HYPER-PLASTIC STRUCTURAL EVOLUTION OF THE PEX DOMAIN - A MODEL OF EVOLUTIONARY EXAPTATION AND NEOFUNCTIONALIZATION AT THE MOLECULAR LEVEL

A DISSERTATION IN
Molecular Biology and Biochemistry
and
Cell Biology and Biophysics

Presented to the Faculty of the University
of Missouri-Kansas City in partial fulfillment of
the requirements for the degree

DOCTOR OF PHILOSOPHY

by
LEE LIKINS

B.S., University of West Florida, Pensacola, FL 1985
M.A., University of Kansas, Lawrence, KS 1996

Kansas City, Missouri
2017

HYPER-PLASTIC STRUCTURAL EVOLUTION OF THE PEX DOMAIN - A MODEL OF EVOLUTIONARY EXAPTATION AND NEOFUNCTIONALIZATION AT THE MOLECULAR LEVEL

Lee Likins, Candidate for the Doctor of Philosophy Degree

University of Missouri – Kansas City, 2017

ABSTRACT

With the advent of sophisticated genetic, biophysical and *in silico* technology an enormous amount of information is being generated regarding the structural, biochemical and physiological aspects of proteins. Tertiary protein structural domains are assumed to be features of proteins whose bio-historical relationships can be traceable over relevant evolutionary space. The phylogenetics of protein domains, including their genesis, duplication, combination, selectively derived loss and potential horizontal capture to derive novel functional rearrangements, are of great interest to molecular evolutionists. Evolutionary and bioinformatic analyses of a considerable collection of variable protein primary, secondary, tertiary, quaternary and biochemical structures has established the principle that, from a functional perspective, many, if not most, proteins are evolutionarily dependent on the functional capacity of readily defined and identifiable, globular components defined as "domains". Therefore, it is reasonable to infer that specific secondary and tertiary folds, or arrangements, represent functional phenotypic *characters* that can be analyzed to provide insights into the evolutionary history of any given protein domain. The evolution of proteins, at the domain level, has a significant impact

on the overall functionality of metabolic pathways in general. Therefore, insights into the evolutionary trajectories of protein domains have the potential to inform the understanding of every aspect in which any given protein has a role of functional significance including, but certainly not limited to basic metabolic equilibrium, potential physical compromise at the phenotypic level, and deeper insights into corruptions that often lead to metabolic dysfunction and potential progression to full blown disease states. The prime focus of this study is to investigate the evolutionary relationships of proteins, across all Kingdoms of life, that contain within their tertiary phenotypic structure the 4-bladed β-propeller domain (the "Hemopexin" or "PEX" domain), towards illuminating the biophysical and biochemical significance of this specific domain's impact on protein functionality. While the phylogenetic relationships of entire proteins that have PEX domains is relatively straight forward, the mutational accumulation in gene sequences that lead to the tertiary structure of the PEX domain itself seems to have partially, if not entirely, obscured the evolutionary history at the domain level.

The phylogenetic analyses presented here allow for several novel conclusions. First, this research demonstrates that a derived primary amino acid sequence in mammalian Hemopexin proteins (the JEN-14 epitope) represents a functional synapomorphy at the molecular level. Secondly, that there is substantial evidence for horizontal gene transfer of PEX domain proteins into specific Fungi. Additionally, there are no proteins containing a PEX domain in Kingdom Archaea. Lastly, it argues that the PEX domain itself represents an evolutionary “spandrel” (*sensu* Gould and Lewontin) with specifically derived functions existing around a core, preserved structural architecture.

The faculty listed below, appointed by the Dean of the School of Graduate Studies, have examined a dissertation titled "Hyper-Plastic Structural Evolution Of The PEX Domain - A Model Of Evolutionary Exaptation And Neofunctionalization At The Molecular Level," presented by Lee Likins candidate for the Doctor of Philosophy degree, and hereby certify that in their opinion it is worthy of acceptance.

Supervisory Committee

Gerald J. Wyckoff, Ph.D. (Committee Chair)
Department of Molecular Biology
and Biochemistry

Alex Idnurm, Ph.D.,
Department of Cell Biology
and Biophysics

Ann Smith, Ph.D.,
Department of Molecular Biology
and Biochemistry

Jakob Waterborg, Ph.D.,
Department of Cell Biology
and Biophysics

Xiaolan Yao, Ph.D.,
Department of Molecular Biology
and Biochemistry

CONTENTS

ILLUSTRATIONS

Figure Page

TABLES

Table Page

ACKNOWLEDGMENTS

I would like to thank my Committee members Dr. Jerry Wyckoff (Chair), Dr. Alex Idnurm, Dr. Ann Smith, Dr. Jakob Waterborg, and Dr. Xiaolan Yao for their mentorship, support and, above all, patience.

I would also like to especially acknowledge and thank, immensely, Dr. Karen Bame, *every graduate student's 6th Committee member*! Her guidance and encouragement have been invaluable to me and, I know, speaking for all of my fellow grad student colleagues, that her professionalism and empathy have been indispensable in successfully coaching us though this extremely challenging endeavor.

Thanks to the lab of Dr. Samuel Boyain, and especially Jessica Kawakami for permitting us access to, and providing excellent instruction in the proper operation of the AKTA FPLC apparatus; and to Dr. Andrew Keightley for running the Mass Spec analyses used in the pursuit of the structure of the PRG4 PEX-domain in Chapter 4.

A deep, heartfelt thank you and sincere appreciation for the entire faculty and staff of the School of Biological Sciences - UMKC; especially those of you whom I have had the honor of learning from in your graduate courses. As a professional teacher I am humbled and inspired by your skills and dedication to the arts of Science Education, as well as your commitments to Biological Research.

To Dean Denis Medeiros and the entire staff of the School of Graduate Studies - UMKC, for their stalwart guidance and perseverance through the long academic and administrative processes, my sincere gratitude.

Also, many thanks to graduate and undergraduate students in the Yao lab: Megan Ehlinger, Jennifer Prashek, Alex Troxel, Tasnuva Nitu, and Cole McMullin for tutoring me in cloning and expression protocols; Ed Bjes, of the Smith lab, for teaching me to be gentle and respectful of the special needs of competent *E.coli* cells; to Peter Hahl for helping me learn text-based commands in PyMOL; and to my fellow "team members" of the Wyckoff lab: Kristi Bishop, Josh Large, Ean Nelson and Ada Solidar, for all the help, in the many guises you have freely and enthusiastically devoted to this project.

A very special thank you to my friend and colleague Dr. Andrew Skaff. He has been extremely patient while working with me on the wet-bench biochemistry in this research. If not for his diligence and incredible skill set, there is no way we could have gotten as far as we have. And I hope he will continue to join me in many an on-going future scientific pursuit.

I would also like to thank my friend and mentor, Dr. Bibie Chronwall for being such a stalwart supporter and an ever-valued colleague.

And, of course, my deepest gratitude, infinite respect, total adoration and unfathomable love for my very best friend in the whole-wide-world, who also happens to be my wife: Debbie Burke Likins.

To the memory and legacy of my dear friend and valued colleague:

Dr. Douglas J. Law,

Whose suggestion led me to continue my pursuit of the PhD.

You are truly, sorely missed Dougie.

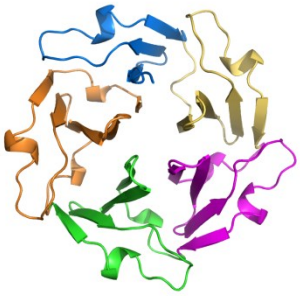
CHAPTER 1

STRUCTURAL ANALYSES OF THE HEMOPEXIN PROTEIN AS A TOOL TO INFER THE EVOLUTION OF FUNCTIONALITY BETWEEN THE PEX DOMAIN CONTAINING PROTEINS

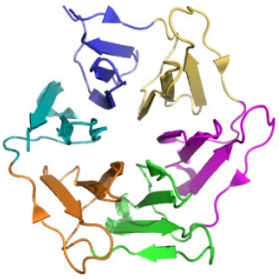
Introduction

From very early in the advent of modern genomics, questions were raised about the observation that many structures at the molecular level, while appearing to be architecturally conserved, were not correlating well with the underlying primary amino acid sequences of the proteins in which they were found. This is notably true when considering the myriad protein domains comprised of the β-propeller fold motif. These domains are constructed by the sequential arrangement of "blades" that are each a single, anti-parallel, β-sheet component, that when combined, make up the propeller configuration. The domains (propellers) may consist of anywhere from 4-8 blades inclusive [1-5]; and there is a single known 10-bladed propeller found in a type I membrane glycoprotein named *sortillin*, that functions in intracellular sorting and endocytotic processes in neurons [6]. Structural models of several protein domains with increasing number of β-propeller blades are available in Research Collaboratory for Structural Bioinformatics Protein Data Base (RCSB PDB - henceforth PDB) [7] **Figure 1**. For more detailed information on these representative proteins see **Table A1**, **Appendix A**.

Hemopexin - N-Domain;
4-bladed

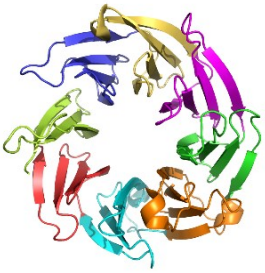


Tachylectin-2;
5-bladed

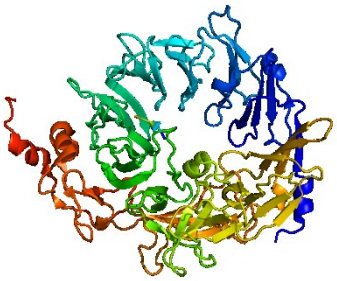


1-RWL;
6-bladed

RACK-1;
7-bladed



4ZOX;
8-bladed



3F6K;
10-bladed

Figure 1. Representative proteins with known variable numbers of β-propeller structural domains. Designations represent PDB ID numbers for each structure.

These analyses will focus exclusively on the evolution of the 4-bladed β-propeller (PEX) domain, exemplified by the dual arrangement seen in the tertiary structure of the Hemopexin (HPX) protein. One principle objective of this study is to use phylogenomic analyses to determine the evolutionary relationship of the PEX domain containing proteins in hopes of uncovering patterns that will help us gain insights into how the structural constraints of the 4-blade β-propeller domain molecular scaffolding has given rise to such variable biochemical functionality. If enzymatic, the active site is often found in the cleft formed in the center of the propeller by loops connecting the successive four-sheet motifs, but the purposes vary considerably between proteins. The evolution of these highly divergent molecular tasks associated with the physically stable architecture of the PEX domain is defined herein as *Functional Hyperplasticity*.

Hemopexin

HPX, a glycoprotein found in the blood plasma of many vertebrate organisms, has been shown to have the highest known binding affinity for heme with a measured Kd $<10^{-12}$ M [8]. The protein is also known as β-1B-glycoprotein, and is a single polypeptide chain of ~439 residues, with a molecular mass of approximately 63 kD.

Many studies have shown that the primary function of Hemopexin is to sequester unbound heme released into the plasma from the breakdown of hemoglobin. The crystal structure of Hemopexin has been solved and the molecule consists of two structurally similar N-terminus and C-terminus β-propeller domains of about 200 residues in length, joined by a flexible, unstructured linker peptide [9]. This linker region is known to be part of the heme

binding site. The gene that codes for Hemopexin (HPX) has been shown to have numerous sequences identified as recognizable hemopexin amino acid repeats (HX) in the primary structure of the protein. The Fe (III) of the heme is coordinated by 2 histidine residues (in *Homo sapiens* NP_000604.1 H213 and H271) and further stabilized by a host of noncovalent interactions provided by a number of invariant aromatic and basic residues [9].

Hemopexin has been shown to play a part as a reactant induced after inflammation during the acute phase of the innate immune response in animals but not in humans [10-12]. It is synthesized primarily in the liver, although it can also be expressed in peripheral nerves, the central nervous system and the retina [13]. Its principal function is to capture heme freed from the breakdown of hemoglobin and subsequently released into the plasma. The need to capture and remove this free heme is vital due to biochemical properties which lead to cell membrane interference and production of free hydroxyl radicals, both properties that can have damaging effects on cellular health and functionality [14, 15]. Hemopexin has been shown to help prevent heme influenced oxidative stress, and to reduce the probability of iron loss associated with heme binding [11, 16, 17].

Additionally, the removal of free heme from circulation reduces the availability of iron that could potentially be used as a resource by invasive pathogens [18]. Several studies have shown that in concert with operating as a plasma heme transporter, hemopexin functions variously in several additional, health-related pathways. These include, but are certainly not limited to, blood iron homeostasis, generalized antioxidant protection, assisting in the regeneration of damaged nerve tissue, and gene expression modulation [19].

After binding the heme, the heme-hemopexin complex is then transported to the liver where is removed from circulation through a receptor-mediated endocytic process [20]. Once

sequestered in the liver cells, the heme is released and degraded by Heme Oxygenase. The freed iron is then stored, used in regulation, or recycled to the bone marrow for inclusion in newly synthesized erythrocyte precursors [20, 21].

In addition to the responses to pathological conditions, a considerable amount of research has indicated that hemopexin constitutively regulates normal physiological iron and heme homeostasis[22]. The heme-hemopexin complex also initiates a variety of important signaling cascades including Protein Kinase C (PKC), Stress-activated Protein Kinase (also known as JNK), Mitogen-activated Protein Kinase 3 (a.k.a. ERK) and Nuclear Factor *kappa* B (NFκB) several of which are known to regulate transcription factors such as Metal Regulatory Transcription Factor 1 (MTF1), and Nuclear Factor, Erythroid 2 like 2 (a.k.a. Nrf2). These transcription factors, in turn, activate genes whose protein products help to protect cells from further oxidative stress [23].

The crystal structure of HPX has been solved and the molecule consists of two structurally similar 4-bladed, N-terminus and C-terminus β-propeller domains joined by a flexible unstructured linker peptide [9]. This linker region is known to be part of the heme binding site and provides one of the two histidine ligands necessary to coordinate the heme iron. Previously delineated HPX repeat locations were accessed and mapped onto the crystal structure using PyMOL. Other sequences of particular interest were also mapped.

Structural mapping has revealed some interesting associations between the functional motifs and tertiary structure of these domains and these relationships may have implications for understanding the evolution of the hemopexin protein (for which the domain is named, but in this work, for clarity, referred to as the *PEX domain*). There is additional potential to elucidate the phylogenetic relationships among several other classes of animal proteins that

contain HPX repeats that comprise PEX domains, including the myriad Matrix Metalloproteinases (MMPs), Proteoglycan-4 (PRG4) and Vitronectin (VTN). All these proteins are associated with the extracellular matrix of multicellular animals, and research has implicated many of them in immune system dynamics [24]. This analysis is a primary outcome of the current study.

PEX domains outside of HPX

The 4-bladed β-propeller (PEX) domain has been discovered to also exist in a variety of proteins found in taxa other than Animalia. In Plantae, several proteins from leguminous members have been discovered to contain PEX domains including lentils (*Lens culinaris*), chick pea (*Cicer arietinum*), grass pea (*Lathyrus sativus*), and cow pea (*Vigna unguiculata*) [25-28]. A PEX fold has also been found in rice (*Oryza sativa*) [29]. These sequences for non-animal PEX domains are used in the phylogenetic analyses found in Chapter 2 of this work. There are also a number of PEX domains that have been identified in prokaryotic proteins [30]. However, the scope of this work will not encompass inclusion of the prokaryotic forms in the analyses with the exception of one chosen to hypothetically help root phylogenies. Pertinently, from an evolutionary perspective, this root likely represents horizontal gene transfer/capture (HGT/C) by a parasitic-pathogenic organism. This result is a primary conclusion of this work (see **Chapter 3**).

In the fungi, a protein with remarkably similar structure to previously described PEX domains from both Animalia and Plantae has been described from the oyster mushroom

(*Pleurotus ostreatus*) and named *ostreopexin* [31]. Other purported PEX domains are also included in the broader phylogenies presented in Chapter 3.

Methods

The basic structure of hemopexin

The Human hemopexin protein has a 462 amino acid primary sequence, with a molecular weight of ~ 60 kDa. The FASTA file of the primary amino acid sequence for the human hemopexin protein was obtained from the National Center for Biotechnology Information (NCBI) database [32]. This file was then imported into the Basic Local Alignment Search Tool (BLAST) and analyzed using the nucleotide BLAST function [33]. **Appendix A**, **Table A2** presents sequence comparisons of hemopexin genes and proteins between representative animal species.

The earliest success at determining the tertiary structure of the PEX domain involved crystallization of the C-terminal domain of hemopexin [34]. Subsequent to that achievement, crystal structures for the entire hemopexin molecule, bound to a heme moiety, in both glycosylated and de-glycosylated forms were derived. The PDB is a public access repository for crystallographic structure analyses of biological macromolecules that has a standardized file format (PDB files) [7]. Each structure is given a unique 4-character alphanumeric designation, the PDB ID. There are two complete structures available for rabbit hemopexin in the PDB:

1QHU - Mammalian blood serum hemopexin *deglycosylated* and in complex with its ligand heme [9], and

1QJS - Mammalian blood serum hemopexin *glycosylated*-native protein and in complex with its ligand heme [9].

For this work, 1QHU was chosen for more detailed structural investigations. The PDB ID file for 1QHU was downloaded and imported into PyMOL, a computer based, 3-dimensional structure visualization program [35]. See **Figures 2 and 3** below for graphic presentations of the entire HPX molecule from various spatial perspectives.

Results

The molecule consists of two structurally similar (but not identical) N-terminus and C-terminus β -propeller domains joined by a flexible unstructured linker peptide, known to be the site of heme binding. Visualizations of the tertiary structure of the hemopexin molecule was accomplished using PyMOL, an *in silico*, molecular visualization system created by Warren Lyford DeLano [35]. The structures are shown with the molecule bound to its prime ligand, the porphyrin heme, shown in green in all views.

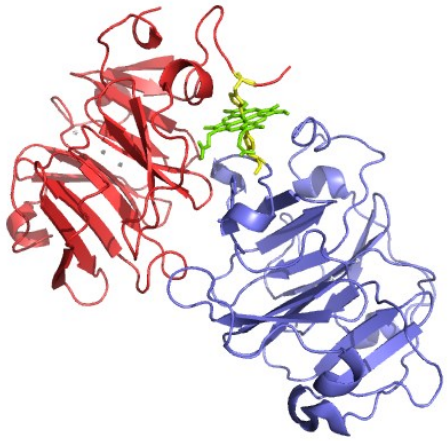


Figure 2. Basic crystal structure of rabbit hemopexin in complex with its ligand, heme. The N-terminus of the complex is shown in red, the C-terminus in blue and the heme moiety green.

Figure 3. Views of the N (left) and C (right) termini, showing the 4-bladed β- propeller architecture of each domain.

Analysis of position of HPX primary sequence repeats within domain structures

There are several important primary amino acid repeat sequences seen in the hemopexin molecule [36-38]. These include the DAAV/F motif and the WD-repeat. The WD repeats form a significant portion of the β-strands of each of the four blades of the propellers. The average number of residues in the WD repeats is approximately 43; the various WD repeats found in the rabbit hemopexin structure are comprised of from 39 to 45 residues and are alternatively referred to as Hemopexin (HPX) Repeats. **Table 1** below shows the 8 HPX repeats found in the HPX protein primary sequence, color-coded as in the graphical representations of the same repeats mapped onto the 3-dimensional structure of the N and C termini of the molecule respectively, shown in **Figures 4 and 5**.

Table 1. Color-coded arrangement of hemopexin repeats in the primary sequence of the hemopexin protein. Colors correspond to those showing positioning of repeats on the tertiary structural model (Figures 4 and 5).

MARVLGAPVALGLWSLCWSLAIATPLPPTSAHGNVAEGETKPDPDVTERCSD
GWSFDATTLDDNGTMLFFKGEFVWKSHKWDRELISERWKNF
PSPVDAAFRQGHNSVFLIKGDKVWVYPPEKKEKGYPKLLQDEFPGI
PSPL *DAAVECHRGECQAEGVLFFQGD* REWFWDLATGTMKERSWPA
VGNCSSALRWLGRYYCFQGNQFLRFDPVRGEVPPRYPRDVRDYFMPC

PGRGHGHRNGTGHGNSTHHGPEYMRCS (heme binding "linker" region)

PHLVLSALTSDNHGATYAFSGTHYWRLDTSRDGWHSWPIAHQWPQG
PSAVDAAFSWEEKLYLVQGTQVYVFLTKGGYTLVSGYPKRLEKEVGTPHGII
LDSVDAAFICPGSSRLHIMAGRRLWWLDLKSGAQATWTELPWP
HEKVDGALCMEKSLGPNSCSANGPGLYLIHGPNLYCYSDVEKLNAAKALPQ
PQNVTSLLGCTH

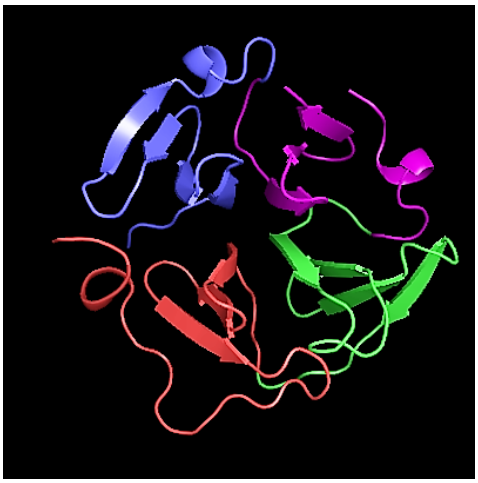


Figure 4. The N-terminal, PEX *domain*, showing the blade structure associated with the 4, tandem hemopexin *repeats* variably colored. Blade 1 = blue; blade 2 = violet; blade 3 = green; blade 4 = red.

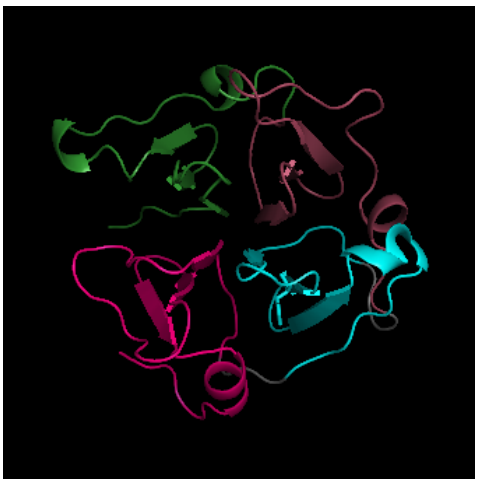


Figure 5. The C-terminal, PEX *domain*, showing the blade structure associated with the 4, tandem hemopexin *repeats* variably colored. Blade 1 = forest green; blade 2 = raspberry; blade 3 = cyan; blade 4 = hot pink.

Each of these repeats is structurally consistent with the "blades" of the 4-bladed β-propeller, or PEX domain. The architecture of each individual blade consists of three or four antiparallel β-strands. The 4 blades, in turn, are connected by short variable length loops and space filling α-helices. The connecting regions between strands differ considerably and account for approximately half the overall structure of the molecule. Although less constrained by structure these loops provide the biochemical features that form the heme-binding mechanism and are responsible for inter-domain functionality.

Recent annotation added to the graphics section for the hemopexin protein (HPX) - *Homo sapiens*, NCBI Reference Sequence: NM_000613.2, shows the positioning of the now established eight hemopexin repeats representing the two 4-bladed β-propeller domains of the molecule. Until this update it was assumed that the protein only contained 4 repeats, 2 of

each associated with the double domain structure. In earlier analyses considerable confusion was generated by the mistaken omission of the additional 4 repeats subsequently added. We have validated the inclusion of these repeats using a domain-based analysis, examining each, utilizing PHI-Blast and Blast2 alignments. As a result, the current domain architecture listed within the NCBI file for Hemopexin should be considered canonical. However, we note that not all putative hemopexin repeats have been thus annotated across NCBI, and our phylogenomic outcome suggests that many more such annotations should be appended. One conclusion of this study is the identification of the necessity for these annotations and to enter them into the appropriate NCBI database upon peer-reviewed publication.

The phenotypic evolution of the hemopexin molecule appears to have involved a genetically predicated structural duplication leading to the similarity in secondary and tertiary architecture of the N and C termini [39]. Paradoxically, the primary structures of these two β-propeller motifs show only a 27% residue sequence identity, and 40% positive biochemical identity**.** However, in spite of this lack of sequence similarity the structures superimpose remarkably well **(Figure 6)**. It was this realization that lead to the fundamental driving evolutionary question of this investigation, namely, how could a structure with such weak primary sequence homology remain so structurally consistent? This, in turn, developed into questions of the evolutionary significance of the concept of functional hyperplasticity defined as the highly divergent molecular functions associated with the evolutionarily constrained physicochemical architecture of the PEX domain.

Figure 6. The N-terminal and C-terminal domains of Hemopexin superimposed, showing a very high degree of symmetry with one another despite having relative weak sequence similarity. The RMSD for this alignment = 1.009 Å.

Not surprisingly, the gene coding for hemopexin has been identified in the Neanderthal genome, although the current state of annotation the Neanderthal Genome site still uses outdated protein domain architecture to identify structural components of the peptide sequence [40, 41]. An annotation and update of the Neanderthal protein will be forthcoming using the current understanding of positioning of the PEX domains within *H. sapiens* hemopexin, and their concomitant pexin-repeat sequences. A PSI-BLAST search was also performed on the recently released Denisovan Hominin (*Homo sapiens* ssp. *Denisova*) genome using the *Homo sapiens* ssp. *sapiens* sequence for hemopexin (NP_000604.3) as the query, but no significant hits were achieved. Given that is most unlikely that this subpopulation of humans did *not* have a gene to code for hemopexin, investigations will continue as the Denisovan Genome is uncovered and reported [42, 43]. For further discussion related to this section, see **Chapter 4**.

CHAPTER 2

DISCOVERY AND CHARACTERIZATION OF THE JEN-14 EPITOPE AS A MOLECULAR SYNAPOMORPHY IN HEMOPEXIN

Introduction

Hemopexin (HPX), a glycoprotein in the blood plasma of many vertebrate organisms, has been shown to have the highest known binding affinity for heme. HPX belongs to the acute phase class of immune reactants and is induced by inflammation. Many studies have shown that the primary function of HPX is to sequester unbound heme released into the plasma from the breakdown of hemoglobin. The need to capture and remove this free heme is vital due to chemical properties that can lead to severe cell and tissue damage, primarily through the production of free hydroxyl radicals.

An epitope, which binds to a monoclonal antibody designated "JEN-14", is a 22 amino acid primary sequence located within the HPX repeat 2 containing blade of the N-terminus of the HPX protein. Hemopexin captures and transports free heme to receptors located on the surface of liver parenchymal cells, that are known to be the low-density lipoprotein receptor related protein 1 (LRP1, a.k.a. "Cluster of Differentiation-91": CD91), a multifunctional receptor which recognizes several ligands including the heme-hemopexin complex [21]. The CD91 receptor is also expressed on the surfaces of several other cell types including macrophages, neurons, and syncytiotrophoblasts [44].

The JEN-14 epitope has been implicated in the ability of the heme-hemopexin complex to recognize and bind to the CD91 receptor since the JEN-14 antibody blocks binding of these complexes to the liver cells [45]. Additionally, it is known that the binding of

the HPX protein to this receptor initiates an endocytotic cascade that results in the heme-HPX being translocated from the blood stream to the interior of the hepatocytes. This eventually leads to the capture and recycling of the potentially cytotoxic heme moiety [45, 46]. The JEN-14 epitope spans residues 123-143 (22 amino acids) of the N-domain of rabbit hemopexin. A dual sequence alignment using BLAST between the human and rabbit primary sequences of the JEN-14 epitope for the two hemopexin proteins is shown below in **Table 2.**

Table 2. A two sequence alignment, with statistics, between the primary sequences of human hemopexin JEN-14 and rabbit hemopexin JEN-14 respectively. The residues highlighted with red are, as explained in the text, positions of missense mutations in the human sequence identified in dbSNP.

Score			**E-value**				**Identities**					**Positives**					**Gaps**							
66.0 bits (148)			**1e-14**				**19/22(86%)**					**21/22(95%)**					**0/22(0%)**							
Jen-14 (Human)	**D**	**A**	**A**	**V**	**E**	**C**	**H**	**R**	**G**	**E**	**C**	**Q**	**A**	**E**	**G**	**V**	**L**	**F**	**F**	**Q**	**G**	**D**	**22**
	|	|	|	|	|	|	|	|	|	|	|	|	**X**	|	|	+	|	|	|	|	|	+	
Jen-14 (Rabbit)	**D**	**A**	**A**	**V**	**E**	**C**	**H**	**R**	**G**	**E**	**C**	**Q**	**D**	**E**	**G**	**I**	**L**	**F**	**F**	**Q**	**G**	**N**	**167**

Methods

The Research Collaboratory for Structural Bioinformatics - Protein Data Bank (PDB) is a public access repository for crystallographic structure analyses of biological macromolecules that has a standardized file format (PDB files) [7]. Each structure is given a unique 4-character alphanumeric designation, the PDB ID. For the purposes of this analysis PDB-ID was used:

1QHU - Mammalian blood serum hemopexin *deglycosylated* and in complex with its ligand heme [9].

The PDB ID file for 1QHU was downloaded and imported into PyMOL, a computer based, 3-dimentional structure visualization program [35]. The primary sequence of the rabbit JEN-14 epitope was superimposed on the entire protein structure at the appropriate location.

See **Figures 7 - 9** for various graphical representations of the position of the JEN-14 epitope superimposed on the crystal structure of rabbit hemopexin (1QHU).

To check for sequence identity, Position-Specific Iterated (PSI) BLASTs were run on the human JEN-14sequence, with an e-value cutoff of e<0.005. Accession numbers for all protein used throughout: **APPENDIX J**.

Results

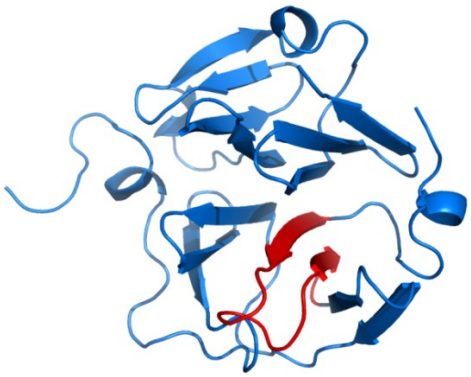


Figure 7. Cartoon representation of the position of the JEN-14 epitope (in red) in the isolated N-PEX domain of the crystal structure from rabbit (PDB-1QHU). The epitope is located in "blade-2" of the domain.

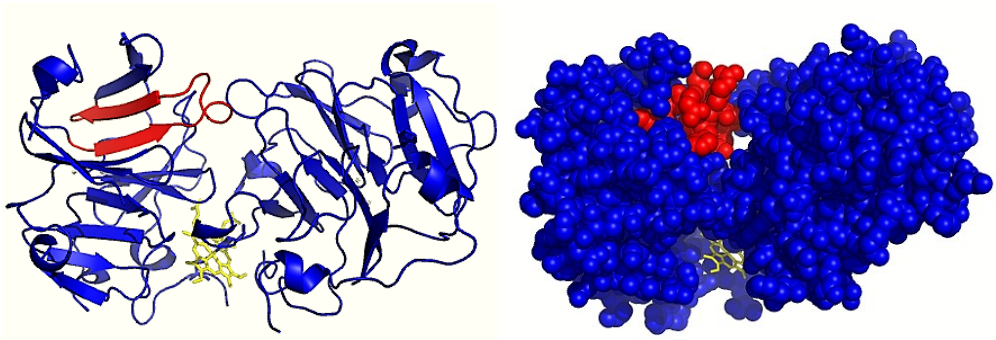


Figure 8. (A) Cartoon representation of the location of the JEN-14 epitope (red) in the N-PEX domain (oriented to the left) superimposed on the crystal structure of rabbit hemopexin (PDB-1QHU); note: bound heme in yellow. (B) Same representation in residue-sphere configuration.

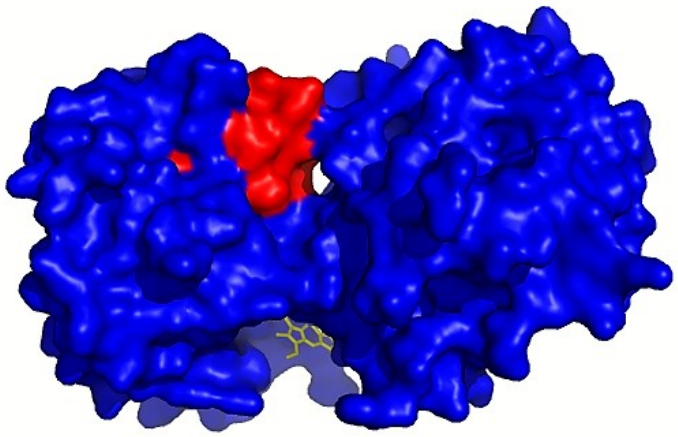


Figure 9. Surface representation of the location of the JEN-14 epitope in the N-domain (oriented to the left) superimposed on the crystal structure of rabbit hemopexin (PDB-1QHU).

Discussion and Conclusions

The human epitope is 100% conserved in sequence identity in common chimp, gorilla, orangutan, gibbon and sooty mangabey (*Cercocebus* sp.). It shares a minimum of 95% biochemical positives with virtually all *Hominoidea* (apes), both *Catarrhini* (old world monkeys) and *Platyrrhini* (new world monkeys). Additionally, it shares at least 86% biochemical identity with all the other primates, down to, and including the lemurs. It shares 91% positives identity with both *Rattus* (Norwegian rat) and *Mus* (mouse), and, furthermore, greater than 85% biochemical identity is conserved for the length of the epitope in mammals as far removed from humans as cats, Guinea pigs, elephants, naked mole rats, giant pandas and flying fox bats. It also shares a respectable 73% biochemical identity with the marsupial short-tailed opossum (*Monodelphis* sp.). Moving "down" the vertebrate phylogeny there is still a greater than 70% identity in birds (chicken), and some reptiles (Alligator) **Table 3**.

Table 3. Comparison of sequence similarity and biochemical identity of the JEN-14 epitope of the N-terminal PEX domain from human hemopexin with other selected species represented in the NCBI database.

SPECIES *Homo sapiens*	(%) similarity JEN-14 primary amino acid sequence	(%) similarity biochemical identity
Common Chimp	100%	100%
Gorilla	100%	100%
Orangutan	100%	100%
Gibbon	100%	100%
Mangabey	100%	100%
Pig	95%	95%
Cat	91%	95%
Rat	91%	95%
Elephant	91%	95%
Mouse	91%	95%
Guinea pig	90%	95%
Naked mole rat	90%	95%
Flying fox (bat)	86%	95%
Rabbit	86%	95%
Giant panda	86%	90%
Possum	73%	81%
Chicken	73%	77%
Alligator	71%	76%

BLAST searches on the sequence associated with the human hemopexin JEN-14 epitope hit exclusively on orthologs of hemopexin, regardless of taxon, and importantly, to the exclusion of all other proteins that contain 4-blade, β-propeller (PEX) domains. A three iteration PSI-BLAST on the human sequence returned 173 hits (at $e < 1e\text{-}04$) all of which were specifically identified as "hemopexin". Additionally the first non-hemopexin hit was a Warm Water Acclimation Protein (WAP1 - *Dicentrarchis labrax*: the European Bass). For

the significance of the evolutionary relationship between hemopexins and WAPs see **Chapters 3 and 4**.

This result is even more remarkable because the sequence is only 22 residues in length. In this short a sequence it might be assumed that random convergence would lead to many more possible matches throughout an exhaustive search of available genomes. However, considering this is not the case, it could be argued that not only is this motif extremely specific, but might actually be maladaptive in situations where the biochemical functionality of the motif could lead to unwanted attraction of the possessor to inappropriate binding of an extremely pervasive and promiscuous molecular actor in the CD91 receptor.

The remarkable level of evolutionary constraint in this epitope is further demonstrated and supported by analyses using the NCBI Single Nucleotide Polymorphism database (*db*SNP) database to check for known human polymorphisms associated with the sequence [47]. Specifically, the 1000 Genomes initiative database was mined for SNP identities [48]. Within the 22 amino acid sequence of the JEN-14 epitope there are SNPs known for 10 positions. However, only 1 has a quantifiable frequency, and that is exceedingly low at 0.2%. Additionally, the residue change from the expected Valine to the substitute Isoleucine is biochemically analogous, both amino acids being hydrophobic/neutral. Furthermore, it is interesting that the substitution at position 16 of Isoleucine for Valine is what is seen when aligning the human and rabbit sequences; **Table 2**, residues highlighted in red.

These data, taken in concert, support the conclusion that the JEN – 14 epitope is an intra-domain (N-terminus) molecular synapomorphy for the family of hemopexin protein within vertebrates (*i.e.* a shared derived character that is assumed to have been inherited from a common ancestor). Given that the phylogenetic analyses presented in chapter 3 have

determined that the hemopexins contain the evolutionarily oldest PEX domains, the JEN-14 epitope most likely evolved as an original component of these molecules and has remained unique to hemopexin. Fascinatingly, there is one 'true' hemopexin (limunectin) found outside of the vertebrates in *Limulus*, the horse shoe crab, an ancient invertebrate. And while *Limulus* does not have iron-protoporphyrin IX associated hemoglobin is does use hemocyanin, a copper oxygen binding protein, and human hemopexin has been shown to bind copper [49]. This establishes a plausible link between limunectin and the hemopexins found in vertebrates, but it is currently hard to explain or test given the lack of a crystal structure for limunectin.

CHAPTER 3

PHYLOGENOMIC ANALYSES PROVIDE INSIGHTS INTO PATTERNS OF FUNCTIONAL DIVERSITY BETWEEN PEX-DOMAIN CONTAINING PROTEINS

Introduction

The PEX domain (originally the "hemopexin" domain) is a tertiary protein structural motif that is characterized by a sequential arrangement of 4 β-propeller blades surrounding a central axis that is the scaffolding site for various metal ions **(Figure 1)**. Aside from the hemopexin (proper) molecule there are several other classes of proteins found in Animalia that contain PEX domains including the Matrix Metalloproteinases (MMPs), Proteoglycan-4 (PRG4) and Vitronectin (VTN). There is also a compendium of proteins that have tertiary structure identified as the 4-bladed β-propeller (PEX) domain that have been identified in Plantae, Fungi and bacterial Prokaryotes. A primary focus of this research is to garner insights into the evolutionary relationships of these various PEX domain containing proteins to highlight the neofunctional diversification that has arisen evolutionarily despite the remarkable architectural constraint exhibited at the tertiary level of structure. What follows is a taxonomic survey of PEX-domain containing proteins.

I. *PEX domains in Animalia:*

1. *HPX* (hemopexin 'proper')

For a detailed description and discussion of the PEX domains in the hemopexin protein (HPX), see **Chapter 1**.

2. *PRG4*

For a detailed description and discussion of the Proteoglycan-4 protein (PRG4), see **Chapter 5**.

3. *VTN*

Vitronectin (VTN) is a multifunctional vertebrate plasma glycoprotein that has been variously known as: Human S-protein, orepibolin, and Serum Spreading Factor. It is secreted into the extra-cellular matrix (ECM) where it recognizes and interacts with an assortment of ligands of structurally diverse architecture. The primary structure of the VTN molecule contains 459 residues and has a weight of approximately 75 kDa, of which about one third is attributed to the carbohydrate moieties [50].

VTN is a highly modular protein consisting of 3 major domains: a Somatomedin B domain (residues 1-39) toward the N-terminus; a domain with sequence similarity to HPX (residues 131-342); and a 102 residue region (347-559), also with identity similar to HPX

toward its C-terminus. This central region of the protein (specifically residues 154-172) is predicted to fold as a 4-bladed β-propeller (PEX) domain [51]. VTN is also known to exist in an alternate two-part form in which cleavage of the apo-protein occurs after the R379 residue and held together by the subsequent formation of a disulfide bond [50]; and is also apparently involved in immune system modulation, cell adhesion and migration, blood coagulation and fibrinolysis, tissue remodeling, and tumor metastasis [52].

4. *MMPs*

The Matrix metalloproteinases (MMPs), also known as 'matrixins', are a family of endopeptidase enzymes primarily responsible for the build-up, degradation, and re-modelling of various components of the animal ECM.

The MMPs are secreted primarily by cells of connective tissues such as endothelial cells, osteoblasts, and fibroblasts; as well as lymphocytes, neutrophils and macrophages, pro-inflammatory cells of the immune system. The activity and/or structural integrity of these proteins are dependent on both calcium and zinc, and they are involved in interactions with ECM proteoglycans, other matrix glycoproteins, gelatin, elastins and collagens. They are also known to be important in ovarian function, and are regulated variously by cytokines, growth factors and hormones [53]. They have a close functional relationship with a family of inhibitory proteins known as the TIMPs (Tissue Inhibitors of Metallo-Proteinases) [54].

The inclusion of the MMPs in these analyses is a consequence of the majority of them having a 4-bladed β-propeller (PEX) domain associated with some aspect of their tertiary structure.

The MMPs are named by numbers according to the sequence in which they were discovered or described. MMPs 4, 5, and 6 were eventually shown to be isoforms of previously describe members and hence those numbers are not used. MMPs 7, 23, and 26 do not contain PEX domains, and so are excluded from these analyses. For a list of the MMPs included in the phylogenies along with a brief description of known functionality, see **Table A3**; **Appendix B**. Sequence similarity at the PEX-domain level for the human proteins is presented in **Table A4**, **Appendix C**.

5. *WAPs*

In the Chordate lineage an orthologous hemopexin gene shows up temporally earliest in the cartilaginous fishes: the Little Skate (*Leucoraja erinacea*), the Nurse Shark (*Ginglymostoma cirratum*) and the White-spotted Bamboo Shark (*Chiloscyllium plagiosum*) [55, 56]. An osteichthyian ortholog has also been identified in the common goldfish (*Carassius auratus*), and was subsequently recognized in several additional bony fish species including, but not limited to, the Channel Catfish (*Ictalurus punctatus*), the common carp (*Cyprinus carpio*), the Green Swordtail (*Xiphophorus helleri*), the Japanese Black Porgy (*Acanthopagrus schlegeli*), the Japanese Puffer (*Takifugu rubripes*), and the Japanese Rice fish (*Oryzias latipes*) [57-62].

The protein coded for by this gene was named "the warm-temperature-acclimation-associated 65-kDa protein (WAP65)", whereas, when the protein was discovered in the chondrichthyes it was called "hemopexin" after its mammalian ortholog. Additional investigations also eventually identified a paralog of the first WAP65 in the boney fishes

resulting in the designation of the separate proteins as WAP65-1 and WAP65-2 respectively. Both paralogs are structurally and functionally similar to one another as well as to their Chondrichthyan and mammalian hemopexin orthologs [55]. However, they do exhibit notable differentiation in spatially specific expression patterns with WAP65-1 found in multiple tissue types, and WAP65-2 being exclusive to the liver [63].

Iron acquisition is known to be a pivotal component of successful bacterial invasion and subsequent infection [64]. Because of the well documented role of hemopexin in capturing free heme and its role in sequestering and recycling iron in mammalian systems, coupled with the structural similarities of the WAP65 proteins to hemopexin, several studies have sought to link similar physiological iron-dynamics functions to the WAP65 proteins of the teleosts [63]. Indeed, both of the paralogous WAP65 proteins have been demonstrated to have multiple functionalities involving iron homeostasis, specific immune response, exposure to heavy metal toxicity, and development and temperature acclimation [59, 60, 63, 65-69].

Recent work has suggested that the multifunctional nature of the divergent WAP65 proteins is representative of an entire gene duplication followed by neofunctionalization that has left the WAP65-2 protein to perform the original adaptive functions of temperature acclimation and immune response (that are similar to what is seen in the extant mammalian hemopexin), while the WAP65-1 has evolved variable new functionalities [55, 63, 70, 71]. Detailed structural comparisons, a functional divergence analyses and identification of points of positive molecular selection during divergence of the two WAP65 paralogs, along with their intra-taxonomic functional distinction from one another, as well as from the mammalian ortholog has been achieved [55]. This study also showed probable strong positive divergent selection, significantly influencing the evolution of these three proteins since the time of the

duplication of the WAP65 paralogs in the teleosts. This study was also able to document marked increased acceleration of positive selection on the mammalian hemopexin, as compared to either of the WAP65's. Additionally, it makes a compelling case that the discernable selection differences in the three proteins were independent of the ability of any of them to bind heme.

6. *PEX domain in invertebrates*

In the horseshoe crab (*Limulus polyphemus*), a 54-kDa protein isolated from amoebocytes and named "limunectin" has been discovered and described [72]. Assays have established that it binds to a phosphocholine and pneumococcal C-polysaccharide in a calcium independent manner, and also participates in several cascades involving activities in the extra-cellular matrix [72]. The binding kinetics are relevant to distinguish limunectin from another immune system protein known as “C-reactive protein” (named for its affinity for the “C”-carbohydrate antigen of *Pneumococcus*), which has been shown to only bind in the presence of free calcium [72, 73]. At the time it was discovered, the researchers establish "limited sequence homologies" to segments of collagenase (now MMP1), gelatinase (now MMP2), and vitronectin (VTN). It is now obvious that the "limited" aspect of the similarity is, in fact, the PEX domains of those respective proteins.

Limunectin is found mainly in the secretory granules of *Limulus* amoebocytes and was predicted to function as an adhesion molecule capable of recognition and neutralization of invasive pathogenic microbes. In its role as an adhesin recognition protein it is thought to play a crucial part in the immune response in *Limulus* [72, 73].

Immunoglobulins are absent and, as such, obviously, the only immunity *Limulus* exhibit is innate. It has been hypothesized, however, that coordination between complement-like proteins, agglutinins and lectins, potentially work in functional synergisms to afford horseshoe crabs with, while not a classically defined adaptive immunity, potentially a "souped-up" version of innate immunity [72, 73]. The evolutionary longevity of genus *Limulus* certainly speaks to the efficacy of such a putative adaptation.

A structure for limunectin has not been reported, so presented here are two homology models from the primary sequence that show remarkable similarity in tertiary structure to mammalian hemopexin, **Figures 10 and 11**.

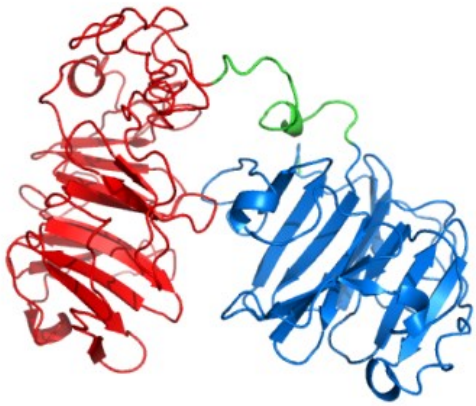
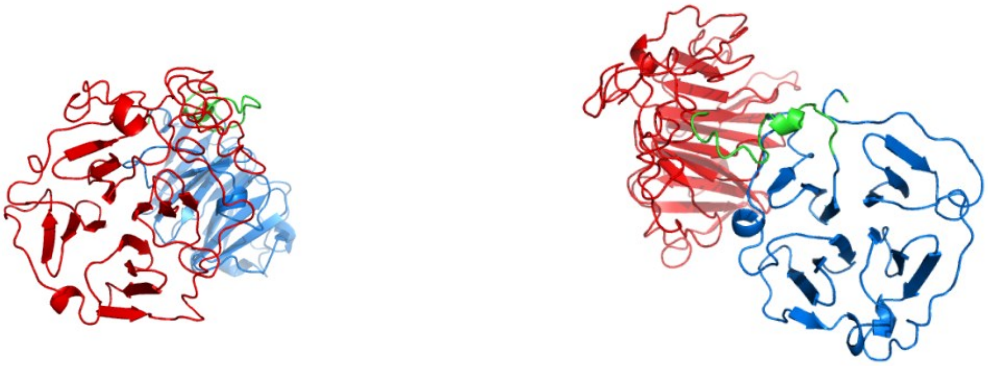


Figure10. Homology model of *Limunectin* from TASSER-I protocol. Lateral view. Note obvious similarity in overall tertiary structure to chordate hemopexin[74-76].



(a) **(b)**

Figure 11. Homology model of *Limunectin* from TASSER-I protocol threaded onto (1QHU). a) face-on view of the N-domain; b) face-on view of the C-domain. Again, note obvious similarity in overall tertiary structure of the β-propeller domains to chordate hemopexin.

The class Xiphosura is thought to have arisen in the Paleozoic Era (570–248 MYBP), and representative fossils have been dated to a minimum of 445 MYBP, therefore this lineage has an extremely deep evolutionary foundation [77, 78]. The presence of limunectin in this arthropod poses several fascinating challenges from an evolutionary standpoint. For instance, due to the ancient nature of any divergence patterns from shared common ancestors with *Limulus*, if limunectin is a homolog of hemopexin, then the ancestor of the hemopexin protein would itself have to have shared this deep evolutionary origin, and pre-date the appearance of phylum Chordata. This seems a very unlikely scenario. However, the alternative is, in many ways, just as improbable, namely that limunectin and hemopexin are examples of evolutionary convergence at the level of the entire protein.

II. *PEX domain in Plantae*

A PEX domain has been identified in a seed albumin protein (PA2) of the garden pea plant (*Pisum sativum*). It has been proposed that the protein structure is comprised, most likely, of a 2 β-barrel globular construction that resembles the classic 4-bladed propeller described for the PEX domain [79]. The presence of this tertiary structure in plants would suggest that the evolution of the PEX domain architecture predates the divergence of plants and animals at the kingdom level. Since the discovery of the PA2 PEX domain, several more have now been seen, not only in multiple species, but in several different plant proteins. It has been suggested that one prime metabolic function of these HPX domain-containing proteins in legumes involves polyamine metabolism and, potentially, protection from heme-induced oxidative stress; and in rice they may be involved in the degradation of chlorophyll [26, 28, 29].

III. *PEX domain in Fungi*

The kingdom level divergence is further evidenced by the discovery, in fungi, specifically the oyster mushroom (*Pleurotus ostreatus*), of a 4-bladed β-propeller domain that is structurally and functionally similar to the PA2 albumins isolated from various leguminous seeds and rice (*Oryza sativa*)[29]. This protein has been named "ostreopexin" [31]. (However, see discussion, **Chapter 5,** for more detailed speculations on the evolutionary origin of this domain in *Pleurotus*).

IV. *PEX domain in Prokaryotes*

1. Domain *Bacteria*

A 4-bladed β-propeller (PEX domain) has been discovered in a bacterium, *Photorhabdus luminescens*. This bacterium is a member of the symbiotic gut flora of nematodes that infect insects, and secretes virulence factors known as "high molecular mass toxin complexes". These toxin complexes are thought to be released by the nematodes into the hemolymph of an infected insect host, and to act to subvert the insect's immune response and to set up a suitable environment for reproduction of both the bacterium and the nematode. This protein was named *photopexin*. During analyses of these virulence factors two open reading frames (ORFs) were discovered which, when expressed, translate into two isoforms of photopexin, ppxA (564 amino acids) and ppxB (340 residues), both with serial repeats and sequence similarity to the PEX domains of interest. The potential roles in pathogenicity of the two photopexins are still not known. However, many pathogenic gram-negative bacteria are known to utilize iron as a metabolic resource, and to possess outer membrane heme-binding and heme-transporting proteins [30]. Therefore, it is possible that, in this case, the bacteria are co-opting a eukaryotic-like heme "scavenging" molecule to hi-jack the host insect's iron for their own metabolic and biochemical needs. Also, importantly, it is known that *Caenorhabditis elegans*, and some parasitic helminths are incapable of synthesizing heme on their own**,** even though they have physiological needs associated with several biochemical pathways involving hemoproteins. It is therefore necessary for these worms to *exogenously acquire* the heme needed for their biological processes [80]. It is

furthermore reasonable to assume that any capability these helminths might co-opt from other organisms that aids in iron acquisition would be of tremendous adaptive significance. More on the potential evolutionary implication of this speculation is presented in **Chapter 5**.

2. Domain *Archaea*

After an exhaustive search of various formulations of PEX domain-specific descriptive primary sequences, these analyses have determined that, to date, there is no reported evidence of a 4-bladed β-propeller structural domain existing in any proteins described within Domain *Archaea*.

Methods

Phylogenies

To assess the evolutionary relationships between the various PEX domain containing proteins phylogenetic analyses were performed at several levels. First, a phylogeny of the mature protein groups from humans (HPX, MMPs, PRG4 and VTN), together with the WAPs and HPXs from fishes, and Limunectin from the horseshoe crab was deduced. The Maximum Likelihood consensus is presented in **Figure 12**. Secondly, the phylogeny for just the isolated PEX domains from the human groups, plus the fish WAPs, and Chondrichthyan HPXs was inferred and shown in **Figure 13.** Finally, a phylogeny of representative PEX domains *across Kingdoms* from Animalia, Plantae, Fungi, and Bacteria is displayed in

Figure 14. Statistical details for all phylogenetic analyses are indicated in the text for each figure. All phylogenetic analyses were conducted using the Molecular Evolutionary Genetics Analysis, Version 7.0 suite of software (MEGA7) [81].

Results

The evolutionary history of the 8 groups of mature PEX domain containing proteins from humans and fishes, rooted with Limunectin, was inferred by using the Maximum Likelihood method based on the Jones, Taylor and Thornton (JTT) matrix-based model [82]. The bootstrap consensus tree inferred from 500 replicates is taken to represent the evolutionary history of the taxa analyzed [83]. Branches corresponding to partitions reproduced in less than 50% of bootstrap replicates are collapsed. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (500 replicates) are shown next to the branches [83]. Initial tree(s) for the heuristic search were obtained automatically by applying Neighbor-Joining and BioNJ algorithms to a matrix of pairwise distances estimated using a JTT model, and selecting the topology with superior log likelihood value. The analysis involved 28 amino acid sequences. All positions containing gaps and missing data were eliminated. There were a total of 240 positions in the final dataset. Evolutionary analyses were conducted in MEGA7 [81] **(Figure 12)**.

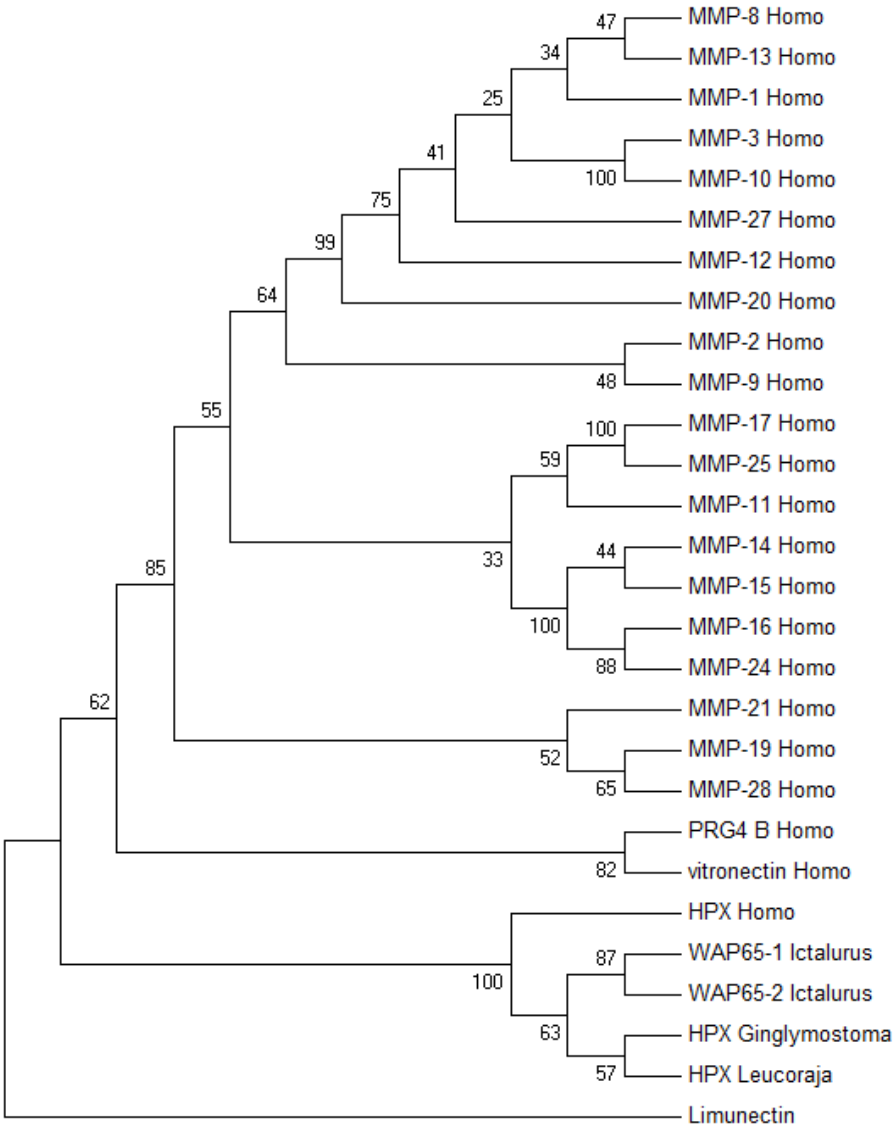


Figure 12. Molecular Phylogenetic analysis of ENTIRE PROTEINS with PEX DOMAINS (*Homo* + WAPs + Limunectin) by Maximum Likelihood method. Numbers represent bootstrap values from 500 iterations.

The evolutionary history of *the isolated PEX domains* from the same suite of proteins as in **Figure 12** was inferred by using the Maximum Likelihood method based on the JTT matrix-based model [82]. The bootstrap consensus tree inferred from 500 replicates is taken to represent the evolutionary history of the taxa analyzed [83]. The percentage of trees in which

the associated taxa clustered together is shown next to the branches. Initial tree(s) for the heuristic search were obtained automatically by applying Neighbor-Joining and BioNJ algorithms to a matrix of pairwise distances estimated using a JTT model, and selecting the topology with superior log likelihood value. The tree is drawn to scale, with branch lengths measured in the number of substitutions per site. The analysis involved 22 amino acid sequences. All positions containing gaps and missing data were eliminated. There were a total of 133 positions in the final dataset. Evolutionary analyses were conducted in MEGA7 [81] **(Figure 13)**.

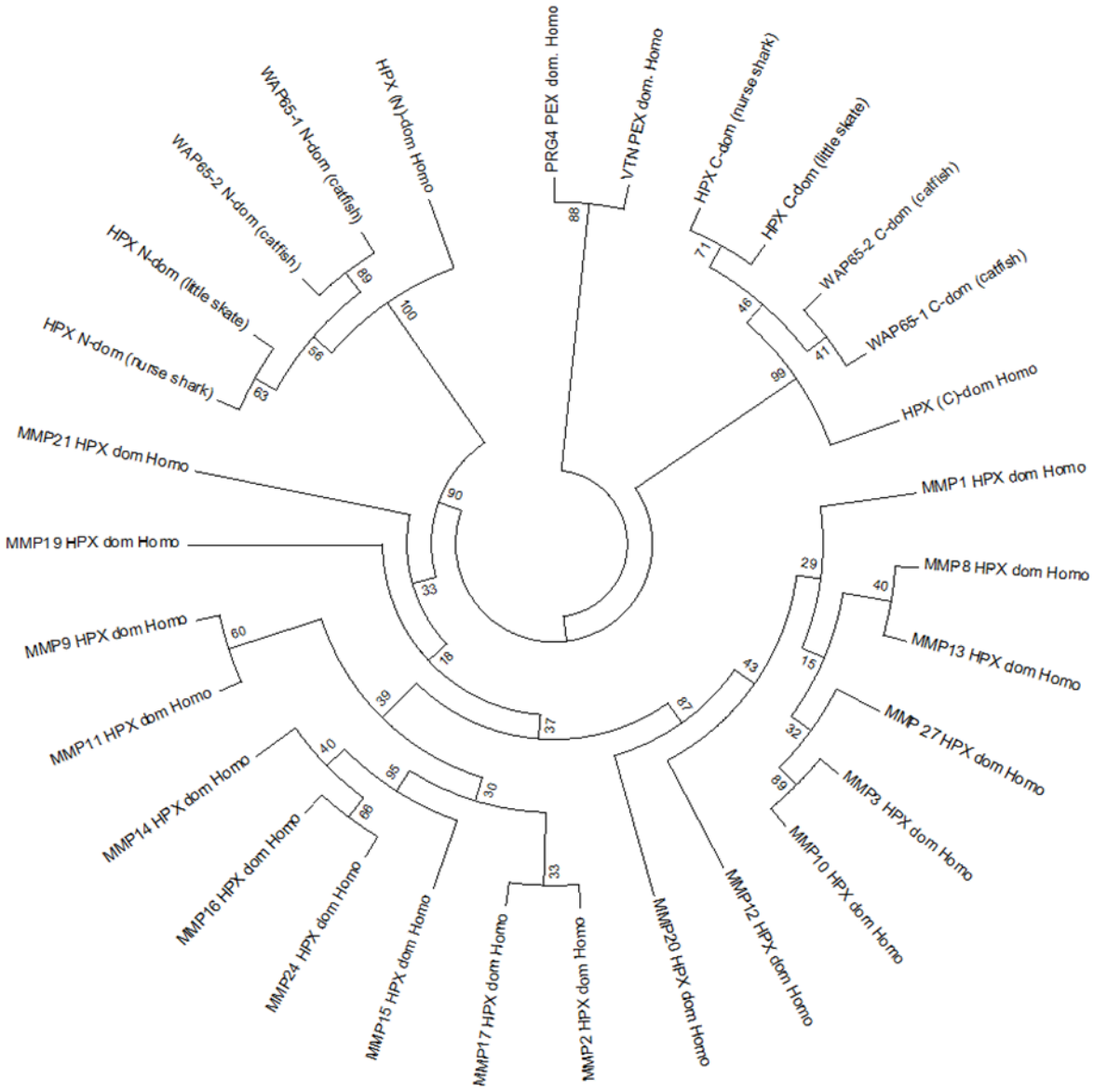


Figure 13. Molecular Phylogenetic analysis of PEX DOMAINS (*Homo* + WAPs) by Maximum Likelihood method.

The evolutionary history of *isolated PEX domains across kingdoms* was inferred by using the Maximum Likelihood method based on the JTT matrix-based model [82]. The bootstrap consensus tree inferred from 500 replicates is taken to represent the evolutionary history of the taxa analyzed [83]. Branches corresponding to partitions reproduced in less than 50% bootstrap replicates are collapsed. Initial tree(s) for the heuristic search were obtained automatically by applying Neighbor-Joining and BioNJ algorithms to a matrix of pairwise distances estimated using a JTT model, and selecting the topology with superior log likelihood value. The analysis involved 54 amino acid sequences. Evolutionary analyses were conducted in MEGA7 [81] **(Figure 14)**.

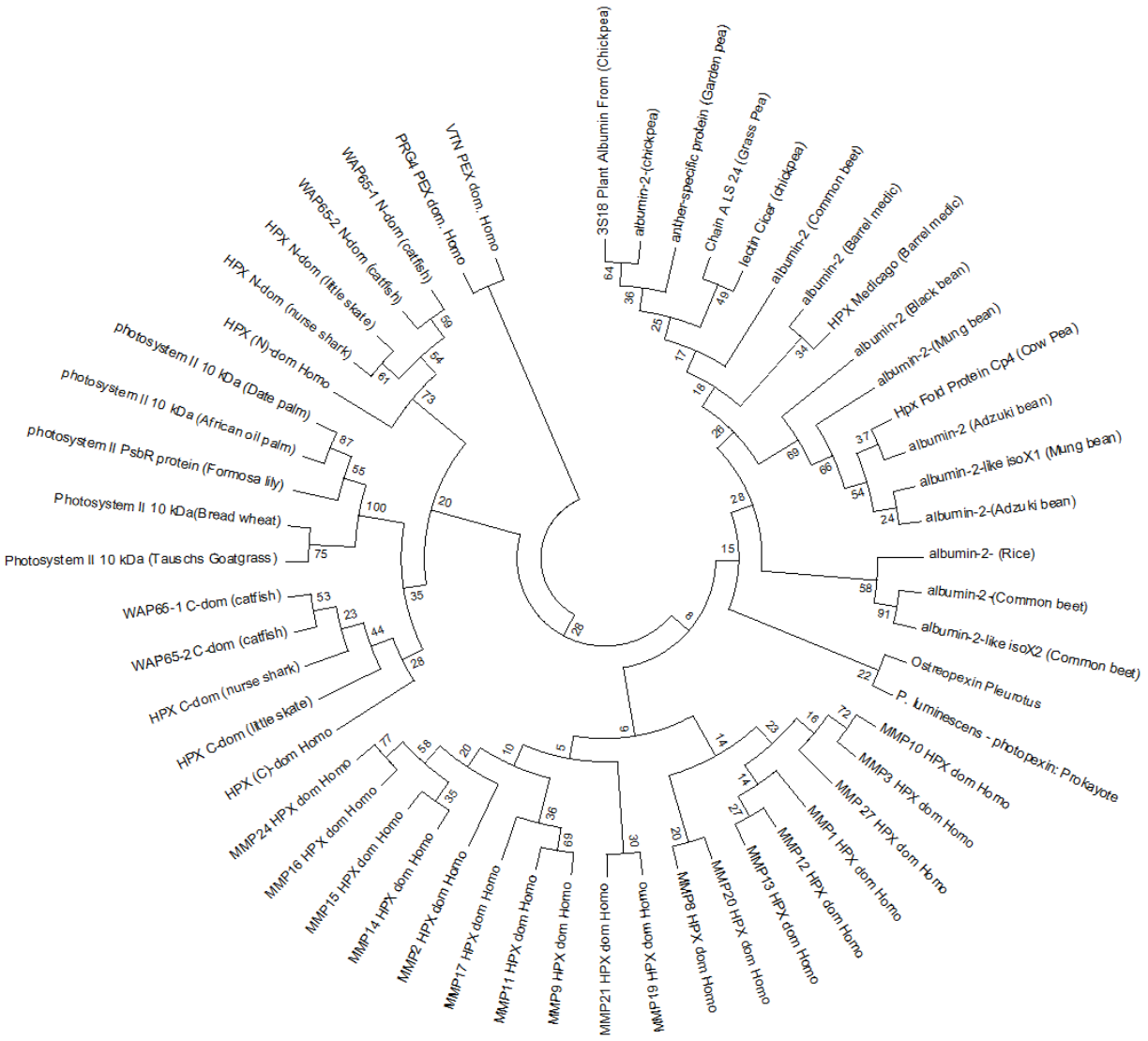


Figure 14. Molecular Phylogenetic analysis by Maximum Likelihood method of isolated PEX domains across kingdoms.

For detailed discussion of results see **Chapter 4**.

CHAPTER 4

DISCUSSION AND CONCLUSIONS - EVOLUTION OF THE 4-BLADED β-PROPELLER (PEX) DOMAIN

Discussion

Implications for human health

One basic premise of these analyses is that understanding the evolutionary foundations of the structural constraints on the PEX domain, and the associated biochemical and cellular functionalities, has direct relevance to issues related to human health. Camptodactyly-Arthropathy-Coxa vara-Pericarditis (CACP) syndrome is a rare autosomal recessive condition that leads to progressive severe deterioration of any bone-to-bone contact that is a result of progressive damage due to the lack of the lubrication normally provided by PRG4 (Lubricin) [84, 85]. As the name suggests, this is a condition that is manifested quite variably at the clinical level, but strong linkages indicate that, regardless of the presentation of pathologies, there is genetic homogeneity as the underlying causative factor. In a recent European cohort study of 13 patients who suffer from CACP, 5 novel mutations in their *PRG4* genes were characterized. Two of these are nonsense mutations in the PEX domain coding region of the *PRG4* gene. Both of these mutations lead to truncated translation of the N-terminal region of the protein and have been shown to disrupt tissue binding [86].

Malfunction in the normal physiological activity of the MMPs have been shown to be associated with a wide-ranging pantheon of human diseases including arthritic suites, osteoporosis, cirrhosis of the liver, and aortic aneurisms, as well as other chronic and acute

cardiovascular conditions [87]. Another consistent pattern has implicated individual or combinations of MMPs in a variety of cancers, and more specifically in the processes of angiogenesis, tissue reorganization and metastases. Consequently, the MMPs, as a class of proteins, have long been considered prime targets for pursuing molecular level treatments for malignant cancer pathologies [87-94]. Many of these studies have explicitly highlighted dysfunction of the normal cellular processes associated with the PEX domains of these proteins as possible mechanistic drivers of the various pathologies. Regrettably, it has been determined through multiple attempts at discovering or producing effective inhibitors for malfunctioning MMP-related pathologies that the extreme functional-plasticity of the hemopexin domains makes accurate modeling of domain specific inhibitors quite problematic [95].

Unfortunately, a series of papers suggesting different functions for hemopexin have been published [96-99]. These studies have been thoroughly refuted [49]. However, the incorrect functions have occasionally made their way into annotation of HPX or HPX-like proteins. This is especially true in cases where automated updating has been used, and persists even in Gene Ontology annotations This has seriously complicated the evolutionary and phylogenetic studies of HPX and related proteins, and has led to incorrect phylogenetic interpretations, especially where weighting for function has been employed.

Recent work has also shown higher expression in the eyes of diabetics: hemopexin is overexpressed in the retinal pigment epithelium (RPE) of diabetic patients with diabetic macular edema (DME) and induces the breakdown of RPE cells which can manifest as diabetes induced loss of vision, leading even to total blindness [100]. There has also been additional evidence that the role of hemopexin in sequestering and recycling free heme is a

prime factor in the control of oxidative stress and ancillary inflammatory maladies. This research indicates that it is worthwhile to consider the pursuit and development of hemopexin-specific therapy protocols that could derive a new level of protection from the adverse effects of oxidative stress-mediated inflammatory conditions caused by overproduction of free heme. Among the more significant human health disorders in this category is atherosclerosis [101].

Evolution of the PEX domain

A divergent evolutionary origin of the PEX domain containing proteins was first proposed in 1987. Early generation gene alignment analyses concluded that nucleotide sequences in the gene that coded for what was at the time known as the "Spreading protein" ('S-protein' now termed: Vitronectin) matched sequences previously described for hemopexin and transin (MMP3) [102, 103]. A "pexin gene/protein" family was proposed as a collective term in recognition of the likelihood of the discovery of additional proteins that would contain recognizable 4-bladed β-propeller (PEX domain) tertiary structure.

There is evidence that in early stages of vertebrate evolution, with the emergence of the fish, blood chemistry pathways arose in response to the necessity of dealing with the removal of deteriorating, senescent and dysfunctional erythrocytes, and the concomitant release into the plasma of hemoglobin protein and its prosthetic heme moieties. Phylogenetic analyses have established that the gene that codes for the protein haptoglobin arose subsequent to the divergence between the urochordata (tunicates) and Osteichthyes (bony fishes). Haptoglobin, a transport protein with very high affinity to bind free hemoglobin,

probably evolved from a protease antecedent (MASP, mannose-binding lectin-associated serine proteinase), a complement-associated protein involved in the innate immune system of the ancestral lineage. Interestingly, haptoglobin subsequently disappears from the genomes of some Lissamphibia (specifically anurans) and Aves. This evolutionary loss may be a consequence of haptoglobin being potentially maladaptive under certain conditions. Nevertheless, there then arose a structurally (and assumed evolutionarily) nonrelated analog to haptoglobin, used for hemoglobin capture and removal in the birds and is known as PIT54 (a.k.a. 18-B) [104, 105]. This hypothesis is further supported by evidence that multimeric haptoglobin evolved independently twice within the mammals, suggesting that these alternate forms might mitigate the uncharacterized "disadvantages" of the original protein [105].

Parallel to the evolution of haptoglobin, several other proteins that target the removal of free heme and other potentially cytotoxic radicals from the plasma appear in the early phylogenic record. Among these ancillary proteins are a1-microglobulin (a.k.a. - Protein HC) and hemopexin. The hypothesis that haptoglobin evolved from an ancestral protease associated with the innate immune system is very intriguing, and a hypothesis of a similar and/or parallel evolutionary origin for hemopexin should be explored [105]. Further hypotheses might suggest, given the high association of the majority of the PEX domain containing proteins, most specifically hemopexin, with the animal immune system, that this suite of blood plasma proteins may have played an integral role in the original emergence and evolution of adaptive immunity in the Vertebrata.

The Warm Water Acclimation Proteins (WAPs - see Chapter 3) in bony fishes have been known for some time, but the evolutionary emergence of the hemopexin protein in vertebrates has been traced back to the cartilaginous fishes [56]. An ortholog has not been

identified, to date, in the genomes of jawless vertebrates or any other "lower deuterostomes" and so it is assumed to have arisen during the second documented round of vertebrate genome-wide duplication and to have appeared, coincidentally, with tetrameric hemoglobin [106]. However, the known presence of *Limunectin* in a protostome (horseshoe crab - *Xiphosura*) makes this interpretation somewhat problematic, in that parsimony would suggest and origin for hemopexin that predates any protostome/deuterostome cladogenisis[72].

Earlier studies had deduced that the hemopexin protein was a complex of sequences that had evolved from ancestral molecular precursors, and as such, is the most derived of the proteins containing the PEX domain [107]. However, the phylogenetic analyses of primary amino acid sequences in this work has determined that hemopexin is, most likely, the evolutionarily ancestral molecule, at least when compared to the domains of the various mammalian MMPs **(see Figure 10)**. If it can be established that the protein *Limunectin*, a potentially major component of the innate immune system in horseshoe crab (*Limulus polyphemus*) is an ortholog of hemopexin, and is not convergent, then this interpretation (hemopexin-proper as ancestral) would be bolstered by the very deep divergence (450+MYBP) of the extant lineages in which it is found.

Interpretations of phylogenies

The phylogeny of the mature proteins in *Homo* (with WAPs, and *Limunectin* for comparisons) follow the patterns already uncovered in previous investigations, and render no new insights [63]. In general the clades are predictable, although meaningful evolutionary explanations remain to be determined, especially regarding the MMPs **(Figure 12)**. It is

notable, that rooting the clade with *Limunectin* (as indicated) supports the contention that the two-domain WAP/HPX proteins are ancestral rather than derived, even though the single domain proteins (PRG4, VTN and MMPs) are structurally "simpler" (due to having only one PEX domain). The clade is rooted with *Limunectin* purely on the phylogenetically historical fact of a deep protostome/deuterostome divergence.

Concerning the phylogeny calculated for the isolated PEX-domains of the individual proteins, the N and C domains of the HPX/WAP proteins are tight and unequivocal **(Figure 13)**. According to this analysis, the WAPs, Chondrichthyan hemopexins and Mammalian hemopexins are homologous, not just at the protein level but also at the level of the two separate individual PEX domains (both N and C terminal positions). Therefore, not surprisingly, the "hemopexins" are homologs. However, it is fascinating that the N- and C-halves of the dual domain proteins are NOT sister clades **(Figure 13)**. This is relevant given the remarkable similarity in 3D structure (RMSD - 1.009 Å) generated by peptides that share only 47% sequence identity **(Figure 6)**. This result suggests a need to rethink the traditional use of primary amino acid sequences *as characters* in phylogenies where the evolutionary questions are exclusive to structural relationships and those relationships may obscured because of how the search algorithms currently work. These observations certainly warrant further investigation.

According to this phylogeny, it is also apparent that Proteoglycan-4 (PRG4) and Vitronectin (VTN) are phylogenetically sister proteins, and interestingly clade together, and away from the PEX domains in other proteins. But, from there, all manner of interpretation possible for the rest of this phylogeny, produced from the primary sequences of the isolated PEX-domains from the same proteins presented in **Figure 12**, is confusing at best. There is

no consistency in the arrangements of the clades within the MMPs. And, more importantly, there is no supportable statistical congruence between the mature proteins phylogeny and that of their respective individual isolated PEX-domains.

The incongruence of these phylogenies suggests an evolutionary trajectory for the PEX domain *independent* of that of the "parent" proteins. This is interpreted as phylogenetic evidence of the supposition of evolutionary *exaptation* at the level of the 4-bladed β-propeller (PEX) domain, discussed later [108, 109].

There are several notable observations concerning the patterns seen in the best calculated inter-kingdom phylogeny **(Figure 14)**. First, and foremost, is that the clades intersperse within the phylogeny and do NOT fall into a predictable separatory branching pattern based on *a priori* expectation of straight inter-kingdom divergence.

On the other hand, taking isolated groupings into consideration, one definitive outcome is that the Photosystem II proteins of the grasses (Viridiplantae: monocots) form a very tightly supported clade, and have a much higher number of substitutions than any other statistically significant grouping **(Figure 14)**. This is biologically consistent, in that it is known that the monocots are the most derived of the extant Viridiplantae [110]. One avenue of future research will be to delve into this pattern deeper to see if it might be possible to better parse the pivotal role of these proteins, and their modifications of photosynthetic efficacy, during the adaptive evolution of the monocots. One aspect of the physics at the molecular level that raises some excitingly fascinating evolutionary possibilities is that heme and chlorophyll are both derived from a common precursor protoporphyrin IX differing in that heme contains iron and chlorophyll contains magnesium.

Potential Horizontal Gene Transfer of the PEX domain

Aside from the apparent lack of predictable divergence patterns, the most striking observation that emerges from the inter-kingdom phylogenetic analysis is the tight relationship indicated between Ostreopexin (Kingdom Fungi) and Photopexin (Kingdom Bacteria) **(Figure 14)**. This may be evidence of a multi-tiered example of inter-kingdom Horizontal Gene Transfer (HGT) or "adaptive genetic capture". HGT (a.k.a. Lateral Gene Transfer) is a well-known phenomenon, but is usually associated with bacterial genomes, the mechanisms of which (transformation, transduction or conjugation) are all very well studied and accepted.

Historically, however, HGT in eukaryotic organisms has been thought to be, from early on "impossible" to more recently "improbable" to currently "proven" [111]. Considerable amounts of evolutionary investigation have now established that, not only is HGT possible in eukaryotic genomes, it is common enough to have now been observed in all three domains of life [112].

One ecological context in which HGT appears to be particularly common, not only between eukaryotic genomes, but potentially between eukaryotes and prokaryotes (and *vice versa*) is in host-parasite interactions, both in Animalia and Plantae [113, 114]**.** Pertinently, there have recently been several analyses that have suggested HGT from host organisms has occurred in parasitic nematodes [115, 116]. Therefore, there is a reasonable likelihood that the PEX domain (photopexin) seen in the *Photorhabdus* protein is a eukaryotic sequence that has been acquired, though HGT, from either a host that the nematode parasitizes, or some other intermediate associate, and is now used to the adaptive advantage of both the bacterium and

its associated nematode symbiont. If the original function of the domain in the hosts (or intermediate) was to capture, sequester, or remove potentially cytotoxic heme and/or iron, then the bacterium now has the capability of essentially turning an infected insect's immune defense against it as a virulence factor. If the mechanism of virulence then is the capacity to rob the host of iron that can subsequently be used for the metabolism and survival of the bacteria, this in turn would aide in the nematode's ability to thrive.

Of course, there is always the possibility that this structure and associated potential functionality has arisen through strict convergence, but the absence of the discovery, to date, of a broad occurrence of the PEX-domain in prokaryotes argues against that possibility. However, if convergence is the case, it could be interpreted in a pseudo-ecological context as an equally interesting example of a protein evolving as a molecular "mimic" of the host's original "model" domain.

The work with ostreopexin suggests that it may participate in intracellular management of metal (II or III)-chelates [31]. The primary function of the fungal protein 'ostreopexin' is not known, but similar to the proteins in Plantae, it shows reversible binding to hemin with moderate affinity [31]. However, it does not bind to polyamines.

BLAST searches (Positon Specific Iterative (PSI) - BLAST with Max Score cut off of minimum = 200) run on the ostreopexin protein with a Kingdom '*Fungi'* specific filter shows the possibility of some sequence homology with three other fungi, *Rhizoctonia solani* (a known plant pathogen), *Hebeloma cylindrosporum* (an ectomycorrhizal basidiomycete), and *Auricularia subglabra* (a jelly fungus related to the edible *Auricularia polytricha,* common name: cloud ear). However, at this point, all three proteins are hypothetical, based exclusively on genome builds from each of those organisms. Additionally, each, while

meeting the Max Score minimums, was in the low range (204-223). The phylogenetic relationships, if any, of these fungi to *Pleurotus* will be investigated further.

Conclusions

When the analyses in this work were initially undertaken, reports in the databases suggested that the hemopexin protein was composed of two PEX domains connected by a flexible linker (which is correct), but that the primary sequence consisted of only 5 pexin sequence repeats. Therefore, earlier structural investigations were premised on the erroneous indication that each PEX domain was composed of two sequence repeats (HPX repeats) corresponding to blades 1 and 2 of each of the 4-bladed β-propeller domains and the fifth (or, actually, 3rd) sequence repeat was the heme-binding linker. This work, however, has discovered that each domain does, in fact, contain 4 primary sequence HPX repeats that match each of the four blades of the individual domains (see Chapter 1). Therefore, the hemopexin protein contains, in reality, 8 sequence repeats. Repeats 1-4 associated with the N-domain, 5- 8 with the C-domain. Contrary to earlier analyses that described the linker region between the two domains, where the heme moiety is bound, as a "hemopexin repeat", it, in fact, is not.

Exaptation and modular neofunctionalization

In 1979 Steven J. Gould and Richard Lewontin proposed a revolutionary new way of attempting to understand phenotypic traits that might have been, and are being, acted upon by

natural selection [108, 117]. They argued that instead of routinely and habitually utilizing the prevailing "adaptationists" interpretations to tell "just-so-stories" about why particular traits are seen in nature, and that, evolutionary biologists needed to begin to actively and aggressively work toward developing more rigorous analytical techniques to investigate the true origins and trajectories of extant phenotypes. This, of course, pre-dated (and perhaps presaged) the emergence of much more sophisticated molecular techniques, and so was meant to address mostly the analyses of gross phenotypic realities, principally at the individual organismic level. The primary novelty of the paper was to propose that particular traits that appear, on the surface, to have arisen solely by natural selection as a function of simply "being adaptive" may be misleading; in other words they argued against the standard interpretation of the time that, the mere FACT that it is seen, indicates it MUST, by definition, be "adaptive" or it wouldn't be here. A major accomplishment of this publication was to usher in a profound, reflective conversation proposing a re-evaluation of the significance and importance of the concepts of evolutionary constraint (biochemical, biophysical, developmental) and Genetic Drift, as being driving mechanisms that could lead to the extant presentation of phenotypic traits.

Additionally, the *Spandrels* paper identified two other constructs that are particularly relevant to the results presented in this work. First, that classic selection, and adaptation, may have been at work, but in multiple, concurrent, temporal and spatial streams that eventually lead to differential phenotypes. Ultimately, in this sense, the issue becomes, that, even though the trait (in this case the PEX domain architecture) is certainly adaptive, and was probably selected for, the ability to distinguish the variable adaptive significance of the differing forms

is obscured; in this instance, the "differing forms" present as the biochemical plasticity of the domain. This is manifested theoretically as the "problem of multiple adaptive peaks" [108].

The second hypothesis arising from the *Spandrels* paper that is directly relevant to these current findings is the idea of the "spandrel" itself. In their paper Gould and Lewontin coopt an architectural term for use as an analogy to the evolutionary principle they describe. In architecture, a spandrel is a roughly triangular space that is the result of the lateral placement of two or more supportive arches. Historically these "found spaces" were then used by artisans as areas in which to place decorative sculpture or other types of art. The point that Gould and Lewontin make is that the spandrel is a *by-product* of the architecturally "adaptive" functionality of the arch, but is then available to be utilized for purposes other than what was originally "selected for".

In the case of phenotypic traits, the nature of the evolutionary spandrel is that the character may have initially been under positive selection because it is (was) adaptive, in its original sense, but subsequent secondary functionality arises from the already existent trait. The secondary function may itself be adaptive, however, the important distinction is that it (the secondary function) was NOT the reason that the trait evolved in the first place. This particular aspect was further codified and expanded upon in a follow-up paper by Gould and Vrba in which they coined more apt biological terminology and labeled these evolutionary spandrels: "exaptations" [109]. This general line of reasoning actually predates the publication of the "*Spandrels*" paper in a very prescient offering by François Jacob entitled *Evolution and Tinkering* [118].

While the phylogenetic analyses presented in this study show relatively consistent and predictable trajectories of evolutionary association within kingdoms, the between

Kingdom analysis that attempts to uncover the deeper evolutionary history of the PEX domain shows a complete lack of consistent patterns of divergence history **(Figure 14)**. Attempts at building a phylogeny with these data using differing modeling protocols (e.g. - Maximum Parsimony) actually resulted in muddling the relationships even worse (data not shown).

The presence of the PEX domain, specifically in *Pleurotus*, raises an extremely interesting possibility, especially in juxtaposition with Photopexin in *Photorhabdus.* It has been well documented for some time now that *Pleurotus* is a carnivorous fungus. Not only that, but their major prey is nematodes. Oyster mushrooms eat nematodes [119]. Again, the phylogenetic analysis suggests a close sequence identity (potentially interpreted as a close evolutionary relationship) between the PEX-domains of *Ostreopexin* and the bacterial *Photopexin* which is found in the gut flora of a host nematode. While no definitive evidence exists in these analyses, the speculation arises that there may be a parasitic/symbiotic mechanistic explanation for the apparent close evolutionary relationship of these particular two PEX-domain containing proteins. Additionally, some very recently published research sheds some extremely pertinent light on the ability of the pathogenic *Pleurotus* to rapidly adapt to changing host mediated environmental selective pressures[120]. If the use of the PEX-domain would be of any adaptive advantage, especially as relates to iron acquisition, to this facultative plant parasite, then it is not improbable that the *Ostreopexin* PEX-domain has been exapted from either a host plant organism, or from its predator-prey relationship to a nematode that had previously exapted it from elsewhere. Again, the lack of an established pattern of divergent heredity within the taxon (Fungi) might be interpreted as evidence for a

HGT event manifested in the genome of the virulently plant-parasitic oyster mushroom (that also happens to prey upon nematodes).

Recent work on the possibility of exaptation at the protein domain level has resulted in some very intriguing modeling showing that there is a fine balance between two critical biophysical traits: binding affinity for target ligands and tertiary folding stability. This research resulted in the hypothesis that at the protein domain level there exist fitness landscapes that give rise to evolutionary coupling between binding strength and folding stability [121]. Using a mathematical model they showed how biophysical protein traits (structure) can develop as exaptations (spandrels) even though they do not impart an "intrinsic fitness advantage" [121].

These discoveries have much more to do than with just esoteric discussions of evolutionary theory. They have real world empirical applications when investigating the role that protein structure may play in, for example, applied medical research:

> …such proteins may have divergent fates, evolving to bind or not bind their targets depending on random mutational events. These observations may explain the abundance of apparently nonfunctional interactions among proteins observed in high-throughput assays. [121]

The results of these analyses combine to lead to the conclusion that the other proteins that have identifiable PEX domain structure have co-opted the 4-bladed β-propeller which has then been modified for variable functionality. This research has verified that the β-

propeller structure has *not* been maintained via consistent amino acid sequences through time. Instead, the propeller structure has remained relatively unchanged even as the component residues of the propellers have been labile over long periods of evolutionary history.

Despite the difficulty in tracing obvious homology, it is most likely that the surviving PEX domains found throughout extant biota share a common evolutionary origin. It is extremely unlikely that the similarities seen at the structural level are a mere coincidental aspect of convergent (or even parallel) evolutionary processes. However, one very concrete conclusion drawn, based on these analyses, is that there has been so much evolutionary time for the primary sequences to have diverged so significantly from one another at the nucleotide and concomitant residue levels, as to be almost "unclade-able".

Because the domain simply does not conform to traditional taxonomic patterning, one conclusion might be that the "adaptive capacity" of the domain lies in its *structure*, and attendant functional plasticity. Consequentially, it is the associated biophysical and biochemical properties that unite the domains at the biological level, not their primary sequences.

CHAPTER 5

DETERMINATION OF THE TERTIARY STRUCTURE OF THE PEX DOMAIN IN THE HUMAN PROTEOGLYCAN-4 (LUBRICIN) PROTEIN

Introduction

Proteoglycan-4 (PRG4) is the name given to the protein encoded by the *PRG4* gene in mammals. In the literature this protein has been variously named: superficial zone protein (SZP), megakaryocyte stimulating factor (MSF), and lubricin [85]. Currently, it is most commonly known as "lubricin", a descriptive name that refers to its primary biological function, which is to act as a lubricant in the synovial joints of vertebrates [122].

PRG4 is a glycoprotein, of 151 kDa approximate size, synthesized by articular chondrocytes and certain cells associated with the tissues of the synovial linings of the articular joints [123]. Dysfunction in this protein is rare, but those that exhibit clinical symptoms suffer from a disease state known as Camptodactyly-Arthropathy-Coxa vara-Pericarditis (CACP) syndrome, an autosomal recessive trait (OMIM 208250) [124]. Patients who inherit any of the mutations that compromise lubricin functionality are born with normal joint activity, but as they age they begin to exhibit a progression of symptoms that starts with hyperplasia of the synoviocytes of the joints [125]. This in turn leads to progressive malformation of the digits of the hands and feet (camptodactyly) and eventually, severe arthropathies that result in premature contracture of their joints. In many, they progress with age to severe hip deformity (coxa vera), and pericarditis [126, 127].

Protein domains are defined as specific portions of proteins that are evolutionarily conserved in primary sequence and tertiary structure [128]. They are also assumed to be able to

exist independently from the parent protein in which they are typically found, and have the ability to obtain their own stable 3-dimensional structure [128]. Many specific domains are found in multiple proteins and can be described as molecular "building blocks" that evolution has available for rearrangement into new and potentially adaptive protein function. The primary structure of the human A-isoform of the PRG4 protein contains a polypeptide of 1,404, residues. It is not atypical that proteins of this size are composed of several potentially functional modular components, or domains [128]. However, even though the structure of PRG4 has yet to be determined, it seems that this particular protein may contain only one identifiable modular domain: the 4-bladed β-propeller PEX domain [2, 129-131] (also see: **Chapter 1**). The remainder of the protein, as best we can determine through homology modeling, is unstructured (**Figures 23 and 24)**.

PRG4 is one of a variety of classes of proteins that contain the PEX domain. This domain is known to also occur in most of the Matrix Metalloproteinases (MMPs) and vitronectin (VTN), and the blood glycoprotein Hemopexin (HPX), for which the domain is named [132-134]. In this section of this study we propose to express and purify the Human PRG4 (proteoglycan-4) PEX domain, obtain the tertiary structure of the domain using nuclear magnetic resonance (NMR) spectroscopy and/or crystallization, and perform phylogenomic structural sequence analyses to aid in placing the PRG4-PEX domain in its proper evolutionary context in relation to the other PEX domain containing proteins.

The principal significance of understanding the evolution, and hence functionality, of this domain is that mutations in the genes that code for, and malfunctions in their subsequently translated protein products, are known to be implicated in the development

and/or exacerbation of a multitude of human diseases including CACP syndrome, rheumatoid arthritis, several types of cancers, cirrhosis of the liver, and multiple sclerosis, among others [87]. We hypothesize this work will help to identify structural components of these proteins at the modular/molecular level that could be potential targets for drug therapy design aimed at mechanistic intervention of the PEX domain function.

Often, the determination of the three-dimensional structure of any given protein domain is extremely valuable (sometimes even crucial) information necessary for understanding the biological functions of the protein molecules in which they are found. Additionally, and most importantly for the purposes of this study, protein domains can evolve. Recently, the analytical approach known as phylogenomics has been developed to provide a framework for evolutionary analyses that can be performed using the previously established 3-dimentional structural aspects of protein domains as character states in cladistic analyses [135]. One fundamental issue is that the structure of the PEX domain of PRG4 has not been determined. While certainly not the only PEX domain that lacks resolved structure, we believe the PRG4 protein PEX domain may provide a critical link in understanding the evolutionary implications of the variation in tertiary structure of this domain between the various protein classes in which it is found.

Previous bioinformatic work has shown that the PRG4 protein contains a single C-terminal PEX structural domain, and most significantly, earlier phylogenetic analyses of genes known to code for proteins containing PEX domains have suggested that the PEX domain seen in PGR4 is the evolutionarily *ancestral* sequence [107, 136]. Therefore, at least in Animalia, the PRG4-PEX domain has been interpreted to be the root of the phylogeny at the domain level. Paradoxically, previous work has suggested that hemopexin is the most derived

of the proteins within the group, while establishing PRG4 as the evolutionary progenitor of the domain[107]. This phylogenomic work, however, sheds some doubt on that interpretation and places the 2 PEX domains of the HPX molecule at the root of the phylogeny, signifying it as the ancestral condition **(Chapter 3)**. A major impetus for deeper understanding of PEX domain evolution will be to build a parallel phylogeny of the proteins that are expressed by these genes to evaluate congruencies and discrepancies between the two hypotheses.

While the 3-D structure is known for the mature hemopexin protein, and several of the MMPs, the structure of the PRG4 PEX domain remains elusive. We propose to use X-ray crystallography and nuclear magnetic resonance (NMR) techniques to establish the 3-dimensional structure of the PGR4 PEX domain. This, in turn, will facilitate building a more robust evolutionary phylogeny of the domain structure, and have the PGR4 representative to bolster the phylogenetic hypotheses.

From a more theoretical perspective, these analyses should shed light with deep implications on the evolution of the highly specialized and adaptive capacity of the exclusively vertebrate acquired (a.k.a.- adaptive) immune system [137]. Whereas, the MMPs are known to function as components of the ancestral innate immune system, the hemopexin protein (proper) belongs to the acute phase class of immune reactants in some animals (although not in humans) and is secreted through induction after an immune system inflammation response, exclusive to the derived adaptive system dynamics [138, 139]. Therefore, it is possible that identifying the origin of the PEX domain will provide a better understanding of the evolution of adaptive immunity and hence a clue to therapeutic approaches targeting immune dysfunction at the molecular level.

Even though PRG4 is a very heavily *O*-glycosylated protein, evidence suggests that very little, if any, glycosylation occurs in the C-terminus PEX domain region [140]. As a consequence, this portion of the protein should be amenable to expression in a prokaryotic system. X-ray crystallography and Nuclear Magnetic Resonance (NMR) are the two techniques most often employed to obtain data that can then be used to calculate the relative positions of atoms, in space, within molecules that can be used in the determination of protein tertiary structure. X-ray diffraction is used on single crystals of purified protein product, and is highly effective at aiding in the determination of 3-D structure in comparatively large mature proteins. In contrast, NMR is frequently more effective at illuminating detail of atomic-resolution for smaller constituent domains or structural motifs of larger mature proteins, and at detecting inter-domain movement and conformational shifts [141, 142].

Solution-NMR is a technique in which indirect information can be obtained that can be used to calculate the three-dimensional structure of proteins. Measurements are made directly on the protein sequence in a near physiological chemical environment [141]. For the purposes of this work, we are interested in determining the structure of the single PEX domain of lubricin (PRG4).

Methods and Results

Cloning, expression and purification

The prime target of this experiment is the recombinant production of the PEX domain from the PGR4 protein. The nucleotide sequence for cloning into an expression vector was obtained from Integrated DNA Technologies (IDT), in the form of their commercial product known as a gBlock® [143]. The model we used to build the synthetic PRG4-PEX domain gene was obtained directly from the primary sequence reported in the NCBI database from Human PRG4 Isoform B (NCBI Reference Sequence: |NP_001121180.2).

gBlocks are synthetic, double-stranded, oligonucleotide gene fragments, the sequences of which are dictated specifically by the investigator and are optimized for codon usage by particular expression systems. In this case, we optimized the sequence for expression in *E. coli*, based on the translational primary structure of the human PRG4-PEX domain. As a consequence, even though actual nucleotide sequences differ somewhat between the human "model" and the synthetic gBlock codons, the primary structure of the peptide is identical. Ultimately, the primary aim was to insure a synthetic sequence that would "mimic" the code for the PRG4- PEX domain region of the human PRG4 gene, but be adjusted for more efficient translation in *E.coli*. The nucleotide sequence for the gBlock used in this work, as ordered, and the primary structure of the polypeptide as expressed are shown in **Appendix E**.

Cloning - Introduction

An optimized series of laboratory protocols for cloning and expression of proteins in *E. coli* was employed to attempt to isolate, purify, stabilize and prepare the peptide sequence for attempts at crystallization and spectroscopic analyses [144]. We determined that we would attempt to clone the PRG4 PEX domain (gBlock) into the pET-44a vector obtained through EMD-Millipore [145]. We decided to insert the gene into the vector between the *Eco*RI and *Xho*I restriction sites. A schematic of the pET-44a vector map is provided in **Appendix F**.

In addition to providing ampicillin resistance in the soon to be transformed *E. coli*, the pET-44a vector is fused with a 495 amino acid protein from *E.coli* called "N-utilization substance protein A" (NusA). This protein is known to function in a variety of cellular and viral transcription termination and antitermination processes [146]. However, the prime reason for using this construct is that NusA is known to have among the highest solubility of all bacterial proteins [147]. Having attempted an expression protocol using a different cloning vector/cell strain combination (vector plasmid pHisGb1 and DH5a competent cells) but not achieving success in purification due the product being highly insoluble (results not shown), we reasoned that by using this vector we might "drag" the target peptide (the PRG4-PEX domain) into solution via the high solubility properties of the Nus-Tag. In addition to the NusA-Tag, this construct has a poly-histidine tag (His-Tag) used for purification by immobilized metal ion affinity chromatography (IMAC) [145]. Between the NusA sequence and the PEX domain is a protease recognition sequence for enterokinase (rEK). After digest with enterokinase (rEK) the resulting protein will have a nine amino acid (SPGARGSGF)

N-terminal tail before the first residue of the PEX domain as a leftover artifact from the fusion protein.

PCR gBlock (PEX domain) amplification for cloning

Polymerase Chain Reaction (PCR) is a well-established methodology for amplification of nucleotide (gene) sequences via *in vitro* enzymatically assisted (DNA polymerase) replication [148]. The sequence optimized gBlock (synthetic gene) for the PRG4 PEX domain and appropriately calculated primers for PCR amplification and cloning were all obtained in lyophilized form from IDT, and suspended in ultra-pure deionized water as per the manufacturer's instructions to achieve 100 µM concentrations of each. Actual volumes and additions are described after sequences below. The sequence of the gene (gBlock) is provided in **Appendix E**. The primers used for PCR amplification were:

Forward primer:

Sequence name: PRG4_*Eco*RI PEX-forward:

Oligo Sequence: 5'- GCT ACA GTC ***GAA TTC*** CCG AAC CAG GGC ATT -3'

Calculated Molecular Weight: 9,176.0; Measured Molecular Weight: 9,175.4

Amount received: 34.6 µmoles + 346 µl = [100 µM] of forward primer.

The "cut" sites for both restriction enzymes are shown in ***RED***:

Reverse primer:

Sequence Name: PRG4_*Xho*I_PEX2-reverse:

Oligo Sequence: 5'- ACG GTC TCG ***CTC GAG*** CTA TGG GCA ATT GTA CCA -3'

Calculated Molecular Weight: 10129.6; Measured Molecular Weight: 10129.60

Amount received: 29.4 µmoles + 294 µl = [100 µM] of reverse primer.

Reactants for the PCR amplification were all mixed into a standard 0.2 mL low retention PCR tube, **Table 4** [149].

Table 4. Reactants and volumes used in the PCR amplification of the PRG4-PEX domain gene (gBlock).

REACTANTS	VOLUME
10X polymerase buffer	5µl
DMSO	3µl
dNTP (deoxyribonucleoside triphosphates)	1µl
5' forward primer	1µl
3' reverse primer	1µl
Template DNA (PEX-gene)	1µl
PFu DNA polymerase	1µl
Ultra-pure deionized water	37µl
TOTAL	50µl

PCR reactions were run on a Perkin Elmer GeneAmp 2400 PCR system [150], using the automated program shown in **Table 5.** The DNA polymerase we chose is an enzyme that has been isolated from hyper-thermophilic archaeon *Pyrococcus furiosus* (abbreviated Pfu). *In vivo* this polymerase not only drives replication of the organism's genome, but it also functions as a proofreader, recognizing and repairing any mis-incorporated nucleotides in the

3' direction of extension. This multiple functionality along with innate superior thermostability ostensibly results in PCR products with fewer errors than those generated by other available retail polymerase products (e.g. Taq) [151]. DMSO was added to assist in keeping the strands of the DNA separated and increasing replication efficiency, especially in G-C rich regions [152].

Table 5. PCR program for amplification of PRG4 PEX domain gene (gBlock).

TEMPERATURE	DURATION	NOTES
95° C.	5 minutes	Hot Start
95° C.	2 minutes	>>>Sequence repeat 30x
52° C.	1 minute	
72° C.	1.5 minutes	
4° C.	Over night	Cool down and storage

Preparation of agarose gel for isolation of PEX domain oligonucleotide (PRG4 PEX gene)

A 1% agarose gel solution was prepared by adding 1g agarose (Midsci: BE-A500, Bullseye agarose, General Purpose [153]) to 100mL of standard TAE (TAE is Tris-Acetate-EDTA) buffer; heating in a microwave oven for 1 minute to dissolve the agarose; cooled to approximately 50° C (just warm to touch). To the cooled but still liquefied gel prep 2µl of GelRed Nucleic Acid Stain was added for eventual visualization of oligonucleotides [154]. The gel was poured into the electrophoresis tray and allowed to sit at room temperature for 20-30

mins, until completely solidified. The gel was placed into the electrophoresis chamber, the unit filled with 1xTAE until the gel was submerged. Two µl of loading buffer/blue dye (5x SDS loading dye = 250mM Tris pH 6.8, 30% glycerol, 10% SDS, 5% β-mercaptoethanol, 0.02% bromophenol blue) was mixed with 8 µl of PCR product (total 10 µl) for well loading accuracy and increasing sample viscosity. A ThermoFisher Scientific 1Kb GeneRuler molecular weight ladder/standard was loaded (2 µl dye + 2 µl ladder + 8µl H_2O = 12 µl total volume) into lane 1 of the gel for size calibration [155]. Replicate samples of the PCR product were loaded into lanes 2 and 3, and run for 30 minutes at 110V to observe the presence of an appropriately sized nucleotide band representative of the target (PEX domain gene - 828 bp: **Appendix E**). Results are presented in **Figure 15**.

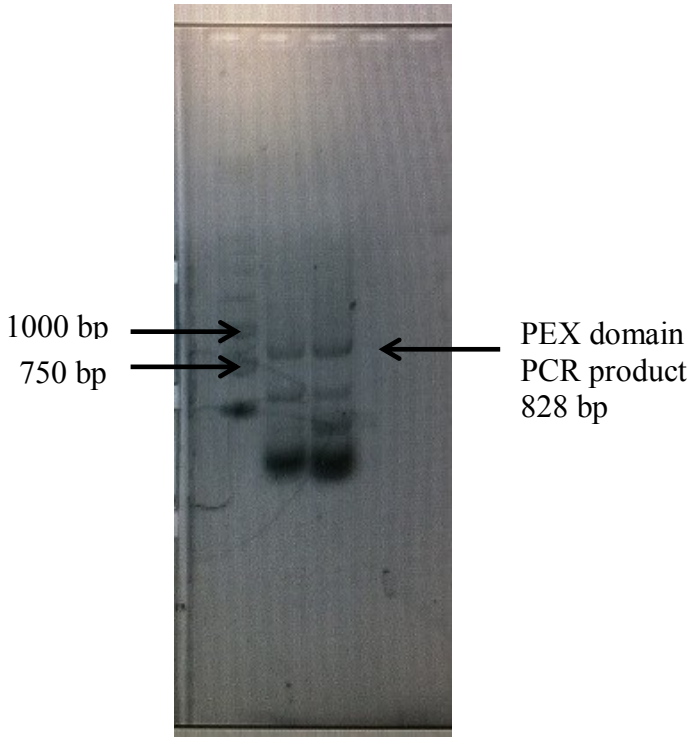


Figure 15. Agarose gel resolving the PEX-domain PCR product, for check on the efficacy of amplification.

The target "gene" is 828 base pairs (bp) in length. Lane 1 is the DNA size ladder/standard. The top band in each of the two sample lanes (lanes 2 and 3) falls between the 750bp and 1000bp marker, and was interpreted as being the target product. The smaller bands further down the gel were not analyzed and are presumed to be by-products of the PCR reaction that arose from secondary priming sites. The lowest "cloud" is most likely primer. Several PCR experiments were performed. The example gel was typical of the pattern consistently observed when amplifying the PEX domain, but represents the clearest result.

Separate gels, for isolation of the plasmid vector (pET-44a) DNA were run following the same protocol (data not shown).

PCR clean up - DNA purification by centrifugation

The gene DNA product was purified using the Wizard SV Gel and PCR Clean-up System – A9281 from Promega [156]. Selecting a visually high quality electrophoresis the best band associated with the PCR product was excised and placed it in a pre-weighed 1.5 ml microcentrifuge tube. The tube plus gel extract was weighed to determine the amount of gel present. The gel slice for the gene fragment weighed 740mg. 740µl of Membrane Binding Solution was added, and the tube was vortexed (agitated) for 20 seconds and set in a 42^{o} C water bath until gel was completely melted. The solution was briefly vortexed for complete mixing and transferred to an SV Minicolumn inserted into a collection tube and was allowed to incubate at room temperature for 1 minute, then centrifuged at 14800 rpm for 1min. Flow through was discarded and 700µl of Membrane Wash Solution (mixed with ethanol) was added to the SV Minicolumn which was centrifuged, again, at 14800 rpm for 1 min. The flow through was discarded. 500µl of Membrane Wash Solution was added to the column and centrifuged at 14800 rpm for 1 min. The flow through was discarded and the empty minicolumn centrifuged again at 14800 rpm for 5 min. The minimal flow through was removed and the tube was re-centrifuged at 14800 rpm for 1 min to facilitate evaporation of residual ethanol. The minicolumn was transferred to a clean 1.5mL microcentrifuge tube. 50 µl of nuclease free (ultra-pure) water was applied to the center of the minicolumn filter, and allowed to equilibrate at room temperature for 1 minute. The tube was centrifuged at 14800 rpm for 1 minute to collect the DNA eluate. The SV Minicolumn was discarded and the DNA eluate (~50µl) stored at 4^{o} C.

Plasmid digest and gel purification

Both insert (PEX domain gene) and the vector (pET-44a plasmid) were digested in separate tubes in preparation for ligation according to the recipes in **Table 6**.

Table 6. Recipes for Insert (PEX domain gene) and Vector (pET-44a) digestion in preparation for DNA Ligation insertion of gene into vector plasmid.

RXN Digest	
Insert (PEX gene)	50µl
NEBuffer 2*	6 µl
Enzyme - *Eco*RI	2 µl
Enzyme - *Xho*I	2 µl
TOTAL VOLUME	60 µl

VECTOR Digest	
Vector (pET-44a)	50µl
NEBuffer 2*	6 µl
Enzyme - *Eco*RI	2 µl
Enzyme - *Xho*I	2 µl
TOTAL VOLUME	60 µl

* The NEBuffer 2 was chosen because it is rated as 100% efficient with both *Eco*RI and *Xho*I enzymes. The biochemical properties of digestion buffer "NEBuffer 2" are indicated in **Table 7**.

Table 7. Biochemical properties of digestion buffer "NEBuffer 2" [157].

1X Buffer Components
50mM NaCl
10mM Tris-HCl
10mM $MgCl_2$
1mM DTT
pH 7.9@25°C

Each digestion tube was spun for 1 minute to mix the solution and incubated in a 37^{o} C water bath overnight.

Purification of digested insert and vector DNA

Both the insert and vector digests were run on a 1% agarose gel, the bands identified, verified for size, excised and gel purified as described above, with the exception that the vector band gel weighed 820 mg and the insert band gel weighed 670 mg. Additions of reagents were adjusted accordingly.

DNA ligation

The T4 DNA Ligation Protocol of Promega corporation was followed [158]. It was determined that the PGR4-PEX gene would be inserted into the pET-44a vector plasmid between the *Eco*RI and *Xho*I restriction sites; see **Appendix F**, for pET-44a vector map and location of restriction enzyme cut sites.

Concentrations of both vector and insert solutions were determined through Optical Density at A_{260} using an Eppendorf BioPhotometer 6131 Spectrophotometer [159]. The concentration of DNA insert (PEX domain gene) was determined as 22.5 ng/µl and that of the vector (pET-44a) as 15.6 ng/µl. Due to the need to account for the difference in size of the two sequences (PEX domain = 828 bp; pET-44a = 7711 bp) conversions were calculated to derive the appropriate volumetric additions of each for both a 3:1 and 9:1 molar ratio ligation mixtures.

The resulting ligation reaction mixtures are indicated in **Table 8**. A Vector Control of 3.2 µl Vector DNA (pET-44a) with 6.8 µl Ultra-pure water (total volume 10 µl) was included.

Table 8. Recipes for ligation reactions at 3:1 and 9:1 molar concentration ratios of insert (PEX domain gene) to vector (pET-44a), note conversion factors in text above.

Ligation RXN 3:1		**Ligation RXN 9:1**	
Insert DNA (PEX gene) (16.7 ng)	0.74 µl	Insert DNA (PEX gene) (50 ng)	2.22 µl
Vector DNA (pET-44a) (50 ng)	3.2 µl	Vector DNA (pET-44a) (50 ng)	3.2 µl
T4 Ligase buffer	10 µl	T4 Ligase buffer	10 µl
T4 Ligase	1 µl	T4 Ligase	1 µl
Ultra-pure water	5.06 µl	Ultra-pure water	3.58 µl
TOTAL VOLUME	20 µl	TOTAL VOLUME	20 µl

The reactions were incubated at 16^{o} C overnight.

Transformation

After ligation was complete three, 20 µl cells/aliquots of DH5a competent *E.coli* cells were removed from the -80^{o} C freezer and allowed to thaw on ice[160]. Two µl of the 3 ligation reactions (3:1 molar ratio; 9:1 molar ratio; and control) were added to the separate tubes of cells and incubated on ice for 30 minutes. The cells were heat shocked by partially submerging the tubes in a 42^{o} C water bath for 40 seconds. The cells were allowed to recover for 5 minutes on ice. The cells were transferred to new microcentrifuge tubes containing 100 µl of Lysogeny Broth (LB - Luria formulation: 10 g tryptone + 5 g yeast extract + 5 g NaCl

suspended in 800 ml deionized water + additional DI H_2O to constitute 1 L volume; autoclaved at 121 °C for 20 mins). Preparations were incubated with 250 rpm agitation at 37° C for 1 hour. Tubes were removed from the incubator and spun at 16.1 xg for 2 minutes at room temperature. Each sample was added separately onto LB plus ampicillin (0.1 µg/ml concentration) agar plates using disposable sterile plastic cell spreaders (one for each) and incubated a 37° C overnight to allow colony growth. Viability of cells was assured by success of colony growth evaluation experiments described below.

Colony growth evaluation

Plates were removed from the incubator and visually assessed for bacterial colony formation. There were 3 colonies on the vector control indicating that most of those bacteria did not acquire ampicillin resistance. This was expected because, even though the bacteria most likely took up the treatment, since the vector was empty (un-ligated) the transformation did not impart ampicillin resistance. There was growth present in both the 3:1 (8 colonies) and 9:1 (29 colonies) molar ratio treatments.

Colony PCR

Colony PCR methodology is a convenient method for rapidly testing whether a plasmid transformation was successful on a per colony basis. This protocol is a high-throughput approach to rapidly test for plasmid uptake without having to run a time

consuming, and potentially unsuccessful, plasmid extraction miniprep on each colony separately. All twelve colonies used were from the 9:1 molar ratio treatment plate.

DNA plasmid extraction from E. coli colonies

For the PCR experiment the protocol described above for the *PCR gBlock (PEX domain) amplification for cloning* was used, except that each PCR reaction tube was inoculated separately with an individual bacterial colony prior to the addition of the PCR master mix, and excluding the PRG4 PEX domain gene addition. The PCR products were analyzed for the presence or absence of the plasmid by standard 1% agarose gel electrophoresis, also described above.

Twelve culture tubes (one for each selected colony) were prepared by adding 4 ml of liquid LB mixed with 4µl ampicillin (0.1 mg/ml concentration) to each. The colonies were visually selected for the colony PCR experiment and sampled with sterile pipet tips, which were used to inoculate both the PCR tubes and the broth in the growth tubes. The results of the gel analysis of the colony PCR amplification experiment using the 12 colonies selected is shown in **Figure 16.**

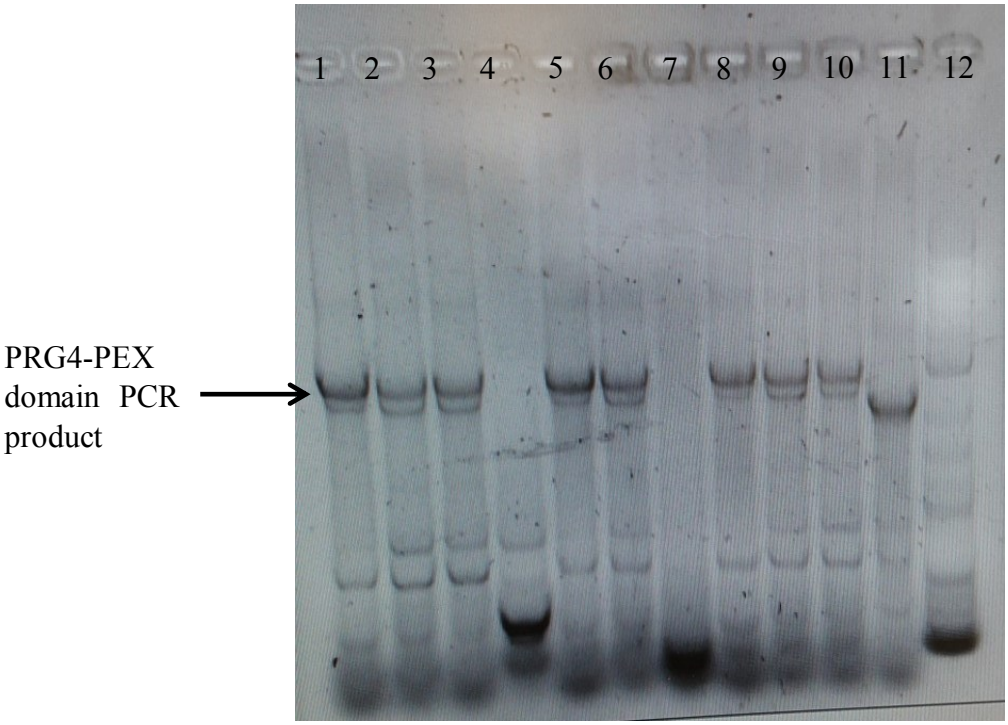


Figure 16. Colony PCR of transformed ligation products of the PRG4 PEX domain-pET-44a construct. Colonies were taken from the 9:1 molar ratio treatment plate. Colonies 4, 7 and 11 (equivocal) appeared not to contain the PEX domain insert. This gel was run without a standard or an "empty" plasmid run, by design, based on information gained from prior experiments concerning expected protein composition and size category of the expression of the PRG4 - PEX domain (see **Figure 15** above).

Based on colony PCR results, samples 1, 5, 8, and 9 were selected for culture growth overnight in a 37^{o} C shaker at 200 rpm. From each culture 1.5 ml was collected into sterile microcentrifuge tubes for further processing by plasmid DNA extraction using the IBI High-Speed Plasmid Mini Kit protocol [161].

Tubes were centrifuged at 14800 rpm for 1 min. to pellet the *E.coli* cells, and the supernatant was discarded. To each tube was added 200 µl of PD1 buffer containing

RNase-A (see protocol for components and properties of buffers used [161]). Cells were resuspended by vortexing. 200 µl of PD2 buffer was added to each tube and gently mixed by 10 times tube inversion. After 2 minutes at room temperature 300 µl of PD3 buffer was added, 10 times inversion mixed and centrifuged for 3 minutes at 14800 rpm. Each supernatant was pipetted off and placed into PD columns set in 2 ml collection tubes and centrifuged at 14800 rpm for 1 minute. Flow throughs were discarded and columns were loaded with 400 µl of W1 buffer. After centrifugation at 14800 rpm for 30 seconds the flow through was discarded. 600 µl of Wash Buffer (with ethanol added) was placed on each PD column and centrifuged at 14800 rpm for 30 seconds. The flow through was removed and the empty PD column/tubes re-centrifuged at 14800 rpm for 3 minutes to dry the PD column matrix. The dried PD column was transferred to a new sterile microcentrifuge tube and 50 µl of Elution Buffer was pipetted onto the center of the PD column matrix. After 3 minutes at room temperature to allow for matrix saturation, columns were centrifuged at 14800 rpm for 2 minutes to elute the plasmid DNA into the collection tube. The yield was approximately 50 µl of eluted DNA from each sample.

Recombinant sequence verification

Samples of purified plasmids were sent to the University of Missouri DNA Sequencing Core Facility for analysis [162]. Colony 8 was found to contain no errors (see **Appendix G**) and was subsequently used for expression tests.

Protein expression and purification

Additional plasmid was purified by miniprep from colony 8 for transformation into *E. coli* strain BL21-DE3 competent cells, for protein expression, following the heat-shock protocol outlined above. Initial attempts at protein expression using standard strength LB medium failed, as insufficient quantity of protein was produced for continued analyses. However, higher yields of the NusA-PEX fusion polypeptide were obtained using double concentrated LB medium as a growth environment.

1 L of 2xLB media was inoculated with a 1:20 overnight culture of transformed bacteria and incubated for 2-3 hours at 37^{o} C, with shaking. Protein expression was induced by addition of 1 mM final concentration of isopropyl β-D-1-thiogalactopyranoside (IPTG), the incubation temperature was lowered to 18^{o} C, and incubation continued for 12-16 hours. Cells were collected by centrifugation at 1000 xg for 15 minutes at 4^{o} C in a Beckman JA -10 rotor. The supernatant was discarded and the cell pellet was frozen at -20^{o} C.

The cell pellet was thawed on ice, re-suspended in 100 mL Lysis buffer (20mM Tris pH 8.0, 500 mM NaCl, 5% glycerol, 5 mM imidazole) with 10 μM leupeptin. Cells were lysed by passing 3x through a micro-fluidizer via gravity load [163]. The lysate was centrifuged at 10 Kxg for 1 hour. The supernatant was passed over a 2.5 cm column packed with 2.5 ml *Ni* Sepharose 6 Fast Flow medium, equilibrated with 25 ml of Lysis buffer [164]. The column was washed with 25 ml of Lysis buffer containing 100 mM imidazole to remove non-specific binding proteins. The NusA-PEX domain protein was eluted by 25 ml of Lysis buffer with 250 mM imidazole (a total of approximately 4-5 mg of protein eluted in the initial 2-3 ml).

As this step was simply a test to determine if using the 2x LB medium was effective at producing more protein than with 1x LB, only one round of gel de-staining was performed. This was sufficient for visualization of target protein as seen in **Figure 17**.

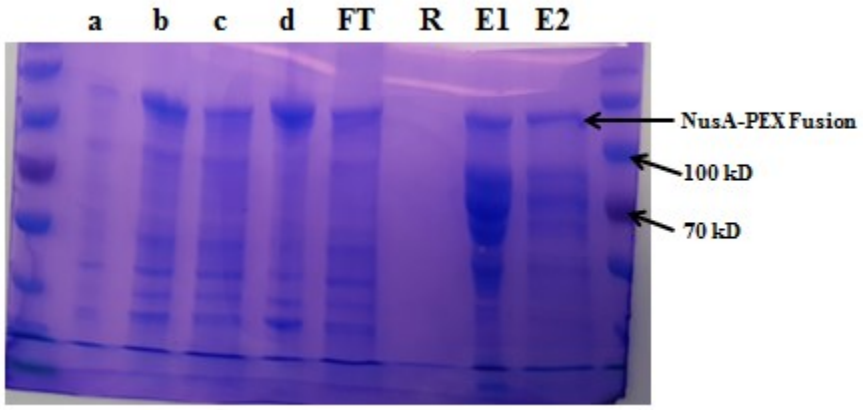


Figure 17. Change in media to double LB - Soluble NusA-PEX fusion protein is now present in the supernatant and the 2nd elution fraction (E2) is considerably cleaner than E1. a = pre-induction; b = crude lysate; c = supernatant; d = pellet; FT = flow through; R = rinse; E = elusions (progressive). The same size standard is run in both lanes 1 and 10 [165].

The molecular weight of the NusA tag is ~54 kDa and the PRG4-PEX domain is ~31.51 kDa [166] . Therefore, the topmost bands that present in the two elution lanes (E1, E2) on the gel above are interpreted to be the NusA-PEX fusion target.

Protein production/elution maximization

Several rounds of expression and elution experiments were run to determine the best conditions for maximization of protein quantity and reduction of non-target background

proteins. One set of tests was designed to establish the most procedurally productive concentration of imidazole for comparatively highest quality of protein elution **(Figure 18)**.

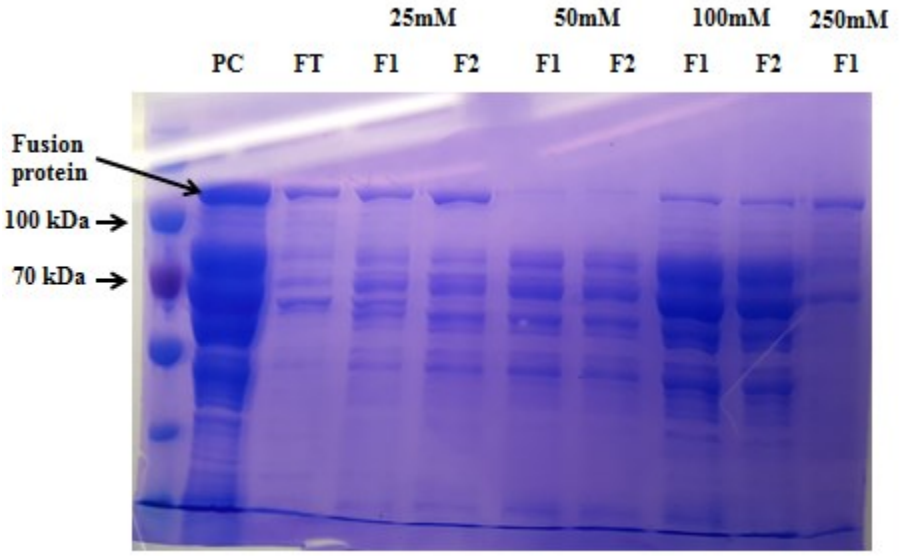


Figure 18 – One example of results from experiments run under the same background conditions as the gel shown in **Figure 17**, but manipulating concentrations of imidazole required to produce the cleanest fraction of NusA-PEX domain fusion. Each lane was loaded with 10µl of materials. PC = Pre-column; FT = Flow Through; F = Fractions

Figure 18 is a representative gel of the expression tests run to determine if changing the concentration of imidazole would affect the quality of protein expression. These experiments were only looking at the relative purity of the expression band for direct comparison of imidazole concentration efficacy. Consistently, the 25mM imidazole concentration produced the cleanest expression patterns and so was initially chosen for future expressions (however, see modifications to procedure associated with subsequent salt addition experiments below.

A second round of experiments was designed and conducted to determine if the addition of various divalent metals to the expression medium would positively affect target

protein production. The logic behind these experiments was that the PEX domain, in all cases known, is a scaffold for the binding of a variety of potentially catalytically active metal ions (Zn, Ca, Na, among others [167]). It was reasoned that addition of these ions might potentially result not only in increased quantity of target expression, but would also possibly aid in post-expression folding of the domain. The series of metal salt additions included $CaCl_2$, $FeSO_4$, $CuSO_4$ and $MgCl_2$. There was no appreciable improvement in target protein production regardless of additive, and therefore, no data are presented from those experiments. From this point, based on the results of this suite of expression tests, a 25mM imidazole concentration preparation was chosen to move forward for use in isolation and purification protocols.

Digestion of target NusA-PEX domain fusion protein

The selected protein fraction (25mM imidazole preparation) was buffer exchanged back to Lysis buffer (10ml G50 column equilibrated with 10x column volume Lysis buffer) to remove the excess imidazole. The sample was diluted to ~0.5 mg/ml and concentrated to 1-2 mg/mL [168]. Concentration was measured by Bradford protein assay at 595 nm absorbance [169]. The fusion protein was digested with 1 unit of recombinant His-tagged enterokinase in 50 µl of 1x rEK cleavage/capture buffer for 24 hours at room temperature, **Table 9** [170].

Table 9. Recipe for digestion reactions involving the NusA-PEX fusion protein.

REACTANTS	VOLUME
10X rEK Cleavage/Capture Buffer	5 µl
NusA-PEX fusion protein (50 µg)	25 µl
Diluted rEK (1 unit/µl)	1 µl
Deionized water	19 µl
Total volume	50 µl

100 µl of *Ni* sepharose beads equilibrated with 1 ml Lysis buffer were added to the solution (30 minutes with *gentle* shaking) to bind the uncut fusion protein, digested NusA protein and enterokinase. *Ni* sepharose beads were separated from the solution by packing in an empty 1 ml (1 cm diameter) column and washing with 200 µl Lysis buffer.

Separation of NusA tag from PEX domain

During the 24 hour digestion cycle, a time-course series of samples were taken at 2, 4, 8 and 19 hours. For each sample 10 µl of digestion solution was removed and the digestion reaction halted by addition of 2 µl of loading dye and heating in a micro-tube heating block for 5 minutes at 95^{o} C. After the reaction was quenched the samples were maintained at room temperature until loading onto the gel. Results are presented in **Figure 19**.

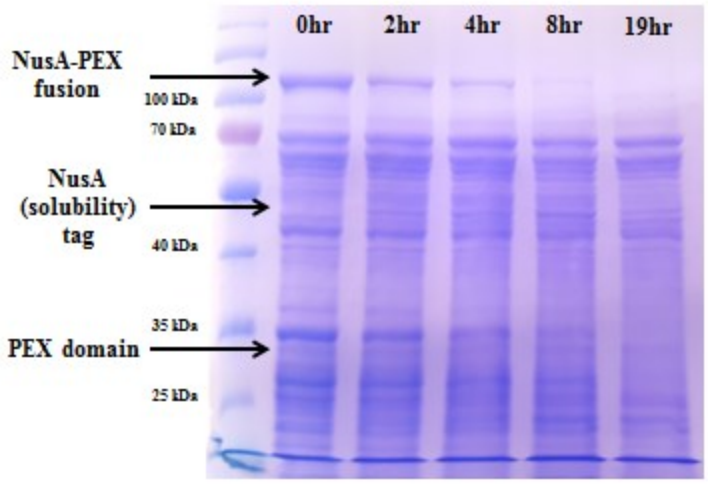


Figure 19. Time-course digest of the 25mM Imidazole fraction with Recombinant Enterokinase (rEK).

The NusA-PEX fusion degrades at a relatively steady rate with the concomitant (albeit weaker) appearance and increase in the amount of both free NusA and free PEX. By 19 hours the NusA-PEX fusion is completely digested. Additionally, all other contaminants appear unaffected by the rEK suggesting they are not truncation variants of the fusion. However, the persistence of these contaminants is indicative of considerable non-target product being expressed.

Considering that the digest shows there is some PEX domain being expressed, but there are still high levels of non-target protein contamination, the previous experiment was repeated and the gel was de-stained multiple rounds to minimize background. The gel was submitted for Mass Spectroscopy analysis in the Mass Spec lab of UMKC-SBS run by Dr. Andrew Keightley. The analysis confirmed that boxes A and C are the NusA-PEX fusion protein; box B is the PEX domain and box D is GroEL, **Figure 20**.

Mass Spectrometry

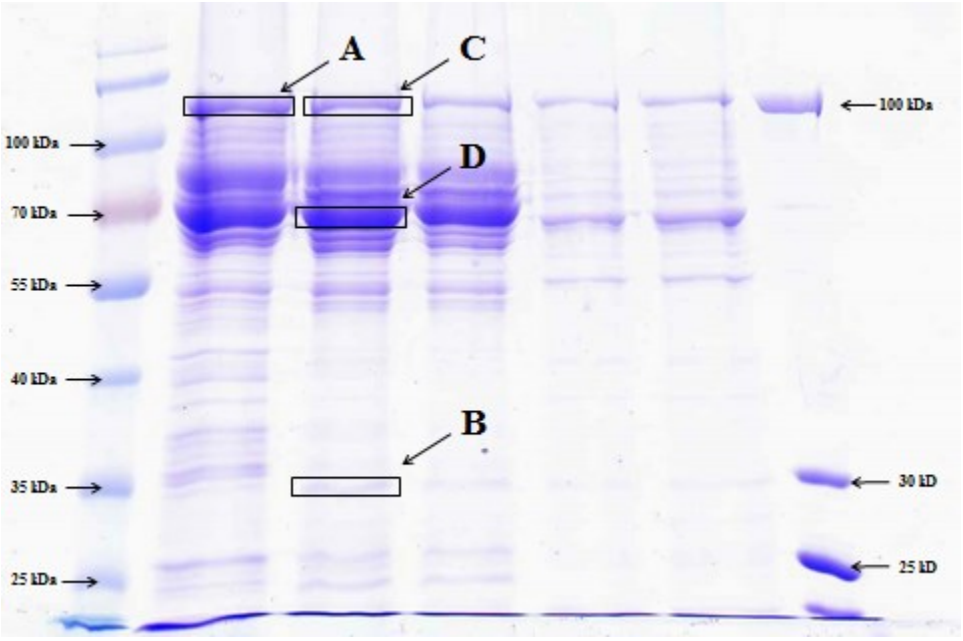


Figure 20. Gel submitted for Mass Spectroscopy analysis. Lane 1: molecular weight ladder/marker. Lane 2: undigested elution from the first round of the *Ni* column purification procedure. Lanes 3-6 are elution fractions, post digest (treated with rEK) from the 2nd round of Ni column purification. Lane 7 is an alternate low range molecular weight ladder used simply as a secondary reference [171].

Comparing lane 2 to 3 on this gel, there is a noticeable lightening of band C, as compared to band A, and the simultaneous appearance of band B, supporting the conclusion that the NusA-PEX fusion is digested by the rEK. The boxes outlined on the figure were positively identified by the Mass Spec analysis as: boxes A and C confirmed to have NusA-PEX fusion protein; box B confirmed to be the PEX domain; and box D is GroEL.

All Mass Spectroscopy analyses were performed in the University of Missouri Kansas City - School of Biological Sciences Proteomics and Mass Spectrometry core Facility

directed by Dr. Andrew Keightley [172]. MS analyses proceeded in two steps. The first step was *in-gel trypsin digestion*, and the second step was *nanoLC-High Resolution Mass Spectrometry (LC-MS)*.

In-gel trypsin digestion

Coomassie stained gel bands were first excised and dehydrated with 100µL acetonitrile. Protein disulfides were reduced by swelling the dehydrated gel with 50µL 50 mM ammonium bicarbonate buffer containing 100 mM DTT (30 minutes). A second dehydration with acetonitrile was performed and subsequently alkylation by re-swelling with 50uL buffer containing 100mM chloroacetamide (30 minutes). The gel with reduced and alkylated protein was dehydrated with acetonitrile and re-swollen with 50mM ammonium bicarbonate buffer (to affect buffer exchange, and dilution/removal of alkylating reagents), dehydrated again with acetonitrile and swollen in a second buffer (50mM H_4HCO_3 at pH7.8 + 20 µg/mL porcine modified Trypsin), and allowed to digest at 37^o C overnight [173].

Peptides were extracted twice with 15µL 70% acetonitrile: 30% water. Extracted peptides were dried down to ~2µL, and diluted in 20µL 0.1% Formic acid (aqueous) for further analysis.

NanoLC-High Resolution Mass Spectrometry (LC-MS) [174]

The extracted peptides were analyzed by capillary nanoflow LC-tandem MS using a 50µm I.D. X 12cm long capillary column packed with Phenomenex Jupiter C18 reversed

phase matrix. Peptides were resolved with a linear gradient of acetonitrile (2-40% acetonitrile over 90 minutes) at a flow rate of 250 nL/min (Eksigent binary LC system) and loaded in 2% acetonitrile, 0.1% Formic acid (aqueous phase). For MS data acquisition, a Q Exactive Plus mass spectrometer, was operated in data-dependent mode in which one mass spectrum (MS, 70,000 resolution) and fifteen dependent HCD spectra (MS2, 35,000 resolution) were acquired per cycle throughout the acquisition [175].

Protein identifications were made using Mascot protein identification software, searching against SwissProt_2017_01 (553,474 protein sequences) with decoy database included (reversed database) [176, 177]. Mass tolerances for the searches were 10 ppm precursor, and 0.01 Dalton fragment mass. Search results were filtered using 95% confidence (chances of no more than 1 in 20 that the match is random), and with a threshold of expected value equal to or below 0.05 imposed. Additional technical specifications of individual MS identifications are presented in **APPENDIX H**.

Ancillary experiments

With the presence of GroEL confirmed, the apparent relatively low yield of unassociated PEX-domain (as evidenced by the weakness of banding at Box 2 - **Figure 20**) could be explained by its association with the bacterial chaperonin protein [178]. While this situation certainly presents a difficulty in attempts at purification of the PEX domain, it does also provide evidence that the domain must be folding to a certain extent to remain bound to the chaperonin.

A high-salt addition purification experiment was run on the NusA-PEX domain fusion protein in an attempt to dislodge the PEX domain from the purported chaperonin (GroEL). All conditions and volumetrics are the same as presented for the previous experiments and outlined on page 73, with the exception that cells were re-suspended after pelleting in Lysis buffer containing 1.5 M NaCl instead of the 0.5 M NaCl used previously. 1.5 M NaCl was maintained in all buffers during subsequent purification attempts **Figure 21**.

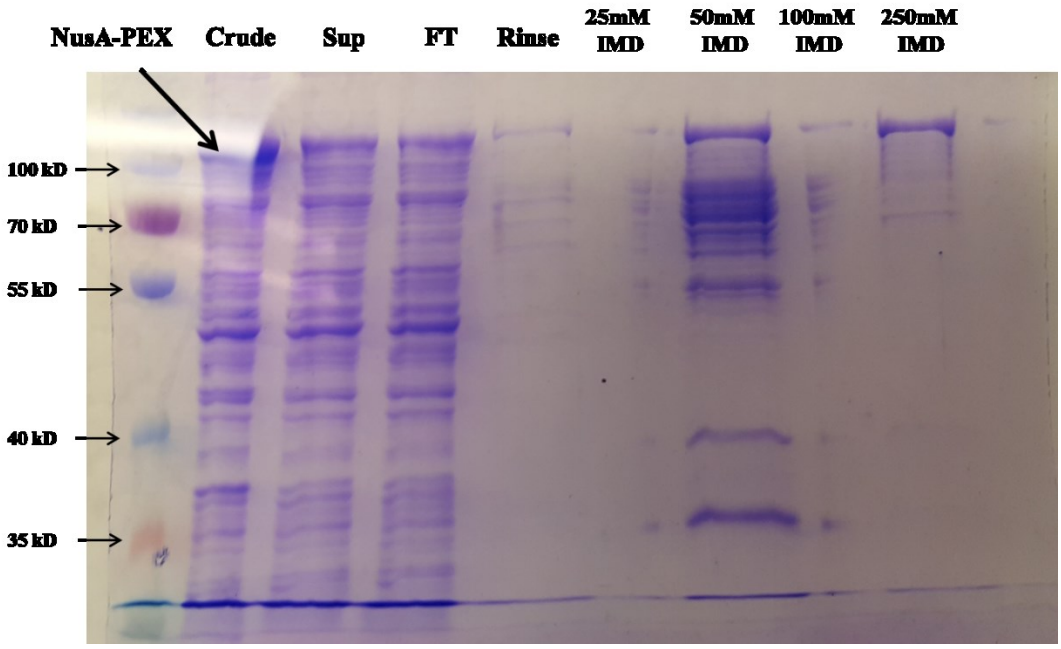


Figure 21. High-salt purification of NusA-PEX domain fusion protein. The addition of 1.5 M NaCl instead of the previously utilized 0.5 M NaCl proved to be more effective than any other experimental manipulation at increasing target protein yield. IMD = Imidazole.

Contrary to earlier protein expression test which determined that 25 mM imidazole would be best, after the addition of the high salt component the 250 mM imidazole concentration treatment produced considerably cleaner elution fractions. The 50 mM

imidazole did show good fusion protein production but the 250 mM concentration was chosen over it due to relative purity.

Fast Protein Liquid Chromatography (FPLC)

In an attempt to purify the minimal amount of post-digest PEX domain protein, the 250 mM imidazole fraction was buffer exchanged back to < 25 mM imidazole and digested with rEK, under the same conditions as outlined above, and an additional purification step using gel filtration via AKTA Fast Protein Liquid Chromatography (FPLC; Bouyain Lab) was performed [179]. 500 µl of protein solution was loaded onto a Superdex 200 10/300 GL prepacked column and eluted in HBS (HEPES buffered saline: 20 mM HEPES/150 mM NaCl), **Figure 22** [180, 181].

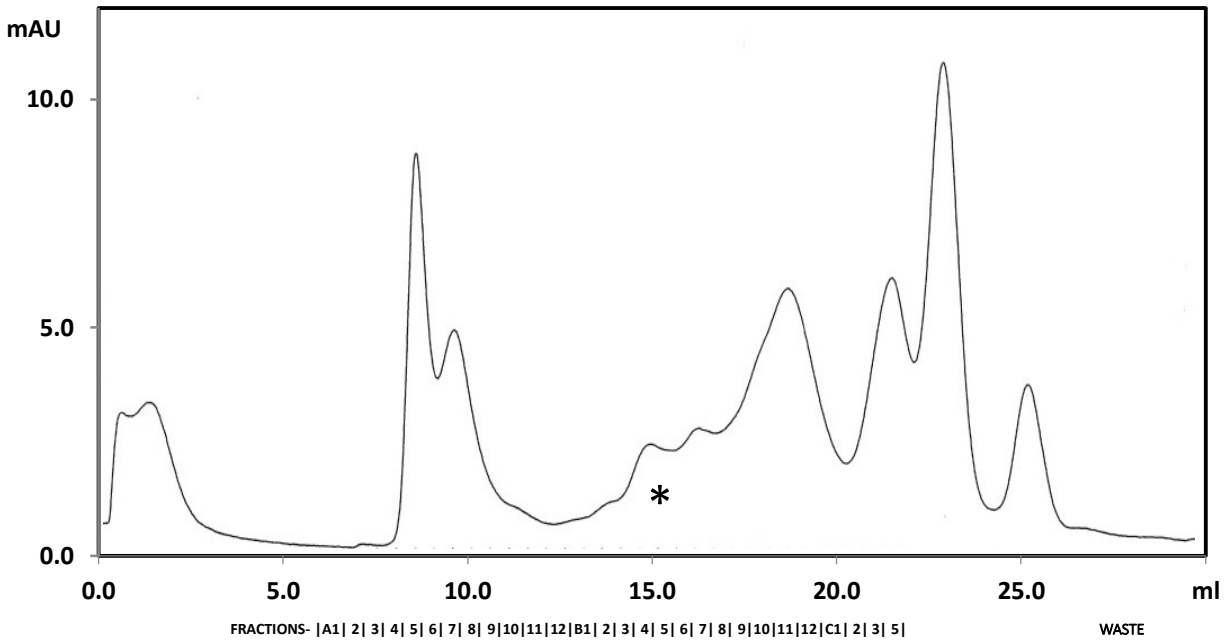


Figure 22. Results of FPLC analysis on the 250 mM imidazole fraction from the high-salt purification step. The asterisk shows the range of fractions (B2-B6) in which the target PEX domain should be eluted based on the molecular weight of approximately 31.51 kDa. mAU = milli absorption units [179].

After fractionation another separation gel was run, with fractions A2-A5; B2-B6 (this is the portion in which the target peptide was expected to be found); B10-11; and C1-C3, based on the obvious peaks associated with the FPLC output. However, concentrations were so low that protein was very weak. A follow-up attempt to highlight the weak banding through silver nitrate staining was unsuccessful [182, 183].

Some fraction of the target protein is soluble (evidenced in weak, but definitive presence of "pure" PEX), and future attempts to adjust the conditions of the expression

experiments (e.g. high-salt, various other additives, temperature changes, etc.) may yield more usable protein.

Homology Modeling of the PRG4-PEX Domain

The primary sequence of the PRG4-PEX domain peptide was submitted to multiple Homology Modeling portals to obtain a conjecture of the potential tertiary structure. And, while fully cognizant of the fact that these models are not experimentally determined structures, some very consistent and interesting data have emerged from the models that will allow further research with several intriguing hypotheses concerning the phylogenetic position of the PRG4-PEX domain in an evolutionary context **(Figures 23 and 24)**. There is also some additional speculation as to the relevance of these data to PEX-domain evolution presented in **Chapter 4**.

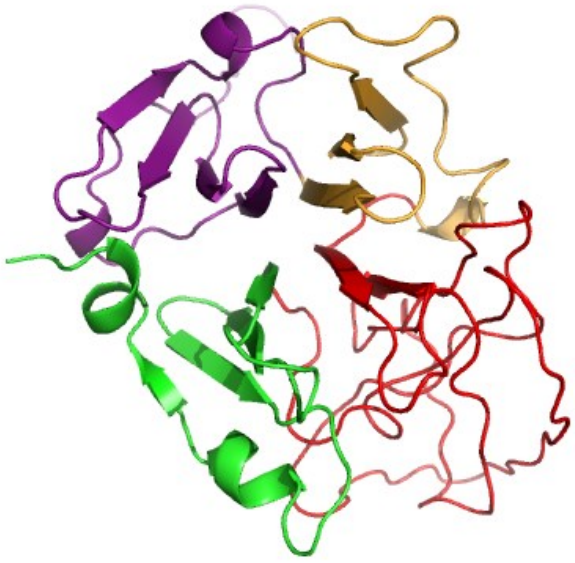


Figure 23. Homology model of the PRG4-PEX domain primary sequence, as returned by the Phyre2 web portal for protein modeling, prediction and analysis [184]. The three most statistically relevant threads were: 1gxdA (MMP2), 1qhuA (HPX N-domain), and 2jxyA (MMP12). β-blades are color coded 1=purple; 2= light orange; 3=red; 4=green. Note the potentially unstructured blade 3 (red).

Figure 24. Homology model of the PRG4-PEX domain primary sequence, as returned by the RaptorX web portal for protein modeling. The three most statistically relevant threads were: 2mqsA (MMP-14); 1genA (MMP-2); and 3ba0A (MMP-12). β-blades are color coded 1=purple; 2= light orange; 3=red; 4=green. Note undetermined structure of blade 3 (red), which is consistent with Figure 23 deduced by PHYRE2 [185-187].

If these models are confirmed by more robust experimental structural determination, then the function, if any, of the unusual (putatively unstructured) 3rd blade will be of interest. One possibility is that the lack of tertiary stability in some portions of the PRG4-PEX domain could be interpreted as a derived LOSS of structure. This in itself should prove interesting if the domain retained biophysical or biochemical functionality. As efforts continue to obtain a viable structure for this domain, experiments will also be conducted, in parallel, to determine its functionality.

HOMOLOGY MODELS: **Complete** *PRG4* protein **(Figures 25 and 26)**

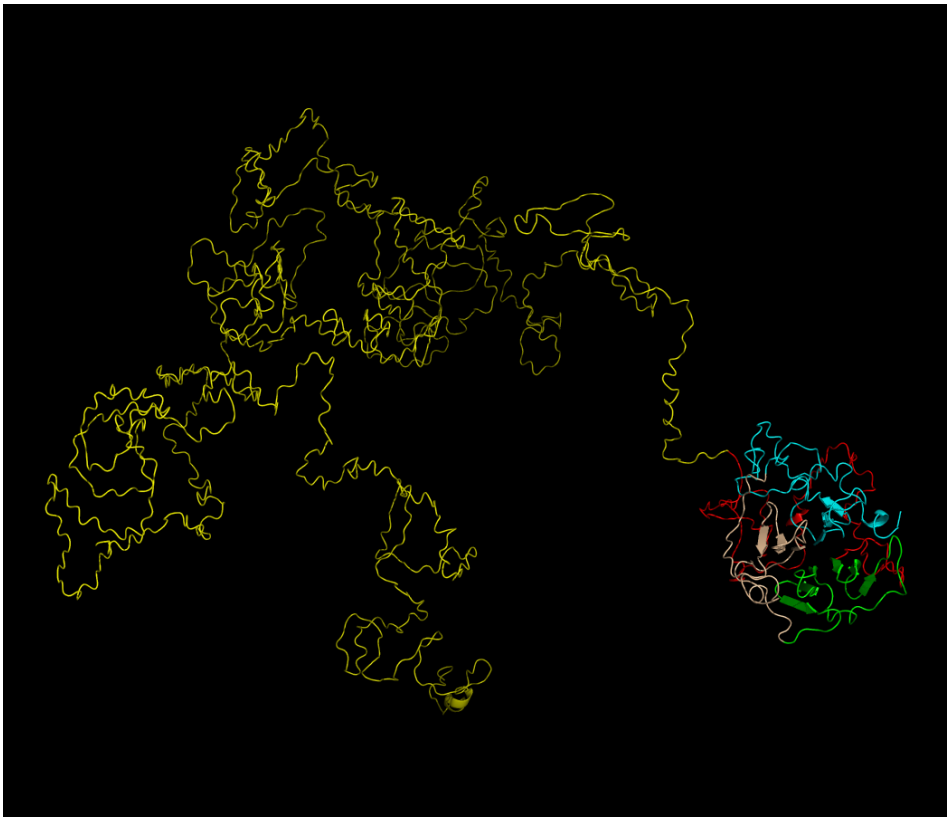


Figure 25. Homology model of the entire PRG4 protein primary sequence, as returned by the web portal I-TASSER for protein modeling, with lateral view of the structured PEX domain [74, 76, 188].

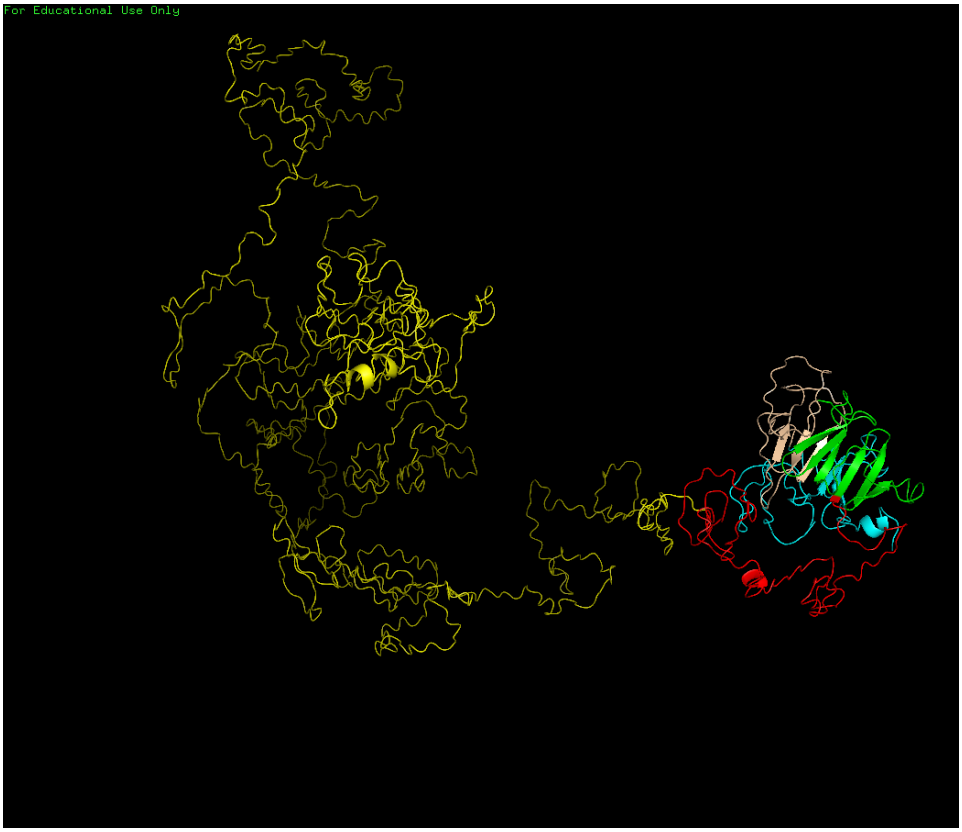


Figure 26. Homology model of the entire PRG4 protein primary sequence, as returned by the web portal I-TASSER for protein modeling, with tangential view of the PEX domain showing link to primary structure of the main "body" of the protein [74, 76, 188]. Note the attachment of the main body of the protein to the PEX domain is via a segment where structure cannot be determined at this time.

Future directions:

Even though sufficient amounts to determine a structure of the PRG4 PEX domain peptide have not been purified to date, using *E. coli* recombinant techniques, experiments will continue until enough protein is produced to attempt crystallization of the domain. There are several protocols under consideration including:

- Attempting the manipulation of experimental conditions to maximize yield of the soluble protein obtained from these preliminary attempts.

- Expression and subsequent isolation from insoluble fractions, refolding and purification [189].

- Attempting to express the protein in a eukaryotic cell system [190].

- Order the full length Lubricin protein and proteolytically cleave off the PEX domain for isolation and purification [191].

Eventually, once sufficient amounts of protein are produced, spectrometry using the School of Biological Sciences High Field NMR Laboratory core facility, located on the first floor of the Spencer Chemistry Building, will be attempted. The available NMR spectrometer is a Varian Inova 600 MHz model.

The data obtained from spectroscopy will be processed with available computer modeling programs NMRPipe and NMRView [192, 193]. Individual and backbone resonance assignments will be determined through the use of established NMR experimental procedures [142, 194].

CHAPTER 6

EVOLUTIONARY ANALYSES OF TARGET GENES IDENTIFIED AS POTENTIAL GENETIC MARKERS OF COMPLICATIONS ASSOCIATED WITH DIET-INDUCED OBESITY

Introduction

A growing body of evidence, collated from variable analyses, has suggested there are numerous discoverable genetic markers that may predispose patients to the onset and maintenance of several human eating disorders associated with clinical obesity.
Studies have been performed that have led to the parsing of differential effects of environmental versus genetic influences on growth statistics including Body Mass Index (BMI). A cross-sectional twin study, involving over 24,000 children from Australia, Europe and North America, found that variability in BMI (along with anatomical weight and height) was highly correlated, and influenced by genetics, across both sexes and throughout the age-span of the participants from 5 months through adolescence (19 yoa) [195].

In response to a defined North American "Obesity Epidemic", a long-term follow-up cohort study was performed in which parents and children who were originally assayed from 1973-1976 were again measured 25-30 years later from 1999-2004. The results suggest that several important parameters including increase in adiposity and BMI levels, were significantly tied to genetic factors [196, 197]. However, in another large sibling-based study, even though there was some evidence for differences in the heritable effects of birth weight in the adolescent years based on sex of the child, there may also be differential effects of intrauterine environmental conditions that remain to be explained [198].

Empirical evidence is mounting on the significant impacts of a variety of genetic modalities on heritable obesity pathologies, including monogenetic, polygenetic and mutational influences. Genetic disruption of energy balance of the leptin/melanocortin pathway and compromised differentiation of the paraventricular nucleus in neurons have been shown to involve defects in at least eight genes with the potential of leading to the development of monogenetic obesity in humans [199]. Studies on inherited polygenic susceptibility to obesity have implicated disruption of normal body weight regulation of CNS function[200]. In whole-exome sequencing analyses within families with a pedigree for two relatively common and severe eating disorders (EDs), *anorexia nervosa* and *bulimia nervosa*, the pathologies were shown to be correlated with rare missense mutations in both the histone deacetylase 4 (HDAC4) gene and the estrogen-related receptor α (ESRRA) gene[201]. Direct secretory biochemical involvement in the development of various obesity syndromes is also well documented. Abnormal levels of dopamine, norepinephrine and serotonin have been noted in clinical experiments with people suffering from a wide spectrum of eating disorder related obesities [201].

The protein product of the *vgf* (non-acronymic) gene is a secreted neuropeptide precursor shown to be synthesized and proteolytically modified in mammalian neurons [202]. A multitude of neuroendocrine roles have been empirically shown to be associated with several specific peptide derivatives of the VGF precursor protein, including TLQP-21 [203-205]. Work with this peptide has uncovered links between metabolic energy budgets and food intake rate in mouse [206-208]. The TLQP-21 peptide is a downstream derivative of its larger TLQP-62 precursor which elicits a dose-dependent increase in energy expenditure in rats [203, 204].

Several other VGF derived peptides including NERP-1, NERP-2 and NERP-4 have been shown to be bioactive, but their respective functions have not been fully elucidated [202, 205, 209].

Experiments with intraventricular infusions of TLQP-21 in Siberian hamsters induced weight loss attributed to decreased food consumption. That same series of experiments also showed that VGF deficient mice exhibited shorter circadian periods, suggesting these secretions play a role in influencing superchiasmatic regulation of food intake patterns [210]. The TLQP-21 peptide has also been correlated with increases in UCP-1 (Uncoupled Protein - 1), a white adipose fat tissue secretion, and concomitantly shown to mitigate some negative effects during high-fat diet experiments with mice [205, 211, 212].

Recent work has used genetic information obtained from the Gene Expression Omnibus (GEO), a public access genomic information repository coordinated and maintained through the National Center for Biotechnology Information (NCBI) [129, 213]. The experiment involved using the GEO subroutine algorithm GEO2R that can be programed to search for expression data based on predesignated correlation parameters [214]. The initial screen in this experiment targeted several genes, including the VGF gene, which had been previously identified as being involved in the regulation of several metabolic processes broadly linked to eating disorders, and more specifically to Night Eating Syndrome (NES). This particular condition is characterized by behaviors that involve eating substantial amounts of post evening meal calories (>25%) either before the patient can fall asleep, or waking to consume the additional food shortly after falling asleep and the inability to return to sleep until the late-night eating episode has occurred [215]. While NES is not manifested in all cases, it is seen in a substantial number of patients with diagnosed clinical obesity, and it has been proposed

that the condition is potentially driven by genetically mediated breakdown in normal metabolic, circadian or neuronal control [216].

The genetic screens were performed under the assumption that metabolic as well as neuronal factors might influence the development of NES [213]. A second level of identification involved parallel screening between multiple gene libraries to determine if any genes registered as significant hits across datasets [213]. This analysis resulted in the identification of 6 genes (in addition to VGF) that tested statistically significant for the original cross correlates[213]. Those 6 genes were: ATPase 1 α-3 (Atp1a3); B930041F14Rik (a murine gene with a human ortholog: FNDC10 - FibroNectin type-III Domain-Containing transmembrane protein); Deformed Epidermal Auto-regulatory Factor 1 (DEAF1); Insulin like Growth Factor Binding Protein 2 (IGFBP2); RNA binding motif protein 3 (Rbm3); and X-box binding protein 1 (Xbp1). This work is a follow-up evolutionary analysis of these 6 gene candidates that are implicated in mechanisms associated with NES.

Genes of interest

During a broad screening search of the Gene Expression Omnibus (GEO) for potential candidate genes associated with a human eating disorder (Night Eating Syndrome - NES), in addition to the primary hits, there were six genes that showed significance across more than one dataset in relation to the established search parameters of involvement in neural development, plus, sleeping/eating disorders [213].

Atp1α3 - ATPase Na^+/K^+ transporting subunit alpha 3 [*Homo sapiens* (human)]

The product of this gene is the alpha-3 component of a two subunit, integral membrane enzyme belonging to the P-type cation transporter ATPases, sub-family Na^+/K^+ - ATPases. Its primary function is to establish and maintain cross-membrane electrochemical gradients of sodium and potassium ions, necessary for electro-excitability of muscle and nerve tissues, sodium coupled molecular transport, and osmoregulation. Malfunctions in this protein have been associated with a suite of human disorders including hyper-excitability of the central nervous system (CNS) and alternating hemiplegia of childhood [217, 218].

DEAF1 - Deformed Epidermal Autoregulatory Factor 1 - a.k.a. - suppressin [*H. sapiens*]

This gene encodes a zinc finger transcription factor. The protein binds to several secondary target genes as well as to its own promoter. It is known to downregulate transcription of the *HNRPA2B1* (Heterogeneous Nuclear Ribonucleoprotein A2/B1) gene [219]. It also binds RARE (the retinoic acid response element), and activates both *PENK* (the preproenkephalin gene) and EIF4G3 (Eukaryotic translation initiation factor 4 gamma 3) [220-222]. In mouse models, defects in availability of DEAF1 were shown to compromise embryogenesis by disrupting skeletal patterning and neural tube closure [223]. Mutations in the SAND (**S**p100, **A**IRE-1, **N**ucP41/75, **D**EAF-1) domain of the *DEAF1* gene have recently been shown to cause a form of an autosomal dominant condition marked by aberrant behavioral patterns, severe speech impairment, and intellectual disability [224]. Additionally,

DEAF1 has been implicated in several pathologies including colorectal cancer, and perhaps most significantly for the purposes of these analyses, type-1 diabetes [225, 226].

FNDC10 - fibronectin type III domain containing 10 [*H. sapiens*]

In the original screen, this hit was identified as a murine gene from *Mus musculus* (house mouse) designated *B930041F14* [213]. It is also known as *C1orf233*, and the human ortholog designation is *FNDC10*. The gene codes for a 226 aa trans-membrane protein. There is no primary literature associated with the functionality of this protein.

IGFBP2 - insulin-like growth factor binding protein 2 [*H. sapiens*]

This gene encodes a protein with several known transcript variants that can either be secreted to bind insulin-like growth factors I and II (IGF-I and IGF-II) in the bloodstream with high affinity or, it can persist intracellularly where it interacts with a variety of differing ligands [227]. It has been shown to be a player in a variety of human pathologies including kidney disease, ventricular stroke, Pallister-Killian Syndrome (PKS), and several cancers including esophageal, breast and colorectal [228-233]. It has also, on a positive note, recently been implicated in the potential moderation of dementia-relevant, age-specific deterioration of cognitive function [234].

Rbm3 - RNA binding motif (RNP1, RRM) protein 3 [*H. sapiens*]

The protein derived from this gene is a member of a family of proteins containing at least one RNA recognition motif domain (RRM). Translation is stimulated via low oxygen tension and/or cold shock [235-237]. There are several splice variants of this gene that lead to the translation of multiple isoforms, and these have been differentially categorized as having myriad effects on human disease progression both positively and negatively [238]. For example, measures of decreased expression have been shown to correlate with tumor growth and poor disease prognosis in patients suffering from urothelial bladder cancer, and increases in nuclear expression have been connected to improvements in prognosis with colorectal cancer [239, 240]. This gene is an X chromosome associate and is known to have a pseudogene copy located on human chromosome 1.

Xbp1 – X-box binding protein 1 [*H. sapiens*]

The functional gene is located on Chromosome 22 in humans and codes for a transcription factor in Basic Leucine Zipper Domain (bZIP) protein family [241, 242]. There exists also a known pseudogene located on chromosome 5[241]. The name refers to the property of the protein binding to a particular promoter sequence known as an X-box [243]. A primary function is the regulation of the Unfolded Protein Response (UPR) during stress events in the endoplasmic reticulum (ER) [244, 245]. The obligatory need for proper expression of this protein during embryogenesis in mouse has been demonstrated, particularly with respect effectual development of heart, liver, endocrine pancreas, salivary gland and secretory tissues [246].

The human XBP1 protein is known to be expressed as two isoforms designated as XBP1 spliced (XBP1s), and XBP1 unspliced (XBP1u) [247]. The two isoforms are differentially but significantly involved in numerous human physiological processes including adipogenesis and lipogenesis [248, 249]. There are also numerous human pathologies in which this protein has been shown to play a crucial role, such as atherosclerosis, ischemia and a variety of cancers, including, but not limited to breast, multiple myeloma (MM), and osteosarcoma [250-253]. Conversely, there is also modeling data to suggest that the functioning of XBP1s during ER stress provides some protection against β-amyloid neurotoxicity in the progression of Alzheimer's Disease (AD) [254]. Additionally, and more significantly for the purposes of this work, mice with an induced deficiency in XBP1 protein became insulin resistant and developed type-2 diabetes, apparently as a function of obesity-linked ER stress [255].

Methods and Results

Evolutionary analyses

Sequences for all available human variants of each of the target genes, and one sequence each from *Macaca,* were accessed through the NIH-NCBI public genomics database repository (henceforth simply: NCBI) [32]. Coding sequences for all genes used in the analyses were verified using the BLAT function of the University of California - Santa Cruz Genome Browser [256, 257].

To assess potential functional consequences of mutation at the protein level, known nucleotide polymorphisms for each of the human target genes were obtained from the *dbSNP* portal of the NCBI, which catalogs research-identified single nucleotide polymorphisms (SNPs)[47]. Each of the genes was subsequently screened using the *SNPNexus* portal of the Barts Cancer Institute at Queen Mary University of London [258]. Individual SNPs are identified using a standard protocol known as "*rs*" numbers in the NCBI *dbSNP* database. The *rs*-numbers were obtained and copied into the analysis window at the *SNPNexus* site and screened through two *in silico* programs *SIFT* and *PolyPhen-2* [259, 260]. Both programs operate through algorithms that are designed to predict the potential functional impact of specific missense amino acid substitutions and score them numerically. The *SIFT* method score the substitutions on a mathematically standardized scale from 0 to 1. If a substitution scores ≤ 0.05 it is considered "damaging". Any score > 0.05 is deemed "tolerated". In the *PolyPhen-2* program, scores are standardized to a range of 0 to 1 but are differentially designated as "probably damaging", "possibly damaging" or "benign". In *PolyPhen-2*, the higher the score, the more potentially damaging is any particular substitution. It is generally agreed in the field, that for a substitution to be considered deleterious with a high level of confidence, it should be predicted as potentially, functionally damaging by at least two independent methods **(Figure 27)**.

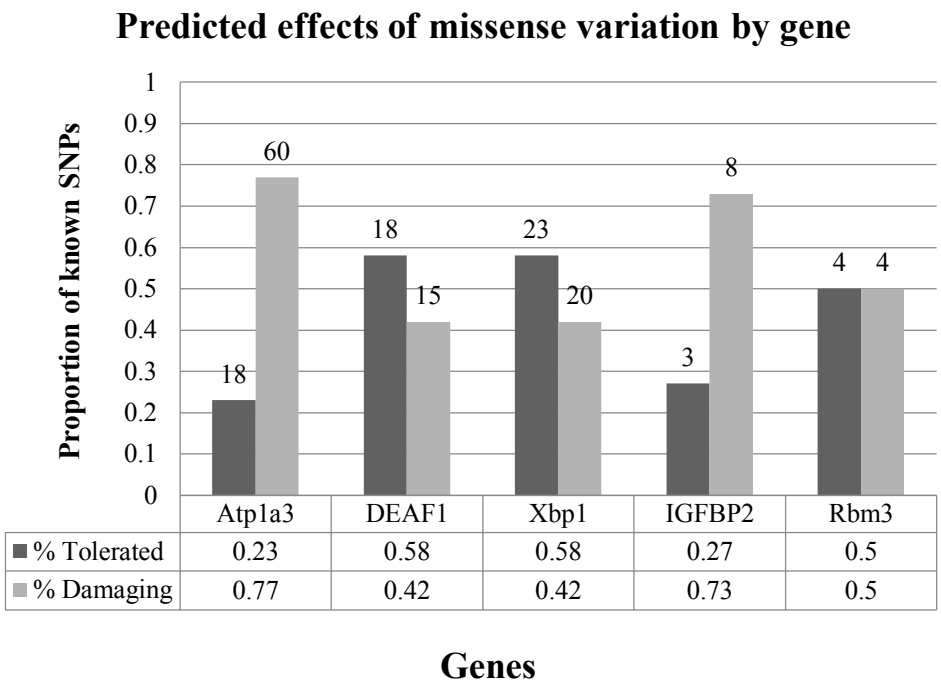


Figure 27. Proportional number of missense (nonsynonymous) mutations effects on the potential functionality of the resulting protein products of each target gene. Categories are defined as either "Tolerated" or "Damaging" when both SIFT and PolyPhen-2 prediction algorithms concur as to the functional consequences of any specific SNP. Ambiguous mutations, and category "unknown", are not represented on this graph.

Atp1a3- There are a total of 78 SNPs currently reported for this gene. SIFT/PolyPhen combined identified 60 (77%) of these as being *damaging*, and 18 (23%) as *tolerated*. This result is consistent with high levels of evolutionary constraint.

DEAF1 - There are a total of 33 SNPs currently reported for this gene. SIFT/PolyPhen combined identified 15 (42%) of these as being *damaging*, and 18 (58%) as *tolerated*. This can be interpreted as this protein having functional plasticity that potentially allows for some level of polymorphic selection.

IGFBP2 - There are a total of 11 SNPs currently reported for this gene. SIFT/PolyPhen combined results identified 8 (73%) of these as being *damaging*, and 3 (27%) as *tolerated*. This suggests constraint, but the small sample size is a concern for proper interpretation.

Xbp1 - There are a total of 43 SNPs currently reported for this gene. SIFT/PolyPhen combined identified 20 (46.5%) of these as being *damaging*, and 23 (53.5%) as *tolerated*. Similar to *DEAF1* this pattern can be interpreted as this protein having functional plasticity that potentially allows for some level of polymorphic selection.

Rbm3 - There are a total of 8 SNPs currently reported for this gene. SIFT/PolyPhen combined identified 4 (50%) of these as being *damaging*, and 4 (50%) as *tolerated*. The even distribution and limited information on this gene render any broad-scale assumptions suspect.

Statistical analyses

To explore comparative evolutionary rates of each of the target genes, the McDonald-Kreitman Test (MKT) was run on each to compare synonymous vs. nonsynonymous amino acid substitutions within human genes, and between humans (*Homo sapiens*) and the Rhesus macaque (*Macaca mullata*) [261]. The Rhesus macaque was chosen for the comparisons because it is a relatively close primate lineage to *Homo*, but with sufficient time since divergence to show a robust number of between-species fixed genetic differences. Additionally, *M. mullata* has a substantial amount of genomic data available for comparisons.

The basal within-species variation for the synonymous and nonsynonymous changes (missense mutations) for the human genes were attained through the *dbSNP* search protocol of the NCBI [32]. They were subsequently validated through the 1000 Genomes Project [48]. The statistical significances of the MKT were evaluated using the Fisher's Exact Test (FET) for a 2x2 contingency table. The between-species comparisons (*Homo x Macaca*), involved aligning the coding nucleotide sequences of both using the MEGA6 package's *ClustalW* function and transferring the alignments to the MKT program for statistical analyses [261-263]. Probability values (p) were considered statistically significant with $\alpha < 0.05$, **Table 10**.

Table 10: Results of MKTs. Statistical significance was assessed by Fisher's Exact Test for a 2x2 contingency table.

NI[a]	*p*-value	Variation	Gene	Nonsynonymous	Synonymous	Total	Interpretation
			Atp1a3				VERY strong constraint
		Within Species Variation [b]		19	10	29	
9.393	$p < 0.001$*	Fixed species differences [c]		20.10	99.38	119.48	
		Total		39.1	109.38	148.48	
			DEAF1				Strong positive selection
		Within Species Variation [b]		25	10	35	
5.973	$p = 0.014$*	Fixed species differences [c]		37.92	3.01	62.92	
		Total		62.92	13.01	75.93	
			FNDC10				Neutral
		Within Species Variation [b]		1	0	1	
NULL	$p = 0.536$	Fixed species differences [c]		84.06	32.20	116.26	
		Total		85.06	32.20	117.26	
			IGFBP2				Weak positive selection?
		Within Species Variation [b]		45	24	69	
0.583	$p = 0.087$	Fixed species differences [c]		113.95	35.44	158.95	
		Total		158.95	59.44	218.39	
			Rbm3				Constraint?
		Within Species Variation [b]		84	22	106	
NULL	$p = 0.006$*	Fixed species differences [c]		0.00	2.04	2.04	
		Total		84	24.04	108.04	
			Xbp1				Neutral
		Within Species Variation [b]		7	2	9	
2.045	$p = 0.400$	Fixed species differences [c]		26.34	15.39	41.73	
		Total		33.34	17.39	50.73	

[a] Neutrality Index (NI) as computed with Jukes & Cantor correction for divergence; [b] all known mutations within humans;
[c] values converted under JC69 model; * statistically significant at $\alpha < 0.05$.

Phylogenetic analyses

Atp1a3

The evolutionary relationships of the Atp1a3 protein within selected primates, rooted by a murine clade, was inferred by using the Maximum Likelihood method based on the JTT matrix-based model [82]. The bootstrap consensus tree inferred from 500 replicates is taken to represent the evolutionary history of the taxa analyzed [83]. Branches corresponding to partitions reproduced in less than 50% bootstrap replicates are collapsed. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (500 replicates) are shown next to the branches [83]. Initial tree(s) for the heuristic search were obtained automatically by applying Neighbor-Join and BioNJ algorithms to a matrix of pairwise distances estimated using a JTT model, and selecting the topology with superior log likelihood value. The analysis involved 13 amino acid sequences. All positions containing gaps and missing data were eliminated. There were a total of 968 potential amino acid positions in the final dataset. Evolutionary analyses were conducted in MEGA7 [81]. Below is the best supported phylogeny for Atp1a3 **(Figure 28)**.

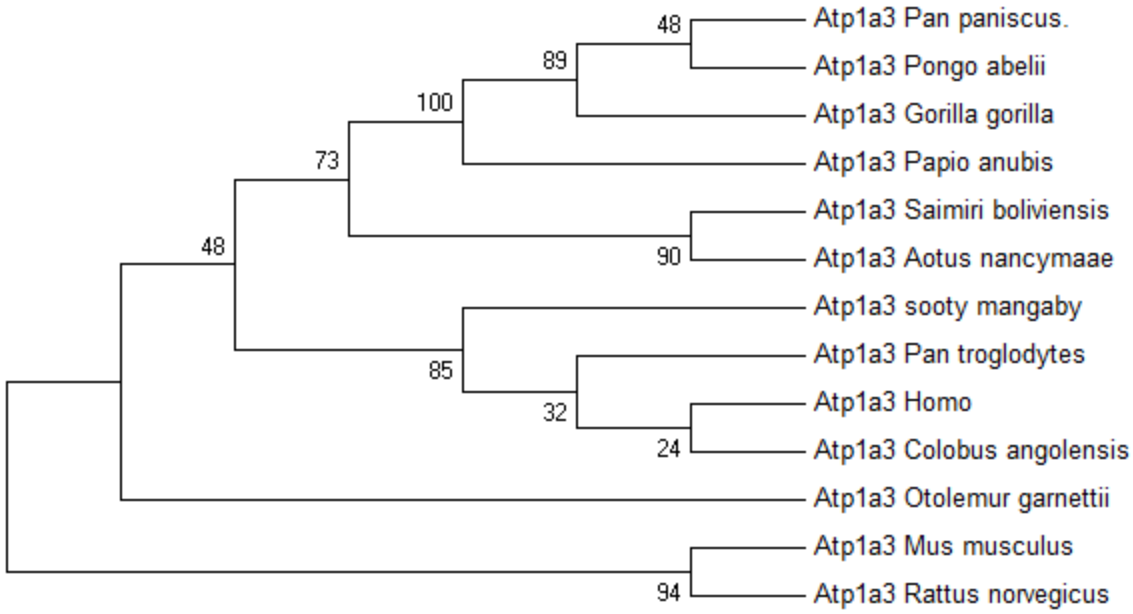


Figure 28. Molecular Phylogenetic analysis by Maximum Likelihood method for Atp1a3 protein.

The placement of *Homo sapiens* on this phylogeny was at first confusing, in that humans would normally be expected to fall in a tight clade with common chimp, bonobo, gorilla and orangutan, the "classic" great apes. However, this result places *P. troglodytes* and *H. sapiens* outside that group and clades them with the sooty mangabey, and *Colobus* (both old world monkeys), and clades bonobo, gorilla, orangutan and *Papio anubis* (the Olive baboon) together, with the baboon being somewhat of an unexpected member there. However, upon detailed investigation of the alignment of the proteins it is very obvious that these clades are real, as also evidenced by the high statistical support for the root branches of both clades. The two clades diverge from one another in several locations along the length of the alignment but the pattern between clades is remarkably consistent, **Table A5, APPENDIX I**.

One hypothesis based on this observation is that this may be an example of a multiple domain level functional convergence in the operations of this otherwise highly constrained protein that has led to a phylogenic "split" in the great apes. While this result seems unconventional, there is a well-documented level of incongruence, or mismatch, between established species-trees and more contemporarily deduced gene-trees to the point that evolutionary relationships within the Hominoidea (apes) still remains somewhat contentious [264-266]. Although, most cladistic analyses done on genomic sequences support the traditional morphological relationships within the clade (*Pan* and *Homo* as sisters, *Gorilla* as outgroup); there are known independent DNA sequence data sets that resolve the clade differentially. Two such sets support a *Pan- Gorilla* clade and one suggests a *Homo-Gorilla* clade [264]. One major significance of this is to lend credence to the phylogeny presented here, as a function of a single gene product (Atp1a3) analysis. This observation will be a major focus of future evolutionary and functional analyses of this specific protein.

DEAF1

The evolutionary history of the DEAF1 protein was inferred by using the Maximum Likelihood method based on the JTT matrix-based model. The bootstrap consensus tree inferred from 500 replicates is taken to represent the evolutionary history of the taxa analyzed [82]. Branches corresponding to partitions reproduced in less than 50% bootstrap replicates are collapsed. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (500 replicates) are shown next to the branches [82]. Initial tree(s) for the heuristic search were obtained automatically by applying Neighbor-Join

and BioNJ algorithms to a matrix of pairwise distances estimated using a JTT model, and selecting the topology with superior log likelihood value. The analysis involved 13 amino acid sequences. All positions containing gaps and missing data were eliminated. There were a total of 176 potential amino acid positions in the final dataset. Evolutionary analyses were conducted using MEGA7 [81].

Orangutan (*Pongo*) was excluded from the DEAF1 analysis due to lack of a full length predicted protein in the database [33]. Not much of scientific value can be inferred from this phylogeny due to the lack of adequate statistical support for several nodes. This is certainly a consequence of the lack of robust information contained currently in the database pertaining to the DEAF1 protein **(Figure 29)**.

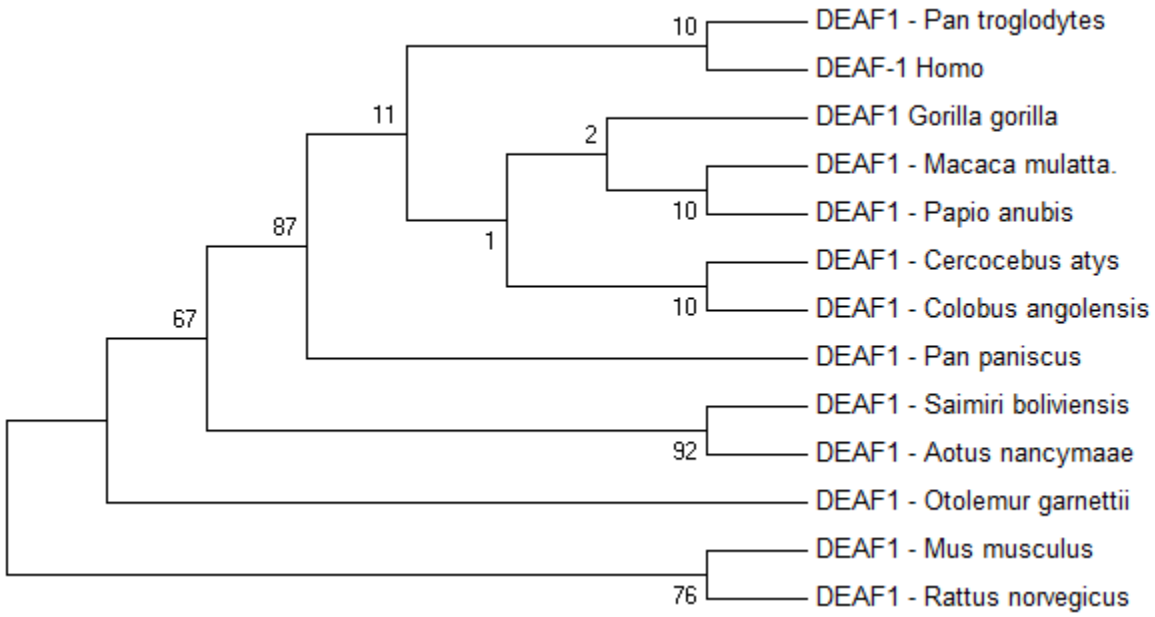


Figure 29. Molecular Phylogenetic analysis of the DEAF1 protein by Maximum Likelihood method.

Although not able to follow up with more in-depth evolutionary analysis, at this time, a puzzling contradiction of the results obtained regarding IGFBP2 is that it falls out in the MKT as being potentially under positive selection (albeit weak) but the SNP analyses suggests a high level of functional intolerance of mutations, which would more typically be associated with evolutionary constraint. However, this may simply be an artifact of there being so few known SNPs (11) for this gene. Again, obviously, more information is needed on the true level of polymorphisms of this gene on a larger population basis to be able to reasonably pursue this observation

Unfortunately, the remaining 4 proteins of interest: FNDC1, IGFBP2, Rbm3, and Xbp1, fall into the same category as DEAF1 in that there is not enough reliable sequence information currently available in the databases to allow for robust phylogenetic analyses. These proteins, therefore will remain a part of future pursuits as regards these principles of association, and await further research, updates and bioinformatic annotations.

Discussion

Six genes were identified in previous work that mined the Gene Expression Omnibus (GEO) database to search for genes that correlated highly with specific input parameters associated with disruption of metabolic processes implicated in, broadly, "eating disorders"; and specifically, with "Night Eating Syndrome" (NES) [129, 213-215]. This survey identified one specific gene (VGF) as being highly correlated with the search parameters. However, also recognized in those analyses there were an additional six genes identified across and between several differing gene libraries, that while not directly implicated in NES, did show high

levels of association with the metabolic and neuronal pathways associated with eating disorders in general [213]. These six genes Atp1a3, C1orf233, DEAF1, IGFBP2, Rmb3, and XBP1 were earmarked for further evolutionary analyses.

As a result of the combined evolutionary and phylogenetic analyses focus has been applied to current and continued efforts on the three genes that appear to be most consequential from an evolutionary stand point: Atp1a3 for apparently being under very strong selective constraint, DEAF1 for its high indication of positive selection, and IGFBP2 for its apparent lack of tolerance to non-synonymous substitutions. Although the remaining genes, for analytical and/or sparse information reasons, are excluded in this current pursuit, deeper evolutionary investigation may be warranted in the future to help establish values of background selection potentially operating on the entire suite of genes associated with evolution driven by changes in human shifts in dietary behavior.

As with the previous study focusing on VGF, it is assumed that eating disorders that lead to obesity are most certainly polygenic conditions. It is therefore important to attempt to understand each of these genes individually, as well as to try and determine if there are any deeper order interactions that may be a consequence of combinations involving allelic variations. One method for attempting to address this complexity would be to set up comparisons between clinically obese individuals that do not exhibit NES with those that are obese and engage in NES behaviors. This could potentially lead to insights regarding interactions and synergisms between and within the various allelic combination of the subject's genomic compliment of SNPs. Although, again, as suggested in the previous work, the "treatment" categories of such an experimental setup would pose significant problems

because behavioral patterns, especially regarding night eating habits, may be very difficult to parse and control for [213].

However, these types of analyses would be of value in potentially establishing emergent relationships between the various genes in question. As a case in point, transcription and expression of VGF in the suprachiasmatic nucleus (SCN) is apparently regulated by circadian modulation [267]. If it could be established that any of these other gene products are under similar regulatory control, then it may be possible to begin to link the combined effects of the various individual players to the phenotypes in question, in this case behaviors associated with eating disorders and the onset of obesity related pathologies.

An underlying working hypothesis for the pursuit of these genetic screens and subsequent evolutionary analyses is that these designated gene candidates are all in some way connected to metabolic modulation and so will potentially negatively influence the metabolic pathways that control "normal" eating cycles, and consequentially lead to weight gain and an eventual obese condition. Deep analysis of patients who can be shown to carry multiple damaging polymorphisms across the entire suite of these implicated genes should give us insights in the genetic foundation of the propensity of these individuals to engage in aberrant eating behaviors.

As with the VGF study there are several cautionary issues to be considered in any analyses of these types going forward. First, the knowledge that these various polymorphisms exist in the target genes must be considered when screening participants for experimental cohorts. The genomes of any individuals considered for control groups must be evaluated to ensure the absence of the allelic mutations. And, as previously noted, this will present a

particularly challenging set of criteria for obesity related diseases due to the probable high level of concurrence within many potential research populations [213, 268].

APPENDIX A

SUPPLEMENTAL INFORMATION ON 4-BLADED β-PROPELLER DOMAIN CONTAINING PROTEINS

Table A1. List of proteins used to describe forms of known multiple β-blade domains for which structures exist to accompany **Figure 1**.

NAME OF PROTEIN	NUMBER OF BLADES	FOUND IN (organisms)	FUNCTION OF DOMAIN
Hemopexin, MMPs, PRG4, VTN, WAPs	4	Bacteria; Eukarya	Variable (see text)
Tachylectin-2	5	Japanese horseshoe crab	Innate immunity: glycol-recognition of pathogens [269]
1-RWL (Sensor domain of the receptor Ser/Thr protein kinase)	6	*Mycobacterium tuberculosis*	Transmembrane receptor Ser/Thr protein kinases [270]
RACK-1 (Receptor of activated protein C kinase 1)	7	Prokaryotes; Eukarya	The Receptor for Activated C Kinase 1 [271]
4ZOX (Ribosome assembly protein SQT1)	8	*Saccharomyces cerevisiae*	Ribosomal assembly Protein [272]
3F6K (Sortilin)	10	*Homo*	Protein transport, primarily neuronal [273]

Table A2. Comparison of sequence identity of human hemopexin with other species represented in the NCBI database.

SPECIES	(%) Identity of protein primary amino acid sequence	(%) Identity of DNA nucleotide sequences
Homo sapiens	100.0	100.0
Common Chimp	99.8	99.7
Dog	81.9	85.9
Cow	73.8	81.1
Rat	77.2	82.6
Mouse	75.9	81.9
Chicken	58.3	62.8
Zebrafish	38.1	50.0

APPENDIX B

LIST OF THE HUMAN MMPs INCLUDED IN THE PHYLOGENIES ALONG WITH A BRIEF DESCRIPTION OF KNOWN FUNCTIONALITY

Table A3. Detailed descriptions of human Matrix Metalloproteinases (MMPs) used in phylogenetic analyses.

Protein (alias)	**Primary known functions**	**Extra/inter cellular**	**Conserved in:** (non-exhaustive)
MMP1 (interstitial collagenase)	Breakdown of interstitial collagens I,II,III,VII, and X, and viral Tat Protein	secreted	Chimpanzee, Bonobo, Rhesus monkey, dog, cow, mouse, rat, chicken
MMP2 (gelatinase A)	Vascular remodeling, angiogenesis, tissue repair, tumor invasion, inflammation, atherosclerotic plaque rupture. Degradation of extracellular matrix proteins and several non-matrix proteins, cleaves KISS at Gly-|-Leu bond. Role in myocardial cell death, cleaves GSK3β, formation of fibrovascular tissues with MMP14, anti-angiogenic and anti-tumor properties, inhibits cell migration and cell adhesion to FGF2 and Vitronectin. Ligand for integrin/β-3 on blood vessel surface.	Intracellular and secreted	Chimpanzee, Bonobo, Rhesus monkey, dog, cow, mouse, rat, chicken, zebrafish, and frog
MMP2 (Isoform)	Mediates proteo-CHUK/IKKA, activation of NF-kappaβ, NFAT, and IRF transcriptional Pathways	not known (N/K)	Chimpanzee, Bonobo, Rhesus monkey, dog, cow, mouse, rat, chicken, zebrafish, and frog
MMP3 (stromelysin-1)	Degradation of fibronectin, collagens III, IV, IX, X, gelatins I,III,IV, and V, and cartilage proteoglycans. Activates pro-collagenase	N/K	Chimpanzee, Bonobo, Rhesus monkey, dog, cow, mouse, rat, chicken, and frog
MMP8 (neutrophil Collagenase)	Degradation of collagens I, II, and III.	N/K	Chimpanzee, Bonobo, Rhesus monkey, dog, cow, mouse, rat and zebrafish.
MMP9 (gelatinase B)	Implicated in local proteolysis of the ECM; leukocyte migration; resorption of bone osteoclasts. Cleavage of KiSS1 at a Gly-|-Leu bond. Cleavage of type IV and type V collagen into 3 C-terminal quarter fragments and 1 shorter N-terminal quarter fragment; degradation of fibronectin (but not: laminin or Pz-peptide)	secreted	Chimpanzee, Bonobo, Rhesus monkey, dog, cow, mouse, rat, chicken, zebrafish, and frog

MMP10 (stromelysin-2)	Degradation of gelatins type I, III, IV, and V; and weakly collagens III, IV, and V. Activates pro-collagenase.	N/K	chimpanzee, Bonobo, Rhesus monkey, mouse, rat and chicken
MMP11 (stromelysin-3)	Cleaves alpha 1-proteinase inhibitor, degrades structural extracellular matrix proteins.	secreted	Chimpanzee, Bonobo, Rhesus monkey, dog, cow, mouse, rat, chicken, zebrafish, and frog
MMP12 (macrophage elastase)	Degradation of soluble and insoluble elastin. Cleavages at 14-Ala-|-Leu-15 and 16-Tyr-|-Leu-17 in the B chain of insulin. Accepts large and small amino acids with preference for leucine at P1' site. Aromatic and hydrophobic residues preferred at the P1 site, small hydrophobic residues (alanine) occupying P3.	N/K	Chimpanzee, Bonobo, Rhesus monkey, dog, cow, mouse, and rat
MMP13 (collagenase 3)	Substrates include Col I, II, III, IV, IX, X, XIV, gelatin	secreted	
MMP14 (transmembrane) MT-MMP 1	Cleaves PTK7. Activates progelatinase A and MMP15	cell surface	Chimpanzee, Bonobo, Rhesus monkey, dog, cow, mouse, rat, zebrafish, fruit fly, mosquito, *C. elegans*, and frog.
MMP15 (transmembrane) MT-MMP 2	Endopeptidase degrades various extracellular matrix. Activates progelatinase A.	cell surface/secreted	Chimpanzee, Bonobo, Rhesus monkey, dog, cow, mouse, rat, chicken, zebrafish.
MMP16 (transmembrane) MT-MMP 3	Endopeptidase degrades various extracellular matrix. Activates progelatinase A. Matrix remodeling. Isoform cleaves fibronectin and collagen III. Degradation of Type I collagen with interaction from CSPG4.	secreted	Chimpanzee, Bonobo, Rhesus monkey, dog, cow, mouse, rat, chicken, zebrafish, and frog
MMP17 (transmembrane) MT-MMP 4	Endopeptidase degrades various extracellular matrix including fibrin. May activate growth factor precursors and inflammatory mediators such as necrosis factor-alpha. May be involved in tumoral process. Does **NOT** hydrolyze collagens I, II, III, IV, and V; gelatin, fibronectin, laminin, decorin, nor alpha1-antitrypsin.	membrane bound	Chimpanzee, Bonobo, Rhesus monkey, dog, cow, mouse, rat, chicken, zebrafish, fruit fly, mosquito, *C. elegans*, and frog.
MMP19 (stromelysin- 4)	Endopeptidase degrades various extracellular matrix components including aggrecan and cartilage oligomeric matrix protein. May take part in neovascularization and angiogenesis. Hydrolyzes Collagen IV, laminin, nidogen, nascin-c isoform, fibronectin and gelatin I.	extracellular	Chimpanzee, Bonobo, Rhesus monkey, dog, cow, mouse, rat, zebrafish, and frog.
MMP20 (enamelysin)	Degrades amelogenin. Major enamel component and macromolecules aggrecan and COMP.	N/K	Chimpanzee, Bonobo, Rhesus monkey, dog, cow, mouse, rat, zebrafish, and frog.
MMP21 (X MMP)	May play role in embryogenesis, especially in neuronal cells, and lymphocyte development and survival.	N/K	Chimpanzee, Bonobo, Rhesus monkey dog, mouse, rat, zebrafish, and frog

MMP24 (MT-MMP 5)	Mediates cleavage of N-cadherin and regulates neuro-immune interactions and neural stem cell quiescence. Involved in cell-cell interactions between nociception and inflammatory hyperalgesia. Mediates cleavage of CDH2. May contribute to axonal growth. Activates progelatinase- A. May be a proteoglycanase involved in degradation of proteoglycans (dermatan sulfate and chondroitin sulfate) partially cleaves gibronectin.	membrane bound	Chimpanzee, Bonobo, Rhesus monkey, dog, cow, mouse, rat, chicken, zebrafish, and frog.
MMP25 (MT-MMP 6)	Degrades casein. Possible role in tissue homeostasis and repair.	membrane bound	Chimpanzee, dog, cow, mouse, zebrafish, mosquito, and frog
MMP27 (MMP-22, C-MMP)	Matrix metalloproteinases degrade extracellular components such as fibronectin, laminin, gelatins, and collagens.	N/K	Chimpanzee, Bonobo, Rhesus monkey, dog, cow, mouse, rat, and chicken
MMP28 (epilysin)	Degrades casein. Possible role in tissue homeostasis and repair.	secreted	chimpanzee, Bonobo, Rhesus monkey, dog, cow, mouse, rat, and chicken

APPENDIX C

PRIMARY SEQUENCE IDENTITY BETWEEN VARIOUS PEX DOMAINS (HPX, MMPs, PRG4, VTN) GENERATED VIA DIRECT PAIRWISE ALIGNMENTS [33]

Table A4. Primary sequence identity between various PEX domains (HPX, MMPs, PRG4, VTN). * In the context of sequence alignments, a "score" is a relative numerical value that describes the overall quality of an alignment. Higher numbers correspond to higher similarity [33].

HPX domain	HPX-N					HPX-C				
	Max score*	Total score*	Query cover	E-Value	Ident	Max score*	Total score*	Query cover	E-Value	Ident
MMP1	52.0	119	87%	1e-12	27%	45.1	137	71%	4e-10	36%
MMP2	46.2	130	96%	2e-10	25%	46.2	144	80%	3e-10	31%
MMP3	51.6	113	97%	2e-12	31%	43.5	147	69%	1e-09	25%
MMP8	52.8	164	98%	8e-13	26%	43.1	138	81%	2e-9	24%
MMP9	32.3	182	93%	9e-06	31%	33.1	96.6	91%	7e-06	25%
MMP10	63.5	108	98%	1e-16	32%	42.4	163	90%	3e-09	24%
MMP11	56.2	101	95%	5e-14	28%	26.6	121	81%	5e-04	29%
MMP12	58.9	90.1	87%	6e-15	30%	41.2	97.0	66%	9e-09	25%
MMP14	45.1	156	95%	4e-10	23%	39.3	128	79%	4e-08	25%
MMP15	53.5	151	96%	7e-13	27%	32.7	167	90%	7e-06	23%
MMP16	53.1	150	95%	7e-13	25%	32.3	122	81%	9e-06	28%
MMP17	57.8	109	99%	2e-14	28%	33.9	101	62%	3e-06	26%
MMP19	57.4	104	95%	2e-14	31%	36.6	119	81%	3e-07	24%
MMP20	45.1	139	96%	3e-10	29%	28.1	92.8	87%	2e-04	24%
MMP21	39.7	140	85%	2e-08	41%	41.2	82.4	64%	1e-08	26%
MMP24	49.7	144	97%	1e-11	26%	30	88.5	97%	6e-05	28%
MMP25	28.9	111	95%	9e-05	24%	35.8	151	95%	6e-07	29%
MMP26	21.9	135	76%	0.011	39%	14.2	14.2	6%	2.4	46%
MMP27	48.9	63.5	87%	2e-11	25%	40.4	83.2	73%	2e-08	24%
MMP28	37.7	123	94%	1e-07	30%	40.0	97.4	81%	2e-08	31%
PRG4	47.4	161	85%	2e-10	29%	27.3	27.3	49%	8e-04	26%
VTN	52.4	177	99%	1e-12	31%	39.7	143	80%	3e-08	29%

APPENDIX D

PRIMARY SEQUENCE AND HOMOLOGY MODEL OF THE PROTEIN LIMUNECTIN FROM HORSESHOE CRAB

Primary sequence:

MFILKGMWTFLLLAILQISTCEVSQTDKTELHSTGMEILQSIFPSIDAVFKWSNGVTYIFKGSCYFRY
EDKTNEISNCRRLSAWGGLTGPVDAVFRWRNGVTYFFQGDCYYRYEDKTDEISKCSPVTAWGGMTGPV
DAVFRWSNGITYFFKEDCYYRYEDKDNKISKCTPITAWGKMTGPIDAVFRWSNGVTYFFKRDCYFRYE
DKPNEISKCRAIALWGASSYQPLDAVFRWNDGVTYFFKGFCYYHNDLKKCKPISAWGGISKPVSAVLL
WNNKETYFFEGKCYHSYEAKNNSISKCIPISTWAKKIRVVDAVFRWSNGITYFFKGDCYYRYEDKTNK
LSQCSPVTEWGGMTGPVDAVFRWSNGATYFFQGNCYYRYDDKNNKLSQCSPVTAWGGMTGPVDAVFRW
SNGATYFFKEDCYMKYEDKPQKLSGCNPISAWGGGIY

APPENDIX E

PRG4 - PEX-DOMAIN DNA NUCLEOTIDE SEQUENCE (SHOWN AS CODONS) USED FOR ORDERING OF THE GBLOCK FOR CLONING, EXPRESSION AND PURIFICATION

```
5'- CCG AAC CAG GGC ATT ATC ATT AAT CCT ATG CTT AGC GAT GAG
ACG AAT ATC TGC AAT GGT AAA CCG GTG GAC GGC TTG ACG ACC CTG
CGC AAC GGT ACT CTC GTG GCG TTC CGC GGC CAT TAC TTT TGG ATG
TTG TCT CCG TTT TCA CCA CCA TCC CCG GCG CGC CGC ATT ACC GAG
GTA TGG GGT ATC CCA AGC CCG ATT GAT ACG GTG TTT ACG CGT TGC
AAC TGT GAA GGT AAA ACT TTT TTC TTT AAA GAC AGC CAG TAT TGG
CGC TTT ACC AAT GAC ATT AAA GAT GCC GGC TAT CCC AAG CCT ATC
TTC AAA GGA TTT GGT GGC CTG ACA GGC CAA ATT GTT GCA GCG CTG
TCT ACG GCT AAA TAC AAA AAC TGG CCG GAA AGT GTC TAC TTT TTT
AAA CGC GGC GGC AGT ATT CAA CAA TAC ATC TAT AAA CAG GAG CCT
GTC CAA AAA TGC CCT GGT CGT CGT CCG GCG CTG AAT TAT CCC GTG
TAT GGA GAG ACA ACG CAG GTT CGC CGC CGT CGC TTC GAA CGC GCC
ATC GGC CCG AGC CAA ACC CAC ACG ATT CGC ATT CAA TAT TCC CCA
GCT CGT CTG GCG TAC CAG GAC AAA GGT GTG CTG CAC AAT GAA GTA
AAA GTT AGC ATC CTG TGG CGT GGC CTG CCG AAT GTG GTG ACG TCT
GCG ATC TCC CTC CCG AAT ATT CGC AAA CCC GAT GGT TAT GAT TAC
TAT GCA TTC TCA AAA GAT CAG TAC TAT AAT ATC GAT GTC CCC TCC
CGT ACA GCT CGT GCG ATC ACC ACT CGG TCA GGT CAA ACG CTG AGC
AAG GTT TGG TAC AAT TGC CCA -3'
```

Primary sequence of translated PRG4-PEX-domain "gene" polypeptide:

```
1   PNQGIIINPMLSDETNICNGKPVDGLTTLRNGTLVAFRGHYFWMLSPFSP 50
51  PSPARRITEVWGIPSPIDTVFTRCNCEGKTFFFKDSQYWRFTNDIKDAGY 100
101 PKPIFKGFGGLTGQIVAALSTAKYKNWPESVYFFKRGGSIQQYIYKQEPV 150
151 QKCPGRRPALNYPVYGETTQVRRRRFERAIGPSQTHTIRIQYSPARLAYQ 200
201 DKGVLHNEVKVSILWRGLPNVVTSAISLPNIRKPDGYDYYAFSKDQYYNI 250
251 DVPSRTARAITTRSGQTLSKVWYNCP 276
```

Total length= 276 residues
Molecular weight = 31.51 kDa

APPENDIX F

Novagen

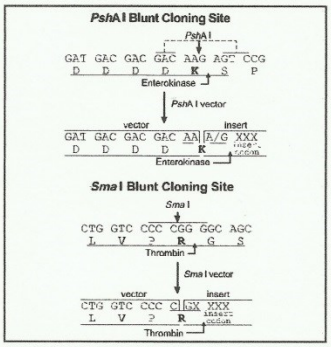
pET-44a-c(+) Vector

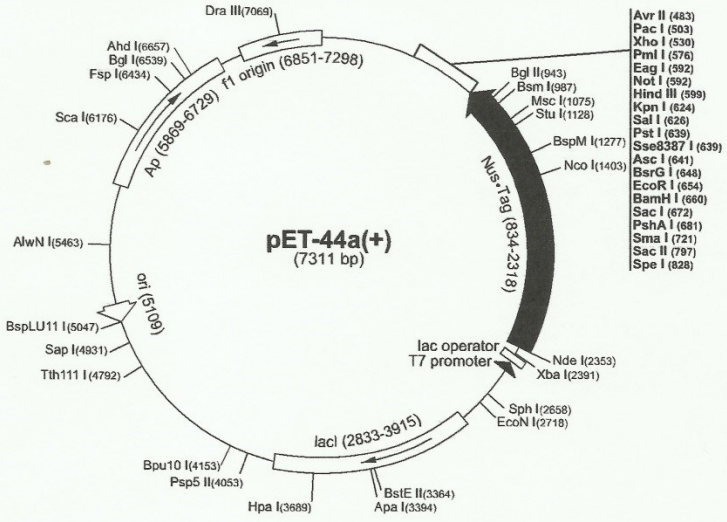
TB330 Rev. A 0502

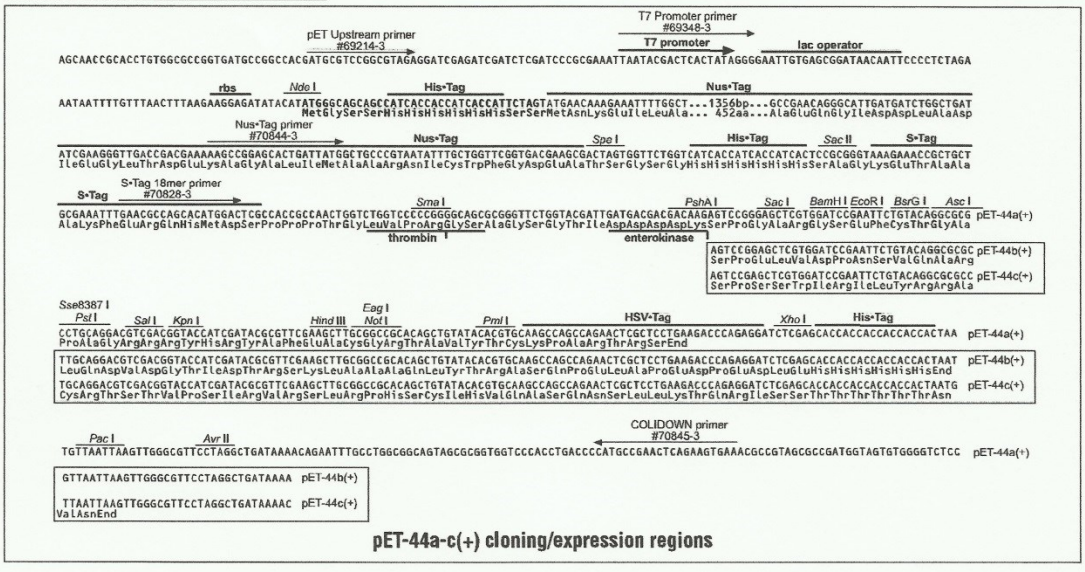
	Cat. No.
pET-44a(+) DNA	71122-3
pET-44b(+) DNA	71123-3
pET-44c(+) DNA	71124-3

pET-44a(+) sequence landmarks	
T7 promoter	2426-2442
T7 transcription start	2425
His•Tag® coding sequence	2325–2342
Nus•Tag™ coding sequence	834–2318
His•Tag® coding sequence	801–818
S•Tag™ coding sequence	747–791
Multiple cloning sites (*Sma* I-*Xho* I)	530–724
HSV•Tag® coding sequence	536–571
His•Tag® coding sequence	512–529
T7 terminator	26–73
lacI coding sequence	2833–3915
pBR322 origin	5109
bla (Ap) coding sequence	5870–6730
f1 origin	6852–7299

The pET-44 vectors are designed for cloning and high-level expression of peptide sequences fused with the 495 aa Nus•Tag™ protein. Compared to the pET-43.1 series, the pET-44 vectors encode an additional N-terminal His•Tag. Unique sites are shown on the circle map. Note that the sequence is numbered by the pBR322 convention, so the T7 expression region is reversed on the circle map. The cloning/expression region of the coding strand transcribed by T7 RNA polymerase is shown below. The f1 origin is oriented so that infection with helper phage will produce virions containing single stranded DNA that corresponds to the coding strand. Therefore, single stranded sequencing should be performed using the COLIDOWN primer (Cat. No. 70845-3). Vector encoded sequence can be completely removed when cloning into the *Psh*A I or *Sma* I sites (as shown below) by cleaving the Nus•Tag fusion protein with enterokinase or thrombin, respectively.







pET-44a-c(+) cloning/expression regions

APPENDIX G

Return of results from MU-DNA LIMS sequencing facility. Submission of sample from Colony 8 from PCR of transformed ligation products of the PRG4 PEX domain-pET-44a construct:

proteoglycan 4 isoform B preproprotein [Homo sapiens]
Sequence ID: ref|NP_001121180.2|Length: 1363Number of Matches: 1
Related Information
Gene-associated gene details
Map Viewer-aligned genomic context
Range 1: 1088 to 1363GenPeptGraphics Next Match Previous Match

Alignment statistics for match #1

Score	Expect	Method	Identities	Positives	Gaps
629 bits (1622)	0.0	Compositional matrix adjust.	276/276 (100%)	276/276 (100%)	0/276 (0%)

```
Query  1     PNQGIIINPMLSDETNICNGKPVDGLTTLRNGTLVAFRGHYFWMLSPFSPPSPARRITEV  60
             PNQGIIINPMLSDETNICNGKPVDGLTTLRNGTLVAFRGHYFWMLSPFSPPSPARRITEV
Sbjct  1088  PNQGIIINPMLSDETNICNGKPVDGLTTLRNGTLVAFRGHYFWMLSPFSPPSPARRITEV  1147

Query  61    WGIPSPIDTVFTRCNCEGKTFFFKDSQYWRFTNDIKDAGYPKPIFKGFGGLTGQIVAALS  120
             WGIPSPIDTVFTRCNCEGKTFFFKDSQYWRFTNDIKDAGYPKPIFKGFGGLTGQIVAALS
Sbjct  1148  WGIPSPIDTVFTRCNCEGKTFFFKDSQYWRFTNDIKDAGYPKPIFKGFGGLTGQIVAALS  1207

Query  121   TAKYKNWPESVYFFKRGGSIQQYIYKQEPVQKCPGRRPALNYPVYGETTQVRRRRFERAI  180
             TAKYKNWPESVYFFKRGGSIQQYIYKQEPVQKCPGRRPALNYPVYGETTQVRRRRFERAI
Sbjct  1208  TAKYKNWPESVYFFKRGGSIQQYIYKQEPVQKCPGRRPALNYPVYGETTQVRRRRFERAI  1267

Query  181   GPSQTHTIRIQYSPARLAYQDKGVLHNEVKVSILWRGLPNVVTSAISLPNIRKPDGYDYY  240
             GPSQTHTIRIQYSPARLAYQDKGVLHNEVKVSILWRGLPNVVTSAISLPNIRKPDGYDYY
Sbjct  1268  GPSQTHTIRIQYSPARLAYQDKGVLHNEVKVSILWRGLPNVVTSAISLPNIRKPDGYDYY  1327

Query  241   AFSKDQYYNIDVPSRTARAITTRSGQTLSKVWYNCP  276
             AFSKDQYYNIDVPSRTARAITTRSGQTLSKVWYNCP
Sbjct  1328  AFSKDQYYNIDVPSRTARAITTRSGQTLSKVWYNCP  1363
```

APPENDIX H

ANALYSES FROM MASS SPECTROMETRY ON PRG4 PEX-DOMAIN

MATRIX SCIENCE **MASCOT Search Results**

Protein View: NUSA_ECOLI

Transcription termination/antitermination protein NusA OS=Escherichia coli (strain K12) GN=nusA PE=1 SV=1

Database:	SwissProt
Score:	1277
Nominal mass (M_r):	55008
Calculated pI:	4.53
Taxonomy:	Escherichia coli K-12

Sequence similarity is available as an NCBI BLAST search of NUSA_ECOLI against nr.

Search parameters

MS data file:	E:\Data\Nov2016\2016Nov0301.raw
Enzyme:	Trypsin: cuts C-term side of KR unless next residue is P.
Fixed modifications:	Carbamidomethyl (C)
Variable modifications:	Oxidation (M)

Protein sequence coverage: 36%

Matched peptides shown in ***bold red***.

```
   1 MNKEILAVVE AVSNEKALPR EKIFEALESA LATATKKYE QEIDVRVQID
  51 RKSGDFDTFR RWLVVDEVTQ PTKEITLEAA RYEDESLNLG DYVEDQIESV
 101 TFDRITTQTA KQVIVQKVRE AERAMVVDQF REHEGEIITG VVKKVNRDNI
 151 SLDLGNNAEA VILREDMLPR ENFRPGDRVR GVLYSVRPEA RGAQLFVTRS
 201 KPEMLIELFR IEVPEIGEEV IEIKAAARDP GSRAKIAVKT NDKRIDPVGA
 251 CVGMRGARVQ AVSTELGGER IDIVLWDDNP AQFVINAMAP ADVASIVVDE
 301 DKHTMDIAVE AGNLAQAIGR NGQNVRLASQ LSGWELNVMT VDDLQAKEQA
 351 EAHAAIDTFT KYLDIDEDFA TVLVEEGFST LEELAYVPMK ELLEIEGLDE
 401 PTVEALRERA KNALATIAQA QEESLGDNKP ADDLLNLEGV DRDLAFKLAA
 451 RGVCTLEDLA EQGIDDLADI EGLTDEKAGA LIMAARNICW PGDEA
```

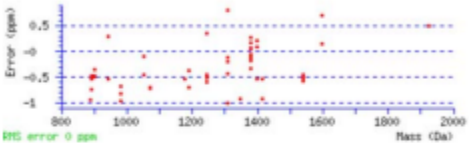
Unformatted sequence string: 495 residues (for pasting into other applications).

Sort peptides by ◉ Residue Number ○ Increasing Mass ○ Decreasing Mass

Show predicted peptides also

Query	Start – End	Observed	Mr(expt)	Mr(calc)	ppm	M	Score	Expect	Rank	U	Peptide
8931	4 – 16	700.8848	1399.7551	1399.7558	-0.53	0	56	3.4e-005	1	U	K.EILAVVEAVSNEK.A
8932	4 – 16	700.8852	1399.7559	1399.7558	0.084	0	53	0.0001	1	U	K.EILAVVEAVSNEK.A
8933	4 – 16	700.8853	1399.7561	1399.7558	0.20	0	53	4.1e-005	1	U	K.EILAVVEAVSNEK.A
8675	38 – 46	590.3035	1178.5925	1178.5931	-0.53	1	60	7.4e-005	1	U	K.KYEQEIDVR.V
8502	39 – 46	526.2561	1050.4977	1050.4982	-0.45	0	58	9.5e-006	1	U	K.YEQEIDVR.V
8503	39 – 46	526.2563	1050.4981	1050.4982	-0.10	0	53	0.00012	1	U	K.YEQEIDVR.V
8529	52 – 60	536.7562	1071.4978	1071.4985	-0.71	1	25	0.01	1	U	R.KSGDFDTFR.R
8530	52 – 60	536.7562	1071.4978	1071.4985	-0.70	1	28	0.013	1	U	R.KSGDFDTFR.R
8010	53 – 60	472.7088	943.4031	943.4036	-0.54	0	44	0.00028	1	U	K.SGDFDTFR.R
8011	53 – 60	472.7092	943.4039	943.4036	0.29	0	45	0.00034	1	U	K.SGDFDTFR.R
8943	62 – 73	707.8818	1413.7491	1413.7504	-0.93	0	70	4.1e-006	1	U	R.WLVVDEVTQPTK.E
8944	62 – 73	707.8821	1413.7497	1413.7504	-0.53	0	58	4.6e-005	1	U	R.WLVVDEVTQPTK.E
7573	74 – 81	451.7505	901.4864	901.4869	-0.49	0	38	0.024	1	U	K.EITLEAAR.Y
7574	74 – 81	451.7505	901.4864	901.4869	-0.47	0	42	0.011	1	U	K.EITLEAAR.Y
7575	74 – 81	451.7505	901.4865	901.4869	-0.35	0	42	0.011	1	U	K.EITLEAAR.Y
8158	124 – 131	491.2443	980.4740	980.4750	-0.97	0	33	0.012	1	U	R.AMVVDQFR.E + Oxidation (M)
8159	124 – 131	491.2444	980.4742	980.4750	-0.81	0	44	0.00086	1	U	R.AMVVDQFR.E + Oxidation (M)
8160	124 – 131	491.2444	980.4743	980.4750	-0.68	0	30	0.015	1	U	R.AMVVDQFR.E + Oxidation (M)
7353	132 – 143	437.5694	1309.6865	1309.6878	-1.01	0	31	0.0042	1	U	R.EHEGEIITGVVK.K
8821	132 – 143	655.8509	1309.6872	1309.6878	-0.42	0	53	5e-005	1	U	R.EHEGEIITGVVK.K
8822	132 – 143	655.8510	1309.6875	1309.6878	-0.19	0	52	0.00011	1	U	R.EHEGEIITGVVK.K
8823	132 – 143	655.8511	1309.6876	1309.6878	-0.13	0	61	1.2e-005	1	U	R.EHEGEIITGVVK.K
7354	132 – 143	437.5702	1309.6888	1309.6878	0.81	0	24	0.0079	1	U	R.EHEGEIITGVVK.K
8759	181 – 191	623.8485	1245.6824	1245.6830	-0.45	0	28	0.017	1	U	R.GVLYSVRPEAR.G
8761	181 – 191	416.2351	1245.6834	1245.6830	0.34	0	15	0.046	1	U	R.GVLYSVRPEAR.G
7510	192 – 199	446.2556	890.4967	890.4974	-0.74	0	51	0.00064	1	U	R.GAQLFVTR.S
7511	192 – 199	446.2557	890.4969	890.4974	-0.54	0	51	0.00053	1	U	R.GAQLFVTR.S
7512	192 – 199	446.2558	890.4970	890.4974	-0.47	0	51	0.00073	1	U	R.GAQLFVTR.S
8864	200 – 210	689.8733	1377.7321	1377.7326	-0.33	0	48	7.2e-005	1	U	R.SKPEMLIELFR.I + Oxidation (M)
8865	200 – 210	689.8733	1377.7321	1377.7326	-0.33	0	55	1.9e-005	1	U	R.SKPEMLIELFR.I + Oxidation (M)
8866	200 – 210	689.8734	1377.7323	1377.7326	-0.18	0	55	1.9e-005	1	U	R.SKPEMLIELFR.I + Oxidation (M)
8867	200 – 210	460.2514	1377.7324	1377.7326	-0.16	0	39	0.00038	1	U	R.SKPEMLIELFR.I + Oxidation (M)
8868	200 – 210	460.2514	1377.7324	1377.7326	-0.13	0	36	0.00044	1	U	R.SKPEMLIELFR.I + Oxidation (M)
8869	200 – 210	689.8735	1377.7325	1377.7326	-0.060	0	42	0.00036	1	U	R.SKPEMLIELFR.I + Oxidation (M)
8870	200 – 210	460.2515	1377.7326	1377.7326	0.040	0	46	7.4e-005	1	U	R.SKPEMLIELFR.I + Oxidation (M)
8871	200 – 210	460.2515	1377.7327	1377.7326	0.075	0	47	0.00017	1	U	R.SKPEMLIELFR.I + Oxidation (M)
8872	200 – 210	460.2515	1377.7328	1377.7326	0.16	0	29	0.0019	1	U	R.SKPEMLIELFR.I + Oxidation (M)

Query	Start - End	Observed	Mr(expt)	Mr(calc)	ppm	M	Score	Expect	Rank	U	Peptide
8873	200 - 210	689.8738	1377.7330	1377.7326	0.28	0	46	4.5e-005	1	U	R.SKPEMLIELFR.I + Oxidation (M)
9087	211 - 224	798.9403	1595.8660	1595.8658	0.14	0	68	3.6e-006	1	U	R.IEVPEIGEEVIEIK.A
9088	211 - 224	798.9407	1595.8669	1595.8658	0.71	0	67	4.8e-006	1	U	R.IEVPEIGEEVIEIK.A
8847	244 - 255	449.5600	1345.6582	1345.6595	-0.93	1	33	0.0023	1	U	K.RIDPVGACVGMR.G + Oxidation (M)
8703	245 - 255	595.7861	1189.5575	1189.5584	-0.70	0	51	0.0001	1	U	R.IDPVGACVGMR.G + Oxidation (M)
8704	245 - 255	595.7863	1189.5579	1189.5584	-0.36	0	32	0.001	1	U	R.IDPVGACVGMR.G + Oxidation (M)
8756	259 - 270	623.3249	1244.6353	1244.6361	-0.60	0	64	2.7e-005	1	U	R.VQAVSTELGGER.I
8757	259 - 270	623.3249	1244.6353	1244.6361	-0.59	0	21	0.021	1	U	R.VQAVSTELGGER.I
8758	259 - 270	623.3250	1244.6354	1244.6361	-0.52	0	73	3.5e-006	1	U	R.VQAVSTELGGER.I
9020	348 - 361	513.9229	1538.7469	1538.7477	-0.58	0	45	0.00024	1	U	K.HQAEAHAAIDTFTK.Y
9021	348 - 361	770.3807	1538.7469	1538.7477	-0.53	0	97	1.2e-009	1	U	K.HQAEAHAAIDTFTK.Y
9022	348 - 361	513.9229	1538.7470	1538.7477	-0.51	0	68	1.5e-006	1	U	K.HQAEAHAAIDTFTK.Y
9023	348 - 361	513.9230	1538.7470	1538.7477	-0.46	0	54	1.2e-005	1	U	K.HQAEAHAAIDTFTK.Y
9209	391 - 407	963.5074	1925.0003	1924.9993	0.50	0	60	3.7e-005	1	U	K.ELLEIEGLDEPTVEALR.E
7494	478 - 486	445.2494	888.4842	888.4851	-0.95	0	46	0.0029	1	U	K.AGALIMAAR.N + Oxidation (M)



```
ID   NUSA_ECOLI              Reviewed;         495 AA.
AC   P0AFF6; P03003; Q2M941;
DT   21-JUL-1986, integrated into UniProtKB/Swiss-Prot.
DT   20-DEC-2005, sequence version 1.
DT   11-MAY-2016, entry version 100.
DE   RecName: Full=Transcription termination/antitermination protein NusA {ECO:0000255|HAMAP-Rule:MF_00945};
DE   AltName: Full=N utilization substance protein A;
DE   AltName: Full=Transcription termination/antitermination L factor;
GN   Name=nusA {ECO:0000255|HAMAP-Rule:MF_00945};
GN   OrderedLocusNames=b3169, JW3138;
OS   Escherichia coli (strain K12).
OC   Bacteria; Proteobacteria; Gammaproteobacteria; Enterobacteriales;
OC   Enterobacteriaceae; Escherichia.
OX   NCBI_TaxID=83333;
RN   [1]
RP   NUCLEOTIDE SEQUENCE [GENOMIC DNA].
RX   PubMed=6326058; DOI=10.1093/nar/12.7.3333;
RA   Ishii S., Ihara M., Maekawa T., Nakamura Y., Uchida H., Imamoto F.;
RT   "The nucleotide sequence of the cloned nusA gene and its flanking
RT   region of Escherichia coli.";
RL   Nucleic Acids Res. 12:3333-3342(1984).
RN   [2]
RP   SEQUENCE REVISION.
RX   PubMed=3027511; DOI=10.1007/BF00430455;
RA   Saito M., Tsugawa A., Egawa K., Nakamura Y.;
RT   "Revised sequence of the nusA gene of Escherichia coli and
RT   identification of nusA11 (ts) and nusA1 mutations which cause changes
RT   in a hydrophobic amino acid cluster.";
RL   Mol. Gen. Genet. 205:380-382(1986).
RN   [3]
RP   SEQUENCE REVISION, PARTIAL PROTEIN SEQUENCE, INDUCTION, AND
RP   MUTAGENESIS OF ARG-104; GLY-181; LEU-183 AND GLU-212.
RX   PubMed=1847365;
RA   Ito K., Egawa K., Nakamura Y.;
RT   "Genetic interaction between the beta' subunit of RNA polymerase and
RT   the arginine-rich domain of Escherichia coli nusA protein.";
RL   J. Bacteriol. 173:1492-1501(1991).
RN   [4]
RP   NUCLEOTIDE SEQUENCE [LARGE SCALE GENOMIC DNA].
RC   STRAIN=K12 / MG1655 / ATCC 47076;
RX   PubMed=9278503; DOI=10.1126/science.277.5331.1453;
RA   Blattner F.R., Plunkett G. III, Bloch C.A., Perna N.T., Burland V.,
RA   Riley M., Collado-Vides J., Glasner J.D., Rode C.K., Mayhew G.F.,
RA   Gregor J., Davis N.W., Kirkpatrick H.A., Goeden M.A., Rose D.J.,
RA   Mau B., Shao Y.;
RT   "The complete genome sequence of Escherichia coli K-12.";
RL   Science 277:1453-1462(1997).
RN   [5]
RP   NUCLEOTIDE SEQUENCE [LARGE SCALE GENOMIC DNA].
RC   STRAIN=K12 / W3110 / ATCC 27325 / DSM 5911;
RX   PubMed=16738553; DOI=10.1038/msb4100049;
RA   Hayashi K., Morooka N., Yamamoto Y., Fujita K., Isono K., Choi S.,
RA   Ohtsubo E., Baba T., Wanner B.L., Mori H., Horiuchi T.;
RT   "Highly accurate genome sequences of Escherichia coli K-12 strains
RT   MG1655 and W3110.";
RL   Mol. Syst. Biol. 2:E1-E5(2006).
RN   [6]
RP   PROTEIN SEQUENCE OF 1-13.
RC   STRAIN=K12 / EMG2;
RX   PubMed=9298646; DOI=10.1002/elps.1150180807;
RA   Link A.J., Robison K., Church G.M.;
RT   "Comparing the predicted and observed properties of proteins encoded
RT   in the genome of Escherichia coli K-12.";
RL   Electrophoresis 18:1259-1313(1997).
RN   [7]
RP   IDENTIFICATION AS L FACTOR, AND INTERACTION WITH N PROTEIN.
RX   PubMed=6154941; DOI=10.1073/pnas.77.4.1991;
```

MASCOT Search Results

Protein View: PRG4_HUMAN

Proteoglycan 4 OS=Homo sapiens GN=PRG4 PE=1 SV=2

Database:	SwissProt
Score:	1125
Nominal mass (M_r):	152238
Calculated pI:	9.53
Taxonomy:	Homo sapiens

Sequence similarity is available as an NCBI BLAST search of PRG4_HUMAN against nr.

Search parameters

MS data file:	E:\Data\Nov2016\2016Nov0301.raw
Enzyme:	Trypsin: cuts C-term side of KR unless next residue is P.
Fixed modifications:	Carbamidomethyl (C)
Variable modifications:	Oxidation (M)

Protein sequence coverage: 10%

Matched peptides shown in ***bold red***.

```
   1 MAMKTLPIYL LLLLSVFVIQ QVSSQDLSSC AGRCGEGYSR DATCNCDYNC
  51 QHYMECCPDF KRVCTAELSC KGRCFESFER GRECDCDAQC KKYDKCCPDY
 101 ESFCAEVHNP TSPPSSKKAP PPSGASQTIK STTKRSPKPP NKKKTKKVIE
 151 SEEITEENSV SENQESSSSS SSSSSSSTIR KISSSKNSAA NRELQKKLKV
 201 KDNKKNRTKK KPTPKPPVVD EAGSGLDNGD FKVTTPDTST TQHNKVSTSP
 251 KITTAKPINP RPSLPPNSDT SKETSLTVNK ETTVETKETT TTNKQTSTDG
 301 KEKTTSAKET QSIEKTSAKD LAPTSKVLAK PTPKAETTTK GPALTTPKEP
 351 TPTTPKEPAS TTPKEPTPTT IKSAPTTPKE PAPTTTKSAP TTPKEPAPTT
 401 TKEPAPTTPK EPAPTTTKEP APTTTKSAPT TPKEPAPTTP KKPAPTTPKE
 451 PAPTTPKEPT PTTPKEPAPT TKEPAPTTPK EPAPTAPKEP APTTPKEPAP
 501 TTPKEPAPTT TKEPSPTTPK EPAPTTTKSA PTTTKEPAPT TTKSAPTTPK
 551 EPSPTTTKEP APTTPKEPAP TTPKEPAPTT PKEPAPTTPK EPAPTTTKKP
 601 APTTPKEPAP TTPKETAPTT PKKLTPTTPE KLAPTTPKEP APTTPKKLAP
 651 TTPKEPTPTT PKEPAPTTPK AAAPNTPKEP APTTPKEPAP TTPKEPAPTT
 701 PKETAPTTPK GTAPTTLKEP APTTPKEPAP KELAPTTTKE PTSTTCDKPA
 751 PTTPKGTAPT TPKEPAPTTP KEPAPTTPKG TAPTTLKEPA PTTPKEPAPK
 801 KLAPTTTKGP TSTTSDKPAP TTPKETAPTT PKEPAPTTPK KPAPTTPETP
 851 PPTTSEVSTP TTTKEPTTIK KSPDESTPEL SAEPTPKALE NSPKEPGVPT
 901 TKTPAATKPE MTTTAKDKTT KRDLRTTPET TTAAPKMTKE TATTTEKTTE
 951 SKITATTTQV TSTTTQDTTP FKITTLKTTT LAPKVTTTKE TITTTEIMNK
1001 PEETAKPKDR ATNSKATTPK PQKPTKAPKK PTSTKKPKTM PRVPKPKTTP
1051 TPRKMTSTMP KLNPTSRIAE AMLQTTTRPN QTPNSKLVEV NPKSEDAGGA
1101 EGETPHMLLR PHVPMPEVTP DMDYLPRVPN QGIIINPMLS DETNICNGKP
1151 VDGLTTLRNG TLVAFRGHYF WMLSPFSPPS PARRITEVWG IPSPIDTVFT
1201 RCNCEGKTFF FKDSQYWRFT NDIKDAGYPK PIFKGFGGLT GQIVAALSTA
1251 KYKNWPESVY FFKRGGSIQQ YIYKQEPVQK CPGRRPALNY PVYGETTQVR
1301 RRRFERAIGP SQTHTIRIQY SPARLAYQDK GVLHNEVKVS ILWRGLPNVV
1351 TSAISLPNIR KPDGYDYYAF SKDQYYNIDV PSRTARAITT RSGQTLSKVW
1401 YNCP
```

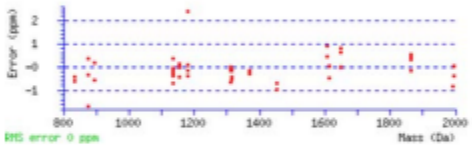
Unformatted sequence string: 1404 residues (for pasting into other applications).

Sort peptides by ◉ Residue Number ○ Increasing Mass ○ Decreasing Mass

Show predicted peptides also

Query	Start – End	Observed	Mr(expt)	Mr(calc)	ppm	M	Score	Expect	Rank	U	Peptide
7402	1159 – 1166	439.2474	876.4803	876.4818	-1.70	0	57	0.00015	1	U	R.NGTLVAFR.G
7404	1159 – 1166	439.2480	876.4815	876.4818	-0.34	0	48	0.0011	1	U	R.NGTLVAFR.G
7405	1159 – 1166	439.2483	876.4821	876.4818	0.34	0	58	0.00018	1	U	R.NGTLVAFR.G
9230	1167 – 1183	664.9851	1991.9336	1991.9352	-0.82	0	49	6.8e-005	1	U	R.GHYFWMLSPFSPPSPAR.R + Oxidation (M)
8616	1225 – 1234	379.2095	1134.6065	1134.6073	-0.69	0	26	0.033	1	U	K.DAGYPKPIFK.G
8617	1225 – 1234	568.3107	1134.6069	1134.6073	-0.38	0	37	0.013	1	U	K.DAGYPKPIFK.G
8618	1225 – 1234	568.3107	1134.6069	1134.6073	-0.37	0	37	0.023	1	U	K.DAGYPKPIFK.G
8620	1225 – 1234	379.2096	1134.6070	1134.6073	-0.28	0	26	0.028	1	U	K.DAGYPKPIFK.G
8621	1225 – 1234	568.3109	1134.6072	1134.6073	-0.16	0	42	0.0075	1	U	K.DAGYPKPIFK.G
8622	1225 – 1234	568.3109	1134.6073	1134.6073	-0.058	0	42	0.0075	1	U	K.DAGYPKPIFK.G
8623	1225 – 1234	568.3111	1134.6077	1134.6073	0.35	0	38	0.0082	1	U	K.DAGYPKPIFK.G
9089	1252 – 1263	536.6015	1606.7827	1606.7820	0.43	1	41	0.0044	1	U	K.YKNWPESVYFFK.R
9090	1252 – 1263	536.6018	1606.7834	1606.7820	0.89	1	41	0.0055	1	U	K.YKNWPESVYFFK.R
8833	1254 – 1263	658.8188	1315.6230	1315.6237	-0.53	0	49	3.2e-005	1	U	K.NWPESVYFFK.R

Query	Start - End	Observed	Mr(expt)	Mr(calc)	ppm	M	Score	Expect	Rank	U	Peptide
8834	1254 - 1263	658.8189	1315.6232	1315.6237	-0.41	0	37	0.0015	1	U	K.MWPESVYFFK.R
8835	1254 - 1263	658.8190	1315.6235	1315.6237	-0.16	0	46	0.00019	1	U	K.MWPESVYFFK.R
7360	1264 - 1274	438.2382	1311.6926	1311.6935	-0.66	1	15	0.043	1	U	K.RGGSIQQYIYK.Q
8824	1264 - 1274	656.8539	1311.6933	1311.6935	-0.14	1	51	0.00037	1	U	K.RGGSIQQYIYK.Q
8825	1264 - 1274	656.8540	1311.6935	1311.6935	-0.020	1	43	0.0026	1	U	K.RGGSIQQYIYK.Q
8646	1265 - 1274	578.8032	1155.5919	1155.5924	-0.42	0	35	0.0021	1	U	R.GGSIQQYIYK.Q
8647	1265 - 1274	578.8035	1155.5924	1155.5924	-0.018	0	47	0.00047	1	U	R.GGSIQQYIYK.Q
8648	1265 - 1274	578.8036	1155.5926	1155.5924	0.13	0	45	0.00075	1	U	R.GGSIQQYIYK.Q
9175	1285 - 1300	932.4891	1862.9636	1862.9639	-0.16	0	50	2.9e-005	1	U	R.RPALNYPVYGETTQVR.R
9176	1285 - 1300	932.4891	1862.9636	1862.9639	-0.13	0	78	2.1e-007	1	U	R.RPALNYPVYGETTQVR.R
9177	1285 - 1300	621.9954	1862.9645	1862.9639	0.32	0	39	0.00023	1	U	R.RPALNYPVYGETTQVR.R
9178	1285 - 1300	621.9955	1862.9647	1862.9639	0.44	0	66	1.5e-006	1	U	R.RPALNYPVYGETTQVR.R
9179	1285 - 1300	621.9955	1862.9647	1862.9639	0.46	0	40	0.00016	1	U	R.RPALNYPVYGETTQVR.R
9180	1285 - 1300	621.9956	1862.9649	1862.9639	0.53	0	67	1e-006	1	U	R.RPALNYPVYGETTQVR.R
8679	1307 - 1317	590.8251	1179.6356	1179.6360	-0.36	0	57	1.8e-005	1	U	R.AIGPSQTHTIR.I
8680	1307 - 1317	394.2192	1179.6358	1179.6360	-0.17	0	42	0.00058	1	U	R.AIGPSQTHTIR.I
8681	1307 - 1317	590.8253	1179.6361	1179.6360	0.062	0	56	3e-005	1	U	R.AIGPSQTHTIR.I
8682	1307 - 1317	394.2202	1179.6389	1179.6360	2.41	0	48	0.00018	1	U	R.AIGPSQTHTIR.I
9232	1307 - 1324	666.0287	1995.0642	1995.0650	-0.39	1	23	0.007	1	U	R.AIGPSQTHTIRIQYSPAR.L
9234	1307 - 1324	499.7735	1995.0651	1995.0650	0.039	1	21	0.01	1	U	R.AIGPSQTHTIRIQYSPAR.L
6748	1318 - 1324	417.7268	833.4390	833.4395	-0.61	0	20	0.047	1	U	R.IQYSPAR.L
6749	1318 - 1324	834.4465	833.4392	833.4395	-0.41	0	37	0.0047	1	U	R.IQYSPAR.L
9099	1325 - 1338	538.6261	1612.8565	1612.8573	-0.46	1	36	0.00071	1	U	R.LAYQDKGVLHNEVK.V
9100	1325 - 1338	538.6264	1612.8573	1612.8573	0.033	1	28	0.006	1	U	R.LAYQDKGVLHNEVK.V
7551	1331 - 1338	448.2532	894.4918	894.4923	-0.53	0	32	0.028	1	U	K.GVLHNEVK.V
7552	1331 - 1338	448.2535	894.4924	894.4923	0.16	0	37	0.0075	1	U	K.GVLHNEVK.V
9120	1345 - 1360	825.9805	1649.9464	1649.9464	-0.0024	0	103	4.7e-010	1	U	R.GLPNVVTSAISLPNIR.K
9121	1345 - 1360	825.9810	1649.9475	1649.9464	0.63	0	89	1.4e-008	1	U	R.GLPNVVTSAISLPNIR.K
9122	1345 - 1360	550.9899	1649.9478	1649.9464	0.82	0	41	0.00013	1	U	R.GLPNVVTSAISLPNIR.K
8969	1361 - 1372	727.3346	1452.6547	1452.6561	-0.98	0	60	2.4e-006	1	U	R.KPDGYDYYAFSK.D
8970	1361 - 1372	727.3348	1452.6551	1452.6561	-0.70	0	71	2e-007	1	U	R.KPDGYDYYAFSK.D
8971	1361 - 1372	485.2257	1452.6551	1452.6561	-0.68	0	20	0.015	1	U	R.KPDGYDYYAFSK.D
8972	1361 - 1372	485.2257	1452.6551	1452.6561	-0.67	0	26	0.0036	1	U	R.KPDGYDYYAFSK.D
8856	1373 - 1383	685.3226	1368.6306	1368.6310	-0.28	0	44	0.00012	1	U	K.DQYYNIDVPSR.T
8857	1373 - 1383	685.3226	1368.6307	1368.6310	-0.19	0	44	8.3e-005	1	U	K.DQYYNIDVPSR.T
8858	1373 - 1383	685.3227	1368.6308	1368.6310	-0.17	0	43	0.00018	1	U	K.DQYYNIDVPSR.T




```
ID   1146   1403       {ECO:0000255|PROSITE-ProRule:PRU00350}.
FT   VAR_SEQ      26     66       Missing (in isoform B, isoform D and
FT                                isoform E).
FT                                {ECO:0000303|PubMed:14702039}.
FT                                /FTId=VSP_016467.
FT   VAR_SEQ     107    199       Missing (in isoform C and isoform D).
FT                                {ECO:0000305}.
FT                                /FTId=VSP_016468.
FT   VAR_SEQ     157    199       Missing (in isoform F).
FT                                {ECO:0000303|PubMed:14976050}.
FT                                /FTId=VSP_016469.
FT   VAR_SEQ     412    841       Missing (in isoform E).
FT                                {ECO:0000303|PubMed:14702039}.
FT                                /FTId=VSP_016470.
FT   VARIANT     180    180       R -> W (in dbSNP:rs2273779).
FT                                {ECO:0000269|Ref.1}.
FT                                /FTId=VAR_024023.
FT   VARIANT    1130   1130       N -> S (in dbSNP:rs10158395).
FT                                /FTId=VAR_051559.
FT   VARIANT    1272   1272       I -> T (in dbSNP:rs1293985).
FT                                /FTId=VAR_051560.
FT   VARIANT    1296   1296       T -> M (in dbSNP:rs12134934).
FT                                {ECO:0000269|Ref.1}.
FT                                /FTId=VAR_051561.
FT   CONFLICT    604    604       T -> A (in Ref. 1; AAB09089).
FT                                {ECO:0000305}.
FT   CONFLICT    746    746       C -> S (in Ref. 1; AAB09089).
FT                                {ECO:0000305}.
FT   CONFLICT   1340   1340       S -> G (in Ref. 4; AAT74746).
FT                                {ECO:0000305}.
FT   CONFLICT   1380   1380       V -> G (in Ref. 4; AAT74746).
FT                                {ECO:0000305}.
FT   CONFLICT   1397   1397       S -> F (in Ref. 4; AAT74746).
FT                                {ECO:0000305}.
SQ   SEQUENCE   1404 AA;  151077 MW;  782A11746B3FDEE5 CRC64;
     MAMKTLPIYL LLLLSVFVIQ QVSSQDLSSC AGRCGEGYSR DATCNCDYNC QNYMECCPDF
     KRVCTAELSC KGRCFESFER GRECDCDAQC KKYDKCCPDY ESFCAEVHNP TSPPSSKKAP
     PPSGASQTIK STTKRSPKPP NKKKTKKVIE SEEITEEHSV SENQESSSSS SSSSSSSTIR
     KIKSSKNSAA NRELQKKLKV KDNKKNRTKK KPTPKPPVVD EAGSGLDNGD FKVTTPDTST
     TQHNKVSTSP KITTAKPINP RPSLPPNSDT SKETSLTVNK ETTVETKETT TTNKQTSTDG
     KEKTTSAKET QSIEKTSAKD LAPTSKVLAK PTPKAETTTK GPALTTPKEP TPTTPKEPAS
     TTPKEPTPTT IKSAPTTPKE PAPTTTKSAP TTPKEPAPTT TKEPAPTTPK EPAPTTTKEP
```

APPENDIX I

SEQUENCE ALIGNMENTS FOR THE Atp1a3 PROTEIN IN PRIMATES ROOTED WITH MURINES

Table A5. Sequence alignment of the Atp1a3 protein in mammals with an emphasis on Primates. The surprise phylogenetic split in the Great Apes is supported by numerous blocks of divergent non-synonymous residue substitutions that are consistent throughout the length of the proteins. The differential blocks for the "Human-Chimp" clade are highlighted in **GREEN** and for the "Bonobo-Gorilla" clade in **RED**. Site # indicates the beginning and end of the specific portions of the alignment presented. The entire alignment is 1031 sites in length

SPECIES	SITE #	SEQUENCE	SITE #
Colobus	1	MGGWEEE**RNGRAKMGRLS--DKKDDKDSPKKNKGKERRD**LDDLKKEVAM	49
Chimp	1	MG-----**SGGSDSYRIATSQDKKDDKDSPKKNKGKERRD**LDDLKKEVAM	49
Human	1	MG----------**-------------DKKDDKDSPKKNKGKERRD**LDDLKKEVAM	49
Mangaby	1	MA-----**R------------DKKDDKDSPKKNKGKERRD**LDDLKKEVAM	49
Bonobo	1	MG-----RGAGREYSPAA--TTAENGGGKKKQKEKE**---**LDELKKEVAM	49
Gorilla	1	MG-----RGAGREYSPAA--TTAENGGGKKKQKEKE**---**LDELKKEVAM	49
Orangutan	1	MG-----RGAGREYSPAA--TTAENGGGKKKQKEKE**---**LDELKKEVAM	49
Baboon	1	MG-----RGAGREYSPAA--TTAENGGGKKKQKEKE---LDELKKEVAM	49
Lemur	1	MG----------DKKDDKGSPKKNKAKERRDLDDLKKEVAM	49
Saimiri	1	MG----------DKKDDKGSPKKNKGKERRDLDDLKKEVAM	49
Aotus	1	MG----------DKKDDKGSPKKNKGKERRDLDDLKKEVAM	49
Rat	1	MG----------DKKDDKSSPKKSKAKERRDLDDLKKEVAM	49
Mouse	1	MG----------DKKDDKSSPKKSKAKERRDLDDLKKEVAM	49

SPECIES	SITE #	SEQUENCE	SITE #
Colobus	50	**TEHKMSVEEVCRKYNTDCVQ**GLTHSKAQEILARDGPNALTPPPTTPEWV	98
Chimp	50	**TEHKMSVEEVCRKYNTDCVQ**GLTHSKAQEILARDGPNALTPPPTTPEWV	98
Human	50	**TEHKMSVEEVCRKYNTDCVQ**GLTHSKAQEILARDGPNALTPPPTTPEWV	98
Mangaby	50	**TEHKMSVEEVCRKYNTDCVQ**GLTHSKAQEILARDGPNALTPPPTTPEWV	98
Bonobo	50	**DDHKLSLDELGRKYQVDLSK**GLTNQRAQDVLARDGPNALTPPPTTPEWV	98
Gorilla	50	**DDHKLSLDELGRKYQVDLSK**GLTNQRAQDVLARDGPNALTPPPTTPEWV	98
Orangutan	50	**DDHKLSLDELGRKYQVDLSK**GLTNQRAQDVLARDGPNALTPPPTTPEWV	98
Baboon	50	**DDHKLSLDELGRKYQVDLSK**GLTNQRAQDVLARDGPNALTPPPTTPEWV	98
Lemur	50	TEHKMSVEEVCRKYNTDCVQGLTHSKAQEILARDGPNALTPPPTTPEWV	98
Saimiri	50	TEHKMSVEEVCRKYNTDCVQGLTHSKAQEILARDGPNALTPPPTTPEWV	98
Aotus	50	TEHKMSVEEVCRKYNTDCVQGLTHSKAQEILARDGPNALTPPPTTPEWV	98
Rat	50	TEHKMSVEEVCRKYNTDCVQGLTHSKAQEILARDGPNALTPPPTTPEWV	98
Mouse	50	TEHKMSVEEVCRKYNTDCVQGLTHSKAQEILARDGPNALTPPPTTPEWV	98

SPECIES	SITE #	SEQUENCE	SITE #
Colobus	410	TTEDQSGTSFDKSSHTWVALSHIAGLCNRAVFKGGQDNIPVLKRDVAGD	458
Chimp	410	TTEDQSGTSFDKSSHTWVALSHIAGLCNRAVFKGGQDNIPVLKRDVAGD	458
Human	410	TTEDQSGTSFDKSSHTWVALSHIAGLCNRAVFKGGQDNIPVLKRDVAGD	458
Mangaby	410	TTEDQSGTSFDKSSHTWVALSHIAGLCNRAVFKGGQDNIPVLKRDVAGD	458
Bonobo	410	TTEDQSGATFDKRSPTWTALSRIAGLCNRAVFKAGQENISVSKRDTAGD	458
Gorilla	410	TTEDQSGATFDKRSPTWTALSRIAGLCNRAVFKAGQENISVSKRDTAGD	458
Orangutan	410	TTEDQSGATFDKRSPTWTALSRIAGLCNRAVFKAGQENISVSKRDTAGD	458
Baboon	410	TTEDQSGATFDKRSPTWTALSRIAGLCNRAVFKAGQENISVSKRDTAGD	458
Lemur	410	TTEDQSGTSFDKSSHTWVALSHIAGLCNRAVFKGGQDNIPVLKRDVAGD	458
Saimiri	410	TTEDQSGTSFDKSSHTWVALSHIAGLCNRAVFKGGQDNIPVLKRDVAGD	458
Aotus	410	TTEDQSGTSFDKSSHTWVALSHIAGLCNRAVFKGGQDNIPVLKRDVAGD	458
Rat	410	TTEDQSGTSFDKSSHTWVALSHIAGLCNRAVFKGGQDNIPVLKRDVAGD	458
Mouse	410	TTEDQSGTSFDKSSHTWVALSHIAGLCNRAVFKGGQDNIPVLKRDVAGD	458

SPECIES	SITE #	SEQUENCE	SITE #
Colobus	466	KCIELSSGSVKLMRERNKKVAEIPFNSTNKYQLSIHETEDPNDNRYLLV	514
Chimp	466	KCIELSSGSVKLMRERNKKVAEIPFNSTNKYQLSIHETEDPNDNRYLLV	514
Human	466	KCIELSSGSVKLMRERNKKVAEIPFNSTNKYQLSIHETEDPNDNRYLLV	514
Mangaby	466	KCIELSSGSVKLMRERNKKVAEIPFNSTNKYQLSIHETEDPNDNRYLLV	514
Bonobo	466	KCIELSCGSVRKMRDRNPKVAEIPFNSTNKYQLSIHEREDSPQS-HVLV	514
Gorilla	466	KCIELSCGSVRKMRDRNPKVAEIPFNSTNKYQLSIHEREDSPQS-HVLV	514
Orangutan	466	KCIELSCGSVRKMRDRNPKVAEIPFNSTNKYQLSIHEREDSPQS-HVLV	514
Baboon	466	KCIELSCGSVRKMRDRNPKVAEIPFNSTNKYQLSIHEREDSPQS-HVLV	514
Lemur	466	KCIELSSGSVKLMRERNKKVAEIPFNSTNKYQLSIHETEDPNDNRYLLV	514
Saimiri	466	KCIELSSGSVKLMRERNKKVAEIPFNSTNKYQLSIHETEDPNDNRYLLV	514
Aotus	466	KCIELSSGSVKLMRERNKKVAEIPFNSTNKYQLSIHETEDPNDNRYLLV	514
Rat	466	KCIELSSGSVKLMRERNKKVAEIPFNSTNKYQLSIHETEDPNDNRYLLV	514
Mouse	466	KCIELSSGSVKLMRERNKKVAEIPFNSTNKYQLSIHETEDPNDNRYLLV	514

SPECIES	SITE #	SEQUENCE	SITE #
Colobus	550	YLELGGLGERVLGFC**HYYLPEEQFPKGFAFDCDDVNFTTDN**LCFVGLMS	598
Chimp	550	YLELGGLGERVLGFC**HYYLPEEQFPKGFAFDCDDVNFTTDN**LCFVGLMS	598
Human	550	YLELGGLGERVLGFC**HYYLPEEQFPKGFAFDCDDVNFTTDN**LCFVGLMS	598
Mangaby	550	YLELGGLGERVLGFC**HYYLPEEQFPKGFAFDCDDVNFTTDN**LCFVGLMS	598
Bonobo	550	YMELGGLGERVLGFC**QLNLPSGKFPRGFKFDTDELNFPTEK**LCFVGLMS	598
Gorilla	550	YMELGGLGERVLGFC**QLNLPSGKFPRGFKFDTDELNFPTEK**LCFVGLMS	598
Orangutan	550	YMELGGLGERVLGFC**QLNLPSGKFPRGFKFDTDELNFPTEK**LCFVGLMS	598
Baboon	550	YMELGGLGERVLGFC**QLNLPSGKFPRGFKFDTDELNFPTEK**LCFVGLMS	598
Lemur	550	YLELGGLGERVLGFCHYYLPEEQFPKGFAFDCDDVNFTTDNLCFVGLMS	598
Saimiri	550	YLELGGLGERVLGFCHYYLPEEQFPKGFAFDCDDVNFTTDNLCFVGLMS	598
Aotus	550	YLELGGLGERVLGFCHYYLPEEQFPKGFAFDCDDVNFTTDNLCFVGLMS	598
Rat	550	YLELGGLGERVLGFCHYYLPEEQFPKGFAFDCDDVNFTTDNLCFVGLMS	598
Mouse	550	YLELGGLGERVLGFCHYYLPEEQFPKGFAFDCDDVNFTTDNLCFVGLMS	598

.

APPENDIX J

ACCESSION NUMBERS FOR ALL PROTEINS USED IN PHYLOGENETIC AND/OR STATISTICAL ANALYSES

Table A6. Accession numbers for all protein sequences used, categorized by analyses.

PROTEIN	ACCESSION NUMBER
PEX-Domain:	
Albumin_From_(Chickpea)	PDB: 3S18_A
Albumin-2_(Adzuki_bean)	XP_017413862.1
Albumin-2_(chickpea)	NP_001296633.1
Albumin-2_(Common_beet)	XP_010671707.1
Albumin-2_(Mung_bean)	XP_014492280.1
Albumin-2_(Adzuki_bean)	XP_017411691.1
Albumin-2_(Black_bean)	ADR30065.1
Albumin-2_(Rice)	XP_015634475.1
Albumin-2_(Barrel_medic)	XP_003590505.1
Albumin-2_(Common_beet)	XP_010667439.1
Albumin-2-like_isoX1_(Mung_bean)	XP_014512929.1
Albumin-2-like_isoX2_(Common_beet)	XP_010692751.1
Anther-specific_protein_(Garden_pea)	AAM12036.1
Chain_A_LS_24_(Grass_Pea)	PDB: 3LP9_A
HPX_(C)-dom_Homo	AAA58678.1
HPX_(N)-dom_Homo	AAA58678.1
HPX_C-dom_(little_skate)	ADU15818.1
HPX_C-dom_(nurse_shark)	ADU15819.1
Hpx_Fold_Protein_Cp4_(Cow_Pea)	PDB: 3OYO_A
HPX_N-dom_(little_skate)	ADU15818.1
HPX_N-dom_(nurse_shark)	ADU15819.1
Lectin_Cicer_(chickpea)	AGL46982.1
Limunectin_C-domain	XP_013779174.1
Limunectin_N-domain	XP_013779174.1

MMP1_HPX_dom_Homo	NP_002412.1
MMP10_PEX_dom_Homo	P09238.1
MMP11_PEX_dom_Homo	P24347.3
MMP12_PEX_dom_Homo	P39900.1
MMP13_PEX_dom_Homo	P45452.1
MMP14_PEX_dom_Homo	P50281.3
MMP15_PEX_dom_Homo	P51511.1
MMP16_HPX_dom_Homo	P51512.2
MMP17_HPX_dom_Homo	Q9ULZ9.4
MMP19_HPX_dom_Homo	Q99542.1
MMP2_HPX_dom_Homo	P08253.2
MMP20_HPX_dom_Homo	O60882.3
MMP21_HPX_dom_Homo	AAM78033.1
MMP24_HPX_dom_Homo	NP_006681.1
MMP27_HPX_dom_Homo	AAQ89112.1
MMP3_HPX_dom_Homo	NP_002413.1
MMP8_HPX_dom_Homo	P22894.1
MMP9_PEX_dom_Homo	P14780.3
Ostreopexin_Pleurotus	AHB89697.1
*P._luminescens*_photopexin:_Prokayote	OCA53183.1
Photosystem_II_10_kDa(Bread_wheat)	EMS53527.1
Photosystem_II_10_kDa_(African_oil_palm)	XP_010926934.1
Photosystem_II_10_kDa_(Date_palm)	XP_008799078.1
Photosystem_II_10_kDa_(Tauschs_Goatgrass)	XP_020201563.1
Photosystem_II_PsbR_protein_(Formosa_lily)	ACC38381.1
PRG4_PEX_dom._Homo	NP_001121180.2
VTN_PEX_dom._Homo	NP_000629.3
WAP65-1_C-dom_(catfish)	NP_001187177.1
WAP65-1_N-dom_(catfish)	NP_001187177.1
WAP65-2_C-dom_(catfish)	ABW07854.1
WAP65-2_N-dom_(catfish)	ABW07854.1

JEN-14:

Common Chimp HPX	XP_508255.1
Gorilla HPX	XP_004050644.1
Orangutan HPX	CAI29597.1
Gibbon HPX	XP_003254905.1
Mangabey HPX	XP_011891896.1
Pig HPX	XP_020957866.1
Cat HPX	XP_003992978.1
Rat HPX	NP_445770.1
Elephant HPX	XP_006885694.1

Mouse HPX	NP_059067.2
Guinea pig HPX	XP_003465366.2
Nekid mole rat HPX	XP_004863448.1
Flying fox HPX	XP_011383482.1
Rabbit HPX	NP_001076229.1
Giant Panda HPX	XP_019661247.1
Short tailed opossum HPX	XP_001380277.2
Chicken HPX	XP_015136422.1
Alligator HPX	XP_006037341.2
European bass WAP1	DAA12504.1
Atp1a3	
Human Atp1a3	NP_689509.1
Common chimp Atp1a3	XP_016789718.1
Bonobo Atp1a3	XP_003811938.1
Gorilla Atp1a3	XP_004027126.1
Orangutan Atp1a3	NP_001125304.1
Olive baboon Atp1a3	XP_009182965.1
Sooty mangabey Atp1a3	XP_011942294.1
Colobus Atp1a3	XP_011791957.1
Saimiri Atp1a3	XP_010329412.1
Aotus Atp1a3	XP_012316962.1
Lemur Atp1a3	XP_003799509.1
Rat Atip1a3	NP_036638.1
Mouse Atp1a3	EDL24296.1
Macaque Atp1a3	NP_001253401.1
DEAF1	
Human DEAF1	NP_066288.2
Common chimp DEAF1	XP_016775505.1
Bonobo DEAF1	XP_003805139.2
Gorilla DEAF1	XP_018891300.1
Orangutan DEAF1	XP_009244260.1
Olive baboon DEAF1	XP_003909385.2
Sooty mangabey DEAF1	XP_011896406.1
Colobus DEAF1	XP_011785530.1
Saimiri DEAF1	XP_010329533.1
Aotus DEAF1	XP_012311445.1
Lemur DEAF1	XP_012668771.1
Rat DEAF1	NP_113989.2
Mouse DEAF1	NP_058570.1
Macaque DEAF1	XP_014968734.1

REFERENCES

1. Jawad, Z.; Paoli, M., Novel sequences propel familiar folds. *Structure* **2002,** *10* (4), 447-54.
2. Kotu, A.; Guruprasad, K., The automatic detection of known beta-propeller structural motifs from protein tertiary structure. *Int J Biol Macromol* **2005,** *36* (3), 176-83.
3. Piccard, H.; Van den Steen, P. E.; Opdenakker, G., Hemopexin domains as multifunctional liganding modules in matrix metalloproteinases and other proteins. *J Leukoc Biol* **2007,** *81* (4), 870-92.
4. Kopec, K. O.; Lupas, A. N., β-Propeller blades as ancestral peptides in protein evolution. *PLoS One* **2013,** *8* (10), e77074.
5. Guruprasad, K.; Dhamayanthi, P., Structural plasticity associated with the beta-propeller architecture. *Int J Biol Macromol* **2004,** *34* (1-2), 55-61.
6. Nykjaer, A.; Willnow, T. E., Sortilin: a receptor to regulate neuronal viability and function. *Trends Neurosci* **2012,** *35* (4), 261-70.
7. Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E., The Protein Data Bank. *Nucleic Acids Res* **2000,** *28* (1), 235-42.
8. Hrkal, Z.; Vodrázka, Z.; Kalousek, I., Transfer of heme from ferrihemoglobin and ferrihemoglobin isolated chains to hemopexin. *Eur J Biochem* **1974,** *43* (1), 73-8.
9. Paoli, M.; Anderson, B. F.; Baker, H. M.; Morgan, W. T.; Smith, A.; Baker, E. N., Crystal structure of hemopexin reveals a novel high-affinity heme site formed between two beta-propeller domains. *Nat Struct Biol* **1999,** *6* (10), 926-31.
10. Baumann, H.; Gauldie, J., The acute phase response. *Immunol Today* **1994,** *15* (2), 74-80.
11. Smith, A.; McCulloh, R. J., Hemopexin and haptoglobin: allies against heme toxicity from hemoglobin not contenders. *Front Physiol* **2015,** *6*, 187.
12. Smith, A., Mechanisms of Cytoprotection by Hemopexin. In *Handbook of Porphyrin Science. Biochemistry of Tetrapyrroles.*, Kadish, K., Smith, KM, Guilard R, Ed. World Scientific Publishing Co. Pte. Ltd.: Singapore, 2011 pp 217-356.
13. Tolosano, E.; Altruda, F., Hemopexin: structure, function, and regulation. *DNA Cell Biol* **2002,** *21* (4), 297-306.
14. Balla, G.; Jacob, H. S.; Eaton, J. W.; Belcher, J. D.; Vercellotti, G. M., Hemin: a possible physiological mediator of low density lipoprotein oxidation and endothelial injury. *Arterioscler Thromb* **1991,** *11* (6), 1700-11.
15. Balla, G.; Vercellotti, G. M.; Muller-Eberhard, U.; Eaton, J.; Jacob, H. S., Exposure of endothelial cells to free heme potentiates damage mediated by granulocytes and toxic oxygen species. *Lab Invest* **1991,** *64* (5), 648-55.
16. Tolosano, E.; Fagoonee, S.; Morello, N.; Vinchi, F.; Fiorito, V., Heme scavenging and the other facets of hemopexin. *Antioxid Redox Signal* **2010,** *12* (2), 305-20.
17. Smith, A.; McCulloh, R., Mechanisms of haem toxicity in haemolysis and protection by the haem-binding protein, haemopexin. *ISBT Sci Ser* **2016,** *0*, 1-15.
18. Lee, B. C., Quelling the red menace: haem capture by bacteria. *Mol Microbiol* **1995,** *18* (3), 383-90.

19. Delanghe, J. R.; Langlois, M. R., Hemopexin: a review of biological aspects and the role in laboratory medicine. *Clin Chim Acta* **2001,** *312* (1-2), 13-23.
20. Smith, A.; Morgan, W. T., Hemopexin-mediated heme transport to the liver. Evidence for a heme-binding protein in liver plasma membranes. *J Biol Chem* **1985,** *260* (14), 8325-9.
21. Smith, A.; Morgan, W. T., Hemopexin-mediated transport of heme into isolated rat hepatocytes. *J Biol Chem* **1981,** *256* (21), 10902-9.
22. Hahl, P.; Davis, T.; Washburn, C.; Rogers, J. T.; Smith, A., Mechanisms of neuroprotection by hemopexin: modeling the control of heme and iron homeostasis in brain neurons in inflammatory states. *J Neurochem* **2013,** *125* (1), 89-101.
23. Eskew, J. D.; Vanacore, R. M.; Sung, L.; Morales, P. J.; Smith, A., Cellular protection mechanisms against extracellular heme. heme-hemopexin, but not free heme, activates the N-terminal c-jun kinase. *J Biol Chem* **1999,** *274* (2), 638-48.
24. Nagase, H.; Visse, R.; Murphy, G., Structure and function of matrix metalloproteinases and TIMPs. *Cardiovasc Res* **2006,** *69* (3), 562-73.
25. Scarafoni, A.; Gualtieri, E.; Barbiroli, A.; Carpen, A.; Negri, A.; Duranti, M., Biochemical and functional characterization of an albumin protein belonging to the hemopexin superfamily from Lens culinaris seeds. *J Agric Food Chem* **2011,** *59* (17), 9637-44.
26. Vigeolas, H.; Chinoy, C.; Zuther, E.; Blessington, B.; Geigenberger, P.; Domoney, C., Combined metabolomic and genetic approaches reveal a link between the polyamine pathway and albumin 2 in developing pea seeds. *Plant Physiol* **2008,** *146* (1), 74-82.
27. Gaur, V.; Qureshi, I. A.; Singh, A.; Chanana, V.; Salunke, D. M., Crystal structure and functional insights of hemopexin fold protein from grass pea. *Plant Physiol* **2010,** *152* (4), 1842-50.
28. Gaur, V.; Chanana, V.; Jain, A.; Salunke, D. M., The structure of a haemopexin-fold protein from cow pea (Vigna unguiculata) suggests functional diversity of haemopexins in plants. *Acta Crystallogr Sect F Struct Biol Cryst Commun* **2011,** *67* (Pt 2), 193-200.
29. Chattopadhyay, T.; Bhattacharyya, S.; Das, A. K.; Maiti, M. K., A structurally novel hemopexin fold protein of rice plays role in chlorophyll degradation. *Biochem Biophys Res Commun* **2012,** *420* (4), 862-8.
30. Crennell, S. J.; Tickler, P. M.; Bowen, D. J.; ffrench-Constant, R. H., The predicted structure of photopexin from Photorhabdus shows the first haemopexin-like motif in prokaryotes. *FEMS Microbiol Lett* **2000,** *191* (1), 139-44.
31. Ota, K.; Mikelj, M.; Papler, T.; Leonardi, A.; Križaj, I.; Maček, P., Ostreopexin: a hemopexin fold protein from the oyster mushroom, Pleurotus ostreatus. *Biochim Biophys Acta* **2013,** *1834* (8), 1468-73.
32. Geer, L. Y.; Marchler-Bauer, A.; Geer, R. C.; Han, L.; He, J.; He, S.; Liu, C.; Shi, W.; Bryant, S. H., The NCBI BioSystems database. *Nucleic Acids Res* **2010,** *38* (Database issue), D492-6.
33. Altschul, S. F.; Gish, W.; Miller, W.; Myers, E. W.; Lipman, D. J., Basic local alignment search tool. *J Mol Biol* **1990,** *215* (3), 403-10.
34. Baker, H. M.; Norris, G. E.; Morgan, W. T.; Smith, A.; Baker, E. N., Crystallization of the C-terminal domain of rabbit serum hemopexin. *J Mol Biol* **1993,** *229* (1), 251-2.

35. DeLano; W.L., The PyMOL Molecular Graphic System, 2002, Version 0.99rc6, Schrodinger, LLC. Copyright 2006, DeLano Scientific, LLC.
36. Morgan, W. T.; Smith, A., Domain structure of rabbit hemopexin. Isolation and characterization of a heme-binding glycopeptide. *J Biol Chem* **1984,** *259* (19), 12001-12006.
37. Faber, H. R.; Groom, C. R.; Baker, H. M.; Morgan, W. T.; Smith, A.; Baker, E. N., 1.8 Å crystal structure of the C-terminal domain of rabbit serum haemopexin. *Structure* **1995,** *3* (6), 551-559.
38. Baker, H. M.; Anderson, B. F.; Baker, E. N., Dealing with iron: common structural principles in proteins that transport iron and heme. *Proc Natl Acad Sci U S A* **2003,** *100* (7), 3579-83.
39. Li, W. H.; Gu, Z.; Cavalcanti, A. R.; Nekrutenko, A., Detection of gene duplications and block duplications in eukaryotic genomes. *J Struct Funct Genomics* **2003,** *3* (1-4), 27-34.
40. Green, R. E.; Krause, J.; Briggs, A. W.; Maricic, T.; Stenzel, U.; Kircher, M.; Patterson, N.; Li, H.; Zhai, W.; Fritz, M. H.; Hansen, N. F.; Durand, E. Y.; Malaspinas, A. S.; Jensen, J. D.; Marques-Bonet, T.; Alkan, C.; Prüfer, K.; Meyer, M.; Burbano, H. A.; Good, J. M.; Schultz, R.; Aximu-Petri, A.; Butthof, A.; Höber, B.; Höffner, B.; Siegemund, M.; Weihmann, A.; Nusbaum, C.; Lander, E. S.; Russ, C.; Novod, N.; Affourtit, J.; Egholm, M.; Verna, C.; Rudan, P.; Brajkovic, D.; Kucan, Z.; Gusic, I.; Doronichev, V. B.; Golovanova, L. V.; Lalueza-Fox, C.; de la Rasilla, M.; Fortea, J.; Rosas, A.; Schmitz, R. W.; Johnson, P. L.; Eichler, E. E.; Falush, D.; Birney, E.; Mullikin, J. C.; Slatkin, M.; Nielsen, R.; Kelso, J.; Lachmann, M.; Reich, D.; Pääbo, S., A draft sequence of the Neandertal genome. *Science* **2010,** *328* (5979), 710-22.
41. Neanderthal Genome Project - Transcript: HPX-001 (ENST00000265983). http://neandertal.ensemblgenomes.org/Homo_sapiens/Transcript/ProteinSummary?db=core.
42. Homo sapiens ssp. Denisova. https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?lvl=0&id=741158.
43. Meyer, M.; Kircher, M.; Gansauge, M. T.; Li, H.; Racimo, F.; Mallick, S.; Schraiber, J. G.; Jay, F.; Prüfer, K.; de Filippo, C.; Sudmant, P. H.; Alkan, C.; Fu, Q.; Do, R.; Rohland, N.; Tandon, A.; Siebauer, M.; Green, R. E.; Bryc, K.; Briggs, A. W.; Stenzel, U.; Dabney, J.; Shendure, J.; Kitzman, J.; Hammer, M. F.; Shunkov, M. V.; Derevianko, A. P.; Patterson, N.; Andrés, A. M.; Eichler, E. E.; Slatkin, M.; Reich, D.; Kelso, J.; Pääbo, S., A high-coverage genome sequence from an archaic Denisovan individual. *Science* **2012,** *338* (6104), 222-6.
44. Hvidberg, V.; Maniecki, M. B.; Jacobsen, C.; Højrup, P.; Møller, H. J.; Moestrup, S. K., Identification of the receptor scavenging hemopexin-heme complexes. *Blood* **2005,** *106* (7), 2572-9.
45. Morgan, W. T.; Muster, P.; Tatum, F. M.; McConnell, J.; Conway, T. P.; Hensley, P.; Smith, A., Use of hemopexin domains and monoclonal antibodies to hemopexin to probe the molecular determinants of hemopexin-mediated heme transport. *J Biol Chem* **1988,** *263* (17), 8220-5.

46. Morgan, W. T.; Muster, P.; Tatum, F.; Kao, S. M.; Alam, J.; Smith, A., Identification of the histidine residues of hemopexin that coordinate with heme-iron and of a receptor-binding region. *J Biol Chem* **1993,** *268* (9), 6256-62.
47. Sherry, S. T.; Ward, M. H.; Kholodov, M.; Baker, J.; Phan, L.; Smigielski, E. M.; Sirotkin, K., dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res* **2001,** *29* (1), 308-11.
48. Khurana, E.; Fu, Y.; Colonna, V.; Mu, X. J.; Kang, H. M.; Lappalainen, T.; Sboner, A.; Lochovsky, L.; Chen, J.; Harmanci, A.; Das, J.; Abyzov, A.; Balasubramanian, S.; Beal, K.; Chakravarty, D.; Challis, D.; Chen, Y.; Clarke, D.; Clarke, L.; Cunningham, F.; Evani, U. S.; Flicek, P.; Fragoza, R.; Garrison, E.; Gibbs, R.; Gümüs, Z. H.; Herrero, J.; Kitabayashi, N.; Kong, Y.; Lage, K.; Liluashvili, V.; Lipkin, S. M.; MacArthur, D. G.; Marth, G.; Muzny, D.; Pers, T. H.; Ritchie, G. R.; Rosenfeld, J. A.; Sisu, C.; Wei, X.; Wilson, M.; Xue, Y.; Yu, F.; Dermitzakis, E. T.; Yu, H.; Rubin, M. A.; Tyler-Smith, C.; Gerstein, M.; Consortium, G. P., Integrative annotation of variants from 1092 humans: application to cancer genomics. *Science* **2013,** *342* (6154), 1235587.
49. Mauk, M. R.; Smith, A.; Mauk, A. G., An alternative view of the proposed alternative activities of hemopexin. *Protein Sci* **2011,** *20* (5), 791-805.
50. Xu, D.; Baburaj, K.; Peterson, C. B.; Xu, Y., Model for the three-dimensional structure of vitronectin: predictions for the multi-domain protein from threading and docking. *Proteins* **2001,** *44* (3), 312-20.
51. Pieper, U.; Webb, B. M.; Dong, G. Q.; Schneidman-Duhovny, D.; Fan, H.; Kim, S. J.; Khuri, N.; Spill, Y. G.; Weinkam, P.; Hammel, M.; Tainer, J. A.; Nilges, M.; Sali, A., ModBase, a database of annotated comparative protein structure models and associated resources. *Nucleic Acids Res* **2014,** *42* (Database issue), D336-46.
52. Schvartz, I.; Seger, D.; Shaltiel, S., Vitronectin. *Int J Biochem Cell Biol* **1999,** *31* (5), 539-44.
53. Sternlicht, M. D.; Werb, Z., How matrix metalloproteinases regulate cell behavior. *Annu Rev Cell Dev Biol* **2001,** *17*, 463-516.
54. Brew, K.; Dinakarpandian, D.; Nagase, H., Tissue inhibitors of metalloproteinases: evolution, structure and function. *Biochim Biophys Acta* **2000,** *1477* (1-2), 267-83.
55. Machado, J. P.; Vasconcelos, V.; Antunes, A., Adaptive functional divergence of the warm temperature acclimation-related protein (WAP65) in fishes and the ortholog hemopexin (HPX) in mammals. *J Hered* **2014,** *105* (2), 237-52.
56. Dooley, H.; Buckingham, E. B.; Criscitiello, M. F.; Flajnik, M. F., Emergence of the acute-phase protein hemopexin in jawed vertebrates. *Mol Immunol* **2010,** *48* (1-3), 147-52.
57. Kinoshita, S., Itoi, S. & Watabe, S., cDNA cloning and characterization of the warm-temperature-acclimation-associated protein Wap65 from carp, Cyprinus carpio. *Fish Physiology and Biochemistry* **(2001),** *24* (125.).
58. Hirayama, M.; Nakaniwa, M.; Ikeda, D.; Hirazawa, N.; Otaka, T.; Mitsuboshi, T.; Shirasu, K.; Watabe, S., Primary structures and gene organizations of two types of Wap65 from the pufferfish Takifugu rubripes. *Fish Physiol Biochem* **2003,** *29* (3), 211-224.
59. Nakaniwa, M.; Hirayama, M.; Shimizu, A.; Sasaki, T.; Asakawa, S.; Shimizu, N.; Watabe, S., Genomic sequences encoding two types of medaka hemopexin-like protein

Wap65, and their gene expression profiles in embryos. *J Exp Biol* **2005,** *208* (Pt 10), 1915-25.
60. Aliza, D.; Ismail, I. S.; Kuah, M.-K.; Shu-Chien, A. C.; Muhammad, T. S. T., Identification of Wap65, a human homologue of hemopexin as a copper-inducible gene in swordtail fish, Xiphophorus helleri. *Fish physiology and biochemistry* **2008,** *34* (2), 129-138 %@ 0920-1742.
61. Choi, C. Y.; An, K. W.; Choi, Y. K.; Jo, P. G.; Min, B. H., Expression of warm temperature acclimation-related protein 65-kDa (Wap65) mRNA, and physiological changes with increasing water temperature in black porgy, Acanthopagrus schlegeli. *Journal of Experimental Zoology Part A: Ecological Genetics and Physiology* **2008,** *309* (4), 206-214 %@ 1932-5231.
62. Takano, T.; Sha, Z.; Peatman, E.; Terhune, J.; Liu, H.; Kucuktas, H.; Li, P.; Edholm, E.-S.; Wilson, M.; Liu, Z., The two channel catfish intelectin genes exhibit highly differential patterns of tissue expression and regulation after infection with Edwardsiella ictaluri. *Developmental & Comparative Immunology* **2008,** *32* (6), 693-705 %@ 0145-305X.
63. Sha, Z.; Xu, P.; Takano, T.; Liu, H.; Terhune, J.; Liu, Z., The warm temperature acclimation protein Wap65 as an immune response gene: its duplicates are differentially regulated by temperature and bacterial infections. *Molecular immunology* **2008,** *45* (5), 1458-1469 %@ 0161-5890.
64. Cherayil, B. J., The role of iron in the immune response to bacterial infection. *Immunologic research* **2011,** *50* (1), 1-9 %@ 0257-277X.
65. Kikuchi, K.; Watabe, S.; Aida, K., Isolation of a 65-kDa protein from white muscle of warm temperature-acclimated goldfish (Carassius auratus). *Comparative Biochemistry and Physiology Part B: Biochemistry and Molecular Biology* **1998,** *120* (2), 385-391 %@ 1096-4959.
66. Hirayama, M.; Kobiyama, A.; Kinoshita, S.; Watabe, S., The occurrence of two types of hemopexin-like protein in medaka and differences in their affinity to heme. *J Exp Biol* **2004,** *207* (Pt 8), 1387-98.
67. Peatman, E.; Baoprasertkul, P.; Terhune, J.; Xu, P.; Nandi, S.; Kucuktas, H.; Li, P.; Wang, S.; Somridhivej, B.; Dunham, R., Expression analysis of the acute phase response in channel catfish (Ictalurus punctatus) after infection with a Gram-negative bacterium. *Developmental & Comparative Immunology* **2007,** *31* (11), 1183-1196 %@ 0145-305X.
68. Peatman, E.; Terhune, J.; Baoprasertkul, P.; Xu, P.; Nandi, S.; Wang, S.; Somridhivej, B.; Kucuktas, H.; Li, P.; Dunham, R., Microarray analysis of gene expression in the blue catfish liver reveals early activation of the MHC class I pathway after infection with Edwardsiella ictaluri. *Molecular immunology* **2008,** *45* (2), 553-566 %@ 0161-5890.
69. Shi, Y. H.; Chen, J.; Li, C. H.; Li, M. Y., Molecular cloning of liver Wap65 cDNA in ayu (Plecoglossus altivelis) and mRNA expression changes following Listonella anguillarum infection. *Molecular biology reports* **2010,** *37* (3), 1523-1529 %@ 0301-4851.
70. Ohta, T., Role of gene duplication in evolution. *Genome* **1989,** *31* (1), 304-310 %@ 0831-2796.
71. Sarropoulou, E.; Fernandes, J. M. O.; Mitter, K.; Magoulas, A.; Mulero, V.; Sepulcre, M. P.; Figueras, A.; Novoa, B.; Kotoulas, G., Evolution of a multifunctional gene: the

warm temperature acclimation protein Wap65 in the European seabass Dicentrarchus labrax. *Molecular phylogenetics and evolution* **2010,** *55* (2), 640-649 %@ 1055-7903.
72. Liu, T.; Lin, Y.; Cislo, T.; Minetti, C. A.; Baba, J. M.; Liu, T. Y., Limunectin. A phosphocholine-binding protein from Limulus amebocytes with adhesion-promoting properties. *J Biol Chem* **1991,** *266* (22), 14813-21.
73. Liu, T. Y.; Minetti, C. A.; Fortes-Dias, C. L.; Liu, T.; Lin, L.; Lin, Y., C-reactive proteins, limunectin, lipopolysaccharide-binding protein, and coagulin. Molecules with lectin and agglutinin activities from Limulus polyphemus. *Ann N Y Acad Sci* **1994,** *712*, 146-54.
74. Zhang, Y., I-TASSER server for protein 3D structure prediction. *BMC Bioinformatics* **2008,** *9*, 40.
75. Roy, A.; Kucukural, A.; Zhang, Y., I-TASSER: a unified platform for automated protein structure and function prediction. *Nat Protoc* **2010,** *5* (4), 725-38.
76. Yang, J.; Yan, R.; Roy, A.; Xu, D.; Poisson, J.; Zhang, Y., The I-TASSER Suite: protein structure and function prediction. *Nat Meth* **2015,** *12* (1), 7-8.
77. Rehm, P.; Pick, C.; Borner, J.; Markl, J.; Burmester, T., The diversity and evolution of chelicerate hemocyanins. *BMC Evol Biol* **2012,** *12*, 19.
78. Rudkin, D. M.; Young, G. A.; Nowlan, G. S., The oldest horseshoe crab: a new Xiphosurid from late Ordovician konservat-lagerstätten deposits, Manitoba, Canada. *Palaeontology* **2008,** *51* (1), 1-9.
79. Jenne, D., Homology of placental protein 11 and pea seed albumin 2 with vitronectin. *Biochem Biophys Res Commun* **1991,** *176* (3), 1000-6.
80. Rao, A. U.; Carta, L. K.; Lesuisse, E.; Hamza, I., Lack of heme synthesis in a free-living eukaryote. *Proc Natl Acad Sci U S A* **2005,** *102* (12), 4270-5.
81. Kumar, S.; Stecher, G.; Tamura, K., MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. *Mol Biol Evol* **2016,** *33* (7), 1870-4.
82. Jones, D. T.; Taylor, W. R.; Thornton, J. M., The rapid generation of mutation data matrices from protein sequences. *Comput Appl Biosci* **1992,** *8* (3), 275-82.
83. Felsenstein, J., Confidence limits on phylogenies: An approach using the bootstrap. *Evolution* **1985,** *39*, 783-791.
84. Faivre, L.; Prieur, A. M.; Le Merrer, M.; Hayem, F.; Penet, C.; Woo, P.; Hofer, M.; Dagoneau, N.; Sermet, I.; Munnich, A.; Cormier-Daire, V., Clinical variability and genetic homogeneity of the camptodactyly-arthropathy-coxa vara-pericarditis syndrome. *Am J Med Genet* **2000,** *95* (3), 233-6.
85. Marcelino, J.; Carpten, J. D.; Suwairi, W. M.; Gutierrez, O. M.; Schwartz, S.; Robbins, C.; Sood, R.; Makalowska, I.; Baxevanis, A.; Johnstone, B.; Laxer, R. M.; Zemel, L.; Kim, C. A.; Herd, J. K.; Ihle, J.; Williams, C.; Johnson, M.; Raman, V.; Alonso, L. G.; Brunoni, D.; Gerstein, A.; Papadopoulos, N.; Bahabri, S. A.; Trent, J. M.; Warman, M. L., CACP, encoding a secreted proteoglycan, is mutated in camptodactyly-arthropathy-coxa vara-pericarditis syndrome. *Nat Genet* **1999,** *23* (3), 319-22.
86. Ciullini Mannurita, S.; Vignoli, M.; Bianchi, L.; Kondi, A.; Gerloni, V.; Breda, L.; Ten Cate, R.; Alessio, M.; Ravelli, A.; Falcini, F.; Gambineri, E., CACP syndrome: identification of five novel mutations and of the first case of UPD in the largest European cohort. *Eur J Hum Genet* **2014,** *22* (2), 197-201.

87. Verma, R. P.; Hansch, C., Matrix metalloproteinases (MMPs): chemical-biological functions and (Q)SARs. *Bioorg Med Chem* **2007,** *15* (6), 2223-68.
88. Kessenbrock, K.; Plaks, V.; Werb, Z., Matrix metalloproteinases: regulators of the tumor microenvironment. *Cell* **2010,** *141* (1), 52-67.
89. Radisky, E. S.; Radisky, D. C., Stromal induction of breast cancer: inflammation and invasion. *Rev Endocr Metab Disord* **2007,** *8* (3), 279-87.
90. Radisky, E. S.; Radisky, D. C., Matrix metalloproteinase-induced epithelial-mesenchymal transition in breast cancer. *J Mammary Gland Biol Neoplasia* **2010,** *15* (2), 201-12.
91. Wieczorek, E.; Jablonska, E.; Wasowicz, W.; Reszka, E., Matrix metalloproteinases and genetic mouse models in cancer research: a mini-review. *Tumour Biol* **2015,** *36* (1), 163-75.
92. Said, A. H.; Raufman, J. P.; Xie, G., The role of matrix metalloproteinases in colorectal cancer. *Cancers (Basel)* **2014,** *6* (1), 366-75.
93. Hua, H.; Li, M.; Luo, T.; Yin, Y.; Jiang, Y., Matrix metalloproteinases in tumorigenesis: an evolving paradigm. *Cell Mol Life Sci* **2011,** *68* (23), 3853-68.
94. Malemud, C. J., Matrix metalloproteinases (MMPs) in health and disease: an overview. *Front Biosci* **2006,** *11*, 1696-701.
95. Coussens, L. M.; Fingleton, B.; Matrisian, L. M., Matrix metalloproteinase inhibitors and cancer: trials and tribulations. *Science* **2002,** *295* (5564), 2387-92.
96. Zhu, L.; Hope, T. J.; Hall, J.; Davies, A.; Stern, M.; Muller-Eberhard, U.; Stern, R.; Parslow, T. G., Molecular cloning of a mammalian hyaluronidase reveals identity with hemopexin, a serum heme-binding protein. *J Biol Chem* **1994,** *269* (51), 32092-7.
97. Bakker, W. W.; Borghuis, T.; Harmsen, M. C.; van den Berg, A.; Kema, I. P.; Niezen, K. E.; Kapojos, J. J., Protease activity of plasma hemopexin. *Kidney Int* **2005,** *68* (2), 603-10.
98. Suzuki, K.; Kato, H.; Sakuma, Y.; Namiki, H., Hemopexins suppress phorbol ester-induced necrosis of polymorphonuclear leucocytes. *Cell Struct Funct* **2001,** *26* (4), 235-41.
99. Suzuki, K.; Kobayashi, N.; Doi, T.; Hijikata, T.; Machida, I.; Namiki, H., Inhibition of Mg2+-dependent adhesion of polymorphonuclear leukocytes by serum hemopexin: differences in divalent-cation dependency of cell adhesion in the presence and absence of serum. *Cell Struct Funct* **2003,** *28* (4), 243-53.
100. Hernández, C.; Garcia-Ramírez, M.; Simó, R., Overexpression of hemopexin in the diabetic eye: a new pathogenic candidate for diabetic macular edema. *Diabetes Care* **2013,** *36* (9), 2815-21.
101. Jeney, V.; Balla, J.; Yachie, A.; Varga, Z.; Vercellotti, G. M.; Eaton, J. W.; Balla, G., Pro-oxidant and cytotoxic effects of circulating heme. *Blood* **2002,** *100* (3), 879-87.
102. Altruda, F.; Poli, V.; Restagno, G.; Argos, P.; Cortese, R.; Silengo, L., The primary structure of human hemopexin deduced from cDNA sequence: evidence for internal, repeating homology. *Nucleic Acids Res* **1985,** *13* (11), 3841-59.
103. Jenne, D.; Stanley, K. K., Nucleotide sequence and organization of the human S-protein gene: repeating peptide motifs in the "pexin" family and a model for their evolution. *Biochemistry* **1987,** *26* (21), 6735-42.

104. Iwasaki, K.; Morimatsu, M.; Inanami, O.; Uchida, E.; Syuto, B.; Kuwabara, M.; Niiyama, M., Isolation, characterization, and cDNA cloning of chicken turpentine-induced protein, a new member of the scavenger receptor cysteine-rich (SRCR) family of proteins. *J Biol Chem* **2001,** *276* (12), 9400-5.
105. Wicher, K. B.; Fries, E., Evolutionary aspects of hemoglobin scavengers. *Antioxid Redox Signal* **2010,** *12* (2), 249-59.
106. Hardison, R. C., A brief history of hemoglobins: plant, animal, protist, and bacteria. *Proc Natl Acad Sci U S A* **1996,** *93* (12), 5675-9.
107. Yang, M. a. W. G. *Dissertation,* Evolutionary Analysis of Hemopexin Domains. University of Missouri Kansas City, Kansas City, Missouri. , 2010.
108. Gould, S. J.; Lewontin, R. C., The spandrels of San Marco and the Panglossian paradigm: a critique of the adaptationist programme. *Proc R Soc Lond B Biol Sci* **1979,** *205* (1161), 581-98.
109. Gould, S. J.; Vrba, E. S., Exaptation-A Missing Term in the Science of Form. *Paleobiology* **1982,** *Vol. 8* (1), 4-15.
110. Ruhfel, B. R.; Gitzendanner, M. A.; Soltis, P. S.; Soltis, D. E.; Burleigh, J. G., From algae to angiosperms-inferring the phylogeny of green plants (Viridiplantae) from 360 plastid genomes. *BMC Evol Biol* **2014,** *14*, 23.
111. Ren, Q.; Wang, C.; Jin, M.; Lan, J.; Ye, T.; Hui, K.; Tan, J.; Wang, Z.; Wyckoff, G. J.; Wang, W.; Han, G. Z., Co-option of bacteriophage lysozyme genes by bivalve genomes. *Open Biol* **2017,** *7* (1).
112. Aminov, R. I., Horizontal gene exchange in environmental microbiota. *Front Microbiol* **2011,** *2*, 158.
113. Wijayawardena, B. K.; Minchella, D. J.; DeWoody, J. A., Hosts, parasites, and horizontal gene transfer. *Trends Parasitol* **2013,** *29* (7), 329-38.
114. Davis, C. C.; Xi, Z., Horizontal gene transfer in parasitic plants. *Curr Opin Plant Biol* **2015,** *26*, 14-9.
115. Mitreva, M.; Smant, G.; Helder, J., Role of horizontal gene transfer in the evolution of plant parasitism among nematodes. *Methods Mol Biol* **2009,** *532*, 517-35.
116. Mayer, W. E.; Schuster, L. N.; Bartelmes, G.; Dieterich, C.; Sommer, R. J., Horizontal gene transfer of microbial cellulases into nematode genomes is associated with functional assimilation and gene turnover. *BMC Evol Biol* **2011,** *11*, 13.
117. Pigliucci, I.; Kaplan, I., The fall and rise of Dr Pangloss: adaptationism and the Spandrels paper 20 years later. *Trends Ecol Evol* **2000,** *15* (2), 66-70.
118. Jacob, F., Evolution and tinkering. *Science* **1977,** *196* (4295), 1161-6.
119. Thorn, R. G.; Barron, G. L., Carnivorous mushrooms. *Science* **1984,** *224* (4644), 76-8.
120. Xiao, Q.; Ma, F.; Li, Y.; Yu, H.; Li, C.; Zhang, X., Differential Proteomic Profiles of Pleurotus ostreatus in Response to Lignocellulosic Components Provide Insights into Divergent Adaptive Mechanisms. *Front Microbiol* **2017,** *8*, 480.
121. Manhart, M.; Morozov, A. V., Protein folding and binding can emerge as evolutionary spandrels through structural coupling. *Proc Natl Acad Sci U S A* **2015,** *112* (6), 1797-802.
122. Swann, D. A.; Slayter, H. S.; Silver, F. H., The molecular structure of lubricating glycoprotein-I, the boundary lubricant for articular cartilage. *J Biol Chem* **1981,** *256* (11), 5921-5.

123. Su, J. L.; Schumacher, B. L.; Lindley, K. M.; Soloveychik, V.; Burkhart, W.; Triantafillou, J. A.; Kuettner, K.; Schmid, T., Detection of superficial zone protein in human and animal body fluids by cross-species monoclonal antibodies specific to superficial zone protein. *Hybridoma* **2001,** *20* (3), 149-57.
124. OMIM, Online Mendelian Inheritance in Man, OMIM ® . McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University (Baltimore, MD), {date}. World Wide Web URL: https://omim.org/.
125. Bahabri, S. A.; Suwairi, W. M.; Laxer, R. M.; Polinkovsky, A.; Dalaan, A. A.; Warman, M. L., The camptodactyly-arthropathy-coxa vara-pericarditis syndrome: clinical features and genetic mapping to human chromosome 1. *Arthritis Rheum* **1998,** *41* (4), 730-5.
126. Martínez-Lavín, M.; Buendía, A.; Delgado, E.; Reyes, P.; Amigo, M. C.; Sabanés, J.; Zghaib, A.; Attíe, F.; Salinas, L., A familial syndrome of pericarditis, arthritis, and camptodactyly. *N Engl J Med* **1983,** *309* (4), 224-5.
127. Bulutlar, G.; Yazici, H.; Ozdoğan, H.; Schreuder, I., A familial syndrome of pericarditis, arthritis, camptodactyly, and coxa vara. *Arthritis Rheum* **1986,** *29* (3), 436-8.
128. Wetlaufer, D. B., Nucleation, rapid folding, and globular intrachain regions in proteins. *Proc Natl Acad Sci U S A* **1973,** *70* (3), 697-701.
129. Edgar, R.; Domrachev, M.; Lash, A. E., Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res* **2002,** *30* (1), 207-10.
130. Rhee, D. K.; Marcelino, J.; Baker, M.; Gong, Y.; Smits, P.; Lefebvre, V.; Jay, G. D.; Stewart, M.; Wang, H.; Warman, M. L.; Carpten, J. D., The secreted glycoprotein lubricin protects cartilage surfaces and inhibits synovial cell overgrowth. *J Clin Invest* **2005,** *115* (3), 622-31.
131. Rhee, D. K.; Marcelino, J.; Al-Mayouf, S.; Schelling, D. K.; Bartels, C. F.; Cui, Y.; Laxer, R.; Goldbach-Mansky, R.; Warman, M. L., Consequences of disease-causing mutations on lubricin protein synthesis, secretion, and post-translational processing. *J Biol Chem* **2005,** *280* (35), 31325-32.
132. Bode, W., A helping hand for collagenases: the haemopexin-like domain. *Structure* **1995,** *3* (6), 527-30.
133. Yoneda, A.; Ogawa, H.; Kojima, K.; Matsumoto, I., Characterization of the ligand binding activities of vitronectin: interaction of vitronectin with lipids and identification of the binding domains for various ligands using recombinant domains. *Biochemistry* **1998,** *37* (18), 6351-60.
134. Das, S.; Mandal, M.; Chakraborti, T.; Mandal, A.; Chakraborti, S., Structure and evolutionary aspects of matrix metalloproteinases: a brief overview. *Mol Cell Biochem* **2003,** *253* (1-2), 31-40.
135. Sjölander, K., Getting started in structural phylogenomics. *PLoS Comput Biol* **2010,** *6* (1), e1000621.
136. Yang, M.; Wyckoff, G. J., Detection of selection utilizing molecular phylogenetics: a possible approach. *Genetica* **2011,** *139* (5), 639-48.
137. Cooper, M. D.; Alder, M. N., The evolution of adaptive immune systems. *Cell* **2006,** *124* (4), 815-22.
138. Parks, W. C.; Wilson, C. L.; López-Boado, Y. S., Matrix metalloproteinases as modulators of inflammation and innate immunity. *Nat Rev Immunol* **2004,** *4* (8), 617-29.

139. Shipulina, N.; Smith, A.; Morgan, W. T., Heme binding by hemopexin: evidence for multiple modes of binding and functional implications. *J Protein Chem* **2000,** *19* (3), 239-48.
140. Ali, L.; Flowers, S. A.; Jin, C.; Bennet, E. P.; Ekwall, A. K.; Karlsson, N. G., The O-glycomap of lubricin, a novel mucin responsible for joint lubrication, identified by site-specific glycopeptide analysis. *Mol Cell Proteomics* **2014,** *13* (12), 3396-409.
141. Wüthrich, K., NMR studies of structure and function of biological macromolecules (Nobel Lecture). *J Biomol NMR* **2003,** *27* (1), 13-39.
142. Sugiki, T.; Yoshiura, C.; Kofuku, Y.; Ueda, T.; Shimada, I.; Takahashi, H., High-throughput screening of optimal solution conditions for structural biological studies by fluorescence correlation spectroscopy. *Protein Sci* **2009,** *18* (5), 1115-20.
143. IDT gBlock; https://www.idtdna.com/pages/products/genes/gblocks-gene-fragments.
144. Geisbrecht, B. V.; Bouyain, S.; Pop, M., An optimized system for expression and purification of secreted bacterial proteins. *Protein Expr Purif* **2006,** *46* (1), 23-32.
145. 71122 | pET-44a(+) DNA - Novagen. http://www.emdmillipore.com/US/en/product/pET-44a%28%2B%29-DNA---Novagen,EMD_BIO-71122 (accessed 03.19.2017).
146. Sigmund, C. D.; Morgan, E. A., Nus A protein affects transcriptional pausing and termination in vitro by binding to different sites on the transcription complex. *Biochemistry* **1988,** *27* (15), 5622-7.
147. Davis, G. D.; Elisee, C.; Newham, D. M.; Harrison, R. G., New fusion protein systems designed to give soluble expression in Escherichia coli. *Biotechnol Bioeng* **1999,** *65* (4), 382-8.
148. Ginzinger_DG_et_al. Quantitative PCR method to enumerate DNA copy number; U.S. Patent No. 6,180,349. 2001.
149. Midsci- 2.0ml_PCR_tubes;. http://shop.midsci.com/productdetail/M50/AVSST/Pryme_PCR_8-Strip_Tubes,_0.2ml,_w/_Attached_Dome_Caps,_Natural,_120/pk/.
150. PerkinElmer_GeneAmp_2400_PCR_system http://www.perkinelmer.com/lab-solutions.
151. Lundberg, K. S.; Shoemaker, D. D.; Adams, M. W.; Short, J. M.; Sorge, J. A.; Mathur, E. J., High-fidelity amplification using a thermostable DNA polymerase isolated from Pyrococcus furiosus. *Gene* **1991,** *108* (1), 1-6.
152. Varadaraj, K.; Skinner, D. M., Denaturants or cosolvents improve the specificity of PCR amplification of a G + C-rich DNA using genetically engineered DNA polymerases. *Gene* **1994,** *140* (1), 1-5.
153. Midsci-Agarose. http://shop.midsci.com/productdetail/M50/BE-A500/Bullseye_agarose_General_Purpose,_500g,_GP2_standard_electrophoresis/.
154. Crisafuli, F. A.; Ramos, E. B.; Rocha, M. S., Characterizing the interaction between DNA and GelRed fluorescent stain. *Eur Biophys J* **2015,** *44* (1-2), 1-7.
155. Thermo Scientific™ a subsidiary of Fisher Scientific; https://www.fishersci.com/shop/products/thermo-scientific-0-2-ml-individual-tubes-14/p-2866179.
156. Promega Corporation: Wizard SV Gel and PCR Clean-up System – A9281. https://www.promega.com/resources/protocols/technical-bulletins/101/wizard-sv-gel-and-pcr-cleanup-system-protocol/.

157. New England BioLabs (NEB, Inc.) - Product NEBuffer 2. https://www.neb.com/products/b7002-nebuffer-2.
158. Promega Corporation, Madison, Wisconsin. https://www.promega.com/-/media/files/resources/protocols/product-information-sheets/g/t4-dna-ligase-blue-white-cloning-qualified-protocol.pdf?la=en.
159. Eppendorf Biotech Company, Hamburg Germany. https://www.eppendorf.com/worldwide/.
160. Thermo Fisher Scientific: MAX Efficiency™ DH5α™ Competent Cells. https://tools.thermofisher.com/content/sfs/manuals/18258012.pdf.
161. MidSci: Midwest Scientific, Saint Louis, Missouri. http://shop.midsci.com/productdetail/M50/IB47101/IBI_MINI_Hi-Speed_Plasmid_Kit,_100_preps/.
162. University of Missouri DNA Sequencing Core Facility, Columbia, MO. http://dnacore.missouri.edu//.
163. M-110L Laboratory Microfluidizer Processor. https://www.bioprocessonline.com/doc/m-110l-laboratory-microfluidizer-processors-h-0001.
164. N*i* Sepharose 6 Fast Flow is a BioProcess medium. http://www.gelifesciences.com/webapp/wcs/stores/servlet/productById/en/GELifeSciences-us/17531801.
165. Thermo Scientific™ PageRuler™ Plus Prestained 10-250kDa Protein Ladder. https://www.fishersci.com/shop/products/pageruler-plus-prestained-10-250kda-protein-ladder/p-4529974.
166. NusA, 1-495aa, E.coli. http://atgenglobal.com/NuS0801.
167. Smith, A.; Rish, K. R.; Lovelace, R.; Hackney, J. F.; Helston, R. M., Role for copper in the cellular and regulatory effects of heme-hemopexin. *Biometals* **2009,** *22* (3), 421-37.
168. Illustra MicroSpin G-50 Columns. http://www.gelifesciences.com/webapp/wcs/stores/servlet/catalog/en/GELifeSciences-us/products/AlternativeProductStructure_17513/27533001.
169. Quick Start™ Bradford Protein Assay. http://www.bio-rad.com/en-us/product/quick-start-bradford-protein-assay?WT.mc_id=170125005456&WT.srch=1&WT.knsh_id=f648167d-144d-4fe1-bd38-7c40fa0cb47c.
170. Cleavage Enzymes for Fusion Protein Purification. https://www.emdmillipore.com/US/en/product/Recombinant-Enterokinase,EMD_BIO-69066.
171. Thermo Scientific™ PageRuler™ Unstained Low Range Protein Ladder. https://www.fishersci.com/shop/products/pageruler-unstained-low-range-protein-ladder/p-4529997.
172. Proteomics and Mass Spectrometry Facility, School of Biological Sciences - UMKC. https://sbs.umkc.edu/research/facilities/.
173. Sequencing Grade Modified Porcine Trypsin. https://www.promega.com/products/mass-spectrometry/proteases-and-surfactants/sequencing-grade-modified-trypsin/?catNum=V5111.

174. Canterbury, J. D.; Yi, X.; Hoopmann, M. R.; MacCoss, M. J., Assessing the dynamic range and peak capacity of nanoflow LC-FAIMS-MS on an ion trap mass spectrometer for proteomics. *Anal Chem* **2008,** *80* (18), 6888-97.
175. Q Exactive™ Plus Hybrid Quadrupole-Orbitrap™ Mass Spectrometer. https://www.thermofisher.com/order/catalog/product/IQLAAEGAAPFALGMBDK?ICID=search-product.
176. Mascot software. http://www.matrixscience.com/.
177. UniProtKB/Swiss-Prot. http://web.expasy.org/docs/swiss-prot_guideline.html.
178. Fenton, W. A.; Horwich, A. L., GroEL-mediated protein folding. *Protein Sci* **1997,** *6* (4), 743-60.
179. Wang, F.; Min, Y.; Geng, X., Fast separations of intact proteins by liquid chromatography. *J Sep Sci* **2012,** *35* (22), 3033-45.
180. Superdex® 200 Increase 10/300 GL. http://www.sigmaaldrich.com/catalog/product/sigma/ge28990944?lang=en&region=US&cm_sp=Insite-_-prodRecCold_xorders-_-prodRecCold2-1.
181. HBS (HEPES buffered saline). http://www2.kumc.edu/soalab/LabLinks/recipes/hbshepes.htm.
182. Dunn, M. J.; Crisp, S. J., Detection of proteins in polyacrylamide gels using an ultrasensitive silver staining technique. *Methods Mol Biol* **1994,** *32*, 113-8.
183. Maurye, P.; Basu, A.; Gupta, A., Simple and cost effective apparatus for silver staining of polyacrylamide gels with sequential reagents addition and real time monitoring. *J Biosci Bioeng* **2014,** *117* (6), 769-74.
184. Kelley, L. A.; Mezulis, S.; Yates, C. M.; Wass, M. N.; Sternberg, M. J. E., The Phyre2 web portal for protein modeling, prediction and analysis. *Nat. Protocols* **2015,** *10* (6), 845-858.
185. Källberg, M.; Wang, H.; Wang, S.; Peng, J.; Wang, Z.; Lu, H.; Xu, J., Template-based protein structure modeling using the RaptorX web server. *Nat. Protocols* **2012,** *7* (8), 1511-1522.
186. Ma, J.; Wang, S.; Zhao, F.; Xu, J., Protein threading using context-specific alignment potential. *Bioinformatics* **2013,** *29* (13), i257-i265.
187. Peng, J.; Xu, J., RaptorX: exploiting structure information for protein alignment by statistical inference. *Proteins* **2011,** *79 Suppl 10*, 161-71.
188. Roy, A.; Kucukural, A.; Zhang, Y., I-TASSER: a unified platform for automated protein structure and function prediction. *Nat. Protocols* **2010,** *5* (4), 725-738.
189. Thomson, C. A.; Olson, M.; Jackson, L. M.; Schrader, J. W., A simplified method for the efficient refolding and purification of recombinant human GM-CSF. *PLoS One* **2012,** *7* (11), e49891.
190. Khan, K. H., Gene expression in Mammalian cells and its applications. *Adv Pharm Bull* **2013,** *3* (2), 257-63.
191. Commercial protein product: Lubricin. https://www.mybiosource.com/prods/Peptide/Lubricin-PRG4/PRG4/datasheet.php?products_id=425914.
192. Delaglio, F.; Grzesiek, S.; Vuister, G. W.; Zhu, G.; Pfeifer, J.; Bax, A., NMRPipe: a multidimensional spectral processing system based on UNIX pipes. *J Biomol NMR* **1995,** *6* (3), 277-93.

193. Johnson, B. A.; Blevins, R. A., NMR View: A computer program for the visualization and analysis of NMR data. *J Biomol NMR* **1994,** *4* (5), 603-14.
194. Cavanagh, J., Fairbrother, W.J., Palmer. A.G.I., Skelton, N.J., Rance, M., *Protein NMR Spectroscopy: Principles and Practice.* Elsevier.: (2006).
195. Dubois, L.; Ohm Kyvik, K.; Girard, M.; Tatone-Tokuda, F.; Pérusse, D.; Hjelmborg, J.; Skytthe, A.; Rasmussen, F.; Wright, M. J.; Lichtenstein, P.; Martin, N. G., Genetic and environmental contributions to weight, height, and BMI from birth to 19 years of age: an international study of over 12,000 twin pairs. *PLoS One* **2012,** *7* (2), e30153.
196. Martin, L. J.; Woo, J. G.; Morrison, J. A., Evidence of shared genetic effects between pre- and postobesity epidemic BMI levels. *Obesity (Silver Spring)* **2010,** *18* (7), 1378-82.
197. Nguyen, D. M.; El-Serag, H. B., The epidemiology of obesity. *Gastroenterol Clin North Am* **2010,** *39* (1), 1-7.
198. Salsberry, P. J.; Reagan, P. B., Effects of heritability, shared environment, and nonshared intrauterine conditions on child and adolescent BMI. *Obesity (Silver Spring)* **2010,** *18* (9), 1775-80.
199. O'Rahilly, S.; Farooqi, I. S., Human obesity as a heritable disorder of the central control of energy balance. *Int J Obes (Lond)* **2008,** *32 Suppl 7*, S55-61.
200. Walley, A. J.; Asher, J. E.; Froguel, P., The genetic contribution to non-syndromic human obesity. *Nat Rev Genet* **2009,** *10* (7), 431-42.
201. Cui, H.; Moore, J.; Ashimi, S. S.; Mason, B. L.; Drawbridge, J. N.; Han, S.; Hing, B.; Matthews, A.; McAdams, C. J.; Darbro, B. W.; Pieper, A. A.; Waller, D. A.; Xing, C.; Lutter, M., Eating disorder predisposition is associated with ESRRA and HDAC4 mutations. *J Clin Invest* **2013,** *123* (11), 4706-13.
202. Canu, N.; Possenti, R.; Ricco, A. S.; Rocchi, M.; Levi, A., Cloning, structural organization analysis, and chromosomal assignment of the human gene for the neurosecretory protein VGF. *Genomics* **1997,** *45* (2), 443-6.
203. Bartolomucci, A.; La Corte, G.; Possenti, R.; Locatelli, V.; Rigamonti, A. E.; Torsello, A.; Bresciani, E.; Bulgarelli, I.; Rizzi, R.; Pavone, F.; D'Amato, F. R.; Severini, C.; Mignogna, G.; Giorgi, A.; Schininà, M. E.; Elia, G.; Brancia, C.; Ferri, G. L.; Conti, R.; Ciani, B.; Pascucci, T.; Dell'Omo, G.; Muller, E. E.; Levi, A.; Moles, A., TLQP-21, a VGF-derived peptide, increases energy expenditure and prevents the early phase of diet-induced obesity. *Proc Natl Acad Sci U S A* **2006,** *103* (39), 14584-9.
204. Bartolomucci, A.; Possenti, R.; Levi, A.; Pavone, F.; Moles, A., The role of the vgf gene and VGF-derived peptides in nutrition and metabolism. *Genes Nutr* **2007,** *2* (2), 169-80.
205. Lewis, J. E.; Brameld, J. M.; Jethwa, P. H., Neuroendocrine Role for VGF. *Front Endocrinol (Lausanne)* **2015,** *6*, 3.
206. Levi, A.; Ferri, G. L.; Watson, E.; Possenti, R.; Salton, S. R., Processing, distribution, and function of VGF, a neuronal and endocrine peptide precursor. *Cell Mol Neurobiol* **2004,** *24* (4), 517-33.
207. Salton, S. R.; Ferri, G. L.; Hahm, S.; Snyder, S. E.; Wilson, A. J.; Possenti, R.; Levi, A., VGF: a novel role for this neuronal and neuroendocrine polypeptide in the regulation of energy balance. *Front Neuroendocrinol* **2000,** *21* (3), 199-219.

208. Brancia, C.; Cocco, C.; D'Amato, F.; Noli, B.; Sanna, F.; Possenti, R.; Argiolas, A.; Ferri, G. L., Selective expression of TLQP-21 and other VGF peptides in gastric neuroendocrine cells and modulation by feeding. *J Endocrinol* **2010,** *207* (3), 329-41.
209. Severini, C.; La Corte, G.; Improta, G.; Broccardo, M.; Agostini, S.; Petrella, C.; Sibilia, V.; Pagani, F.; Guidobono, F.; Bulgarelli, I.; Ferri, G. L.; Brancia, C.; Rinaldi, A. M.; Levi, A.; Possenti, R., In vitro and in vivo pharmacological role of TLQP-21, a VGF-derived peptide, in the regulation of rat gastric motor functions. *Br J Pharmacol* **2009,** *157* (6), 984-93.
210. Wisor, J. P.; Takahashi, J. S., Regulation of the vgf gene in the golden hamster suprachiasmatic nucleus by light and by the circadian clock. *J Comp Neurol* **1997,** *378* (2), 229-38.
211. Sasaki, K.; Osaki, T.; Minamino, N., Large-scale identification of endogenous secretory peptides using electron transfer dissociation mass spectrometry. *Mol Cell Proteomics* **2013,** *12* (3), 700-9.
212. Kim, J. W.; Rhee, M.; Park, J. H.; Yamaguchi, H.; Sasaki, K.; Minamino, N.; Nakazato, M.; Song, D. K.; Yoon, K. H., Chronic effects of neuroendocrine regulatory peptide (NERP-1 and -2) on insulin secretion and gene expression in pancreatic β-cells. *Biochem Biophys Res Commun* **2015,** *457* (2), 148-53.
213. Sabbagh, U.; Mullegama, S.; Wyckoff, G. J., Identification and Evolutionary Analysis of Potential Candidate Genes in a Human Eating Disorder. *Biomed Res Int* **2016,** *2016*, 7281732.
214. Barrett, T.; Wilhite, S. E.; Ledoux, P.; Evangelista, C.; Kim, I. F.; Tomashevsky, M.; Marshall, K. A.; Phillippy, K. H.; Sherman, P. M.; Holko, M.; Yefanov, A.; Lee, H.; Zhang, N.; Robertson, C. L.; Serova, N.; Davis, S.; Soboleva, A., NCBI GEO: archive for functional genomics data sets--update. *Nucleic Acids Res* **2013,** *41* (Database issue), D991-5.
215. Allison, K. C.; Lundgren, J. D.; O'Reardon, J. P.; Geliebter, A.; Gluck, M. E.; Vinai, P.; Mitchell, J. E.; Schenck, C. H.; Howell, M. J.; Crow, S. J.; Engel, S.; Latzer, Y.; Tzischinsky, O.; Mahowald, M. W.; Stunkard, A. J., Proposed diagnostic criteria for night eating syndrome. *Int J Eat Disord* **2010,** *43* (3), 241-7.
216. Gallant, A. R.; Lundgren, J.; Drapeau, V., The night-eating syndrome and obesity. *Obes Rev* **2012,** *13* (6), 528-36.
217. Clapcote, S. J.; Duffy, S.; Xie, G.; Kirshenbaum, G.; Bechard, A. R.; Rodacker Schack, V.; Petersen, J.; Sinai, L.; Saab, B. J.; Lerch, J. P.; Minassian, B. A.; Ackerley, C. A.; Sled, J. G.; Cortez, M. A.; Henderson, J. T.; Vilsen, B.; Roder, J. C., Mutation I810N in the alpha3 isoform of Na+,K+-ATPase causes impairments in the sodium pump and hyperexcitability in the CNS. *Proc Natl Acad Sci U S A* **2009,** *106* (33), 14085-90.
218. Heinzen, E. L.; Swoboda, K. J.; Hitomi, Y.; Gurrieri, F.; Nicole, S.; de Vries, B.; Tiziano, F. D.; Fontaine, B.; Walley, N. M.; Heavin, S.; Panagiotakaki, E.; Fiori, S.; Abiusi, E.; Di Pietro, L.; Sweney, M. T.; Newcomb, T. M.; Viollet, L.; Huff, C.; Jorde, L. B.; Reyna, S. P.; Murphy, K. J.; Shianna, K. V.; Gumbs, C. E.; Little, L.; Silver, K.; Ptáček, L. J.; Haan, J.; Ferrari, M. D.; Bye, A. M.; Herkes, G. K.; Whitelaw, C. M.; Webb, D.; Lynch, B. J.; Uldall, P.; King, M. D.; Scheffer, I. E.; Neri, G.; Arzimanoglou, A.; van den Maagdenberg, A. M.; Sisodiya, S. M.; Mikati, M. A.; Goldstein, D. B.; Consortium, E. A. H. o. C. A. G.; Consortium, B. e. R. C. p. l. E. A. I. B. A.;

Consortium, E. N. f. R. o. A. H. E. f. S. a. M.-s. E. S., De novo mutations in ATP1A3 cause alternating hemiplegia of childhood. *Nat Genet* **2012,** *44* (9), 1030-4.
219. Garayoa, M.; Man, Y. G.; Martínez, A.; Cuttitta, F.; Mulshine, J. L., Downregulation of hnRNP A2/B1 expression in tumor cells under prolonged hypoxia. *Am J Respir Cell Mol Biol* **2003,** *28* (1), 80-5.
220. Huggenvik, J. I.; Michelson, R. J.; Collard, M. W.; Ziemba, A. J.; Gurley, P.; Mowen, K. A., Characterization of a nuclear deformed epidermal autoregulatory factor-1 (DEAF-1)-related (NUDR) transcriptional regulator protein. *Mol Endocrinol* **1998,** *12* (10), 1619-39.
221. Spiro, C.; McMurray, C. T., Transcriptional regulation of the human proenkephalin gene by conformational switching: implications for decoy design. *Antisense Nucleic Acid Drug Dev* **1998,** *8* (2), 159-65.
222. Gradi, A.; Imataka, H.; Svitkin, Y. V.; Rom, E.; Raught, B.; Morino, S.; Sonenberg, N., A novel functional human eukaryotic translation initiation factor 4G. *Mol Cell Biol* **1998,** *18* (1), 334-42.
223. Hahm, K.; Sum, E. Y.; Fujiwara, Y.; Lindeman, G. J.; Visvader, J. E.; Orkin, S. H., Defective neural tube closure and anteroposterior patterning in mice lacking the LIM protein LMO4 or its interacting partner Deaf-1. *Mol Cell Biol* **2004,** *24* (5), 2074-82.
224. Vulto-van Silfhout, A. T.; Rajamanickam, S.; Jensik, P. J.; Vergult, S.; de Rocker, N.; Newhall, K. J.; Raghavan, R.; Reardon, S. N.; Jarrett, K.; McIntyre, T.; Bulinski, J.; Ownby, S. L.; Huggenvik, J. I.; McKnight, G. S.; Rose, G. M.; Cai, X.; Willaert, A.; Zweier, C.; Endele, S.; de Ligt, J.; van Bon, B. W.; Lugtenberg, D.; de Vries, P. F.; Veltman, J. A.; van Bokhoven, H.; Brunner, H. G.; Rauch, A.; de Brouwer, A. P.; Carvill, G. L.; Hoischen, A.; Mefford, H. C.; Eichler, E. E.; Vissers, L. E.; Menten, B.; Collard, M. W.; de Vries, B. B., Mutations affecting the SAND domain of DEAF1 cause intellectual disability with severe speech impairment and behavioral problems. *Am J Hum Genet* **2014,** *94* (5), 649-61.
225. Manne, U.; Gary, B. D.; Oelschlager, D. K.; Weiss, H. L.; Frost, A. R.; Grizzle, W. E., Altered subcellular localization of suppressin, a novel inhibitor of cell-cycle entry, is an independent prognostic factor in colorectal adenocarcinomas. *Clin Cancer Res* **2001,** *7* (11), 3495-503.
226. Yip, L.; Su, L.; Sheng, D.; Chang, P.; Atkinson, M.; Czesak, M.; Albert, P. R.; Collier, A. R.; Turley, S. J.; Fathman, C. G.; Creusot, R. J., Deaf1 isoforms control the expression of genes encoding peripheral tissue antigens in the pancreatic lymph nodes during type 1 diabetes. *Nat Immunol* **2009,** *10* (9), 1026-33.
227. Rajaram, S.; Baylink, D. J.; Mohan, S., Insulin-like growth factor-binding proteins in serum and other biological fluids: regulation and functions. *Endocr Rev* **1997,** *18* (6), 801-31.
228. Ding, H.; Kharboutli, M.; Saxena, R.; Wu, T., Insulin-like growth factor binding protein-2 as a novel biomarker for disease activity and renal pathology changes in lupus nephritis. *Clin Exp Immunol* **2016,** *184* (1), 11-8.
229. Carter, S.; Capoulade, R.; Arsenault, M.; Bédard, É.; Dumesnil, J. G.; Mathieu, P.; Pibarot, P.; Picard, F., Relationship Between Insulin-Like Growth Factor Binding Protein-2 and Left Ventricular Stroke Volume in Patients With Aortic Stenosis. *Can J Cardiol* **2015,** *31* (12), 1447-54.

230. Izumi, K.; Kellogg, E.; Fujiki, K.; Kaur, M.; Tilton, R. K.; Noon, S.; Wilkens, A.; Shirahige, K.; Krantz, I. D., Elevation of insulin-like growth factor binding protein-2 level in Pallister-Killian syndrome: implications for the postnatal growth retardation phenotype. *Am J Med Genet A* **2015,** *167* (6), 1268-74.
231. Myers, A. L.; Lin, L.; Nancarrow, D. J.; Wang, Z.; Ferrer-Torres, D.; Thomas, D. G.; Orringer, M. B.; Lin, J.; Reddy, R. M.; Beer, D. G.; Chang, A. C., IGFBP2 modulates the chemoresistant phenotype in esophageal adenocarcinoma. *Oncotarget* **2015,** *6* (28), 25897-916.
232. Catsburg, C.; Gunter, M. J.; Tinker, L.; Chlebowski, R. T.; Pollak, M.; Strickler, H. D.; Cote, M. L.; Page, D. L.; Rohan, T. E., Serum IGFBP-2 and Risk of Atypical Hyperplasia of the Breast. *J Cancer Epidemiol* **2015,** *2015*, 203284.
233. French, C. L.; Ye, F.; Revetta, F.; Zhang, B.; Coffey, R. J.; Washington, M. K.; Deane, N. G.; Beauchamp, R. D.; Weaver, A. M., Linking patient outcome to high throughput protein expression data identifies novel regulators of colorectal adenocarcinoma aggressiveness. *F1000Res* **2015,** *4*, 99.
234. Royall, D. R.; Bishnoi, R. J.; Palmer, R. F., Serum IGF-BP2 strongly moderates age's effect on cognition: a MIMIC analysis. *Neurobiol Aging* **2015,** *36* (7), 2232-40.
235. Derry, J. M.; Kerns, J. A.; Francke, U., RBM3, a novel human gene in Xp11.23 with a putative RNA-binding domain. *Hum Mol Genet* **1995,** *4* (12), 2307-11.
236. Danno, S.; Nishiyama, H.; Higashitsuji, H.; Yokoi, H.; Xue, J. H.; Itoh, K.; Matsuda, T.; Fujita, J., Increased transcript level of RBM3, a member of the glycine-rich RNA-binding protein family, in human cells in response to cold stress. *Biochem Biophys Res Commun* **1997,** *236* (3), 804-7.
237. Wellmann, S.; Bührer, C.; Moderegger, E.; Zelmer, A.; Kirschner, R.; Koehne, P.; Fujita, J.; Seeger, K., Oxygen-regulated expression of the RNA-binding proteins RBM3 and CIRP by a HIF-1-independent mechanism. *J Cell Sci* **2004,** *117* (Pt 9), 1785-94.
238. Lleonart, M. E., A new generation of proto-oncogenes: cold-inducible RNA binding proteins. *Biochim Biophys Acta* **2010,** *1805* (1), 43-52.
239. Boman, K.; Segersten, U.; Ahlgren, G.; Eberhard, J.; Uhlén, M.; Jirström, K.; Malmström, P. U., Decreased expression of RNA-binding motif protein 3 correlates with tumour progression and poor prognosis in urothelial bladder cancer. *BMC Urol* **2013,** *13*, 17.
240. Hjelm, B.; Brennan, D. J.; Zendehrokh, N.; Eberhard, J.; Nodin, B.; Gaber, A.; Pontén, F.; Johannesson, H.; Smaragdi, K.; Frantz, C.; Hober, S.; Johnson, L. B.; Påhlman, S.; Jirström, K.; Uhlen, M., High nuclear RBM3 expression is associated with an improved prognosis in colorectal cancer. *Proteomics Clin Appl* **2011,** *5* (11-12), 624-35.
241. Liou, H. C.; Eddy, R.; Shows, T.; Lisowska-Grospierre, B.; Griscelli, C.; Doyle, C.; Mannhalter, J.; Eibl, M.; Glimcher, L. H., An HLA-DR alpha promoter DNA-binding protein is expressed ubiquitously and maps to human chromosomes 22 and 5. *Immunogenetics* **1991,** *34* (5), 286-92.
242. Hurst, H. C., Transcription factors 1: bZIP proteins. *Protein Profile* **1995,** *2* (2), 101-68.
243. Liou, H. C.; Boothby, M. R.; Finn, P. W.; Davidon, R.; Nabavi, N.; Zeleznik-Le, N. J.; Ting, J. P.; Glimcher, L. H., A new member of the leucine zipper class of proteins that binds to the HLA DR alpha promoter. *Science* **1990,** *247* (4950), 1581-4.

244. Yoshida, H.; Nadanaka, S.; Sato, R.; Mori, K., XBP1 is critical to protect cells from endoplasmic reticulum stress: evidence from Site-2 protease-deficient Chinese hamster ovary cells. *Cell Struct Funct* **2006,** *31* (2), 117-25.
245. Winnay, J. N.; Boucher, J.; Mori, M. A.; Ueki, K.; Kahn, C. R., A regulatory subunit of phosphoinositide 3-kinase increases the nuclear accumulation of X-box-binding protein-1 to modulate the unfolded protein response. *Nat Med* **2010,** *16* (4), 438-45.
246. Reimold, A. M.; Etkin, A.; Clauss, I.; Perkins, A.; Friend, D. S.; Zhang, J.; Horton, H. F.; Scott, A.; Orkin, S. H.; Byrne, M. C.; Grusby, M. J.; Glimcher, L. H., An essential role in liver development for transcription factor XBP-1. *Genes Dev* **2000,** *14* (2), 152-7.
247. Yoshida, H., ER stress and diseases. *FEBS J* **2007,** *274* (3), 630-58.
248. Ning, J.; Hong, T.; Ward, A.; Pi, J.; Liu, Z.; Liu, H. Y.; Cao, W., Constitutive role for IRE1α-XBP1 signaling pathway in the insulin-mediated hepatic lipogenic program. *Endocrinology* **2011,** *152* (6), 2247-55.
249. Sha, H.; He, Y.; Chen, H.; Wang, C.; Zenno, A.; Shi, H.; Yang, X.; Zhang, X.; Qi, L., The IRE1alpha-XBP1 pathway of the unfolded protein response is required for adipogenesis. *Cell Metab* **2009,** *9* (6), 556-64.
250. Yang, J.; Cheng, D.; Zhou, S.; Zhu, B.; Hu, T.; Yang, Q., Overexpression of X-Box Binding Protein 1 (XBP1) Correlates to Poor Prognosis and Up-Regulation of PI3K/mTOR in Human Osteosarcoma. *Int J Mol Sci* **2015,** *16* (12), 28635-46.
251. Martinet, W.; Croons, V.; Timmermans, J. P.; Herman, A. G.; De Meyer, G. R., Nitric oxide selectively depletes macrophages in atherosclerotic plaques via induction of endoplasmic reticulum stress. *Br J Pharmacol* **2007,** *152* (4), 493-500.
252. Zeng, L.; Zampetaki, A.; Margariti, A.; Pepe, A. E.; Alam, S.; Martin, D.; Xiao, Q.; Wang, W.; Jin, Z. G.; Cockerill, G.; Mori, K.; Li, Y. S.; Hu, Y.; Chien, S.; Xu, Q., Sustained activation of XBP1 splicing leads to endothelial apoptosis and atherosclerosis development in response to disturbed flow. *Proc Natl Acad Sci U S A* **2009,** *106* (20), 8326-31.
253. Chen, X.; Iliopoulos, D.; Zhang, Q.; Tang, Q.; Greenblatt, M. B.; Hatziapostolou, M.; Lim, E.; Tam, W. L.; Ni, M.; Chen, Y.; Mai, J.; Shen, H.; Hu, D. Z.; Adoro, S.; Hu, B.; Song, M.; Tan, C.; Landis, M. D.; Ferrari, M.; Shin, S. J.; Brown, M.; Chang, J. C.; Liu, X. S.; Glimcher, L. H., XBP1 promotes triple-negative breast cancer by controlling the HIF1α pathway. *Nature* **2014,** *508* (7494), 103-7.
254. Casas-Tinto, S.; Zhang, Y.; Sanchez-Garcia, J.; Gomez-Velazquez, M.; Rincon-Limas, D. E.; Fernandez-Funez, P., The ER stress factor XBP1s prevents amyloid-beta neurotoxicity. *Hum Mol Genet* **2011,** *20* (11), 2144-60.
255. Ozcan, U.; Cao, Q.; Yilmaz, E.; Lee, A. H.; Iwakoshi, N. N.; Ozdelen, E.; Tuncman, G.; Görgün, C.; Glimcher, L. H.; Hotamisligil, G. S., Endoplasmic reticulum stress links obesity, insulin action, and type 2 diabetes. *Science* **2004,** *306* (5695), 457-61.
256. Kent, W. J.; Sugnet, C. W.; Furey, T. S.; Roskin, K. M.; Pringle, T. H.; Zahler, A. M.; Haussler, D., The human genome browser at UCSC. *Genome Res* **2002,** *12* (6), 996-1006.
257. Kent, W. J., BLAT--the BLAST-like alignment tool. *Genome Res* **2002,** *12* (4), 656-64.
258. Dayem Ullah, A. Z.; Lemoine, N. R.; Chelala, C., A practical guide for the functional annotation of genetic variations using SNPnexus. *Brief Bioinform* **2013,** *14* (4), 437-47.

259. Kumar, P.; Henikoff, S.; Ng, P. C., Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat Protoc* **2009,** *4* (7), 1073-81.
260. Adzhubei, I. A.; Schmidt, S.; Peshkin, L.; Ramensky, V. E.; Gerasimova, A.; Bork, P.; Kondrashov, A. S.; Sunyaev, S. R., A method and server for predicting damaging missense mutations. *Nat Methods* **2010,** *7* (4), 248-9.
261. McDonald, J. H.; Kreitman, M., Adaptive protein evolution at the Adh locus in Drosophila. *Nature* **1991,** *351* (6328), 652-4.
262. Tamura, K.; Stecher, G.; Peterson, D.; Filipski, A.; Kumar, S., MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. *Mol Biol Evol* **2013,** *30* (12), 2725-9.
263. Thompson, J. D.; Higgins, D. G.; Gibson, T. J., CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* **1994,** *22* (22), 4673-80.
264. Ruvolo, M., Molecular phylogeny of the hominoids: inferences from multiple independent DNA sequence data sets. *Mol Biol Evol* **1997,** *14* (3), 248-65.
265. Lockwood, C. A.; Kimbel, W. H.; Lynch, J. M., Morphometrics and hominoid phylogeny: Support for a chimpanzee-human clade and differentiation among great ape subspecies. *Proc Natl Acad Sci U S A* **2004,** *101* (13), 4356-60.
266. Prado-Martinez, J.; Sudmant, P. H.; Kidd, J. M.; Li, H.; Kelley, J. L.; Lorente-Galdos, B.; Veeramah, K. R.; Woerner, A. E.; O'Connor, T. D.; Santpere, G.; Cagan, A.; Theunert, C.; Casals, F.; Laayouni, H.; Munch, K.; Hobolth, A.; Halager, A. E.; Malig, M.; Hernandez-Rodriguez, J.; Hernando-Herraez, I.; Prüfer, K.; Pybus, M.; Johnstone, L.; Lachmann, M.; Alkan, C.; Twigg, D.; Petit, N.; Baker, C.; Hormozdiari, F.; Fernandez-Callejo, M.; Dabad, M.; Wilson, M. L.; Stevison, L.; Camprubí, C.; Carvalho, T.; Ruiz-Herrera, A.; Vives, L.; Mele, M.; Abello, T.; Kondova, I.; Bontrop, R. E.; Pusey, A.; Lankester, F.; Kiyang, J. A.; Bergl, R. A.; Lonsdorf, E.; Myers, S.; Ventura, M.; Gagneux, P.; Comas, D.; Siegismund, H.; Blanc, J.; Agueda-Calpena, L.; Gut, M.; Fulton, L.; Tishkoff, S. A.; Mullikin, J. C.; Wilson, R. K.; Gut, I. G.; Gonder, M. K.; Ryder, O. A.; Hahn, B. H.; Navarro, A.; Akey, J. M.; Bertranpetit, J.; Reich, D.; Mailund, T.; Schierup, M. H.; Hvilsom, C.; Andrés, A. M.; Wall, J. D.; Bustamante, C. D.; Hammer, M. F.; Eichler, E. E.; Marques-Bonet, T., Great ape genetic diversity and population history. *Nature* **2013,** *499* (7459), 471-5.
267. Pizarro, A.; Hayer, K.; Lahens, N. F.; Hogenesch, J. B., CircaDB: a database of mammalian circadian gene expression profiles. *Nucleic Acids Res* **2013,** *41* (Database issue), D1009-13.
268. Organization, W. H. Obesity: Situation and Trends. http://www.who.int/gho/ncd/ risk factors/obesity text/en/.
269. Beisel, H. G.; Kawabata, S.; Iwanaga, S.; Huber, R.; Bode, W., Tachylectin-2: crystal structure of a specific GlcNAc/GalNAc-binding lectin involved in the innate immunity host defense of the Japanese horseshoe crab Tachypleus tridentatus. *EMBO J* **1999,** *18* (9), 2313-22.
270. Good, M. C.; Greenstein, A. E.; Young, T. A.; Ng, H. L.; Alber, T., Sensor domain of the Mycobacterium tuberculosis receptor Ser/Thr protein kinase, PknD, forms a highly symmetric beta propeller. *J Mol Biol* **2004,** *339* (2), 459-69.

271. Adams, D. R.; Ron, D.; Kiely, P. A., RACK1, A multifaceted scaffolding protein: Structure and function. *Cell Commun Signal* **2011,** *9*, 22.
272. Pausch, P.; Singh, U.; Ahmed, Y. L.; Pillet, B.; Murat, G.; Altegoer, F.; Stier, G.; Thoms, M.; Hurt, E.; Sinning, I.; Bange, G.; Kressler, D., Co-translational capturing of nascent ribosomal proteins by their dedicated chaperones. *Nat Commun* **2015,** *6*, 7494.
273. Quistgaard, E. M.; Grøftehauge, M. K.; Madsen, P.; Pallesen, L. T.; Christensen, B.; Sørensen, E. S.; Nissen, P.; Petersen, C. M.; Thirup, S. S., Revisiting the structure of the Vps10 domain of human sortilin and its interaction with neurotensin. *Protein Sci* **2014,** *23* (9), 1291-300.

VITA

Lee Likins was born in Hattiesburg, Mississippi on June 5th, 1957. His father was a career U.S. Army officer and as a consequence in the early 1970s the family was stationed in the town of Bamberg, formerly West Germany, which led to Lee graduating from Nürnberg American High School, located in the town of Fürth, in June of 1975. He was awarded a Naval ROTC scholarship and returned to his stateside home town of Auburn, Alabama to attend Auburn University. He left Auburn in 1978 to enter into active military service in the United States Navy where he served for 2 years as an Aerographer's Mate (Weather Forecasting Technician). Upon his discharge from the Navy he used the G.I. Bill to finish his undergraduate work at the University of West Florida, in Pensacola, Florida, where he earned a Bachelor's of Science in Marine Biology in June of 1985. After graduation, he took several jobs with state and federal agencies finally landing in the Washington, D.C. area working as a Marine Biologist/Fisheries Statistician for a private company contracted through the National Marine Fisheries Service (NMFS) branch of the National Oceanic and Atmospheric Administration (NOAA). He worked for NOAA until 1992 when he left to return to graduate school. He enrolled in the graduate program at the University of Kansas (KU), Lawrence, where he received his Master of Arts degree in Ecology and Evolutionary Biology in 1996. After graduating from KU he took several teaching positions as an adjunct faculty member at KU, Johnson County Community College (JCCC) and Kansas City Kansas Community College (KCKCC) until he was hired onto the faculty of the School of Biological Sciences at the University of Missouri Kansas City (SBS-UMKC), where he is currently employed as a Research Instructor and pursuing the iPhD in Molecular Biology and Biochemistry/Cellular

Biology and Biophysics. Upon graduation he plans to remain on the Teaching Faculty of SBS-UMKC. He is a member of the National Science Teachers Association (NSTA) - College Teaching Division, and the American Association for the Advancement of Science (AAAS).