Investigating the duplication and evolution of essential fertilization proteins

Alberto Marcos Rivera

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington
2022

Reading Committee:
Willie J. Swanson, Chair
Barbara Wakimoto
Evan E. Eichler

Program Authorized to Offer Degree:
Genome Sciences

University of Washington

Abstract

Investigating the duplication and evolution of essential fertilization proteins

Alberto Marcos Rivera

Chair of Supervisory Committee:

Professor Willie J. Swanson

Genome Sciences

Duplication processes such as whole gene duplication and tandem domain expansion are important for the evolution and functional diversification of essential protein families. While whole gene duplications are well established sources of new genes and biological novelty, less attention has been paid to how domain level duplications also allow proteins to neofunctionalize. Genes encoding fertilization proteins are also some of the most rapidly evolving in the genome, which could enable the neofunctionalization of duplicated domains within these genes. Chapter 1 of this dissertation reviews multiple well known gene families (Izumo, DCST, ZP, and the TFP superfamily) that arose from gene duplication. ZPs and TFPs also demonstrate tandem domain duplication and functional diversification.

In chapter 2 of this dissertation, we present research into the evolutionary history of domain duplication and neofunctionalization within the Zona pellucida amino (ZP-N) terminal domain. A large scale phylogenetics analysis across vertebrates revealed a divide between two classes of ZP-N domains: those that are paired with the ZP-C domain in the terminal ZP module (modular), and those outside of this module (free). This suggests that there was an initial ZP-N duplication event in vertebrates which then produced a wide array of functional diverse ZP-N

domains. Machine learning classification also reveals that modular domains are more conserved at the level of both sequence and structure. In contrast, free domains are more divergent and some in ZP2 show evidence of positive selection. While modular ZP-Ns domains may be conserved for a structural role, free ZP-Ns have experienced a history of duplications and neofunctionalization in fertilization.

Chapter 3 of this dissertation outlines transcriptomic research in abalone ovaries. Much of this research has been motivated by earlier sperm proteomic work from the Swanson lab, as well as advancements in sequencing technology. The aim is to identify ovary expressed genes that play important roles in fertilization. We have identified multiple ZP proteins, that show homology with previously sequenced ZP pseudogenes. One of the newly described ZP proteins has a duplicated ZP-N domain, and phylogenetics suggests this occurred independently from vertebrate ZP-N expansions. This transcriptome analysis also identified five abalone ovary TFPs, which may have experienced structural modifications relative to published TFPs. Taken together, the research findings suggest a history of recurrent independent co-option, structural modification, and functional diversification of fertilization proteins. In chapter 4, we discuss several possible extensions of this research including more extensive positive selection analyses and co-evolutionary analyses of the transcriptome data.

Contents

Acknowledgments

I want to thank everyone who made the completion of dissertation possible. I would like to thank my advisor, Willie Swanson, along with the rest of the Swanson lab: Jan Aagaard, Damien Wilburn, Emily Killingbeck, and Jolie Carlisle. Willie was always incredibly supportive and provided a good balance of advice, encouragement, and freedom. I want to thank Damien Wilburn who taught me a great deal about protein biochemistry and machine learning, but perhaps more importantly he shared multiple streaming service passwords with me. Jan was also a delightful presence in the lab, and Emily helped hone my eyes for PowerPoint slide making. I also thank Jolie, who was a good co-worker and friend for the duration of my dissertation.

I also have to thank the Genome Sciences department for all the help and support I have received. I thank the members of my committee: Evan Eichler, Christine Queitsch, Mike MacCoss, Jim Bruce, and Barbara Wakimoto. They all provided useful feedback during my committee meetings. PacBio long read sequencing was performed with the help of the Eichler lab, and mass spectrometry was performed with help from the MacCoss lab. The department also provided opportunities for conference travel including Society for Molecular Biology and Evolution meetings in Manchester, UK and Yokohama, Japan. I also owe a great deal of thanks to GSIT for their patience with me, as well as Brian Giebel for answering all of my questions. I also acknowledge the funding I received, which included the Genome Training Grant, and NIH grants awarded to Willie and Damien.

Thank you to many of the other acquaintances and friends I made during my time in graduate school. When I moved to this city, the GS2016 cohort provided me a great sense of community and social cohesion. I have also been lucky enough to befriend several graduate students in the years above and below me. These ties have warmed my heart and have been someone of the greatest rewards from my time in genome sciences. I would want to individually

thank all the graduate students who were important to me, but it would take many pages and I would be mortified if I forgot to mention someone. I will just say that I hope some of the friendships I made here last a very long time.

I also need to thank the people who made it possible for me to get to this point in the first place. None of this would have been possible without my parents Ana and Aurelio. For starters, I am their biological progeny so this literally would have been impossible without them. To most people higher education and scientific research seem so far out of reach, but my parents made sure I understood that my academic goals were possible and admirable. They encouraged my scientific interests for as long as I can remember and have also provided three decades of emotional support. I also thank my older brother Adrian, who might be the most compassionate person I know. As a child, he looked out for me and helped me build my sense of self-confidence. Even as an adult, I still feel him looking out for me. I also have a large extended family of aunts, uncles, and cousins, on whom I can always rely. I won't thank them all individually, but my cousins really were my first friends. I am so thankful for the family and friends who have cared for me. I don't know if I'm the luckiest man alive, but I sure feel that way sometimes.

Dedication

I dedicate this dissertation to my parents Ana and Aurelio, and to my older brother Adrian, who have been there for me every step of this journey.

“I’d like to thank the streets that drove me crazy and all the televisions out there that raised me”
– Lupe Fiasco

Chapter 1: The Importance of Gene Duplication and Domain Repeat Expansion for the Function and Evolution of Fertilization Proteins

A version of this chapter was previously published as:
Rivera AM and Swanson WJ, 2022. The Importance of Gene Duplication and Domain Repeat Expansion for the Function and Evolution of Fertilization Proteins. Frontier in Cell and Developmental Biology (10):827455

The process of gene duplication followed by gene loss or evolution of new functions has been studied extensively, yet the role gene duplication plays in the function and evolution of fertilization proteins is underappreciated. Gene duplication is observed in many fertilization protein families including Izumo, DCST, ZP, and the TFP superfamily. Molecules mediating fertilization are part of larger gene families expressed in a variety of tissues, but gene duplication followed by structural modifications has often facilitated their co-option into a fertilization function. Repeat expansions of functional domains within a gene also provide opportunities for the evolution of novel fertilization protein. ZP proteins with domain repeat expansions are linked to species-specificity in fertilization and TFP proteins that experienced domain duplications were co-opted into a novel sperm function. This review outlines the importance of gene duplications and repeat domain expansions in the evolution of fertilization proteins.

1.1 Introduction to duplication in fertilization proteins

The fertilization of oocytes by sperm is an essential function in sexual reproduction, and multiple stages of the fertilization cascade have been described (Vacquier 1998). First the sperm is drawn to the egg through chemotaxis (Ramírez-Gómez et al. 2019), and it then binds to the egg and releases proteins stored in the acrosome. The sperm then passes through the

glycoproteinaceous egg coat (Monne et al. 2008; Wilburn and Swanson 2016) (named Zona Pellucida in mammals), and proceeds to the oocyte cell membrane to initiate fusion (Siu et al. 2021). Understanding fertilization requires knowledge of both these broad steps of the fertilization cascade and the molecular mechanism underlying them. Research into the evolution and function of gametic proteins has implications for the development of novel contraception or treatments for unexplained human infertility (Gelbaya et al. 2014).

Many fertilization proteins are members of gene families that result from whole gene duplication events, which is a common mechanism for gene birth (Hughes 1994). There has been extensive research into the relationship between gene duplication and other aspects of reproductive biology, including the neuroendocrine control of reproduction (Dufour et al. 2020), protease activity in the female reproductive tract (Kelleher et al. 2007; Kelleher and Markow 2009), the resolution of sexual conflict (Gallach et al. 2010; Connallon and Clark 2011; Gallach et al. 2011; Gallach and Betrán 2011), and hybridization barriers (Ting et al. 2004). This review specifically focuses on our growing knowledge of duplicated protein families implicated in fertilization. These proteins include the Izumo1 and Juno pair of interacting proteins, which each arose from independent gene duplication events and are essential gamete membrane fusion function in mammals(Bianchi et al. 2014). DCST1 and DCST2 are paralogous proteins expressed in the sperm membrane of some bilateral animals, that are essential for fertilization (Inoue, Hagihara, et al. 2021: 1). Other duplicated proteins that act in fertilization include ADAMs (Primakoff and Myles 2000; Civetta 2003; Finn and Civetta 2010), CRISPs (Busso et al. 2007; Da Ros et al. 2008; Gibbs et al. 2011; Maldera et al. 2014), Catspers (Clapham and Garbers 2005; Navarro et al. 2008; Speer et al. 2021), and PKDREJ on the male side (Sutton et al. 2008), and tetraspanins (CD9,CD81) (Le Naour François et al. 2000: 9; Miyado Kenji et al. 2000; Frolikova et al. 2018) and EBR1 on the female side (Kamei and Glabe 2003; Hart 2013).Genomic resources suggests that most of these families (ADAMs, tetraspanins, EBR,

PKRDEJ, and Catsper) have orthologs in other bilateral animals, while CRISP has orthologs in animals and in yeast (Howe et al. 2021).

Duplicated genes can experience further structural diversification, such as the duplication of individual functional protein domains. Proteins containing tandemly duplicated domains constitute a small, but significant portion of the genome (Han et al. 2007; Nacher et al. 2010). Independent tandem duplications of individual functional domains is also a recurrent trend in some protein families (TFP and ZP) (Galindo et al. 2002; Aagaard et al. 2010; Doty et al. 2016). There are several families of reproductive proteins on both the sperm and egg that show a history of being co-opted from non-reproductive functions (Fig 1.1). Three finger proteins (TFPs) have been frequently co-opted for fertilization including SPACA4 in tetrapods, Bouncer in fish, and multiple classes of sperm proteins in plethodontid salamanders (PMF, SPFs) (Doty et al. 2016; Fujihara et al. 2021). Salamander SPFs have a duplicated three finger protein domain, and have evolved structural modifications to those domains (Doty et al. 2016). Similarly, the family of ZP proteins (named after the Zona Pellucida), essential components of egg coats across vertebrates and invertebrates (Wilburn and Swanson 2016), show evidence of independent expansions of ZP-N domains in different lineages (Liang and Dean 1993; Galindo et al. 2002). These highlight the role of gene duplication and repeat domain expansions in fertilization. An observed trend is rapid sequence evolution in reproductive proteins (Swanson and Vacquier 2002), and newly duplicated domains can provide novel substrates for evolving new functions at multiple stages of the fertilization cascade.

The role of duplications in genome evolution is well documented across the tree of life. (Kondrashov et al. 2002; Conant and Wolfe 2008). Gene duplication (Ponting 2008) is an important source for new genetic material that facilitates biological innovation. The duplication and differentiation of genomic regions has been linked to the evolution of modularity in organisms (Wagner et al. 2007). Modularity is an abstract concept in which part of an organism (such as a network of protein interactions) functions largely autonomously relative to other

aspects of the organism's biology (Wagner and Altenberg 1996; West-Eberhard 2005). Duplicated genes can participate in existing modular protein interaction networks, which facilitates increasing biological complexity of these networks (Wagner et al. 2007). Such increases in modular network complexity through gene duplication has been linked to adaptations in humans (Perry et al. 2007). Duplicated functional domains can similarly contribute to the evolution of biological complexity. This review will discuss both whole gene duplications and within gene domain duplications, and their role in the evolution of reproductive functions.

When genes duplicate they experience one of three possible fates: pseudogenization, subfunctionalization, and neofunctionalization (Walsh 2003; Innan 2009). Due to redundancies in function, the duplicated gene may no longer experience conservation and accumulate silencing mutations, resulting in a non-coding "pseudogene" (Fig 1.2). New mutations are frequently deleterious, so pseudogenization is hypothesized to be the most common fate of duplicated genes (Lynch and Conery 2000). However, the other two fates of duplicated genes (subfunctionalization and neofunctionalization) are common mechanisms for biological innovation. Under neofunctionalization, one gene copy maintains its original function while the other experiences positive selection and evolves a novel function. While under subfunctionalization, both copies parse the original function, and neither gene is sufficient (Walsh 2003; Innan 2009).

Tandem duplications of individual protein domains within a gene can add greater complexity to the duplication process. Paralogous genes experiencing relaxed selection can have greater freedom for tandem domain duplications. There is strong research interest in the mechanisms underlying domain repeat expansions and how they affect the evolution of protein families (Björklund et al. 2005; Vogel et al. 2005; Björklund et al. 2006; Weiner 3rd et al. 2006; Moore et al. 2008; Buljan and Bateman 2009). Repeats can experience concerted evolution where they maintain a high degree of sequence identity (Elder and Turner 1995; Liao 1999),

through unequal recombination and gene conversion (Schimenti 1999). Under this scenario, the repeat expansion of highly identical domains is itself an innovation that could allow proteins to evolve novel functions. A repeat domain expansion could also affect dosage or protein interaction networks. Repeated domains could similarly differentiate in amino acid sequence, leading to neofunctionalization or subfunctionalization with the original domain. There are many possible orders and combinations of whole gene duplications and domain duplications that can contribute to the expansion of gene families (Fig 1.2). The process by which duplicate genes are maintained and experience subfunctionalization or neofunctionalization has been characterized under the duplication-degeneration-complementation model (DDC) (Force et al. 1999). While most classical population genetics models (Walsh 2003; Innan 2009) primarily discuss the effect of silencing or beneficial mutations on coding regions, the DDC model focuses on the effect of mutations on regulatory regions and subfunctionalization. Essentially, mutations that can silence certain regulatory regions in a duplicate gene can lead to the two genes partitioning expression and eventually function (Force et al. 1999). Other models have suggested subfunctionalization is primarily important as a transition phase to neofunctionalization (Rastogi and Liberles 2005). The mechanisms of subfunctionalization and neofunctionalization remain a subject of rich debate, and concepts like the DDC model could have ramifications for protein evolution.

Subfunctionalization and neofunctionalization are foundational to the evolution of increased complexity in genomes and protein networks, and it is worth examining their particular importance in fertilization. Fertilization proteins are some of the most rapidly evolving proteins in genomes, as evidenced by high amino acid divergence (Swanson and Vacquier 2002). Their rapid evolution is likely driven by factors such as sexual conflict and molecular arms race dynamics between gametes, which can also contribute to the maintenance of fertilization barriers between species (Gavrilets and Waxman 2002; Gavrilets 2014). The general trend of rapid evolution in reproductive proteins could facilitate the subfunctionalization or neofunctionalization of domains.

1.2 Families of duplicated fertilization genes

1.2.1 Izumo/Juno

The fusion of sperm and egg is necessary for fertilization, but there are only a few known pairs of interacting gametic proteins identified at this stage (Wilburn and Swanson 2016). After years of research the interacting pair Izumo1 and Juno were identified in mammals (Bianchi et al. 2014). Izumo1 is the sperm expressed protein that mediates fusion (Inoue et al. 2005), and it interacts with the egg surface bound folate receptor 4 (known as Juno) (Bianchi and Wright 2014). Izumo1 and Juno are each part of protein families with multiple paralogues, but only the Izumo1/Juno pair is capable of interacting (Bianchi et al. 2014). There are four members of both the Izumo (Ellerman et al. 2009) and folate receptor families (FOLR) in mammals (Elwood 1989; Shen et al. 1994; Spiegelstein et al. 2000; Petronella and Drouin 2014). Despite being part of the folate receptor family, Juno does not actually bind folate, exemplifying how a single member of this gene family has been co-opted for a novel reproductive function (Bianchi et al. 2014).

While Juno represents a clear co-option into fertilization, the evolution of the Izumo gene family could also present an interesting example of neofunctionalization. Izumo1-4 all have a highly structurally conserved Izumo domain, but Izumo1 and Izumo4 have a shared pair of β-strands extending from this domain. Izumo1 experienced further structural modifications, as its β-strand extensions act as a hinge between the Izumo domain and a co-opted immunoglobulin-like domain (Aydin et al. 2016; Ohto et al. 2016). Such substantial structural changes could be important for the protein's ability to bind Juno. Research into other Izumo proteins suggests their involvement in fertilization. Izumo1-3 are transmembrane testis expressed proteins (Ellerman et al. 2009), while Izumo4 lacks a transmembrane domain and is expressed in the acrosome (Guasti et al. 2020). Izumo3 shows evidence of positive selection (Grayson and Civetta 2012), and is necessary for sperm acrosome formation (Inoue, Satouh, et al. 2021). The parallel

histories of structural modifications in Izumo1 and Juno allowed for this essential interaction to evolve.

The relationship between Izumo1, Juno and their paralogs is highlighted by our phylogeny (Fig 1.3), which contains a long branch leading to Juno (FOLR4). This could reflect the rapid accumulation of mutations in the Juno branch as it was co-opted to bind Izumo1 during gametic membrane fusion. Crystal structures confirm that 1:1 binding complexes form between Izumo1 and Juno (Aydin et al. 2016; Ohto et al. 2016). The adhesion of Izumo1 and Juno is conserved in mammals, and after the adhesion event Juno is released from the egg's surface in vesicles and may act to bind and neutralize acrosome reacted sperm (Bianchi et al. 2014). In mammals, this interaction functions as a block against polyspermy (Bianchi and Wright 2014). Blocks to polyspermy are essential, because eggs that fuse with multiple sperm are not viable and mammalian blocks to polyspermy exist at both the cell membrane (Evans 2020) and egg coat (Fahrenkamp et al. 2020).

Mutations to residues conserved in mammals greatly reduce binding, highlighting that particular changes to amino acid sequence and protein structure facilitated the neofunctionalization of Juno (Aydin et al. 2016). The more variable structural features (Ohto et al. 2016) in Juno may be important for the species-specificity of its binding to Izumo1 (Bianchi et al. 2014; Bianchi and Wright 2015; Han et al. 2016). Comparative genetic analyses identify positive selection in a subset of mammals (Laurasiatheria) (Grayson and Civetta 2012), and that Juno is likely rapidly co-evolving with Izumo1, which contributes to the specificity of their interactions (Grayson 2015). This specific binding is essential to both Juno's function in initiating membrane fusion, and the post-fusion neutralization of acrosome-reacted sperm (Wright and Bianchi 2016).

1.2.2 DCST

While Izumo1 and Juno are thought to initiate the complex molecular process of gametic membrane fusion in mammals, recent transgenic experiments and complementation studies have demonstrated that DCST1 and DCST2 are also essential (Inoue, Hagihara, et al. 2021). The DCST1/2 proteins are expressed on the sperm surface, and contain variable (4-6) transmembrane helical domains (DC-STAMP) (Inoue, Hagihara, et al. 2021: 1). DC-STAMP (dendritic cell specific transmembrane protein) refers to both the name of the domain and one of the proteins that contains this domain (Hartgers et al. 2000). The originally identified DC-STAMP protein has four transmembrane domains (Hartgers et al. 2001), and it is highly expressed in myeloid dendrocytes (Hartgers et al. 2000; Hartgers et al. 2001; Eleveld-Trancikova et al. 2005; Eleveld-Trancikova et al. 2008). The expression of DC-STAMP has been induced in macrophages (Staege et al. 2001) and osteoclasts(Nomiyama et al. 2005). This broad array of functions has motivated much research into the molecular mechanisms of DC-STAMP interactions, which has supported a role in osteoclast fusion (Kukita et al. 2004; Yagi et al. 2005; Jansen et al. 2009). There is also evidence of DC-STAMP related signaling in immune response (Nair et al. 2016). Along with these other diverse functions, it seems that DC-STAMP domains have been co-opted into an essential role in sperm-egg membrane fusion.

DCST1/2 are the first known essential fertilization factors that are conserved in both vertebrates and invertebrates (Inoue, Hagihara, et al. 2021). DCST1/2 orthologues have been identified in both *Caenorhabditis* and *Drosophila* (Kroft et al. 2005; Wilson et al. 2006; Wilson et al. 2018), which is the first known example sperm related factors being conserved this broadly across vertebrates and invertebrates (Inoue, Hagihara, et al. 2021: 1). However, there has been extensive structural diversification of these DCST1/2 across animals (Fig 1.4), especially between invertebrates and vertebrates. The low sequence identity of DCST1/2 proteins across animals, makes the conservation of reproductive function all the more remarkable. The ubiquitin

ligase activity of DCST1 (Nair et al. 2016) raises questions about the function of DCST1/2 in sperm. There is intense research interest into the signal activity of long non-coding RNA produced by DCST1 and its effect on cancer cell progression (Hu et al. 2020; Ai et al. 2021: 1; Wang et al. 2021). More investigation is necessary to understand the function of DC-STAMP domains in a broad range of signaling networks, and how they were neofunctionalized in sperm DCST1/2.

1.2.3 ZP Domains

ZP proteins are an essential class of egg coat proteins. An important feature of ZP proteins is the ZP module that consists of two domains, ZP-N and ZP-C, named after their relative N-terminal and C-terminal positioning. ZP-N and ZP-C domain are immunoglobular domains with characteristic patterns of disulfide bonding and β-sheets (Bokhove and Jovine 2018), and likely resulted from an ancestral domain duplication. The variability in amino acid sequence, disulfide placement, and loop structures between ZP-N and ZP-C (Lin et al. 2011) suggests differences in their biological function and evolutionary history.

ZP-N domains are of particular interest, because they form asymmetric dimers with their β-sandwich edges which are believed to promote polymerization between ZP modules (Jovine et al. 2002; Wilburn and Swanson 2017; Bokhove and Jovine 2018). There are several ZP proteins identified in vertebrates (ZP1-4, ZPAX and ZPD), and there appears to be a history of lineage specific gain and loss of ZP proteins among vertebrates (Galindo et al. 2002; Conner et al. 2005; Goudet et al. 2008; Claw and Swanson 2012; Meslin et al. 2012; Shu et al. 2015; Killingbeck and Swanson 2018). Like other families discussed in this review, there also multiple ZP proteins with non-reproductive functions (e.g. uromodulin and tectorin-alpha) (Legan et al. 1997; Brunati et al. 2015; Bokhove et al. 2016). This may be another example of domains being

co-opted into a reproductive function, and ZP-N polymerization domains may be important for egg coat assembly and structure.

Not only has gene duplication produced an assortment of ZP proteins, there are also examples of independent repeat expansions of ZP-N in both vertebrates and invertebrate egg coat proteins (Fig 1.5). Some have only one additional ZP-N domain, but there are more dramatic repeat expansion like mammalian ZP2 (4 ZP-Ns) and abalone VERL (23 ZP-Ns) (Galindo et al. 2002). This process of domain duplications helped contribute to the diversity of ZP proteins. Given the ability of ZP-N domains to dimerize (Jovine et al. 2002; Bokhove and Jovine 2018; Litscher and Wassarman 2020), their duplications could create opportunities to evolve novel binding functions. Proteins with duplicated ZP-N domains, such as mammalian ZP2 and abalone VERL, are thought to be essential for species-specific in fertilization (Avella et al. 2013; Avella et al. 2014; Raj et al. 2017). Species-specificity in abalone is associated with the coevolution between VERL and the sperm protein lysin (Galindo et al. 2003; Clark et al. 2009), suggesting a cooption of ZP-Ns in sperm-egg interactions during egg coat dissolution.

Neofunctionalization of ZP-N domains can also drive new interactions between ZP proteins, such as the evolution of essential intermolecular crosslinks (Nishimura et al. 2019), which affect the physical assemblage of proteins in the supramolecular structure of the egg coat. Indeed, mouse research has suggested the importance of egg coat supramolecular structure in fertilization (Rankin et al. 2003; Avella et al. 2013). The structure of the egg coat is also important for the oocyte's ability to block polyspermy. Protein cleavage of ZP2 is thought to initiate other egg coat structural modifications, which "harden" the egg coat and prevent sperm binding (Bleil et al. 1981; Gahlay et al. 2010; Fahrenkamp et al. 2020). Gene and domain duplications have produced a family of ZP proteins that contribute to the egg coat supramolecular structure, and are involved in both sperm recognition and polyspermy avoidance.

1.2.4 TFP superfamily

Three finger proteins are defined by their TFP domains, which have a characteristic disulfide bonding pattern and fold (Galat 2008; Galat et al. 2008). The broader TFP protein superfamily also includes proteins with structurally modified TFP-like domains(Galat 2015). While TFPs were originally identified in snake toxins (Low et al. 1976; Tsernoglou and Petsko 1977), members of the TFP superfamily have been to co-opted for reproductive functions in sperm (SPACA4, PMFs, and SPFs), egg (Bouncer), and pheromones (PMFs, and SPFs) (Doty et al. 2016; Fujihara et al. 2021; Wilburn et al. 2022) (Fig 1.6). Bouncer plays a role in species-specific sperm-egg fusion in teleost fish (Herberg et al. 2018), which raises questions about how other TFPs may function in fertilization. The TFP superfamily includes both soluble and membrane bound proteins, and has great functional diversity across many tissues and taxa (Alape-Girón et al. 1999; Tsetlin 1999; Kini 2002; Nirthanan et al. 2003; Kessler et al. 2017). Similar to ZP proteins, we observe a history of gene duplication, repeat expansion of domains, and functional diversification of TFP domain containing proteins.

An ancestral TFP protein experienced gene duplication to produce an assortment of single TFP-like domain proteins (1D-TFPs). One of these TFP genes experienced a tandem domain expansion to produce the ancestor of proteins with two TFP-like domains (2D-TFPs). Three independent co-option events have produced TFPs in gametes (Fig 1.6). A co-option of 1D-TFPs occurred in the ancestor of tetrapods and produced both Bouncer in fish, and SPACA4 in amniotes (Fig 1.6). Despite their protein homology, Bouncer is egg expressed while SPACA4 is sperm expressed and it is implicated in interactions between the sperm and egg coat (Fujihara et al. 2021), highlighting the functional diversification of TFPs. Another independent co-option of 1D-TFPs resulted in the sperm expressed plethodontid modulating factor (PMFs) salamanders, which extensively duplicated producing a diverse family of reproductive molecules (Wilburn et al. 2012; Wilburn et al. 2014; Doty et al. 2016; D. Wilburn et al. 2017). Salamander

PMFs are hypervariable proteins expressed in multiple tissues, and while they are structurally similar to other TFPs, they differ in loop length and disulfide bridge patterning, and show evidence of persistent diversification and positive selection (Palmer et al. 2010; Wilburn et al. 2012; Wilburn et al. 2014)

Among 2D-TFPs there was independent co-option into the sodefrin precursor-like factors (SPFs) of salamander sperm. SPFs then experienced their own history of gene duplications and radiation (Palmer, Watts, et al. 2007). Both PMFs and SPFs experienced disulfide bond reshuffling relative to the canonical 1D-TFP and 2D-TFP binding patterns, and these changes reflect the neofunctionalization of these molecules (Doty et al. 2016). These striking examples of independent gene duplications and neofunctionalization for reproductive functions raises questions as to whether there are additional unknown co-options of TFPs, and whether some protein domains are more susceptible to co-option in diverse biological contexts.

Both PMFs and SPFs are highly duplicated protein families, with some members being co-opted into pheromone function and others for sperm expression (Doty et al. 2016; Wilburn et al. 2022). As the sperm paralogs of PMFs and SPFs have only recently been discovered, functional studies have not yet been conducted. Male salamanders produce large number of PMFS and SPFs within their mental glands which promote ritual courtship behavior in females (Doty et al. 2016). Duplications of secreted male-expressed proteins could have provided an evolutionary substrate to evolve new pheromones (Wilburn et al. 2022). Structural changes in PMFs and SPFs, such as disulfide shuffling, may contribute to new functions in both sperm and pheromones. The TFP’s superfamily’s history of gene duplication, domain duplication, and neofunctionalization provides a unique model for the evolution of large gene families involved in fertilization.

1.3 Discussion and Conclusion

In this review, we discussed examples of duplicated gene families with roles in fertilization. Gene duplication and neofunctionalization are essential processes for the evolution of greater genomic and functional complexity in organisms. Duplicated paralogous genes have been co-opted into both sperm (Izumo1, DCST1/2) and egg (Juno) proteins involved in gamete membrane fusion (Bianchi et al. 2014; Inoue, Hagihara, et al. 2021: 1). Domain duplications within paralogs are also observed in the TFP superfamily and ZPs and have allowed both groups of genes to adopt novel functions at multiple stages of fertilization. As seen with TFPs, duplication events are often followed by notable protein structural changes (Doty et al. 2016) which may be tied to their co-option for novel fertilization functions. It is intriguing to consider hypotheses that account for these patterns of gene family expansion and diversification common in reproductive molecules.

Duplication events can facilitate the rapid evolution and neofunctionalization observed in many families of fertilization proteins. This rapid evolution can also be influenced by multiple factors such as sexual conflict, polyspermy avoidance, or genetic drift (Vacquier et al. 1997). The necessity of pathogen avoidance or blocks to polyspermy can drive oocytes to evolve reduced sperm binding ability. The sperm would then co-evolutionarily “chase” the egg, which can contribute to the rapid sequence evolution of gametic proteins, and to the species-specificity of these protein interactions (Gavrilets and Waxman 2002; Gavrilets 2014). The rapid evolution of reproductive proteins is explored in terms of amino acid mutations, but the repeat expansion of domains could also be part of this trend. Proteins with repeated domains could experience drift resulting in ever-changing molecular target, that interacting proteins must co-evolutionarily chase (Vacquier et al. 1997).

Duplications of reproductive proteins can also contribute to the phenomenon of functional redundancy, in which two duplicated genes have partially overlapping functions and

can compensate for each other's loss (Kafri et al. 2009). Functional redundancy has been observed in the CRISP family of reproductive proteins (Curci et al. 2020), and this property could emerge in other large protein families. While functional redundancy seems like it would be temporary as duplicated genes subfunctionalized or neofunctionalized, it can be a surprisingly evolutionarily stable property. Functional redundancy could confer fitness advantages by maintaining the robusticity of protein interaction networks in spite of stochasticity of expression between cells (Kafri et al. 2009). The rapid evolution of other reproductive proteins in these networks could place even greater value on robustness and stability of essential functions. Robusticity in these protein networks is believed to reduce the fitness cost of new mutations, which would increase the "evolvability" of these proteins and facilitate functional innovation(Kirschner and Gerhart 2008). The concepts of functional redundancy and robusticity of function may also apply to domain repeat expansions like the ZP-N domains of VERL. The processes of gene duplication, repeat domain expansion, structural modification, and neofunctionalization have been fundamental to the evolution of reproductive molecules across life.

1.4 Acknowledgments

We thank Damien B. Wilburn for sharing his code for visualizing transmembrane proteins, and fellow lab members Jolie Carlisle and Jan Aagaard for engaging in discussions. The lab is funded by NIH grant HD105025 awarded to Willie J. Swanson.

1.5 Figures

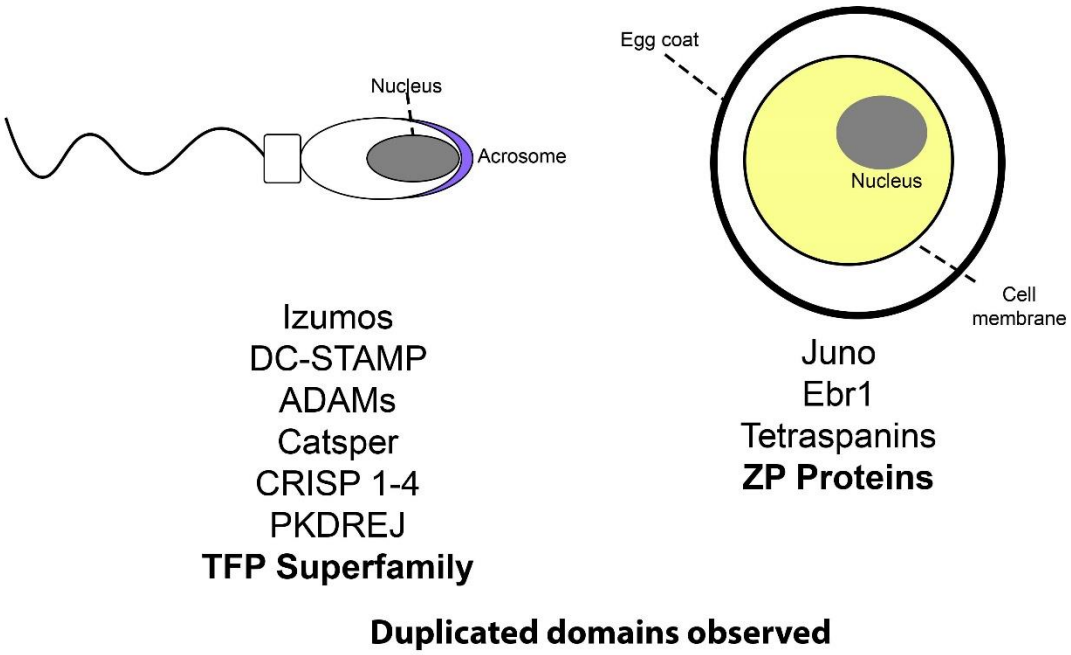


Figure 1.1: Overview of some duplicated gene families in fertilization. A cartoon schematic lists several protein families involved in reproduction. Those with notable repeat expansions are bolded.

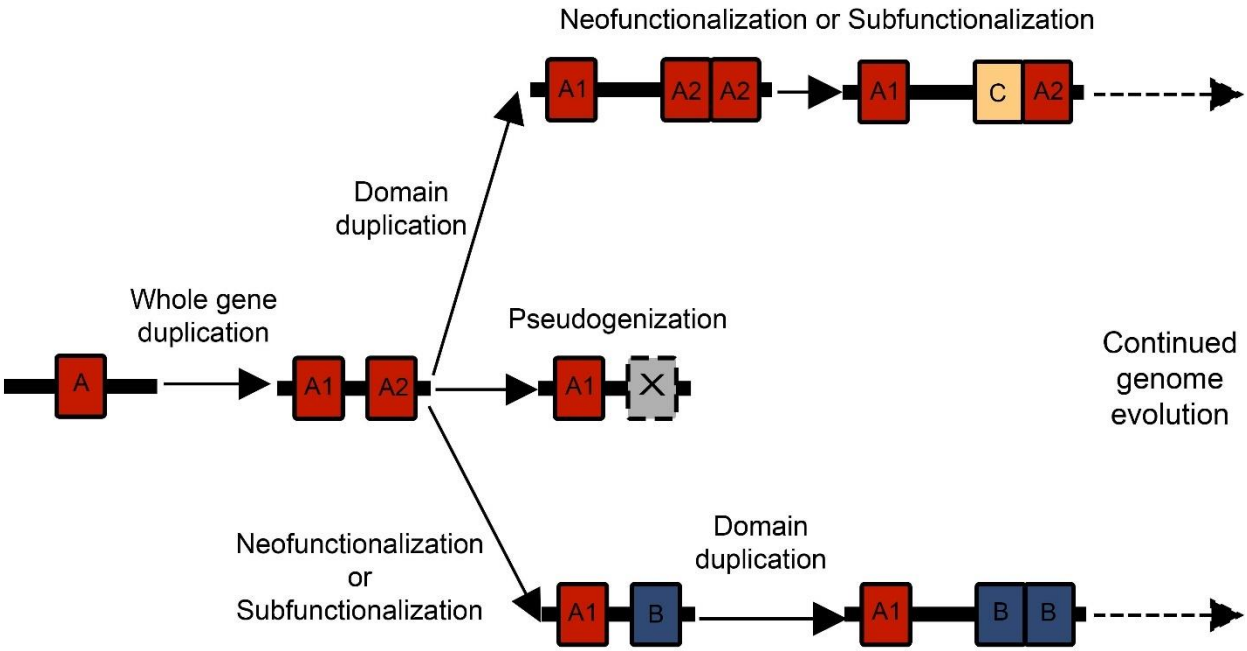


Figure 1.2: Schematic of duplication processes underlying gene family expansion. There are multiple possible combinations of whole gene and domain duplications that can birth new genes and functional domains. Often a whole gene duplication begins the process, then one of the gene duplicates experiences a domain expansion. These genes can then act as substrates for further duplication and neofunctionalization or subfunctionalization events.

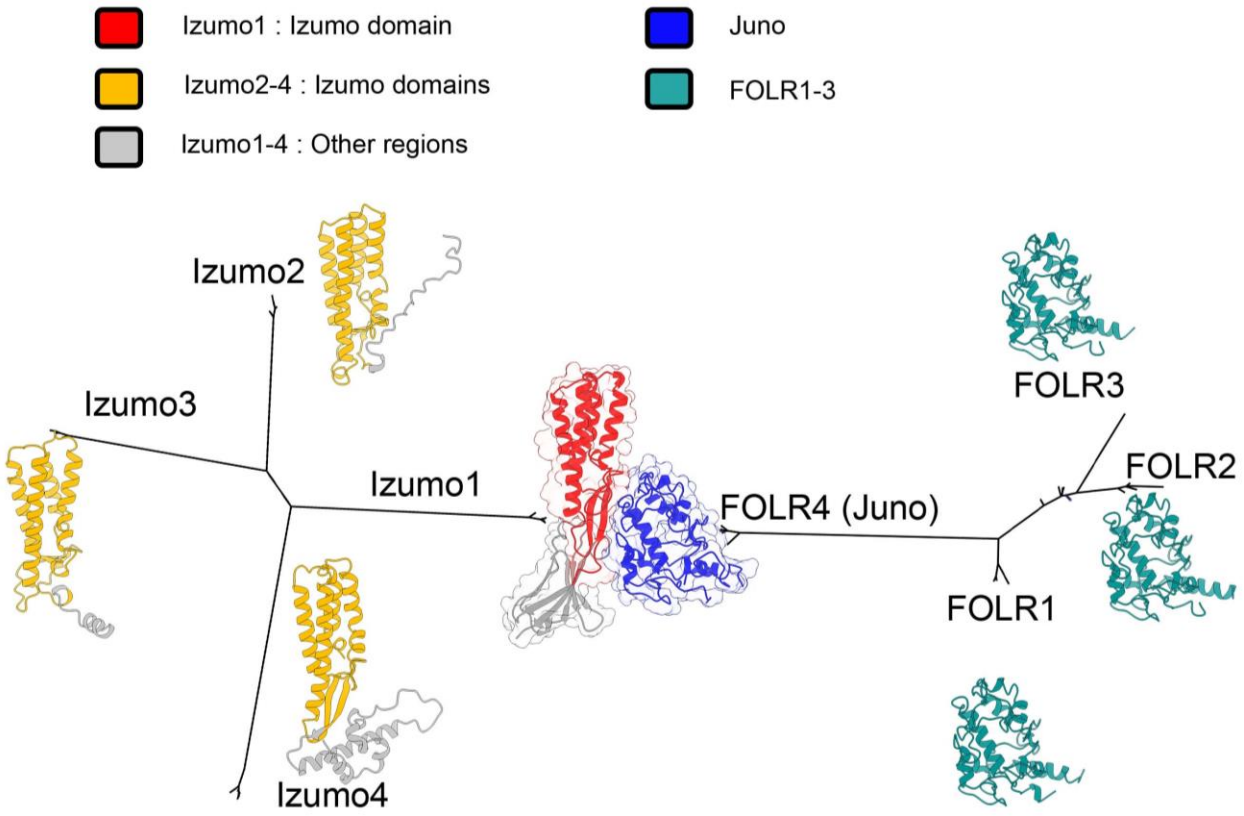


Figure 1.3: Phylogenies of Izumo and FOLR protein families. We constructed unrooted maximum likelihood phylogenies for *Izumo* and *FOLR* gene families in a subset of primates, based on multiple sequence alignments(Katoh and Standley 2013; Kozlov et al. 2019). Both gene families independently duplicated, but *FOLR4* was co-opted to bind *Izumo1*. Crystal structures have been obtained for the Izumo1-Juno complex (Aydin et al. 2016). For other proteins, alphafold predicted structures were used (Jumper et al. 2021). Using predictions of signal peptides and transmembrane domains, and secondary structural alignments, we identified shared izumo domains (Sonnhammer et al. 1998; Krogh et al. 2001; Almagro Armenteros et al. 2019).

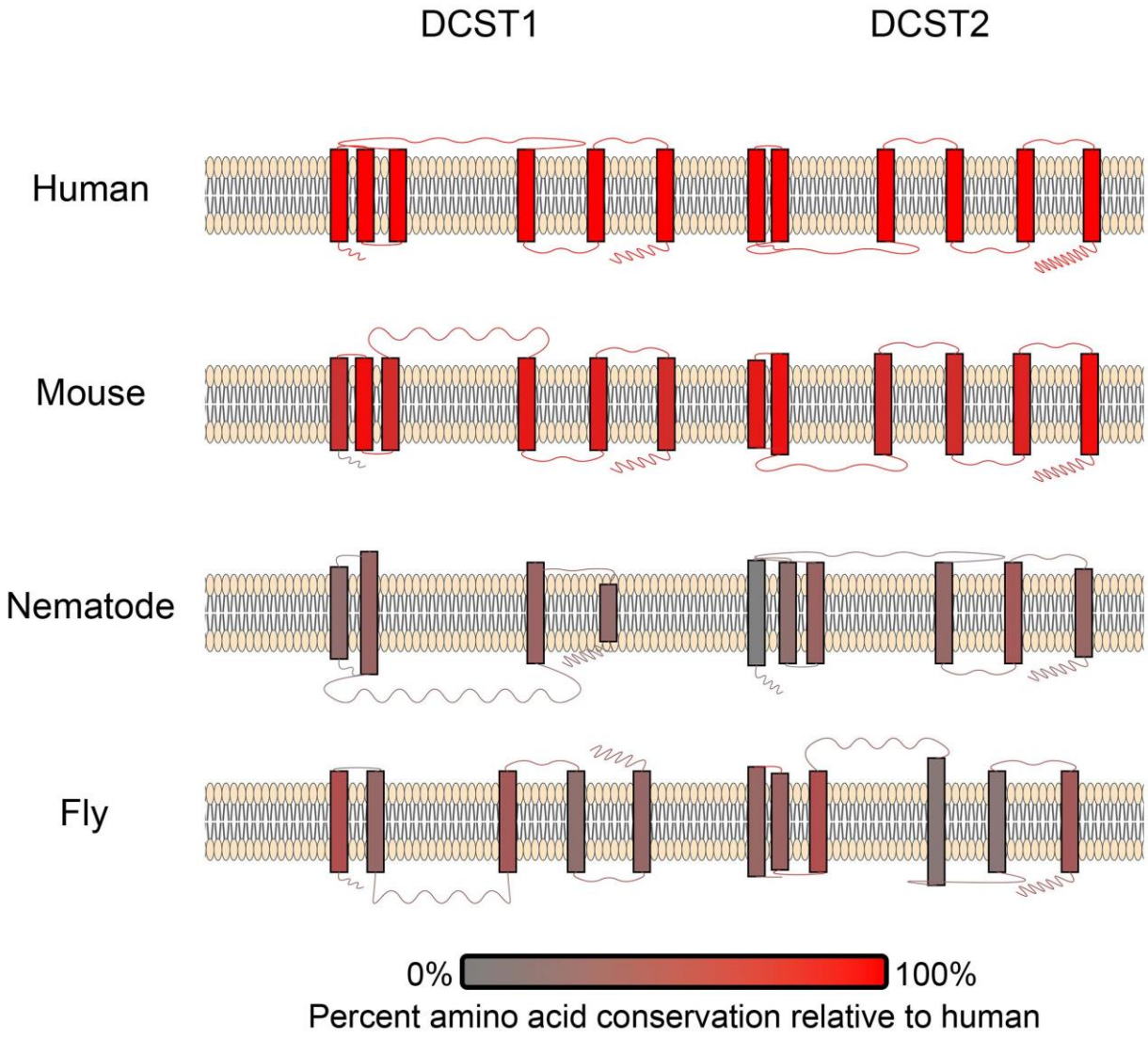


Figure 1.4: Schematic of membrane bound DCST1/2 proteins. We show here DCST1/2 proteins in multiple model species. The number of transmembrane domains and loop lengths differ across species. Transmembrane domains and loops are colored based on conservation (Pei et al. 2008), where red coloration signifies amino acid conservation relative to humans. Therefore, the human examples are all red.

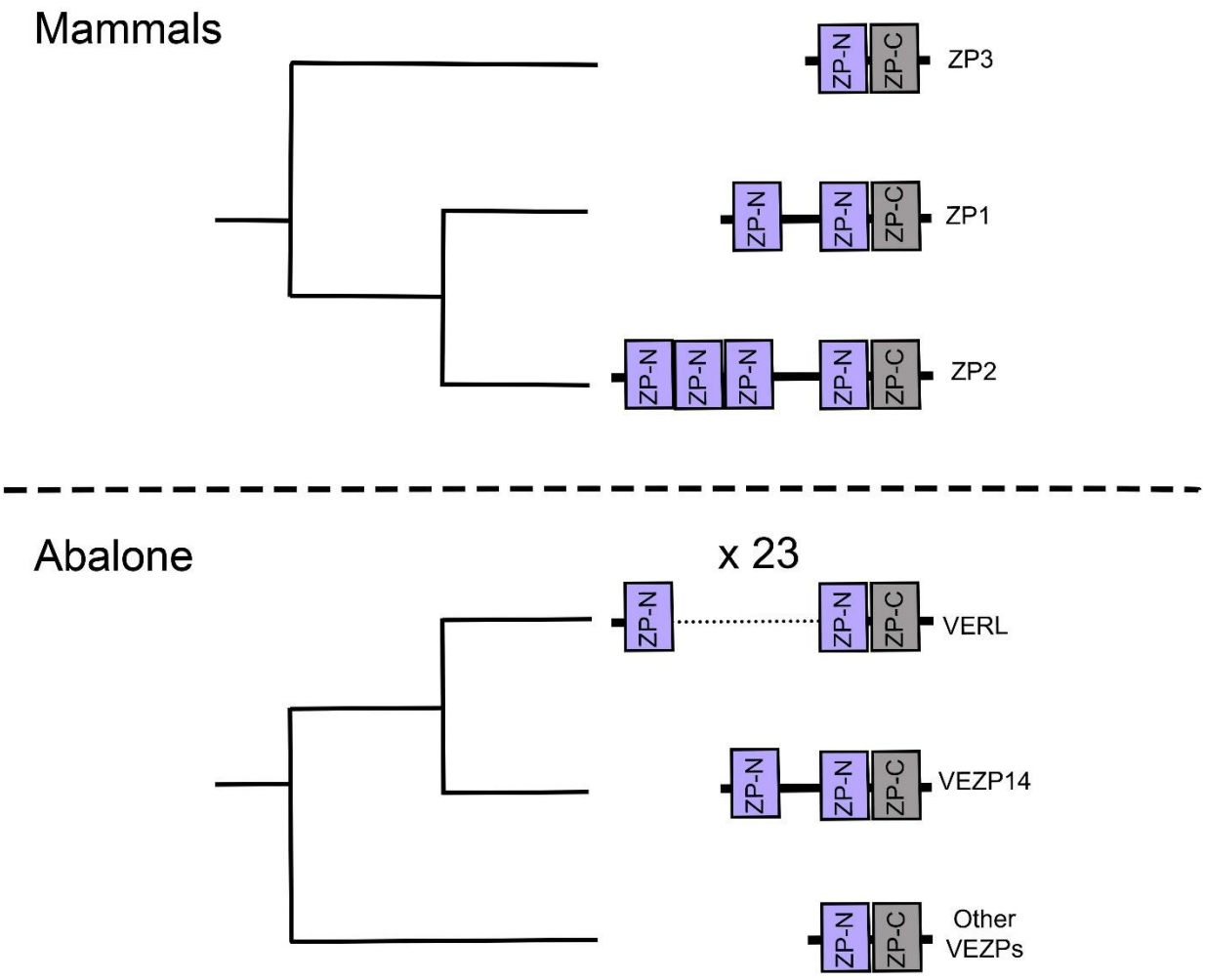


Figure 1.5: Evolution of ZP repeat expansions in mammals and abalone. Cladograms of ZP-N proteins are based on phylogenies from the literature (Aagaard et al. 2010; Claw and Swanson 2012). These suggest independent repeat expansion of the ZP-N domain in both abalone and human egg coat genes.

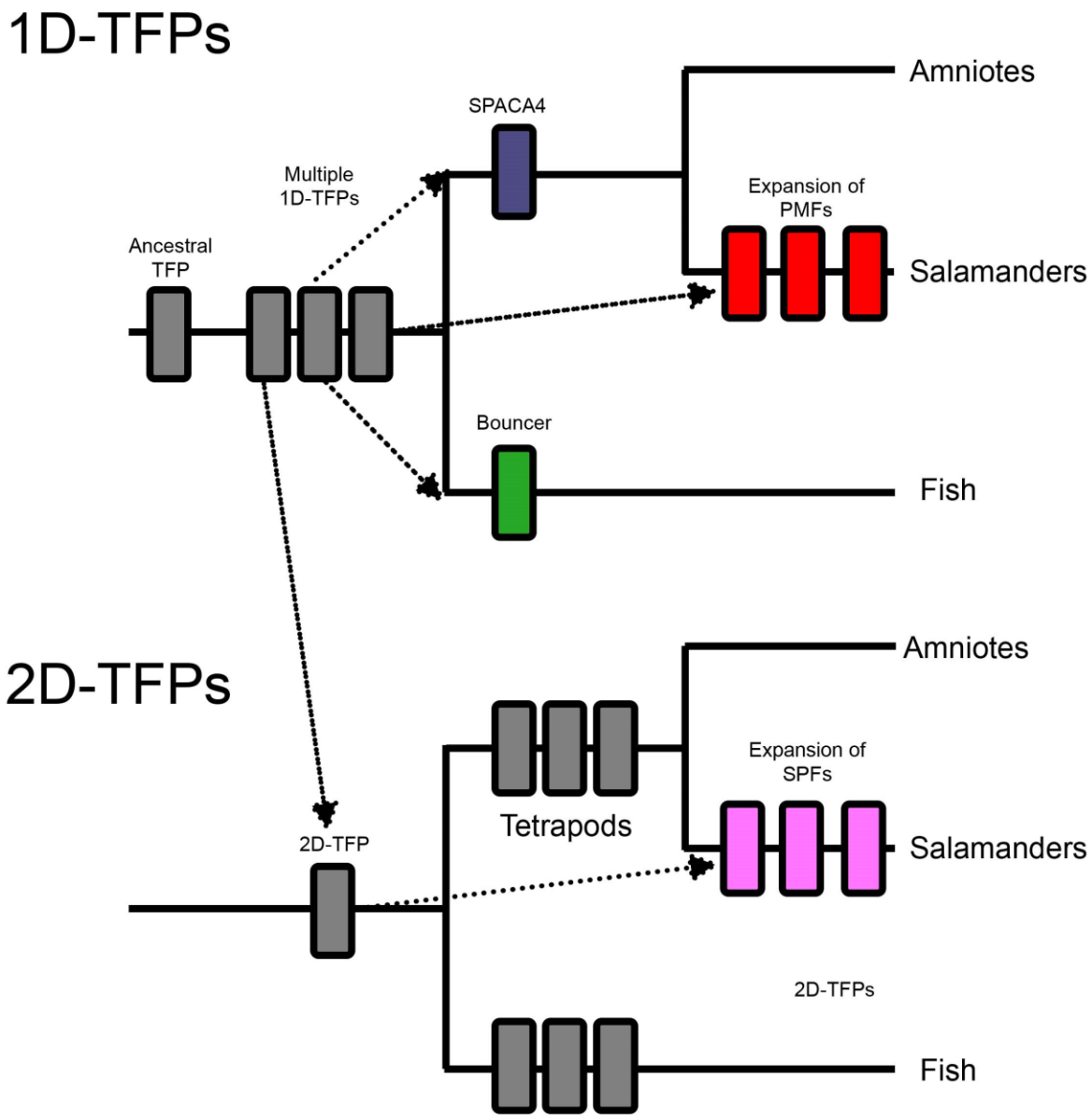


Figure 1.6: Evolutionary history of domain duplication and structural modifications in gametic TFPs. These two cladograms outline the whole gene and domain duplications within the three finger protein superfamily (TFPs) and their expansions into reproductive systems. An ancestral single domain TFP (1D-TFP), duplicated into multiple vertebrate 1D-TFPs, and also had a domain level duplication which created a lineage of two TFP domain proteins (2D-TFPs). The 1D-TFPs produced tetrapod SPACA4, fish Bouncer, and multiple salamander PMFs. The 2D-TFPs also duplicated throughout vertebrates including salamander SPFs. Both salamander PMF and SPF protein families include both sperm and pheromone expressed members (Wilburn et al. 2022).

Chapter 2: Domain expansion and functional diversification in vertebrate reproductive proteins

A version of this chapter was previously published as:
Rivera AM, Wilburn DB, and Swanson WJ, 2022. Domain expansion and functional diversification in vertebrate reproductive proteins. Molecular Biology and Evolution 39(5):msac105.

The rapid evolution of fertilization proteins has generated remarkable diversity in molecular structure and function. Glycoproteins of vertebrate egg coats contain multiple zona pellucida (ZP)-N domains (1–6 copies) that facilitate multiple reproductive functions, including species-specific sperm recognition. In this report, we integrate phylogenetics and machine learning to investigate how ZP-N domains diversify in structure and function. The most C-terminal ZP-N domain of each paralog is associated with another domain type (ZP-C), which together form a “ZP module.” All modular ZP-N domains are phylogenetically distinct from nonmodular or free ZP-N domains. Machine learning–based classification identifies eight residues that form a stabilizing network in modular ZP-N domains that is absent in free domains. Positive selection is identified in some free ZP-N domains. Our findings support that strong purifying selection has conserved an essential structural core in modular ZP-N domains, with the relaxation of this structural constraint allowing free N-terminal domains to functionally diversify.

2.1 Introduction

Protein structural domains are a major type of molecular building block which multimerize into higher order assemblies and provide the architectural foundation for nearly all cellular features, including organelles and extracellular matrices. Within molecular complexes, structural domains function as interlocking modules with specific, well-defined binding surfaces. Consequently, structural proteins commonly experience intense purifying selection to preserve their three-dimensional conformations, which can lead to extreme sequence conservation

between diverse taxa (e.g. actin is 89% identical between yeast and humans) (Rivero and Cvrčková 2007). The modularity of structural domains makes them prime templates for duplication within a genome, taking the form of both whole gene duplication to produce new paralogs as well as formation of tandem domain arrays within a single gene (Rivera and Swanson 2022). Redundancy of duplicated domains can relax purifying selection to allow for diversification and neofunctionalization, as is observed for the mechanosensitive tandem domains of cadherins that function in the inner ear(Jaiganesh et al. 2018). However, little is known as to how positive selection can shape structural domain diversification within rapidly evolving systems.

Within animal genomes, many of the fastest evolving genes are associated with fertilization (Swanson and Vacquier 2002). While often considered paradoxical, reproductive proteins evolve at extraordinary rates in part due to differences in male and female optimal mating rates that can drive sexual arms races, especially in gamete recognition proteins that initially mediate sperm-egg interactions (Wilburn and Swanson 2016; Wilburn et al. 2019). Fertilization of an egg by multiple sperm will fail to form a zygote – a phenomenon known as pathological polyspermy – and oocytes possess multiple reproductive barriers to modulate the rate of sperm entry (Frank 2000; Carlisle and Swanson 2021). One such barrier in vertebrate oocytes is an elevated glycoprotein envelope with clade-specific names: the zona pellucida (ZP) in mammals, the chorion in fishes, and the vitelline membrane in amphibians, reptiles, and birds (Wilburn and Swanson 2018). Named after the mammalian version, all vertebrate egg coat proteins contain a pair of immunoglobulin-like domains, ZP-N and ZP-C, that together form a polymerization unit called a ZP module (Jovine et al. 2002; Wilburn and Swanson 2017; Bokhove and Jovine 2018). The last common ancestor of vertebrates possessed six paralogous genes (*zp1*, *zp2*, *zp3*, *zp4*, *zpd*, and *zpax*) that have experienced clade-specific birth and death events. Consequently, the egg coat of each major vertebrate class has a different composition

of ZP module-containing proteins (Conner et al. 2005; Wong and Wessel 2005; Goudet et al. 2008; Meslin et al. 2012; Shu et al. 2015; Wassarman and Litscher 2016; Killingbeck and Swanson 2018). ZP modules are also found in non-reproductive proteins that form extracellular matrices, such as uromodulin (UMOD) which protects against urinary pathogens (Brunati et al. 2015; Bokhove et al. 2016; Devuyst and Pattaro 2018) and tectorin alpha (TECTA) which function in inner ear organization (Bokhove et al. 2016; Kim et al. 2019).

While both ZP-N and ZP-C are immunoglobulin-like domains with a core β-sandwich (Bokhove and Jovine 2018), they are evolutionarily distinct domains that have low amino acid sequence identity, unique disulfide patterns, and variable loop structures (Lin et al. 2011). Independent ZP-C domains outside of the ZP module have been identified in *C. elegans* (Weadick 2020), and four of the egg coat proteins (ZP1, ZP2, ZP4, and ZPAX) contain additional ZP-N domains independent of the ZP-N/ZP-C pair in the ZP module (Figure 2.1A). We do not know of non-reproductive proteins that contain duplicated ZP-N domains. We refer to ZP-N domains in the ZP module as “modular” and the N-terminal repeats as “free” domains. As ZP-N domains can form asymmetric dimers through their β-sandwich edges (Jovine et al. 2002; Bokhove and Jovine 2018; Litscher and Wassarman 2020), they have been considered the major driver of ZP module polymerization. While free ZP-N domains may similarly function as polymerization units, recent structural studies support that they may have acquired novel functions: the free ZP-N domains of ZP1 form intermolecular cross-links important for egg coat structure (Nishimura et al. 2019), while N-terminal domains in ZP2 (Avella et al. 2013; Avella et al. 2014) and ZP4 (Dilimulati et al. 2022) have been implicated in sperm-egg binding. The functional diversification of duplicated ZP-N domains seems to play an important role in the evolution of species-specific interactions. Despite their functional significance, the evolutionary history of ZP-N domains within and between these many paralogous proteins has not been examined. Our combination of phylogenetic and machine learning approaches addresses how a

complex history of whole gene and tandem domain duplications followed by structural adaptation produced the current diversity of ZP proteins.

2.2 Results and Discussion

2.2.1 Free and Modular ZP-N domains are phylogenetically distinct.

We investigated the evolutionary history of vertebrate ZP-N domains by extracting a total of 2405 ZP-N domain sequences from ZP module containing genes of 247 species with both reproductive (*zp1*, *zp2*, *zp3*, *zp4*, *zpax*, *zpd*) and non-reproductive (*umod*, *tecta*, *cuzd1*) functions (Table S2.1). While modular and free ZP-N sequences were found to share little sequence identity beyond 4 conserved cysteine residues that form stabilizing disulfide bonds, both domain types are highly similar in three-dimensional structure (Fig 2.1A). As such, we used a structure-based sequence alignment (Pei et al. 2008) to perform phylogenetic analysis. Maximum likelihood-based phylogenies indicated that the free ZP-N domains form a single clade distinct from the ZP-C associated modular ZP-N domains (Fig 2.1B), and this separation was robust to amino acid substitution matrices (LG, WAG, and JTT) (Fig S2.1). The topology of the modular ZP-N clade was broadly consistent with previously published gene trees based on the complete ZP module with both ZP-N and ZP-C (Claw and Swanson 2012; Feng et al. 2018). The topology of the free ZP-N clade supports that the initial duplication gave rise to the first repeat of the tandem array shared by ZP1, ZP2, ZP4, and ZPAX, which was followed by lineage specific repeat expansions of free ZP-Ns in *ZP2* and *ZPAX* (Fig 2.1C). Free ZP-Ns have only been identified in proteins associated with the egg coat.

2.2.2 Machine learning reveals conserved residues and structural features in modular ZP-N domains.

The phylogenetic separation of modular and free ZP-N domains using a structure-based alignment suggests important structural differences between the two domain types, but their high sequence divergence complicated manual identification of such characteristics. Machine learning methods have been applied to various aspects of protein biology such as function prediction (Yang et al. 2018; Bonetta and Valentino 2020) and the classification of membrane bound proteins (Guo et al. 2019). Here, we used a machine learning-based classification strategy to identify what structural features distinguish free and modular types of ZP-N domains. We applied a logistic regression model to the structurally aligned ZP-N domain sequences, where the probability of being a modular vs free ZP-N type was estimated for each of the 20 amino acids at each position in the alignment. Given the large number of parameters in this model (9321), we combined elastic net regularization and cross-validation to identify the most parsimonious model (i.e. the fewest non-zero parameters) within the 95% confidence interval of the highest scoring model (Fig 2.2A-B). Through this regularization strategy, we identified 8 modular-associated and 2 free-associated residues that were sufficient to predict whether a given ZP-N sequence was modular or free with 100% accuracy (Fig 2.2B). The greater number of modular-associated residues and their greater probabilistic weight suggests greater sequence conservation of modular domains (Fig 2.2B). Further examination of individual clades of modular and free ZP-Ns demonstrate the substantial sequence conservation of our residues identified by machine learning (Fig 2.2C).

Examination of the residues associated with either ZP-N type in the context of three-dimensional structures suggest differences in both function and quaternary structural dynamics. ZP-N monomers have an immunoglobulin-like β-sandwich fold with the 4- and 3-membered β-strands connected by a disulfide bridge on each edge of the molecule. Biochemical and

crystallographic studies support that modular ZP-N domains form asymmetric dimers through the molecular edge that includes the most N- and C-terminal β-strands (Jovin(Jovine et al. 2006)e, et al. 2006; Bokhove, et al. 2016). Free ZP-N domains do not appear to dimerize through this N/C-terminal edge, and have experienced functional diversification of the outer edge of the molecule to perform additional protein binding functions (Raj, et al. 2017; Nishimura, et al. 2019). When the modular-associated sites were mapped onto their respective structures, we observed that modular-associated residues form an integrated network of mostly hydrophobic stabilizing contacts that interlock between the β-sheets around the outer edge of the molecules (Fig 2.2D, Fig S2.2). The phylogenetic clustering of free ZP-N domains (Fig 2.1C), along with molecular dynamics support the loss of dimerization activity along the free ZP-N lineage, which could have facilitated their evolution of new binding partners (Fig S2.3). The stabilizing contacts along the outer edge of the modular ZP-N domains are consistent with these domains principally having structural roles, while in free domains this edge has diversified to allow functional innovation. Further subdivision of free ZP-N domains by their major clades (the first repeat versus internal repeats in ZP2 and ZPAX) largely supports our initial findings (Fig S2.4). Consequently, our sequence-based machine learning classifier identified conserved residues underlying structural differences between the two domain types that have implications on their respective functions.

2.2.3 Free ZP-N domains display signatures of sequence divergence and rapid evolution.

The difference in relative conservation of modular domain structures motivated additional analysis of the sequence evolution of these ZP-N domains. Here we focused on mammalian ZP genes (*zp1*, *zp2*, *zp3* ,*zp4*, *umod*, *tecta*, and *cuzd1*) due to both higher genomic assembly quality and to avoid synonymous substitution saturation that may occur when considering greater phylogenetic breadth (Anisimova and Liberles 2012). Measures of

sequence diversity within and between ZP-N groups reveal that modular domains are less diverse overall, and that free ZP-Ns are just as dissimilar to one another as they are to modular domains (Fig 2.3A).

These findings motivated molecular evolutionary analyses on 12 mammalian ZP-N domains, and only ZP2-N1 and ZP2-N2 showed evidence of positive selection (Table S2.2). These are notably the two domains with the lowest within group similarity (diagonal of Fig 2.3A). Positively selected sites in ZP2-N1 are far from the homodimerization edge and physically closer to the network of modular biased residues (Fig 2.3B). Analyses that detect positive selection in free ZP-Ns may reveal complementary information to the high conservation of residues in modular ZP-Ns. Protein regions associated with structural stability in modular domains may rapidly evolve in free domains that gain a new binding interface. Positively selected sites also constituted a substantial portion of the solvent exposed surface area (34% in ZP2-N1 and 24% in ZP2-N2), potentially facilitating their evolution of novel functions and protein interactions. The rapid evolution of ZP2-N1 is consistent with its role in species-specific sperm recognition (Avella, et al. 2014) and may reflect sexual co-evolution with its sperm receptor (whose identity is currently unknown). Remarkably, these positively selected sites cluster near a region associated with species-specific sperm protein binding in free invertebrate ZP-N domains (Raj et al. 2017). However, based on expansion and retraction of loop lengths outside the core β-sandwich, we believe that these invertebrate free ZP-N domains evolved independently of the free ZP-N domains of vertebrates, suggesting that the expansion of ZP-N arrays for species-specific sperm recognition is a convergent phenomenon that has arisen multiple times throughout metazoan evolution. Previously, positive selection in ZP3 had been detected in certain mammalian clades (W. Swanson et al. 2001; Turner and Hoekstra 2006). However, selective pressures related to reproduction (eg, mating system) can vary across taxa, and this can affect why a broad search of Boreoeutherian mammals did not detect positive selection in

ZP3 when compared to previous analyses on limited taxon datasets. Similar observations have been made in the fertilization proteins Izumo and Juno (Grayson 2015).

In summary, our combined phylogenetic, machine learning classification, and positive selection analyses illustrated a clear distinction between modular and free ZP-N domains. These two classes of domains experienced different evolutionary trajectories, as modular ZP-Ns likely retained a conserved structural role while free ZP-Ns neofunctionalized to serve different reproductive functions. These findings are of relevance to the evolution of species-specificity in fertilization, as the ZP-N domain expansion of ZP2 provided substrates to evolve novel species-specific interactions. Structural changes within free ZP-Ns could result in of a dimerization edge and the evolution of a new sperm binding loop. As these domains are co-opted into a reproductive context, co-evolution (Clark et al. 2009; Hart et al. 2018) and sexual conflict (Gavrilets and Waxman 2002) with sperm proteins could contribute to their rapid evolution. This reflects the evolutionary dynamics that drive structural diversification and neofunctionalization of duplicated domains. Our combined phylogenetic and machine learning approach outlined here can be applied to other essential gene families with complex duplication histories.

2.3 Materials and Methods

2.3.1 Multiple Sequence Alignment

Sequences for multiple ZP-N containing proteins were curated from the Ensembl database (release 104) (Howe et al. 2021). Sequences were preliminarily labelled as one of the ZP genes of interest based on PSI-BLAST e-value scores (Altschul et al. 1997). Sets of orthologous genes were aligned with MAFFT (Katoh and Standley 2013) and then trimmed to individual ZP-N domains. Groups of orthologous ZP-N domains were deemed “orthogroups”. Sequences with ambiguous characters were removed, and then sets of orthologous ZP-N sequences were realigned with MAFFT. A full multiple sequence alignment was generated by

concatenating orthogroup alignments together using a representative paralog alignment: individual representative sequences were selected from each orthgroup, and aligned using the structural based PROMALS tool (Pei et al. 2008). This approach was used because of the low sequence identity, but high structural similarity between paralogous Z-N domains. A custom script was used to algorithmically add gaps to orthogroup alignments to form a full multiple sequence alignment. For phylogenetics, CD-Hit was used to remove highly cluster highly similar sequences (>90% identity) (Li and Godzik 2006; Fu et al. 2012), in order to improve computing speed, and because this study was not concerned with very recent evolutionary splits. A full dataset was used for machine learning training because those methods are less computationally strained by large alignments and can gain greater sensitivity with a high depth of taxonomic sampling.

2.3.2 Phylogenetics

Maximum likelihood phylogenies were built using RAxML-NG(Kozlov et al. 2019), and multiple different amino acid substitution matrices were tested (LG+G, JTT+G,WAG+G), to evaluate the robustness of the deepest phylogenetic divide. The maximum likelihood tree was selected from 100 replicate runs using different starting trees. Nodal support was calculated with transfer bootstrap expectation (Lemoine et al. 2018), a modified form of bootstrapping that is more effective at detecting deep phylogenetic relationships in datasets with large number of taxa. Sequence labels were initially based on BLAST results but later refined based on phylogenetic clustering, (e.g. ZP1-N1, ZP2-N1, ZP4-N1).

2.3.3 Machine Learning

A basic machine learning algorithm using mean squared regression and regularization was coded in Python to distinguish the two free and modular groups of ZP-N domains. Logistic regression models are well suited for these classifications, because their outputs are bounded

between 0 and 1, which can be interpreted as probability that a given domain is modular (Bewick et al. 2005). The multiple sequence alignment was identical to that used for phylogenetic analysis. The alignment was split into a testing (25%) and training set (75%), and logistic regression modelling with cross-validation was performed on the training set using five-way cross validation. The final model scores were based on performance in the testing dataset.

For machine learning analysis, aligned ZP-N sequences were one-hot encoded: each position in the sequence was converted into a vector of twenty digits, corresponding to the twenty amino acids. The value was set to 1 for the entry in the vector corresponding to that residue, and all other values are set to 0. Gapped sites were set to a vector of twenty 0's. Thus, the classifier was trained using (1+20n) features (there is an additional intercept term), where n is the alignment length. Each of these features has a parameter associated with it and the value of the parameter indicates how informative that feature is, and whether it supports a modular ZP-N or free ZP-N classification. There are a large number of possible parameters in this model (9321 including the intercept), which introduces a risk for "overfitting"(Hawkins 2004), and motivates our regularization strategies.

To determine the minimal number of highly informative parameters, elastic net regularization was employed to penalize overparameterization and reduce overfitting (Zou and Hastie 2005). In our sci-kit learn implementation (Pedregosa et al. 2011), both the strength of regularization and the L1/L2 penalty ratio between the two penalty types were optimized by grid search. The highest scoring model was identified according to the negative mean-squared error scoring metric. In order to choose a suitable sparse model (i.e. fewest non-zero parameters), we adapted the one standard error rule common in machine learning (Hastie et al. 2009), where the sparsest model that is still within one standard error of the highest scoring model is selected. For this analysis we used 95% confidence intervals (~1.96 standard errors), to identify the sparsest model (fewest non-zero parameters) that is not statistically different from the highest

scoring model sampled. Raw parameter values were plotted in the style of sequence LOGO plots (Schneider and Stephens 1990). The sum of the raw parameter values for matching amino acids in the alignment (and the intercept term), are equivalent to the log odds that a given sequence is classified as modular. For simplicity, each parameter is described as the log odds associated with a particular residue. In addition to the initial binary classification (free vs modular), our analysis was repeated using a three-way multiclassification (first N-terminal, internal, and modular). This procedure used alignments, hyperparameter grid searching, and regularization strategies in the same manner as the binary classification.

2.3.4 Sequence Divergence and Positive Selection Analyses

Our analyses of sequence divergence and positive selection was performed on a set of Boreoeutherian mammals, and we used the mammalian ZP-N domains coming from *zp1*, *zp2*, *zp3*, *zp4*, *umod*, *tecta*, and *cuzd1*. Boreoeutherian sequences were mined from Ensembl (Howe et al. 2021). Sequences from a given Boreoeutherian species were included in these analyses, if 10 or more of the ZP-N domains from our phylogenetic analyses were represented in that species. Phylogenetic distances both within and between orthogroups were calculated in MEGA using Poisson estimation with a gamma distribution of variation between sites (Kumar et al. 2016; Kumar et al. 2018).

Evidence of positive selection was measured using PAML analyses (Yang et al. 2005; Yang 2007) on the same sets of ZP-N domains from the sequence divergence estimation. A likelihood ratio test between a model allowing positive selection (M8) and a neutral model (M8a), was used to determine which domains showed evidence of positive selection. Likelihood ratio tests were performed by comparing M8 and M8a, using a chi-squared distribution with one degree of freedom. We also performed a Benjamini-Hochberg p-value correction to account for multiple testing (Benjamini and Hochberg 1995). Positively selected sites were visualized on a published crystal structure (ZP2-N1) (Raj et al. 2017), or the alpha-fold predicted structure

(Jumper et al. 2021) when this did not exist (ZP2-N2). Sites were labelled if they had a posterior probability of being positively selected > 75% according Bayes Empirical Bayesian (BEB) analysis.

2.3.5 Visualization and Other methods

When protein structures were not available Alpha-Fold2 tertiary structure prediction was used (Jumper et al. 2021), and three-dimensional protein structures were visualized using either pymol (Schrödinger 2015) or ChimeraX (Pettersen et al. 2004). Docking simulations of homodimerization for ZP2-N1 and ZP3-N were performed using Rosetta 3.5 (Chaudhury and Gray 2008; Sircar et al. 2010). Briefly, each template structure was energy minimized in Rosetta using the relax function, each structure was duplicated, aligned to the dimeric ZP-N structure of uromodulin (PDB 4wrn), 10000 independent docking simulations performed, and interface scores analyzed for the top 5% lowest energy structures.

2.3.6 Data Availability

We are sharing a link to a github repository that contains our maximum likelihood phylogeny and relevant alignments and code. The repository link is https://github.com/amrivera526/ZPN_Evolution

2.4 Acknowledgments

This research was funded by the following NIH grants: R21HD105025 awarded to WJS and K99HD090201 awarded to DBW. We also thank fellow lab members Jolie Carlisle and Jan Aagaard for participating in scientific discussions.

2.5 Figures

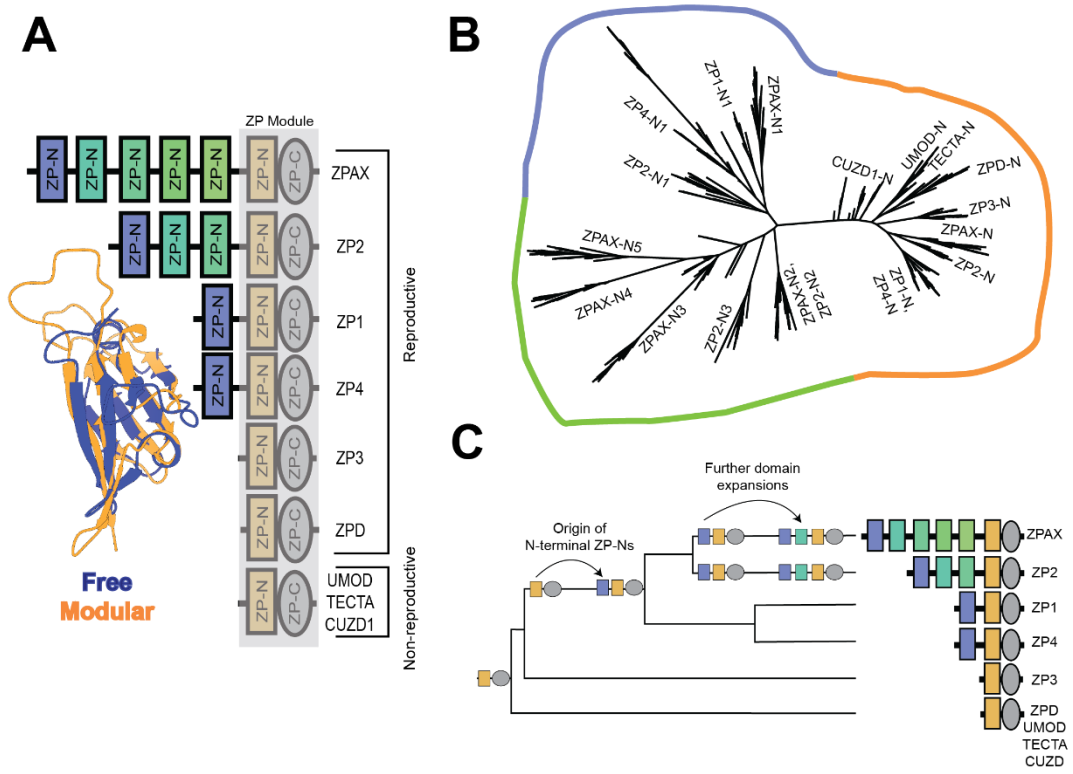


Figure *2.1*: Phylogenetic analysis of ZP-N domain duplication history. A) A structural alignment of Mouse ZP2-N1 and ZP3-N highlights the broad structural conservation of these two classes of ZP-N domains (RMSD = ~4.7 Å) despite only ~18% amino acid sequence identity. The protein schematics summarize the ZP proteins included in this analysis. B) Phylogenetic analysis (Kozlov et al. 2019) of ZP-N sequences (shown as a maximum likelihood tree) supports an ancestral separation between free and modular ZP-N domains (~78% support). C) A summary of ZP-N domain evolution based on the gene tree in panel B. The ancestral protein contained a ZP module with a C-terminal ZP-N and ZP-C domains, and duplication of the ZP-N produced the most N-terminal domain found in ZP1, ZP4, ZP2, and ZPAX. Later duplication events within ZP2 and ZPAX gave rise to multiple additional ZP-N domains between ZP-N1 and the ZP module.

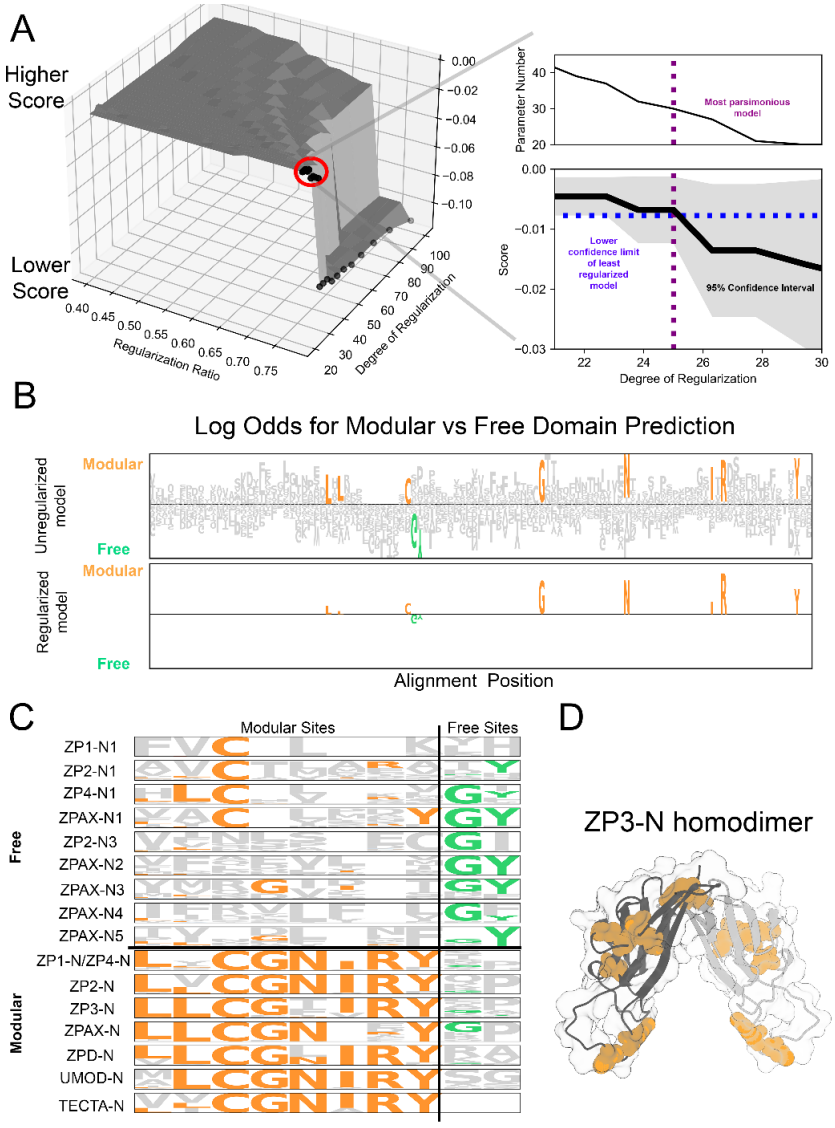


Figure *2.2*: Machine learning based inference of sequence features that distinguish modular and free ZP-N domains. A logistic regression model with elastic net regularization was trained on the ZP-N multiple sequence alignment generated as part of the phylogenetic analysis, with the data partitioned for training and testing (75% and 25%, respectively), with 5-way cross validation of the training data employed to estimate the error distribution of the score function. We defined our optimal model as the most parsimonious model (*i.e.,* the fewest parameters) within the estimated 95% confidence interval of the unregularized model. (A) The space of regularization hyperparameters was explored during model optimization, plotted as a 3D surface (left). The score is the negative mean squared error, and dots correspond to the two-dimensional cross section shown on the right, with the blue line denoting the intersection between the lower confidence limit of the unregularized model to its intersection with the score as a function of regularization strength. B) Comparison of the unregularized and optimal logistic regression models as LOGO plots with the height of each amino acid at each position corresponding to its parameter weight, with colored amino acids denoting parameters retained in the regularized model (orange for modular, green for free). Each parameter weight approximating the logs odd ratio for a modular domain prediction, when a residue is present at that position. C) Sequence LOGOs were constructed for individual clades within the phylogeny. They emphasize the conservation of residues within the modular ZP-N clade. There is also greater conservation of a characteristic ZP-N disulfide bond in the most N-terminal ZP-Ns compared to other free domains. D) Mapping highly predictive sites onto ZP-N protein models suggest differences in structural properties between free and modular domain. The available crystal structure ZP3-N (3d4c) was used and modelled as a dimer for spatial context. Modular-associated sites are generally buried along the outer edge of the homodimer.

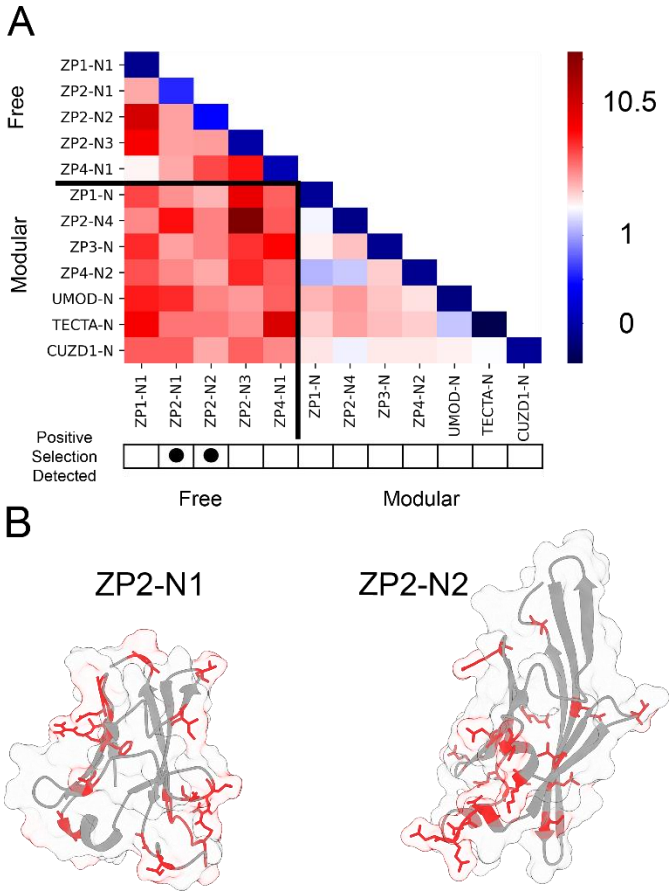


Figure *2.3*: Amino acid diversity and tests of positive selection in modular and free ZP-N domains. A) A heatmap showing the within group and between group mean phylogenetic distances for orthologous groups of ZP-N domains (Kumar et al. 2018). B) Positively selected sites in mammalian ZP2-N1 and ZP2-N2 were identified through maximum likelihood analysis and mapped onto protein models (4wrn for ZP2-N1, and an AlphaFold prediction for ZP2-N2) (Yang 2007).

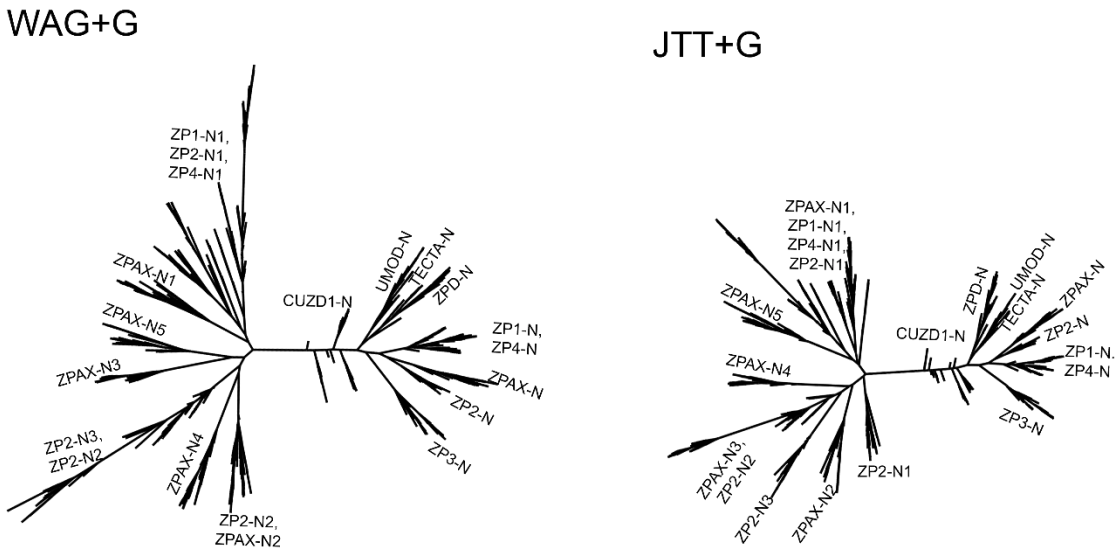
2.6 Supplemental Tables and Figures

	Mammals	Birds	Reptiles	Amphibians	Fish
ZP1-N1	50 (30)	28 (11)	12 (10)	0 (1)	0 (19)
ZP1-N	53 (30)	28 (9)	11 (8)	0 (0)	28 (20)
ZP2-N1	70 (46)	35 (20)	8 (4)	2 (2)	36 (34)
ZP2-N2	74 (46)	44 (19)	14 (10)	1 (1)	19 (21)
ZP2-N3	75 (45)	41 (18)	15 (11)	2 (2)	3 (3)
ZP2-N	76 (38)	40 (7)	14 (10)	1 (1)	15 (14)
ZP3-N	72 (36)	40 (13)	19 (12)	2 (2)	53 (38)
ZP4-N1	63 (37)	33 (11)	12 (10)	2 (2)	34 (23)
ZP4-N	52 (29)	35 (8)	12 (8)	1 (1)	61 (42)
ZPAX-N1	2 (2)	22 (9)	12 (11)	2 (2)	67 (50)
ZPAX-N2	2 (2)	24 (9)	13 (9)	2 (2)	67 (53)
ZPAX-N3	2 (2)	25 (9)	12 (9)	2 (2)	64 (51)
ZPAX-N4	2 (2)	20 (8)	12 (7)	2 (2)	69 (56)
ZPAX-N5	2 (2)	24 (12)	15 (11)	2 (2)	68 (55)
ZPAX-N	2 (2)	13 (5)	8 (8)	2 (2)	59 (44)
ZPD-N	0 (0)	38 (9)	18 (10)	2 (2)	69 (54)
UMOD-N	75 (41)	0 (0)	6 (6)	1 (1)	8 (6)
TECTA-N	43 (1)	4 (1)	11 (0)	1 (0)	49 (2)
CUZD1-N	62 (38)	44 (23)	17 (13)	2 (2)	51 (37)

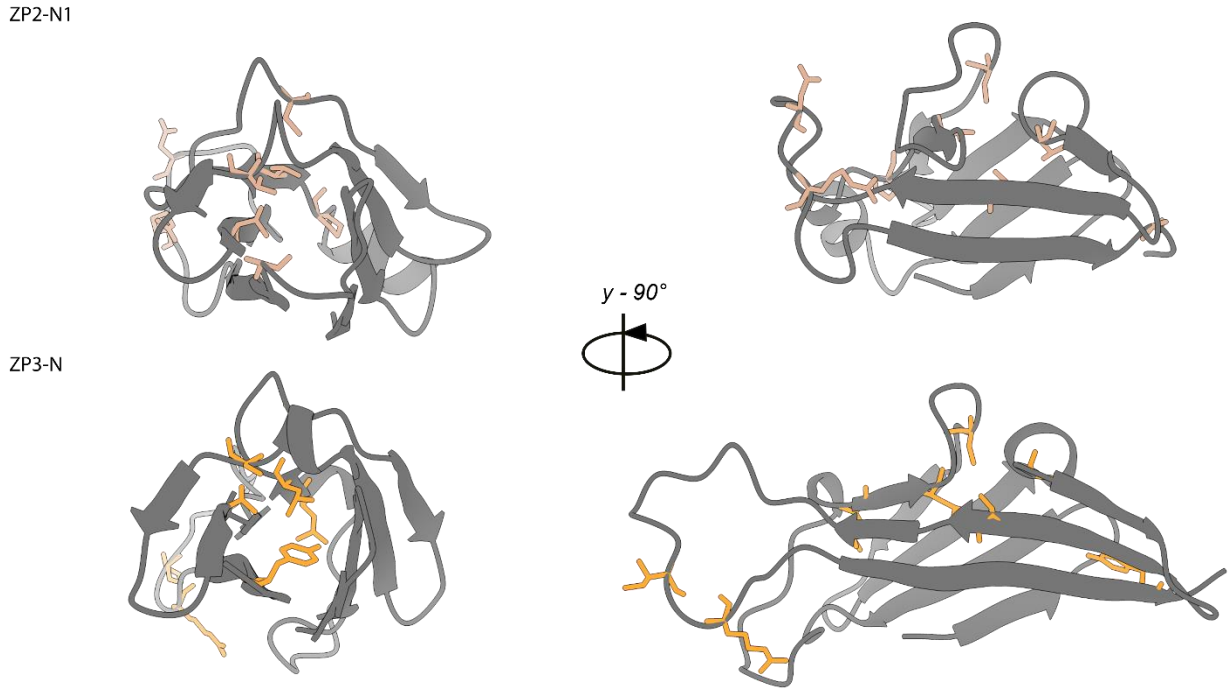
Supplemental Table 2.1: Summary of species sampled in phylogeny. This table summarizes the sequences included for the machine learning classification and phylogenetic analysis (these values are in parenthesis). There were 2405 ZP-N sequences across 247 species included in the machine learning analysis and the phylogenetics included 1488 ZP-N sequences from 210 species. The phylogeny was filtered for sequences at greater than 90% identity, unlike the machine learning dataset and there are additional differences due to manual filtering of alignments. Here the amphibians only included frogs due to genomic availability reasons. The fish class includes all non-tetrapod vertebrates, and is non-monophyletic. All labels in this table are based on BLAST results Labels in the final phylogeny are based on phylogenetic clustering.

Domain	M8a	M8	-2ΔlogL	*p* value	*p* value corrected
CUZD1-N	p_0 = 0.97, p_1= 0.03, p = 0.64, q=2.1, ω = 1	p_0 = 0.99, p_1= 0.01, p = 0.63, q=1.9, ω = 2.5	2.4	0.06	0.25
TECTA-N	p_0 = 0.98, p_1= 0.02, p = 0.05, q=0.99, ω = 1	p_0 = 0.99, p_1= 0.01, p = 0.05, q=0.97, ω = 1.9	1.0	0.16	0.24
UMOD-N	p_0 = 0.98, p_1= 0.02, p = 0.29, q=0.61, ω = 1	p_0 = 0.98, p_1= 0.02, p = 0.29, q=0.61, ω = 1	0.0	0.50	0.57
ZP1-N1	p_0 = 0.74, p_1= 0.26, p = 0.74, q=1.4, ω = 1	p_0 = 0.99, p_1= 0.01, p = 0.54, q=0.51, ω = 3.5	1.9	0.08	0.19
ZP1-N2	p_0 = 1, p_1= 0, p = 0.46, q=0.68, ω = 1	p_0 = 1, p_1= 0, p = 0.46, q=0.68, ω = 1	0.0	0.50	0.57
ZP2-N1	p_0 = 0.43, p_1= 0.57, p = 2.7, q=8.0, ω = 1	p_0 = 0.65, p_1= 0.35, p = 1.2, q=1.4, ω = 1.5	6.7	4.7E-03	0.03 *
ZP2-N2	p_0 = 0.42, p_1= 0.58, p = 29.5, q=99, ω = 1	p_0 = 0.70, p_1= 0.30, p = 1.6, q=1.5, ω = 1.9	18.8	7.2E-06	8.6E-5 *
ZP2-N3	p_0 = 0.67, p_1= 0.33, p = 1.6, q=7.9, ω = 1	p_0 = 0.85, p_1= 0.15, p = 0.65, q=1.5, ω = 1.4	1.5	0.11	0.22
ZP2-N	p_0 = 0.87, p_1= 0.13, p = 0.76, q=3.4, ω = 1	p_0 = 0.87, p_1= 0.13, p = 0.76, q=3.4, ω = 1	0.0	0.50	0.57
ZP3-N	p_0 = 0.98, p_1= 0.02, p = 0.55, q=1.2, ω = 1	p_0 = 1, p_1= 0, p = 0.52, q=1.21 ω = 2.3	0.0	0.50	0.57
ZP4-N1	p_0 = 0.58, p_1= 0.42, p = 1.3, q=4.3, ω = 1	p_0 = 0.93, p_1= 0.07, p = 0.64, q=0.70, ω = 1.9	2.3	0.065	0.19
ZP4-N	p_0 = 0.82, p_1= 0.18, p = 0.40, q=1.1, ω = 1	p_0 = 0.94, p_1= 0.06, p = 0.37, q=0.67, ω = 1.5	1.1	0.15	0.26

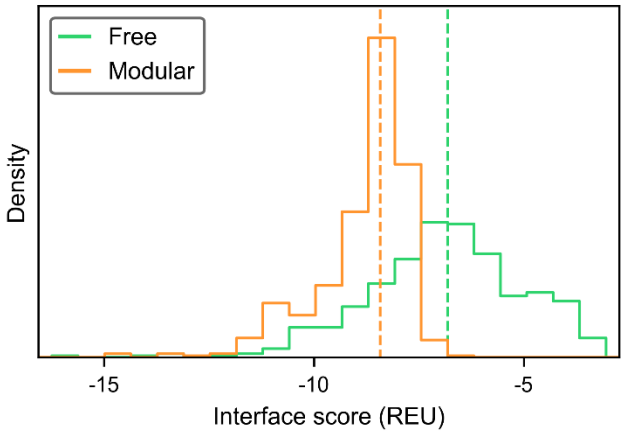
Supplemental Table 2.2: Summary of PAML output. This table summarizes the results from PAML analysis (Yang 2007). Here we compared a neutral model (M8a) to a model that allows positive selection (M8). A * denotes statistically significant p values after Benjamini-Hochberg multiple testing correction (Benjamini and Hochberg 1995).



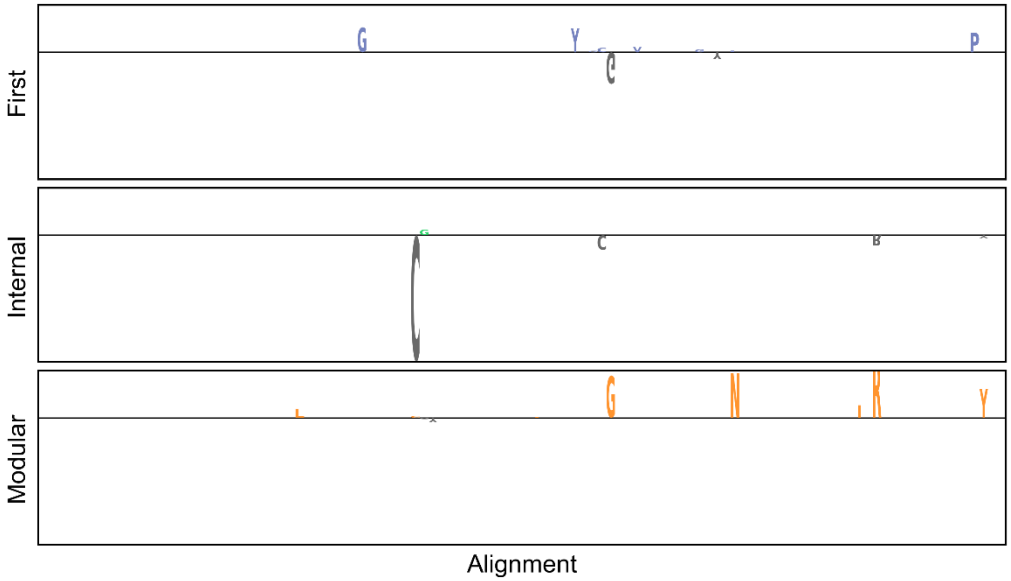
Supplemental Figure 2.1: Phylogenies from alternative substitution matrices. These two trees were produced through RAxML-NG (Kozlov et al. 2019). Major aspects of the tree are conserved, specifically the monophyletic grouping of free ZP-N domains.



Supplemental Figure 2.2: Modular-biased residues in the context of domain tertiary structures. This figure shows the modular biased sites in the context of both ZP2-N1 (Raj et al. 2017) and ZP3-N (Monne et al. 2008) crystal structures. The brighter orange corresponds to the modular-biased sites according to our machine learning model, while the duller orange on ZP2-N1 represent the structural homolog of those sites according to a pymol (Schrödinger 2015) structural alignment. We observe a clustering of these sites within the core of the domain in ZP3-N and not ZP2-N1. The two positions of each are 90° rotations around the y-axis.



Supplemental Figure 2.3: Rosetta docking simulation of dimerization. This is a histogram of interface scores from rosetta docking simulations of dimers for both free (ZP2-N1) and modular (ZP3-N) ZP-N domains.



Supplemental Figure 2.4: Multiclass machine-learning ZP-N domain classification. Since we observed monophyletic grouping of the most N-terminal ZP-N domain, we performed a multiclass variation of our machine learning analysis. The data was split into three classes: modular, first (*i.e.,* most N-terminal free domain), and internal domains (all other free ZP-N domains). In this multiclass analysis the model is fit three times, each time producing a classifier that distinguishes one of the classes from the rest of the data. The first row is the unregularized model from our two-class analysis, and our regularized multiclass models are summarized in the other three rows, where positive values suggest a bias towards that class. Our modular classifier mostly recapitulates our earlier results, because it is in essence still comparing modular ZP-Ns versus all free ZP-Ns. The first ZP-N domain recapitulates the two free associated residues and has a few amino acids associated with it with relatively low parameter values. The internal ZP-N domains seem to have a bias against the cys 2-3 bond, but that likely is a reflection of the conservation of the second cysteine in the first ZP-Ns and the third cysteine in modular domains

Chapter 3: A transcriptomic and comparative genomic investigation of abalone ovaries

Identifying potential functional ovary genes provides insight into the molecules mediating fertilization. Previous abalone sperm transcriptomic work demonstrated that high throughput genomic approaches can identify previously unknown essential fertilization genes. We leveraged advancements in sequencing technology, such as PacBio Isoseq, which has been used to identify new genes and genetic variants in other species. The transcriptome reads clustered into 30,579 putative genes. To focus on the most likely candidate genes, we looked at three features: 1) likely sperm accessibility as determined by presence of a signal peptide or transmembrane domain, 2) homology to known fertilization genes, and 3) signatures of positive selection. We identified 4 new abalone ZPs and 5 new abalone TFPs. The ZPs show evidence of domain duplication alongside other examples of structural divergence, such as the absence of a transmembrane domain. Phylogenetic analyses suggest that ZP-N domain expansion in fertilization proteins occurred independently in vertebrates and invertebrates. Two newly identified tentative abalone TFPs have more cysteines (12) than previously described TFPs, which have 8 or 10 cysteines. TFPs have been co-opted multiple times for fertilization and these abalone TFPs could represent an additional examples of lineage specific structural modification and neofunctionalization. Positive selection analyses were used to further identify potential candidate proteins, including a putative lipoprotein receptor homolog. Many candidates described in this chapter warrant further functional characterization with the goal of identifying new fertilization proteins.

3.1 Introduction

The molecular interactions between sperm and egg proteins are essential to life, yet the identification of interacting fertilization remains understudied (Herberg et al. 2018). Despite essential proteins often experiencing purifying selection, proteins related to fertilization and

reproduction are some of the most rapidly evolving in the genome (Swanson and Vacquier 2002; Wilburn and Swanson 2016). This has motivated extensive research into the evolutionary dynamics of reproductive proteins. Specific molecular interactions between gametes may underlie hybridization barriers and reproductive isolation, illustrating how fundamental fertilization research is to our understanding of how populations evolve and speciate.

Marine invertebrates have historically been attractive models of fertilization (Loeb 1915; Lillie 1919; Lewis et al. 1982). Molluscs such as abalone, are broadcast spawners that release billions of gametes (Leighton and Lewis 1982) providing large quantities of gametes that can be purified and isolated for biochemical analyses. The overall large-scale steps of the fertilization cascade (Vacquier 1998), such as sperm-egg fusion and egg coat dissolution are similar processes even in species distant as humans and abalone (Wilburn and Swanson 2016). This may be surprising given how rapidly fertilization proteins evolve and diverge (Swanson and Vacquier 2002). Abalone is also a broadly relevant fertilization model that illustrates how molecular processes can be conserved with rapid divergence of the underlying molecular mechanisms. While fertilization proteins may differ greatly in amino acid sequences among distant species, conservation of protein structure could also contribute to conservation of large-scale fertilization processes.

Egg coats, or vitelline envelopes in abalone, consist of glycoproteins which form a complex extracellular matrix (Rankin et al. 2003). The arrangement of proteins within the matrix, or its supramolecular structure, is likely essential to its function. Abalone egg coat genes represent a compelling example of widespread whole gene duplication and internal tandem domain duplication of the ZP-N domain. Abalone have approximately 30 identified egg coat genes that contain a ZP module (Aagaard et al. 2010), which are known as VEZPs (Vitelline Envelope ZP containing proteins). Some VEZP genes have duplications of ZP-N domains, including VERL (Vitelline Envelope Receptor for Lysin) which has a remarkable 23 ZP-N

domains (Galindo et al. 2002).. Considering how rapidly fertilization proteins evolve, this extensive gene duplication and ZP-N domain expansion may have provided opportunities for the birth and neofunctionalization of new egg coat proteins (Rivera et al. 2022).

More broadly, multiple aspects of abalone reproductive biology have been studied extensively. In particular, abalone have been a historic model for species-specific sperm-egg interactions (Vacquier et al. 1990; Raj et al. 2017). Remarkably, the mating ranges and habitats of eight species of Pacific abalone overlap, yet distinct species remain in close proximity (Galindo et al. 2002). Hybridization can be achieved under certain conditions (Leighton and Lewis 1982), which raises questions regarding the molecular mechanisms of species-specificity. Are hybrids viable in the wild? And if so, what are the molecular mechanisms that maintain clear species boundaries, in spite of this hybrid viability? Research has focused on how the affinity and dynamics of interactions of protein interactions involving VERL ZP-N repeats (Raj et al. 2017; Carlisle and Swanson 2021) affect species-specific fertilization. Indeed, abalone ZP-N domains show clear species-specific binding preferences in *in vitro* assays (Raj et al. 2017). This observed species-specific binding preference could contribute to hybridization barriers in Pacific abalone.

Repeat expansions of ZP-N domains may relate to the function of lysin-VERL interactions in species-specific fertilization. Increasing the number of ZP-N domains could increase the specificity of sperm-egg interactions by requiring more successful lysin-VERL binding events. Lysin is one of the most abundant abalone sperm proteins (Lewis et al. 1982; Kresge et al. 2000) which could reflect a co-evolutionary arms race with the large number of ZP-N domains in VERL. It has been hypothesized that abalone sperm and the egg surface may be in sexual conflict, which is when the sexes evolve with conflicting fitness benefits (Kokko and Jennions 2014). An egg may evolve in a way which reduces sperm binding ability, because fusion with multiple sperm (polyspermy) does not result in a viable zygote (Gilbert 2000) and

reducing sperm binding ability may allow time for activation of blocks to polyspermy. The sperm may then evolve to improve egg binding ability, resulting in a co-evolutionary chase between lysin and VERL (Rankin et al. 2003; Avella et al. 2013; Nishimura et al. 2019). Mathematical models predict that these co-evolutionary dynamics contribute to the rapid evolution of fertilization proteins, maintenance of reproductive barriers, and eventual speciation (Gavrilets and Waxman 2002).

It is important to note that other ZP proteins have not been researched as extensively as VERL. To date, only a few other ZP-N domain-containing proteins have attracted research interest, including abalone VEZP14, because of its repeat expansion and negative co-evolutionary signature with sperm lysin (Aagaard et al. 2013). Investigations into other ZP proteins implicated multiple vertebrate ZPs in species-specific sperm-egg interactions and the supramolecular structure of the egg coat (Rankin et al. 2003; Okumura et al. 2004; Avella et al. 2013; Nishimura et al. 2019). The complexity of the egg coat motivates research into evolution and function of multiple ZP-N containing proteins. Research on the mammalian ZP family motivated ovary transcriptomic analysis to discover additional putative reproductive proteins in abalone. Abalone egg coat components are synthesized in the ovaries, unlike fish which synthesize their egg coats in the liver (Jovine et al. 2005; Litscher and Wassarman 2007; Sano et al. 2013; Wassarman and Litscher 2016). This research complements earlier red abalone testis transcriptomics (Palmer et al. 2013). One newly identified rapidly evolving sperm protein (Palmer et al. 2013) was later denoted Fuzzy Interaction Transient Zwitterions Anionic Partner (FITZAP) (Wilburn et al. 2019). FITZAP is a highly abundant rapidly evolving sperm protein that helps pack acrosomal proteins. Despite the abundance of FITZAP, it is difficult to elute and visualize on SDS-PAGE gels (Wilburn et al. 2019), demonstrating how genomic investigation can uncover components that are difficult to discover by biochemical approaches.

Long read cDNA sequencing technology facilitated genomic discoveries, such as identification of structural variants (Kronenberg et al. 2018), and even the completion of whole genomes (Nurk et al. 2022). We used long sequencing technology to conduct transcriptome sequencing with the goal of identifying ovary-expressed genes potentially important for sperm-egg interactions. Bioinformatics tools can identify features such as signal peptides (Almagro Armenteros et al. 2019) and transmembrane domains (Sonnhammer et al. 1998; Krogh et al. 2001; Möller et al. 2001; Tusnády and Simon 2001), which suggest that an ovary protein is sperm accessible. Signal peptides are part of a highly conserved pathway that allows proteins to be secreted (Holland 2004; Pohlschröder et al. 2004; Pohlschröder et al. 2005). Transmembrane domains, which are typically hydrophobic alpha helices that span the cell membrane (Shao and Hegde 2011; Guna and Hegde 2018), are a common feature of ZPs (Zona Pellucida Module Containing Proteins) (Gupta 2018), and these ZPs are integrated into the egg plasma membrane prior to cleavage and incorporation into the extracellular egg coat (Jovine et al. 2002). Comparative genomics between the transcriptome and abalone draft genomes allow for positive selection analyses as well, which can also be important for identifying which we used to identify rapidly evolving candidate fertilization genes.

Advancements in long read sequencing has improved our ability to discover potentially relevant ovary-expressed genes. The PacBio Isoseq approach outlined in this chapter was used to sequence transcripts more accurately and with higher throughput than strategies for gene discovery that relied on expressed sequence tags (Swanson et al. 2001). It is worth noting that newly discovered genes could encode proteins that function at any stage of the fertilization cascade. Properties such as likely sperm accessibility, rapid evolution, or homology to known fertilization proteins could help hint at the site of action and function of candidate fertilization proteins.

3.2 Results and Discussion

3.2.1 Overview of Ovary Transcriptome

Identifying functional fertilization proteins expressed in both sperm and egg can help us better understand the molecular mechanisms underlying fertilization. Previous abalone testis transcriptomic work that discovered an essential acrosomal protein in red abalone (Palmer et al. 2013), motivated a complementary analysis on ovaries. We performed PacBio Isoseq on cDNA from gravid red abalone ovaries, with the goal of identifying potentially functionally important fertilization proteins.

The transcriptome contained over 2.9 million reads, and included over 2.6 million full length non-concatenated reads (FLNC), which cluster into 30,579 predicted genes (Table 3.1). Because of the length of PacBio reads, a genome assembly is not required to generate the coding sequence of predicted genes and proteins. To obtain accurate sequences, long high-fidelity (HiFi) reads were obtained using a circular consensus sequencing (CCS) method that uses hairpin adaptors to pass over each cDNA multiple times. For example, experiments that optimized CCS in human genomic DNA produced HiFi reads with a predicted average per nucleotide accuracy of 99.8%. (Wenger et al. 2019). We performed transcriptome read clustering and *de novo* gene prediction using the isONclust program (Sahlin and Medvedev 2019). This analysis predicted at least 30,579 genes that are expressed in the red abalone ovary transcriptome, with 19,967 genes having more than one read supporting it.

From each of the transcript reads, we algorithmically determined longest non-overlapping reading frame and the predicted protein, and then assigned the best representative proteins and reads for each predicted gene cluster based on length and read count. Representative proteins were ranked based on a heuristic of log of read count multiplied by the protein length. This

metric was optimized to balance the importance of abundance and length. These representative proteins and reads were used in further downstream analyses.

3.2.2 Newly identified ZP proteins illustrate gene duplication and domain expansion

Since we are particularly interested in the evolution of ZPs in fertilization (Rivera et al. 2022), we sought to identify which of our abundant transcripts are likely ZPs. Homology searches against the SWISS Protein database (Bairoch and Apweiler 1996; The UniProt Consortium 2021) and a previous red abalone ovary express sequence tag (EST) library (Aagaard et al. 2010) identified multiple predicted novel VEZP (Vitelline Envelope ZP containing proteins) (Fig 3.1). We identified 9 tentative abalone ZP proteins in our top 100 most abundant gene clusters, according to transcript count (Fig 3.1A). This includes likely representatives of VEZP2, VEZP4, and VEZP8, as well as partial VERL sequences. Four of these abundant ZPs seem to be newly identified proteins with previously predicted ZP-N containing pseudogenes (ZPF, ZPS, ZPW, and ZPX) (Aagaard et al. 2010) (Table 3.2). As these ZPs did not have predicted coding homologs, they represent newly identified ZPs in red abalone ovaries. We refer to these proteins in this chapter according to their most similar pseudogene (e.g., ZPF-like protein 1). These newly identified genes are likely close paralogs of these ZP pseudogenes. This expands upon our knowledge of the pattern of duplication and gene loss thus far described for the family of ZP proteins (Killingbeck and Swanson 2018). There are transcribed ZPs that include a duplicated free ZP-N domain (Fig 3.2) reminiscent of abalone egg receptor VERL (Raj et al. 2017), mouse ZP2 (Avella et al. 2013; Avella et al. 2014), and bovine ZP4 (Dilimulati et al. 2022), all of which regulate species specific sperm-egg interaction (Avella et al. 2014). This further highlights the importance of whole gene duplication and domain expansion events in the evolution of this essential family of fertilization proteins (Rivera et al. 2022; Rivera and Swanson 2022). This history of duplication and potential gene loss also mirrors similar evolutionary processes in sperm acrosomal proteins (Carlisle et al. 2022).

Previous phylogenetic work analyzed the duplication and evolution of ZP-N domains in vertebrates (Rivera et al. 2022). Placing the newly predicted red abalone ZP module sequences in the context of other vertebrate ZP sequences obtained from Ensembl (Cunningham et al. 2022), can expand on our previous work and provide greater insight into the evolution of ZP across animalia. Our RAxML maximum likelihood phylogenies included all 48 of our predicted abalone transcriptome ZPs and vertebrate ZP module sequences from other vertebrate species (human, chimpanzee, stickleback, zebrafish, chicken, tropical tree frog, and anole) obtained from Ensembl (Cunningham et al. 2022). The topology of ZP module sequences largely supports previous ZP genes trees (Claw and Swanson 2012), as well as our ZP-N domain phylogenetics (Rivera et al. 2022). There is a phylogenetic separation between abalone ZPs from vertebrate ZPs (Fig 3.1B). This strongly suggests that any ZP-N domain expansions that we observed in abalone occurred independently from the repeat expansions that occurred in vertebrates (Rivera et al. 2022). Our unrooted phylogeny suggests that neither abalone nor vertebrate ZP modules are monophyletic (Fig 3.1B). This suggests the existence of an ancient divide between ZP3 and other gametic ZPs that predate the separation between vertebrates and invertebrates. There could be meaningful structural and functional differences between ZP proteins on different sides of this phylogenetic divide in both vertebrates and invertebrates.

Repeat expansions of ZP-N domains have been important in the evolution and functional diversification of vertebrate ZP proteins (Rivera et al. 2022). In this ZP phylogeny, vertebrate proteins with ZP-N domain expansions (ZP2 and ZPAX) do not cluster with abalone VERL. This illustrates an example of recurrent expansions of ZP-N domains in fertilization proteins across animalia. It is striking that this domain has been independently expanded at least twice in genes involved in species-specific fertilization. We suggest this represents a case of convergent evolution of morphological cell structures (egg coats) by duplication of similar protein domains.

Particular features of ZP-N biology and evolutionary history could have contributed to these independent expansion events. The trend of rapid evolution in reproduction proteins could be accompanied with multiple independent repeat expansions of ZP-N domains, as well as the neofunctionalization of these domains. The ability for ZP-Ns to act as polymerization domains (Callebaut et al. 2007; Wilburn and Swanson 2017) also provide opportunities of ZPs to evolve new binding partners during neofunctionalization. We observe the diversity of binding function in egg coat ZPs such as crosslinking (vertebrate ZP1) (Nishimura et al. 2019) and sperm binding (mammalian ZP2, bovine ZP4, and abalone VERL) (Avella et al. 2014: 4; Raj et al. 2017; Dilimulati et al. 2022). This diversity of ZP functions supports the notion that structural modifications and the evolution of new binding interactions may be important in egg coat evolution.

At least one ZP newly identified in this study, which we named ZPF-like protein 1 (ZPFL1) has a duplicated ZP-N domain that may be an example of an independent duplication and neofunctionalized domain (Fig 3.2). The AlphaFold2 (Jumper et al. 2021; Mirdita et al. 2022) predicted structure includes one duplicated ZP-N outside of the module. Surprisingly, the structure lacks a predicted transmembrane domain which is a feature common to known egg coat ZPs. Transmembrane domains are thought to target ZPs to the cell membrane, after which they are cleaved and free to incorporate into the extracellular matrix of the egg coat (Jovine et al. 2002; Gupta 2018). While there could be other unknown mechanisms for egg coat ZP-N incorporation, it is also possible that ZPFL1 is not a component of the abalone egg coat. If functional, it could potentially act a decoy to block polyspermy, by binding and sequestering excess sperm or sperm proteins (Wessel et al. 2001). This mechanism could be similar to Juno, a membrane protein which is shed after fertilization and acts as block to polyspermy (Bianchi and Wright 2014). Such functional diversity of ZPs would not be surprising, given the existence of other ZPs with non-reproductive functions (TECTA, UMOD, and CUZD1) (Brunati et al. 2015;

Bokhove et al. 2016; Devuyst and Pattaro 2018; Kim et al. 2019). While several ZPN-containing proteins are known to have essential roles in multiple biological systems, we do not know if ancestral ZP-N domains had a function related to fertilization.

3.2.3 Newly discovered abalone ovary TFPs show evidence of potential structural diversity.

Other fertilization protein families including the three finger protein superfamily (TFPs), have extensive histories of repeat domain expansion alongside protein structural and functional diversity (Doty et al. 2016; Rivera and Swanson 2022; Wilburn et al. 2022). To identify other potential fertilization proteins, the ovary EST library was searched using queries of well characterized reproductive proteins found in multiple model organisms and humans. Queries included Juno, Izumo, Bouncer, CD9, DCST1, DC-STAMP, Integrin beta-1 (Bianchi and Wright 2014; Frolikova et al. 2018: 9; Inoue, Hagihara, et al. 2021). Both sperm and egg genes were used as queries, because bouncer/SPACA4 are known to have switched expression in male vs female gametes in evolutionary history (Herberg et al. 2018; Fujihara et al. 2021). Our analysis revealed five proteins likely belonging to the TFP superfamily (Three-finger-protein domain-like) (Galat 2008). Known examples of TFPs acting in fertilization the gametic proteins bouncer and SPACA4 (Herberg et al. 2018; Fujihara et al. 2021). TFPs in salamanders also show evidence of tandem domain duplication and structural modification (Wilburn et al. 2014; Doty et al. 2016; Wilburn et al. 2022). In the class of plethodontid proteins with two TFP domains (2D-TFPs) we observe a different disulfide bonding pattern than the canonical disulfide bonding pattern present in TFPs with a single domain (1D-TFP) (Palmer, Hollis, et al. 2007; Wilburn et al. 2012; Wilburn et al. 2014; D.B. Wilburn et al. 2017). The canonical TFP disulfide bonding pattern has 8 cysteines, as opposed to 10 in this other class of domains (Doty et al. 2016).

Given our focus on identifying new fertilization proteins, we employed additional sequenced based methods for identifying potential proteins. A regular expression based on multiple sequence alignments of other TFPs (Garza-Garcia et al. 2009) was used to detect an additional five putative TFPs (Fig 3.3). Regular expressions can search for specific patterns of characters in a string of text (Thompson 1968), such as amino acid sequence features based on known TFP sequences (Garza-Garcia et al. 2009). We will refer to these as AOTFP1-5 (Abalone Ovary TFP), ranked based on transcript abundance (with AOTFP1 being the most abundant). While we are using transcript abundance as a heuristic in the search for candidate fertilization genes, it is worth noting that Isoseq data is not quantitative. In other quantitative sequencing experiments, long-read Isoseq data can be supplemented with more quantitative short read RNA-seq data (Huang et al. 2021). There are overall similarities in the tentatively identified TFPs from this transcriptome, but there is a considerable diversity in their length and number of cysteines. Differences in loop length and disulfide bonding patterns could affect protein structures and binding function.

The two most abundant AOTFPs both have 12 cysteines in comparison to 1D-TFPs (8 cysteines) and 2D-TFPs (10 cysteines). The AlphaFold2 predicted model of AOTFP1 had a similar disulfide bonding pattern to the second domain in 2D-TFPs, but with an additional disulfide bond within the third finger of the TFP domain, which makes it mirror the first finger of the domain. An additional disulfide bond could stabilize this finger of the TFP, which could be important if this loop is especially functionally important. There are also other possible ways possible ways disulfide pattern can reshuffle that could have substantial functional consequences. Any disulfide bonding pattern similarity between 2D-TFPs and abalone TFPs could reflect homology or potentially even convergent evolution. These proteins require more structural and functional verification, but these modifications of TFP loop structures seem plausible given the functional diversification of these proteins. AlphaFold2 predictions are based

on multi-sequence alignments of existing data (Jumper et al. 2021), and the true disulfide bonding patterns of these AOTFPs could differ completely from previously identified TFPs. If such disulfide bonding patterns were characterized, they could reflect another example of structural and functional diversification in TFPs. To the best of our knowledge, these would be the first known 12 cysteine TFPs if validated, and they could even represent an invertebrate specific offshoot of the TFP superfamily.

3.2.4 Comparative genomics can uncover rapidly evolving ovary genes.

Evidence of rapid evolution is an important characteristic when identifying functional proteins in fertilization. While many essential proteins could be conserved, others may rapidly evolve as evidence of recent adaptive evolution. We analyzed for positive selection in genes expressed in the abalone ovary transcriptome by comparing the transcripts to published Pacific abalone draft genomes (Masonbrink et al. 2019), with the goal of identifying rapidly evolving genes that may function in fertilization. Genes that are rapidly evolving may serve an essential function in the ovary, and features such as signal peptides or transmembrane domains suggest that a protein is extracellular and sperm accessible. Signal peptides allow for proteins to be secreted (Holland 2004), and transmembrane domains allow for proteins to be membrane bound (Guna and Hegde 2018). It is believed that ZP proteins are cleaved at the egg plasma membrane and later incorporated into the egg coat (Jovine et al. 2002). Proteins that had a signal peptide and/or transmembrane domain were initially considered in this analysis. Genomic variant data from other Pacific abalone species were used to reconstruct the orthologous gene sequences for five Pacific abalone species: green (*H. fulgens*), pink (*H. corrugata*), pinto (*H. kamtschatkana*), black (*H. cracherodii*), and white (*H. sorenseni*) abalone. We focused on non-singleton gene clusters, which we defined as having more than one read supporting the gene cluster. This was done to improve our statistical power by reducing the number of tests conducted. Our site-wise nucleotide substitution rate analysis (Yang 2007) identified proteins

from our abundant sperm-accessible candidates that are rapidly evolving (Table 3.3). We calculated site-wise dN/dS values for these nucleotides, and positively selected genes were decided based on likelihood ratio testing between a neutral model (M8a), and one that allows positive selection (M8). BLAST homology searches provide clues on the identity of these proteins, but these distant homology relations might not reflect the true function of these proteins. One protein from this positive selection analysis was identified as a putative homolog of low-density lipoprotein receptor (Fig 3.4). Its AlphaFold2 predicted structure contains multiple barrels that form a cylindrical shape, which supports a receptor function. It is worth confirming protein structures experimentally, as Alphafold2 prediction quality varies along the protein sequence. Receptor proteins could potentially be co-opted for a role in fertilization.

The transcriptomic analysis has provided a starting point to investigate the role of newly discovered proteins and protein families in red abalone egg biology. There are multiple newly discovered examples of proteins belonging to ZPs and the TFP superfamily, which are two groups of proteins previously implicated in the egg coat and fertilization. Neofunctionalization of proteins could cause effects such as the loss of transmembrane domains in ZPs, or a change in TFP disulfide bonding patterning. Positive selection analyses can also provide a way to identify ovary-expressed genes that are rapidly evolving fertilization proteins, such as the putative low-density lipoprotein receptor. Candidate fertilization should be validated with mass spectrometry (Canterbury et al. 2014) to confirm their abundance in proteome. Going forward, putative fertilization genes can be functionally analyzed with biochemical binding assays such as surface plasmon resonance (SPR) (Douzi 2017). This research illustrates the importance of integrating multiple methods to identify candidate fertilization proteins.

3.3 Materials and Methods

3.3.1 Sequencing Ovary Transcriptome

Red abalone ovary tissues were isolated from gravid female red abalone samples purchased from The Abalone Farm in Cayous, California (http://www.abalonefarm.com). RNA was isolated from abalone ovary tissue samples using a modification of the guanidinium isothiocyanate isolation procedure (MacDonald et al. 1987). Ovary tissue was homogenized in a 4M guanidinium isothiocyanate solution (containing 0.1 M Tris-HCl pH 7.5 + BetaMercaptoethanol - BME), and RNA was isolated by ultracentrifuging the sample in a 5.7M cesium chloride gradient containing 0.1M EDTA. Centrifugation was performed at 30,000 rpm (rotation per minute) for 23 hours at 20°C, using a Beckman SW41Ti rotor. RNA pellets were washed three times with 70% ethanol and stored in 70% ethanol at -80°C prior to sequencing.

We used a standard PacBio Isoseq procedure for red abalone Ovary RNA. Initially, reverse transcription primers are annealed to the RNA sample to produce single stranded DNA (ssDNA). After ssDNA was generated, we performed reverse transcription with template switching primers, in order to generate double stranded (dsDNA), which was bead purified prior to cDNA amplification. This entailed optimization of reverse transcription of mRNA generate cDNA. The use of poly-T primers allowed for specific amplification of mRNA. The number of PCR cycles for reverse transcription was optimized at 24 cycles, with the goal of amplifying our transcripts without introducing a high risk of PCR artifacts. Then the cDNA sample was purified. Reverse transcription has a size bias for smaller transcripts that can be mitigated by the use of ampure beads which size select molecules during nucleic acid purification. Purifying nucleic acid samples at lower volumes of beads selects for larger molecules (Quail et al. 2009). One fifth of our cDNA sample was cleaned at x1 ampure bead volume (not size-selected), and the remaining 80% was cleaned with x0.4 ampure bead volume (size selection for larger cDNAs).

After cDNA is amplified and purified, DNA damage repair was performed. The cDNA was mixed with NAD, water, and PacBio “DNA Prep Buffer” and “DNA Damage Repair Mix v2”, and incubated at 37°C for 30 minutes. On ice, the cDNA sample is then mixed with Isoseq “End Prep Mix” and is placed on a thermocycler for 30 minutes at 20°C, and 30 minutes at 65°C. After DNA repair, the adapters are ligated mixing the PacBio Isoseq kit ligation reagents, and then putting the mixture at 20°C for one hour. The addition of circular adaptors allows the sequencing molecule multiple times to create subreads, which are then averaged into a circular consensus read, with much higher per-base accuracy. The sequencing was performed using Sequel II PacBio system, with SMRT-cell (Single Molecule Real Time) technology.

3.3.2 Transcriptome Analysis

Gene clusters were predicted from full length non-concatenated reads using isonclust, which is a tool designed for *de novo* long read clustering (Sahlin and Medvedev 2019). We used the standard isonclust settings for PacBio Isoseq data. For each gene cluster all non-identical reads were identified and largest non-overlapping open reading frames (ORF) were identified for each read. The initial transcriptomic analysis considered one representative protein for each gene cluster for computational efficiency when performing phylogenetics and positive selection analysis. Similarly, one representative ORF, and one representative read was chosen for each cluster. In order to select a representative protein, we used a quality score for each predicted protein that was the protein length multiplied by the log of the read count, in an effort to balance the importance of abundance and length when identifying real transcripts. Similarly, ORFs and reads were also scored based on their length multiplied by the log of their count. Representative ORFs were the highest scoring ORFs that codes for the cluster’s representative protein, and representative read is the highest scoring read corresponding to that ORF. Signal peptides and transmembrane domains were identified using SignalP and HMMTOP (Tusnády and Simon 2001; Almagro Armenteros et al. 2019).

We searched for other well-established classes of fertilization protein domains, such as ZPs and TFPs (Three finger proteins) with the goal of discovering potential new fertilization proteins. ZP modules were identified based on BLAST (Altschul et al. 1997) homology searches of the best representative protein from each gene cluster against red ovary express sequence tag (EST) libraries (Aagaard et al. 2010). TFP-like domain containing proteins were identified using a regular expression based on a TFP multiple sequence alignment (Garza-Garcia et al. 2009).

3.3.3 Phylogenetics

Our ZP maximum likelihood phylogeny was performed on our predicted transcriptome ZP sequences along with other vertebrate ZP sequences from the following proteins: ZP1, ZP2, ZP3, ZP4, ZPAX, ZPD, UMOD, CUZD1, and TECTA, obtained from the following species: *D. rerio*, *G. aculeatus*, *X. tropicalis*, *A. carolinensis*, *G. gallus*, and *P. troglodytes*. Our TFP phylogeny contained five predicted TFPs along with SPACA4 sequences from *D. rerio*, *G. aculeatus*, *H. sapien*, and *P. troglodytes*. These vertebrate ZP and TFP sequences were obtained from Ensembl (Cunningham et al. 2022). For both phylogenies, we used RAxML-NG (Kozlov et al. 2019) with the LG+G substitution model with transfer bootstrap support values calculated (Lemoine et al. 2018).

3.3.4 Positive Selection Testing

In order to test for positive selection, we established a mapping of the representative reads to an abalone draft genome (Masonbrink et al. 2019) using minimap2 (Li 2018). This established coordinates for each of the genes of interest. Using variant call files (VCF) we identified variants that exist between red abalone and other pacific abalone, which allowed for the reconstruction of nucleotide sequences for those species. The predicted sequences from these other genomes did not require additional alignment, because the short read sequences

were also aligned to the red abalone genome. Our analysis considered abundant proteins that had at least one read supporting the best representative protein and did not include insertion-deletion variants to make the search for rapidly evolving genes more conservative. Using PAML (Yang 2007) we performed site-wise dN/dS ratio analysis to detect for positive selection. We then performed a likelihood ratio test by comparing a neutral model (M8a) and one allowing for positive selection (M8). Benjamini-Hochberg multiple testing correction was used (Benjamini and Hochberg 1995) to identify which proteins have statistically significant support for rapid evolution ($p<0.05$).

3.4 Acknowledgments

This research was funded by the following NIH grants: R21HD105025 awarded to WJS and K99HD090201 awarded to DBW. We also thank members of the lab past and present for contributions to discussions (Damien Wilburn, Jolie Carlisle, and Jan Aagaard).

3.5 Figures and Tables

Number	Category
2,903,189	Total number of reads
2,682,257	Reads with 5’ and 3’ primers
2,661,779	Non-concatemer reads with 5’ and 3’ primers
2,658,784	Non-concatemer reads with 5’ and 3’ primers and poly-A Tail
1,886	Mean length of full-length non-concatemer reads (FLNC)
30,579	Number of predicted gene clusters
19,967	Number of non-singleton gene clusters

Table 3.1: Summary of the reads obtained from the transcriptome of *Haliotis rufescens.*

Gene Name Used	Closest Abalone Homolog	Cluster Rank	TPM (Transcripts per million)
ZPSL1	**ZPS**	12	5759.0
VERL	VERL	18	5154.6
VEZP8	VEZP8	20	4600.2
ZPFL1	**ZPF**	42	2157.8
ZPWL1	**ZPW**	46	2069.0
VEZP2	VEZP2	53	1866.6
VEZP4	VEZP4	55	1837.7
ZPFL2	**ZPF**	69	1588.7
ZPXL1	**ZPX**	95	1326.9

Table 3.2: Nine most abundantly expressed putative abalone ZPs. ZP identity was based on homology searches. Some of these homologs are pseudogenes (ZPS, ZPF, ZPW, ZPX), and our transcripts are putative coding paralogs that follow the naming convention (ZPF-like 1). Genes whose closest abalone homologs are pseudogenes are shown in bold, to indicate that these are likely newly identified protein coding genes.

Close Abalone Homolog	Transcript Rank	M8a	M8	-2ΔlogL	*p* value	*p* value corrected
Nucleolar protein 14-like	218	p_0=0.846, p_1=0.154, p=0.005, q=87.3, ω=1	p_0=0.997, p_1=0.003, p=0.005, q=99, ω=306	17.7	1.28E-05	1.17E-03
Phospholipid scramblase 2-like	12466	p_0=0.83, p_1=0.17, p=0.005, q=76.1, ω=1	p_0=0.997, p_1=0.003, p=0.005, q=99, ω=636	19.2	5.85E-06	3.22E-03
ATP-dependent RNA helicase DDX55-like	1104	p_0=1, p_1=0, p=0.005, q=99, ω=1	p_0=0.998, p_1=0.002, p=0.005, q=2.27, ω=158	14.9	5.68E-05	3.91E-03
--	10373	p_0=0.534, p_1=0.466, p=0.005, q=51.3, ω=1	p_0=0.967, p_1=0.033, p=99, q=0.005, ω=999	14.3	7.68E-05	4.70E-03
Palmitoyl protein thioesterase 1-like	686	p_0=0.819, p_1=0.181, p=0.005, q=92.1, ω=1	p_0=0.969, p_1=0.031, p=0.005, q=7.04, ω=14.5	12.4	2.19E-04	0.012
Low density lipoprotein receptor-like	1707	p_0=0.702, p_1=0.298, p=0.005, q=3.43, ω=1	p_0=0.985, p_1=0.015, p=0.005, q=11.9, ω=59.4	11.2	4.07E-04	0.016
mitotic checkpoint serine/threonine-protein kinase BUB1-like	1010	p_0=0.910, p_1=0.09, p=0.005, q=99, ω=1	p_0=0.994, p_1=0.006, p=0.005, q=99, ω=52.6	11.5	3.46E-04	0.017
Lupus la-like protein	109	p_0=0.974, p_1=0.026, p=0.005, q=0.032, ω=1	p_0=0.997, p_1=0.003, p=0.005, q=4.06, ω=999	10.6	5.67E-04	0.021
uncharacterized protein LOC124111153	10313	p_0=0.920, p_1=0.08, p=0.005, q=99, ω=1	p_0=0.994, p_1=0.006, p=0.005, q=29.6, ω=38.5	10.4	6.25E-04	0.022
Ceramide synthase 2-like	10067	p_0=0.818, p_1=0.182, p=0.005, q=1.73, ω=1	p_0=0.997, p_1=0.003, p =0.005, q=5.25, ω=999	10.2	7.08E-04	0.022

Table 3.3: Top results for rapidly evolving predicted fertilization proteins. Predicted proteins from the red abalone transcriptome are predicted as sperm accessible if they have a signal peptide or transmembrane domain. Positive selection was determined through site-based dN/dS analysis, with likelihood ratio tests based on comparisons between a neutral model (M8a) and one that allows for positive selection (M8). The top 10 results are shown, ranked according to Benjamini-Hochberg corrected p-values. Closest abalone predicted homologs are listed according to BLASTP homology searches. These BLAST protein designations are based on sequence homology and may not reflect the true function.

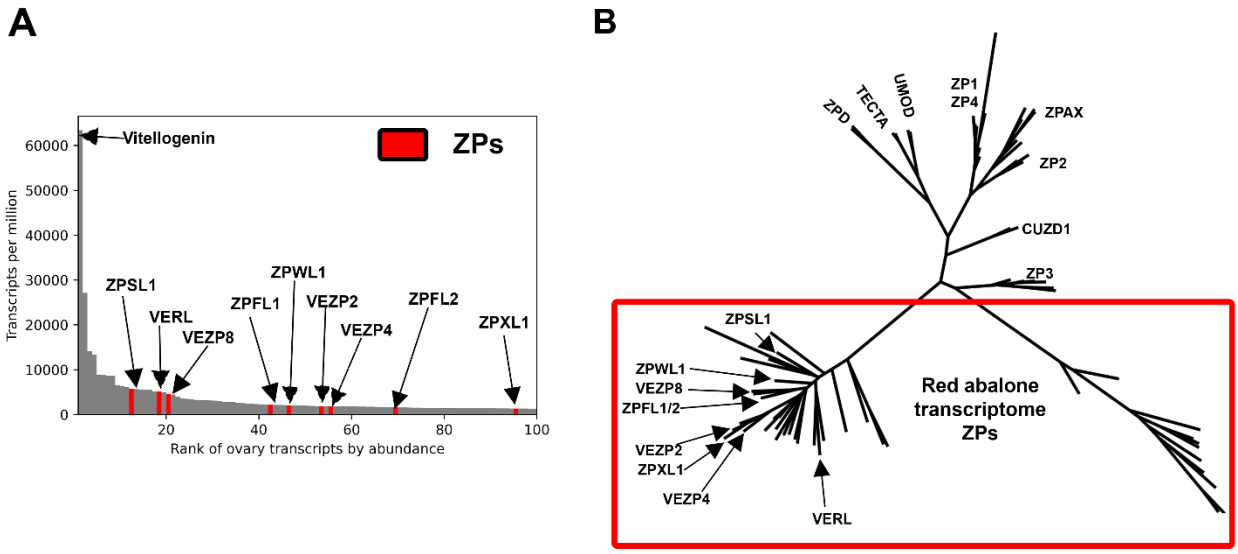


Figure 3.1: Phylogenetic context of red abalone transcriptome ZPs. A) Abundance was measured by transcripts per million for the top 100 most gene clusters in the transcriptome. BLAST homology searches indicated 9 of these top 100 clusters as coding for ZP containing proteins. B) A maximum likelihood phylogeny included all ZP module sequences from vertebrates and predicted abalone ZP proteins. Red abalone transcripts form two clades that are shown in the red box. The 9 most abundant ZPs are labelled on the phylogeny.

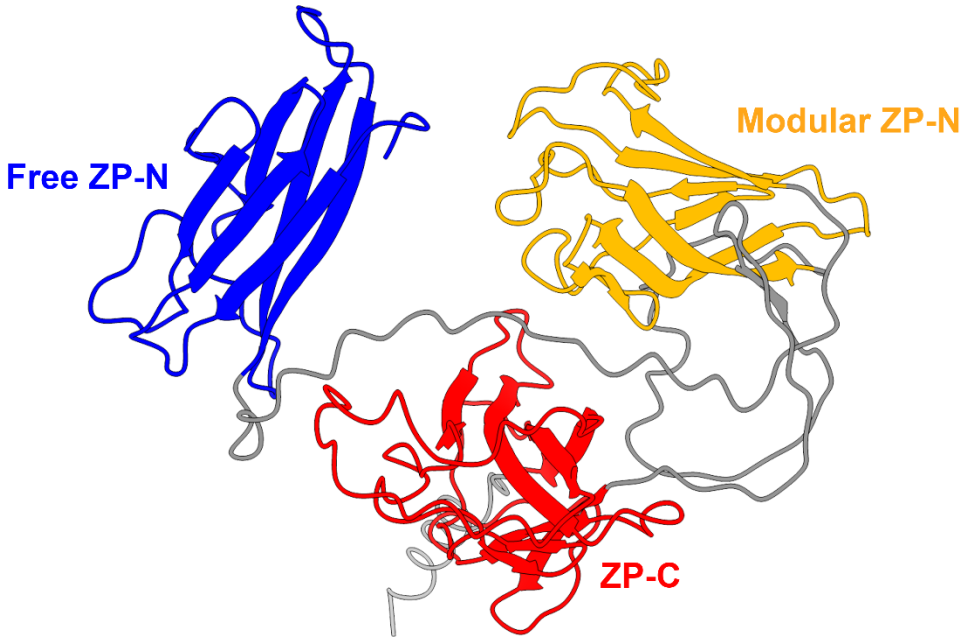


Figure 3.2: Predicted tertiary structure of ZPF-like protein 1. This putative ZP protein is referred to as ZPF-like protein 1 (ZPFL1). An AlphaFold2 predicted tertiary structure of ZPFL1 contains a free duplicated ZP-N domain as well as a complete ZP module (ZP-C paired with ZP-N), while lacking a predicted transmembrane domain, which is typical of VEZPs.

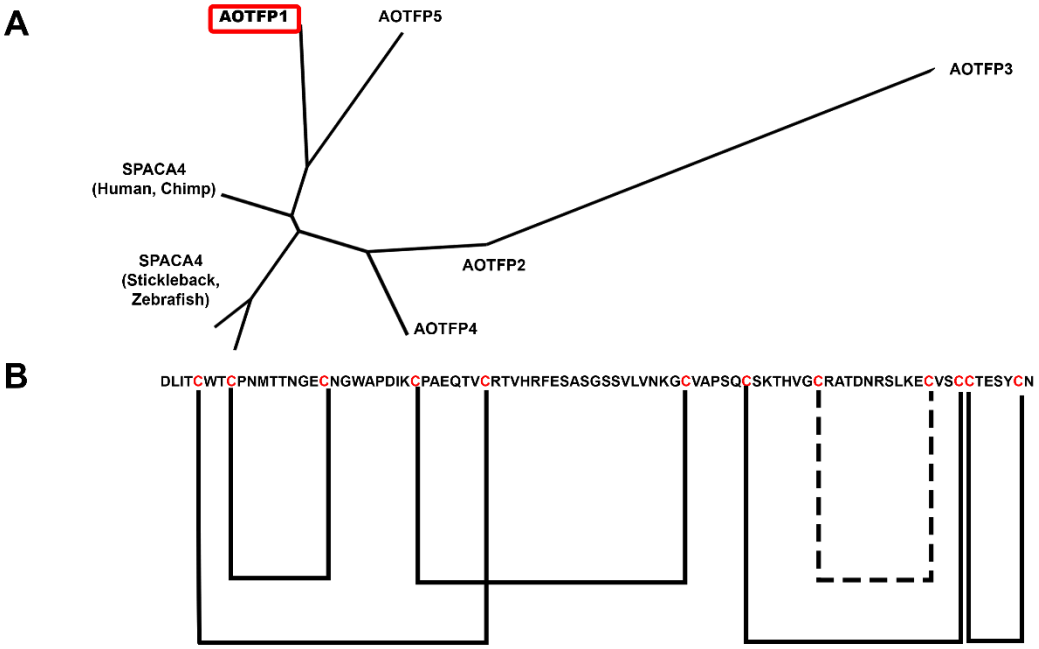


Figure 3.3: Evolution and predicted disulfide bonding patterns of AOTFP1. A) Phylogeny of five identified abalone ovary TFPs (AOTFPs), along with SPACA4 sequences from apes and fish. They are numbered one to five based on their relative transcript abundance (AOTFP1 having the most reads supporting it). B) Potential 12-cysteine disulfide bonding pattern in AOTFP1. the canonical 2D-TFP disulfide bonding pattern was predicted by AlphaFold2, along with a hypothetical additional disulfide bond shown in a dotted bracket.

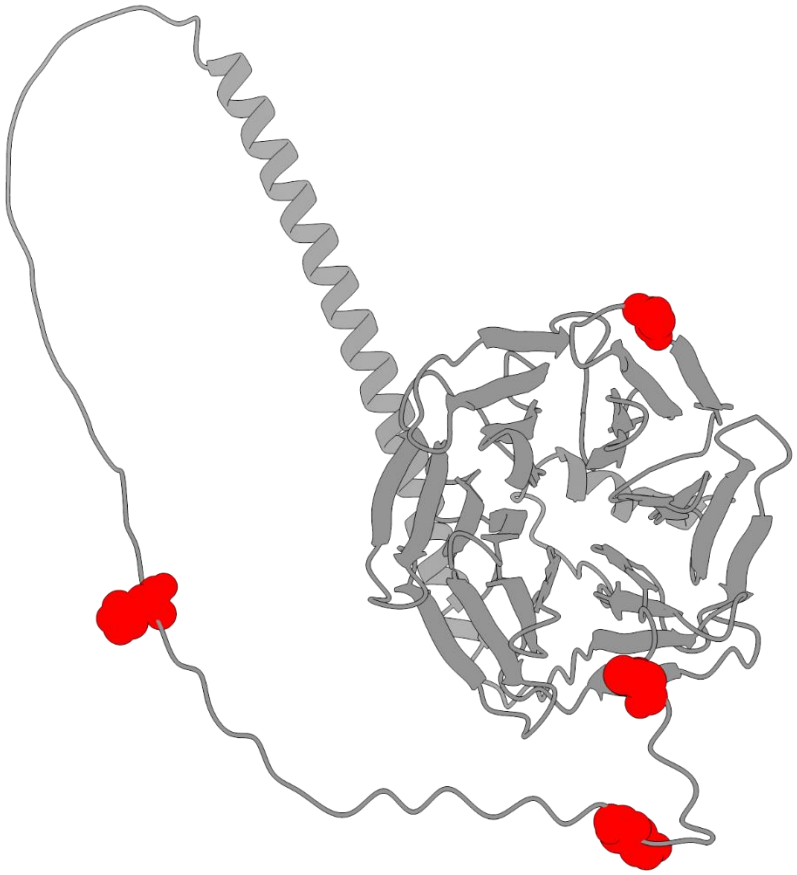


Figure 3.4: Predicted tertiary of rapidly evolving ovary protein (LDLR-like). The AlphaFold2 predicted structure for a putative homolog of a low-density lipoprotein receptor (LDLR-like). The function of this protein could differ from that of its predicted receptor homolog. Red sites are those predicted to be rapidly evolving by PAML dN/dS analysis Structural predicted quality varies along the protein structure, as the long unstructured c-terminal end has less sequence data informing its structural prediction.

Chapter 4: Future Directions and Conclusions

4.1 Extensions of Previous Research

There are multiple avenues of research that naturally arise from the ZP-N evolutionary characterization and transcriptomic analysis of abalone ovaries. Additional positive selection analysis can be performed, especially by relaxing our constraints on searching only the most abundant proteins. We initially performed positive selection analyses on gene clusters with more than one read supporting them. While transcript abundance was used as a heuristic to identify likely functional proteins, there RNA expression levels do not necessarily predict protein abundance (Schrimpf et al. 2009). Certain essential fertilization proteins may have relatively low abundance. Expansion of our positive selection analysis to less abundantly expressed products would allow for a more comprehensive assessment of ovary proteins. There is also the example of five predicted AOTFPs, which are not particularly abundant at the transcript level, but they could be more highly expressed at the protein level. Expression patterns between mRNAs and proteins could differ substantially due various post-transcriptional and translational mechanisms (Greenbaum et al. 2003; de Sousa Abreu et al. 2009; Vogel and Marcotte 2012). This encourages the relaxation of transcript abundance constraints in future positive selection analyses. We can perform positive selection analysis on all gene clusters, including low abundance transcripts, like our tentative TFP sequences.

The transcriptomic analysis identified candidates for functional egg coat genes, but validation of protein expression can be performed on existing proteomic datasets. We previously performed DDA (data dependent acquisition) tandem mass spectrometry (MS/MS) on a red abalone sample of isolated egg coats. DDA performs the MS/MS acquisition on the most abundant peptides from the first round of mass spectrometry (Canterbury et al. 2014). In contrast, DIA (data independent acquisition) based mass spectrometry systematically scans the entire initial mass spectra, with each set of peptides falling within a mass-to-charge (m/z)

window being processed for MS/MS (Canterbury et al. 2014). Because it does not rely on precursor ion abundance, DIA could be more sensitive for detecting lower abundance proteins. Proteins that show evidence of egg coat expression could be prioritized as candidates for future functional characterization. We could further characterize these proteins by fractionating egg into egg coats and cell membrane fractions, allowing for potential functional insights. For example, an egg receptor for fusion, such as Juno (Bianchi and Wright 2014; Bianchi and Wright 2015), might be present in a cell membrane fraction. Expression of essential egg genes can also vary over time, such as Juno expression increases in oocyte cell membranes during their maturation (Jean et al. 2019). Taking into consideration such developmental effects when analyzing abalone ovaries can further elucidate the molecular mechanisms of fertilization.

So far, the transcriptomic research outlined here has focused on identifying individual candidate proteins. However, co-evolution between interacting proteins is critical to understanding the molecular mechanisms of fertilization. Co-evolutionary analyses between transcripts present in abalone ovaries and sperm could provide evidence of fertilization functions in particular protein pairs. Evolutionary rate covariation (ERC) (Findlay et al. 2014) is a useful tool for identifying co-evolving sperm and egg proteins. ERC analyses compare multiple species trees constructed with different protein sequences. If branch lengths correlate between two different trees, it indicates that those corresponding proteins are co-evolving (Clark et al. 2012). A signature of co-evolution could reflect a direct biochemical interaction between two gametic proteins, or it could be evidence of factors such as indirect sexual selection (Wilburn et al. 2019). One of the benefits of computational prediction of interactors by ERC is that it provides explicit hypotheses to test experimentally. For instance, predicted co-evolving ovary and sperm proteins can be further characterized functionally. Identified gametic proteins can be further investigated with biochemical binding assays such as immunoprecipitation, affinity chromatography or direct biophysical measurements of interaction such as SPR (Douzi 2017).

There are also more complicated possible evolutionary dynamics. Under indirect sexual selection, an abalone protein might not directly interact with sperm, but it could be co-evolving with a different egg coat protein that does engage in such gametic interactions. This could similarly be characterized if two co-evolving proteins both display binding affinity with a third protein.

4.2 New Research Directions

The research described here focused primarily on the evolution of fertilization proteins, such as those in the ZP family, with the goal of addressing the limits to our understanding of ZP-N evolution. Much of our research also required developing and applying machine learning algorithms to analyze the evolution of large sets of genomic data. A project that has naturally arisen from the work described in this dissertation, has been to develop a ZP-N detection algorithm. Much of my research into ZP-N domain evolution, has been motivated by the lack of information about ZP-N domain sequence and structure. Current protein databases such as pfam (Mistry et al. 2021), only include full ZP modules and exclude free ZP-N domains such as those in ZP2. Our previous ZP-N phylogenetics work collected over 2000 ZP-N domain sequences from Ensembl (Howe et al. 2021), which can be used to construct a training dataset for ZP-N model detection. This training data can also be supplemented with several invertebrate ZP-N sequences such as cuticlin from beetles or dumpy from *Drosophila* in order to provide greater sensitivity to detect functionally diverse ZP-N sequences (Bork and Sander 1992). This training dataset can be expanded to include even more diverse sequences, such as the fungal sag1p (Wilburn and Swanson 2016), which contains a domain that is structurally homologous to ZP-N domain and is part of the yeast agglutination system. A model trained on only vertebrate

sequences could prove sufficient, but adding various invertebrate or fungi sequences could improve our sensitivity to detect ZP-Ns in other species.

We are constructing our ZP-N domain detection algorithm using a residual neural network (He et al. 2016). Residual neural networks are often used for complicated pattern recognition problems (Hanif and Bilal 2020), which make them appropriate for the recognition of patterns underlying these hypervariable ZP-N domains. This model was trained using unaligned ZP-N sequences from previously published phylogenetic and machine learning research (Rivera et al. 2022). The goal of this algorithm is to identify the probability a given sequence contains a ZP-N domain followed by the location of the domain within a sequence. A sliding window approach could allow us to provide specific coordinates for the ZP-N domain. Because the model needs to provide ZP-N coordinates on any individual sequence rather than an alignment, it is important that we train on unaligned sequences.

Preliminary targeted analysis with human protein predicted that mammalian OOSP2 (oocyte secreted protein 2) contains a previously undetected ZP-N domain. This is consistent with the cysteine patterning in OOSP2 and its AlphaFold2 predicted structures (Jumper et al. 2021; Mirdita et al. 2022). OOSP2 has female reproductive system expression in mice (Abbasi et al. 2020), as well as evidence of testis expression in humans (Fagerberg et al. 2014). This could provide an example for ZP-N domains switching to the sperm side of reproduction. Such co-option of reproductive proteins by the opposite gamete has been proposed for bouncer/SPACA4 (Fujihara et al. 2021).

These preliminary results illustrate the power of the ZP-N detection algorithm. We can continue to add invertebrate sequences to the training data to improve our specificity and further optimize the hyperparameters of our residual neural network. This algorithm could detect new human ZPs, which would warrant further functional research. Identifying more functional mammalian ZPs could greatly advance our understanding of fertilization overall, and could

provide targets for research into contraception or infertility. The identification of ZP-N domains in the genomes of other model or non-model organisms could would elucidate their evolutionary history. For instance, detecting a ZP in plants, would make the ZP-N domain substantially more ancient than previously proposed (Wilburn and Swanson 2016). Applying a ZP-N detection algorithm to existing genomes will be a fruitful source of biological inquiry.

Other future projects include investigations into VERL, an abalone egg coat gene with 23 ZP-N domains. Analyzing these duplicated sequences and performing comparative genomics may be helpful for understanding the importance of this large repeat array. The large number of highly duplicated domains make VERL a particularly difficult to sequence, but long read sequencing advancements such as telomere-to-telomere sequencing of human chromosomes (Nurk et al. 2022) have made this goal appear more feasible. Several attempts at sequencing VERL have only obtained partial VERL sequences. One approach utilized PCR amplification followed by PacBio long-read sequencing, but PCR produced multiple repeat related artifacts. Direct RNA sequencing of VERL with Oxford Nanopore MinION was also preformed, with the hope of skipping PCR would reduce artifacts. However, only incomplete VERL sequences were obtained, likely due to issues processing the entire molecule through the sequencing pore, potentially due to RNA secondary structure (Mathews et al. 2004; Gruber et al. 2008; Lorenz et al. 2011). Long read sequencing methods of native DNA without amplification present a powerful alternative for obtaining full length VERL. There are PacBio no-amplification sequencing methods that rely on targeted CRISPR/Cas9 digestion to enrich for a genomic region without PCR amplification (Hafford-Tear et al. 2019). Targeted digestion of a gene followed by a pull-down allows for enrichment of specific genomic regions for long-range sequencing. Whole genome sequencing can also be a powerfully alternative for obtaining full length VERL sequences, especially given recent advancements (Nurk et al. 2022). However,

more targeted sequencing approaches could be more scalable for large population genetic analyses across multiple species of abalone.

Once a sequencing methodology is established, VERL can be sequenced in multiple species of Pacific abalone so that the evolutionary history of these dramatic domain expansions can be investigated. This could entail VERL sequencing at the population level in multiple Pacific abalone species. Is the repeat number conserved within or between species? If it is conserved within species, but not between species it could represent the importance of VERL in particular reproductive isolation and speciation events. Alternatively, mechanisms such as non-homologous recombination could make repeat number hypervariable both within and between species, as observed in oyster bindin (Moy et al. 2008). If VERL ZP-N domain expansions exist to increase species-specific fertilization preference by demanding more successful binding events between ZP-N domains and lysin, having a large number of ZP-Ns could improve specificity. Comparative genomics between different abalone species, could help identify specific rapidly evolving ZP-N residues that are functionally important. Given the importance of VERL in mediating species-specificity, population genetics on VERL may help us understand the evolution of essential fertilization proteins.

4.3 Concluding Remarks

The work outlined here provides a foundation for further research into fertilization as well as other biological systems. We have primarily discussed the evolution of protein domains in terms of neofunctionalization of distantly related genes, but subfunctionalization may have played an important role in the evolutionary history of highly duplicated gene families. Future research can focus on addressing the differentiation of neofunctionalization and subfunctionalization. Analysis of differential RNA expression profiles has been used to classify recently duplicated *Drosophila* genes as neofunctionalized or subfunctionalized (Assis and

Bachtrog 2013). This approach considers how differential expression in time and space is modelled as way genes can subfunctionalize (Force et al. 1999). When studying duplicated gene families there are ways to form hypotheses regarding neofunctionalization or subfunctionalization. For instance, the gain or loss of a particular binding domain could suggest a change in function. Functions for proteins of interest can be investigated through binding assays and analysis of co-evolution. Observed and predicted protein structures for existing proteins, and predicted protein structures of ancestral proteins could also provide clues regarding functional evolution of duplicated proteins. Identifying possible subfunctionalization or neofunctionalization events in a phylogenetic context can elucidate the evolution of essential protein families.

However, the exact distinction between neofunctionalization and subfunctionalization is a complicated question in biology. At what point does a differentially expressed subfunctionalized gene become divergent enough to be considered as having a novel function? Is such divergence inevitable at longer timescales as differentially expressed genes face different selective pressures? Some have described subfunctionalization as a transitory state to neofunctionalization (Rastogi and Liberles 2005). An alternative hypothesis is that strong evolutionary conservation could allow for preservation of subfunctionalized gene copies across greater evolutionary timescales. The greater availability of genomic data and advancements in sequencing technology could make these difficult questions more tractable. Combining different methodologies such as RNA single cell sequencing (Hwang et al. 2018), phylogenomics, co-evolutionary analyses, and protein structural modelling could provide valuable insight into distinctions between neofunctionalization and subfunctionalization.

Our integration of phylogenetics, machine learning and protein structural analysis can be applied to other essential protein families. It is also worth considering how this approach can be modified. In order to align large numbers of diverse ZP-N domains (Wilburn and Swanson 2016) we used a two-stage structurally informed alignment (Pei et al. 2008). A simpler alignment strategy could be less likely to introduce computational artifacts, especially for less diverged sets of protein sequences. The conclusions we draw regarding ZP-N evolution (Rivera et al. 2022) are based on our multiple sequence alignments of vertebrate sequence data obtained from Ensembl (Howe et al. 2021). We are also able to model ancient divides in ZP evolution using vertebrate and abalone transcriptome sequences. The inclusion of outgroups for phylogenetic rooting can also be an important tool for making inferences about evolutionary history.

Different inferences can be drawn from analyzing protein sequence data at varying taxonomic depths. Our analyses of positive selection in ZP-N domains was limited to boreoeutherian mammals (Rivera et al. 2022), to avoid the saturation of mutations in rapidly evolving sequences (Anisimova and Liberles 2012). However, analyzing positive selection at differing taxonomic scales has produced different conclusions regarding ZP-N evolution (Turner and Hoekstra 2006). Whether you investigate ZP-N domains or other protein families, these considerations of alignment size, number of sequences, and diversity can affect your understanding of evolutionary history.

We outlined our strategy for using machine learning classifier on phylogenetically defined ZP-N clades, but this approach can be also be modified in many ways. In our supplemental materials, we briefly discuss a three-way alternative classification scheme that warrants further investigation (Rivera et al. 2022). If this analysis were applied to phylogenies of other protein families, different numbers of classes can be considered. Binary classifications can be more useful for understanding more ancient evolutionary events. However, defining multiple

classes of protein domain sequences could allow for the targeted investigation of specific clades, such as groups of orthologous sequences. Results from models obtained with different number of classes can also be combined to understand evolution at varying timescales.

We used a supervised logistic regression classifier (Bewick et al. 2005) in our analysis of ZP-N domains, because outputs can be interpreted as classification probabilities, and the parameters can be interpreted as probabilistic weights associated with particular residues as particular sites. This focus on biological interpretability proved useful, but other machine learning can be explored when investigating protein domain evolution. Alternative methods for classifying sequence data include support vector machines (SVMs) and random tree algorithms. SVMs are capable of complex biological classification problems and can provide highly parsimonious solutions, but interpretations of the models can prove complex (Devos et al. 2009). However, methods exist to select for SVMs that rely on relatively few datapoints for classification (Devos et al. 2009). Defining a small set of sequences that define a biological classification (such as free vs modular ZP-Ns), could lead to meaningful biological inquiries. Identifying the position of these sequences within a phylogeny could highlight important evolutionary divides and mutations throughout history.

Random tree algorithms are often used for biological classification problems such as gene prediction (Díaz-Uriarte and Alvarez de Andrés 2006). Model optimization methods exist to reduce the number of features in random tree algorithms (Kursa 2014), which could be useful for obtaining sparser solutions where the relevant biological features can be investigated. As outlined in the future direction section of this chapter, other machine learning applications exist such as the identification of protein domains of interest from genomic data, using residual neural networks (Hanif and Bilal 2020). Neural networks are exceptional at complex pattern recognition, but the interconnectivity between nodes and layers within a model complicates biological interpretations of parameters (Schmidhuber 2015). While the parameters would be

less easily interpretable than the values from a logistic regression model, a convolutional neural network (Gu et al. 2018) could capture epistatic effects between neighboring sites. The combination of different machine learning classification systems could enrich our understanding of protein domain evolution.

There can be more positively selected proteins in our transcriptome that warrant co-evolutionary analyses and subsequent functional characterization. Sequencing an essential ZP protein with multiple duplicated domains can help us better understand the evolutionary dynamics underlying their repeat expansions. ZP-N detection algorithms could help identify novel functional ZPs in numerous genomes, which could reveal a more complicated history of ZP-N domain duplication. Integrating some of these future projects with our existing research could help elucidate the intersection of genomic duplication events, neofunctionalization, and reproductive biology.

Bibliography

Aagaard J, Springer S, Soelberg S, Swanson W. 2013. Duplicate Abalone Egg Coat Proteins Bind Sperm Lysin Similarly, but Evolve Oppositely, Consistent with Molecular Mimicry at Fertilization. PLoS Genet. 9(2).

Aagaard J, Vacquier V, MacCoss M, Swanson W. 2010. ZP Domain Proteins in the Abalone Egg Coat Include a Paralog of VERL under Positive Selection That Binds Lysin and 18-kDa Sperm Proteins. Mol Biol Evol. 27(1):193–203.

Abbasi F, Kodani M, Emori C, Kiyozumi D, Mori M, Fujihara Y, Ikawa M. 2020. CRISPR/Cas9-Mediated Genome Editing Reveals Oosp Family Genes are Dispensable for Female Fertility in Mice. Cells. 9(4). doi:10.3390/cells9040821.

Ai Y, Liu S, Luo H, Wu S, Wei H, Tang Z, Li X, Zou C. 2021. lncRNA DCST1-AS1 Facilitates Oral Squamous Cell Carcinoma by Promoting M2 Macrophage Polarization through Activating NF-κB Signaling. J Immunol Res. 2021:5524231–5524231. doi:10.1155/2021/5524231.

Alape-Girón A, Persson B, Cederlund E, Flores-Díaz M, Gutiérrez JM, Thelestam M, Bergman T, Jörnvall H. 1999. Elapid venom toxins: multiple recruitments of ancient scaffolds. European Journal of Biochemistry. 259(1–2):225–234. doi:10.1046/j.1432-1327.1999.00021.x.

Almagro Armenteros JJ, Tsirigos KD, Sønderby CK, Petersen TN, Winther O, Brunak S, von Heijne G, Nielsen H. 2019. SignalP 5.0 improves signal peptide predictions using deep neural networks. Nature Biotechnology. 37(4):420–423. doi:10.1038/s41587-019-0036-z.

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research. 25(17):3389–3402. doi:10.1093/nar/25.17.3389.

Anisimova M, Liberles D. 2012. Detecting and understanding natural selection.

Assis R, Bachtrog D. 2013. Neofunctionalization of young duplicate genes in Drosophila. Proceedings of the National Academy of Sciences. 110(43):17409–17414. doi:10.1073/pnas.1313759110.

Avella MA, Baibakov B, Dean J. 2014. A single domain of the ZP2 zona pellucida protein mediates gamete recognition in mice and humans. J Cell Biol. 205(6):801–809. doi:10.1083/jcb.201404025.

Avella MA, Xiong B, Dean J. 2013. The molecular basis of gamete recognition in mice and humans. Mol Human Reprod. 19(5):279–289.

Aydin H, Sultana A, Li S, Thavalingam A, Lee JE. 2016. Molecular architecture of the human sperm IZUMO1 and egg JUNO fertilization complex. Nature. 534(7608):562–565. doi:10.1038/nature18595.

Bairoch A, Apweiler R. 1996. The SWISS-PROT Protein Sequence Data Bank and Its New Supplement TREMBL. Nucleic Acids Research. 24(1):21–25. doi:10.1093/nar/24.1.21.

Benjamini Y, Hochberg Y. 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society: Series B (Methodological). 57(1):289–300. doi:10.1111/j.2517-6161.1995.tb02031.x.

Bewick V, Cheek L, Ball J. 2005. Statistics review 14: Logistic regression. Crit Care. 9(1):112–118. doi:10.1186/cc3045.

Bianchi E, Doe B, Goulding D, Wright GJ. 2014. Juno is the egg Izumo receptor and is essential for mammalian fertilization. Nature. 508(7497):483–487. doi:10.1038/nature13203.

Bianchi E, Wright GJ. 2014a. Izumo meets Juno: preventing polyspermy in fertilization. Cell Cycle. 13(13):2019–2020. doi:10.4161/cc.29461.

Bianchi E, Wright GJ. 2014b. Izumo meets Juno. Cell Cycle. 13(13):2019–2020. doi:10.4161/cc.29461.

Bianchi E, Wright GJ. 2015. Cross-species fertilization: the hamster egg receptor, Juno, binds the human sperm ligand, Izumo1. Philosophical Transactions of the Royal Society B: Biological Sciences. 370(1661):20140101. doi:10.1098/rstb.2014.0101.

Björklund ÅK, Ekman D, Elofsson A. 2006. Expansion of Protein Domain Repeats. PLOS Computational Biology. 2(8):e114. doi:10.1371/journal.pcbi.0020114.

Björklund ÅK, Ekman D, Light S, Frey-Skött J, Elofsson A. 2005. Domain Rearrangements in Protein Evolution. Journal of Molecular Biology. 353(4):911–923. doi:10.1016/j.jmb.2005.08.067.

Bleil JD, Beall CF, Wassarman PM. 1981. Mammalian sperm-egg interaction: Fertilization of mouse eggs triggers modification of the major zona pellucida glycoprotein, ZP2. Developmental Biology. 86(1):189–197. doi:10.1016/0012-1606(81)90329-8.

Bokhove M, Jovine L. 2018. Structure of Zona Pellucida Module Proteins. Current Topic in Developmental Biology. 130(In Press):413–442.

Bokhove M, Nishimura K, Brunati M, Han L, de Sanctis D, Rampoldi L, Jovine L. 2016. A structured interdomain linker directs self-polymerization of human uromodulin. Proc Natl Acad Sci USA. 113(6):1552. doi:10.1073/pnas.1519803113.

Bonetta R, Valentino G. 2020. Machine learning techniques for protein function prediction. Proteins: Structure, Function, and Bioinformatics. 88(3):397–413. doi:10.1002/prot.25832.

Bork P, Sander C. 1992. A large domain common to sperm receptors (Zp2 and Zp3) and TGF-β type III receptor. FEBS Letters. 300(3):237–240. doi:10.1016/0014-5793(92)80853-9.

Brunati M, Perucca S, Han L, Cattaneo A, Consolato F, Andolfo A, Schaeffer C, Olinger E, Peng J, Santambrogio S, et al. 2015. The serine protease hepsin mediates urinary secretion and polymerisation of Zona Pellucida domain protein uromodulin. Elife. 4:e08887–e08887. doi:10.7554/eLife.08887.

Buljan M, Bateman A. 2009. The evolution of protein domain families. Biochemical Society Transactions. 37(4):751–755. doi:10.1042/BST0370751.

Busso D, Goldweic NM, Hayashi M, Kasahara M, Cuasnicú PS. 2007. Evidence for the Involvement of Testicular Protein CRISP2 in Mouse Sperm-Egg Fusion1. Biology of Reproduction. 76(4):701–708. doi:10.1095/biolreprod.106.056770.

Callebaut I, Mornon J-P, Monget P. 2007. Isolated ZP-N domains constitute the N-terminal extensions of Zona Pellucida proteins. Bioinformatics. 23(15):1871–1874. doi:10.1093/bioinformatics/btm265.

Canterbury JD, Merrihew GE, MacCoss MJ, Goodlett DR, Shaffer SA. 2014. Comparison of data acquisition strategies on quadrupole ion trap instrumentation for shotgun proteomics. J Am Soc Mass Spectrom. 25(12):2048–2059. doi:10.1007/s13361-014-0981-1.

Carlisle JA, Glenski MA, Swanson WJ. 2022. Recurrent Duplication and Diversification of Acrosomal Fertilization Proteins in Abalone. Frontiers in Cell and Developmental Biology. 10. https://www.frontiersin.org/articles/10.3389/fcell.2022.795273.

Carlisle JA, Swanson WJ. 2021. Molecular mechanisms and evolution of fertilization proteins. Journal of Experimental Zoology Part B: Molecular and Developmental Evolution. 336(8):652–665. doi:10.1002/jez.b.23004.

Chaudhury S, Gray JJ. 2008. Conformer Selection and Induced Fit in Flexible Backbone Protein–Protein Docking Using Computational and NMR Ensembles. Journal of Molecular Biology. 381(4):1068–1087. doi:10.1016/j.jmb.2008.05.042.

Civetta A. 2003. Positive Selection Within Sperm-Egg Adhesion Domains of Fertilin: An ADAM Gene with a Potential Role in Fertilization. Molecular Biology and Evolution. 20(1):21–29. doi:10.1093/molbev/msg002.

Clapham DE, Garbers DL. 2005. International Union of Pharmacology. L. Nomenclature and Structure-Function Relationships of CatSper and Two-Pore Channels. Pharmacol Rev. 57(4):451. doi:10.1124/pr.57.4.7.

Clark N, Gasper J, Sekino M, Springer S, Aquadro C, Swanson W. 2009. Coevolution of Interacting Fertilization Proteins. PLoS Genet. 5(7):e1000570.

Clark NL, Alani E, Aquadro CF. 2012. Evolutionary rate covariation reveals shared functionality and coexpression of genes. Genome Res. 22(4):714–720. doi:10.1101/gr.132647.111.

Claw KG, Swanson WJ. 2012. Evolution of the Egg: New Findings and Challenges. Annu Rev Genom Hum Genet. 13(1):109–125. doi:10.1146/annurev-genom-090711-163745.

Conant GC, Wolfe KH. 2008. Turning a hobby into a job: How duplicated genes find new functions. Nature Reviews Genetics. 9(12):938–950. doi:10.1038/nrg2482.

Connallon T, Clark AG. 2011. The Resolution of Sexual Antagonism by Gene Duplication. Genetics. 187(3):919–937. doi:10.1534/genetics.110.123729.

Conner SJ, Lefièvre L, Hughes DC, Barratt CLR. 2005. Cracking the egg: increased complexity in the zona pellucida. Human Reproduction. 20(5):1148–1152. doi:10.1093/humrep/deh835.

Cunningham F, Allen JE, Allen J, Alvarez-Jarreta J, Amode MR, Armean IM, Austine-Orimoloye O, Azov AG, Barnes I, Bennett R, et al. 2022. Ensembl 2022. Nucleic Acids Research. 50(D1):D988–D995. doi:10.1093/nar/gkab1049.

Curci L, Brukman NG, Weigel Muñoz M, Rojo D, Carvajal G, Sulzyk V, Gonzalez SN, Rubinstein M, Da Ros VG, Cuasnicú PS. 2020. Functional redundancy and compensation: Deletion of multiple murine Crisp genes reveals their essential role for male fertility. The FASEB Journal. 34(12):15718–15733. doi:10.1096/fj.202001406R.

Da Ros VG, Maldera JA, Willis WD, Cohen DJ, Goulding EH, Gelman DM, Rubinstein M, Eddy EM, Cuasnicu PS. 2008. Impaired sperm fertilizing ability in mice lacking Cysteine-RIch Secretory Protein 1 (CRISP1). Developmental Biology. 320(1):12–18. doi:10.1016/j.ydbio.2008.03.015.

Devos O, Ruckebusch C, Durand A, Duponchel L, Huvenne J-P. 2009. Support vector machines (SVM) in near infrared (NIR) spectroscopy: Focus on parameters optimization and model interpretation. Chemometrics and Intelligent Laboratory Systems. 96(1):27–33. doi:10.1016/j.chemolab.2008.11.005.

Devuyst O, Pattaro C. 2018. The UMOD Locus: Insights into the Pathogenesis and Prognosis of Kidney Disease. J Am Soc Nephrol. 29(3):713–726. doi:10.1681/ASN.2017070716.

Díaz-Uriarte R, Alvarez de Andrés S. 2006. Gene selection and classification of microarray data using random forest. BMC Bioinformatics. 7(1):3. doi:10.1186/1471-2105-7-3.

Dilimulati K, Orita M, Yonahara Y, Imai FL, Yonezawa N. 2022. Identification of Sperm-Binding Sites in the N-Terminal Domain of Bovine Egg Coat Glycoprotein ZP4. International Journal of Molecular Sciences. 23(2). doi:10.3390/ijms23020762.

Doty KA, Wilburn DB, Bowen KE, Feldhoff PW, Feldhoff RC. 2016. Co-option and evolution of non-olfactory proteinaceous pheromones in a terrestrial lungless salamander. Journal of Proteomics. 135:101–111. doi:10.1016/j.jprot.2015.09.019.

Douzi B. 2017. Protein–Protein Interactions: Surface Plasmon Resonance. In: Journet L, Cascales E, editors. Bacterial Protein Secretion Systems: Methods and Protocols. New York, NY: Springer New York. p. 257–275. https://doi.org/10.1007/978-1-4939-7033-9_21.

Dufour S, Quérat B, Tostivint H, Pasqualini C, Vaudry H, Rousseau K. 2020. Origin and Evolution of the Neuroendocrine Control of Reproduction in Vertebrates, With Special Focus on Genome and Gene Duplications. Physiological Reviews. 100(2):869–943. doi:10.1152/physrev.00009.2019.

Elder John F, Turner BJ. 1995. Concerted Evolution of Repetitive DNA Sequences in Eukaryotes. The Quarterly Review of Biology. 70(3):297–320. doi:10.1086/419073.

Eleveld-Trancikova D, Janssen RAJ, Hendriks IAM, Looman MWG, Moulin V, Jansen BJH, Jansen JH, Figdor CG, Adema GJ. 2008. The DC-derived protein DC-STAMP influences differentiation of myeloid cells. Leukemia. 22(2):455–459. doi:10.1038/sj.leu.2404910.

Eleveld-Trancikova D, Triantis V, Moulin V, Looman MWG, Wijers M, Fransen JAM, Lemckert AAC, Havenga MJE, Figdor CG, Janssen RAJ, et al. 2005. The dendritic cell-derived protein

DC-STAMP is highly conserved and localizes to the endoplasmic reticulum. Journal of Leukocyte Biology. 77(3):337–343. doi:10.1189/jlb.0804441.

Ellerman DA, Pei J, Gupta S, Snell WJ, Myles D, Primakoff P. 2009. Izumo is part of a multiprotein family whose members form large complexes on mammalian sperm. Mol Reprod Dev. 76(12):1188–1199. doi:10.1002/mrd.21092.

Elwood PC. 1989. Molecular Cloning and Characterization of the Human Folate-binding Protein cDNA from Placenta and Malignant Tissue Culture (KB) Cells. Journal of Biological Chemistry. 264(25):14893–14901. doi:10.1016/S0021-9258(18)63786-X.

Evans JP. 2020. Preventing polyspermy in mammalian eggs—Contributions of the membrane block and other mechanisms. Molecular Reproduction and Development. 87(3):341–349. doi:10.1002/mrd.23331.

Fagerberg L, Hallström BM, Oksvold P, Kampf C, Djureinovic D, Odeberg J, Habuka M, Tahmasebpoor S, Danielsson A, Edlund K, et al. 2014. Analysis of the Human Tissue-specific Expression by Genome-wide Integration of Transcriptomics and Antibody-based Proteomics *. Molecular & Cellular Proteomics. 13(2):397–406. doi:10.1074/mcp.M113.035600.

Fahrenkamp E, Algarra B, Jovine L. 2020. Mammalian egg coat modifications and the block to polyspermy. Molecular Reproduction and Development. 87(3):326–340. doi:10.1002/mrd.23320.

Feng J, Tian H, Hu Q-M, Meng Y, Xiao H-B. 2018. Evolution and multiple origins of zona pellucida genes in vertebrates. Biology Open. 7:bio036137. doi:10.1242/bio.036137.

Findlay GD, Sitnik JL, Wang W, Aquadro CF, Clark NL, Wolfner MF. 2014. Evolutionary Rate Covariation Identifies New Members of a Protein Network Required for Drosophila melanogaster Female Post-Mating Responses. PLOS Genetics. 10(1):e1004108. doi:10.1371/journal.pgen.1004108.

Finn S, Civetta A. 2010. Sexual Selection and the Molecular Evolution of ADAM Proteins. Journal of Molecular Evolution. 71(3):231–240. doi:10.1007/s00239-010-9382-7.

Force A, Lynch M, Pickett FB, Amores A, Yan YL, Postlethwait J. 1999. Preservation of duplicate genes by complementary, degenerative mutations. Genetics. 151(4):1531–1545.

Frank S. 2000. Sperm competition and female avoidance of polyspermy mediated by sperm-egg biochemistry. Evolutionary Ecology Research. 2:613–625.

Frolikova M, Manaskova-Postlerova P, Cerny J, Jankovicova J, Simonik O, Pohlova A, Secova P, Antalikova J, Dvorakova-Hortova K. 2018. CD9 and CD81 Interactions and Their Structural Modelling in Sperm Prior to Fertilization. Int J Mol Sci. 19(4):1236. doi:10.3390/ijms19041236.

Fu L, Niu B, Zhu Z, Wu S, Li W. 2012. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics. 28(23):3150–3152. doi:10.1093/bioinformatics/bts565.

Fujihara Y, Herberg S, Blaha A, Panser K, Kobayashi K, Larasati T, Novatchkova M, Theussl H-C, Olszanska O, Ikawa M, et al. 2021. The conserved fertility factor SPACA4/Bouncer has divergent modes of action in vertebrate fertilization. Proc Natl Acad Sci USA. 118(39):e2108777118. doi:10.1073/pnas.2108777118.

Gahlay G, Gauthier L, Baibakov B, Epifano O, Dean J. 2010. Gamete recognition in mice depends on the cleavage status of an egg's zona pellucida protein. Science. 329(5988):216–219. doi:10.1126/science.1188178.

Galat A. 2008. The three-fingered protein domain of the human genome. Cellular and Molecular Life Sciences. 65(21):3481–3493. doi:10.1007/s00018-008-8473-8.

Galat A. 2015. Multidimensional Drift of Sequence Attributes and Functional Profiles in the Superfamily of the Three-Finger Proteins and Their Structural Homologues. J Chem Inf Model. 55(9):2026–2041. doi:10.1021/acs.jcim.5b00322.

Galat A, Gross G, Drevet P, Sato A, Ménez A. 2008. Conserved structural determinants in three-fingered protein domains. The FEBS Journal. 275(12):3207–3225. doi:10.1111/j.1742-4658.2008.06473.x.

Galindo B, Moy G, Swanson W, Vacquier V. 2002. Full-length sequence of VERL, the egg vitelline envelope receptor for abalone sperm lysin. Gene. 288(1–2):111–7.

Galindo BE, Vacquier VD, Swanson WJ. 2003. Positive selection in the egg receptor for abalone sperm lysin. Proc Natl Acad Sci U S A. 100(8):4639–4643. doi:10.1073/pnas.0830022100.

Gallach M, Betrán E. 2011. Intralocus sexual conflict resolved through gene duplication. Trends in Ecology & Evolution. 26(5):222–228. doi:10.1016/j.tree.2011.02.004.

Gallach M, Chandrasekaran C, Betrán E. 2010. Analyses of Nuclearly Encoded Mitochondrial Genes Suggest Gene Duplication as a Mechanism for Resolving Intralocus Sexually Antagonistic Conflict in Drosophila. Genome Biology and Evolution. 2:835–850. doi:10.1093/gbe/evq069.

Gallach M, Domingues S, Betrán E. 2011. Gene Duplication and the Genome Distribution of Sex-Biased Genes. Kulathinal R, editor. International Journal of Evolutionary Biology. 2011:989438. doi:10.4061/2011/989438.

Garza-Garcia A, Harris R, Esposito D, Gates P, Driscoll P. 2009. Solution Structure and Phylogenetics of Prod1, a Member of the Three-Finger Protein Superfamily Implicated in Salamander Limb Regeneration. PloS one. 4:e7123. doi:10.1371/journal.pone.0007123.

Gavrilets S. 2014. Is sexual conflict an "engine of speciation"? Cold Spring Harb Perspect Biol. 6(12):a017723–a017723. doi:10.1101/cshperspect.a017723.

Gavrilets S, Waxman D. 2002. Sympatric speciation by sexual conflict. PNAS. 99(16):10533–10538.

Gelbaya TA, Potdar N, Jeve YB, Nardo LG. 2014. Definition and Epidemiology of Unexplained Infertility. Obstetrical & Gynecological Survey. 69(2). https://journals.lww.com/obgynsurvey/Fulltext/2014/02000/Definition_and_Epidemiology_of_Unexplained.17.aspx.

Gibbs GM, Orta G, Reddy T, Koppers AJ, Martínez-López P, Luis de la Vega-Beltràn J, Lo JCY, Veldhuis N, Jamsai D, McIntyre P, et al. 2011. Cysteine-rich secretory protein 4 is an inhibitor of

transient receptor potential M8 with a role in establishing sperm function. Proc Natl Acad Sci USA. 108(17):7034. doi:10.1073/pnas.1015935108.

Gilbert SF. 2000. Gamete Fusion and the Prevention of Polyspermy. In: Developmental Biology. 6th ed. Sunderland, MA: Sinauer Associates. https://www.ncbi.nlm.nih.gov/books/NBK10033/?log$=activity.

Goudet G, Mugnier S, Callebaut I, Monget P. 2008. Phylogenetic Analysis and Identification of Pseudogenes Reveal a Progressive Loss of Zona Pellucida Genes During Evolution of Vertebrates1. Biology of Reproduction. 78(5):796–806. doi:10.1095/biolreprod.107.064568.

Grayson P. 2015. Izumo1 and Juno: the evolutionary origins and coevolution of essential sperm–egg binding partners. Royal Society Open Science. 2(12):150296. doi:10.1098/rsos.150296.

Grayson P, Civetta A. 2012. Positive Selection and the Evolution of izumo Genes in Mammals. Kulathinal R, editor. International Journal of Evolutionary Biology. 2012:958164. doi:10.1155/2012/958164.

Greenbaum D, Colangelo C, Williams K, Gerstein M. 2003. Comparing protein abundance and mRNA expression levels on a genomic scale. Genome Biology. 4(9):117. doi:10.1186/gb-2003-4-9-117.

Gruber AR, Lorenz R, Bernhart SH, Neuböck R, Hofacker IL. 2008. The Vienna RNA Websuite. Nucleic Acids Research. 36(suppl_2):W70–W74. doi:10.1093/nar/gkn188.

Gu J, Wang Z, Kuen J, Ma L, Shahroudy A, Shuai B, Liu T, Wang X, Wang G, Cai J, et al. 2018. Recent advances in convolutional neural networks. Pattern Recognition. 77:354–377. doi:10.1016/j.patcog.2017.10.013.

Guasti PN, Souza FF, Scott C, Papa PM, Camargo LS, Schmith RA, Monteiro GA, Hartwig FP, Papa FO. 2020. Equine seminal plasma and sperm membrane: Functional proteomic assessment. Theriogenology. 156:70–81. doi:10.1016/j.theriogenology.2020.06.014.

Guna A, Hegde RS. 2018. Transmembrane Domain Recognition during Membrane Protein Biogenesis and Quality Control. Current Biology. 28(8):R498–R511. doi:10.1016/j.cub.2018.02.004.

Guo L, Wang S, Li M, Cao Z. 2019. Accurate classification of membrane protein types based on sequence and evolutionary information using deep learning. BMC Bioinformatics. 20(25):700. doi:10.1186/s12859-019-3275-6.

Gupta SK. 2018. Chapter Twelve - The Human Egg’s Zona Pellucida. In: Litscher ES, Wassarman PM, editors. Current Topics in Developmental Biology. Vol. 130. Academic Press. p. 379–411. https://www.sciencedirect.com/science/article/pii/S0070215318300012.

Hafford-Tear NJ, Tsai Y-C, Sadan AN, Sanchez-Pintado B, Zarouchlioti C, Maher GJ, Liskova P, Tuft SJ, Hardcastle AJ, Clark TA, et al. 2019. CRISPR/Cas9-targeted enrichment and long-read sequencing of the Fuchs endothelial corneal dystrophy–associated TCF4 triplet repeat. Genetics in Medicine. 21(9):2092–2102. doi:10.1038/s41436-019-0453-x.

Han J-H, Batey S, Nickson AA, Teichmann SA, Clarke J. 2007. The folding and evolution of multidomain proteins. Nature Reviews Molecular Cell Biology. 8(4):319–330. doi:10.1038/nrm2144.

Han L, Nishimura K, Sadat Al Hosseini H, Bianchi E, Wright GJ, Jovine L. 2016. Divergent evolution of vitamin B9 binding underlies Juno-mediated adhesion of mammalian gametes. Curr Biol. 26(3):R100–R101. doi:10.1016/j.cub.2015.12.034.

Hanif MS, Bilal M. 2020. Competitive residual neural network for image classification. ICT Express. 6(1):28–37. doi:10.1016/j.icte.2019.06.001.

Hart M, Stover D, Guerra V, V Mozaffari S, Ober C, Mugal C, Kaj I. 2018. Positive selection on human gamete-recognition genes.

Hart MW. 2013. Structure and evolution of the sea star egg receptor for sperm bindin. Molecular Ecology. 22(8):2143–2156. doi:10.1111/mec.12251.

Hartgers FC, Looman MWG, van der Woning B, Merkx GFM, Figdor CG, Adema GJ. 2001. Genomic organization, chromosomal localization, and 5′ upstream region of the human DC-STAMP gene. Immunogenetics. 53(2):145–149. doi:10.1007/s002510100302.

Hartgers FC, Vissers JLM, Looman MWG, Zoelen C van, Huffine C, Figdor CG, Adema GJ. 2000. DC-STAMP, a novel multimembrane-spanning molecule preferentially expressed by dendritic cells. European Journal of Immunology. 30(12):3585–3590. doi:https://doi.org/10.1002/1521-4141(200012)30:12<3585::AID-IMMU3585>3.0.CO;2-Y.

Hastie T, Tibshirani R, Friedman J. 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Second Edition. New York: Springer (Springer Series in Statistics).

Hawkins DM. 2004. The Problem of Overfitting. J Chem Inf Comput Sci. 44(1):1–12. doi:10.1021/ci0342472.

He K, Zhang X, Ren S, Sun. 2016. Deep Residual Learning for Image Recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). p. 770–778.

Herberg S, Gert KR, Schleiffer A, Pauli A. 2018. The Ly6/uPAR protein Bouncer is necessary and sufficient for species-specific fertilization. Science. 361(6406):1029. doi:10.1126/science.aat7113.

Holland IB. 2004. Translocation of bacterial proteins—an overview. Biochimica et Biophysica Acta (BBA) - Molecular Cell Research. 1694(1):5–16. doi:10.1016/j.bbamcr.2004.02.007.

Howe KL, Achuthan P, Allen James, Allen Jamie, Alvarez-Jarreta J, Amode MR, Armean IM, Azov AG, Bennett R, Bhai J, et al. 2021. Ensembl 2021. Nucleic Acids Research. 49(D1):D884–D891. doi:10.1093/nar/gkaa942.

Hu S, Yao Y, Hu X, Zhu Y. 2020. LncRNA DCST1-AS1 downregulates miR-29b through methylation in glioblastoma (GBM) to promote cancer cell proliferation. Clinical and Translational Oncology. 22(12):2230–2235. doi:10.1007/s12094-020-02363-1.

Huang KK, Huang J, Wu JKL, Lee M, Tay ST, Kumar V, Ramnarayanan K, Padmanabhan N, Xu C, Tan ALK, et al. 2021. Long-read transcriptome sequencing reveals abundant promoter diversity in distinct molecular subtypes of gastric cancer. Genome Biology. 22(1):44. doi:10.1186/s13059-021-02261-x.

Hughes AL. 1994. The evolution of functionally novel proteins after gene duplication. Proceedings of the Royal Society of London Series B: Biological Sciences. 256(1346):119–124. doi:10.1098/rspb.1994.0058.

Hwang B, Lee JH, Bang D. 2018. Single-cell RNA sequencing technologies and bioinformaticspipelines. Experimental & Molecular Medicine. 50(8):1–14. doi:10.1038/s12276-018-0071-8.

Innan H. 2009. Population genetic models of duplicated genes. Genetica. 137(1):19. doi:10.1007/s10709-009-9355-1.

Inoue N, Hagihara Y, Wada I. 2021. Evolutionarily conserved sperm factors, DCST1 and DCST2, are required for gamete fusion. Elife. 10:e66313. doi:10.7554/eLife.66313.

Inoue N, Ikawa M, Isotani A, Okabe M. 2005. The immunoglobulin superfamily protein Izumo is required for sperm to fuse with eggs. Nature. 434(7030):234–238. doi:10.1038/nature03362.

Inoue N, Satouh Y, Wada I. 2021. IZUMO family member 3, IZUMO3, is involved in male fertility through the acrosome formation. Molecular Reproduction and Development. 88(7):479–481. doi:10.1002/mrd.23520.

Jaiganesh A, Narui Y, Araya-Secchi R, Sotomayor M. 2018. Beyond Cell-Cell Adhesion: Sensational Cadherins for Hearing and Balance. Cold Spring Harb Perspect Biol. 10(9):a029280. doi:10.1101/cshperspect.a029280.

Jansen BJH, Eleveld-Trancikova D, Sanecka A, van Hout-Kuijer M, Hendriks IAM, Looman MGW, Leusen JHW, Adema GJ. 2009. OS9 interacts with DC-STAMP and modulates its intracellular localization in response to TLR ligation. Molecular Immunology. 46(4):505–515. doi:10.1016/j.molimm.2008.06.032.

Jean C, Haghighirad F, Zhu Y, Chalbi M, Ziyyat A, Rubinstein E, Gourier C, Yip P, Wolf JP, Lee JE, et al. 2019. JUNO, the receptor of sperm IZUMO1, is expressed by the human oocyte and is essential for human fertilisation. Human Reproduction. 34(1):118–126. doi:10.1093/humrep/dey340.

Jovine L, Darie CC, Litscher ES, Wassarman PM. 2005. ZONA PELLUCIDA DOMAIN PROTEINS. Annu Rev Biochem. 74(1):83–114. doi:10.1146/annurev.biochem.74.082803.133039.

Jovine L, Janssen WG, Litscher ES, Wassarman PM. 2006. The PLAC1-homology region of the ZP domain is sufficient for protein polymerisation. BMC Biochemistry. 7(1):11. doi:10.1186/1471-2091-7-11.

Jovine L, Qi H, Williams Z, Litscher E, Wassarman PM. 2002. The ZP domain is a conserved module for polymerization of extracellular proteins. Nature Cell Biology. 4(6):457–461. doi:10.1038/ncb802.

Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A, et al. 2021 Jul 15. Highly accurate protein structure prediction with AlphaFold. Nature. doi:10.1038/s41586-021-03819-2. https://doi.org/10.1038/s41586-021-03819-2.

Kafri R, Springer M, Pilpel Y. 2009. Genetic Redundancy: New Tricks for Old Genes. Cell. 136(3):389–392. doi:10.1016/j.cell.2009.01.027.

Kamei N, Glabe CG. 2003. The species-specific egg receptor for sea urchin sperm adhesion is EBR1,a novel ADAMTS protein. Genes Dev. 17(20):2502–2507. doi:10.1101/gad.1133003.

Katoh K, Standley DM. 2013. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Molecular Biology and Evolution. 30(4):772–780. doi:10.1093/molbev/mst010.

Kelleher ES, Markow TA. 2009. Duplication, Selection and Gene Conversion in a Drosophila mojavensis Female Reproductive Protein Family. Genetics. 181(4):1451–1465. doi:10.1534/genetics.108.099044.

Kelleher ES, Swanson WJ, Markow TA. 2007. Gene Duplication and Adaptive Evolution of Digestive Proteases in Drosophila arizonae Female Reproductive Tracts. PLOS Genetics. 3(8):e148. doi:10.1371/journal.pgen.0030148.

Kessler P, Marchot P, Silva M, Servent D. 2017. The three-finger toxin fold: a multifunctional structural scaffold able to modulate cholinergic functions. Journal of Neurochemistry. 142(S2):7–18. doi:10.1111/jnc.13975.

Killingbeck EE, Swanson WJ. 2018. Chapter Fourteen - Egg Coat Proteins Across Metazoan Evolution. In: Litscher ES, Wassarman PM, editors. Current Topics in Developmental Biology. Vol. 130. Academic Press. p. 443–488. https://www.sciencedirect.com/science/article/pii/S0070215318300486.

Kim D-K, Kim JA, Park J, Niazi A, Almishaal A, Park S. 2019. The release of surface-anchored α-tectorin, an apical extracellular matrix protein, mediates tectorial membrane organization. Sci Adv. 5(11):eaay6300–eaay6300. doi:10.1126/sciadv.aay6300.

Kini RM. 2002. Molecular moulds with multiple missions: Functional sites in three-finger toxins. Clinical and Experimental Pharmacology and Physiology. 29(9):815–822. doi:10.1046/j.1440-1681.2002.03725.x.

Kirschner MW, Gerhart JC. 2008. The Plausibility of Life: Resolving Darwin?s Dilemma. Yale University Press. https://doi.org/10.12987/9780300128673.

Kokko H, Jennions MD. 2014. The relationship between sexual selection and sexual conflict. Cold Spring Harb Perspect Biol. 6(9):a017517. doi:10.1101/cshperspect.a017517.

Kondrashov FA, Rogozin IB, Wolf YI, Koonin EV. 2002. Selection in the evolution of gene duplications. Genome Biology. 3(2):research0008.1. doi:10.1186/gb-2002-3-2-research0008.

Kozlov AM, Darriba D, Flouri T, Morel B, Stamatakis A. 2019. RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. Bioinformatics. 35(21):4453–4455. doi:10.1093/bioinformatics/btz305.

Kresge N, Vacquier V, Stout C. 2000. 1.35 and 2.07 A resolution structures of the red abalone sperm lysin monomer and dimer reveal features involved in receptor binding. Acta Cyrstallogr D Biol Crystallogr. 56(Pt 1):34–41.

Kroft TL, Gleason EJ, L'Hernault SW. 2005. The spe-42 gene is required for sperm–egg interactions during C. elegans fertilization and encodes a sperm-specific transmembrane protein. Developmental Biology. 286(1):169–181. doi:10.1016/j.ydbio.2005.07.020.

Krogh A, Larsson B, von Heijne G, Sonnhammer ELL. 2001. Predicting transmembrane protein topology with a hidden markov model: application to complete genomes11Edited by F. Cohen. Journal of Molecular Biology. 305(3):567–580. doi:10.1006/jmbi.2000.4315.

Kronenberg ZN, Fiddes IT, Gordon D, Murali S, Cantsilieris S, Meyerson OS, Underwood JG, Nelson BJ, Chaisson MJP, Dougherty ML, et al. 2018. High-resolution comparative analysis of great ape genomes. Science. 360(6393):eaar6343. doi:10.1126/science.aar6343.

Kukita T, Wada N, Kukita A, Kakimoto T, Sandra F, Toh K, Nagata K, Iijima T, Horiuchi M, Matsusaki H, et al. 2004. RANKL-induced DC-STAMP Is Essential for Osteoclastogenesis. Journal of Experimental Medicine. 200(7):941–946. doi:10.1084/jem.20040518.

Kumar S, Stecher G, Li M, Knyaz C, Tamura K. 2018. MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. Molecular Biology and Evolution. 35(6):1547–1549. doi:10.1093/molbev/msy096.

Kumar S, Stecher G, Tamura K. 2016. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. Molecular Biology and Evolution. 33(7):1870–1874. doi:10.1093/molbev/msw054.

Kursa MB. 2014. Robustness of Random Forest-based gene selection methods. BMC Bioinformatics. 15(1):8. doi:10.1186/1471-2105-15-8.

Le Naour François, Rubinstein Eric, Jasmin Claude, Prenant Michel, Boucheix Claude. 2000. Severely Reduced Female Fertility in CD9-Deficient Mice. Science. 287(5451):319–321. doi:10.1126/science.287.5451.319.

Legan PK, Rau A, Keen JN, Richardson GP. 1997. The Mouse Tectorins: MODULAR MATRIX PROTEINS OF THE INNER EAR HOMOLOGOUS TO COMPONENTS OF THE SPERM-EGG ADHESION SYSTEM *. Journal of Biological Chemistry. 272(13):8791–8801. doi:10.1074/jbc.272.13.8791.

Leighton DL, Lewis CA. 1982. Experimental hybridization in abalones. null. 5(5):273–282. doi:10.1080/01651269.1982.10553479.

Lemoine F, Domelevo Entfellner J-B, Wilkinson E, Correia D, Dávila Felipe M, De Oliveira T, Gascuel O. 2018. Renewing Felsenstein's phylogenetic bootstrap in the era of big data. Nature. 556(7702):452–456. doi:10.1038/s41586-018-0043-0.

Lewis Cindy A, Talbot CF, Vacquier V. 1982. A Protein from Abalone Sperm Dissolves the Egg Vitelline Layer by a Nonenzymatic Mechanism. Dev Biol. 92:227–239.

Lewis Cindy A., Talbot CF, Vacquier VD. 1982. A protein from abalone sperm dissolves the egg vitelline layer by a nonenzymatic mechanism. Developmental Biology. 92(1):227–239. doi:10.1016/0012-1606(82)90167-1.

Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 34(18):3094–3100. doi:10.1093/bioinformatics/bty191.

Li W, Godzik A. 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics. 22(13):1658–1659. doi:10.1093/bioinformatics/btl158.

Liang L-F, Dean J. 1993. Conservation of Mammalian Secondary Sperm Receptor Genes Enables the Promoter of the Human Gene to Function in Mouse Oocytes. Developmental Biology. 156(2):399–408. doi:10.1006/dbio.1993.1087.

Liao D. 1999. Concerted Evolution: Molecular Mechanism and Biological Implications. The American Journal of Human Genetics. 64(1):24–30. doi:10.1086/302221.

Lillie FR. 1919. Problems of Fertilization. University of Chicago Press (University of Chicago science series). https://books.google.com/books?id=p5ANAAAAYAAJ.

Lin S, Hu Y, Zhu J, Woodruff T, Jardetzky T. 2011. Structure of betaglycan zona pellucida (ZP)-C domain provides insights into ZP-mediated protein polymerization and TGF- binding. Proceedings of The National Academy of Sciences - PNAS. 108:5232–5236. doi:10.1073/pnas.1010689108.

Litscher ES, Wassarman PM. 2007. Egg extracellular coat proteins: from fish to mammals. Histol Histopathol. 22(3):337–347. doi:10.14670/HH-22.337.

Litscher ES, Wassarman PM. 2020. Zona Pellucida Proteins, Fibrils, and Matrix. Annu Rev Biochem. 89(1):695–715. doi:10.1146/annurev-biochem-011520-105310.

Loeb J. 1915. On the Nature of the Conditions Which Determine or Prevent the Entrance of the Spermatozoon Into the Egg. The American Naturalist. 49(581):257–285.

Lorenz R, Bernhart SH, Höner zu Siederdissen C, Tafer H, Flamm C, Stadler PF, Hofacker IL. 2011. ViennaRNA Package 2.0. Algorithms for Molecular Biology. 6(1):26. doi:10.1186/1748-7188-6-26.

Low BW, Preston HS, Sato A, Rosen LS, Searl JE, Rudko AD, Richardson JS. 1976. Three dimensional structure of erabutoxin b neurotoxic protein: inhibitor of acetylcholine receptor. Proc Natl Acad Sci U S A. 73(9):2991–2994. doi:10.1073/pnas.73.9.2991.

Lynch M, Conery JS. 2000. The Evolutionary Fate and Consequences of Duplicate Genes. Science. 290(5494):1151–1155. doi:10.1126/science.290.5494.1151.

MacDonald RJ, Swift GH, Przybyla AE, Chirgwin JM. 1987. Isolation of RNA using guanidinium salts. Methods Enzymol. 152:219–227. doi:10.1016/0076-6879(87)52023-7.

Maldera JA, Weigel Muñoz M, Chirinos M, Busso D, GE Raffo F, Battistone MA, Blaquier JA, Larrea F, Cuasnicu PS. 2014. Human fertilization: epididymal hCRISP1 mediates sperm–zona pellucida binding through its interaction with ZP3. Molecular Human Reproduction. 20(4):341–349. doi:10.1093/molehr/gat092.

Masonbrink RE, Purcell CM, Boles SE, Whitehead A, Hyde JR, Seetharam AS, Severin AJ. 2019. An Annotated Genome for Haliotis rufescens (Red Abalone) and Resequenced Green, Pink, Pinto, Black, and White Abalone Species. Genome Biology and Evolution. 11(2):431–438. doi:10.1093/gbe/evz006.

Mathews DH, Disney MD, Childs JL, Schroeder SJ, Zuker M, Turner DH. 2004. Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. Proceedings of the National Academy of Sciences. 101(19):7287–7292. doi:10.1073/pnas.0401799101.

Meslin C, Mugnier S, Callebaut I, Laurin M, Pascal G, Poupon A, Goudet G, Monget P. 2012. Evolution of Genes Involved in Gamete Interaction: Evidence for Positive Selection, Duplications and Losses in Vertebrates. PLOS ONE. 7(9):e44548. doi:10.1371/journal.pone.0044548.

Mirdita M, Schütze K, Moriwaki Y, Heo L, Ovchinnikov S, Steinegger M. 2022. ColabFold: making protein folding accessible to all. Nature Methods. 19(6):679–682. doi:10.1038/s41592-022-01488-1.

Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer ELL, Tosatto SCE, Paladin L, Raj S, Richardson LJ, et al. 2021. Pfam: The protein families database in 2021. Nucleic Acids Research. 49(D1):D412–D419. doi:10.1093/nar/gkaa913.

Miyado Kenji, Yamada Gen, Yamada Shuichi, Hasuwa Hidetoshi, Nakamura Yasuhiro, Ryu Fuminori, Suzuki Kentaro, Kosai Kenichiro, Inoue Kimiko, Ogura Atsuo, et al. 2000. Requirement of CD9 on the Egg Plasma Membrane for Fertilization. Science. 287(5451):321–324. doi:10.1126/science.287.5451.321.

Möller S, Croning MDR, Apweiler R. 2001. Evaluation of methods for the prediction of membrane spanning regions. Bioinformatics. 17(7):646–653. doi:10.1093/bioinformatics/17.7.646.

Monne M, Han L, Schwend T, Burendahl S, Jovine L. 2008. Crystal structure of the ZP-N domain of ZP3 reveals the core fold of animal egg coats. Nature. 456(7222):653–7.

Moore AD, Björklund ÅK, Ekman D, Bornberg-Bauer E, Elofsson A. 2008. Arrangements in the modular evolution of proteins. Trends in Biochemical Sciences. 33(9):444–451. doi:10.1016/j.tibs.2008.05.008.

Moy G, Springer S, Adams S, Swanson W, Vacquier V. 2008. Extraordinary intraspecific diversity in oyster sperm bindin. Proc Natl Acad Sci USA. 105(6):1993–8.

Nacher JC, Hayashida M, Akutsu T. 2010. The role of internal duplication in the evolution of multi-domain proteins. Biosystems. 101(2):127–135. doi:10.1016/j.biosystems.2010.05.005.

Nair S, Bist P, Dikshit N, Krishnan MN. 2016. Global functional profiling of human ubiquitome identifies E3 ubiquitin ligase DCST1 as a novel negative regulator of Type-I interferon signaling. Scientific Reports. 6(1):36179. doi:10.1038/srep36179.

Navarro B, Kirichok Y, Chung J-J, Clapham DE. 2008. Ion channels that control fertility in mammalian spermatozoa. Int J Dev Biol. 52(5–6):607–613. doi:10.1387/ijdb.072554bn.

Nirthanan S, Gopalakrishnakone P, Gwee MCE, Khoo HE, Kini RM. 2003. Non-conventional toxins from Elapid venoms. Toxicon. 41(4):397–407. doi:10.1016/S0041-0101(02)00388-4.

Nishimura K, Dioguardi E, Nishio S, Villa A, Han L, Matsuda T, Jovine L. 2019. Molecular basis of egg coat cross-linking sheds light on ZP1-associated female infertility. Nat Commun. 10(1):3086–3086. doi:10.1038/s41467-019-10931-5.

Nomiyama H, Egami K, Wada N, Tou K, Horiuchi M, Matsusaki H, Miura R, Yoshie O, Kukita T. 2005. Short Communication: Identification of Genes Differentially Expressed in Osteoclast-like Cells. Journal of Interferon & Cytokine Research. 25(4):227–231. doi:10.1089/jir.2005.25.227.

Nurk S, Koren S, Rhie A, Rautiainen M, Bzikadze AV, Mikheenko A, Vollger MR, Altemose N, Uralsky L, Gershman A, et al. 2022. The complete sequence of a human genome. Science. 376(6588):44–53. doi:10.1126/science.abj6987.

Ohto U, Ishida H, Krayukhina E, Uchiyama S, Inoue N, Shimizu T. 2016. Structure of IZUMO1–JUNO reveals sperm–oocyte recognition during mammalian fertilization. Nature. 534(7608):566–569. doi:10.1038/nature18596.

OKUMURA H, KOHNO Y, IWATA Y, MORI H, AOKI N, SATO C, KITAJIMA K, NADANO D, MATSUDA T. 2004. A newly identified zona pellucida glycoprotein, ZPD, and dimeric ZP1 of chicken egg envelope are involved in sperm activation on sperm–egg interaction. Biochemical Journal. 384(1):191–199. doi:10.1042/BJ20040299.

Palmer CA, Hollis DM, Watts RA, Houck LD, McCall MA, Gregg RG, Feldhoff PW, Feldhoff RC, Arnold SJ. 2007. Plethodontid modulating factor, a hypervariable salamander courtship pheromone in the three-finger protein superfamily. The FEBS Journal. 274(9):2300–2310. doi:10.1111/j.1742-4658.2007.05766.x.

Palmer CA, Watts RA, Hastings AP, Houck LD, Arnold SJ. 2010. Rapid Evolution of Plethodontid Modulating Factor, a Hypervariable Salamander Courtship Pheromone, is Driven by Positive Selection. Journal of Molecular Evolution. 70(5):427–440. doi:10.1007/s00239-010-9342-2.

Palmer CA, Watts RA, Houck LD, Picard AL, Arnold SJ. 2007. EVOLUTIONARY REPLACEMENT OF COMPONENTS IN A SALAMANDER PHEROMONE SIGNALING COMPLEX: MORE EVIDENCE FOR PHENOTYPIC-MOLECULAR DECOUPLING. Evolution. 61(1):202–215. doi:10.1111/j.1558-5646.2007.00017.x.

Palmer MR, McDowall MH, Stewart L, Ouaddi A, MacCoss MJ, Swanson WJ. 2013. Mass spectrometry and next-generation sequencing reveal an abundant and rapidly evolving abalone sperm protein. Molecular Reproduction and Development. 80(6):460–465. doi:10.1002/mrd.22182.

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, et al. 2011. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research. 12:2825–2830.

Pei J, Kim B-H, Grishin NV. 2008. PROMALS3D: a tool for multiple protein sequence and structure alignments. Nucleic Acids Res. 36(7):2295–2300. doi:10.1093/nar/gkn072.

Perry GH, Dominy NJ, Claw KG, Lee AS, Fiegler H, Redon R, Werner J, Villanea FA, Mountain JL, Misra R, et al. 2007. Diet and the evolution of human amylase gene copy number variation. Nature Genetics. 39(10):1256–1260. doi:10.1038/ng2123.

Petronella N, Drouin G. 2014. Purifying selection against gene conversions in the folate receptor genes of primates. Genomics. 103(1):40–47. doi:10.1016/j.ygeno.2013.10.004.

Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, Ferrin TE. 2004. UCSF Chimera—a visualization system for exploratory research and analysis. Journal of computational chemistry. 25(13):1605–1612.

Pohlschröder M, Dilks K, Hand NJ, Wesley Rose R. 2004. Translocation of proteins across archaeal cytoplasmic membranes. FEMS Microbiology Reviews. 28(1):3–24. doi:10.1016/j.femsre.2003.07.004.

Pohlschröder M, Giménez MI, Jarrell KF. 2005. Protein transport in Archaea: Sec and twin arginine translocation pathways. Current Opinion in Microbiology. 8(6):713–719. doi:10.1016/j.mib.2005.10.006.

Ponting CP. 2008. The functional repertoires of metazoan genomes. Nature Reviews Genetics. 9(9):689–698. doi:10.1038/nrg2413.

Primakoff P, Myles DG. 2000. The ADAM gene family: surface proteins with adhesion and protease activity. Trends in Genetics. 16(2):83–87. doi:10.1016/S0168-9525(99)01926-5.

Quail MA, Swerdlow H, Turner DJ. 2009. Improved protocols for the illumina genome analyzer sequencing system. Curr Protoc Hum Genet. Chapter 18:Unit 18.2. doi:10.1002/0471142905.hg1802s62.

Raj I, Sadat Al Hosseini H, Dioguardi E, Nishimura K, Han L, Villa A, de Sanctis D, Jovine L. 2017. Structural Basis of Egg Coat-Sperm Recognition at Fertilization. Cell. 169(7):1315-1326.e17. doi:10.1016/j.cell.2017.05.033.

Ramírez-Gómez HV, Tuval I, Guerrero A, Darszon A. 2019. Chapter 21 - Analysis of sperm chemotaxis. In: Hamdoun A, Foltz KR, editors. Methods in Cell Biology. Vol. 151. Academic Press. p. 473–486. https://www.sciencedirect.com/science/article/pii/S0091679X18301821.

Rankin T, Coleman J, Epifano O, Hoodbhoy T, Turner S, Castle P, Lee E, Gore-Langton R, Dean J. 2003. Fertility and Taxon-Specific Sperm Binding Persist after Replacement of Mouse Sperm Receptors with Human Homologs. Dev Cell. 5(1):33–43.

Rastogi S, Liberles DA. 2005. Subfunctionalization of duplicated genes as a transition state to neofunctionalization. BMC Evolutionary Biology. 5(1):28. doi:10.1186/1471-2148-5-28.

Rivera AM, Swanson WJ. 2022. The Importance of Gene Duplication and Domain Repeat Expansion for the Function and Evolution of Fertilization Proteins. Frontiers in Cell and Developmental Biology. 10. https://www.frontiersin.org/article/10.3389/fcell.2022.827454.

Rivera AM, Wilburn DB, Swanson WJ. 2022. Domain Expansion and Functional Diversification in Vertebrate Reproductive Proteins. Molecular Biology and Evolution. 39(5):msac105. doi:10.1093/molbev/msac105.

Rivero F, Cvrčková F. 2007. Origins and Evolution of the Actin Cytoskeleton. In: Eukaryotic Membranes and Cytoskeleton: Origins and Evolution. New York, NY: Springer New York. p. 97–110. https://doi.org/10.1007/978-0-387-74021-8_8.

Sahlin K, Medvedev P. 2019. De Novo Clustering of Long-Read Transcriptome Data Using a Greedy, Quality-Value Based Algorithm. In: Cowen LJ, editor. Research in Computational Molecular Biology. Cham: Springer International Publishing. p. 227–242.

Sano K, Kawaguchi M, Watanabe S, Nagakura Y, Hiraki T, Yasumasu S. 2013. Inferring the evolution of teleostean zp genes based on their sites of expression. J Exp Zool B Mol Dev Evol. 320(5):332–343. doi:10.1002/jez.b.22507.

Schimenti JC. 1999. Mice and the Role of Unequal Recombination in Gene-Family Evolution. The American Journal of Human Genetics. 64(1):40–45. doi:10.1086/302220.

Schmidhuber J. 2015. Deep learning in neural networks: An overview. Neural Networks. 61:85–117. doi:10.1016/j.neunet.2014.09.003.

Schneider TD, Stephens RM. 1990. Sequence logos: a new way to display consensus sequences. Nucleic Acids Research. 18(20):6097–6100. doi:10.1093/nar/18.20.6097.

Schrimpf SP, Weiss M, Reiter L, Ahrens CH, Jovanovic M, Malmström J, Brunner E, Mohanty S, Lercher MJ, Hunziker PE, et al. 2009. Comparative Functional Analysis of the Caenorhabditis elegans and Drosophila melanogaster Proteomes. PLOS Biology. 7(3):e1000048. doi:10.1371/journal.pbio.1000048.

Schrödinger L. 2015. The PyMOL Molecular Graphics System, Version 1.8.

Shao S, Hegde RS. 2011. Membrane Protein Insertion at the Endoplasmic Reticulum. Annu Rev Cell Dev Biol. 27(1):25–56. doi:10.1146/annurev-cellbio-092910-154125.

Shen F, Ross JF, Wang X, Ratnam M. 1994. Identification of a novel folate receptor, a truncated receptor, and receptor type .beta. in hematopoietic cells: cDNA cloning, expression, immunoreactivity, and tissue specificity. Biochemistry. 33(5):1209–1215. doi:10.1021/bi00171a021.

Shu L, Suter MJ-F, Räsänen K. 2015. Evolution of egg coats: linking molecular biology and ecology. Molecular Ecology. 24(16):4052–4073. doi:10.1111/mec.13283.

Sircar A, Chaudhury S, Kilambi KP, Berrondo M, Gray JJ. 2010. A generalized approach to sampling backbone conformations with RosettaDock for CAPRI rounds 13–19. Proteins: Structure, Function, and Bioinformatics. 78(15):3115–3123. doi:10.1002/prot.22765.

Siu KK, Serrão VHB, Ziyyat A, Lee JE. 2021. The cell biology of fertilization: Gamete attachment and fusion. Journal of Cell Biology. 220(10). doi:10.1083/jcb.202102146. [accessed 2021 Nov 26]. https://doi.org/10.1083/jcb.202102146.

Sonnhammer EL, von Heijne G, Krogh A. 1998. A hidden Markov model for predicting transmembrane helices in protein sequences. Proc Int Conf Intell Syst Mol Biol. 6:175–182.

de Sousa Abreu R, Penalva LO, Marcotte EM, Vogel C. 2009. Global signatures of protein and mRNA expression levels. Mol Biosyst. 5(12):1512–1526. doi:10.1039/b908315d.

Speer KF, Allen-Waller L, Novikov DR, Barott KL. 2021. Molecular mechanisms of sperm motility are conserved in an early-branching metazoan. Proc Natl Acad Sci USA. 118(48):e2109993118. doi:10.1073/pnas.2109993118.

Spiegelstein O, Eudy JD, Finnell RH. 2000. Identification of two putative novel folate receptor genes in humans and mouse. Gene. 258(1):117–125. doi:10.1016/S0378-1119(00)00418-2.

Staege H, Brauchlin A, Schoedon G, Schaffner A. 2001. Two novel genes FIND and LIND differentially expressed in deactivated and Listeria-infected human macrophages. Immunogenetics. 53(2):105–113. doi:10.1007/s002510100306.

Sutton KA, Jungnickel MK, Florman HM. 2008. A polycystin-1 controls postcopulatory reproductive selection in mice. Proc Natl Acad Sci USA. 105(25):8661. doi:10.1073/pnas.0800603105.

Swanson W, Vacquier V. 2002. Rapid evolution of reproductive proteins. Nature Review Genetics. 3:137–144.

Swanson W, Yang Ziheng, Wolfner Mariana F., Aquadro Charles F. 2001. Positive Darwinian selection drives the evolution of several female reproductive proteins in mammals. Proceedings of the National Academy of Sciences. 98(5):2509–2514. doi:10.1073/pnas.051605998.

Swanson WJ, Clark AG, Waldrip-Dail HM, Wolfner MF, Aquadro CF. 2001. Evolutionary EST analysis identifies rapidly evolving male reproductive proteins in Drosophila. Proceedings of the National Academy of Sciences. 98(13):7375–7379. doi:10.1073/pnas.131568198.

The UniProt Consortium. 2021. UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Research. 49(D1):D480–D489. doi:10.1093/nar/gkaa1100.

Thompson K. 1968. Programming Techniques: Regular Expression Search Algorithm. Commun ACM. 11(6):419–422. doi:10.1145/363347.363387.

Ting C-T, Tsaur S-C, Sun S, Browne WE, Chen Y-C, Patel NH, Wu C-I. 2004. Gene duplication and speciation in Drosophila: Evidence from the Odysseus locus. Proc Natl Acad Sci U S A. 101(33):12232. doi:10.1073/pnas.0401975101.

Tsernoglou D, Petsko GA. 1977. Three-dimensional structure of neurotoxin a from venom of the Philippines sea snake. Proc Natl Acad Sci U S A. 74(3):971–974. doi:10.1073/pnas.74.3.971.

Tsetlin V. 1999. Snake venom α-neurotoxins and other ‘three-finger’ proteins. European Journal of Biochemistry. 264(2):281–286. doi:10.1046/j.1432-1327.1999.00623.x.

Turner LM, Hoekstra HE. 2006. Adaptive Evolution of Fertilization Proteins within a Genus: Variation in ZP2 and ZP3 in Deer Mice (Peromyscus). Molecular Biology and Evolution. 23(9):1656–1669. doi:10.1093/molbev/msl035.

Tusnády GE, Simon I. 2001. The HMMTOP transmembrane topology prediction server. Bioinformatics. 17(9):849–850. doi:10.1093/bioinformatics/17.9.849.

Vacquier VD. 1998. Evolution of Gamete Recognition Proteins. Science. 281(5385):1995–1998. doi:10.1126/science.281.5385.1995.

Vacquier VD, Carner KR, Stout CD. 1990. Species-specific sequences of abalone lysin, the sperm protein that creates a hole in the egg envelope. Proc Natl Acad Sci USA. 87(15):5792. doi:10.1073/pnas.87.15.5792.

Vacquier VD, Swanson WJ, Lee Y-H. 1997. Positive darwinian selection on two homologous fertilization proteins: what is the selective pressure driving their divergence? Journal of Molecular Evolution. 44(1):S15–S22. doi:10.1007/PL00000049.

Vogel C, Marcotte EM. 2012. Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. Nature Reviews Genetics. 13(4):227–232. doi:10.1038/nrg3185.

Vogel C, Teichmann SA, Pereira-Leal J. 2005. The Relationship Between Domain Duplication and Recombination. Journal of Molecular Biology. 346(1):355–365. doi:10.1016/j.jmb.2004.11.050.

Wagner GP, Altenberg L. 1996. PERSPECTIVE: COMPLEX ADAPTATIONS AND THE EVOLUTION OF EVOLVABILITY. Evolution. 50(3):967–976. doi:10.1111/j.1558-5646.1996.tb02339.x.

Wagner GP, Pavlicev M, Cheverud JM. 2007. The road to modularity. Nature Reviews Genetics. 8(12):921–931. doi:10.1038/nrg2267.

Walsh B. 2003. Population-Genetic Models of the Fates of Duplicate Genes. Genetica. 118(2):279–294. doi:10.1023/A:1024194802441.

Wang J, Lei C, Shi P, Teng H, Lu L, Guo H, Wang X. 2021. LncRNA DCST1-AS1 Promotes Endometrial Cancer Progression by Modulating the MiR-665/HOXB5 and MiR-873-5p/CADM1 Pathways. Frontiers in Oncology. 11:3112. doi:10.3389/fonc.2021.714652.

Wassarman PM, Litscher ES. 2016. Chapter Thirty-One - A Bespoke Coat for Eggs: Getting Ready for Fertilization. In: Wassarman PM, editor. Current Topics in Developmental Biology. Vol. 117. Academic Press. p. 539–552. https://www.sciencedirect.com/science/article/pii/S0070215315001167.

Weadick CJ. 2020. Molecular Evolutionary Analysis of Nematode Zona Pellucida (ZP) Modules Reveals Disulfide-Bond Reshuffling and Standalone ZP-C Domains. Genome Biology and Evolution. 12(8):1240–1255. doi:10.1093/gbe/evaa095.

Weiner 3rd J, Beaussart F, Bornberg-Bauer E. 2006. Domain deletions and substitutions in the modular protein evolution. The FEBS Journal. 273(9):2037–2047. doi:10.1111/j.1742-4658.2006.05220.x.

Wenger AM, Peluso P, Rowell WJ, Chang P-C, Hall RJ, Concepcion GT, Ebler J, Fungtammasan A, Kolesnikov A, Olson ND, et al. 2019. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. Nature Biotechnology. 37(10):1155–1162. doi:10.1038/s41587-019-0217-9.

Wessel GM, Brooks JM, Green E, Haley S, Voronina E, Wong J, Zaydfudim V, Conner S. 2001. The biology of cortical granules. In: International Review of Cytology. Vol. 209. Academic Press. p. 117–206. https://www.sciencedirect.com/science/article/pii/S007476960109012X.

West-Eberhard MJ. 2005. Developmental plasticity and the origin of species differences. Proc Natl Acad Sci USA. 102(suppl 1):6543. doi:10.1073/pnas.0501844102.

Wilburn D, Arnold S, Houck L, Feldhoff P, Feldhoff R. 2017. Gene Duplication, Co-option, Structural Evolution, and Phenotypic Tango in the Courtship Pheromones of Plethodontid Salamanders. Herpetologica. 73:206–219. doi:10.1655/Herpetologica-D-16-00082.1.

Wilburn D, Swanson W. 2016. From molecules to mating: Rapid evolution and biochemical studies of reproductive proteins. J Proteomics. 135:12–25.

Wilburn DB, Bowen KE, Doty KA, Arumugam S, Lane AN, Feldhoff PW, Feldhoff RC. 2014. Structural Insights into the Evolution of a Sexy Protein: Novel Topology and Restricted Backbone Flexibility in a Hypervariable Pheromone from the Red-Legged Salamander, Plethodon shermani. PLOS ONE. 9(5):e96975. doi:10.1371/journal.pone.0096975.

Wilburn DB, Bowen KE, Gregg RG, Cai J, Feldhoff PW, Houck LD, Feldhoff RC. 2012. PROTEOMIC AND UTR ANALYSES OF A RAPIDLY EVOLVING HYPERVARIABLE FAMILY OF VERTEBRATE PHEROMONES. Evolution. 66(7):2227–2239. doi:10.1111/j.1558-5646.2011.01572.x.

Wilburn DB, Doty KA, Chouinard AJ, Eddy SL, Woodley SK, Houck LD, Feldhoff RC. 2017. Olfactory effects of a hypervariable multicomponent pheromone in the red-legged salamander, Plethodon shermani. PLOS ONE. 12(3):e0174370. doi:10.1371/journal.pone.0174370.

Wilburn DB, Kuncel CL, Feldhoff RC, Feldhoff PW, Searle BC. 2022. Recurrent cooption and recombination of cytokine-like and three finger-like proteins throughout salamander reproduction as sperm proteins and courtship pheromones. Frontiers in Cell and Developmental Biology. 10(Special issue on Dynamics and Mechanisms of Sperm-Egg Interaction).

Wilburn DB, Kunkel CL, Feldhoff RC, Feldhoff PW, Searle BC. 2022 Jan 1. Recurrent co-option and recombination of cytokine and three finger proteins in multiple reproductive tissues throughout salamander evolution. bioRxiv.:2022.01.04.475003. doi:10.1101/2022.01.04.475003.

Wilburn DB, Swanson WJ. 2017. The “ZP domain” is not one, but likely two independent domains. Molecular Reproduction and Development. 84(4):284–285. doi:10.1002/mrd.22781.

Wilburn DB, Swanson WJ. 2018. Gamete Structure: Egg, Comparative Vertebrate. In: Skinner MK, editor. Encyclopedia of Reproduction (Second Edition). Oxford: Academic Press. p. 204–209. https://www.sciencedirect.com/science/article/pii/B9780128096338205578.

Wilburn DB, Tuttle LM, Klevit RE, Swanson WJ. 2019. Indirect sexual selection drives rapid sperm protein evolution in abalone. Wittkopp PJ, Neuweiler H, Neuweiler H, Jaudzems K, editors. eLife. 8:e52628. doi:10.7554/eLife.52628.

Wilson KL, Fitch KR, Bafus BT, Wakimoto BT. 2006. Sperm plasma membrane breakdown during Drosophila fertilization requires Sneaky, an acrosomal membrane protein. Development. 133(24):4871–4879. doi:10.1242/dev.02671.

Wilson LD, Obakpolor OA, Jones AM, Richie AL, Mieczkowski BD, Fall GT, Hall RW, Rumbley JN, Kroft TL. 2018. The Caenorhabditis elegans spe-49 gene is required for fertilization and encodes a sperm-specific transmembrane protein homologous to SPE-42. Molecular Reproduction and Development. 85(7):563–578. doi:10.1002/mrd.22992.

Wong JL, Wessel GM. 2005. Defending the Zygote: Search for the Ancestral Animal Block to Polyspermy. In: Current Topics in Developmental Biology. Vol. 72. Academic Press. p. 1–151. https://www.sciencedirect.com/science/article/pii/S0070215305720019.

Wright GJ, Bianchi E. 2016. The challenges involved in elucidating the molecular basis of sperm–egg recognition in mammals and approaches to overcome them. Cell and Tissue Research. 363(1):227–235. doi:10.1007/s00441-015-2243-3.

Yagi M, Miyamoto T, Sawatani Y, Iwamoto K, Hosogane N, Fujita N, Morita K, Ninomiya K, Suzuki T, Miyamoto K, et al. 2005. DC-STAMP is essential for cell–cell fusion in osteoclasts and foreign body giant cells. Journal of Experimental Medicine. 202(3):345–351. doi:10.1084/jem.20050645.

Yang KK, Wu Z, Bedbrook CN, Arnold FH. 2018. Learned protein embeddings for machine learning. Bioinformatics. 34(15):2642–2648. doi:10.1093/bioinformatics/bty178.

Yang Z. 2007. PAML 4: Phylogenetic Analysis by Maximum Likelihood. Molecular Biology and Evolution. 24(8):1586–1591. doi:10.1093/molbev/msm088.

Yang Z, Wong WSW, Nielsen R. 2005. Bayes Empirical Bayes Inference of Amino Acid Sites Under Positive Selection. Molecular Biology and Evolution. 22(4):1107–1118. doi:10.1093/molbev/msi097.

Zou H, Hastie T. 2005. Regularization and Variable Selection via the Elastic Net. Journal of the Royal Statistical Society Series B (Statistical Methodology). 67(2):301–320.