation of package into analysis phases

The rationale behind separating the bmeTools package into analysis phases 1 - 5 (Figure 2.1) was primarily to aid users in understanding the distinct procedures involved in selective pressure analysis and to provide more advanced users with a flexible and adaptable pipeline. All functions within a phase analyze the same input type (e.g. sequences, BLAST output, etc.), only specific functions can be combined and the overall output of a phase produces a new data type.

2.6 General overview of bmeTools

There are five separate analysis phases and two analysis pipelines in the bmeTools package, a basic pipeline for single gene orthologs (SGOs) and an advanced pipeline for both SGOs and multi-gene families (MGFs). The basic pipeline was designed to bypass the phylogenetic reconstruction techniques (phase 3) by inferring a gene phylogeny from a user defined species phylogeny. Usage of the basic pipeline is only recommended if the genes are confirmed SGOs. See Figure 2.1 for additional details on the pipeline. The output of each phase in the bmeTools package requires an analysis step that must be completed by the user with third-party software (Figure 2.1). These analyses are not automated by bmeTools for two reasons: i) these analyses are far too computationally intensive, and ii) the submission process for these programs may differ from user to user. In addition, software updates may create bugs within the pipeline.

Figure 2.1: Overview of the bmeTools package.



Each phase indicates the functions (white boxes) and the order in which they are invoked. Optional functions are indicated by 'or' and may be skipped. Dark

boxes indicate third-party programs. (a) Phase 1 (Section 2.6) is the data preparation phase and includes the functions: ensembl_clean/clean (Section 2.6.1), translate (Section 2.6.2), create_database (Section 2.6.3), and gene_selection (Section 2.6.4). This phase ends with the requirements for sequence similarity searching. (b) Phase 2 (Section 2.7) is the similarity group creation phase and includes the following functions: similarity_groups (Section 2.7.2), reciprocal_groups (Section 2.7.2) and best_reciprocal_groups (Section 2.7.3). This phase results in the creation of requirements for multiple sequence alignment (MSA). (c) Phase three (Section 2.8) is the alignment assessment stage and includes both a basic pipeline (on the left) for MSA files that contain only single gene orthologous (SGOs) and an advanced pipeline (on the right) for unconfirmed MSA files. The phase includes the following functions: metal_compare (Section 2.8.1), protest_setup (Section 2.8.2), protest_reader (Section 2.8.2), and mrbayes_setup (Section 2.8.3). This phase results in either: i) a phylogenetic trees of the MSAs for the advanced pipeline or ii) selected MSAs for the basic pipeline. (d) Phase four (Section 2.9) is the selective pressure phase and continues the basic pipeline and advanced pipeline of the previous phase. The phase four basic pipeline includes: map_alignment (Section 2.9.1), infer_genetree (Section 2.9.2), create_branch (Section 2.9.6), and setup_codeml (Section 2.9.3). The phase four advanced pipeline includes: create_subtrees (Section 2.9.4), mrbayes_reader (Section 2.9.5), create_branch (Section 2.9.6), and setup_codeml (Section 2.9.3). This phase results in the input requirements for selective pressure analysis by codeML. (e) The final phase (Section 2.10) includes the function codeml_reader (Section 2.10.1) that analyzes the results of the codeML analysis.

2.7 Phase 1 – Data Preparation

The data preparation phase was included in bmeTools for users with limited bioinformatics training. The phase prepares downloaded genomes for homology searching using the two bmeTools supported homology search tools: BLAST [Altschul *et al.*, 1990] and HMMER [Eddy, 1998]. Many of the included functions were designed to operate in various circumstances by incorporating additional options. The phase also includes a number of supplementary functions not required for either pipeline shown in Figure 2.1 but rather to aid inexperienced users in homology searching.

2.7.1 Functions: clean and ensembl_clean

The basic 'clean' function was designed as a quality control (QC) filter for downloaded nucleotide sequences and/or genomes (Figure 2.2a). Each sequence is confirmed as protein coding by using a conditional statement to verify that the nucleotide sequence encompasses only complete codons (Figure 2.2b). This is an essential step to confirm gene annotation quality and permit the codon substitution models of codeML [Yang, 2007]. Only sequences that pass QC are retained (Figure 2.2c).

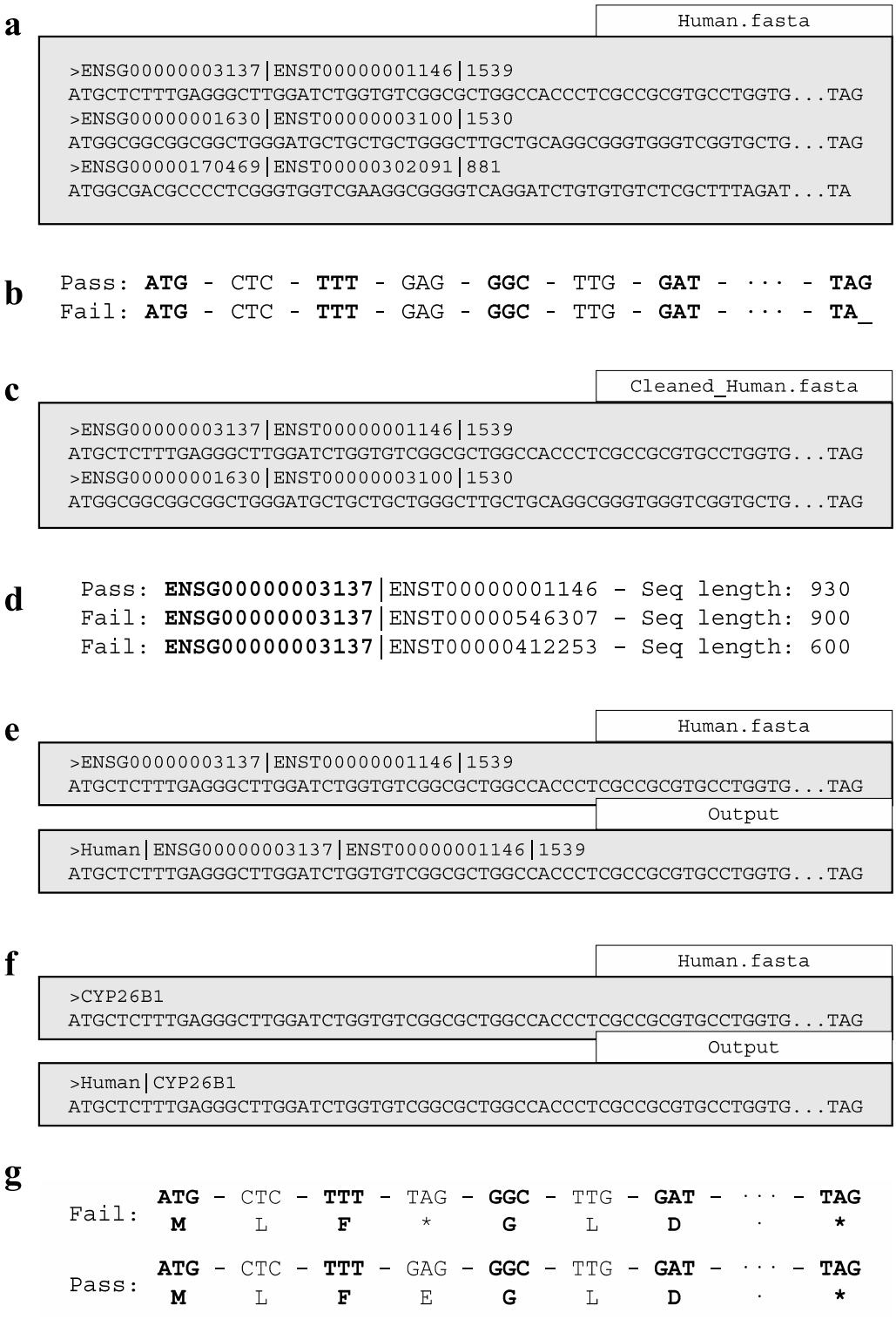
The more advanced 'ensembl_clean' function was designed to identify the longest nucleotide (canonical) transcript within an Ensembl nucleotide genome that passed the above QC step. This is achieved by exploiting the pattern of ensembl sequence identifiers, which consistently begin with the gene identifier followed by the transcript identifier (Figure 2.2d). The longest transcript is then identified for each ensembl gene identifier and saved within the output file.

Shorter transcripts along with the sequences that failed the QC filter are reported in a separate log file.

2.7.1.1 Additional options of 'clean' and 'ensembl_clean'

Both clean functions have a single enabled option ('rm_internal_stop') and two disabled options ('label_filename' and 'infer_ensembl_species') that may be manually configured by the user. The option 'rm_internal_stop' will remove sequences if they contain an internal stop codon (Figure 2.2g), those removed will be reported in the log file. It should be noted that while 'rm_internal_stop' is configurable, codeML does not permit nonsense mutations and the option should be enabled if the toolkit is being used for that purpose. The options 'label_filename' and 'infer_ensembl_species' alter sequence headers (i.e. Ensembl gene and transcript identifiers) by adding an additional identifier at the beginning of the header: 'infer_ensembl_species' adds the common species name of the respective Ensembl identifier (Figure 2.2e) and 'label_filename' adds the filename (without the file extension) (Figure 2.2f). It should be noted that executing a labeling option is required for enabling bmeTools to automate the creation of gene trees and setup of the codeML branch-site models (for details see Section 2.9.6 for automation and Section 1.3.2.2 for branch-site models).

Figure 2.2: Overview of 'clean' and 'ensembl_clean' functions.



FastA formatted files are shown as grey boxes and the associated white boxes show the filename. Data confirmation steps shown as readout beneath each example indicates if the results passed the check. The following QC checks are

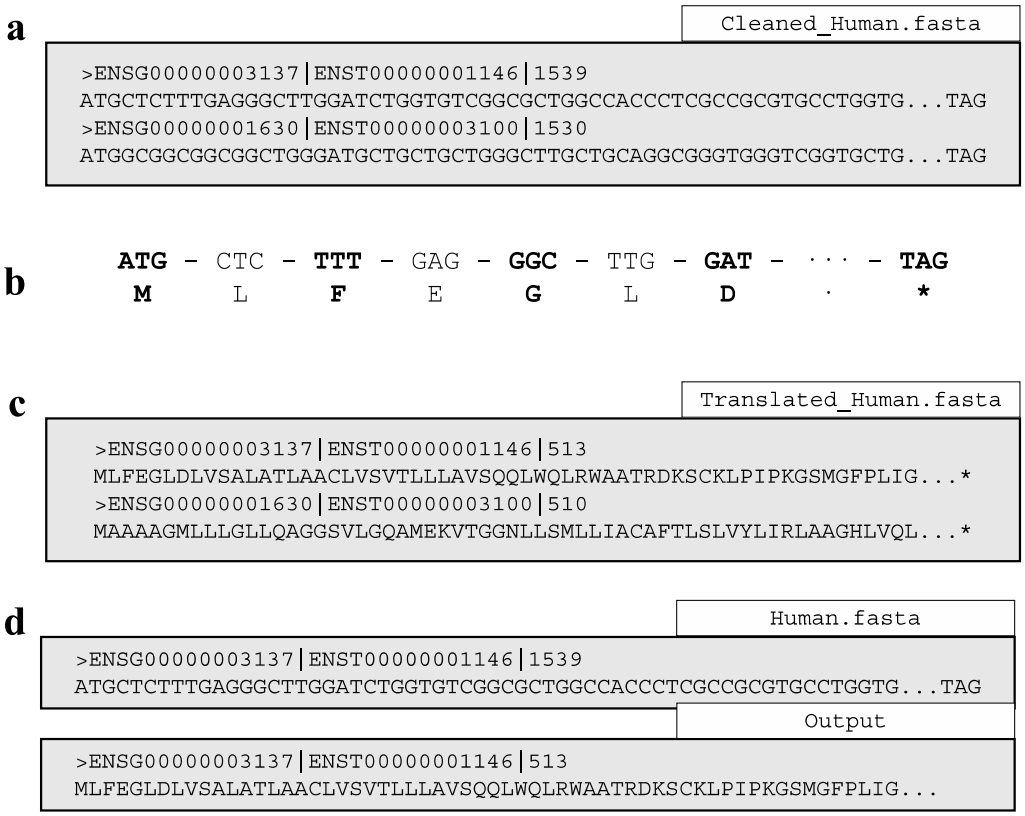
illustrated here: (a) Cleaning an input file, (b) initiates with codon confirmation, (c) only sequences that pass are saved in the output. If the ‘ensembl_clean’ function is invoked, in addition to codon confirmation, each transcript of an ensembl gene undergoes (d) a longest transcript confirmation and only the longest transcript is saved in the output. Two options are available to append a prefix to sequence headers: (e) ‘infer_ensembl_species’ to append the Ensembl genome, or (f) ‘label_filename’ to append the input filename. Invoking (g) ‘rm_internal_stop’ will remove genes that fail stop codon confirmation.

2.7.2 Function: translate

The ‘translate’ function translates nucleotide sequences that passed previous QC (Figure 2.3a). The function operates by splitting the nucleotide sequence into codons and then translating them into amino acids (Figure 2.3b). Translation is a mandatory step to produce alignments permitted by the codon substitution models of codeML (see Section 2.9.1) [Yang, 2007]. The resulting protein sequences are then saved (Figure 2.3c). If non-coding sequences (incomplete codons or internal stop codons) were not removed prior to invoking the ‘translate’ function, the function will produce a warning message. The warning reports that the function is designed to only translate protein-coding sequences and non-coding sequences will be removed from the pipeline and will be recorded within a separate log file.

The ‘translate’ function incorporates a single unique option ‘cleave_terminal’ and the previously described options of the clean functions (Section 2.6.1.1). If not manually configured, ‘cleave_terminal’ is enabled by default and is designed to cleave the terminal stop codon of each sequence (Figure 2.3d). The function and default status of the remaining options are detailed in Section 2.6.1.1.

Figure 2.3: Overview of 'translate' function.



FastA formatted files are shown as grey boxes and their filenames are given in white boxes. (a) Translating an input file using 'translate' initiates the translation procedure by separating the sequence (as in (b)) into each codon to determine the respective amino acid, (c) translated sequences are saved in the 'Translated' output file. (d) If the 'cleave_terminal' option is invoked, terminal stop codons will be removed from each applicable sequence.

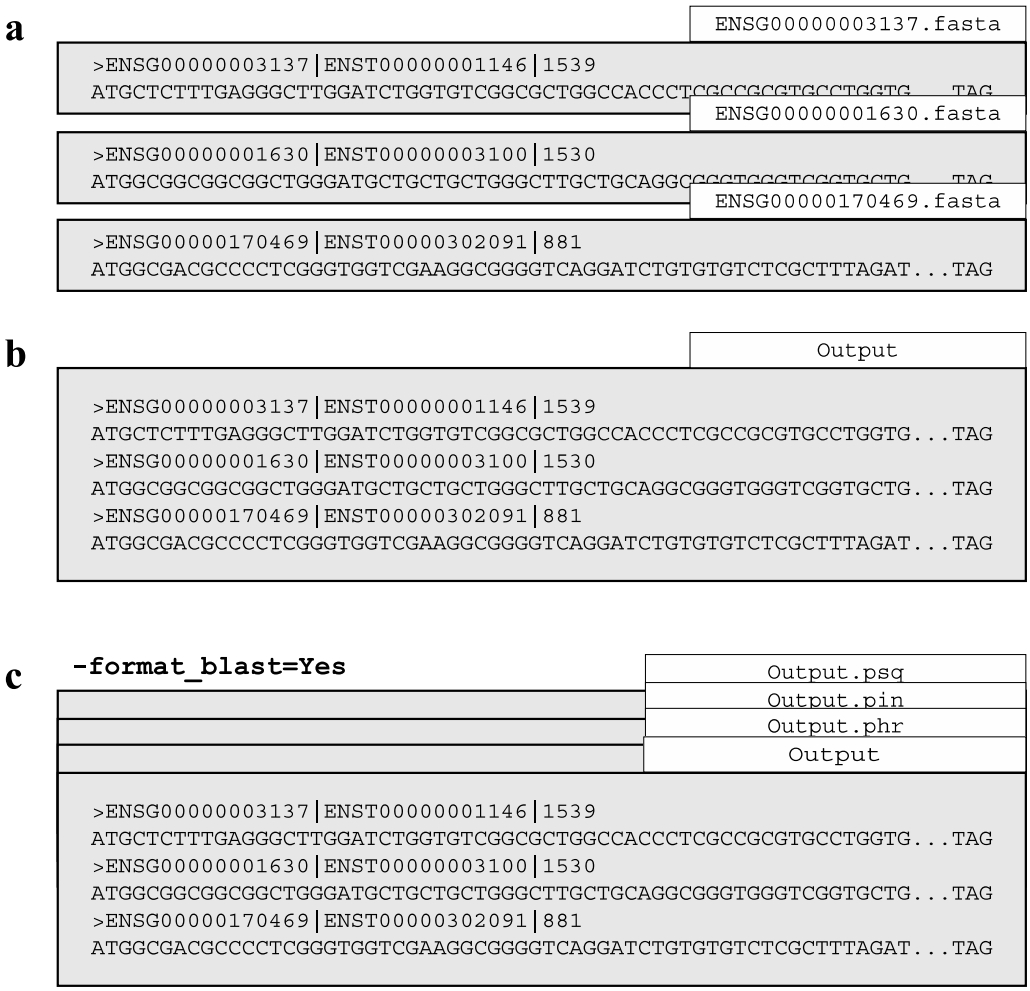
2.7.3 Function: create_database

The ‘create_database’ function was designed for inexperienced users to concatenate multiple genomes into the single database required for homology searching. The function operates by building the database a single sequence at a time (Figure 2.4a and b). The command-line version of BLAST requires additional commands to create a BLAST-formatted database. If the user enables the option ‘format_blast’ and BLAST is installed on the system the function will attempt to automate the additional steps required for producing a BLAST-ready database (Figure 2.4c). If ‘create_database’ is unable to create the BLAST-formatted database, a warning message will be produced.

2.7.4 Function: gene_selection

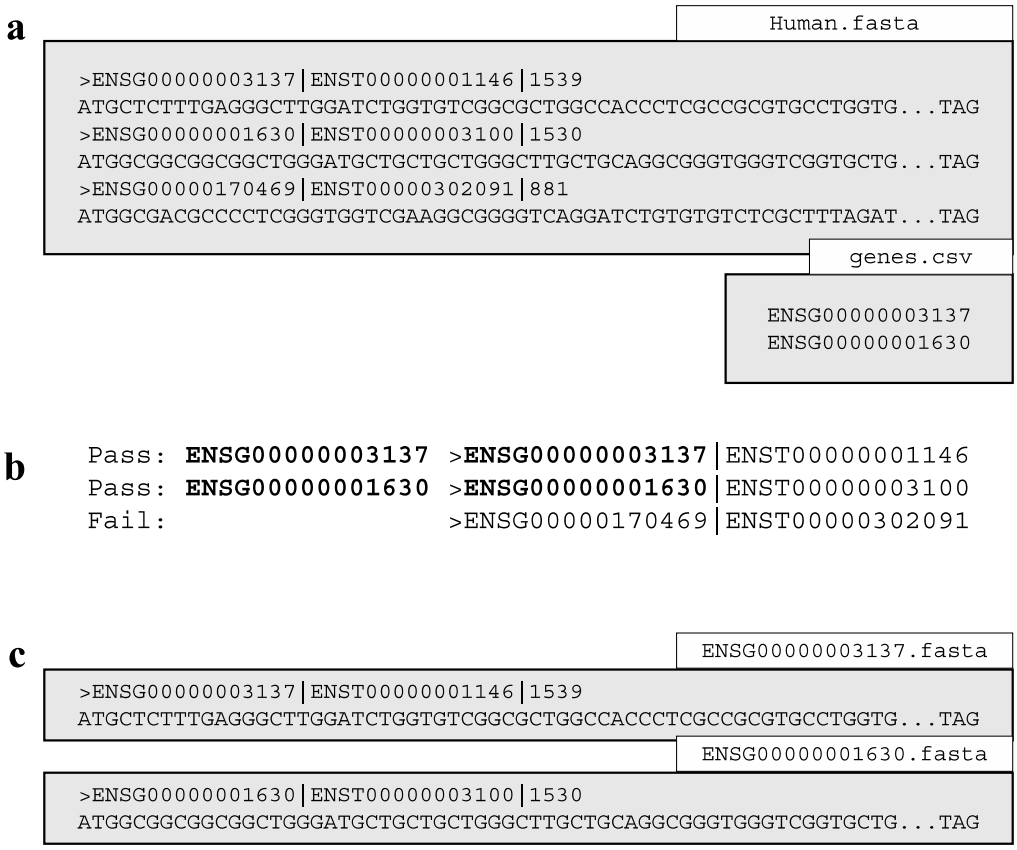
If the user is only interested in a subset of genes, the ‘gene_selection’ function was designed to enable the user to search a database for gene identifiers specified in a separate file. The function operates by searching the sequence headers of the database for matches with the user specified gene identifiers (Figure 2.5a). The matching process only requires the user-specified identifiers to match a portion of the database sequence headers (Figure 2.5b). The function saves a single sequence file for each matched identifier (Figure 2.5c). If a user-specified identifier matches more than a single sequence header in the database, or indeed no sequence in the database, the function will produce a warning message.

Figure 2.4: Overview of 'create_database' function.



FastA formatted files are shown as grey boxes and their filenames in white boxes. Invoking the 'create_database' function (a) combines numerous sequence files into (b) a single sequence database file. (c) Shows the 'format_blast' option that will generate the required database files for BLAST [Altschul *et al.,* 1990].

Figure 2.5: Overview of 'gene_selection' function.



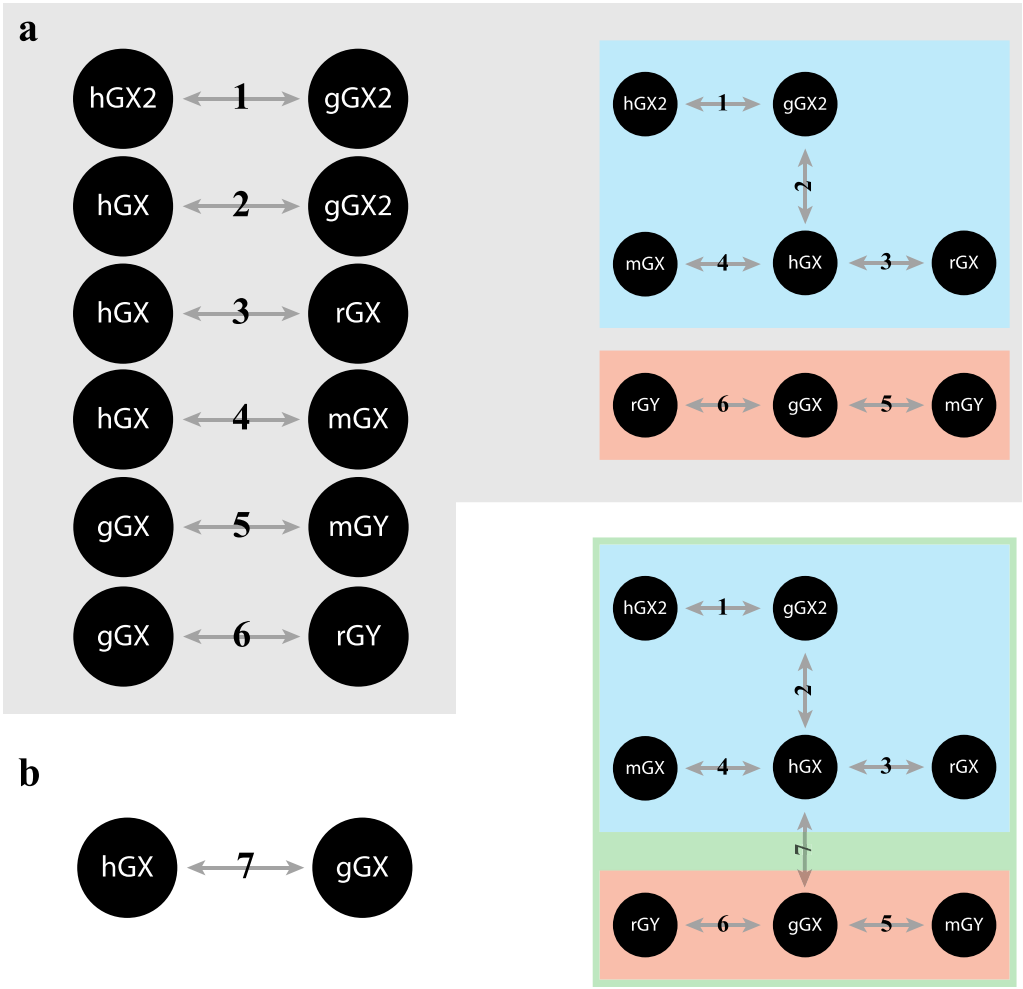
FastA formatted files are shown as grey boxes and their filenames in white boxes. Data confirmation steps indicate if the results passed the check. (a) The 'gene_selection' function requires two files to operate: a database (Human.fasta) and a user specified gene identifiers file (genes.csv). (b) The function operates using header confirmation to identify sequences in the database that match to those specified by the user. (c) The output of the function is a single sequence file for each user specified genes found.

2.8 Phase 2: Homology searching

The second phase of bmeTools is concerned with identifying homologous groups of sequences from either BLAST or HMMER searches. A homologous group of sequences is restricted to those that are found by a recursive connection function (Figure 2.6). Three types of homology are recognized by bmeTools: non-reciprocal (unidirectional), reciprocal (bidirectional), and best-reciprocal. "Non-reciprocal similarity" is characterized by sequence similarity that is only detected by one of the pair of sequences, commonly resultant of an E-value near the threshold. Non-reciprocal similarity is generally distantly related sequences. "Reciprocal similarity" is similarity identified by both sequences in the pair. Reciprocal similarity is typically closely related orthologs or paralogs. "Best-reciprocal similarity" requires that the sequences pass two criteria: (i) they are sequences from different species, and (ii) in the pair-wise connection each sequence finds no other sequence in the respective species with a lower E-value. These requirements limit identification to orthologs (non-orthologs may be identified due to identical E-values).

Each type of similarity connection is invoked using a separate function and will generate the families specific to that connection type. Each function is required to be linked to a protein sequence database (Section 2.6.3). The database is used to produce an output file of each similarity group containing the protein sequences of each member. Each protein sequence file then undergoes multiple sequence alignment using bmeTools currently supported methods (MUSCLE and PRANK). More experienced users may wish to use unsupported methods – the package is flexible enough to permit these changes.

Figure 2.6: Recursive homology group creation function



Genes are represented as circles and sequence similarity as grey arrows. Each of the similarity functions of bmeTools ('similarity_groups', 'reciprocal_groups', and 'best_reciprocal_groups') uses the same method for group creation. (a) The function begins by reading BLAST/HMMER input and sequentially identifying pairwise connections (connection requirements differ for each function, see Sections 2.7.2 and 2.7.3 for details). The arrow numbers indicate the sequential position (i.e. order of identification) of the respective pairwise connection. (b) The program will then generate the respective similarity groups using these connections. For example, two groups are created in the example above, one group contains five members (highlighted in blue) and another has three

members (highlighted in red). (c) The program will eventually encounter a new pairwise connection related to two previously generated similarity groups. (d) Such events will join the previous similarity groups – shown highlighted in blue and red – and a union of the groups – highlighted in green – will be reported in the results.

2.8.1 Core options

Each function within Phase 2 includes three threshold options (default = disabled). The three options enable the user to define threshold values for the E-value, alignment length, and percentage identity of each homology connection. Enabled thresholds must be passed for a pair-wise homology connection to be used. If the user has not enabled an E-value threshold, each function is designed to only incorporate E-values < 1, otherwise warning message is printed.

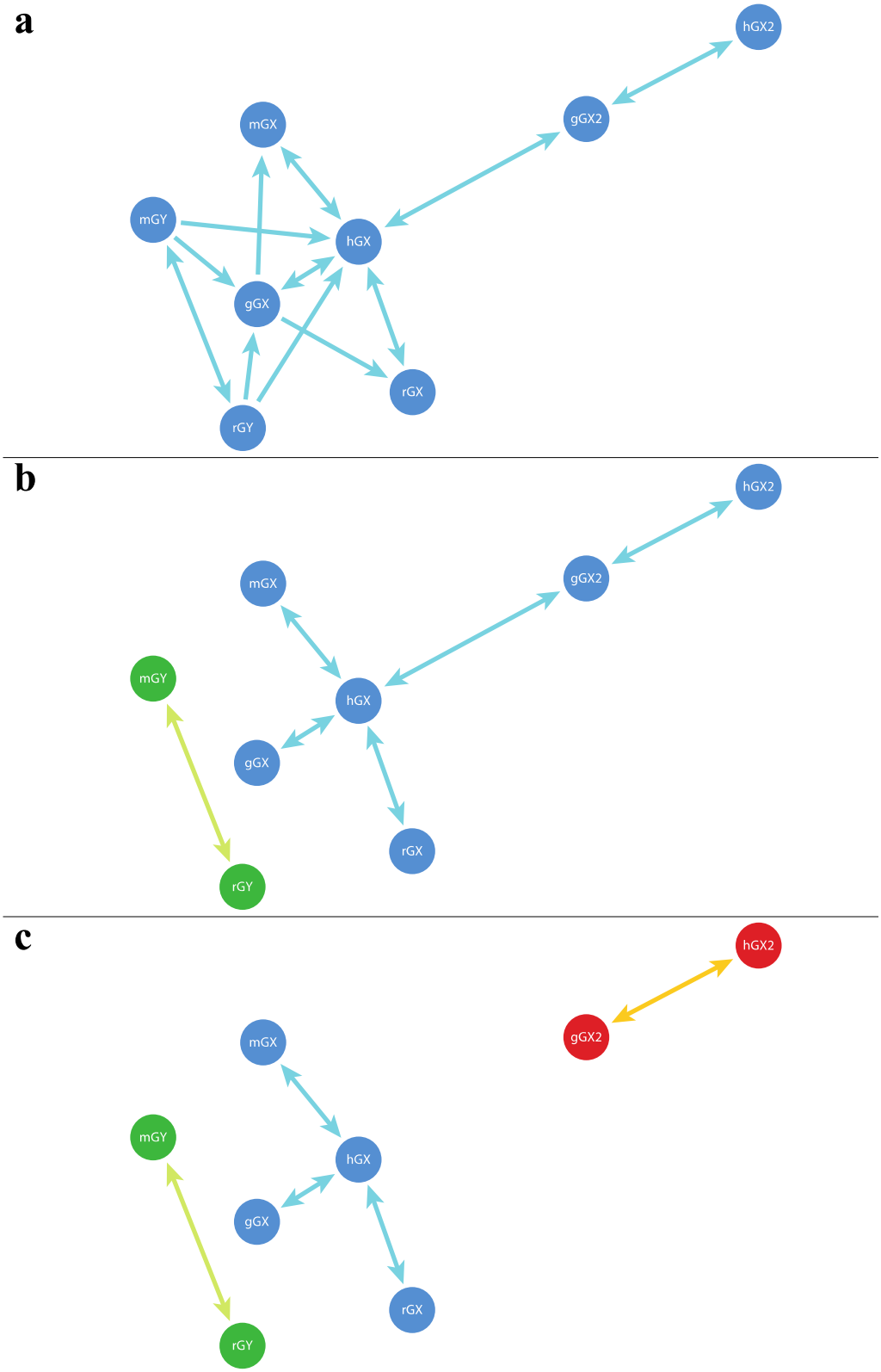
2.8.2 Functions: similarity_groups and reciprocal_groups

The ‘similarity_groups’ and ‘reciprocal_groups’ functions both construct sequence homology groups using a similar approach. Both functions iteratively read a single line of input (BLAST or HMMER output) and record only the name of the query and subject if they pass enabled thresholds. Limiting the recorded data of the homology search to sequence names and their respective role (query or subject) results in reduced computational requirements, increased function speed, and permits the function to parse larger BLAST or HMMER input files. Both functions are able to recognize and record input that denotes reciprocal homology of a previously recorded entry. Once each function has completed processing the input, the pair-wise homologs are used to build families (Figure 2.6). The ‘similarity_groups’ function allows both non-reciprocal and reciprocal connections within a sequence group (Figure 2.7a) whereas ‘reciprocal_groups’ is restricted to reciprocal connection within a sequence group (Figure 2.7b).

2.8.3 Function: best_reciprocal_groups

The ‘best_reciprocal_groups’ function constructs sequence homology groups by iteratively reading each line of input and storing the record within a database in reference to the query sequence. Once the function has completed parsing the input, the database is used to determine the best-homolog for each query sequence. This is achieved by identifying which subject sequence has the best E-value for each designated species. The designated best-hit for each query are then parsed to determine if the relationship is reciprocal (i.e. the subject sequence [as a query] identifies the query [as a subject]). If a query and subject are identified as best-reciprocal homology hits, they are used to create families (Figure 2.7c).

Figure 2.7: Similarity groups created by functions



The families created using (a) ‘similarity_groups’, (b) ‘reciprocal_groups’, and (c) ‘best_reciprocal_groups’. Shorter lines represent better E-values between two

sequences (circles). Lines with a single arrow represent non-reciprocal or unidirectional similarity connections. Lines with arrows on both sides represent reciprocal or bidirectional similarity connections. Sequence identifiers are shown for each sequence, different species are designated in this figure by lowercase letter at the beginning of each sequence identifier – h (human), m (mouse), r (rat), g (gorilla). (a) the 'similarity_groups' function connects all sequences as they are connected by either unidirectional or bidirectional similarity connections. (b) the 'reciprocal_groups' function creates two groups because the sequences mGY and rGY only exhibit a bidirectional similarity connection with each other. (c) the 'best_reciprocal_groups' function creates a three groups as the gorilla GX2 (gGX2) exhibits a stronger (i.e. lower e-value) bidirectional similarity connection with human GX2 (hGX2) than human GX (hGX).

2.9 Phase 3: Alignment assessment and phylogeny reconstruction

Phase 3 of bmeTools combines multiple third-party programs to automate the assessment of protein MSAs and enable simplified phylogenetic reconstruction. Alignment error is reported to cause high rates of false positives in a selective pressure analysis (Section 1.4.1). Therefore, bmeTools incorporates third-party programs for MSA comparison and scoring. A complete analysis of the MSAs from each method is recommended. The next step in this phase is the selection of the empirical model of evolution that best-fits each MSA [Foster, 2004; Keane *et al.,* 2006]. Phase 3 concludes with an automated method for phylogenetic reconstruction using the selected MSA and model of evolution. As bmeTools was created with inexperienced users in mind, the functions of this phase are primarily designed to interface with selected third-party programs. However, each step of this phase has been made optional if the user has different preferences.

2.9.1 Function: metal_compare

The 'metal_compare' function is designed to fully automate MSA comparison and scoring. The function operates using the third-party program MetAl [Blackburne and Whelan, 2012] to compare two protein MSAs (as described in Section 1.4.1). If MetAl indicates that the two MSAs are dissimilar, the function employs the third-party program noRMD [Thompson *et al.,* 2001] (as described in Section 1.4.1) to score each protein MSA using column-based similarity. The MSA with the highest noRMD (i.e. column-based similarity) score is then selected for subsequent analysis.

The ‘metal_compare’ function incorporates two additional options (‘metal_cutoff’ and ‘alignment_preference’) that may be configured by the user. The ‘metal_cutoff’ option assigns the numeric threshold determining MSA dissimilarity and by default is fixed at 5% (0.05). Alignment methods that yield MetAl scores lower than 0.05 are considered comparable, and in that case the ‘alignment_preference’ option may be used to specify an alignment method preference. If ‘alignment_preference’ is not configured the function by default will select the MSA from the first alignment method.

2.9.2 Functions: prottest_setup and prottest_reader

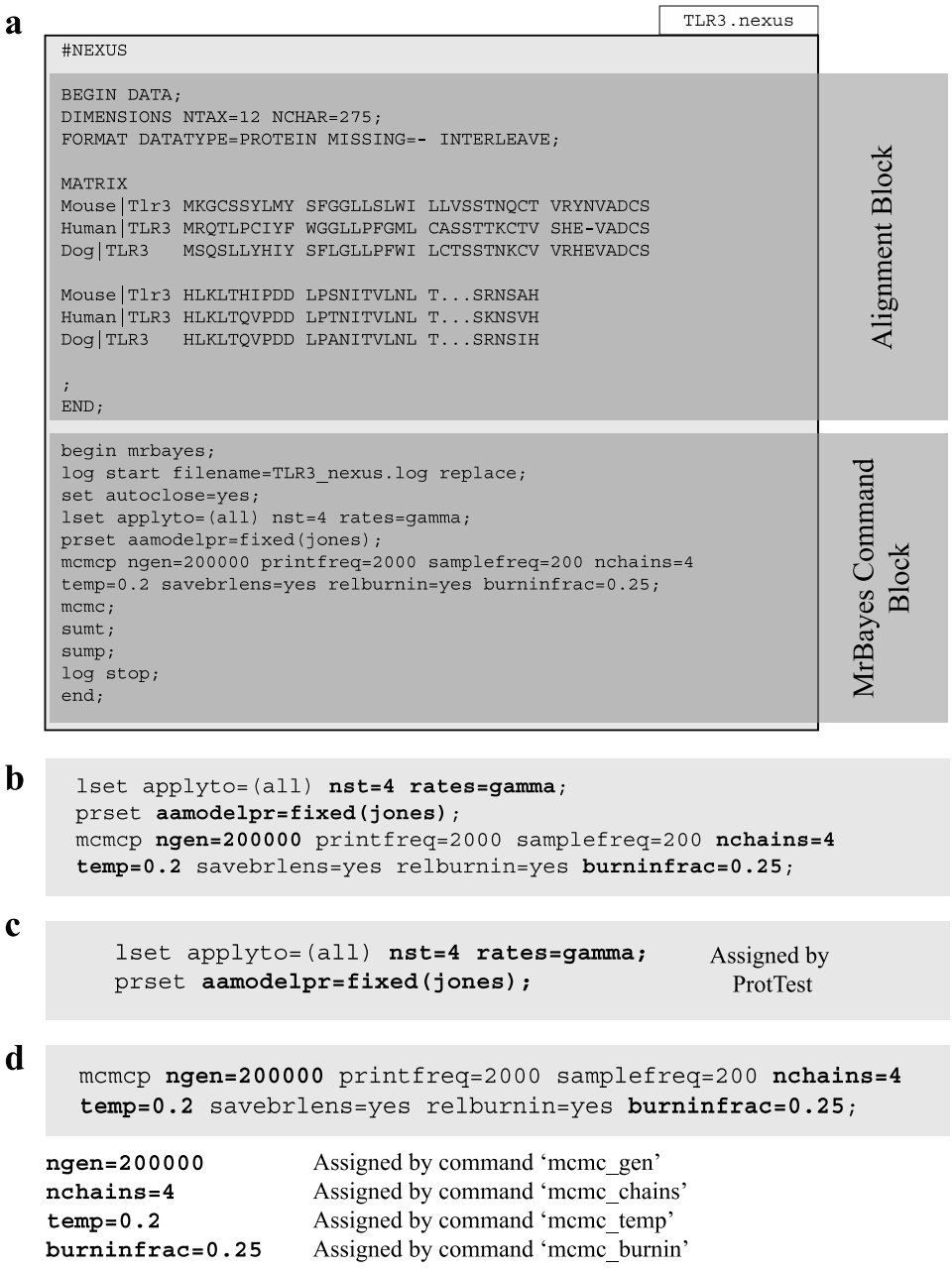
The ‘prottest_setup’ function is designed to automate the process of identifying the best-fit model of amino acid replacement for a specified protein alignment using the third-party program ProtTest3 [Darriba *et al.*, 2011]. The function is designed to test each amino acid replacement model in both the presence and absence of rate-heterogeneity (i.e. invariant sites, gamma categories). The ‘prottest_reader’ function automates the process of reading the output of ProtTest3. The function creates two output files: best_models.csv and best_supported_models.csv. The best models file reports the best-fit model of amino acid replacement (± rate-heterogeneity) reported by ProtTest3 whereas the best supported file reports the best-fit model of amino acid replacement (± rate-heterogeneity) supported by the third-party phylogenetic reconstruction program MrBayes [Ronquist and Huelsenbeck, 2003]. The two output files are given to enable the user to use different phylogenetic reconstruction software if desired.

2.9.3 Function: mrbayes_setup

The ‘mrbayes_setup’ function is designed to simplify the process of phylogenetic reconstruction using the third-party program MrBayes [Ronquist and Huelsenbeck, 2003]. The function begins by converting each protein MSA into the nexus format (Figure 2.8a). Each nexus-formatted MSA is then appended with a standardized MrBayes command block that defines the variables required for phylogenetic reconstruction (Figure 2.8b-d) (Section 1.3.2.3), they include the number of MCMC generations, the number of chains (trees) to be examined per generation, the temperature of the heated chain, the burn-in percentage, and the best-fit model of amino acid replacement (Section 2.8.2).

The ‘mrbayes_setup’ function incorporates multiple options (‘mcmc_gen’, ‘mcmc_chains’, ‘mcmc_temp’, ‘mcmc_burnin’) for permitting the user to alter variables within the MrBayes command block (Figure 2.8b-d). The ‘mcmc_gen’ option sets the number of generations for the phylogenetic reconstruction and should be increased from the default value of 200,000 if previous attempts failed to converge. The remaining options have the following recommended settings by default: ‘mcmc_chains’ i.e. the number of chains (default = 4), ‘mcmc_temp’ i.e. the temperature of the heated chain (default = 0.2), and ‘mcmc_burnin’, i.e. the burn-in percentage respectfully (default = 0.25).

Figure 2.8: Overview of 'mrbayes_setup function.



(a) The NEXUS file is separated into two blocks, a sequence alignment block and a MrBayes command block. (b) The specific commands within the MrBayes command block are each assigned default values (in bold) based on recommend values and previous commands. (c) The commands lset and prset by default are assigned from the 'best_supported_models.csv' file generated in Section 2.8.2. (d) The remaining commands are assigned based on recommended values, but may configured by the user is desired.

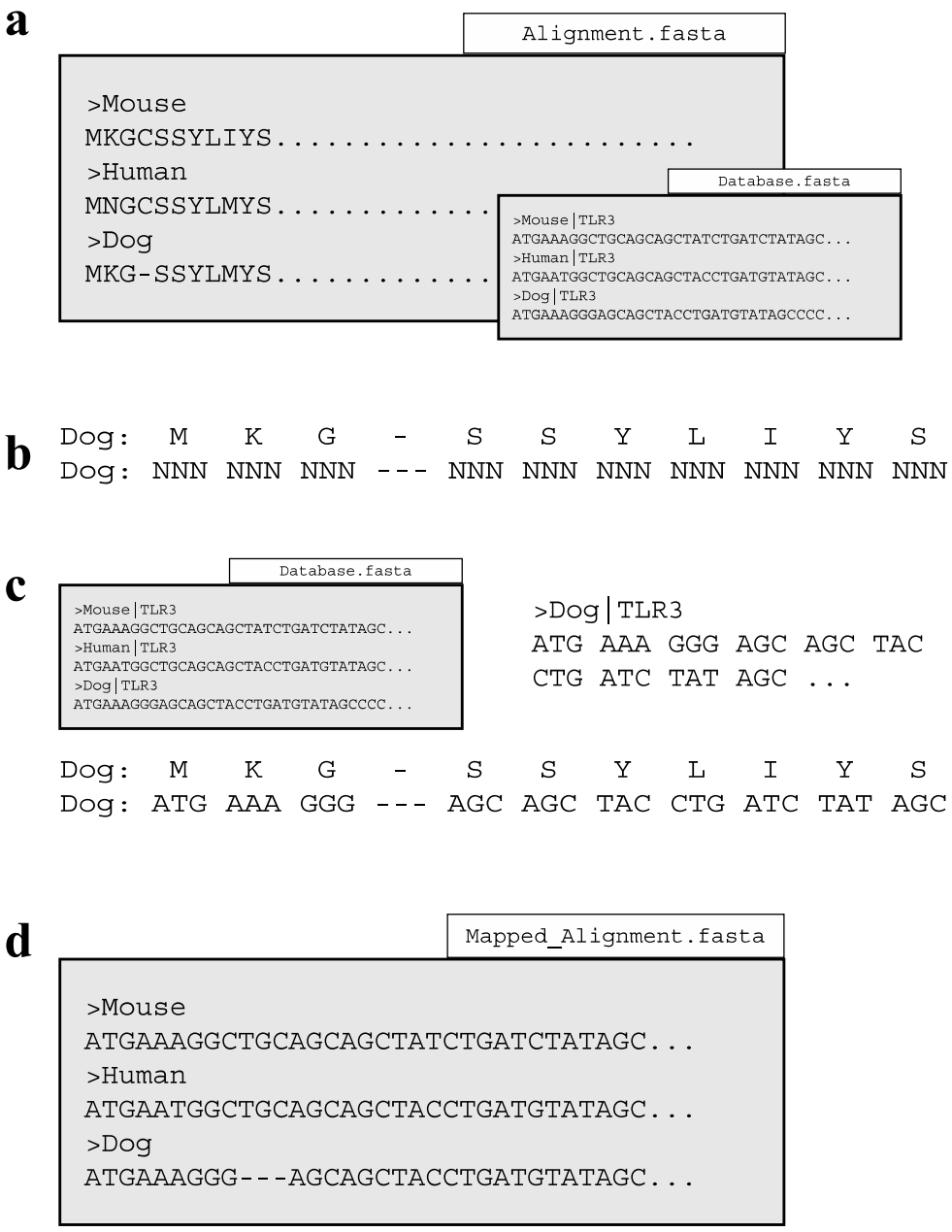
2.10 Phase 4: Selection analysis

Phase 4 of bmeTools automates selective pressure analysis using codeML from the PAML package [Yang, 2007]. Phase 4 is characterized by specific commands for the simple and advanced pipeline options (Figure 2.1). These pipeline-associated functions are designed to process the specific input of each pipeline into a standardized file format for the common functions used by both pipelines. Following standardization, bmeTools automates the normally labor-intensive process of creating the necessary files and directory structures for codeML. Phase 4 also incorporates a single optional function 'branch-label table' (Section 2.9.6) that may be invoked to enable the branch-site models of codeML.

2.10.1 Function: map_alignments

The 'map_alignments' function is designed to automate the conversion of protein MSAs to nucleotide MSAs. This process is mandatory - codon substitution models of codeML require nucleotide alignments. Protein-MSA guided nucleotide MSAs are generated rather than directly generating nucleotide MSAs because: i) each column within the protein MSA represents aligned codons and therefore avoids aligning incomplete codons or frame-shift mutations, and ii) protein MSAs represent a comparison of the phenotype-producing elements of protein-coding sequences (Figure 2.9a). The function begins by reading the protein MSA to map the non-gap position of each codon within the inferred nucleotide alignment (Figure 2.9b). The sequence of the mapped codons is then inferred using the nucleotide dataset (Figure 2.9c). If the mapping process results in no errors, the respective nucleotide MSA is created (Figure 2.9d). All errors detected by the function will be returned within a separate log file.

Figure 2.9: Overview of the 'map_alignments' function



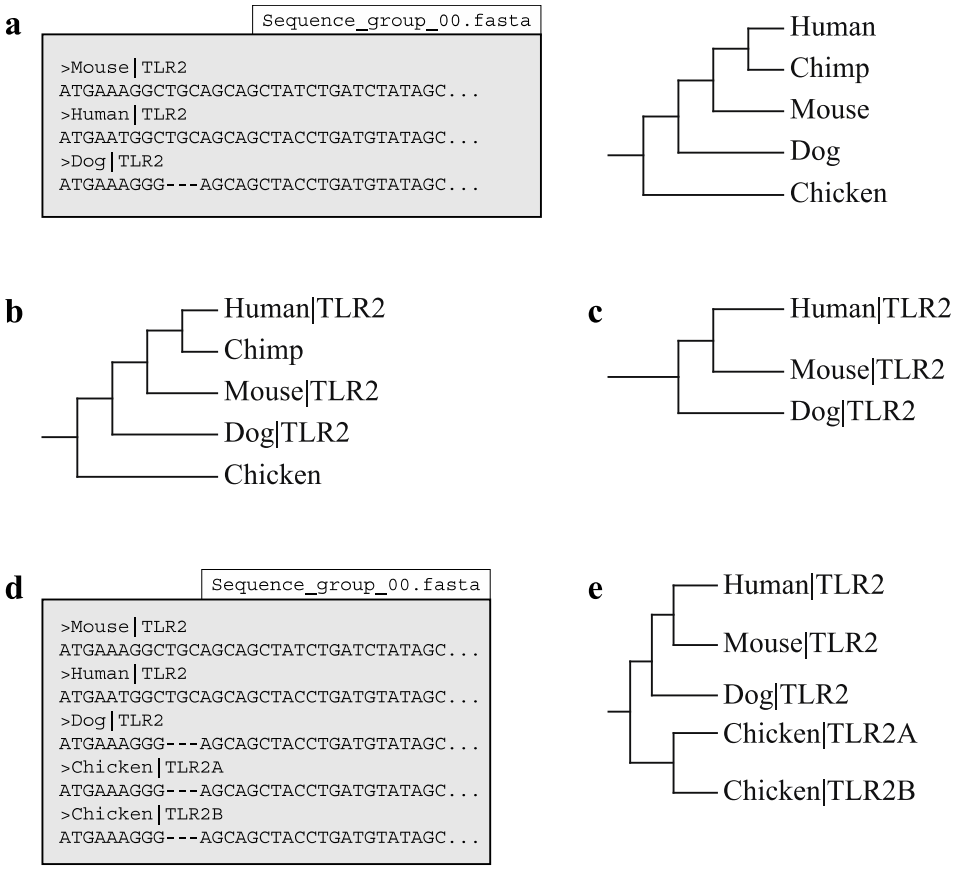
Sequence files are shown above as grey boxes indicating the sequences and white boxes indicating the filename. The 'map_alignments' function requires (a) two files to operate: a protein alignment (Alignment.fasta) and a nucleotide sequence database (Database.fasta). The function initiates by (b) mapping the gaps of the nucleotide alignment. (c) The nucleotide sequence of each alignment is then mapped using the sequence database to produce (d) the completely mapped output file.

2.10.2 Function: infer_genetree

The ‘infer_genetree’ function is designed to automate the creation of the corresponding gene tree for a user-specified MSA. This is achieved by associating the taxa specified on a user-defined species tree with the headers created by ‘label_filename’ and ‘infer_ensembl_species’ (Section 2.6.1.1) within the MSA. The function operates by first creating a copy of the species tree with the species names (Figure 2.10a). The species names are then replaced with their associated MSA headers (Figure 2.10b). If any species names remain after the MSA associating phase, the taxa and their respective branches are removed from the tree to create the finished gene tree (Figure 2.10c). It should be noted that the ‘infer_genetree’ function incorporates the non-standard python library dendropy [Sukumaran *et al.*, 2010], details on this requirement can be found in Section 2.11.2.

The ‘infer_genetree’ function incorporates a single option ‘allow_paralogs’ that is disabled by default. Normally, ‘infer_genetree’ is designed to only allow a single MSA header to associate with a species name (Figure 2.10d). If multiple headers are found to associate with a species name, bmeTools will produce a warning message. The ‘allow_paralogs’ may be enabled in these situations if the association error(s) are caused by within-species paralogs, in this case a gene tree will be created with associated headers shown as within-species paralogs (Figure 2.10e).

Figure 2.10: Overview of the 'infer_genetree' function.

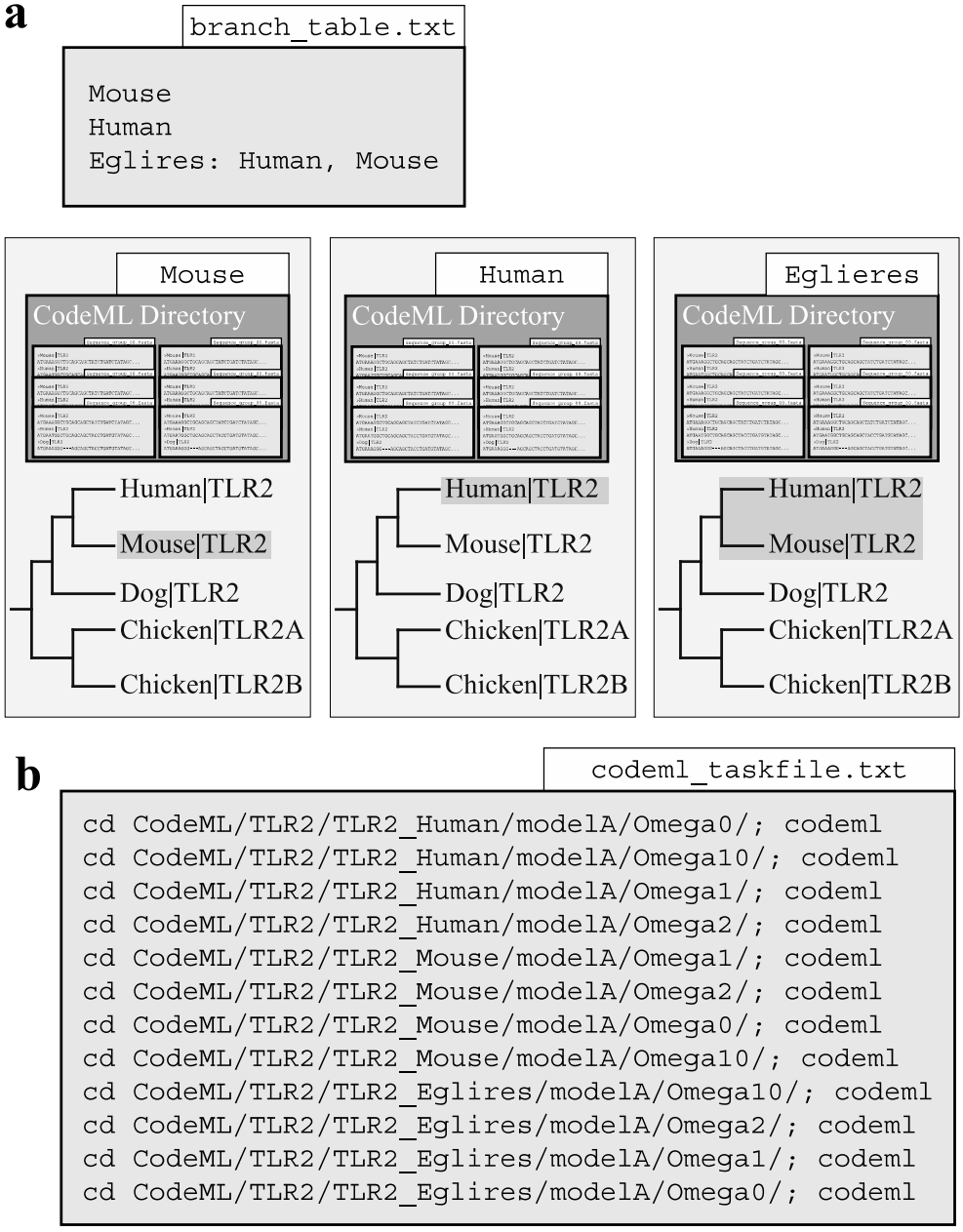


(a) The 'infer_genetree' function requires two files to operate: a nucleotide alignment (Sequence_group_00.fasta) and a species phylogeny to determine the phylogenetic relationship of the sequences within the alignment in relation to the species phylogeny. (b) The function begins by replacing each species name within the phylogeny with their respective gene identifier (i.e. Human → Human|TLR2) located in the nucleotide alignment. (c) The function then creates the gene phylogeny by removing the species that have not been replaced by a gene identifier. (d) If the nucleotide alignment specified by the user contains paralogs (Chicken TLR2A and TLR2B) bmeTools will produce an error message. (e) If the 'allow_paralogs' option is enabled the function will create a new branch to house the paralogs with the original species acting as an ancestral node.

2.10.3 Function: setup_codeml

The ‘setup_codeml’ function is designed to simplify the creation of the complex codeML directory structure. This is achieved by incorporating previously written in-house software ‘GenerateCodemlWorkspace.pl’ written by Dr. Thomas Walsh to produce the codeML directory structure [Walsh, 2013]. The purpose of automating the program ‘GenerateCodemlWorkspace.pl’ via ‘setup_codeml’ was to simplify input requirements and enable high-throughput analyses. The function requires only a protein-inferred nucleotide MSA (Section 2.9.1) and an associated phylogenetic tree (Section 2.9.6) to construct the directory structure for the codeML site-specific models [Walsh, 2013]. However, if the user has created the optional branch-label table (Section 2.9.6) and enabled the ‘label_table’ option the function will create the directory structure for the codeML branch-site models (Section 1.3.2.2 for description of models). Automating the branch-site models requires a specific directory for each species and/or lineage specified by the user in the optional branch-label table (Figure 2.11a). Next the ‘setup_codeml’ function will produce a codeML “taskfile” that contains each codeML command line command to be computed (Figure 2.11b). Following creation of the taskfile, a separate log file reporting the branch-site models that cannot be tested (due to missing taxa) is produced.

Figure 2.11: Overview of the 'setup_codeml' function.

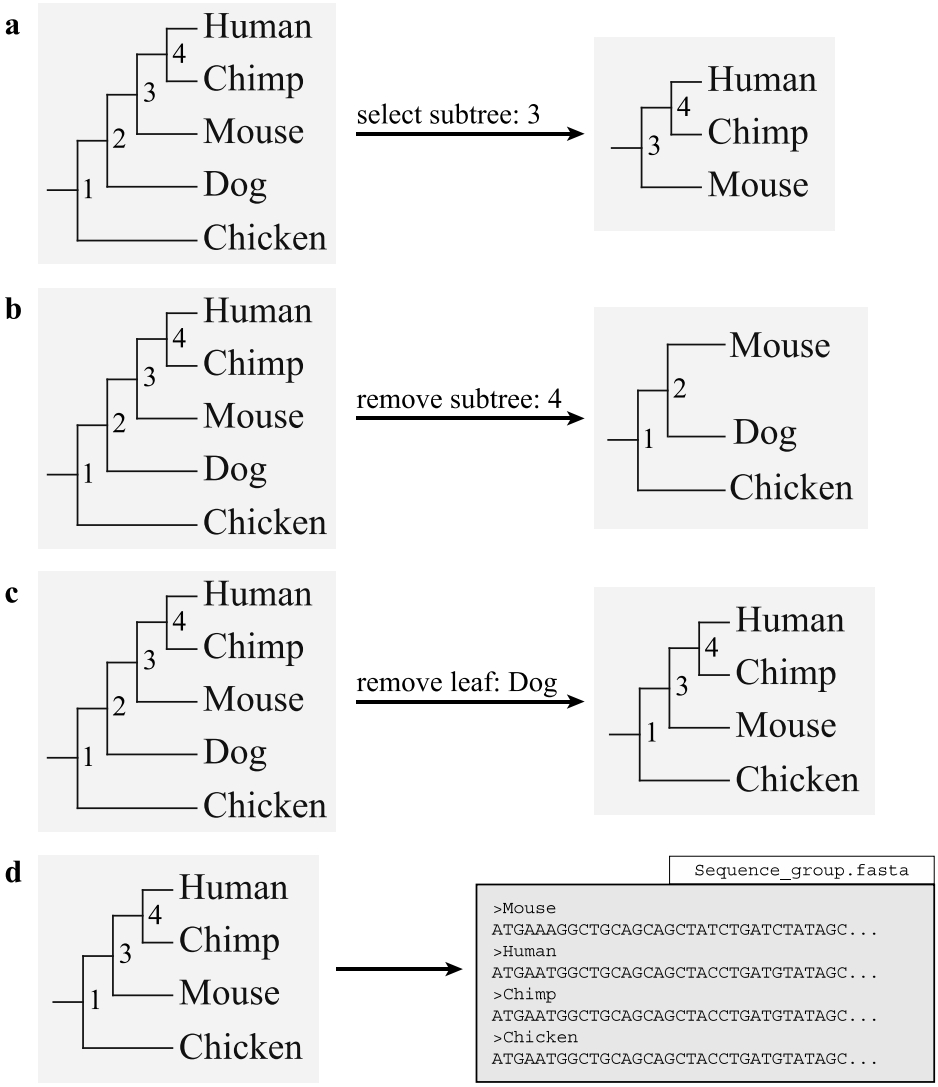


(a) Using the branch-label table (branch_table.txt) the function produces species-labelled (highlighted) phylogenies for each species or ancestral node specified and then automates the production of the codeML directory for the branch-site models. (b) The function terminates by producing a codeML taskfile with all the codeML command line commands required to complete the job.

2.10.4 Function: create_subtrees

The ‘create_subtrees’ function is designed for high throughput tree pruning. This optional step is often required to prune very large multigene family phylogenies into smaller phylogenies. User phylogenies may require pruning due to feasibility concerns as subfamilies increase data manageability and decrease computational requirements. Users may require this option for pruning out SGOs for selection analyses that are focused on single genes. The function operates by displaying the current phylogeny with a set of pruning commands/options. The user is then prompted to select of the four commands: ‘select subtree’, ‘remove subtree’, ‘remove leaf’, or ‘keep original’. If either ‘select subtree’ or ‘remove subtree’ is selected, the user is prompted to select a single node (numbered on the displayed phylogeny) for selection or removal respectively (Figure 2.12a/b). If ‘remove leaf’ is selected, the user is prompted to select a leaf label (sequence header) for removal (Figure 2.12c). If ‘keep original’ is selected the tree manipulation step is skipped. The ‘create_subtrees’ function will produce a protein sequence file of the remaining nodes in the phylogeny (Figure 2.12d). The protein sequence file is then required to undergo re-alignment and it proceeds from Phase 3 through the remainder of the pipeline (Figure 2.1). The ‘create_subtrees’ function will also produce a separate log file of the original phylogeny, the selected command, and the resulting phylogeny. The ‘create_subtrees’ function incorporates the non-standard python library dendropy [Sukumaran *et al.,* 2010] (Section 2.11.2).

Figure 2.12: Overview of the 'create_subtrees' function.



(a-c) shows an example of the node-labelled phylogeny displayed for the user is shown on the left for each option. (a) If the user specifies the 'select subtree' option along with a node, the function will create the subtree by separating the specified node from its next common ancestor node and returning the requested subtree. (b) The 'remove subtree' options functions identically to 'select subtree' except that requested subtree is discarded and rather the subtree containing the common ancestor node is returned. (c) The 'remove leaf' option will remove the specified taxa from the phylogeny. (d) The function terminates by creating sequence files for each pruned phylogeny.

2.10.5 Function: mrbayes_reader

If phylogenetic reconstruction has been performed by MrBayes then the 'mrbayes_reader' function is designed to replace 'infer_genetree' [Ronquist and Huelsenbeck, 2003]. The function operates by converting the nexus-formatted phylogeny into the newick format supported by bmeTools and codeML. If the function is unable to locate the original amino acid fasta-formatted MSA required by 'mrbayes_setup' (Section 2.8.3) the nexus-formatted MSA will be converted and placed with the newick-formatted phylogeny. It should be noted that 'mrbayes_reader' is unable to check phylogenies for convergence. Instead users are directed to confirm convergence using the third party program Tracer [Rambaut *et al.*, 2014].

2.10.6 Function: create_branch

The 'create_branch' function is designed to simplify the creation of the branch-label table required for the optional branch-site models of codeML (Section 1.3.2.2). The branch-label table (previously shown in Figure 2.11a) indicates the lineages or 'branches' that will undergo lineage-specific selection analysis, i.e. designation of the "foreground lineages" for codeML. Each line indicates one lineage, either a species or an ancestral node. Ancestral nodes (uniquely named by user [i.e. Eglires]) are followed by a list of descendant (extant) species (Figure 2.11a). The function operates by displaying a user-specified species phylogeny and promoting the user to select the species and/or ancestral nodes (numbered on the displayed phylogeny) of interest for the study (identical display methodology as described in Section 2.9.4 - see phylogeny in Figure 2.12a for example). When the user has finished their selection, the function will automatically produce the

branch-label table. It should be noted that this function is completely optional as the branch-label table may be easily created by hand. The ‘create_branch’ function incorporates the non-standard python library dendropy [Sukumaran *et al.*, 2010] (Section 2.11.2).

2.11 Phase 5: Selection analysis assessment

2.11.1 Function: codeml_reader

The ‘codeml_reader’ function is designed to parse the complex codeML directory structure and create simplified results for inexperienced users. This is achieved by incorporating in-house software ‘CreateSummaryReport.pl’ written by Dr. Thomas Walsh [Walsh, 2013] to produce the majority of the codeML results. In addition to automating ‘CreateSummaryReport.pl’, ‘codeml_reader’ produces supplementary output files that are designed for detection of false positives. If the user specifies a branch-label table (Section 2.9.6) ‘codeml_reader’ will produce codeML MSAs, these MSAs are characterized by the addition of i) the putative positively selected sites, and ii) the codons/amino acids that are positively selected in the respective lineage/s (Figure 2.13).

2.12 General requirements of the software package

2.12.1 Core functions

Each phase of bmeTools incorporates multiple core functions that are designed to minimize code redundancies. Primary core functions include: a log file creator, a sequence reader, a sequence writer, general sequence tools (translator, labeller, stop-codon detector, and length calculator), homology connection reader, homology group creator, general output creators, and sequence/alignment

verifiers. Each core function is designed with high flexibility to allow general use.

2.12.2 Software dependencies

The bmeTools software package is designed to minimize potential software dependencies, as additional software requirements may be difficult for inexperienced users to install on their systems. Currently, the non-standard python library dendropy [Sukumaran *et al.*, 2010] is the only dependency that remains in bmeTools. Dendropy incorporates numerous functions for storing phylogenetic information and simplifying tree-based analyses. Removal of dendropy would require substantial development time and the design of numerous core functions. However, installation of dendropy is simple and only requires a single command to be invoked by the user. If the user invokes a dendropy-dependent function, bmeTools is designed to print a warning message detailing the installation process of dendropy if the software is not installed.

2.13 Case study

2.13.1 Project overview

The feasibility of the bmeTools software package for simplifying large-scale selective pressure analysis was explored in our international collaboration on the bowhead whale genome project. From a biological perspective this investigation was of particular interest given the unusually long lifespan of bowhead whales and their apparent lack of cancer [George *et al.*, 1999; Caulin and Maley, 2011; de Magalhaes, 2013]. We compared the bowhead, minke, and orca to 20 other placental mammals, with marsupial, and monotreme outgroups (Figure 2.13).

This unique opportunity to explore cetacean molecular evolution was made possible by the recent publication of the minke whale genome [Yim *et al.,* 2013], our collaboration with the bowhead and orca whale genome-sequencing efforts and our bmeTools software package.

2.13.2 Analysis Pipeline

Here we applied the simple SGO option in the bmeTools package (Figure 2.1), this was mostly due to time constraints imposed by the consortium. A genome-wide analysis of the protein coding elements of all genomes (as per Sections 2.6.1 – 2.6.3) was performed by carrying out an all-vs-all BLASTp with an E-value cutoff of 10^{-7}. BLASTp results were then examined and 866 reciprocal SGOs were identified (as per Sections 2.7.2). Protein MSAs were created using PRANK to account for the high levels of indels observed in the bowhead and minke genomes. CodeML analysis focused on the branch-site models of each extant cetacean lineage and the two ancestral lineages (i.e. the most recent common ancestor (MRCA) of all cetaceans and the MRCA of baleen whales) using phylogenies inferred from a mammal species tree and protein-inferred nucleotide MSAs (Sections 2.9.1 - 2.9.3 and 2.9.6). CodeML results were subsequently verified for potential alignment-based false positives using codeML-enhanced alignments (Section 2.10.1).

2.13.3 Overview of Original Findings

Examining the three extant cetacean genomes for the number of SGOs exhibiting lineage-specific positive selection resulted in the following: bowhead (112 SGOs), minke (112 SGOs), and orca (28 SGOs). It should be noted that data

quality varied greatly across the cetacean genomes and may contribute to elevated values observed in the baleen whales (bowhead and minke). A functional survey by the genome-sequencing queried the putative cases of positive selection in the bowhead whale for previous published links to longevity and resulted in the following genes of interest: COQ6 (coenzyme Q6 monooxygenase), ERCC1 (Excision repair cross-complementation group 1), TP53TG5 (TP53 [tumor protein p53] target 5), TTI1 (TELO2 [telomere maintenance 2] interacting protein 1), and XRCC2 (X-ray repair complementing defective repair in Chinese hamster cells 2).

2.13.4 Data Quality Concerns and Importance of filters

In-depth analysis of the bowhead sequences of COQ6, ERCC1, TP53TG5, TTI1, and XRCC2 by the genome-sequencing effort identified evidence of potential annotation error. The genes of interest were subsequently re-annotated to eliminate the possibility of false positives (Section 1.4). Repeating the codeML analysis with the re-annotated bowhead COQ6, ERCC1, TP53TG5, TTI1, and XRCC2 resulted in no evidence of lineage-specific positive selection.

To minimize other potential false positives in the selection results, the MSAs of the 866 SGOs underwent strict data-quality filtering. The first imposed filter prohibited the presence of gaps in the MSA if created by unique insertions in either Bowhead or Minke sequences. The second imposed filter required unaligned Bowhead or Minke sequences to be at least half the length of their respective MSA. These two filters reduced the number of testable SGOs to 319. Examining the refined SGOs for evidence of lineage-specific positive selection

resulted in the following: bowhead (14 SGOs), minke (10 SGOs), and orca (6 SGOs), (Figure 2.13 and Table 2.1).

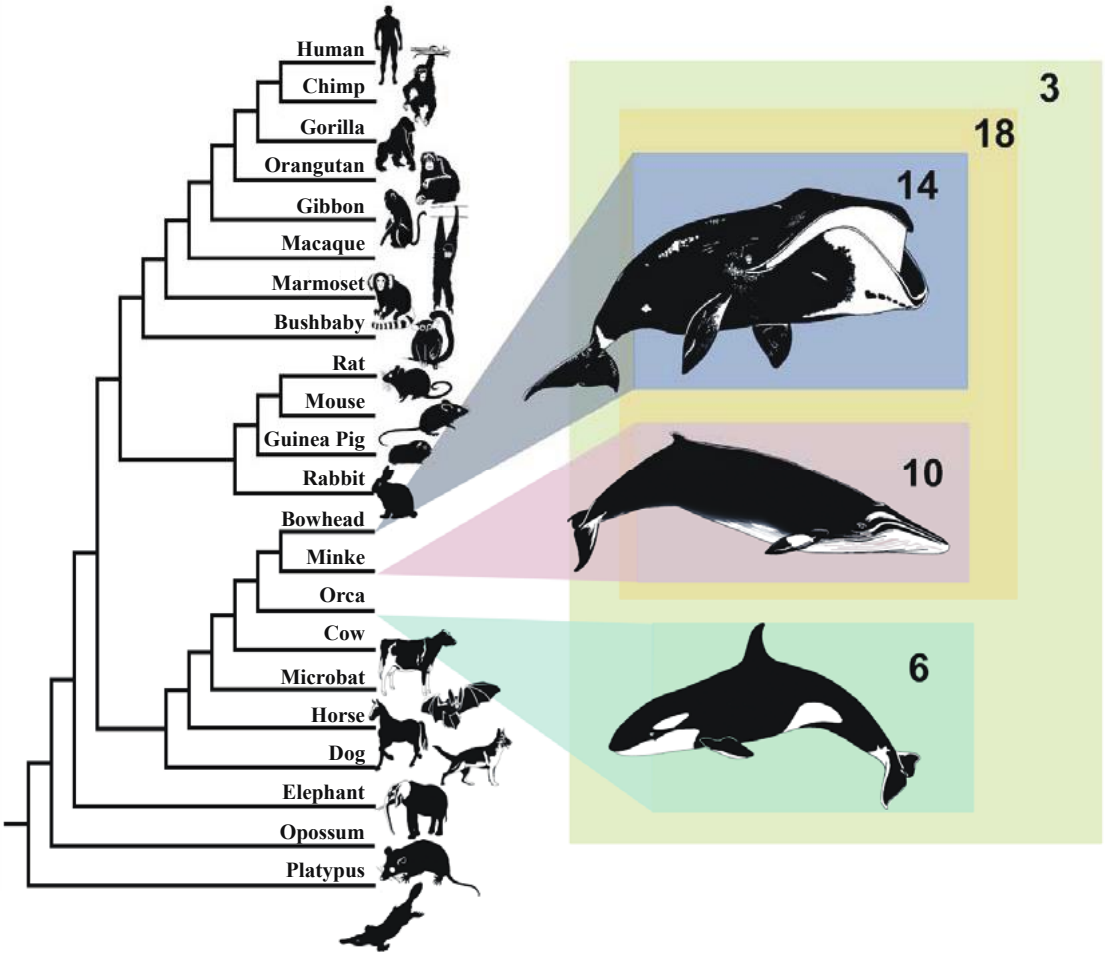
Both the re-annotation effort that led to the identification of the false positives in the bowhead lineage (i.e. COQ6, ERCC1, TP53TG5, TTI1, and XRCC2) and the substantial impact of data-quality filtering highlight the importance of genome quality and annotation in selective pressure analyses. We can only reliably predict selective pressure variation if each protein coding sequence is accurate. For this reason, high-quality genomes with high quality annotations are required for an accurate inference of positive selection and are strongly recommended for use with this pipeline.

2.13.5 Feasibility of bmeTools

Employing bmeTools in the analysis of the bowhead whale resulted in a streamlined and simple analysis that could be completed within the imposed time constraints. Data manipulation techniques that would have typically been time consuming and error-prone in an analysis of this size have been completely automated. For example, the inference of nearly one thousand gene phylogenies from a single species phylogeny was completed in less than five minutes without error. In addition, the interpretation of the codeML results and identification of a great number of alignment-based false positives (Section 1.4.1) was greatly simplified by employing the bmeTools-produced alignments. The inability of bmeTools to directly detect the annotation errors of the bowhead and minke whale indicate that the pipeline is only suitable for users experienced in identifying high-quality genomes with high quality annotations. For this reason,

bmeTools will remain a software package for internal laboratory (and trusted collaborators) use for the time being. Updates to bmeTools will be made in the near future to verify the suitability of the data for analysis to remedy this shortcoming (in addition to other improvements).

Figure 2.13: Phylogeny of mammals used in comparison of selective pressure variation.



The figure shows the mammals used in the selective pressure analysis conducted with bmeTools. The number of candidate genes under positive selection on each extant species – bowhead (blue), minke (red), and orca (dark green) – and ancestral lineages – baleen whales (i.e. bowhead and minke [in yellow]) and – cetaceans (i.e. orca, bowhead, and minke [in light green]) is indicated within the color boxes.

Table 2.1: Proteins with evidence of lineage-specific positive selection

Lineage	Protein	Position of Positively Selected Residues
Bowhead	BAMBI	BEB: 204
Bowhead	C19orf38	BEB: 88
Bowhead	CCDC181	BEB: 65, 89, 91, and 265
Bowhead	TBC1D19	BEB: 469
Bowhead	ZNF646	BEB: 211, 608, 702, 796, and 829
Bowhead	ARL6IP6	BEB: 18
Bowhead	DMP1	BEB: 322 and 386
Bowhead	IFI30	BEB: 316
Bowhead	GAPT	BEB: 197**, 228**, 265**, 269**, 271*, and 278**
Bowhead	SPP2	BEB: 15
Bowhead	C22orf15	BEB: 46* and 75
Bowhead	ZGLP1	BEB: 333**
Bowhead	LIME1	BEB: 148 and 330
Bowhead	Gm15440	BEB: 117
Cetacea	SRRD	BEB: 176 and 289
Cetacea	TMEM119	BEB: 102 and 340**
Cetacea	C12orf68	BEB: 124 and 150**
Minke	NDC80	BEB: 193**
Minke	CRHBP	BEB: 331**
Minke	MLEC	BEB: 15** and 115*
Minke	AKAP12	BEB: 160, 217, 250, 268, 295, 592, 609, 627, 821, 1065, 1563, and 1923**
Minke	DMP1	BEB: 24, 84, 145, 153**, 184, 214, 241, 261, 279, 292, 322, 342**, and 360
Minke	PIGV	BEB: 232
Minke	GPLD1	BEB: 690, 831**, and 865
Minke	C3orf49	BEB: 237
Minke	MN1	BEB: 162, 1089**, and 1152
Minke	LIME1	BEB: 140, 257**, and 326
Mysticeti	BAMBI	BEB: 24**, 127**, and 181
Mysticeti	ESPL1	BEB: 166, 1054**, 1081, 1220, 1748, and 2005
Mysticeti	SRRD	BEB: 148, 316, and 319

Table 2.1: Proteins with evidence of lineage-specific positive selection

Lineage	Protein	Position of Positively Selected Residues
Mysticeti	IGSF6	BEB: 65**, 69, 106, and 207
Mysticeti	ODF1	BEB: 41**, 64, 76**, and 99**
Mysticeti	C4orf29	BEB: 35, 56, 317, 342, 344, and 375
Mysticeti	DNALI1	BEB: 95 and 355
Mysticeti	BBS12	NEB: 535*
Mysticeti	TC2N	BEB: 109, 154, 301, and 399
Mysticeti	C17orf59	BEB: 57** and 243
Mysticeti	PPP1R15A	BEB: 75, 113, 624*, and 691
Mysticeti	PODXL	BEB: 774 and 779
Mysticeti	BRICD5	BEB: 19**
Mysticeti	CXCL17	BEB: 17, 21**, 28, 45, 110, 113, and 124
Mysticeti	C22orf15	BEB: 36
Mysticeti	HPS6	BEB: 14, 152**, 187, 248, 420, 557**, and 574
Mysticeti	C9orf131	BEB: 251**, 584, 713, 1028, 1048, and 1053
Mysticeti	MBLAC1	BEB: 22, 23**, 218, and 255**
Orca	AP5M1	BEB: 15, 93**, 130, 406, 439, 532**, 533**, 534**, and 535**
Orca	IL33	BEB: 37, 45**, 77, 173, and 217
Orca	GAPT	BEB: 173, 192**, 193, 195, 225, 271, 272*, and 276**
Orca	APOA2	BEB: 33**, 64, and 94
Orca	G0S2	BEB: 22**, 54, 56**, and 101
Orca	LIME1	BEB: 205, 252, 256, 274, 276, and 290

The table shows the proteins identified under lineage-specific positive selection. The proteins are sorted into the respective lineage in which the positive selection was observed. Each protein is reported with the positions of the residues identified as positively selected and the empirical Bayes model used – BEB or NEB – to calculate the posterior probability of the residues being under positive selection (see Section 1.3.2.2 for details). The probability of the residues being under selection is also reported at three cutoffs: residues with a probability only

greater than 0.50 are unmarked, residues with a probability only greater than 0.90 are marked with a single asterisk (*), and residues with a probability greater than 0.95 are marked with a double asterisk (**). See table of abbreviations for the name of each protein shown above.

Chapter 3: Evolutionary immunology: exploring the potential of identifying species-specific innate immune responses from sequence data

3.1 Chapter Aim

Human and mouse last shared a common ancestor approximately 100 million years ago. Therefore their systems such as the innate immune system have approximately 200 million years of independent evolution. The primary aim of chapter three was to assess what level of lineage specific positive selection had occurred in the human and mouse lineages since they diverged and whether these positively selected residues were fixed or variable in modern populations. Then we wished to determine if positive selection had occurred whether it was confined to specific pathways or interacting proteins. And finally, with the growing body of literature on human mouse innate immune discordance we wished to determine if positive selection identified in the innate immune system correlated with known phenotypic discordance in immune response between human and mouse.

To achieve these goals, the bmeTools package developed in chapter two was applied as we carried out a genome-wide lineage-specific selective pressure analysis in the human, mouse, and multiple ancestral lineages of the Euarchontoglires clade (primate, rodent, and murinae).

3.2 Introduction

The study of innate immunology relies predominately upon the mouse as a model organism. However, it is becoming increasingly evident from the literature that there are inconsistencies between human and our model organisms in terms of response to pathogenic infection [Mestas and Hughes, 2004]. For example, a key receptor in innate immunity is toll-like receptor (TLR) 4, upon detecting gram-negative bacteria TLR4 triggers the activation of factors (MYD88, TIRAP, and TRAF6) that regulate immune and inflammatory responses [Kawai and Akira, 2010]. Coding sequence mutations in TLR4 have been shown to produce nickel sensitivity in human but not mouse [Schmidt *et al.,* 2010]. This unpredictability of immune response between human and our model organisms is particularly problematic in clinical trials as outlined in Section 1.1.4 [Stebbings *et al.,* 2007]. Discordance is a significant issue contributing to the unpredictability of modelling human disease [Mestas and Hughes, 2004]. The completion of a large variety of vertebrate genomes, including the recently completed 1000 Human and 17 Mouse genome projects [Keane *et al.,* 2011; Abecasis *et al.,* 2012], and the increased quality attained for the Neandertal [Green *et al.,* 2010], provide us with a unique opportunity to approach the problem of discordance in immune response from an evolutionary perspective, permitting us to address the molecular underpinnings of known discordance cases along with predicting novel discordance candidates at the molecular level.

The "Red Queen hypothesis" describes the evolutionary arms race between host and pathogen [Van Valen, 1973]. This dynamic results in signatures of positive

selection (synonymous with protein functional shift) detectable at the molecular level in immune related genes [Sawyer *et al.,* 2005]. It is not surprising therefore that genes of the immune system tend to have the highest levels of positive selection in comparison to other functional categories of genes [Kosiol *et al.,* 2008]. Indeed it is known that this can also occur at the species-specific level [Sawyer *et al.,* 2005]. The relationship between positive selection and protein functional shift has also recently been elucidated using rational mutagenesis of positively selected residues in a human enzyme called myeloperoxidase (MPO) that is produced by neutrophils [Loughran *et al.,* 2012]. MPO exhibits a novel chlorination activity among the mammalian heme peroxidase family [Loughran *et al.,* 2012]. Three positively selected residues (N496, Y500, and L504) were attributed to conferring chlorination activity as mutating them to their respective ancestral peroxidase residues abolished novel function [Loughran *et al.,* 2012]. These and other recent studies have shown a clear relationship between positive selection and protein functional shift [Farhat *et al.,* 2013; Moury and Simon, 2011]. In this chapter we set out to determine if orthologous innate immune proteins function in precisely the same way across different species using an *in silico* approach. We also set out to predict which innate immune proteins are most likely to have altered function in a given species thereby contributing to observed discordance in response to infection.

Here we have used patterns of conservation and variation to map regions of possible discordance in the innate immune system of human, mouse, and their closest MRCAs. We have combined comparative genomics, molecular evolution, structural modeling and population data analyses to identify both human and

mouse-specific adaptive evolutionary events. Not only do our results correlate with known discordance cases from the literature - thereby showing the value of our approach - but these findings also provide us with a platform for the prediction of novel molecular and phenotypic discordance of relevance to modeling of human disease.

3.3 Materials and Methods

3.3.1 Generating the vertebrate innate immune gene dataset

The full list of documented innate immune genes was downloaded from InnateDB [Lynn *et al.,* 2008] (Appendix 3.1). The InnateDB database is manually curated and requires experimental evidence for all entries [Lynn *et al.*, 2008]. The InnateDB dataset was filtered for documented evidence of a human ortholog, which restricted our dataset to 725 Ensembl gene identifiers. This filter was imposed as our primary interest is in discordance between human and mouse innate immune response and due to InnateDB housing data for human, mouse, and bovine.

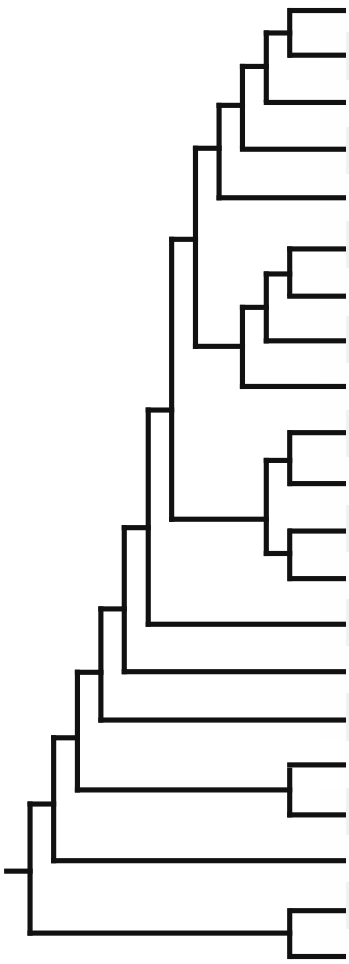
The 21 high coverage (>6x) vertebrate genomes were downloaded from Ensembl BioMart (Ensembl Gene 60) [Kinsella *et al.,* 2011] (Table 3.1). The longest transcripts for all protein coding genes were taken for each genome [Kinsella *et al.,* 2011]. The downloaded genomes underwent a simple quality check using the 'clean' function within the program 'FastaTools.py' (Appendix 3.2 – function in bmeTools package [Section 2.6.1]). This function tested the integrity of the protein coding sequences by ensuring they had complete codons. Transcripts then underwent translation into proteins using the 'translate' function from the

program ‘FastaTools.py’ (Appendix 3.2 - function in bmeTools package [Section 2.6.2]).

Homologs for the 725 innate immune genes were identified from the 21 vertebrate genomes using standalone BLASTp (v2.2.23+) [Altschul *et al.*, 1990]. The query sequences required by BLAST were provided by the ‘GeneSelect’ function within ‘FastaTools.py’ (Appendix 3.2 - function in bmeTools package [Section 2.6.4]). The database to query was created using the ‘database’ function within ‘FastaTools.py’ (Appendix 3.2 - function in bmeTools package [Section 2.6.3]). BLASTp was run with an E-value threshold of 1.0e-10.

Initial clustering of the BLAST results using MCL produced large multigene families containing more than one innate immune gene each. Therefore, a conservative E-value threshold of 1.0e-100 was imposed along with an alignment length threshold of approximately 85-87%. These thresholds were applied as they maximized the number of gene families that contained only orthologous genes while accounting for the possible presence of species-specific gene duplications, which are known to affect selective pressure variation [Zhang, 2003]. A best reciprocal BLAST approach was deemed unsuitable, as the approach is unable to account for gene duplications. Gene families were created with these thresholds using the ‘brc’ or Reciprocal Check function in ‘BLASTer.py’ (Appendix 3.2 - function in bmeTools package [Section 2.7.2]). Gene families that contained six or more members were retained for further analysis.

Table 3.1: Details on the vertebrate genomes used in this study.



Species	Assembly	Coverage	Contig N50
Human	GRCh37	High	38Mb
Chimpanzee	CHIMP2.1.4	6	50kb
Gorilla	gorGor3.1	2.1 & 35	11kb
Orangutan	PPYG2	6	15kb
Marmoset	culJac3	6	29kb
Mouse	NCBIM37	High	32Mb
Rat	RGSC3.4	6	52kb
Guinea Pig	cavPor3	6.79	80kb
Rabbit	OryCun2	7	64kb
Dog	CanFam3.1	7.6	267kb
Horse	EquCab2	6.79	112kb
Cow	Btau_4.0	7	78kb
Pig	Sscrofa9	24	69kb
Elephant	loxAfr3	7	69kb
Opossum	MonDom5	7.33	108kb
Platypus	OANA5	6	11kb
Chicken	WASHUC2	7.1	45kb
Zebrafinch	taeGut3.2.4	6	39kb
Xenopus	JGI41	7.6	22kb
Fugu	FUGU4.0	8.5	52kb
Zebrafish	Zv9	7.5	1Mb

The phylogenetic relationship, common name, genome assembly, fold coverage, and contig N50 are given for each the 21 vertebrate genomes used in this chapter. All details given were gathered from Ensembl [Kinsella *et al.*, 2011], NCBI Assembly, and respective genome publications.

3.3.2 Selection of multiple sequence alignment method

MSAs were created for each gene family using two distinct alignment methods under standard conditions: MUSCLE (v3.8.31) [Edgar, 2004] and PRANK (v100802) [Loytynoja and Goldman, 2005]. MUSCLE was selected due to the reported accuracy and efficiency of the algorithm [Edgar, 2004]. PRANK was selected due to the unique ability of the algorithm to distinguish insertions from deletions within an alignment [Loytynoja and Goldman, 2005]. Excluding the +F option for PRANK designates that the algorithm should not align inferred insertions, but the algorithm still correctly distinguishes insertions from deletions but inferred insertions may be aligned [Loytynoja and Goldman, 2008]. Reports indicate that including the +F option improves sequence alignments and downstream analyses in comparison to traditional alignment methods [Loytynoja and Goldman, 2008]. However, the +F option is only recommend if the phylogeny can be fully trusted [Loytynoja and Goldman, 2008; Loytynoja and Goldman, 2010]. For this reason, the PRANK alignments were constructed without the +F option.

The MUSCLE and PRANK MSAs for each family were subsequently compared by MetAl (v1.1.0) [Blackburne and Whelan, 2012] using the function 'scoreMetAl' from the 'metalman.py' program (Appendix 3.2 - function in bmeTools package [Section 2.8.1]). MetAl measures variation between MSAs produced by different alignment methodologies. Using the default metric (d-pos) a percentage is returned indicating the difference between the alignment methods [Blackburne and Whelan, 2012]. MSAs were treated as identical if the returned percentage was ≤ 0.05 (5%) [Communication with Blackburne and Whelan]. If

alignments were reported as identical, we chose the MUSCLE alignment. If the alignments were >5% different they were subsequently compared using noRMD (v1.2) [Thompson *et al.*, 2001] with the function 'noRMDchk' from the 'metalman.py' program (Appendix 3.2 - function in bmeTools package [Section 2.8.1]). MSAs with greater column-based similarity will return higher noRMD scores. The methods that returned the higher noRMD was used, if the noRMD scores were identical, again MUSCLE alignments were favoured.

3.3.3 Selecting the best-fit model of protein evolution using ProtTest3

ProtTest3 (v3.0) [Darriba *et al.*, 2011] was selected for identifying the best-fit substitution model. Using the function 'SetupProtTest' in the program 'MUSCLEman.py' (Appendix 3.2 - function in bmeTools package [Section 2.8.2]), each MSA was assessed by ProtTest3 with a subset of the substitution models (JTT [Jones *et al.*, 1990], Dayhoff [Dayhoff *et al.*, 1978], Blosum62 [Henikoff and Henikoff, 1992], VT [Muller and Vingron, 2000], and WAG [Whelan and Goldman, 2001]), see per command below. Limiting the number of substitution models because software used in subsequent steps of our analysis pipeline only used these models.

Substitution models were also assessed in the presence of different variables of rate-heterogeneity, including: invariable sites (+I) [Reeves, 1992], variable rate categories (+G/Γ) [Yang, 1993], and a combination of these two factors (+I+G/Γ). Using a maximum likelihood approach ProtTest3 determines the likelihood of each substitution model (± rate-heterogeneity), and compares these scores using the Bayesian Information Criterion (BIC) [Schwarz, 1978]. Using

the function ‘CheckProtTest’ in ‘MUSCLEman.py’ (Appendix 3.2 - subsequently incorporated into bmeTools.py [See 2.5.2.3]), the substitution model (± rate-heterogeneity) for each MSA with the highest overall BIC value was identified as the best-fit model of protein evolution.

3.3.4 Phylogenetic reconstruction by MrBayes

Phylogenetic reconstruction was carried out for all gene families using MrBayes (v3.1.2) [Ronquist and Huelsenbeck, 2003]. The function ‘SetupMrBayes’ in ‘MUSCLEman.py’ (Appendix 3.2 - function in bmeTools package [Section 2.8.3]) was used to automate the generation of nexus formatted alignments and to append the MrBayes command-block, see Figure 2.8 for example. The function ‘SetupMrBayes’ reads the output of ProtTest3 and assigns two parameters within the command-block: i) the substitution model, and ii) rate-heterogeneity (Figure 2.8).

For each gene family, the standard number of four Markov chain Monte Carlo (MCMC) chains were set. As is standard practice, three of the chains acted as “heated” chains to better allow the regular sampling "cold" chain to escape local maxima/peaks [Ronquist *et al.*, 2011]. Each MCMC chain ran for a minimum of 10^6 generations or until MrBayes reported convergence and the average standard deviation of split frequencies reflected convergence < 0.01 [Ronquist *et al.*, 2011]. Chains were sampled every 200 generations with a standard burn-in of 0.25 (25%) to remove the initial generations prior to likelihood stabilization [Ronquist *et al.*, 2011].

3.3.5 Extracting SGOs from multigene family

A number of phylogenies contained gene duplicates. Where appropriate the program ‘nodes_picker.py’ (Appendix 3.2 - function in bmeTools package [Section 2.9.4]) was used to remove (or prune) the SGO of interest alongside its respective orthologous genes (Figure 2.12). Following such extraction the reduced gene family were realigned (Section 3.2.2) and continued through the analysis.

3.3.6 Selective pressure analysis

Selective pressure analyses were performed using codeML from the PAML package (v4.4e) [Yang, 2007]. CodeML examines nested codon-based models of evolution in a Maximum Likelihood framework to determine ω [Yang, 2007] (See Section 1.3.2.2 for more details). We employed branch-site specific models to scan for positive selection unique to a specific foreground lineage [Yang and dos Reis, 2011]. Using the function ‘SetupCodeml’ in ‘MUSCLEman.py’ (Appendix 3.2 - function in bmeTools package [Section 2.9.3]), the codeML input for each homologous group was generated: i) a nucleotide MSAs inferred form the previously selected protein MSAs (Appendix 3.3) (Section 2.9.2), and ii) the labeled phylogenetic trees required for the branch-site specific models for the extant species (human and mouse) and ancestral lineages (primates, murinae, and rodents) of interest. Using the programs ‘GenerateCodemlWorkspace.pl’ and ‘SetupCodemlTaskfarm.pl’ written by Dr. Thomas Walsh, the nucleotide MSAs and labeled phylogenetic trees were assessed for the modelA branch-site specific model.

The codeML results were interpreted by implementing likelihood ratio tests (LRTs) using the program 'CreateCodemlReports.pl' written by Dr. Thomas Walsh [Walsh, 2013]. The LRT test statistic approximates the chi-squared (χ^2) distribution critical value with degrees of freedom equal to the number of additional free parameters in the alternative model. If branch-site specific models passed LRT and positive selection is inferred, the posterior probability of the positively selected site is estimated by 'CreateCodemlReports.pl' using two calculations: NEB or BEB [Yang, 2007]. If both BEB and NEB are predicted, we used the BEB results as they have been reported to be more statistically robust [Yang, 2005]. The candidate positively selected sites were then compared to UniProt protein entries using the program 'swissAlign.py' (Appendix 3.2).

3.3.7 Identifying evidence of recombination breakpoints

Detection of recombination breakpoints was performed on the nucleotide alignments of each putative positively selected gene using RDP3 [Martin *et al.*, 2010]. RDP3 implements a wide range of independent recombination detection methods: RDP [Martin and Rybicki, 2000], BOOTSCAN [Martin *et al.,* 2005], GENECONV [Padidam *et al.,* 1999], MAXCHI [Smith, 1992], CHIMAERA [Posada and Crandall, 2001], SISCAN [Gibbs *et al.,* 2000], and 3SEQ [Boni *et al.,* 2007]. Implementing multiple detection methods in tandem has been recommended to more accurately detect recombination [Posada and Crandall, 2001]. Additionally, these methods are broadly categorised into two distinct detection approaches: phylogeny-based (RDP, Bootscan, and SISCAN) and substitution-based (CHIMAERA, MAXCHI, GENECONV, and 3SEQ). A past comparative study of recombination detection methods reported GENECONV,

CHIMAERA, and MAXCHI as the most powerful substitution-based methods, with CHIMAERA and MAXCHI performing best overall [Posada and Crandall, 2001]. The more recent substitution based detection method, 3SEQ has been reported to be comparable in power to CHIMAERA [Boni *et al.,* 2007]. Comparisons of the phylogeny-based detection methods have found BOOTSCAN and SISCAN to perform well [Martin *et al.,* 2005; Posada and Crandall, 2001].

The program 'recombReader.py' (Appendix 3.2) was then used to parse the RDP3 output for the nucleotide alignment (created in 3.2.6) for each gene with signatures of positive selected. Due to the inherent difficulty of identifying recombination breakpoints, all recombination events were required to be statistically significant for at least one substitution-based and one phylogenetic-based detection method.

3.3.8 Structural analysis of TLR3

Three-dimensional structures of mouse and human toll-like receptor 3 (TLR3) ectodomain were modeled using MODELLER [Eswar *et al.,* 2008]. We obtained 100% sequence identity between the target sequences and template structure for mouse (PDB id: 3CIG) and human (PDB id: 2A0Z). The dynamic flexibility index (*dfi*) [Gerek *et al.,* 2013] was then computed. The *dfi* score quantifies the dynamic properties of individual residues in the protein structure and the stability change caused by mutating residues.

Certain positions in a 3-D structure are more susceptible to perturbation, showing high fluctuation responses and high *dfi* values, whereas other positions with low *dfi* values are stable and the stability of the proteins does not deviate significantly upon perturbation. Therefore, higher *dfi* means greater chance that this mutation does not alter the 3D structure and by assuming that structure and function are tightly linked [Lee *et al.,* 2007], these high *dfi* scores indicate non-function altering mutations.

Using this approach, stability change is estimated by the protein folding free energy (ΔΔG). A ΔΔG value for each amino acid substitution (mutant) is calculated from ΔG_{mutant} - ΔG_{wt} where ΔG_{wt} is the wild type free energy of unfolding. To compute the ΔΔG values for each amino acid substitution, we applied the FoldX method [Guerois *et al.,* 2002; Schymkowitz *et al.,* 2005] that uses empirical potential combining both physical force fields and free parameters fitted with known experimental data. If the value of ΔΔG is greater than 0, the mutation has a destabilizing effect on the protein structure, while ΔΔG<0, the mutation is stabilizing, we used 1kcal/mol as a threshold. First, we computed the stability change of positively selected residues in the mouse structure and then the corresponding orthologous residues in the human structure. Finally, we estimated ΔΔG for all possible amino acid substitutions in the human structure, including disease-associated sites in human. Disease-associated sites were obtained from the Human Gene Mutation Database (HGMD) [Stenson *et al.,* 2003]. We also estimated ΔΔG caused by all possible amino acid substitutions for randomly selected sites.

3.3.9 Fixation of positively selected sites in populations

To determine if the positively selected sites for each positively selected gene were fixed within their respective populations, variation data was downloaded from Ensembl Biomart (Ensembl Variation 72) [Kinsella *et al.*, 2011]. Human variation data was limited to validated single nucleotide polymorphisms (SNPs) to increase accuracy of the assessment; this limitation was not imposed on mouse variation due to the small number of validated mouse SNPs. Using the program 'buildVariation.py' (Appendix 3.2) a total of 559 SNPs were mapped onto the protein sequence of 29 positively selected genes (Human 2, Mouse 27), see Table 3.2. Using the program 'buildFixationData.py' (Appendix 3.2) the SNP-mapped protein sequences were combined with positive selection data (Section 3.2.6) to identify the positively selected sites that were not fixed in the population.

3.3.10 Assessing positively selected genes for evidence of selection within human population data:

To determine if population data corroborated the findings from the species-level comparisons for positive selection in human lineage, variation data was downloaded from the 1000 Genomes Project website, with each individual consisting of two chromosomal samples [Abecasis *et al.*, 2012]. Using the program '1000Reader.py' (Appendix 3.2) the SNP data from the 1000 Genomes Project for each was separated into individuals. Population sequence alignments for each positively selected gene were created using 'genemapper.py' (Appendix 3.2), this program built alignments in two steps: i) it created the genomic sequence of the individual by mapping their respective SNPs onto the reference

genome [Kinsella *et al.,* 2011], and ii) it added each completed individual nucleotide sequence to the overall alignment. Using DnaSP [Librado and Rozas, 2009], we calculated Tajima's D [Tajima 1989] and Fay and Wu's H [Fay and Wu, 2000] for these population alignments. Fay and Wu's H requires an outgroup sequence, to satisfy this requirement the respective chimpanzee genomic sequences (with 1kb of flanking DNA) for each positively selected gene was obtained from Ensembl Biomart (Ensembl Gene 72) [Kinsella *et al.,* 2011]. Chimpanzee sequences were aligned to the human population data using MUSCLE [Edgar, 2004] and chimpanzee-flanking sequence was cleaved using TrimAl [Capella-Gutiérrez, 2009]. To determine the significance of Tajima's D and Fay and Wu's H, 10,000 coalescence simulations were conducted for each gene [Hudson, 2002].

To determine if the regions surrounding the genes identified as positively selected in human from the species-level comparative analysis exhibited evidence of selective sweep, we created 1kb alignments between 100kb upstream and 100kb downstream of each positively selected gene. Population sequence alignments for the regions surrounding the positively selected genes were created using a modification to 'neutralSetup.py' (Appendix 3.2); alignments were created as described above with the reference sequence. DnaSP [Librado and Rozas, 2009] was used to calculate Tajima's D [Tajima, 1989] for each window. Significance was determined for Tajima's D [Tajima, 1989] for each window as described above. All graphs were created using the matplotlib python library [Hunter, 2007].

Table 3.2: Genes tested for positively selected site fixation in their population.

Gene	Species	SNP Sites
CARD6	Human	38
IRF9	Human	9
Stat2	Mouse	58
C6	Mouse	53
Nlrp6	Mouse	53
Il4ra	Mouse	51
Plcg2	Mouse	51
Lrrfip1	Mouse	49
C8b	Mouse	36
Lbp	Mouse	33
Rnf31	Mouse	24
Tcf4	Mouse	20
C1ra	Mouse	16
Grn	Mouse	16
Ltb4r1	Mouse	16

Gene	Species	SNP Sites
Nlrp14	Mouse	16
Lgals3	Mouse	15
Ifit2	Mouse	14
Irf5	Mouse	14
Il1rapl2	Mouse	11
Adipoq	Mouse	9
Cfh	Mouse	8
Cd63	Mouse	6
Card6	Mouse	4
Tlr3	Mouse	4
Atg9a	Mouse	3
Trif	Mouse	3
Snap23	Mouse	2
Ecsit	Mouse	1

The table shows the gene name, species, and the number of available non-synonymous SNPs from Ensembl Biomart (Ensembl Variation 72) [Kinsella *et al.*, 2011] for each of the 29 genes under positive selection. The SNP site column indicates the number of non-synonymous SNP that mapped to protein coding residues.

3.4 Results

3.4.1 Selection analysis reveals species-specific adaptation in mouse and human innate immune genes:

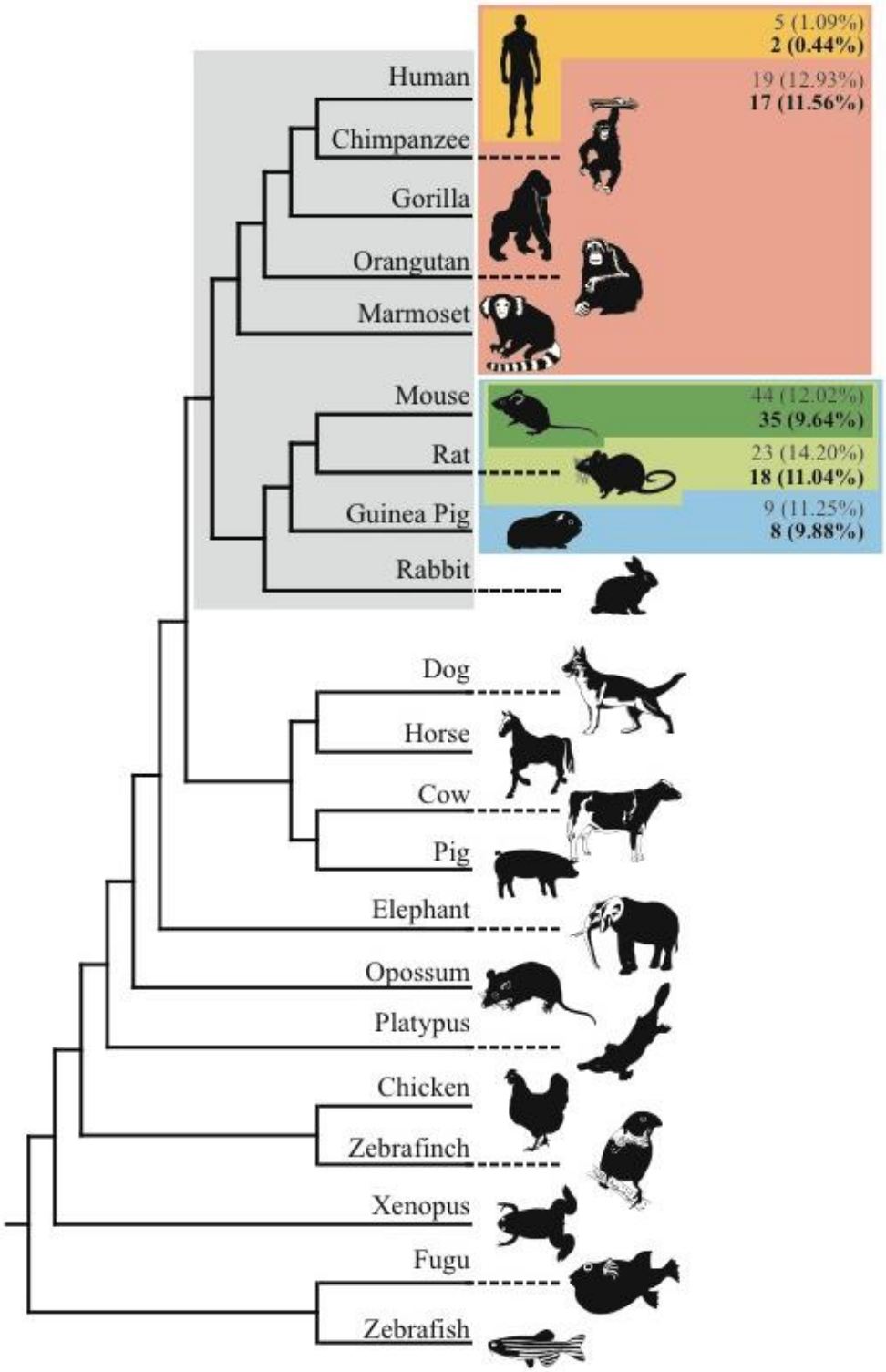
A total of 457 protein coding single gene orthologous families were analyzed. As functional divergence could have emerged prior to the divergence of modern human and mouse all ancestral lineages across the Euarchontoglires clade were tested (Figure 3.1), i.e. the ancestral primate, murinae and rodent branches, as well as modern human and mouse. Following LRT analyses [Yang, 2007] five candidate genes were identified as under positive selection in the human lineage (LAP2, CARD6, C1RL, INPP5D, and IRF9). Analysis of the lineage leading to modern mouse revealed 44 genes under positive selection this is out of a total of 366 gene families that had a mouse ortholog (Appendix 3.4). The branch leading to primates, murinae and rodents showed evidence of positive selection specific to these lineages in 19 (12.93%), 23 (14.20%), and 9 (11.25%) of testable genes respectively (Appendix 3.4). These results emphasize the heterogeneity in selective pressures on innate immunity in different lineages.

3.4.2 Filtering for false positives due to recombination removes potential candidate genes from the positively selected gene set:

To reduce the level of false positive detection for adaptive evolution we applied a number of subsequent filters on the results outlined above. The first was a test for recombination that has previously been associated with potential false positives [Anisimova *et al.*, 2003]. Any putative positively selected gene that had evidence of recombination proximal to positively selected sites was removed from our list of positively selected genes (Figure 3.1). In total, 12 positively selected genes

were identified with evidence of recombination, three human genes (C1RL, LAP2, and INPP5D) and nine mouse genes (Cd22, Csf2rb, Itgam, Ptk2, Sirpa, Tlr8, Traf5, Tyro3, and Zp3r), these were subsequently removed from the list of genes under positive selection (Table 3.3). See Table 3.4 for an updated list of the positively selected genes. From the analysis of the ancestral branches there were a total of three murinae genes (Igf1r, Itgam, and Tyro3), two primate genes (BCAR1 and NLRP9), and one rodent gene (Tyro3) removed (Table 3.5).

Figure 3.1. Phylogeny of species included in this study and summary of lineage-specific positive selection results.



The lineages within the Euarchontoglires clade (denoted by grey box) tested for species-specific selective pressure variation are shown in colored boxes: human

(yellow), ancestral primate lineage – human, chimpanzee, gorilla, orangatun, and marmoset – (red), mouse (dark green), ancestral murinae lineage – mouse, rat – (light green), and ancestral rodent lineage – mouse, rat, and guinea pig – (blue). Dashed lines are provided to increase clarity of the species and are unrelated to the branches of the phylogeny. The initial number and percentage of genes displaying evidence of species-specific positive selection are shown in grey. Totals after the filter for recombination are shown in black.

Table 3.3: Recombination within human and mouse positively selected genes.

Gene Name	Positively Selected Residue Positions	Recombination Event	Internal Recombination Sites	Proximal Recombination Sites
	Genes under positive selection specifically in the human lineage			
C1RL	103	127 - 247		103
INPP5D	72, 223, 227*, 228**, 235, 239, 241, 244, 249, a nd 250	218 - 248	223, 227*, 228**, 235, 239, 241, and 244	249 and 250
LAP2	714 and 1194	1197 - 1261		1194
	Genes under positive selection specifically in the mouse lineage			
Cd22	137, 190, 271, 474, a nd 812	1 - 68		
Csf2rb	169, 271, 288, 473, 536, 569, a nd 576	68 - 161		169
Itgam	23, 45, 820, 831, 844, 1035, 1089, 1092, a nd 1131	1110 - 1206	1131	
Ptk2	390** and 800	366 - 447	390**	
Sirpa	23, 51, 52**, 69, 77, 83, 91, 193, 202, 221, 224, 226, 237, 238, 250, 276, 297, 305, 307, 338**, 344, a nd 490	6 - 74	23, 51, 52**, and 69	77
Tlr8	44, 751, 764*, 778, 802*, 864, a nd 1003	23 - 80	44	
Traf5	133, 308, 309, and 339	223 - 302	308, 309, and 339	
Tyro3	85, 101, 825, and 826	0 - 88 & 700 - 787	85	
Zp3r	3, 7, 39, 110, 174, 179, 206, 212, 219, 244, 31 1, 454, 460, 461, 470, 485, 488, 510, 514, 518**, 524*, 528, 537, 541, a nd 546	396 - 445	454, 460, and 461	

The gene name and positively selected residues are given for each gene exhibiting a recombination event. The location of the recombination event is given alongside the sites within the event (internal) and within close proximity of the event (proximal).

Table 3.4: Positively selected genes identified in this study.

Gene name	Position of positively selected residues
	Genes under positive selection specifically in the human lineage
CARD6	264, 346, 382, 750, 767, 805, 818, 903, 916, 937, 998, 1010, a nd 1031
IRF9	119, 129, and 333
	Genes under positive selection specifically in the mouse lineage
Adipoq	25, 27, 29, and 82
Atg9a	634 and 662*
C1inh	332*, 365, 468, 479
C1ra	468, 520, 574, 631, 633, a nd 634*
C6	220, 233, 319, 353, 378, 408, 419, 430, 554, 655, 681, 703, 792, and 930
C8b	242*, 263, 278, 383*, and 488
Card6	394, 501, and 702
Cd200	129 and 177
Cd63	31, 118, 143, 184*, 194, and 203
Cfh	209, 243, 474, 767, 1005, 1068, 1074, 1104, 1181, and 1227
Ecsit	10, 12, 75, 82, 176, 325, 330**, 348, and 371
Eif2ak2	136, 155, 181, 182*, 344, and 345
F12	45, 65, 166, 243**, and 454
Grn	18, 101, 198, 303, 375*, 382, 411, 549, and 597
Ifit2	191, 402, and 420
Il1rapl2	566, 628, and 666*
Il2rb	4, 13, 31, 55, 174, 202, 347*, 402, 418, 491, 496, a nd 516
Il4ra	47, 67, 308, 330, and 626
Irf5	232, 259, and 262
Lbp	24, 40 and 329
Lgals3	22, 92, 94, and 260
Lrrfip1	328, 449, 468, 480, a nd 571
Ltb4r1	53, 101, and 175
Nlrp14	77, 79*, 186, 212, 219, 254, 257, 263, 272, 281, 284, 291, 294, 315, 319, 333, 358, 393, 415, 424, 453, 465, 530, 549 , 552, 553, 584, 613, 657, 679, 684, 685, 687, 696, 782, 810, 814, 829, 846, 848, 902, 908, 912, 931, 953, 956, 958, 978, 982, 984, and 986

Table 3.4: Positively selected genes identified in this study.

Gene name	Position of positively selected residues
	Genes under positive selection specifically in the mouse lineage
Nlrp6	22, 25, 72, 77, 80, 81, 85, 96, 101, 113, 114, 190, 192, 251, 260**, 329, 344, 479, 488, 515, 553, 571, 628, 657, 727, 737, 739, 744, 771, 775*, 776, 793, 807, 865, 877, and 880
Oas2	55, 56, 139, 171, 199, 211, 221, 298, 481, 549, and 711
Plcg2	461 and 594*
Ptpn2	166, 206, 319, 321, and 329
Rnf31	203, 431, and 1025
Sirt1	107, 537, 698, and 701
Snap23	109, 133, and 197
Stat2	21, 130, 149, 157, 195, 205, 218, 354, 623*, 869, 871, 874, 876, and 877
Tcf4	139
Tlr3	266, 297, and 603
Trif	18, 327, 338, 388, 482, 556, and 711

The table shows the proteins identified under lineage-specific positive selection that showed no evidence of recombination proximal to positively selected residues. The proteins are sorted into the respective lineage in which the positive selection was observed. Each protein is reported alongside the positions that were identified by the BEB empirical Bayes model (see Section 1.3.2.2 for details). See table of abbreviations for the name of each protein shown above.

Table 3.5: Recombination within the ancestral lineages.

Gene Name	Positively Selected Residue Positions	Recombination Event	Internal Recombination Sites	Proximal Recombination Sites
	Genes under positive selection specifically in the primate lineage			
BCAR1	475** and 602	84 - 916	475** and 602	
IFNGR2	99, 126, 170, 195, 197, and 253	2 - 45		
NLRP9	37, 362, and 502	94 - 106, 466 - 499, & 558 - 728		502
TLR8	142, 188, 193*, 212, 213, 255, 309, 312, 335, 349, 386, 387, 432, 562, 580, 605, 634, 641, 690, 722, 793, 835, 850, 923, 975, 1013, 1087, 1089, and 1103	41 - 98		
TRIM5	433, 520, 523, 538, 568, 584, 654, 767, 810, 842, 848, and 859	242 - 276		
	Genes under positive selection specifically in the rodent lineage			
Mst1r	736 and 964*	267 - 376 & 814 - 880		
Tyro3	91*	1 - 89		91*
	Genes under positive selection specifically in the murinea lineage			
Ccdc88a	195, 221, 443, 601, 613, 635, 638, 675, 995, 1021, 1031, 1037, 1219, 1230, 1235, 1607, 1704, 1728, 1730, 1733, 1817, and 1827	89 - 113		
Mst1r	97, 268, 291, 553, 769, and 1041	267 - 376 & 814 - 880	268 and 291	
Traf6	16*	275 - 313		
Igf1r	155 and 257**	39 - 227	155	
Itgam	98, 132, 214, 226, 468, and 784	1 - 1111	98, 132, 214, 226, 468, and 784	
Tyro3	101, 105, 127, 202, 223, 240, 406*, and 445	1 - 89		101 and 105

The gene name and positively selected residues are given for each gene exhibiting a recombination event. The location of the recombination event is given alongside the sites within the event (internal) and within close proximity of the event (proximal).

3.4.3 A subset of mouse innate immune pathways are enriched for adaptive evolution:

Assessment of the two candidate genes in the human lineage (CARD6 and IRF9) revealed no evidence of protein-protein interaction among these proteins and no evidence for enrichment in particular innate immune pathways. Conversely, the 35 candidates in mouse exhibited a number of proteins involved in direct interaction with one another (Figure 3.2) and in the same pathways: 5 genes in the complement cascade (C1ra, C1inh, C6, C8b, and Cfh - Figure 3.2a), 4 genes in the TLR signalling pathway (Irf5, Lbp, Tlr3, and Trif - Figure 3.2b), 3 genes in the JAK-STAT pathway (Stat2, Il2rb, and Il4ra), and a single genes in the MAPK signalling pathway (Ecsit). Of particular interest was the interaction between Trif and Tlr3. Trif exhibits positive selection within the Tlr3 interaction interface (Figure 3.2) [Oshiumi *et al.*, 2003].

3.4.4 The Ancestral nodes have unique subsets of genes under positive selection:

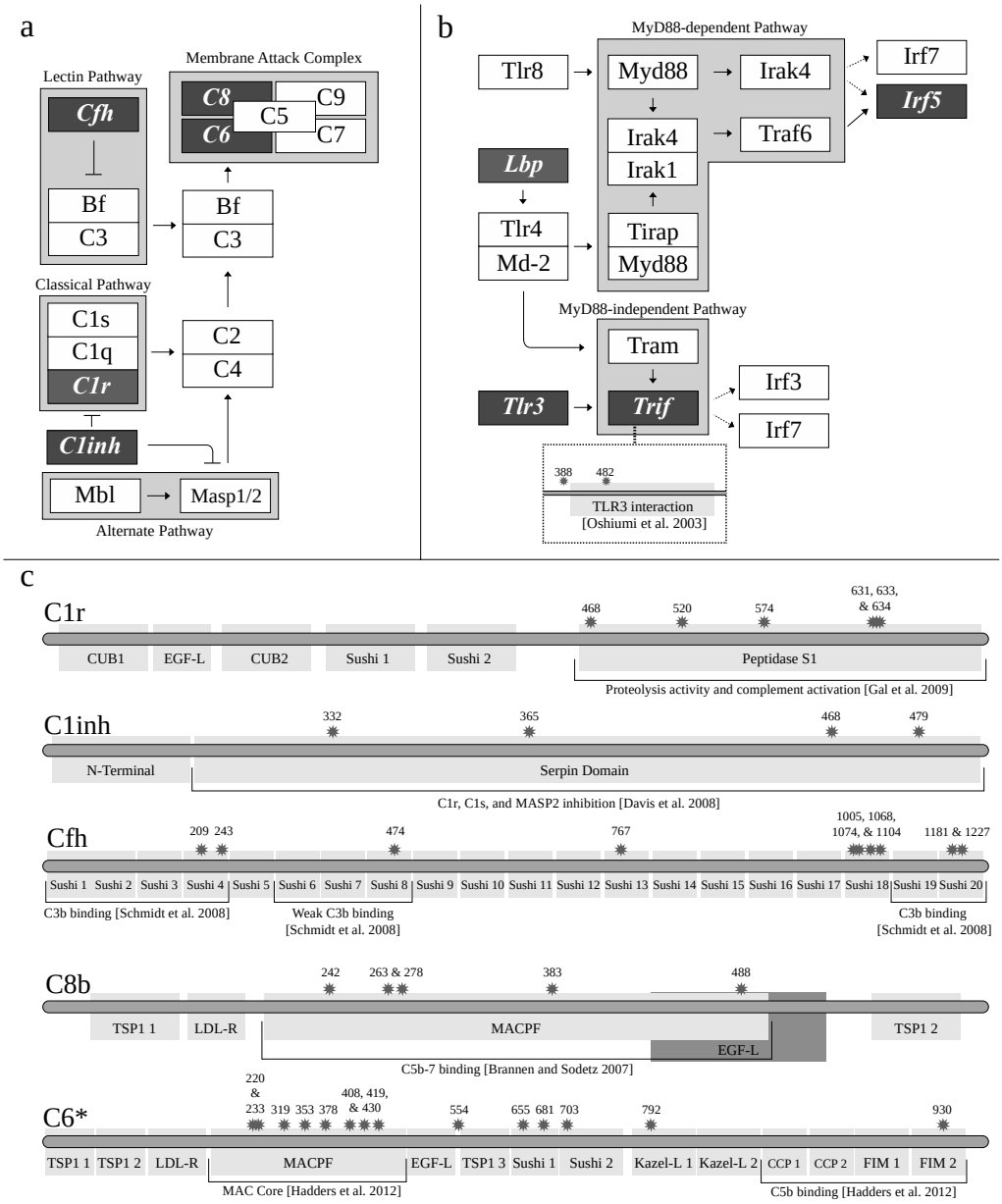
We identified 18 genes with evidence of positive selection in the ancestral primate lineage alone and nowhere else on the tree. In terms of the functional classification of these genes they fell broadly into five categories: Nod-Like Receptors (NLRP1, NLRP5, NLRP8, and NLRP9), TRIM receptors (TRIM5 and TRIM25), Interferon gamma receptors (IFNGR1 and IFNGR2), TLR (TLR8), and one breast cancer related gene (BCAR1) (Appendix 3.4). Within this set of genes there is one known protein-protein interaction between IFNGR1 and IFNGR2 associated with the JAK-STAT pathway [Kanehisa and Goto, 2000]. Analyses of the ancestral rodent and murinae branches identified 7 and 14 genes

respectively as under positive selection (Appendix 3.4). There was limited information on function and associated pathways for these genes and there were no reports of direct protein-protein interactions in these subsets of the data. The C1inh protein from the complement cascade also featured as under positive selection in the ancestral murinae branch, and there was evidence for a murinae unique positive selection event in the TRAF6 protein, a gene known to interact with the TLR signaling pathway [Kanehisa and Goto, 2000].

3.4.5 Positively selected residues map to essential functional domains:

To determine potential functional effects of positive selection we compared the positively selected residues identified in this study to functional data available on SwissProt [The UniProt Consortium, 2012]. Assessment of the 35 candidate genes under positive selection in the mouse lineage alone identified numerous examples of positively selected residues within known functional domains (Appendix 3.4). Where data permitted additional functional assessment was performed on the genes of the complement cascade (Figure 3.2c).

Figure 3.2: Innate immune pathways containing positively selected genes.



The positively selected genes in (a) the complement system and (b) the TLR signaling pathway are shown as darkened rectangles. Signaling cascades are depicted as arrows and inhibitors are depicted as blunt-ended lines. Defined pathways and complexes are shown in grey boxes with the given name. (c) Positively selected residues of the complement system alongside information on domain structure. Information on the function of these domains is also given. See table of abbreviations for the name of each protein shown above.

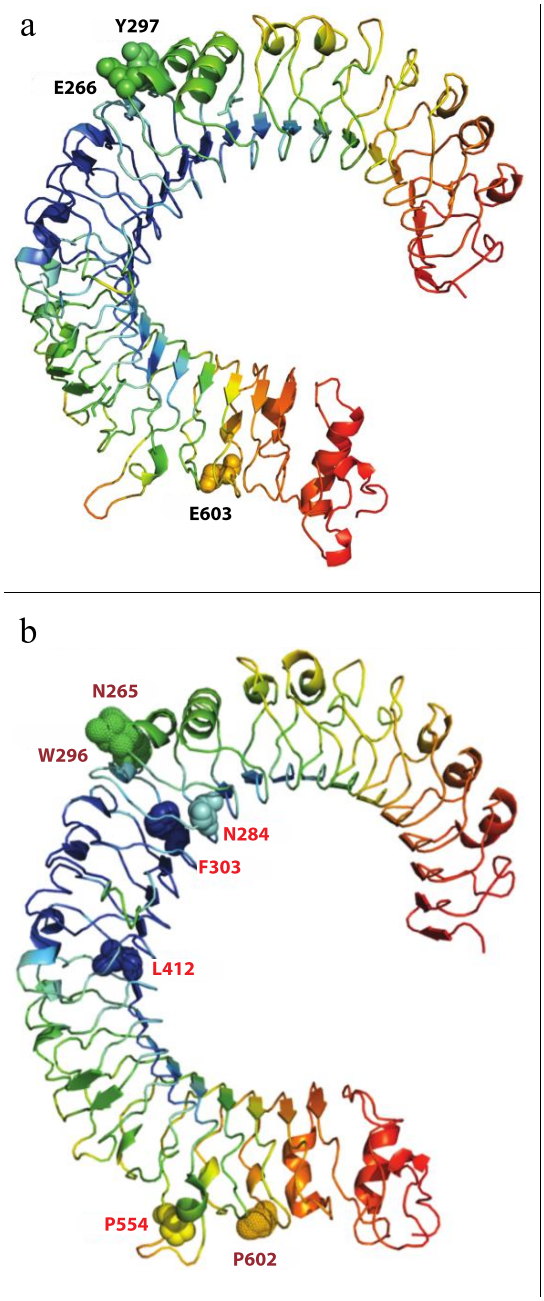
3.4.6 Positively selected residues in mouse TLR3 have predicted effects on structural stability:

The dynamic flexibility index (*dfi*) measures the contribution of each residue to the overall structural dynamics and stability of the protein [Gerek *et al.*, 2013]. The *dfi* approach measures the spatial fluctuation of each residue in response to perturbing residues along the peptide. *Dfi* values indicate the resilience of each residue to perturbations; a low *dfi* indicates a residue essential for dynamic stability as they absorb the transfer of perturbation (i.e. structurally inflexible) whereas high *dfi* implies the residue is prone to perturbation (i.e. structurally flexible) [Gerek *et al.*, 2013]. *Dfi* values are reported to significantly correlate with known neutral variants (high *dfi* values) and residues strongly linked with genetic disease (low *dfi* values) [Gerek *et al.*, 2013]. This correlation of *dfi* to biological function was also reported to have a greater ability to discern functionally critical from non-critical sites as compared to the solvent accessible surface area (ASA) metric frequently used to assess functional significance [Gerek *et al.*, 2013].

LR3 was chosen for this analysis as it is well characterized and displayed lineage specific positive selection in mouse but not human. The human TLR3 disease-associated residues displayed low dfi values [Stenson *et al.*, 2003] (Figure 3.3) whereas the positively selected residues from mouse TLR3 (i.e. E266, Y297, and E604) and their counterparts in human (i.e. N265, W296, and P602) had moderate to high dfi values indicating structural flexibility (Figure 3.3). These findings are in keeping with the red queen hypothesis and the potential role of these sites in pathogen recognition or binding at the surface of the TLR3

ectodomain. We then calculated the change in protein stability for all possible amino acid substitutions at each position in the ectodomain of human and mouse TLR3 (Figure 3.3). We found no evidence of a destabilizing effect for the positively selected residues in the ectodomain of mouse TLR3 (E266, Y297, and E603) (Figure 3.3), indicating that these positions in mouse are tolerant to mutation and are structurally flexible. Mutating the homologous residues in human (N265, W296, and P602) we found two sites (W296 and P602) that exhibited an effect on folding stability. More specifically, of the 19 possible mutations of W296 and P602, destabilizing effects were observed in 13 and 18 substitutions, respectively (Figure 3.3). Substitutions of W296 and P602 exhibit dfi profiles comparable to disease-associated sites, leading us to propose that in human these positions may also be fundamental for protein stability and structurally inflexible. Testing the effect of mutating randomly chosen residues we found no obvious patterns in protein destabilizing effects (Figure 3.3). Taken together, these data suggest that two of the positively selected residues in mouse TLR3 and their homologous positions in human TLR3 contribute differently to overall protein stability in these two species.

Figure 3.3: Dynamic flexibility index of human TLR3 ectodomain.



c

Residues	Substitutions ΔΔG>1 kcal/mol
Mouse positively selected	
E266	0
Y297	7
E603	0
Human homologs	
N265	0
W296	13
P602	18
Disease-associated sites	
N284	19
F303	19
L412	18
P554	16
Randomly selected sites	
A35	11
R65	1
Q146	3
I250	16
L360	17
T400	3
W429	11
S500	13
R544	4

Ribbon diagrams of the crystal structure of the TLR3 ectodomain of (a) human (PDB id: 2A0Z) and (b) mouse (PDB id: 3CIG). (a) and (b) are colored with a spectrum of red-yellow-green-cyan-blue representing the dynamic flexibility index (*dfi*), where red indicates the highest *dfi* values down to blue which indicates the lowest values. **(c)** The stability change for all possible substitutions was computed for: the positively selected sites in mouse (E266, Y297, and

E603), their human homologs (N265, W296 and P602), known human disease-associated sites (N284, F303, L412, and P554) [Stenson *et al.*, 2003], and randomly selected sites. Except for the randomly selected sites, sites have been indicated on the respective ribbon diagrams in the following colors: mouse positively selected sites in black, human homologs in brown, and human disease sites in red.

3.4.7 The majority of positively selected residues are fixed within human and mouse populations:

Positive selection rapidly drives advantageous alleles to fixation within a population [Haldane, 1927]. Depending on the age of the event, effective population size (N_e) and the strength of selection, positively selected sites will either be fully fixed within a population or will have some degree of variability [Sabeti *et al.*, 2006]. We sought to determine if the positively selected residues identified in human and mouse are fixed in their respective populations or whether they are variable (Table 3.6). We gathered all available SNP data for all positively selected genes, i.e. two human candidates and 27 mouse candidates. The majority of positively selected sites in both human (15/16 or 94%) and mouse (207/214 or ~97%) were entirely fixed in their lineage. The exceptions were the CARD6 gene in human and six mouse genes (C6, C8b, Ecsit, Il4ra, Nlrp14, and Stat2) (Table 3.6). Of the total of 8 positively selected residues across all 7 genes showing variability at the population level, four SNPs resulted in the ancestral residue present at the homologous position in other species, they were as follows: human CARD6 (G264E) and mouse Ecsit (S75L); Nlrp6 (R744K); and Stat2 (L874M). In addition, there were two substitutions at positively selected residues in mouse genes that resulted in amino acids with similar physicochemical properties as the homologous position in another species: C6 (L554) and Il4ra (G626) (Table 3.6).

The positively selected residues in human CARD6 and IRF9 genes were compared to the recently released Neanderthal genomes [Green *et al.*, 2010]. The

same positively selected residues that are fixed in the modern human lineage were also found in Neanderthal.

3.4.8 Population level data shows no ongoing selective sweep in modern humans:

Within a population, positive selection not only leads to the fixation of an advantageous allele, but a notable reduction in variation in the surrounding region [Sabeti *et al.*, 2006]. The regions identified in the species level analyses as positively selected were tested to determine if they are evolving neutrally in modern human populations, this was done using Tajima's D statistic [Tajima, 1989]. Tajima's D is a scaled measurement of the difference between the number of segregating sites and average nucleotide diversity, the value of D is expected to be close to zero if sequences are evolving neutrally [Tajima, 1989]. Therefore, significant deviations from zero indicate sequences that are evolving non-neutrally. To determine if our two candidate genes from human (CARD6 and IRF9) were evolving neutrally, their Tajima's D was calculated (Figure 3.4a). Both had a negative Tajima's D, indicating non-neutral evolution. These values were found to be statistically significant.

We wished to determine if this "non-neutral" signal from the Tajima's D statistic for CARD6 and IRF9 in human was due to positive selective pressure or purifying selective pressure. Fay and Wu's H test was applied, as it accounts for derived alleles using an outgroup sequence [Fay and Wu, 2000]. Derived alleles are non-ancestral recent mutations that are expected to be at lower frequencies.

Table 3.6: Fixation of human and mouse positively selected genes.

Genes	PS Residues	Coding SNPs	Residue(s) not fixed	Details
Genes under positive selection in the human lineage				
CARD6	13	38	G264E	Ancestral residue
IRF9	3	9	None	
Genes under positive selection in the mouse lineage				
Adipoq	4	9	None	
Atg9a	1	3	None	
C1ra	6	16	None	
C6	14	53	R554L	Similar physicochemical properties
C8b	5	36	M263I	
Card6	3	4	None	
Cd63	6	6	None	
Cfh	10	8	None	
Ecsit	9	1	S75L	Ancestral residue
Grn	9	16	None	
Ifit2	3	14	None	
Il1rapl2	3	11	None	
Il4ra	5	51	F47S & D626G	D626G: Similar physicochemical properties
Irf5	3	14	None	
Lbp	2	33	None	
Lgals3	4	15	None	
Lrrfip1	5	49	None	
Ltb4r1	3	16	None	
Nlrp14	51	16	A613S	
Nlrp6	36	53	None	Ancestral residue
Plcg2	1	51	None	
Rnf31	3	24	None	
Snap23	3	2	None	
Stat2	14	58	L874M	Ancestral residue
Tcf4	1	20	None	
Tlr3	3	4	None	
Trif	7	3	None	

The gene name, number of positively selected residues, number of protein coding SNPs, and unfixed residues for each positively selected gene with variation data. If a coding SNP produces a positively selected residue that is not fixed within the

population, the respective polymorphism and position are given. If the polymorphism is to an ancestral residue present at the homologous position in other species, it is designated as: Ancestral residue. If the polymorphism is to a residue that shares similar physicochemical properties to the residue as the homologous position in another species, it is designated as: Similar physicochemical properties.

However, derived alleles within close proximity of an advantageous allele may spread by hitchhiking and thereby become high frequency [Sabeti *et al.,* 2006]. Assessment of CARD6 and IRF9 by Fay and Wu's H test resulted in negative values that failed to reach significance, indicating that these regions are evolving neutrally (Figure 3.4a).

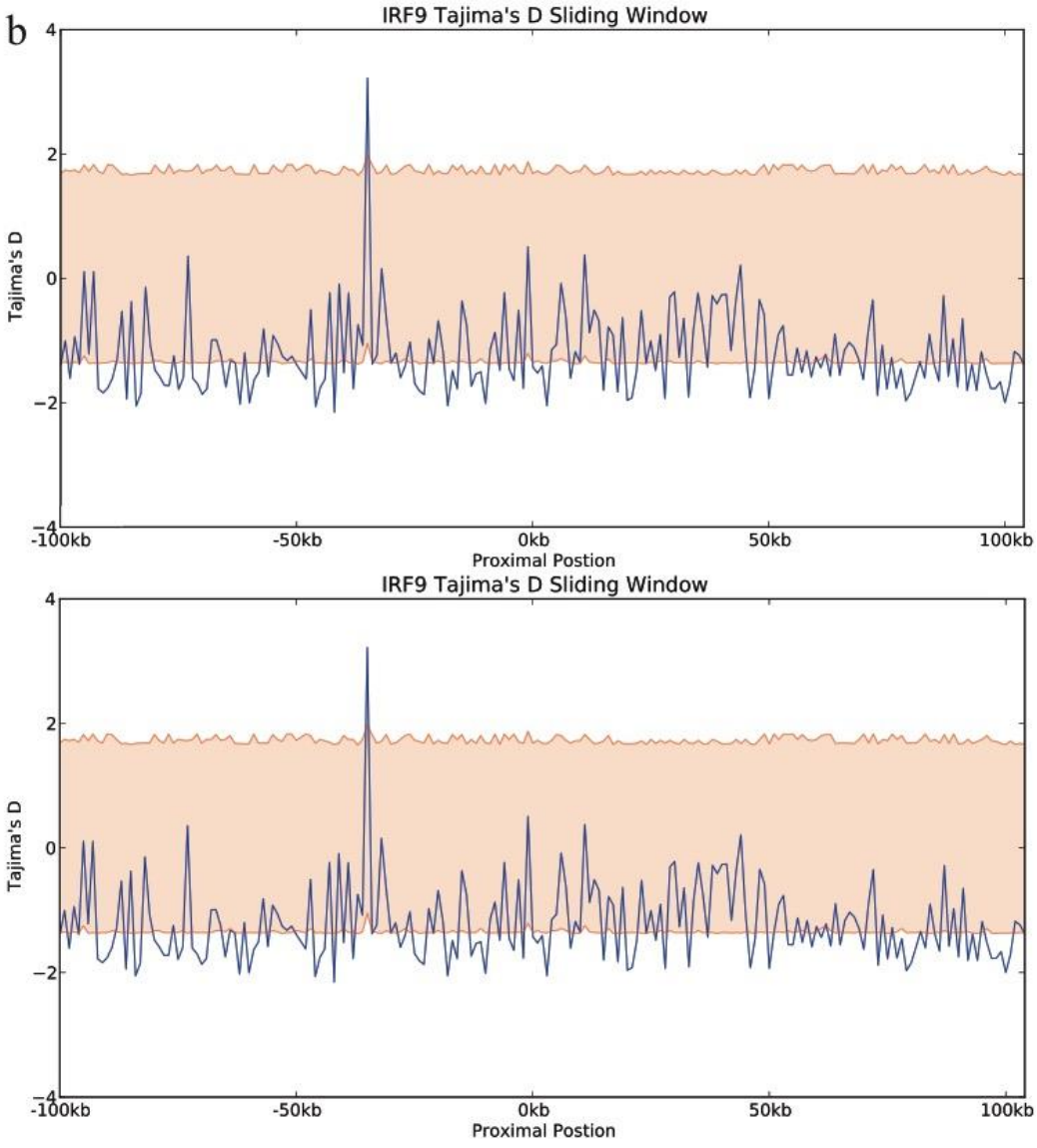
The rapid fixation of an advantageous allele within a population results in a notable and skewed decrease in variation at linked neutral sites, termed a selective sweep [Sabeti *et al.,* 2006]. To determine if our candidates from the species-level analyses exhibited evidence of reduced variability, a 1kb sliding window analysis of allele frequency was carried out incorporating 100kb of upstream and downstream sequence for each candidate gene. Departure from neutrality was measured in each window [Tajima, 1989], graphical representations of the results can be found in Figure 3.4b. The analysis of IRF9 and CARD6 identified comparable levels of variation for all windows assessed. The results for the human candidate genes IRF9 and CARD6, are consistent with a potential positive selective pressure/sweep to fixation in the ancestral human lineage prior to the divergence of Neanderthal, and a relaxation of selective pressure in the current human population.

Figure 3.4: Neutrality tests for positively selected genes in the human lineage.

a

Gene	Tajima's D	Significant	Fay and Wu's H[1]	Significant
IRF9	-2.079	Yes	-0.0069	No
CARD6	-1.7586	Yes	-1.1437	No

b



(a) Results of calculating Tajima's D and Fay and Wu's H and their respective 95% confidence intervals from coalescent simulations for IRF9 and CARD6. (b) Sliding window analysis of Tajima's D of the positively selected genes identified in human. The analysis was conducted using a window size of 1kb within 100kb

upstream and downstream of each gene. The 95% confidence interval is shown as red highlighted region.

3.5 Discussion

In agreement with previous large-scale surveys of positive selection [Kosiol *et al.*, 2008], we find a high frequency of positive selection in innate immune genes in the mouse lineage as compared to human. These high levels of species-specific positive selection in human and in mouse were not observed in ancestral mammal lineages, suggesting that these sites are responding to more recent selective pressures.

The candidate genes of both mouse and human exhibited some evidence of recombination breakpoints. It has been reported that this may produce molecular signatures indistinguishable from those of positive selection [Anisimova *et al.*, 2003] and genes with signatures for recombination were removed.

The positively selected genes in mouse are components of well-known innate immune pathways, and are involved upstream and downstream in these pathways. In pathways enriched for positively selected members (e.g. TLR signaling pathways and complement system), more than half the components are involved directly or indirectly with initiation (Cfh, C1r, C1inh, Tlr3, and Lbp), indicating that upstream positions may be subjected to stronger positive selective pressure. Recent studies have reported a relaxation in selective constraint as you progress downstream through a pathway [Ramsay *et al.*, 2009] whereas other reports detail patterns similar to those presented here [Alvarez-Ponce *et al.*, 2009]. The enrichment for positive selection that we observe in proteins functioning at the start of the pathways is most likely because many of these proteins interact directly with pathogens (C6, C8, and Lbp); bind to pathogens

for immunological defense (Cfh); or are pathogen recognition receptors (Tlr3) and are most likely under increased positive selective pressure from the pathogen [Van Valen, 1973]. The absence of human-specific positively selected genes within the TLR signaling pathways and complement system may be partially explained by a previous study of the evolution of TLRs in primates [Wlasiuk and Nachman, 2010] where they found primate TLRs exhibit an episodic pattern of evolution. In agreement with this pattern, we identified positive selection in TLR8 in the ancestral primate lineage.

As positively selected sites have been documented to result in protein functional shifts [e.g. Loughran *et al.,* 2012; Saywer *et al.,* 2005], our results may implicate positively selected residues responsible for the emergence of divergent protein functions between human and mouse. We identified positively selected genes in the alternative, classical, and lectin pathways of the complement system in mouse indicating possible functional shift in these pathways in the mouse lineage alone. The complement system is reported to neutralize herpes simplex virus (HSV) in rat, mouse, and human [Wakimoto *et al.,* 2002]. However, complement activation proceeds uniquely for each species: via the lectin pathway in mouse, via the alternative and lectin pathways in rat, and in human via the classical pathway [Wakimoto *et al.,* 2002]. The exclusive use of the lectin pathway in mouse is of interest due to the presence of positively selected genes in the alternative (C1inh), classical (C1r and C1inh), and lectin (Cfh) pathways that offer potential molecular markers of *in vitro* study of the observed species-specific functional shift.

The complement system is also reported to display functional discordance in response to Acanthamoeba infection. Both human and mouse are reported to initiate the complement system by the alternative pathway leading to Acanthamoeba binding with C9 of the membrane attack complex (MAC) [Pumidonming *et al.,* 2011]. By contrast, the MAC of mouse is unable to lyse Acanthamoeba [Pumidonming *et al.,* 2011]. Here we present evidence of positively selected residues in the C6 and C8b proteins of the complement cascade unique to the modern mouse lineage. Both the C6 and C8b proteins are essential to MAC formation and cell lysis [Aleshin *et al.,* 2012], therefore these residues provide the community with a molecular target for the observed phenotypic discordance between human and mouse in their MAC activity.

In comparison to innate immune genes with phenotypic discordance and a known molecular cause, our results were found to be moderately successful. The restriction factor TRIM5α is reported to confer a species-specific resistance to HIV-1 in rhesus macaque but not humans [Stremlau *et al.,* 2004]. Selective pressure analysis performed on TRIM5α identified an 11- to 13-amino acid segment of the SPRY domain responsible for species-specific retroviral restriction [Sawyer *et al.,* 2005]. In contrast, our analysis was able to identify TRIM5α as under positive selection within the ancestral primate lineage (Appendix 3.4), but was unable to identify the reported causative region. It should be noted that our analysis did not include the rhesus macaque and this could partially explain the variation in the results. Another example of a known innate immune discordance is TLR8, which is reported as being able to confer NF-κB activation in response to multiple RNA ligands in human but not mouse

[Jurk *et al.,* 2002; Forsbach *et al.,* 2008; Liu *et al.,* 2010]. The cause of this species-specific activation was determined to be a 5-amino acid motif – RQSYA – that was not present within mouse TLR9 [Liu *et al.,* 2010]. Our analysis was able to identify TLR8 as being under positive selection in the ancestral primate lineage but the causative motif was not identified as being under positive selection.

The preferred approach to test functional effect of positive selection is by *in vitro* analysis. The same approach (albeit with a less sophisticated branch-site model) has been shown to result in functional effects of positive selection in the innate immune related protein myeloperoxidase [Loughran *et al.,* 2012]. Providing this human mouse discordance data will lead to further *in vitro* studies.

In keeping with our computational approach we performed *in silico* structural modeling on the TLR3 protein. The difference in stability effects between human and mouse residues of the TLR3 protein suggests a major structural and functional discordance between human and mouse in their ability to detect double stranded RNA in viral infections. Indeed these findings for TLR3 are of particular interest given reports of the restricted anti-viral role of human TLR3 in comparison to mouse Tlr3 where LPS up-regulation of Tlr3 is seen in mouse but not in human macrophages [Ariffin and Sweet, 2013]. This is an important finding given the difficulty in modeling the human innate immune system [Mestas and Hughes, 2004]. This application of *in silico* modeling highlights its ability to predict molecular level signatures of species phenotypic discordance that warrant future *in vitro* study.

While the lineage leading to modern human and indeed Neanderthal display signatures of positive selection in two genes (CARD6 and IRF9), there is no evidence for an ongoing selective sweep in modern human in these genes [Tajima, 1989]. It should be mentioned that this approach is sensitive to the age of the adaptive event, because signatures of selective sweep are eventually lost. The outer limit of these analyses in human is approximately 250,000 years [Sabeti *et al.,* 2006]. The conservation of these positively selected residues between modern human and Neanderthal may be due to shared ancestry making these adaptive events at minimum ~400,000 to 600,000 years old [Scally and Durbin, 2012].

The number of positively selected sites where the residue was variable within human or mouse was rare, in total 94% (in human) to 97% (in mouse) of positively selected sites were completely fixed in the modern human or mouse populations. There were a total of 8 sites where the positively selected residue was not fixed. All but one of these were from mouse population data and often encoded for ancestral residues observed in the corresponding position in other species (4 of 7) or that were physicochemically similar residues to those in other species in the alignment (2 of 7). It is important to note that the mouse population data is based on only 21 inbred laboratory mice strains, and likely does not represent true population structure due to their artificial selection histories [Keane *et al.,* 2011].

There was a single unfixed positively selected site in our cohort of positively selected genes in human: position (G264E) of CARD6. This replacement SNP (rs61757657) is documented in only 3% of Africans and 1% of Americans (2% of Puerto Ricans) [Kinsella *et al.,* 2011; 1000 Genomes Project Consortium, 2012], suggesting that the putative positively selected residue is not completely fixed in these populations. Subsequent inspection of the multiple species alignment identified the codon of the unfixed E264 (GAA) to be homologous to the other great apes (chimpanzee, gorilla, and orangutan) within the alignment, suggesting multiple functional alternatives may be tolerated at this position. Disease association data would also prove valuable to determine if this polymorphism is slightly deleterious or neutral.

The combined use of new data such as the recently completed 1000 human genomes and Neanderthal genomes with phylogenetic analyses of selective pressure, has potential for advancing our understanding of the molecular underpinnings of species-specific response to disease. Distinct species-specific selective forces are acting on components of the innate immune system, they are detectable at the molecular level and they align with known phenotypic discordance, thereby providing a predictive tool for the identification of currently unknown discordance cases of the immune system and beyond.

Chapter 4: A non-phylogenetic approach to determine gene organization and domain sharing within vertebrate protein coding regions

4.1 Chapter Aim

Domain rearrangements in have been directly implicated in the creation of novel proteins, including the establishment of species-specific proteins. While characterizing and identifying such proteins would be highly beneficial to understanding functional discordance, traditional phylogenetic approaches are unable to fully characterize the events due to non-vertical nature of domain rearrangement. The primary aim of chapter four was to explore the ability of networks to: i) identify species-specific proteins create by domain rearrangements and ii) understand the properties of domain rearrangements in multi-domain proteins. These goals were achieved by constructing a network of Pfam-A domains to explore the ability of Pfam-A domains to co-occur within a gene. The co-occurrence network was then used to identify multi-domain exhibiting species-specific domain combinations. The network was also used to identify the governing principals of multi-domain proteins to determine if these principals differed in proteins exhibiting species-specific fusion events.

4.2 Introduction

Phylogenetic trees have proven to be an invaluable tool for the field of evolutionary biology. Inferred from evidence of vertical descent, phylogenetic trees are commonly used to explore the evolutionary relationships among species, genes, and populations. While advances in phylogenetic reconstruction have led to complex models of evolution, this framework alone cannot fully explore the process of modular rearrangement in multi-domain (i.e. modular) proteins due to their non-vertical nature [Bapteste *et al.,* 2013].

An established view of protein modularity is the existence of functional modules – or domains – that are analogous to independently folding elements [Moore *et al.,* 2008; Coulson and Moult, 2002]. Support for this claim has been documented from reports of mutations that affect the function of a particular domain but not the other domains of a protein [Tjoelker *et al.,* 2000]. Domains are also reported to belong to domain families, which are collections of domains that share similar structural profiles and/or evolutionary histories [Andreeva *et al.,* 2008; Finn *et al.,* 2014]. Specific protein families, such as fibronectin III and kinases are widely used in the genome and found within a multitude of proteins [Little *et al.,* 1994; Manning *et al.,* 2002]. In addition, research indicates that several of these domain families are common to most species, indicating they may be ancient elements [Apic *et al.,* 2001]. Considering the age and frequency of these domains families, it may not be surprising that research has shown that domain rearrangement have resulted in the creation of novel proteins [Bashton and Chothia, 2007]. As domain rearrangement may result in the creation of novel proteins, the identification of species-specific domain rearrangements is of

particular interest. For example, a number of species-specific composite genes (*i.e.* fusion of two or more genes) have been found to result in the establishment of unique species-specific functional properties [Thomson *et al.,* 2000; Rogers *et al.,* 2010; Molero *et al.,* 2013]. Therefore, characterizing the governing characteristics of modular proteins – the domains combinations and rearrangements permitted – in addition to the identification of species-specific domain rearrangement may lead to a better understanding of protein evolution and the establishment of new function.

In this chapter non-vertical evolution of modular proteins is explored by employing graph theory to accurately characterize the properties of modularity. In comparison to phylogenetic trees, graphs or "networks" represent biological data as unrestricted pairwise connections and allow genetic material to have multiple sources [Halary *et al.,* 2010]. Therefore, a network considers the independent domains of modular proteins as individual genetic sources, and enables the exploration of modularity in proteins [Moore *et al.,* 2008]. Previous applications of graph theory have found that biological graphs share many features with technological and social networks [Newman, 2003; Barabasi and Oltvai, 2004], highlighting the potential for universal laws of networks [Barabasi and Oltvai, 2004]. Beyond simply enabling the visualization of modular proteins, applying the concepts and techniques of network theory to biological data facilitates the exploration, discovery, and description of previously unknown properties and mechanisms of protein evolution.

To date, a considerable amount of research has been conducted on the modular rearrangements of proteins in a variety of species [Wuchty and Almaas, 2005; Moore *et al.,* 2008; Kersting *et al.,* 2012]. From these studies we have learned that rearrangement event such as domain fusion and fission, as well as large-scale chromosomal events such as intragenic duplications, intergenic repeats, and exon relocation all play an important role in the evolution of modular proteins [Moore *et al.,* 2008]. In addition, single-species networks have been successfully applied to understand the evolutionary impact of domain modularity [Wuchty, 2001; Wuchty and Almaas, 2005]; nonetheless, much work remains to fully understand the impact of domain modularity and species-specific rearrangement in vertebrates.

The data used in this chapter consisted of the CDSs from 30 vertebrate genomes, and the corresponding constructed Pfam domain co-occurrence network (Appendix 4.1). Here our goal is to determine the global properties of the domains combinations and rearrangements in the evolution of modular proteins in vertebrates including the discovery of species-specific domain rearrangement events that may result in potential functional discordance.

4.3 Materials and Methods

4.3.1 Bipartite graph and co-occurrence unipartite-projection of Pfam-A data:

The profile hidden markov models (profile HMMs) of 14,831 Pfam-A domain families were downloaded from the Pfam website (v27.0) [Finn *et al.,* 2014]. A total of 30 high-coverage genomes were downloaded from Ensembl Gene 73 on

Ensembl BioMart [Flicek *et al.*, 2014] (Table 4.1). These annotated genomes together with the HMMscan function within standalone HMMER (v3.1b1) [Eddy, 1998] provided the sequence information required to identify homology of domains, i.e. Pfam-A motifs. Sequences homologous to Pfam-A domains were only reported if they passed a conservative E-value threshold of 1.0E-20. HMMscan homology connections were then filtered to account for the possibility of erroneously assigning Pfam-A domains to a gene due to the presence of either a composite domain or a domain family member.

Composite Pfam-A domains are comprised of smaller Pfam-A domains (i.e. component domains) and may generate false domain combination preferences if component domains are not removed from the database. Domains were classified as component domains if they found within a larger Pfam-A domain. To account for the imperfect nature of determining exact alignment positions [Eddy, 2010], a component was allowed to have 5% of its sequence to be partially unaligned (either 5' or 3') to the composite domain. Identified components were removed from subsequent analysis. If multiple tiers of composites/components were identified only the largest overall composite was included in subsequent analyses, if multiple composites were found to share the largest length, the composite with the lowest e-value was selected.

Domain families are characterized by possessing multiple Pfam-A domains with similar sequence motifs and may generate false domain combination preferences due to multiple Pfam-A domains aligning to the same position within a protein. To account for this potential error, if 80% of a Pfam-A domain was found to

overlap with another domain on a protein, only the Pfam-A domain with the lowest e-value was included in subsequent analyses. If multiple tiers of overlaps were identified only the Pfam-A domain with the lowest e-value was included in subsequent analyses. It should be noted that if there was evidence of domain families within a composite Pfam-A domain, the composite method of domain selection was preferred.

The filtered homology connections (filtered using the program 'Pfam_Checker.py' [Appendix 4.2]) were then used to construct a bipartite graph consisting of edges between Pfam-A domains and protein sequences (i.e. Pfam-A homology graph) (Appendix 4.1) using the program 'Pfam_Checker.py' (Appendix 4.2). The bipartite graph was subsequently separated into connected components to create Pfam-A homology sub-graph. A Pfam-A domain unipartite-projection was generated from the bipartite graph by removing each protein sequence node and inferring the connections between the Pfam-A domains based on the removed protein sequences (i.e. Pfam-A co-occurrence graph) (Appendix 4.1). Where possible Pfam-A co-occurrence sub-graphs were created.

Table 4.1: Details on the vertebrate genomes used in this study.

Species	Assembly	Coverage	Contig N50
Anole Lizard	AnoCar2.0	7	79kb
Cat	Felis_catus_6.2	2 & 12	20kb
Chicken	Galgal4	12	279kb
Chimpanzee	CHIMP2.1.4	6	50kb
Coelacanth	LatCha1	77.5	12kb
Cow	UMD3.1	9	96kb
Dog	CanFam3.1	7.6	267kb
Elephant	loxAfr3	7	69kb
Fugu	FUGU4.0	8.5	52kb
Gibbon	Nleu1.0	5.6	35kb
Gorilla	gorGor3.1	2.1 & 35	11kb
Guinea Pig	cavPor3	6.79	80kb
Horse	EquCab2	6.79	112kb
Human	GRCh37.p12	High	36Mb
Macaque	MMUL 1.0	5	25kb
Marmoset	C_jacchus3.2.1	6.6	29kb
Microbat	Myoluc2.0	7	64kb
Mouse	GRCm38.p1	High	32Mb
Opossum	MonDom5	7.33	108kb
Orangutan	PPYG2	6	15kb
Panda	ailMel1	56	39kb
Platypus	OANA5	6	11kb
Platyfish	Xipmac4.4.2	19.6	22kb
Rabbit	OryCun2.0	7	64kb
Rat	Rnor_5.0	3 & 6	52kb
Stickleback	BROAD S1	11	n.s.
Turkey	Turkey_2.01	17	12kb
Xenopus	JGI41	7.6	22kb
Zebrafinch	taeGut3.2.4	6	39kb
Zebrafish	Zv9	7.5	1Mb

The common name, genome assembly, fold coverage, and contig N50 are given for each the 30 vertebrate genomes used in this chapter. All details given were gathered from Ensembl [Flicek *et al.,* 2014], NCBI Assembly, and respective genome publications. The contig N50 of the stickleback was not specified (n.s.).

4.3.2 Pfam-A domain co-occurrence graph centrality:

The following centrality measurements: degree, closeness, and betweenness, were calculated for each node within the Pfam-A co-occurrence sub-graphs using the program 'General_Stats.py' (Appendix 4.2). The calculation of degree centrality was independent of both in-degree and out-degree measurements as the sub-graphs were undirected. Closeness centrality values were normalized by the total number of remaining nodes $(n-1)$. Betweenness centrality values were normalized by maximum number of pairs of nodes not including the node of interest $\left(\frac{2}{(n-1)(n-2)}\right)$ (as per Section 1.6.2.1).

4.3.3 Node removal within unipartite-projected co-occurrence graph:

Removing the 50 nodes with the highest degree, betweenness, or closeness centrality values from the Pfam-A co-occurrence sub-graphs allowed us to determine the role these nodes play in the structure of the graph. Using the program 'node_deletion.py' (Appendix 4.2), the impact was measured by calculating graph transitivity (see Section 1.5.2.4) [Luce and Perry, 1949] and average clustering (see Section 1.5.2.4) [Watts and Strogatz, 1998] pre- and post-removal of nodes. Nodes were also removed at random from the graph and the same calculations were made. The process of selection and removal of random nodes from the graph was repeated 100 times.

4.3.4 Pfam-A co-occurrence graph assortativity:

Assortativity of the Pfam-A co-occurrence graph was visualized by plotting each edge of the graph by the degree (K) of its respective nodes. Assortativity was measured by: i) the linear regression of the average degree of nearest neighbors

for a node (<Knn>) plotted against K, and ii) the assortativity coefficient (r) [Newman, 2002; Newman, 2003], this was implemented in the program 'Network_Assortativity.py' (Appendix 4.2). A confidence interval of the assortativity coefficient can be determined by creating randomized graphs that share the same degree distribution of the graph in question [Foster *et al.*, 2010]. To determine the confidence interval, the program "Random_Assortativity" (Appendix 4.2) generated 10,000 randomized networks and computed the 95% confidence interval.

4.3.5 Identification of domain co-occurrence communities:

Pfam-A co-occurrence communities were identified from the largest Pfam-A co-occurrence sub-graph using the NeMo plugin [Rivera *et al.*, 2010] from the cytoscape package [Shannon *et al.*, 2003]. NeMo identifies communities using a hierarchical method that permits the detection of internal sub-communities. The NeMo communities produced were evaluated as either independent (*i.e.* with sub-communities) or combined (*i.e.* without sub-communities). The remaining Pfam-A co-occurrence sub-graphs were automatically classified as independent communities as they lacked connections elsewhere in the graph.

4.3.6 GO term associations and relevance in Pfam-A co-occurrence communities:

Pfam-A co-occurrence communities were evaluated in relation to gene and domain GO terms to determine if community structure correlated with function. This was achieved using the programs 'NeMo_Gstats.py' (Appendix 4.2) for gene level analysis and 'NeMo_Dstats.py' (Appendix 4.2) for domains. As the

Pfam-A co-occurrence communities exclude gene nodes, genes were associated using the initial Pfam-A homology sub-graphs. Genes were only associated with a community if they did not possess homology outside the community. To circumvent potential species-based biases in GO-term annotation, only human genes were allowed to be associated with a community. Gene GO terms were downloaded from Ensembl Gene 76 on Ensembl BioMart [Flicek *et al.,* 2014] and Pfam-A GO terms were downloaded from the gene ontology website [Ashburner *et al.,* 2000]. GO term enrichment was evaluated using Fisher's exact test and chi-squared test. Where both calculations were made, Fisher's exact test was favored. To determine if communities were accurately displaying evidence of GO term enrichment (i.e. not due to due to gene- or domain-specific GO terms only found once in the network) we calculated the recall and precision accuracy of each GO term associated with a community. Recall is the ratio of the specific GO term displayed by community members to the remainder of the network. Precision is the ratio of community members with the GO term to the community members without.

4.3.7 Enrichment of innate immunity in Pfam-A co-occurrence communities:

Pfam-A communities were also evaluated for the presence of genes involved in innate immunity using the program 'NeMo_IIstats.py' (Appendix 4.2). GO terms associated with innate immune response were downloaded from AMIGO [Carbon *et al.,* 2009] and were used to filter the previously downloaded gene GO terms (described in Section 4.2.6). Using similar methodologies to Section 4.2.6, each community was tested for enrichment of innate immune specific GO terms

using the Fisher's exact and chi-squared tests. The accuracy of functional enrichment was determined by calculating recall and precision as before (see Section 4.2.6.).

4.3.8 Identification of species-specific domain combinations:

Human, mouse, and dog orthologous gene families were downloaded from Ensembl Gene 76 on Ensembl BioMart [Flicek *et al.*, 2014]. Orthologous gene families were only included if all genes were present within the same Pfam-A homology sub-graph. Presence or absence of the Pfam-A domain was determined for each gene from the bipartite similarity graph using the programs 'ortho_domain_networker.py' (Appendix 4.2). Orthologous families that contained a member that had either a gain or loss of domain were subsequently aligned using PRANK [Loytynoja and Goldman, 2005]. The domain gain or loss events identified were then assessed at the alignment level to identify false positives due to sequence polymorphisms, this was achieved using the program 'domain_checker.py' (Appendix 4.2). MSAs not flagged as false positives were confirmed using Ensembl BLAST by searching for the identified domain gain or loss [Flicek *et al.*, 2014].

4.4 Results

4.4.1 Construction of the domain co-occurrence graph

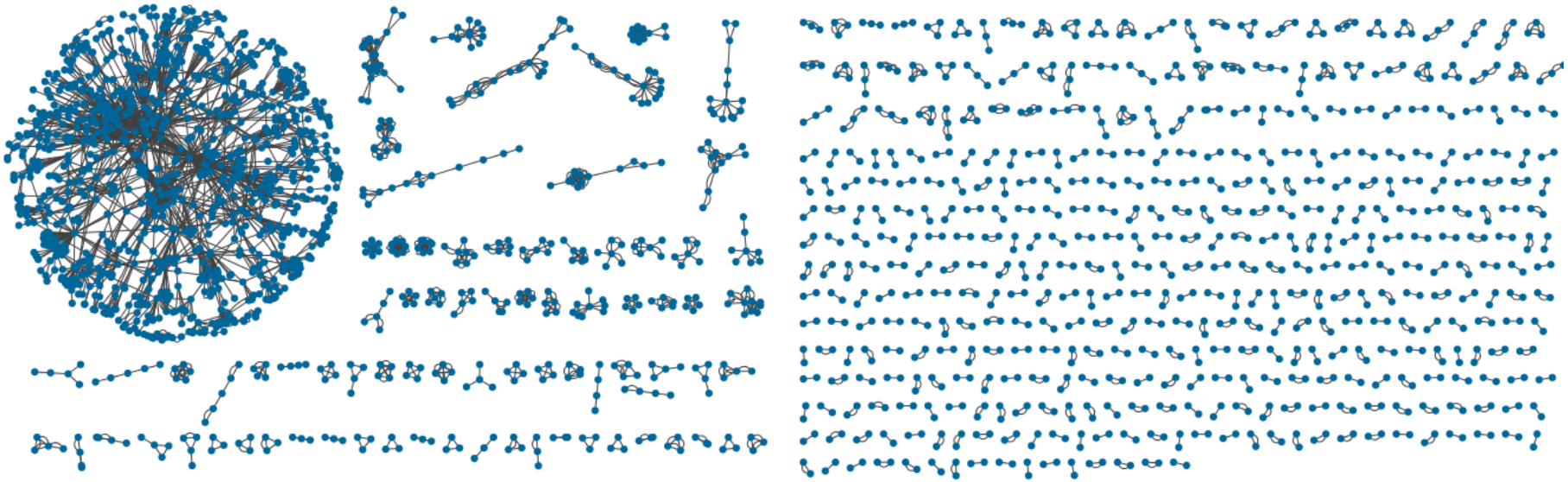
Modular proteins are typically characterized by encompassing multiple functional domains [Moore *et al.*, 2008]. However, the functional domains that are capable of residing or co-occurring within a given modular proteins have been proposed to be limited, suggesting that a number of domains function

unaccompanied by other domains [Tordai *et al.*, 2005]. To identify the functional domain combinations permitted in the evolution of vertebrate modular proteins, we constructed a bipartite graph (Section 1.5.1) from two independent datasets: the protein sequences of 30 vertebrate genomes and the HMM profiles from the Pfam-A domain database. Sequence homology provided connections between the nodes of these independent datasets. Accurate assessment of the Pfam-A domain database required our analysis to account for both composite domains (i.e. Pfam-A domains that are comprised of smaller Pfam-A domains) and domain families (i.e. multiple Pfam-A domains with similar sequence motifs) as both categories bias results by creating false signatures of domain co-occurrence. See Section 4.2.1 for details on removing biases from composite domains and domain families.

Investigating the structure of the bipartite graph revealed 3,336 connected components, with a single giant connected component comprising 36.14% of all homology connections and 11.93% of all Pfam-A domains. The presence of this giant connected component in addition to other smaller components with multiple domains indicate that approximately 40% of all known functional domains are functionally promiscuous (*i.e.* they may co-occur alongside at least one other functional domain in a modular protein) and there are specific functional domains that reside within multiple modular proteins. The majority of connected components (2,816) exhibited only a single Pfam-A domain, indicating that the majority (60%) of known biological domains are functionally exclusive and therefore cannot occur within a modular protein.

Although the majority of vertebrate proteins are not modular in nature, an accurate global description of protein evolution requires us to account for those proteins that are modular. Protein kinases form a large gene family, they are also modular and frequent within our dataset, but the kinases display preferential functional domain combinations [Manning *et al.*, 2002]. And so domains of high frequency in our dataset may not necessarily be domains that are permitted in many combinations in modular proteins. To determine the relationship between the function of a domain and its presence in modular proteins we measured the unbiased functional permissiveness of each domain in vertebrates. We constructed Pfam-A unipartite projections (Figure 4.1) by inferring connections between two Pfam-A domains if they co-occur on the same gene, following the process described in Section 1.5.1. Constructing the unipartite projections resulted in 520 Pfam-A unipartite graphs, with a single giant connected component consisting of 43.60% of all co-occurrence connections and 28.76% of the co-occurring Pfam-A domains. The presence of approximately a quarter of all co-occurring functional domains in a single component indicates that specific functional domains are able to co-occur in numerous combinations and therefore are like "glue" that holds the connected component together. In the next section (Section 4.3.2) we explore the properties of this component to determine if kinases (in addition to other domains) are representative of the "glue" pattern. Excluding the giant connected component, the remaining 519 components exhibit an average of 3 ± 2 domains. The small average of the remaining 519 components indicates that these domains may co-occur but in limited combinations.

Figure 4.1: Visiualization of the Pfam-A co-occurrence graph.



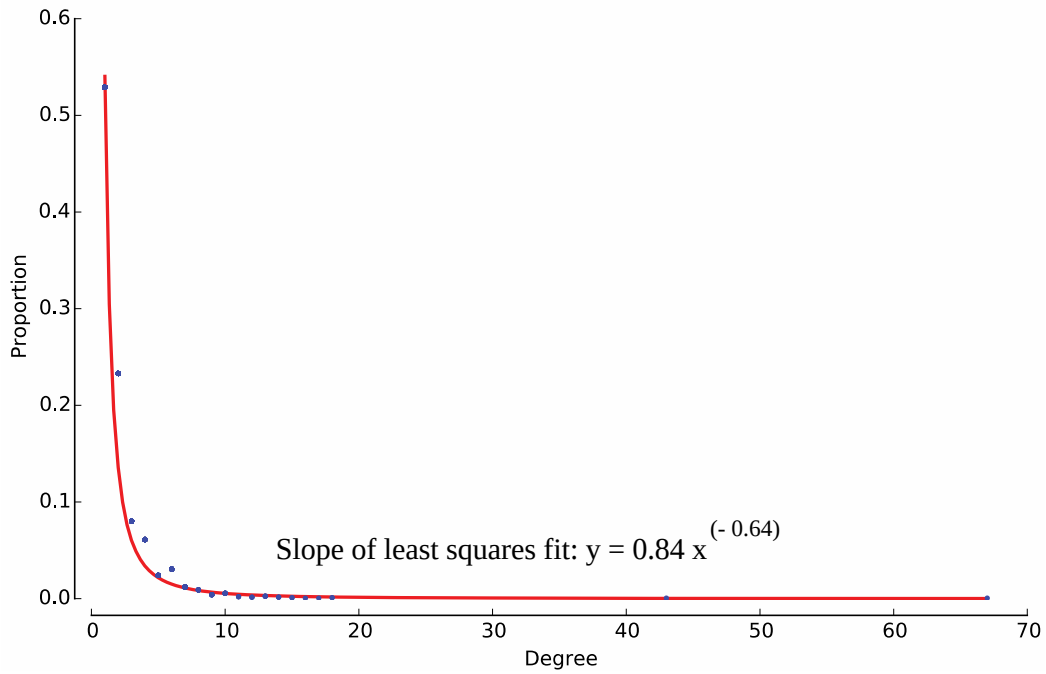
The Pfam-A co-occurrence network shows each of the 1996 Pfam-A domains found in modular proteins (i.e. nodes) as blue dots. Grey lines (i.e. edges) indicate that the connected Pfam-A domains are found to both reside within a single modular protein (i.e. co-occurrence). Co-occurring domains are separated into 520 connected components (i.e. groups of nodes connected by edges that share no edges elsewhere in the network). The presence of the large connected component (large circle) indicates that some domains are able to co-occur in numerous combinations and hold the connected component together.

4.4.2 Highly central Pfam-A domains are most functionally permissive

The properties of the Pfam-A co-occurrence network are characteristic of a scale-free graph (Figure 4.2). This is expected for biological graphs [Barabasi and Oltvai, 2004] and indicates that modular proteins exhibit a small number of functional domains that are capable of functioning in a multitude of combinations, *i.e.* the "glue" or hub nodes, whereas the majority of the functional domains are limited in their functional combinations [Barabasi and Oltvai, 2004]. To identify which of the functional domains act as hub nodes, and therefore which of the domains have the most influence in modular proteins we calculated degree, closeness, and betweenness centrality for each node in the Pfam-A co-occurrence graph (See Section 1.5.2.1 for details on the centrality measurements). Calculating the centrality measurements identified several domains that strongly influenced (greater than expected by random chance) the possible functional combinations exhibited by modular proteins (Figure 4.3). Of these domains, Pkinase was found to have the greatest overall influence as the domain exhibited the highest values for degree, closeness, and betweenness centrality. This finding is perhaps unsurprising given the wide range of functional domains reported alongside kinase domains in the human genome [Manning *et al.*, 2002]. Many of the other influential domains – Ank2, SH2, 7tm_1, and RhoGAP – are motifs frequently observed in biology and found in proteins with a multitude of functions [Pierce *et al.*, 2002; Mosavi *et al.*, 2004; Tcherkezian and Lamarche-Vane, 2007; Filippakopoulos *et al.*, 2008] (Table 4.2). The observed impact of removing central domains (i.e. high degree, closeness, and betweenness) was measured using two graph-clustering measurements: average clustering and transitivity (See Sections 1.5.2.4 for

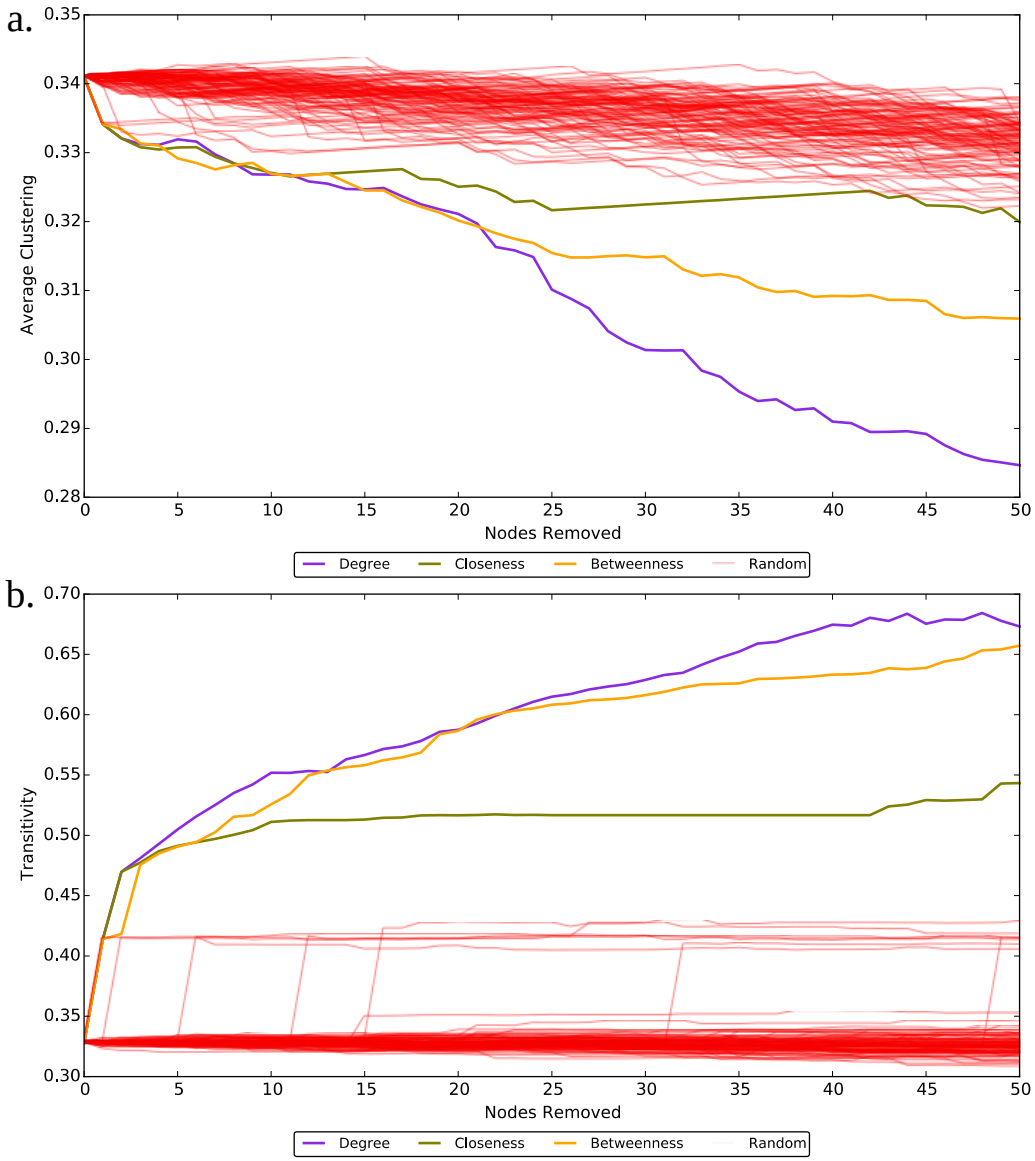
details on measurements). The observed decrease in average clustering upon removing central domains – especially domains with high degree centrality – indicates a decrease in individual domains displaying clique-structure with their nearest neighbors (including individual nodes becoming detached from the network altogether). In contrast, the observed increase in transitivity upon removing central domains indicates an increase in cliques among connected triplets (i.e. three nodes connected by at least two edges) (Figure 4.3). Taken together, these results indicate that removing central domains is resulting in stronger community/clique structure among the network by losing individual sparsely connected domains that frequently form triplets but infrequently form cliques. This interpretation is supported by a continuous decrease of triangles (*i.e.* triplets displaying clique structure) in the transitivity calculation after each deletion event, demonstrating a much greater decrease in the total number of triplets to result in the continuous increase in transitivity (Appendix 4.3). From a biological perspective, these central or "glue-like" domains are highly promiscuous, they are present in many different modular proteins, and without these central domains there would be no other relationship between these proteins.

Figure 4.2: The degree distribution of the Pfam-A co-occurrence graph is scale-free



The chart shows the proportion of domains in the co-occurrence network on the y-axis that exhibit the degree centrality value (i.e. the number edges possessed by the node) on the x-axis. The graph indicates that the majority of domains in the co-occurrence network exhibit low degree centrality values with approximately 50% of domains exhibiting a degree centrality of one (i.e. only exhibit a single edge). Additionally, the graph shows that a much smaller number of domains exhibit high degree centrality values (≥10). In graph theory this degree distribution pattern is termed scale-free as it adheres to a power law slope $\left(P(k) \sim k^{-\gamma}\right)$ (see Section 1.5.1 for details). This interpretation is supported by the slope of the least squares following a power law ($y = 0.84x^{-0.64}$). Scale-free graphs are characterized by having a small number of nodes that possess far more connections than the other nodes within the network. These highly connected nodes are termed "hubs" and highly influential within the network.

Figure 4.3: Changes to average clustering and transitivity upon removal of domains.



The impact removing nodes with high centrality on two measures of graph clustering: (a) average clustering, and (b) transitivity. Both (a) and (b) show their respective measurement on the y-axis for each domain removed (on the x-axis). The order in which domains were removed from the network was either based on centrality values (see Section 1.5.2.1) – degree (violet), closeness (green), or betweenness (orange) – or by random selection (red). Random selection was repeated 100 times for an accurate sampling of random domains. See Table 4.2

for the list of domains removed by centrality values. (a) illustrates the affect of node deletion on the average clustering, a measure of clique structure among each given node and the nearest neighbors of that given node within the network (see Section 1.5.2.4 for details). The chart indicates that removing domains based on degree or betweenness centrality decreased the average clustering of the network far more than expected at random. In comparison, removing domains based on closeness centrality decreased the average clustering only slightly more than expected at random. In both instances the decrease in average clustering shows a decrease in the number of individual domains displaying clique-structure with their nearest neighbors within the network. (b) illustrates the affect of node deletion on transitivity, a measure of clique structure among connected triplets (i.e. three nodes connected by edges) within the network (see Section 1.5.2.4 for details). All three centrality measurements were found to increase the transitivity of the network far more than expected at random. This indicates that the removal of central domains is increasing the number of cliques among connected triplets.

Table 4.2: The Pfam-A domains removed from the co-occurrence graph to measure average clustering and transitivity.

	Degree Centrality		Closeness Centrality		Betweeness Centrality	
#[1]	Node Deleted	Value	Node Deleted	Value[2]	Node Deleted	Value[2]
1	Pkinase	67	Pkinase	0.31209	Pkinase	0.56188
2	Pkinase_Tyr	43	Pkinase_Tyr	0.29145	RabGAP-TBC	0.21173
3	VWA	18	RhoGEF	0.27016	Pkinase_Tyr	0.19535
4	GVQW	18	Ank_2	0.26826	Ank_2	0.18470
5	7tm_1	17	SH2	0.26503	SNF2_N	0.12246
6	Ank_2	16	F5_F8_type_C	0.26321	Myosin_head	0.11926
7	SNF2_N	15	C2	0.26105	RhoGAP	0.11140
8	RhoGEF	15	Fz	0.26081	7tm_1	0.09532
9	Laminin_G_2	14	Myosin_head	0.26045	Filament	0.08818
10	Trypsin	14	RabGAP-TBC	0.25569	Trypsin	0.08358
11	RhoGAP	14	Kringle	0.25489	RhoGEF	0.08199
12	SPRY	13	RBD	0.25421	GVQW	0.07881
13	7tm_2	13	SH3_1	0.25376	SH2	0.07775
14	Cadherin	13	F_actin_bind	0.25376	HECT	0.07246
15	Y_phosphatase	13	I-set	0.25365	NACHT	0.07044
16	VWD	12	Inhibitor_Mig-6	0.25354	F5_F8_type_C	0.06922
17	FERM_M	12	Miro	0.25132	C2	0.06664
18	RabGAP-TBC	12	ANF_receptor	0.25088	Bromodomain	0.06518
19	MAM	11	OLF	0.25033	VWA	0.06453
20	CUB	11	Guanylate_cyc	0.24956	DEAD	0.06102
21	C2	11	Death	0.24902	Y_phosphatase	0.05927
22	HECT	10	CNH	0.24859	Fz	0.05354
23	Bromodomain	10	Ephrin_lbd	0.24752	SET	0.05225
24	Myosin_head	10	EphA2_TM	0.24752	PRY	0.05183
25	Fz	10	Recep_L_doma	0.24741	ANF_receptor	0.05152
26	DUF3497	10	SAM_1	0.24741	Exo_endo_phos	0.04814
27	ANF_receptor	10	Furin-like	0.24741	SPRY	0.04783
28	AAA	10	GF_recep_IV	0.24741	MHC_I	0.04776
29	Kinesin	9	Sema	0.24741	PARP	0.04709
30	ABC_tran	9	DUF4071	0.24730	RVT_1	0.04522
31	NACHT	9	GTPase_bindin	0.24730	C1-set	0.04467
32	UCH	9	DCX	0.24720	UCH	0.04143
33	Laminin_G_1	9	ApoL	0.24720	Cadherin	0.03908

Table 4.2: The Pfam-A domains removed from the co-occurrence graph to measure average clustering and transitivity.

	Degree Centrality		Closeness Centrality		Betweeness Centrality	
#[1]	Node Deleted	Value	Node Deleted	Value[2]	Node Deleted	Value[2]
34	F5_F8_type_C	9	PBD	0.24720	PH	0.03772
35	SH2	9	KSR1-SAM	0.24720	PWWP	0.03735
36	DEAD	8	POLO_box	0.24720	MAM	0.03584
37	FERM_N	8	Mst1_SARAH	0.24720	Sec7	0.03445
38	PARP	8	Focal_AT	0.24720	Transposase_22	0.03436
39	SET	8	DUF1908	0.24720	MIT	0.03435
40	RVT_1	8	PKK	0.24720	I-set	0.03241
41	NTR	8	Ig_Tie2_1	0.24720	Guanylate_kin	0.03182
42	Filament	7	PH_3	0.24720	PABP	0.03097
43	SRCR	7	Trypsin	0.24635	VWD	0.02988
44	I-set	7	Filament	0.24603	NTR	0.02923
45	FERM_C	7	MAM	0.24096	Laminin_G_2	0.02817
46	Guanylate_kin	7	Guanylate_kin	0.24076	AAA	0.02774
47	Laminin_N	7	RGS	0.24055	RRM_1	0.02773
48	RRM_1	7	MIT	0.23995	CUB	0.02750
49	A2M_recep	7	7tm_1	0.23945	OLF	0.02732
50	PWWP	7	PX	0.23895	7tm_2	0.02656

[1]Deletion iteration
[2]Rounded values

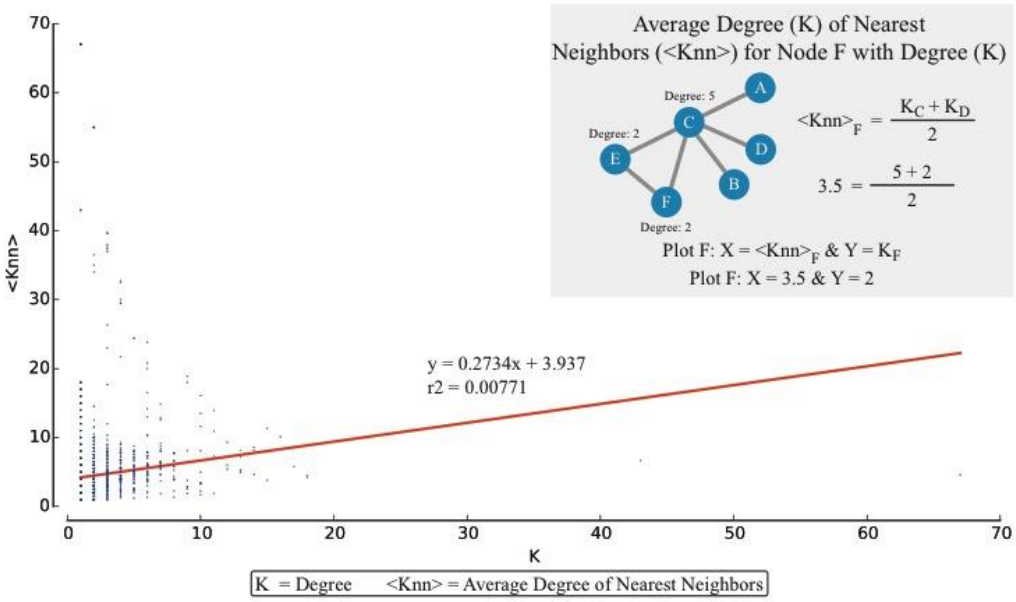
The order of removal of nodes, the functional domain removed, and centrality score for each of the centrality measurement assessed is given. The centrality scores used for selection were based on the original Pfam-A co-occurrence graph, the scores given for closeness and betweenness are rounded to increase legibility of the table.

4.4.3 Modular proteins exhibit a preference for domains with similar functional permissiveness

To determine the influence of specific domain combinations on modular proteins we calculated the degree assortativity (i.e. preference of nodes to share edges due to similar degree centrality values of the respective nodes) of each co-occurrence connection within the Pfam-A co-occurrence graph (See Section 1.5.2.2 for details on calculating assortativity). The Pfam-A co-occurrence graph was determined to be assortative from a neighbor connectivity plot, indicating a preference of nodes to attach if they share exhibit degree values (Figure 4.4). It is possible that the large number of low degree nodes may have unrealistically influenced this measurement, and so we also obtained the degree assortativity coefficient for the network – which we calculated as 0.03301 and was determined to be significant using 10,000 randomized graphs with the same degree distribution [Foster *et al.,* 2010] (Appendix 4.4). Together these network metrics indicate that Pfam-A co-occurrence is an assortative network comparable for example to the patterns observed in co-authorship networks, which are characterized by a preferential attachment of nodes with similar degree centrality values [Newman, 2003]. This finding is in direct contrast to the disassortative mixing patterns (i.e. preferential attachment of high-degree nodes with low-degree nodes) that characterize most biological networks [Newman, 2002]. Therefore, in our exploration of vertebrate modular protein evolution we see that domains with similar domain centrality values are more likely to co-occur. For example, the modular protein HACE1 (HECT domain and ankyrin repeat containing E3 ubiquitin protein ligase 1) [Anglesio *et al.,* 2004; Flicek *et al.,* 2014] contains the HECT and Ank_2 (Ankyrin 2) domains whereas TNKS2

(TRF1-interacting ankyrin-related ADP-ribose polymerase 2) [Lyons *et al.*, 2001; Flicek *et al.*, 2014] contains the Ank_2 and PARP (Poly ADP ribose polymerase) domains, in both cases the pair of domains have similar degree centrality values. The results indicate that modular proteins are preferentially constructed with domains exhibiting similar co-occurrence possibilities (i.e. permissiveness to reside with other domains).

Figure 4.4: Assortativity of the Pfam-A co-occurrence graph.



One method for measuring network assortativity is generating a neighbor connectivity plot by plotting the average degree centrality of nearest neighbors <Knn> on the y-axis for a node with degree K on the x-axis. As illustrated in the box above, <Knn> is the average degree centrality of the nodes that share an edge with the given node (i.e. nearest neighbors of the given node), for example, the <Knn> for node F is the average degree centrality of C and E $\left(\frac{K_C + K_E}{2}\right)$ and is plotted against the K of F. The chart shows that the Pfam-A graph displays assortative mixing patterns as the positive slope (y = 0.2734x + 3.937) of the trend line (i.e. linear regression of K and <Knn>) indicates that the degree of a given node (K) increases alongside the degree of its nearest neighbors <Knn>. It should be noted that the large number of nodes with low degree values may have unrealistically influenced the linear regression – causing the low r-squared value of 0.00771 – and indicates that additional confirmation is required.

4.4.4 Functional domain combinations are influenced by function

Many Pfam-A domains, such as Pkinase and Pkinase_Tyr, have been associated with specific biological functions [Manning *et al.*, 2002; Scheeff and Bourne, 2005]. Kinase domains in particular are documented to frequently co-occur with a variety of functional domains to construct a functional modular protein [Manning *et al.*, 2002]. We wished to determine if Pfam-A domain combinations were associated with function at three distinct levels: i) domain functional enrichment (do domain combinations occur more often between domains of similar function?), ii) gene functional enrichment (do specific domain combinations occur more often in genes of similar function?), and iii) biological pathway functional enrichment (do domain combinations occur more frequently in specific pathways?). The innate immune system was selected as the set of biological pathways for analysis of functional enrichment.

Pfam-A communities were generated from each of the 518 co-occurrence graphs to identify frequently connected domain combinations (see Section 1.5.2.3 for description of network communities). Each community was assessed for potential functional enrichment using GO terms associated with either the Pfam-A domains (in the case of domain functional enrichment analysis) or the associated human proteins (in the case of gene and pathway functional enrichment analyses). Identifying functional enrichment was achieved by using a one-tailed (enrichment required) Fisher's exact test. In total, there were 1,376 domains, 26,843 proteins, and 147 innate immune GO terms identified as significantly enriched ($P < 0.05$). However, this approach includes GO terms that are only associated with a single protein or domain and were not found elsewhere

in the Pfam-A network. Such GO terms do not accurately reflect enrichment due to their small sampling. To more accurately quantify community enrichment, two precision measurements (precision and recall) were also calculated. Using a precision threshold of 0.5 (i.e. requiring 50% of the community to possess the GO term) refined the results down to 243 domains, 4,363 proteins, and 20 innate immune genes that had GO terms that were significantly enriched ($P < 0.05$). Table 4.3 contains details on the communities that exhibited the highest precision and recall scores (i.e. fraction of GO term within the community vs. the entire network). Communities exhibiting high recall possess functional enrichment that is unique and not frequently observed elsewhere in the network. Indicating that the particular domain combination is not frequently observed in other combinations. Communities exhibiting higher precision possess more members (i.e. domains and genes) exhibiting a specific functional enrichment, indicating that more of the domain combination is required for the function. Therefore, communities exhibiting both high recall and precision indicate that the majority of the domain combination is required for a unique function. In summary, we show that modular proteins occasionally require a specific combination of domains to function independently or within a pathway such as innate immunity, or they may require a combination of functionally similar domains to function. For a full list of the enrichment findings see Appendix 4.5.

From our analysis we found that multiple functionally enriched communities displayed consistently low degree centrality values with an average of 1.8 and a standard deviation of 1.2 (Appendix 4.6). Low centrality would indicate that these communities contain domain combinations that have limited functional

combinations. Degree centrality was also determined to be similar among members of the same community, with an average standard deviation across all communities of 0.17 (Appendix 4.6). The lack of deviation in degree that we observe in domains that make up the vertebrate modular proteins is also supported by the assortativity of the network. Therefore, it appears that functionally enriched domain combinations are restricted in their ability to operate in other combinations and favor combinations of domains with similar restrictions.

Table 4.3: Community structure and gene GO-term, domain GO-term, and innate immune functional enrichment.

	Term	Group Size	Members w/Term	Non-group w/Term	Recall	Precision	Fisher's Exact odds ratio	Fisher's Exact P-value
Protein Enrichment	cation-transporting ATPase activity	27	27	0	1.000	1.000	inf	7.10E-85
	connexon complex	22	22	0	1.000	1.000	inf	4.35E-71
	neurotransmitter:sodium symporter activity	21	21	0	1.000	1.000	inf	2.82E-68
	diacylglycerol kinase activity	15	15	0	1.000	1.000	inf	6.14E-51
	calcium-dependent cysteine-type endopeptidase activity	15	15	0	1.000	1.000	inf	6.14E-51
	protein-glutamine gamma-glutamyltransferase activity	9	9	0	1.000	1.000	inf	1.45E-32
	ribose phosphate diphosphokinase activity	5	5	0	1.000	1.000	inf	2.01E-19
	glucokinase activity	5	5	0	1.000	1.000	inf	2.01E-19
	phosphopyruvate hydratase activity	5	5	0	1.000	1.000	inf	2.01E-19
	phosphopyruvate hydratase complex	5	5	0	1.000	1.000	inf	2.01E-19

Table 4.3 Community structure and gene GO-term, domain GO-term, and innate immune functional enrichment.

	Term	Group Size	Members w/Term	Non-group w/Term	Recall	Precision	Fisher's Exact odds ratio	Fisher's Exact P-value
Domain Enrichment	DNA ligase (ATP) activity	4	4	0	1.000	1.000	inf	1.52E-12
	thiamine pyrophosphate binding	3	3	0	1.000	1.000	inf	7.56E-10
	negative regulation of transcription, DNA-dependent	3	3	0	1.000	1.000	inf	7.56E-10
	S-adenosylmethionine biosynthetic process	3	3	0	1.000	1.000	inf	7.56E-10
	methionine adenosyltransferase activity	3	3	0	1.000	1.000	inf	7.56E-10
	protein-arginine deiminase activity	3	3	0	1.000	1.000	inf	7.56E-10
	folic acid-containing compound biosynthetic process	3	3	0	1.000	1.000	inf	7.56E-10
	intramolecular transferase activity, phosphotransferases	3	3	0	1.000	1.000	inf	7.56E-10
	arginyl-tRNA aminoacylation	3	3	0	1.000	1.000	inf	7.56E-10
	arginine-tRNA ligase activity	3	3	0	1.000	1.000	inf	7.56E-10

Table 4.3 Community structure and gene GO-term, domain GO-term, and innate immune functional enrichment.

	Term	Group Size	Members w/Term	Non-group w/Term	Recall	Precision	Fisher's Exact odds ratio	Fisher's Exact P-value
Inante Immune Enrichment	RIG-I signaling pathway	3	2	1	0.667	0.667	28598	8.80E-08
	positive regulation of type I interferon-mediated signaling pathway	3	2	5	0.286	0.667	5718	6.16E-07
	positive regulation of innate immune response	2	2	8	0.2	1.000	inf	4.40E-07
	type I interferon-mediated signaling pathway	9	9	88	0.093	1.000	inf	2.08E-20
	interferon-gamma-mediated signaling pathway	9	9	126	0.067	1.000	inf	4.54E-19
	innate immune response	24	16	485	0.032	0.667	56.88247423	2.31E-18
	innate immune response	17	10	491	0.02	0.588	40.13674716	3.97E-11
	innate immune response	4	4	497	0.008	1.000	inf	1.49E-06
	innate immune response	2	2	499	0.004	1.000	inf	1.22E-03
	innate immune response	2	2	499	0.004	1.000	inf	1.22E-03

The 10 communities with the highest precision and recall for gene/protein, domain, and innate immune system functional enrichment. Calculations were determined using the size of the community (Group Size), the number of community members with the term in question

(Members w/Term), and the number of non-community members with the term (Non-group w/Term). The results of the Fisher's exact test (odds ratio and P-value) are given. An odds ratio of “inf” (infinity) indicates a large difference.

4.4.5 Species-specific domain combinations exhibit unique properties

Species-specific domain combinations in modular proteins may represent novel functional domain combinations. We wished to identify if there were species-specific combinations present in the vertebrate modular protein network and if these novel combinations displayed similar characteristic to our previous findings of the entire network. To this end we identified species-specific domain combinations in the Pfam-A co-occurrence network using orthologous genes from human, mouse, and dog.

The domains responsible for the establishment of the species-specific combinations were classified by protein position (5', internal, and 3') and most likely mechanism that created the event. Three creation mechanisms were observed within the Pfam-A network: (i) complete domain events (CDEs) characterized by indels that contain an entire Pfam-A domain motif, (ii) incomplete domain events (IDEs) characterized by indels that only contain a fraction of a Pfam-A domain motif, and (iii) composite gene events (CGEs) characterized by a single gene that contains the coding sequence of two nonallelic genes (Figure 4.4a).

Identification of species-specific domains were limited to human and mouse to minimize false positives due to poor assemblies and low alternative transcript counts (Appendix 4.7). Analysis of human and mouse orthologs resulted in the identification of 122 potential species-specific domain combinations. Manual inspection and Ensembl BLAST [Flicek *et al.*, 2014] identified 113 false positives present due to poor alignment and alternative transcripts. Of the

remaining 9 events there were 2 IDEs, 2 CDEs, and 5 CGEs (Table 4.4). Calculating the assortativity of the modular proteins exhibiting species-specific combinations alone (the edges of species-specific modular proteins) resulted in an assortativity coefficient of -0.1807 (Appendix 4.8), indicating disassortative mixing patterns. Therefore in this small subset of species-specific combinations there is a preferential attachment of high-degree nodes with low-degree nodes, similar to most biological networks [Newman, 2002]. This is in direct contrast to the assortative mixing patterns observed for the entire network that indicated a preference for modular proteins to incorporate domains with similar combination possibilities. In terms of modular proteins, the domains responsible for species-specific combinations are frequently found in modular proteins with dissimilar co-occurrence ability.

For each of the 9 species-specific events identified, the dog ortholog (where available) was used to determine the genetic mechanism behind the event and determine if the event was a species-specific domain gain or loss. Dog was selected as an output due to being the closest non-euarchontoglires mammal with the highest frequency of alternative transcripts (Appendix 4.7). This analysis of the data and additional confirmation by Ensembl BLAST (i.e verification with genomic DNA) provided evidence for four genetic mechanisms that have generated the observed species-specific combination events (Figure 4.4b). These mechanisms comprised: (i) the gain or loss of exons, (ii) the extension of exons by indels with an additional 5' splice site, (iii) the partial gain or loss of an exon, and (iv), transcription readthrough events (Table 4.4).

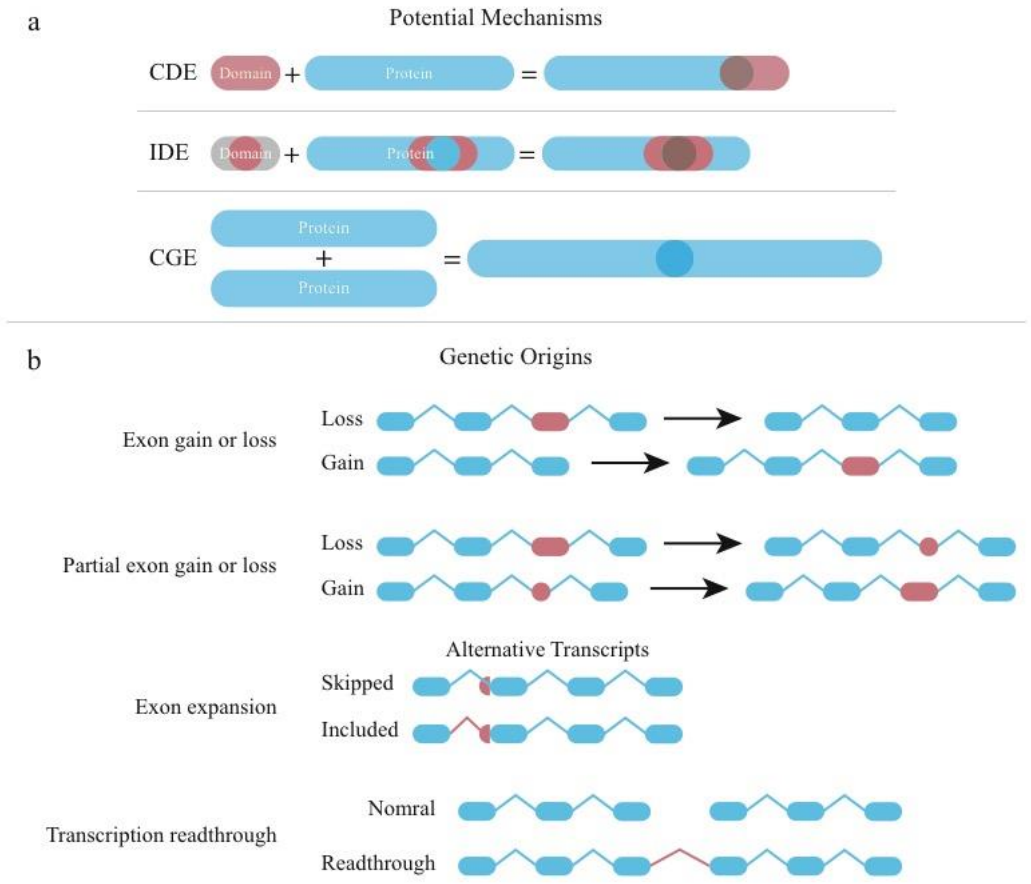
Table 4.4: Details on species-specific domain combinations identified from the Pfam-A domain co-occurrence graph.

Species	Gene Name	Domain Name(s)	Location	Mechanism	Genetic Origin	Species Status (Present or Absent) *Hs* | *Mm* | *Cf*
Mouse	Kxd1	Ribosomal_L40e, ubiquitin	3'	CGE	Readthrough	− | + | −
Human	FIP1L1	Pkinase_Tyr	3'	CGE	Readthrough	+ | − | −
Human	PPAN-P2RY11	7tm_1	3'	CGE	Readthrough	+ | − | −
Human	PRR5-ARHGAP8	HbrB	3'	CGE	Readthrough	+ | − | −
Human	IQCJ-SCHIP1	IQ-like	5'	CGE	Readthrough	+ | − | −
Human	AZIN2	Orn_DAP_Arg_deC	Internal	IDE	Exon expansion	+ | − | +
Mouse	Olfr1260	7tm_1	Internal	IDE	Partial exon indel	− | + | +
Human	MMP13	GVQW	3'	CDE	Exon indel	+ | − | −
Human	NLRP5	PYRIN	Internal	CDE	Exon indel	+ | − | +

The species, gene name, domain name, and proximal protein location (3', 5', and internal) are given for the postulated domain(s) responsible for the species-specific combinations. The table indicates the proposed mechanisms and genetic origin of each event. In addition, the current status

of the postulated domain(s) are given in the human (*Hs*), mouse (*Mm*), and dog (*Cf*) Ensembl genome assemblies, status of the postulated domain(s) are either present within assembly (+) or absent (-).

Figure 4.4: Schematic of the mechanisms and genetic origins of species-specific domain combinations in the network of human, mouse and dog.



(a) The three introgressive mechanisms found by the Pfam-A co-occurrence network. CDEs involve the gain or loss of a complete Pfam-A domain. IDEs involve the gain or loss of an incomplete Pfam-A domain into a protein that contains sequence that completes the Pfam-A domain. CGEs involve the fusing of two or more non-allelic proteins to become a single protein. (b) Schematic of four possible genetic origins responsible for the introgressive events reported in the Pfam-A co-occurrence network. Exon gain or loss: In this instance, a unique domain combination is created by the fusion (gain) or fission (loss) of an exon that contained a Pfam domain. Partial exon gain or loss: In this instance, a unique domain combination is created by the fusion (gain) or fission (loss) of a sub-sequence of an exon that contained a Pfam domain. Exon-expansion: In this

instance, a unique domain combination is created by the fusion (gain) or fission (loss) of a sequence that contains a splice site that either extends (gain) or reduces (loss) an exon sequence. Transcription readthrough: In this instance, a unique domain combination is created by the fusion (gain) of an entire gene by a readthrough event.

4.5 Discussion

In this chapter, we have used graph theory to globally characterize and identify the species-specific domain combination events that underpin the evolution of modular proteins in vertebrates. In agreement with previous co-occurrence network analyses that used only single-species [Wuchty, 2001; Wuchty and Almaas, 2005], our results indicate that domain co-occurrence (i.e. the ability of two domains to exist on a common gene) is characteristically scale-free in vertebrate evolution. In biological terms, these findings indicate that only a small number of domains (i.e. hub nodes, see degree centrality in Table 4.2) are tolerated in the construction of most modular proteins, whereas the majority of functional domains are restricted in terms of the combinations of modular proteins where they are found. The topology of the network was also found to exhibit patterns of assortative mixing, indicating that modular proteins favour combinations of domains with similar degree centrality values, this is a previously undocumented characteristic of modular proteins. The presence of such preferences suggests that domains that are restricted in their functional combinations are incompatible with domains that have numerous functional combinations.

Restrictions in domain combinations in modular proteins is further explained by the identification of graph communities, indicating that modular proteins have a biological preference for specific combinations of domains beyond just similar degree centrality values. A number of graph communities were found to display evidence of functional enrichment, indicating that these specific domain combinations may be favored for functional reasons. One community of four

Pfam-A domains (DNA_ligase_IV, DNA_ligase_A_C, DNA_ligase_A_M, and DNA_ligase_A_N) was identified as significantly enriched for DNA ligase activity (Table 4.3). All four domains exhibited low degree centrality values (Appendix 4.6) in keeping with the assortative mixing of domain co-occurrence. The combination of functional enrichment and low centrality was frequently observed in our results (Appendix 4.6) and may not be so surprising as functions such as DNA ligase activity and RNA polymerase activity (Appendix 4.5) are essential biological functions that require specialized domains and may be disrupted or become harmful in combinations with domains able to function in or interact with a wider array of modular proteins.

The possible domain combinations of modular proteins are further complicated by the presence of species-specific domain combinations. Such events represent the creation of an altered modular protein due to the gain or loss of a domain [Bapteste *et al.,* 2012]. The network topology for our human, mouse and dog comparison indicated that humans exhibits more species-specific combinations than mouse, but this could be partially explained by the higher frequency of alternative transcripts in the human genome assembly (Appendix 4.7). In comparison to the assortative mixing patterns of modular proteins, these species-specific domain combinations are characterized by disassortative mixing patterns. The disassortative mixing patterns may be partially explained by the non-vertical mechanisms that generated the event, and suggests that assortativity could potentially enable the global identification of species-specific domain combinations from network topology alone.

The biological implication of species-specific domain combinations such as those identified in this chapter is the potential alteration of the function of a modular protein, this is plausible particularly when considering the observed association of functional enrichment with low centrality and the disassortative mixing patterns of the species-specific domain combinations. A single human-specific combination was due to the putative binding domain GVQW [Finn *et al.,* 2014]. The presence of GVQW could therefore potentially alter the binding capabilities of the modular protein in which it is contained. The mouse-specific combination between the YchF-GTPase_C domain (hypothesized to be required for signal transduction or ribosome function [Caldon *et al.,* 2001]) and the Nrf1 gene (transcription factor linked to regulating cellular growth, respiration, heme biosynthesis, and mtDNA transcription and replication [Pruitt *et al.,* 2014]) could potentially be responsible for various species-specific functions. To properly investigate the biological implications of these putative species-specific domain combinations events, *in vitro* functional assessment would be essential to determine that the gene and protein products of these putative species-specific events are expressed, and if they are expressed, to characterize and compare their functions.

In closing, network biology offers a powerful tool for studying species-specific domain combinations in the evolution of modular proteins. The identification of species-specific combinations and determining their characteristics is an important step to understand the causes of species-specific functions. The network topology and network analyses carried out in this chapter have enabled

us to uncover previously unknown characteristics that are unique to species-specific events in vertebrate modular proteins.

Chapter 5: Discussion

In this thesis the innate immune system has been employed as a model system to better understand phenotypic discordance at the molecular level that is potentially governed by protein functional shift. Our approach was to use large-scale screens for molecular signatures of positive selection between species as a proxy for functional discordance, an approach that has proven successful in previous studies of individual proteins [Sawyer *et al.*, 2005; Loughran *et al.*, 2012, Moury and Simon, 2011; Farhat *et al.*, 2013]. We identified a large number of genes displaying species-specific positive selection in the extant mouse lineage in particular. Examining these results alongside reported instances of functional discordance there are a number of positively selected genes implicated in known cases of phenotypic discordance (Section 3.4) [Wakimoto *et al.*, 2002; Pumidonming *et al.*, 2011].

Using signatures of positive selection to predict potential functional discordance has particular relevance in deepening our understanding of the relationship between genotype and phenotype. In the future, this approach could direct the choice of model organism in which a drug will be tested, but it could also be used to determine which model organism will produce the closest mimic of a human genetic disorder. Using the software designed in this thesis, this analysis could easily be expanded in future projects to include lineage tests on other model organisms such as hamster and rat, and indeed to test non-model organisms that are gaining significance in biomedical research [Kim *et al.*, 2011]. Our findings in Chapter 3 provide a number of potential molecular candidates to assist in the current attempt to "humanize" the mouse model for the immune system [Garcia and Freitas, 2012; Ito *et al.*, 2012].

Large-scale analyses of selective pressure variation is not straight-forward as evidenced by the complexity of the software designed in Chapter 2, and also there were many difficulties and limitations encountered due to data quality. The poor quality of sequences and assemblies resulted in unacceptably high levels of false positives in our analysis of the bowhead whale genome (Section 2.12) and spurred us to use very high quality genomes in subsequent large-scale analysis presented in Chapter 3.

Initially there were 112 genes estimated as positively selected in the bowhead lineage alone, upon close inspection and following manual filtering for gene annotation and sequencing errors this total was adjusted to just 14 genes. The bowhead whale study highlighted the importance of genome and alignment quality for the accurate identification of positive selection [Schneider *et al.*, 2009]. Unfortunately this limits selective pressure analyses to species that have genomes of suitably high quality (at minimum > 6X coverage) (as in Chapter 3). Our analysis was also limited by the absence of genome-wide population data for the vast majority of currently sequenced species. We implemented population level data analytics for the genes displaying positive selection in the human lineage [Tajima, 1989; Fay and Wu, 2000]. But these analyses were not possible for our genes displaying positive selection in the mouse lineage as we currently lack population data (we did attempt analyses using the 17 mouse genomes but these are laboratory strains and are two few in number). A greater amount of data on disease-associated mutations would allow us to link more precisely functional discordance to molecular signatures.

There are of course many other causes of phenotypic discordance along with point mutational processes studied in Chapters 2 and 3, they include but are not limited to: regulatory differences [Prud'homme *et al.,* 2006; McLean *et al.,* 2011; Wittkopp and Kalay, 2012], differences in gene duplication strategies/gene family repertoires [Blanc and Wolfe, 2004; Sullivan *et al.,* 2009; Brown *et al.,* 2010; Dennis *et al.,* 2012; Abascal *et al.,* 2013], copy number variation [Dumas *et al.,* 2007; Perry *et al.,* 2008], and epigenomic differences [Feng and Jacobsen, 2011; Zeng *et al.,* 2012] between species. During the process of identifying protein families for analysis in Chapters 2 and 3, we noticed patterns in the protein coding sequences that were suggestive of introgression or gene remodeling (e.g. domain shuffling) playing an important role in vertebrate protein evolution. This observation spurred us to explore non-linear patterns that lead to novel protein coding genes in vertebrate evolution (Chapter 4).

In Chapter 4 we applied graph theory to study the prevalence and role of introgressive events (gene remodeling) in the emergence of novel genes. Our analysis revealed not only multiple species-specific introgressive events in the evolution of vertebrate modular proteins, but it also determined the unique evolutionary principles that govern remodeling in vertebrate protein coding space. We discovered that vertebrate modular proteins are more likely to be composed of domains that share similar promiscuity levels. In addition we found that there was a preference for proteins exhibiting unique functions to incorporate domains with limited promiscuity levels (Section 4.3.3 and Section 4.3.4). These discoveries suggest that introgressive events do not strictly adhere to these same principles (Section 4.3.5). It should be stated that our graph theory

approach was not without challenges, including the detection of false positives primarily due to differences in the frequency of alternative transcripts between species. This again demonstrates that low quality assemblies result in unacceptably high levels of false positives.

While the application of graph theory to evolutionary biology is still in its infancy, this approach has already led to advances in important theoretical concepts such as the ortholog conjecture [Haggerty *et al.*, 2014]. The application of network theory to detecting and characterizing non-linear gene remodeling is providing important insights into the complex nature of protein change over time and is contributing to theoretical advancements in the field of evolutionary biology [Bapteste *et al.*, 2012; Bapteste *et al.*, 2013; Haggerty *et al.*, 2014]. The domain shuffling described in Chapter 4 generates "partially orthologous" sequences that are divergent in function and also are potentially lineage-specific [Gharib and Robinson-Rechavi, 2011; Haggerty *et al.*, 2014]. These results warrant further study at the *in silico* and *in vitro* level to determine the functional impact of species-specific gene remodeling.

Conclusion:

The research conducted in this thesis employed the innate immune system to elucidate the evolution of unique protein function within vertebrates. We successfully developed a high throughput pipeline that greatly simplified the large-scale analysis of protein coding sequence datasets. Examining the human and mouse innate immune systems identified a number of genes with species-specific signatures of positive selection. Our approach was found to be able to

accurately identify functional discordance from sequence data for known cases of phenotypic discordance. Our investigation of gene remodeling by domain shuffling revealed how frequent this mechanism of protein evolution is and what the rules for gene remodeling are (e.g. which domains are compatible and incompatible). Our analysis revealed the prevalence of species-specific gene remodeling events across these vertebrate species and highlighted the importance of domain shuffling for the introduction of novel proteins into the innate immune system and indeed into the vertebrate species tested.

Chapter 6: Bibliography

Abascal F, Corpet A, Gurard-Levin ZA, Juan D, Ochsenbein F, Rico D, Valencia A, Almouzni G. 2013. Subfunctionalization via adaptive evolution influenced by genomic context: the case of histone chaperones ASF1a and ASF1b. Molecular Biology and Evolution 30:1853-1866.

Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT, McVean GA. 2012. An integrated map of genetic variation from 1,092 human genomes. Nature 491:56-65.

Abi-Rached L, Jobin MJ, Kulkarni S, McWhinnie A, Dalva K, Gragert L, Babrzadeh F, Gharizadeh B, Luo M, Plummer FA, *et al.* 2011. The shaping of modern human immune systems by multiregional admixture with archaic humans. Science 334:89-94.

Alcaide M, Edwards SV. 2011. Molecular evolution of the toll-like receptor multigene family in birds. Molecular Biology and Evolution 28:1703-1715.

Aleshin AE, Schraufstatter IU, Stec B, Bankston LA, Liddington RC, DiScipio RG. 2012. Structure of Complement C6 Suggests a Mechanism for Initiation and Unidirectional, Sequential Assembly of Membrane Attack Complex (MAC). Journal of Biological Chemistry 287:10210-10222.

Alexopoulou L, Holt AC, Medzhitov R, Flavell RA. 2001. Recognition of double-stranded RNA and activation of NF-kappaB by Toll-like receptor 3. Nature 413:732-738.

Altenhoff AM, Studer RA, Robinson-Rechavi M, Dessimoz C. 2012. Resolving the ortholog conjecture: orthologs tend to be weakly, but significantly, more similar in function than paralogs. PLoS Computational Biology 8:e1002514.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. Journal of Molecular Biology 215:403-410.

Alvarez-Ponce D, Aguade M, Rozas J. 2009. Network-level molecular evolutionary analysis of the insulin/TOR signal transduction pathway across 12 Drosophila genomes. Genome Research 19:234-242.

Andersen-Nissen E, Smith KD, Bonneau R, Strong RK, Aderem A. 2007. A conserved surface on Toll-like receptor 5 recognizes bacterial flagellin. The Journal of Experimental Medicine 204:393-403.

Andolfatto P. 2001. Adaptive hitchhiking effects on genome variability. Current Opinion in Genetics & Development 11:635-641.

Andreeva A, Howorth D, Chandonia JM, Brenner SE, Hubbard TJ, Chothia C, Murzin AG. 2008. Data growth and its impact on the SCOP database: new developments. Nucleic Acids Research 36:D419-425.

Anglesio MS, Evdokimova V, Melnyk N, Zhang L, Fernandez CV, Grundy PE, Leach S, Marra MA, Brooks-Wilson AR, Penninger J, *et al.* 2004. Differential expression of a novel ankyrin containing E3 ubiquitin-protein ligase, Hace1, in sporadic Wilms' tumor versus normal kidney. Human Molecular Genetics

13:2061-2074.

Anisimova M, Bielawski JP, Yang Z. 2002. Accuracy and power of bayes prediction of amino acid sites under positive selection. Molecular Biology and Evolution 19:950-958.

Anisimova M, Nielsen R, Yang Z. 2003. Effect of recombination on the accuracy of the likelihood method for detecting positive selection at amino acid sites. Genetics 164:1229-1236.

Apic G, Gough J, Teichmann SA. 2001. Domain combinations in archaeal, eubacterial and eukaryotic proteomes. Journal of Molecular Biology 310:311-325.

Ariffin JK, Sweet MJ. 2013. Differences in the repertoire, regulation and function of Toll-like Receptors and inflammasome-forming Nod-like Receptors between human and mouse. Current Opinion in Microbiology 16:303-310.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, *et al.* 2000. Gene Ontology: tool for the unification of biology. Nature Genetics 25:25-29.

Attarwala H. 2010. TGN1412: From Discovery to Disaster. Journal of Young Pharmacists 2:332-336.

Bapteste E, Lopez P, Bouchard F, Baquero F, McInerney JO, Burian RM. 2012. Evolutionary analyses of non-genealogical bonds produced by introgressive descent. Proceedings of the National Academy of Sciences of the United States of America 109:18266-18272.

Bapteste E, van Iersel L, Janke A, Kelchner S, Kelk S, McInerney JO, Morrison DA, Nakhleh L, Steel M, Stougie L, *et al.* 2013. Networks: expanding evolutionary thinking. Trends in Genetics 29:439-441.

Barabasi AL, Oltvai ZN. 2004. Network biology: understanding the cell's functional organization. Nature Reviews Genetics 5:101-113.

Bashton M, Chothia C. 2007. The generation of new protein functions by the combination of domains. Structure 15:85-99.

Bauer S, Kirschning CJ, Hacker H, Redecke V, Hausmann S, Akira S, Wagner H, Lipford GB. 2001. Human TLR9 confers responsiveness to bacterial DNA via species-specific CpG motif recognition. Proceedings of the National Academy of Sciences of the United States of America 98:9237-9242.

Begun DJ, Holloway AK, Stevens K, Hillier LW, Poh YP, Hahn MW, Nista PM, Jones CD, Kern AD, Dewey CN, *et al.* 2007. Population genomics: whole-genome analysis of polymorphism and divergence in Drosophila simulans. PLoS Biology 5:e310.

Berry A, Pogorelcnik R, Simonet G. 2010. An Introduction to Clique Minimal Separator Decomposition. Algorithms 3:197–215.

Blackburne BP, Whelan S. 2012. Measuring the distance between multiple sequence alignments. Bioinformatics 28:495-502.

Blanc G, Wolfe KH. 2004. Functional divergence of duplicated genes formed by polyploidy during Arabidopsis evolution. Plant Cell 16:1679-1691.

Boni MF, Posada D, Feldman MW. 2007. An exact nonparametric method for inferring mosaic structure in sequence triplets. Genetics 176:1035-1047.

Borgatti SP. 2005. Centrality and network flow. Social Networks 27:55-71.

Borgatti SP, Everett MG. 2006. A graph-theoretic perspective on centrality. Social Networks 28:466-484.

Brannen CL, Sodetz JM. 2007. Incorporation of human complement C8 into the membrane attack complex is mediated by a binding site located within the C8beta MACPF domain. Molecular Immunology 44:960-965.

Bridgham JT, Carroll SM, Thornton JW. 2006. Evolution of hormone-receptor complexity by molecular exploitation. Science 312:97-101.

Brown CA, Murray AW, Verstrepen KJ. 2010. Rapid Expansion and Functional Divergence of Subtelomeric Gene Families in Yeasts. Current Biology 20:895-903.

Butenko S, Wilhelm WE. 2006. Clique-detection models in computational biochemistry and genomics. European Journal of Operational Research 173:1-17.

Caamano J, Hunter CA. 2002. NF-kappaB family of transcription factors: central regulators of innate and adaptive immune functions. Clinical Microbiology Reviews 15:414-429.

Caceres EF, Hurst LD. 2013. The evolution, impact and properties of exonic splice enhancers. Genome Biology 14:R143.

Caldon CE, Yoong P, March PE. 2001. Evolution of a molecular switch: universal bacterial GTPases regulate ribosome function. Molecular Microbiology 41:289-297.

Capella-Gutierrez S, Silla-Martinez JM, Gabaldon T. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25:1972-1973.

Carbon S, Ireland A, Mungall CJ, Shu S, Marshall B, Lewis S. 2009. AmiGO: online access to ontology and annotation data. Bioinformatics 25:288-289.

Caulin AF, Maley CC. 2011. Peto's Paradox: evolution's prescription for cancer prevention. Trends in Ecology & Evolution 26:175-182.

Cervantes JL, Weinerman B, Basole C, Salazar JC. 2012. TLR8: the forgotten relative revindicated. Cell Molecular Immunology 9:434-438.

Charlesworth J, Eyre-Walker A. 2006. The rate of adaptive evolution in enteric bacteria. Molecular Biology and Evolution 23:1348-1356.

Chen R, Jeong SS. 2000. Functional prediction: identification of protein orthologs and paralogs. Protein Science 9:2344-2353.

Chun S, Fay JC. 2011. Evidence for hitchhiking of deleterious mutations within the human genome. PLoS Genetics 7:e1002240.

Coats SR, Do CT, Karimi-Naser LM, Braham PH, Darveau RP. 2007. Antagonistic lipopolysaccharides block E-coli lipopolysaccharide function at human TLR4 via interaction with the human MD-2 lipopolysaccharide binding site. Cellular Microbiology 9:1191-1202.

Comeron JM. 1999. K-Estimator: calculation of the number of nucleotide substitutions per site and the confidence intervals. Bioinformatics 15:763-764.

Conant GC, Wolfe KH. 2008. Turning a hobby into a job: How duplicated genes find new functions. Nature Reviews Genetics 9:938-950.

Consortium TU. 2014. Activities at the Universal Protein Resource (UniProt). Nucleic Acids Research 42:D191-198.

Coulson AF, Moult J. 2002. A unifold, mesofold, and superfold model of protein fold use. Proteins 46:61-71.

Creevey CJ, McInerney JO. 2002. An algorithm for detecting directional and non-directional positive selection, neutrality and negative selection in protein coding DNA sequences. Gene 300:43-51.

Creevey CJ, McInerney JO. 2003. CRANN: detecting adaptive evolution in protein-coding DNA sequences. Bioinformatics 19:1726.

Darriba D, Taboada GL, Doallo R, Posada D. 2011. ProtTest 3: fast selection of best-fit models of protein evolution. Bioinformatics 27:1164-1165.

Darwin C. 1859. The Origin of Species. London: John Murray.

Davis AE, 3rd, Mejia P, Lu F. 2008. Biological activities of C1 inhibitor. Molecular Immunology 45:4057-4063.

Davis MM. 2008. A Prescription for Human Immunology. Immunity 29:835-838.

Dayhoff MO, Schwartz RM, Orcutt BC. 1978. A model of evolutionary change in proteins. Dayhoff, M.O. (Ed.). Atlas of Protein Sequence and Structure. Washington, DC: National Biomedical Research Foundation. p. 345–352.

de Magalhaes JP. 2013. How ageing processes influence cancer. Nature Reviews Cancer 13:357-365.

Delport W, Poon AF, Frost SD, Kosakovsky Pond SL. 2010. Datamonkey 2010: a suite of phylogenetic analysis tools for evolutionary biology. Bioinformatics 26:2455-

2457.

Dennis MY, Nuttle X, Sudmant PH, Antonacci F, Graves TA, Nefedov M, Rosenfeld JA, Sajjadian S, Malig M, Kotkiewicz H, *et al.* 2012. Evolution of Human-Specific Neural SRGAP2 Genes by Incomplete Segmental Duplication. Cell 149:912-922.

Dessimoz C, Gabaldon T, Roos DS, Sonnhammer ELL, Herrero J, Consortium QO. 2012. Toward community standards in the quest for orthologs. Bioinformatics 28:900-904.

Deutsch M, Long M. 1999. Intron-exon structures of eukaryotic model organisms. Nucleic Acids Research 27:3219-3228.

Dumas L, Kim YH, Karimpour-Fard A, Cox M, Hopkins J, Pollack JR, Sikela JM. 2007. Gene copy number variation spanning 60 million years of human and primate evolution. Genome Research 17:1266-1277.

HMMER User's. 2010. Available from: ftp://selab.janelia.org/pub/software/hmmer3/3.1b1/Userguide.pdf

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

Edgar RC. 2004. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics 5:113.

Ehrnthaller C, Ignatius A, Gebhard F, Huber-Lang M. 2011. New insights of an old defense system: structure, function, and clinical relevance of the complement system. Molecular Medicine 17:317-329.

Fairbrother WG, Yeh RF, Sharp PA, Burge CB. 2002. Predictive identification of exonic splicing enhancers in human genes. Science 297:1007-1013.

Fares MA. 2004. SWAPSC: sliding window analysis procedure to detect selective constraints. Bioinformatics 20:2867-2868.

Farhat MR, Shapiro BJ, Kieser KJ, Sultana R, Jacobson KR, Victor TC, Warren RM, Streicher EM, Calver A, Sloutsky A, *et al.* 2013. Genomic analysis identifies targets of convergent positive selection in drug-resistant Mycobacterium tuberculosis. Nature Genetics 45:1183-1189.

Fay JC, Wu CI. 2000. Hitchhiking under positive Darwinian selection. Genetics 155:1405-1413.

Felsenstein J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. Journal of Molecular Evolution 17:368-376.

Feng SH, Jacobsen SE. 2011. Epigenetic modifications in plants: an evolutionary perspective. Current Opinion in Plant Biology 14:179-186.

Filippakopoulos P, Kofler M, Hantschel O, Gish GD, Grebien F, Salah E, Neudecker P, Kay LE, Turk BE, Superti-Furga G, *et al.* 2008. Structural coupling of SH2-

kinase domains links Fes and Abl substrate recognition and kinase activation. Cell 134:793-803.

Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, Heger A, Hetherington K, Holm L, Mistry J, *et al.* 2014. Pfam: the protein families database. Nucleic Acids Research 42:D222-230.

Fitch WM. 1970. Distinguishing homologous from analogous proteins. Systematic Zoology 19:99-113.

Flajnik MF, Kasahara M. 2010. Origin and evolution of the adaptive immune system: genetic events and selective pressures. Nature Reviews Genetics 11:47-59.

Fletcher W, Yang Z. 2010. The effect of insertions, deletions, and alignment errors on the branch-site test of positive selection. Molecular Biology and Evolution 27:2257-2267.

Fletcher W, Yang ZH. 2010. The Effect of Insertions, Deletions, and Alignment Errors on the Branch-Site Test of Positive Selection. Molecular Biology and Evolution 27:2257-2267.

Flicek P, Amode MR, Barrell D, Beal K, Billis K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fitzgerald S, *et al.* 2014. Ensembl 2014. Nucleic Acids Research 42:D749-755.

Force A, Lynch M, Pickett FB, Amores A, Yan YL, Postlethwait J. 1999. Preservation of duplicate genes by complementary, degenerative mutations. Genetics 151:1531-1545.

Forsbach A, Nemorin JG, Montino C, Muller C, Samulowitz U, Vicari AP, Jurk M, Mutwiri GK, Krieg AM, Lipford GB, *et al.* 2008. Identification of RNA sequence motifs stimulating sequence-specific TLR8-dependent immune responses. The Journal of Immunology 180:3729-3738.

Forslund K, Pekkari I, Sonnhammer EL. 2011. Domain architecture conservation in orthologs. BMC Bioinformatics 12:326.

Fortier ME, Kent S, Ashdown H, Poole S, Boksa P, Luheshi GN. 2004. The viral mimic, polyinosinic:polycytidylic acid, induces fever in rats via an interleukin-1-dependent mechanism. The American Journal of Physiology - Regulatory, Integrative and Comparative Physiology 287:R759-766.

Foster JG, Foster DV, Grassberger P, Paczuski M. 2010. Edge direction and the structure of networks. Proceedings of the National Academy of Sciences of the United States of America 107:10815-10820.

Foster PG. 2004. Modeling compositional heterogeneity. Systematic Biology 53:485-495.

Fourment M, Gillings MR. 2008. A comparison of common programming languages used in bioinformatics. BMC Bioinformatics 9.

Franceschini A, Szklarczyk D, Frankild S, Kuhn M, Simonovic M, Roth A, Lin J, Minguez P, Bork P, von Mering C, *et al.* 2013. STRING v9.1: protein-protein interaction networks, with increased coverage and integration. Nucleic Acids Research 41:D808-815.

Freeman LC. 1979. Centrality in Social Networks Conceptual Clarification. Social Networks 1:215-239.

Gabaldon T, Dessimoz C, Huxley-Jones J, Vilella AJ, Sonnhammer EL, Lewis S. 2009. Joining forces in the quest for orthologs. Genome Biology 10:403.

Gal P, Dobo J, Zavodszky P, Sim RB. 2009. Early complement proteases: C1r, C1s and MASPs. A structural insight into activation and functions. Molecular Immunology 46:2745-2752.

Galtier N, Duret L. 2007. Adaptation or biased gene conversion? Extending the null hypothesis of molecular evolution. Trends in Genetics 23:273-277.

Garcia S, Freitas AA. 2012. Humanized mice: current states and perspectives. Immunology Letters 146:1-7.

George JC, Bada J, Zeh J, Scott L, Brown SE, O'Hara T, Suydam R. 1999. Age and growth estimates of bowhead whales (Balaena mysticetus) via aspartic acid racemization. Canadian Journal of Zoology-Revue Canadienne De Zoologie 77:571-580.

Gharib WH, Robinson-Rechavi M. 2013. The Branch-Site Test of Positive Selection Is Surprisingly Robust but Lacks Power under Synonymous Substitution Saturation and Variation in GC. Molecular Biology and Evolution 30:1675-1686.

Gharib WH, Robinson-Rechavi M. 2011. When orthologs diverge between human and mouse. Briefings in Bioinformatics 12:436-441.

Gibbs MJ, Armstrong JS, Gibbs AJ. 2000. Sister-scanning: a Monte Carlo procedure for assessing signals in recombinant sequences. Bioinformatics 16:573-582.

Gillespie JH. 1998. Population genetics: a concise guide. Baltimore, Md: The Johns Hopkins University Press.

Girvan M, Newman ME. 2002. Community structure in social and biological networks. Proceedings of the National Academy of Sciences of the United States of America 99:7821-7826.

Goh KI, Cusick ME, Valle D, Childs B, Vidal M, Barabasi AL. 2007. The human disease network. Proceedings of the National Academy of Sciences of the United States of America 104:8685-8690.

Grabiec A, Meng GX, Fichte S, Bessler W, Wagner H, Kirschning CJ. 2004. Human but not murine Toll-like receptor 2 discriminates between tri-palmitoylated and tri-lauroylated peptides. Journal of Biological Chemistry 279:48004-48012.

Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, Patterson N, Li H,

Zhai W, Fritz MH, *et al.* 2010. A draft sequence of the Neandertal genome. Science 328:710-722.

Guerois R, Nielsen JE, Serrano L. 2002. Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. Journal of Molecular Biology 320:369-387.

Hadders MA, Bubeck D, Roversi P, Hakobyan S, Forneris F, Morgan BP, Pangburn MK, Llorca O, Lea SM, Gros P. 2012. Assembly and regulation of the membrane attack complex based on structures of C5b6 and sC5b9. Cell Reports 1:200-207.

Haggerty LS, Jachiet PA, Hanage WP, Fitzpatrick DA, Lopez P, O'Connell MJ, Pisani D, Wilkinson M, Bapteste E, McInerney JO. 2014. A Pluralistic Account of Homology: Adapting the Models to the Data. Molecular Biology and Evolution 31:501-516.

Hahn MW. 2008. Toward a selection theory of molecular evolution. Evolution 62:255-265.

Halary S, Leigh JW, Cheaib B, Lopez P, Bapteste E. 2010. Network analyses structure genetic diversity in independent genetic worlds. Proceedings of the National Academy of Sciences of the United States of America 107:127-132.

Haldane JB. 1927. A Mathematical Theory of Natural and Artificial Selection. V. Selection and mutation. Proceedings of the Cambridge Philosophical Society 23:838-844.

Hamblin MT, Thompson EE, Di Rienzo A. 2002. Complex signatures of natural selection at the Duffy blood group locus. The American Journal of Human Genetics 70:369-383.

Hanson-Smith V, Kolaczkowski B, Thornton JW. 2010. Robustness of Ancestral Sequence Reconstruction to Phylogenetic Uncertainty. Molecular Biology and Evolution 27:1988-1999.

Harms MJ, Thornton JW. 2010. Analyzing protein structure and function using ancestral gene reconstruction. Current Opinion in Structural Biology Biology 20:360-366.

Harris H. 1966. Enzyme polymorphisms in man. Proceedings of the Royal Society of London. Series B 164:298-310.

Hasan U, Chaffois C, Gaillard C, Saulnier V, Merck E, Tancredi S, Guiet C, Briere F, Vlach J, Lebecque S, *et al.* 2005. Human TLR10 is a functional receptor, expressed by B cells and plasmacytoid dendritic cells, which activates gene transcription through MyD88. Journal of Immunology 174:2942-2950.

Hedges SB, Dudley J, Kumar S. 2006. TimeTree: a public knowledge-base of divergence times among organisms. Bioinformatics 22:2971-2972.

Henikoff S, Henikoff JG. 1992. Amino acid substitution matrices from protein blocks. Proceedings of the National Academy of Sciences of the United States of

America 89:10915-10919.

Hennig W. 1966. Phylogenetic systematics. Urbana: University of Illinois Press.

Henricson A, Forslund K, Sonnhammer EL. 2010. Orthology confers intron position conservation. BMC Genomics 11:412.

Hudson RR. 2002. Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics 18:337-338.

Huelsenbeck JP, Larget B, Miller RE, Ronquist F. 2002. Potential applications and pitfalls of Bayesian inference of phylogeny. Systematic Biology 51:673-688.

Huerta-Cepas J, Dopazo J, Huynen MA, Gabaldon T. 2011. Evidence for short-time divergence and long-time conservation of tissue-specific expression after gene duplication. Briefings in Bioinformatics 12:442-448.

Hughes AL. 1999. Adaptive evolution of genes and genomes. New York: Oxford University Press.

Hughes AL. 2007. Looking for Darwin in all the wrong places: the misguided quest for positive selection at the nucleotide sequence level. Heredity 99:364-373.

Hunter JD. 2007. Matplotlib: A 2D graphics environment. Computing in Science & Engineering 9:90-95.

Hurst LD. 2002. The Ka/Ks ratio: diagnosing the form of sequence evolution. Trends in Genetics 18:486.

Hurst LD, Pal C. 2001. Evidence for purifying selection acting on silent sites in BRCA1. Trends in Genetics 17:62-65.

Ito R, Takahashi T, Katano I, Ito M. 2012. Current advances in humanized mouse models. Cell Molecular Immunology 9:208-214.

Jachiet PA, Pogorelcnik R, Berry A, Lopez P, Bapteste E. 2013. MosaicFinder: identification of fused gene families in sequence similarity networks. Bioinformatics 29:837-844.

Janeway CA, Jr., Medzhitov R. 2002. Innate immune recognition. Annual Review of Immunology 20:197-216.

Jault C, Pichon L, Chluba J. 2004. Toll-like receptor gene family and TIR-domain adapters in Danio rerio. Molecular Immunology 40:759-771.

Jones DT, Taylor WR, Thornton JM. 1992. The rapid generation of mutation data matrices from protein sequences. Computer Applications in the Biosciences 8:275-282.

Jordan MS, Singer AL, Koretzky GA. 2003. Adaptors as central mediators of signal transduction in immune cells. Nature Immunology 4:110-116.

Jurk M, Heil F, Vollmer J, Schetter C, Krieg AM, Wagner H, Lipford G, Bauer S. 2002. Human TLR7 or TLR8 independently confer responsiveness to the antiviral compound R-848. Nature Immunology 3:499.

Kaluza P, Kolzsch A, Gastner MT, Blasius B. 2010. The complex network of global cargo ship movements. Journal of The Royal Society Interface 7:1093-1103.

Kanehisa M, Goto S. 2000. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Research 28:27-30.

Karp RM. 1972. Reducibility among Combinatorial Problems. Boston, MA: Springer US.

Katoh M. 2007. WNT signaling pathway and stem cell signaling network. Clinical Cancer Research 13:4042-4045.

Katzman S, Capra JA, Haussler D, Pollard KS. 2011. Ongoing GC-biased evolution is widespread in the human genome and enriched near recombination hot spots. Genome Biology Evolution 3:614-626.

Kawai T, Akira S. 2010. The role of pattern-recognition receptors in innate immunity: update on Toll-like receptors. Nature Immunology 11:373-384.

Kawai T, Akira S. 2007. TLR signaling. Seminars in Immunology 19:24-32.

Keane TM, Creevey CJ, Pentony MM, Naughton TJ, McInerney JO. 2006. Assessment of methods for amino acid matrix selection and their use on empirical data shows that ad hoc assumptions for choice of matrix are not justified. BMC Evolutionary Biology 6.

Keane TM, Goodstadt L, Danecek P, White MA, Wong K, Yalcin B, Heger A, Agam A, Slater G, Goodson M, *et al.* 2011. Mouse genomic variation and its effect on phenotypes and gene regulation. Nature 477:289-294.

Kersting AR, Bornberg-Bauer E, Moore AD, Grath S. 2012. Dynamics and adaptive benefits of protein domain emergence and arrangements during plant genome evolution. Genome Biology Evolution 4:316-329.

Kim EB, Fang XD, Fushan AA, Huang ZY, Lobanov AV, Han LJ, Marino SM, Sun XQ, Turanov AA, Yang PC, *et al.* 2011. Genome sequencing reveals insights into physiology and longevity of the naked mole rat. Nature 479:223-227.

Kimura M. 1968. Evolutionary Rate at Molecular Level. Nature 217:624-626.

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 (Oxford) 2011:bar030.

Kosakovsky Pond SL, Frost SD. 2005. Not so different after all: a comparison of methods for detecting amino acid sites under selection. Molecular Biology and Evolution 22:1208-1222.

Kosiol C, Vinar T, da Fonseca RR, Hubisz MJ, Bustamante CD, Nielsen R, Siepel A. 2008. Patterns of positive selection in six Mammalian genomes. PLoS Genetics 4:e1000144.

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, *et al.* 2001. Initial sequencing and analysis of the human genome. Nature 409:860-921.

Lee D, Redfern O, Orengo C. 2007. Predicting protein function from sequence and structure. Nature Reviews Molecular Cell Biology 8:995-1005.

Lee SMY, Kok KH, Jaume M, Cheung TKW, Yip TF, Lai JCC, Guan Y, Webster RG, Jin DY, Peiris JSM. 2014. Toll-like receptor 10 is involved in induction of innate immune responses to influenza virus infection. Proceedings of the National Academy of Sciences of the United States of America 111:3793-3798.

Lemaitre B, Nicolas E, Michaut L, Reichhart JM, Hoffmann JA. 1996. The dorsoventral regulatory gene cassette spatzle/Toll/cactus controls the potent antifungal response in Drosophila adults. Cell 86:973-983.

Leslie M. 2014. Biomedical research. Inflammation debate reignites. Science 345:607.

Leulier F, Lemaitre B. 2008. Toll-like receptors - taking an evolutionary approach. Nature Reviews Genetics 9:165-178.

Levasseur A, Gouret P, Lesage-Meessen L, Asther M, Asther M, Record E, Pontarotti P. 2006. Tracking the connection between evolutionary and functional shifts using the fungal lipase/feruloyl esterase A family. BMC Evolutionary Biology 6.

Lewontin RC, Hubby JL. 1966. A molecular approach to the study of genic heterozygosity in natural populations. II. Amount of variation and degree of heterozygosity in natural populations of Drosophila pseudoobscura. Genetics 54:595-609.

Li WH, Ellsworth DL, Krushkal J, Chang BH, Hewett-Emmett D. 1996. Rates of nucleotide substitution in primates and rodents and the generation-time effect hypothesis. Molecular Phylogenetics and Evolution 5:182-187.

Li WH, Wu CI, Luo CC. 1985. A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. Molecular Biology and Evolution 2:150-174.

Liang H, Zhou W, Landweber LF. 2006. SWAKK: a web server for detecting positive selection in proteins using a sliding window substitution rate analysis. Nucleic Acids Research 34:W382-384.

Librado P, Rozas J. 2009. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. Bioinformatics 25:1451-1452.

Lipman DJ, Pearson WR. 1985. Rapid and sensitive protein similarity searches. Science 227:1435-1441.

Liu J, Xu C, Hsu LC, Luo Y, Xiang R, Chuang TH. 2010. A five-amino-acid motif in the undefined region of the TLR8 ectodomain is required for species-specific ligand recognition. Molecular Immunology 47:1083-1090.

Loughran NB, Hinde S, McCormick-Hill S, Leidal KG, Bloomberg S, Loughran ST, O'Connor B, O'Fagain C, Nauseef WM, O'Connell MJ. 2012. Functional consequence of positive selection revealed through rational mutagenesis of human myeloperoxidase. Molecular Biology and Evolution 29:2039-2046.

Loughran NB, O'Connor B, O'Fagain C, O'Connell MJ. 2008. The phylogeny of the mammalian heme peroxidases and the evolution of their diverse functions. BMC Evolution Biology 8.

Loytynoja A, Goldman N. 2005. An algorithm for progressive multiple alignment of sequences with insertions. Proceedings of the National Academy of Sciences of the United States of America 102:10557-10562.

Loytynoja A, Goldman N. 2008. Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. Science 320:1632-1635.

Loytynoja A, Goldman N. 2010. webPRANK: a phylogeny-aware multiple sequence aligner with interactive alignment browser. BMC Bioinformatics 11.

Luce RD, Perry AD. 1949. A method of matrix analysis of group structure. Psychometrika 14:95-116.

Lundberg AM, Drexler SK, Monaco C, Williams LM, Sacre SM, Feldmann M, Foxwell BM. 2007. Key differences in TLR3/poly I:C signaling and cytokine induction by human primary cells: a phenomenon absent from murine cell systems. Blood 110:3245-3252.

Lynch M. 2007. The frailty of adaptive hypotheses for the origins of organismal complexity. Proceedings of the National Academy of Sciences of the United States of America 104 (Suppl 1):8597-8604.

Lynn DJ, Winsor GL, Chan C, Richard N, Laird MR, Barsky A, Gardy JL, Roche FM, Chan TH, Shah N, *et al.* 2008. InnateDB: facilitating systems-level analyses of the mammalian innate immune response. Molecular Systematic Biology 4:218.

Lyons RJ, Deane R, Lynch DK, Ye ZS, Sanderson GM, Eyre HJ, Sutherland GR, Daly RJ. 2001. Identification of a novel human tankyrase through its interaction with the adaptor protein Grb14. Journal of Biological Chemistry 276:17172-17180.

Manning G, Whyte DB, Martinez R, Hunter T, Sudarsanam S. 2002. The protein kinase complement of the human genome. Science 298:1912-1934.

Marais G, Mouchiroud D, Duret L. 2001. Does recombination improve selection on codon usage? Lessons from nematode and fly complete genomes. Proceedings of the National Academy of Sciences of the United States of America 98:5688-5692.

Martin D, Rybicki E. 2000. RDP: detection of recombination amongst aligned

sequences. Bioinformatics 16:562-563.

Martin DP, Lemey P, Lott M, Moulton V, Posada D, Lefeuvre P. 2010. RDP3: a flexible and fast computer program for analyzing recombination. Bioinformatics 26:2462-2463.

Martin DP, Posada D, Crandall KA, Williamson C. 2005. A modified bootscan algorithm for automated identification of recombinant sequences and recombination breakpoints. AIDS Research and Human Retroviruses 21:98-102.

Massingham T, Goldman N. 2005. Detecting amino acid sites under positive selection and purifying selection. Genetics 169:1753-1762.

Matsumoto M, Funami K, Tanabe M, Oshiumi H, Shingai M, Seto Y, Yamamoto A, Seya T. 2003. Subcellular localization of Toll-like receptor 3 in human dendritic cells. The Journal of Immunology 171:3154-3162.

Matsumoto M, Oshiumi H, Seya T. 2011. Antiviral responses induced by the TLR3 pathway. Reviews in Medical Virology 21:67–77.

McCormack JE, Faircloth BC, Crawford NG, Gowaty PA, Brumfield RT, Glenn TC. 2012. Ultraconserved elements are novel phylogenomic markers that resolve placental mammal phylogeny when combined with species-tree analysis. Genome Research 22:746-754.

McDonald JH, Kreitman M. 1991. Adaptive protein evolution at the Adh locus in Drosophila. Nature 351:652-654.

McLean CY, Reno PL, Pollen AA, Bassan AI, Capellini TD, Guenther C, Indjeian VB, Lim X, Menke DB, Schaar BT, *et al.* 2011. Human-specific loss of regulatory DNA and the evolution of human-specific traits. Nature 471:216-219.

Medzhitov R, Janeway C, Jr. 2000. Innate immunity. The New England Journal of Medicine 343:338-344.

Medzhitov R, Preston-Hurlburt P, Janeway CA, Jr. 1997. A human homologue of the Drosophila Toll protein signals activation of adaptive immunity. Nature 388:394-397.

Meng J, Drolet JR, Monks BG, Golenbock DT. 2010. MD-2 residues tyrosine 42, arginine 69, aspartic acid 122, and leucine 125 provide species specificity for lipid IVA. Journal of Biological Chemistry 285:27935-27943.

Mestas J, Hughes CC. 2004. Of mice and not men: differences between mouse and human immunology. The Journal of Immunology 172:2731-2738.

Mestas J, Hughes CCW. 2004. Of mice and not men: Differences between mouse and human immunology. Journal of Immunology 172:2731-2738.

Mogensen TH. 2009. Pathogen recognition and inflammatory signaling in innate immune defenses. Clinical Microbiology Reviews 22:240-273.

Molero C, Petrenyi K, Gonzalez A, Carmona M, Gelis S, Abrie JA, Strauss E, Ramos J, Dombradi V, Hidalgo E, *et al.* 2013. The Schizosaccharomyces pombe fusion gene hal3 encodes three distinct activities. Molecular Microbiology 90:367-382.

Moore AD, Bjorklund AK, Ekman D, Bornberg-Bauer E, Elofsson A. 2008. Arrangements in the modular evolution of proteins. Trends Biochem Sci 33:444-451.

Moresco EM, LaVine D, Beutler B. 2011. Toll-like receptors. Current Biology 21:R488-493.

Morgan CC, Creevey CJ, O'Connell MJ. 2014. Mitochondrial data are not suitable for resolving placental mammal phylogeny. Mammalian Genome 25:636-647.

Morgan CC, Foster PG, Webb AE, Pisani D, McInerney JO, O'Connell MJ. 2013. Heterogeneous models place the root of the placental mammal phylogeny. Molecular Biology and Evolution 30:2145-2156.

Morgan CC, Loughran NB, Walsh TA, Harrison AJ, O'Connell MJ. 2010. Positive selection neighboring functionally essential sites and disease-implicated regions of mammalian reproductive proteins. BMC Evolution Biology 10:39.

Mosavi LK, Cammett TJ, Desrosiers DC, Peng ZY. 2004. The ankyrin repeat as molecular architecture for protein recognition. Protein Science 13:1435-1448.

Moury B, Simon V. 2011. dN/dS-Based Methods Detect Positive Selection Linked to Trade-Offs between Different Fitness Traits in the Coat Protein of Potato virus Y. Molecular Biology and Evolution 28:2707-2717.

Muller J, Creevey CJ, Thompson JD, Arendt D, Bork P. 2010. AQUA: automated quality improvement for multiple sequence alignments. Bioinformatics 26:263-265.

Muller T, Vingron M. 2000. Modeling amino acid replacement. Journal of Computational Biology 7:761-776.

Murray SA, Mihali TK, Neilan BA. 2011. Extraordinary conservation, gene loss, and positive selection in the evolution of an ancient neurotoxin. Molecular Biology and Evolution 28:1173-1182.

Nehrt NL, Clark WT, Radivojac P, Hahn MW. 2011. Testing the Ortholog Conjecture with Comparative Functional Genomic Data from Mammals. Plos Computational Biology 7.

Nei M, Li WH. 1979. Mathematical model for studying genetic variation in terms of restriction endonucleases. Proceedings of the National Academy of Sciences of the United States of America 76:5269-5273.

Nevin Gerek Z, Kumar S, Banu Ozkan S. 2013. Structural dynamics flexibility informs function and evolution at a proteome scale. Evolutionary Applications 6:423-433.

Newman ME. 2002. Assortative mixing in networks. Physical Review Letters 89:208701.

Newman ME. 2003. Mixing patterns in networks. Physical Review E: Statistical, Nonlinear, and Soft Matter Physics 67:026126.

Newman MEJ. 2002. Assortative mixing in networks. Physical Review Letters 89.

Newman MEJ. 2003. Mixing patterns in networks. Physical Review E 67.

Newman MEJ. 2003. The structure and function of complex networks. Siam Review 45:167-256.

Nielsen R, Yang Z. 1998. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. Genetics 148:929-936.

Ohno S. 1970. Evolution by gene duplication. London, New York,: Allen & Unwin; Springer-Verlag.

Ohta T. 1973. Slightly deleterious mutant substitutions in evolution. Nature 246:96-98.

Ohta T, Gillespie JH. 1996. Development of neutral and nearly neutral theories. Theoretical Population Biology 49:128-142.

Opsahl T, Agneessens F, Skvoretz J. 2010. Node centrality in weighted networks: Generalizing degree and shortest paths. Social Networks 32:245-251.

Oshiumi H, Matsumoto M, Funami K, Akazawa T, Seya T. 2003. TICAM-1, an adaptor molecule that participates in Toll-like receptor 3-mediated interferon-beta induction. Nature Immunology 4:161-167.

Padidam M, Sawyer S, Fauquet CM. 1999. Possible emergence of new geminiviruses by frequent recombination. Virology 265:218-225.

Parmley JL, Chamary JV, Hurst LD. 2006. Evidence for purifying selection against synonymous mutations in mammalian exonic splicing enhancers. Molecular Biology and Evolution 23:301-309.

Parmley JL, Urrutia AO, Potrzebowski L, Kaessmann H, Hurst LD. 2007. Splicing and the evolution of proteins in mammals. PLoS Biology 5:e14.

Pastor-Satorras R, Vespignani A. 2001. Epidemic spreading in scale-free networks. Physical Review Letters 86:3200-3203.

Pedersen JS, Bejerano G, Siepel A, Rosenbloom K, Lindblad-Toh K, Lander ES, Kent J, Miller W, Haussler D. 2006. Identification and classification of conserved RNA secondary structures in the human genome. Plos Computational Biology 2:251-262.

Perry GH, Yang F, Marques-Bonet T, Murphy C, Fitzgerald T, Lee AS, Hyland C, Stone AC, Hurles ME, Tyler-Smith C, *et al.* 2008. Copy number variation and evolution in humans and chimpanzees. Genome Research 18:1698-1710.

Pertel T, Hausmann S, Morger D, Zuger S, Guerra J, Lascano J, Reinhard C, Santoni FA, Uchil PD, Chatel L, *et al.* 2011. TRIM5 is an innate immune sensor for the retrovirus capsid lattice. Nature 472:361-365.

Peterson ME, Chen F, Saven JG, Roos DS, Babbitt PC, Sali A. 2009. Evolutionary constraints on structural similarity in orthologs and paralogs. Protein Science 18:1306-1315.

Pierce KL, Premont RT, Lefkowitz RJ. 2002. Seven-transmembrane receptors. Nature Reviews Molecular Cell Biology 3:639-650.

Pinhasi R, Higham TF, Golovanova LV, Doronichev VB. 2011. Revised age of late Neanderthal occupation and the end of the Middle Paleolithic in the northern Caucasus. Proceedings of the National Academy of Sciences of the United States of America 108:8611-8616.

Pond SL, Frost SD, Muse SV. 2005. HyPhy: hypothesis testing using phylogenies. Bioinformatics 21:676-679.

Pontarollo RA, Rankin R, Babiuk LA, Godson DL, Griebel PJ, Hecker R, Krieg AM, van Drunen Littel-van den Hurk S. 2002. Monocytes are required for optimum in vitro stimulation of bovine peripheral blood mononuclear cells by non-methylated CpG motifs. Veterinary Immunology and Immunopathology 84:43-59.

Posada D, Crandall KA. 2002. The effect of recombination on the accuracy of phylogeny estimation. Journal of Molecular Evolution 54:396-402.

Posada D, Crandall KA. 2001. Evaluation of methods for detecting recombination from DNA sequences: computer simulations. Proceedings of the National Academy of Sciences of the United States of America 98:13757-13762.

Prechelt L. 2000. An empirical comparison of seven programming languages. Computer 33:23-29.

Prud'homme B, Gompel N, Rokas A, Kassner VA, Williams TM, Yeh SD, True JR, Carroll SB. 2006. Repeated morphological evolution through cis-regulatory changes in a pleiotropic gene. Nature 440:1050-1053.

Pruitt KD, Brown GR, Hiatt SM, Thibaud-Nissen F, Astashyn A, Ermolaeva O, Farrell CM, Hart J, Landrum MJ, McGarvey KM, *et al.* 2014. RefSeq: an update on mammalian reference sequences. Nucleic Acids Research 42:D756-763.

Pumidonming W, Walochnik J, Dauber E, Petry F. 2011. Binding to complement factors and activation of the alternative pathway by Acanthamoeba. Immunobiology 216:225-233.

Purcell MK, Smith KD, Hood L, Winton JR, Roach JC. 2006. Conservation of Toll-Like Receptor Signaling Pathways in Teleost Fish. Comparative Biochemistry and Physiology - Part D: Genomics and Proteomics 1:77-88.

Raetz CR, Whitfield C. 2002. Lipopolysaccharide endotoxins. Annual Review of

Biochemistry 71:635-700.

Rambaut A, Suchard M, Xie D, Drummond A. 2014. Tracer v1.6, Available from http://beast.bio.ed.ac.uk/Tracer.

Ramsay H, Rieseberg LH, Ritland K. 2009. The correlation of evolutionary rate with pathway position in plant terpenoid biosynthesis. Molecular Biology and Evolution 26:1045-1053.

Rankin R, Pontarollo R, Ioannou X, Krieg AM, Hecker R, Babiuk LA, van Drunen Littel-van den Hurk S. 2001. CpG motif identification for veterinary and laboratory species demonstrates that sequence recognition is highly conserved. Antisense and Nucleic Acid Drug Development 11:333-340.

Ratnakumar A, Mousset S, Glemin S, Berglund J, Galtier N, Duret L, Webster MT. 2010. Detecting positive selection within genomes: the problem of biased gene conversion. Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences 365:2571-2580.

Reeves JH. 1992. Heterogeneity in the substitution process of amino acid sites of proteins coded for by mitochondrial DNA. Journal of Molecular Evolution 35:17-31.

Rehli M. 2002. Of mice and men: species variations of Toll-like receptor expression. Trends Immunology 23:375-378.

Rivera CG, Vakil R, Bader JS. 2010. NeMo: Network Module identification in Cytoscape. BMC Bioinformatics 11 (Suppl 1):S61.

Roach JC, Glusman G, Rowen L, Kaur A, Purcell MK, Smith KD, Hood LE, Aderem A. 2005. The evolution of vertebrate Toll-like receptors. Proceedings of the National Academy of Sciences of the United States of America 102:9577-9582.

Rogers RL, Bedford T, Lyons AM, Hartl DL. 2010. Adaptive impact of the chimeric gene Quetzalcoatl in Drosophila melanogaster. Proceedings of the National Academy of Sciences of the United States of America 107:10943-10948.

Romiguier J, Ranwez V, Delsuc F, Galtier N, Douzery EJ. 2013. Less is more in mammalian phylogenomics: AT-rich genes minimize tree conflicts and unravel the root of placental mammals. Molecular Biology and Evolution 30:2134-2144.

Ronquist F. 2011. Draft MrBayes version 3.2 Manual: Tutorials and Model Summaries [Internet]. Available from: http://mrbayes.sourceforge.net/mb3.2_manual.pdf

Ronquist F, Huelsenbeck JP. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19:1572-1574.

Sabeti PC, Schaffner SF, Fry B, Lohmueller J, Varilly P, Shamovsky O, Palma A, Mikkelsen TS, Altshuler D, Lander ES. 2006. Positive natural selection in the human lineage. Science 312:1614-1620.

Sarma JV, Ward PA. 2011. The complement system. Cell Tissue Research 343:227-235.

Sawyer SL, Wu LI, Emerman M, Malik HS. 2005. Positive selection of primate TRIM5alpha identifies a critical species-specific retroviral restriction domain. Proceedings of the National Academy of Sciences of the United States of America 102:2832-2837.

Scally A, Durbin R. 2012. Revising the human mutation rate: implications for understanding human evolution. Nature Reviews Genetics 13:745-753.

Scheeff ED, Bourne PE. 2005. Structural evolution of the protein kinase-like superfamily. PLoS Computational Biology 1:e49.

Schmid K, Yang Z. 2008. The trouble with sliding windows and the selective pressure in BRCA1. PLoS One 3:e3746.

Schmidt CQ, Herbert AP, Kavanagh D, Gandy C, Fenton CJ, Blaum BS, Lyon M, Uhrin D, Barlow PN. 2008. A new map of glycosaminoglycan and C3b binding sites on factor H. The Journal of Immunology 181:2610-2619.

Schmidt M, Raghavan B, Muller V, Vogl T, Fejer G, Tchaptchet S, Keck S, Kalis C, Nielsen PJ, Galanos C, *et al.* 2010. Crucial role for human Toll-like receptor 4 in the development of contact allergy to nickel. Nature Immunology 11:814-819.

Schneider A, Souvorov A, Sabath N, Landan G, Gonnet GH, Graur D. 2009. Estimates of positive Darwinian selection are inflated by errors in sequencing, annotation, and alignment. Genome Biology Evolution 1:114-118.

Schwarz G. 1978. Estimating Dimension of a Model. Annals of Statistics 6:461-464.

Schymkowitz JW, Rousseau F, Martins IC, Ferkinghoff-Borg J, Stricher F, Serrano L. 2005. Prediction of water and metal binding sites and their affinities by using the Fold-X force field. Proceedings of the National Academy of Sciences of the United States of America 102:10147-10152.

Seok J, Warren HS, Cuenca AG, Mindrinos MN, Baker HV, Xu WH, Richards DR, McDonald-Smith GP, Gao H, Hennessy L, *et al.* 2013. Genomic responses in mouse models poorly mimic human inflammatory diseases. Proceedings of the National Academy of Sciences of the United States of America 110:3507-3512.

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 Research 13:2498-2504.

Shi Y, Yokoyama S. 2003. Molecular analysis of the evolutionary significance of ultraviolet vision in vertebrates. Proceedings of the National Academy of Sciences of the United States of America 100:8308-8313.

Shimazu R, Akashi S, Ogata H, Nagai Y, Fukudome K, Miyake K, Kimoto M. 1999. MD-2, a molecule that confers lipopolysaccharide responsiveness on Toll-like receptor 4. Journal of Experimental Medicine 189:1777-1782.

Short KM, Cox TC. 2006. Subclassification of the RBCC/TRIM superfamily reveals a

novel motif necessary for microtubule binding. Journal of Biological Chemistry 281:8970-8980.

Simonsen KL, Churchill GA, Aquadro CF. 1995. Properties of statistical tests of neutrality for DNA polymorphism data. Genetics 141:413-429.

Smith JM. 1992. Analyzing the mosaic structure of genes. Journal of Molecular Evolution 34:126-129.

Smith JM, Haigh J. 1974. The hitch-hiking effect of a favourable gene. Genetics Research 23:23-35.

Smith NG, Eyre-Walker A. 2002. Adaptive protein evolution in Drosophila. Nature 415:1022-1024.

Somerville JE, Cassiano L, Bainbridge B, Cunningham MD, Darveau RP. 1996. A novel Escherichia coli lipid a mutant that produces an antiinflammatory lipopolysaccharide. Journal of Clinical Investigation 97:359-365.

Stajich JE, Block D, Boulez K, Brenner SE, Chervitz SA, Dagdigian C, Fuellen G, Gilbert JGR, Korf I, Lapp H, *et al.* 2002. The bioperl toolkit: Perl modules for the life sciences. Genome Research 12:1611-1618.

Stebbings R, Findlay L, Edwards C, Eastwood D, Bird C, North D, Mistry Y, Dilger P, Liefooghe E, Cludts I, *et al.* 2007. "Cytokine storm" in the phase I trial of monoclonal antibody TGN1412: better understanding the causes to improve preclinical testing of immunotherapeutics. The Journal of Immunology 179:3325-3331.

Steeghs L, Keestra AM, van Mourik A, Uronen-Hansson H, van der Ley P, Callard R, Klein N, van Putten JP. 2008. Differential activation of human and mouse Toll-like receptor 4 by the adjuvant candidate LpxL1 of Neisseria meningitidis. Infection and Immunity 76:3801-3807.

Stenson PD, Ball EV, Mort M, Phillips AD, Shiel JA, Thomas NS, Abeysinghe S, Krawczak M, Cooper DN. 2003. Human Gene Mutation Database (HGMD): 2003 update. Human Mutat 21:577-581.

Stremlau M, Owens CM, Perron MJ, Kiessling M, Autissier P, Sodroski J. 2004. The cytoplasmic body component TRIM5alpha restricts HIV-1 infection in Old World monkeys. Nature 427:848-853.

Stremlau M, Perron M, Welikala S, Sodroski J. 2005. Species-specific variation in the B30.2(SPRY) domain of TRIM5alpha determines the potency of human immunodeficiency virus restriction. Journal of of Virology 79:3139-3145.

Sukumaran J, Holder MT. 2010. DendroPy: a Python library for phylogenetic computing. Bioinformatics 26:1569-1571.

Sullivan C, Charette J, Catchen J, Lage CR, Giasson G, Postlethwait JH, Millard PJ, Kim CH. 2009. The Gene History of Zebrafish tlr4a and tlr4b Is Predictive of Their Divergent Functions. Journal of Immunology 183:5896-5908.

Suntharalingam G, Perry MR, Ward S, Brett SJ, Castello-Cortes A, Brunner MD, Panoskaltsis N. 2006. Cytokine storm in a phase 1 trial of the anti-CD28 monoclonal antibody TGN1412. The New England Journal of Medicine 355:1018-1028.

Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123:585-595.

Takao K, Miyakawa T. 2014. Genomic responses in mouse models greatly mimic human inflammatory diseases. Proceedings of the National Academy of Sciences of the United States of America. Published ahead of print August 4, 2014.

Tamura T, Yanai H, Savitsky D, Taniguchi T. 2008. The IRF family transcription factors in immunity and oncogenesis. Annual Review of Immunology 26:535-584.

Tcherkezian J, Lamarche-Vane N. 2007. Current knowledge of the large RhoGAP family of proteins. Biology Cell 99:67-86.

Tennessen JA. 2008. Positive selection drives a correlation between non-synonymous/synonymous divergence and functional divergence. Bioinformatics 24:1421-1425.

Thompson JD, Plewniak F, Ripp R, Thierry JC, Poch O. 2001. Towards a reliable objective function for multiple sequence alignments. Journal of Molecular Biology 314:937-951.

Thomson TM, Lozano JJ, Loukili N, Carrio R, Serras F, Cormand B, Valeri M, Diaz VM, Abril J, Burset M, *et al.* 2000. Fusion of the human gene for the polyubiquitination coeffector UEV1 with Kua, a newly identified gene. Genome Research 10:1743-1756.

Tordai H, Nagy A, Farkas K, Banyai L, Patthy L. 2005. Modules, multidomain proteins and organismic complexity. FEBS Journal 272:5064-5078.

van der Walt S, Colbert SC, Varoquaux G. 2011. The NumPy Array: A Structure for Efficient Numerical Computation. Computing in Science & Engineering 13:22-30.

Van Valen L. (r03423 co-authors). 1973. A new evolutionary law. Evolutionary Theory 1:1-30.

Wakimoto H, Ikeda K, Abe T, Ichikawa T, Hochberg FH, Ezekowitz RA, Pasternack MS, Chiocca EA. 2002. The complement response against an oncolytic virus is species-specific in its activation pathways. Molecular Therapy 5:275-282.

Walsh TA. 2013. The evolution of the mammal placenta - a computational approach to the identification and analysis of placenta-specific genes and microRNAs. PhD Thesis. Dublin: Dublin City University.

Watterson GA, Guess HA. 1977. Is the most frequent allele the oldest? Theoretical Population Biology 11:141-160.

Watts DJ, Strogatz SH. 1998. Collective dynamics of 'small-world' networks. Nature 393:440-442.

Webb B, Sali A. 2014. Protein Structure Modeling with MODELLER. Protein Structure Prediction, 3rd Edition 1137:1-15.

Werling D, Jann OC, Offord V, Glass EJ, Coffey TJ. 2009. Variation matters: TLR structure and species-specific pathogen recognition. Trends Immunology 30:124-130.

Whelan S, Goldman N. 2001. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Molecular Biology and Evolution 18:691-699.

Wiens J. 2004. The role of morphological data in phylogeny reconstruction. Systematic Biology 53:653-661.

Wittkopp PJ, Kalay G. 2012. Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. Nature Reviews Genetics 13:59-69.

Wlasiuk G, Nachman MW. 2010. Adaptation and constraint at Toll-like receptors in primates. Molecular Biology and Evolution 27:2172-2186.

Woolfe A, Mullikin JC, Elnitski L. 2010. Genomic features defining exonic variants that modulate splicing. Genome Biology 11:R20.

Wuchty S. 2001. Scale-free behavior in protein domain networks. Molecular Biology and Evolution 18:1694-1702.

Wuchty S, Almaas E. 2005. Evolutionary cores of domain co-occurrence networks. BMC Evolution Biology 5:24.

Yamamoto M, Sato S, Hemmi H, Sanjo H, Uematsu S, Kaisho T, Hoshino K, Takeuchi O, Kobayashi M, Fujita T, *et al.* 2002. Essential role for TIRAP in activation of the signalling cascade shared by TLR2 and TLR4. Nature 420:324-329.

Yang Z. 2002. Inference of selection from multiple species alignments. Current Opinion in Genetics & Development 12:688-694.

Yang Z. 1998. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. Molecular Biology and Evolution 15:568-573.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Molecular Biology and Evolution 24:1586-1591.

Yang Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. Computer Applications in the Biosciences 13:555-556.

Yang Z, Bielawski JP. 2000. Statistical methods for detecting molecular adaptation. Trends in Ecology & Evolution 15:496-503.

Yang Z, dos Reis M. 2011. Statistical properties of the branch-site test of positive selection. Molecular Biology and Evolution 28:1217-1228.

Yang Z, Nielsen R. 2002. Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. Molecular Biology and Evolution 19:908-917.

Yang Z, Nielsen R, Goldman N, Pedersen AM. 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics 155:431-449.

Yang Z, Nielsen R, Hasegawa M. 1998. Models of amino acid substitution and applications to mitochondrial protein evolution. Molecular Biology and Evolution 15:1600-1611.

Yang Z, Rannala B. 2012. Molecular phylogenetics: principles and practice. Nature Reviews Genetics 13:303-314.

Yang Z, Wong WS, Nielsen R. 2005. Bayes empirical bayes inference of amino acid sites under positive selection. Molecular Biology and Evolution 22:1107-1118.

Yang ZH. 1993. Maximum-Likelihood-Estimation of Phylogeny from DNA-Sequences When Substitution Rates Differ over Sites. Molecular Biology and Evolution 10:1396-1401.

Yang ZH, Rannala B. 1997. Bayesian phylogenetic inference using DNA sequences: A Markov Chain Monte Carlo method. Molecular Biology and Evolution 14:717-724.

Yap MW, Nisole S, Stoye JP. 2005. A single amino acid change in the SPRY domain of human Trim5alpha leads to HIV-1 restriction. Current Biology 15:73-78.

Yim HS, Cho YS, Guang XM, Kang SG, Jeong JY, Cha SS, Oh HM, Lee JH, Yang EC, Kwon KK, *et al.* 2014. Minke whale genome and aquatic adaptation in cetaceans. Nature Genetics 46:88-92.

Yokoyama S. 2012. Synthesis of Experimental Molecular Biology and Evolutionary Biology: An Example from the World of Vision. Bioscience 62:939-948.

Yokoyama S. 2013. Synthetic biology of phenotypic adaptation in vertebrates: the next frontier. Molecular Biology and Evolution 30:1495-1499.

Yokoyama S, Radlwimmer FB. 2001. The molecular genetics and evolution of red and green color vision in vertebrates. Genetics 158:1697-1710.

Yokoyama S, Tada T, Zhang H, Britt L. 2008. Elucidation of phenotypic adaptations: Molecular analyses of dim-light vision proteins in vertebrates. Proceedings of the National Academy of Sciences of the United States of America 105:13480-13485.

Zeng J, Konopka G, Hunt BG, Preuss TM, Geschwind D, Yi SV. 2012. Divergent Whole-Genome Methylation Maps of Human and Chimpanzee Brains Reveal Epigenetic Basis of Human Regulatory Evolution. The American Journal of

Human Genetics 91:455-465.

Zhang DK, Zhang GL, Hayden MS, Greenblatt MB, Bussey C, Flavell RA, Ghosh S. 2004. A toll-like receptor that prevents infection by uropathogenic bacteria. Science 303:1522-1526.

Zhang J. 2004. Frequent false detection of positive selection by the likelihood method with branch-site models. Molecular Biology and Evolution 21:1332-1339.

Zhang J, Nielsen R, Yang Z. 2005. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. Molecular Biology and Evolution 22:2472-2479.

Zhang JZ. 2003. Evolution by gene duplication: an update. Trends in Ecology & Evolution 18:292-298.

Zuckerkandl E, Pauling L. 1965. Evolutionary divergence and convergence in proteins. pages. Evolving Genes and Proteins, edited by V. Bryson and H.J. Vogel. Academic Press, New York.

Publications

1) Morgan CC, Shakya K, Webb A, Walsh TA, Lynch M, Loscher CE, Ruskin HJ, O'Connell MJ. 2012. Colon cancer associated genes exhibit signatures of positive selection at functionally significant positions. BMC Evolutionary Biology 12:114. doi:10.1186/1471-2148-12-114. PMID: 22788692.

Contributions by Webb A: Homolog identification, selective pressure analyses, and contributed to the drafting of the manuscript.

2) Morgan CC, Foster PG, Webb AE, Pisani D, McInerney JO, O'Connell MJ. 2013. Heterogeneous models place the root of the placental mammal phylogeny. Molecular Biology and Evolution 30:2145-2156. doi: 10.1093/molbev/mst117. PMID: 23813979.

Contributions by Webb AE: Single gene ortholog identification and contributed to the drafting of the manuscript.

3) Keane M, Semeiks J, Webb AE, Li YI, Quesada V, Craig T, Madsen LB, van Dam S, Brawand D, Marques PI, et al. 2015. Insights into the evolution of longevity from the bowhead whale genome. Cell Reports 10:112-122. doi: 10.1016/j.celrep.2014.12.008. PMID: 25565328.

Contributions by Webb AE: Data analysis (homolog identification and selective pressure analyses) and contributed to the drafting of the manuscript.