Norwegian University of Life Sciences
Faculty of Environmental Sciences
and Natural Resource Management

Philosophiae Doctor (PhD)
Thesis 2022:8

Understanding the structure-function relationship of honey bee Vitellogenin

Forståelse av forholdet mellom struktur og funksjon til Vitellogenin i honningbia

Vilde Leipart

Understanding the structure-function relationship of honey bee Vitellogenin

Forståelse av forholdet mellom struktur og funksjon til Vitellogenin i honningbia

Philosophiae Doctor (PhD) Thesis

Vilde Leipart

Norwegian University of Life Sciences
Faculty of Environmental Sciences and Natural Resource Management

Ås (2022)

Thesis number 2022:8
ISSN 1894-6402
ISBN 978-82-575-1885-1

Supervisors and evaluation committee

Main supervisor:
Professor Gro V. Amdam,
Faculty of Environmental Sciences and Natural Resource Management (MINA) and
School of Life Sciences, Arizona State University, AZ, USA

Co-supervisor:
Professor Øyvind Halskau (University of Bergen) and Dr. Jane Ludvigsen, MINA/NMBU

First opponent: Dr. Vicky Higman, University of Leicester
Second opponent: Professor Leonard Foster, University of British Colombia
Committee coordinator: Professor Tone Birkemoe, Faculty of Environmental Sciences and Natural Resource Management, NMBU

Acknowledgments

I would like to thank my main supervisor, Gro V. Amdam, for supporting and believing in me. Your enthusiasm and endless knowledge is motivating and inspiring, and I will forever be grateful for this opportunity and your excellent mentoring.
I would also like to thank my co-supervisors:
Øyvind Halskau – your massive support, encouragement, and sharing of your wisdom on structural biology has been greatly appreciated.
Jane Ludvigsen – your insight and expertise have been crucial – I have learned so much!
My sincere gratitude also goes to Claus Kreibich for teaching me about honey bees and always helping in any situation.

I would also like to thank all my collaborators at NMBU and the University of Bergen. Thanks to the many beekeepers for the time and effort invested in helping me! Thanks to all my fellow PhD students and postdocs at MINA for supporting me and contributing to creating a wonderful working environment.

A big thanks to all my friends and family for supporting and encouraging me through my ups and downs. A special thanks to Eivind for always being there and for helping me along the way!

Finally, I would like to thank the Research Council of Norway for my funding, which made this research possible. A special thanks also goes to BioCat (the National Graduate School in Biocatalysis) for providing me with funding for extra travels and conferences.

Table of Contents

Abbreviations

Cryo-EM	Cryo-electron microscopy
DUF	Domain of unknown function
ICP-MS	Inductively coupled plasma mass spectrometry
LLTPs	Large lipid transfer proteins
MSA	Multiple sequence alignment
ND	N-terminal domain
NMR	Nuclear magnetic resonance
nsSNPs	Non-synonymous single nucleotide polymorphisms
TGIP	Trans-generational immune priming
Vg	Vitellogenin
vWF	von Willebrand factor

List of papers

Paper I:

Vilde Leipart, Mateu Montserrat-Canals, Eva S Cunha, Hartmut Luecke, Elías Herrero-Galán, Øyvind Halskau, Gro V. Amdam. **Structure prediction of honey bee vitellogenin: a multi-domain protein important for insect immunity.** *FEBS Open Bio* (2021), doi: 10.1002/2211-5463.13316

Paper II:

Vilde Leipart, Øyvind Enger, Diana Cornelia Turcu, Olena Dobrovolska, Finn Drabløs, Øyvind Halskau, Gro V. Amdam. **Where Honey Bee Vitellogenin may Bind Zn^{2+}-Ions** *Manuscript in preparation for Insect Molecular Biology*

Paper III:

Vilde Leipart, Jane Ludvigsen, Matthew Kent, Simen Sandve, Thu-Hien To, Mariann Árnyasi, Claus D Kreibich, Bjørn Dahle, Gro V. Amdam. **Identification of 121 variants of honey bee Vitellogenin protein sequence with structural differences at functional sites.** *Manuscript in preparation for PLoS Biology*

Paper IV:

Vilde Leipart, Øyvind Halskau, Gro V. Amdam. **How honey bee Vitellogenin holds lipid cargo: A role of the C-terminal.** *Editorially Accepted Research Topic for submission to: "In Celebration of Women in Science: Structural Biology" Frontiers in Molecular Biosciences Structural Biology*

Abstract

This thesis focuses on the structure and molecular function of Vitellogenin (Vg) from honey bees (*Apis mellifera*). Vg is an ancient protein found in animals. Most biological processes depend on proteins' activities, and the structural shape of proteins determines what they can do and how they work. It is important to understand the shape and associated functional properties of honey bee Vg, as honey bees are important pollinators in our natural environment and agricultural food system. A yolk-protein that transports nutrients like lipids and zinc, Vg is necessary for honey bee reproduction, and the protein also regulates social behavior and has immune-related functions. Paper I presents a full-length protein structure for honey bee Vg, generated using computational structure prediction. For the first time, we describe the complete structural fold of the protein, revealing previously unknown structural features. In Paper II, I use structural- and sequence-data analysis to identify seven potential zinc-binding sites at different protein regions. Element analysis of purified Vg shows that, on average, three zinc-sites are occupied per molecule – a ratio not reported before. Paper III explores the Vg structure from the perspective of allelic variation on the honey bee *vg*-gene. We used amplicon Nanopore sequencing with barcoded primers to identify 121 Vg variants. With these data, I found that the domains and subdomains of Vg are characterized by different levels of variation. While some of these patterns were expected, my results also provide new insights on possible structure-function relationships. I use findings from Papers I, II, and III in Paper IV to develop a novel explanatory model for how Vg holds its lipid load. In sum, this thesis presents a detailed structural study that contributes toward understanding the multifunctional role of honey bee Vg.

Norsk sammendrag

Denne avhandlingen fokuserer på strukturen og funksjonen til Vitellogenin (Vg) hos honningbier (*Apis mellifera*). Vg er et gammelt protein som finnes i mange dyr. De fleste biologiske prosesser er avhengige av proteiners aktivitet, og den strukturelle formen til et protein bestemmer hva det kan gjøre og hvordan det fungerer. De er viktig å forstå formen og de assosierte funksjonelle egenskapene til Vg i honningbia, ettersom honningbier er viktige pollinatorer i vårt naturlige miljø og for matproduksjon i landbruk. Vg er nødvendig for reproduksjon i honningbier som et egg-protein, ved å transportere næringsstoffer som lipider og sink, men proteinet regulerer også sosial adferd og har immunrelaterte funksjoner. Paper I presenterer en full-lengde proteinstruktur av Vg i honningbia, generert ved å bruke beregningsmessig protein-prediksjon. Vi beskriver en fullstendig strukturell form av proteinet for første gang, som avdekker nye strukturelle egenskaper. I Paper II, bruker jeg struktur- og sekvensdata-analyser til å identifisere syv potensielle sink-bindingsseter på ulike områder i proteinet. Element-analyse av renset Vg viser at tre sink-seter, i snitt, er bundet per molekyl – en ratio som ikke har blitt rapportert tidligere. Paper III utforsker Vg strukturen fra et genetisk variasjonsperspektiv i *vg*-genet til honningbia. Vi bruker amplikon Nanopore-sekvensering med seriekodede primere for å identifisere 121 Vg-varianter. Med disse data fant jeg ut at domener og subdomer i Vg karakteriseres av variasjonsnivå. Noen av disse mønstrene var forventet, men mine resultater bidrar også til ny innsikt i forholdet mellom Vgs struktur og funksjon. Jeg bruker funnene fra Paper I, II, og III i Paper IV for å utlede en ny forklaringsmodell for hvordan Vg bærer sin lipidlast. Min avhandling representerer en detaljert strukturell studie som tar viktige steg mot å forstå den flerfunksjonelle rollen til Vg i honningbia.

Synopsis

Introduction

Proteins

Proteins are essential molecular building blocks in living organisms [1]. Proteins come in many shapes and sizes, with a variety of specialized functions [2]. Some are long and thin and can create muscle movement, while others have a spherical shape and contain metal ions that can transport oxygen. Proteins in the immune system can defend against damaging substances, while other proteins can regulate or control the expression of genes. They can also work together with other substances inside cells to make factories that produce new proteins. All proteins are composed of amino acids. Genes in the genetic material provide the main instructions for making unique proteins, and copies of this information are delivered to the factories. The factories build proteins by linking amino acids together in the instructed order. The twenty standard amino acids, each having a unique side chain, consist of a basic structure: a carbon atom, an amino group, and a carboxyl group [3]. The amino group from one amino acid is combined with the carboxyl group from another amino acid to form a covalent peptide bond. A sequence of covalent peptide bonds makes up the protein's primary structure (polypeptide), which is its backbone (Figure 1). The sequence quickly folds into secondary structures, mainly caused by hydrogen bonds formed in the backbone. Interactions between the side chains create the tertiary shape. Water molecules or other proteins are additional factors that can influence the folding. For example, hydrophobic side chains avoid contact with water and clump together. Proteins may sometimes need the assistance of other proteins to fold correctly. Occasionally, a metal ion or a modification might be inserted or bound to the protein structure so that the protein can function accurately. The multiple interactions in and between the backbone and the side changes result in a loss of free energy and create a stable structure [3]. Longer polypeptide chains can sometimes fold into two or more compact regions, called domains, that can usually be stable enough to exist independently. Finally, several polypeptides can interact, making a quaternary structure.

The amino acid types and order in the protein sequence dictate the protein’s structural shape, and the structure of the protein determines its function. The final form is precisely folded so that the protein can complete its specific tasks. Although the protein structure typically persists if there are minor changes to the amino acid sequence, changing a single amino acid can sometimes disrupt the structure. Such disruptions can result in lost, gained, or altered functions [2, 3]. How often changes occur may depend on the proteins’ function. For example, proteins with many interacting partners or proteins produced at several locations tend to have few changes, while updates can sometimes be necessary when the recognized binding

partner keeps on changing [4]. Determining a protein amino acid composition and describing the resulting protein structure can help us better understand most biological processes.

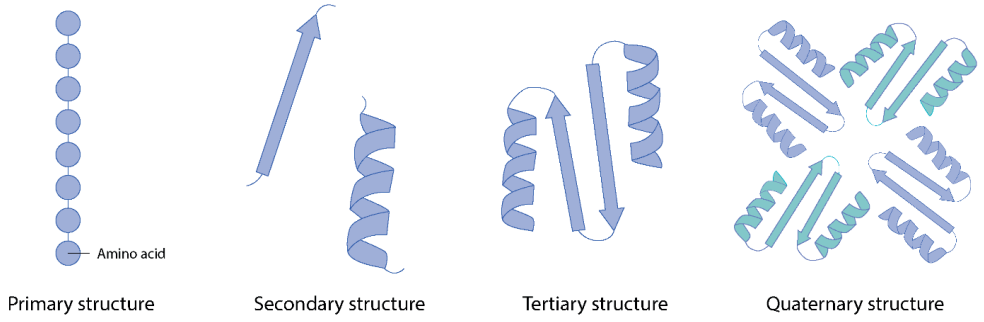


Figure 1: The primary structure of a protein is the polypeptide chain of covalently linked amino acids (blue circles). The chain folds into secondary structural elements, for example, a β-strand (arrow) or an α-helix (spiral). The elements pack together to create the tertiary structure. Several polypeptides can create a quaternary protein structure.

Honey bees

Wild or domesticated honey bees (*Apis mellifera*) are ecologically and economically important social insects; they are efficient pollinators and primary producers of materials like honey and beeswax. Honey bees are under natural or artificial selection, and their labor is critical for the agricultural food system, helping flower plants and ensuring the survival of the animals that feed on these plants [5]. Honey bees are also a good model organism for studying sophisticated behaviors [6]. These social insects live together in a well-organized colony [5]. A colony contains an egg-laying queen, tens of thousands of functionally sterile workers, and male drones. Worker bees conduct a variety of tasks like collecting pollen and nectar, cleaning the colony, or feeding developing larvae. The drones' main role is to mate with queens. A queen mates with several drones early in her life and stores the sperm. She fertilizes her eggs by supplying sperm, which gives rise to female worker bees or new queens (unfertilized eggs become drones). Worker bees can be full sisters with the same father or be half-sisters with different fathers, increasing the colony's genetic diversity [7].

This “superorganism” is an evolutionarily successful strategy and an example of thousands of individuals successfully working together for common goals [8]. For example, teamwork is important for locating pollen and nectar. Honey bees are heavily dependent on the environment close to the colony for food, and the quantity and quality of food sources in the local environment can vary significantly. When a foraging bee has identified a rich food source, she communicates its direction and distance from the hive to her sisters by dancing

[5]. Honey bee workers also collaborate during cold weather or winter to keep the queen and the food warm and safe by clustering and producing heat [9]. Furthermore, honey bees fight infections as a group. They do this through so-called social immunity, which involves behaviors to kill pathogens and prevent transmission. The food brought into the colony (during foraging seasons) can have side passengers like bacteria, viruses, fungi, or toxins. Honey bees live in dense populations, putting them at high risk for infectious diseases. To combat this risk, the bees groom each other to inspect for and remove potential parasites. If this is not sufficient, individuals who become infected or die are removed from the colony by other workers [10]. Taken together, honey bees are one of the most studied social insects on our planet [5, 6]. The available genome information, their global presence, and the low cost of obtaining many individuals make them a practical study system. We have much to learn from honey bees, and my thesis brings the research field one step further in this endeavor.

Vitellogenin

The egg yolk precursor protein Vitellogenin (Vg) provides lipids and other nutrients to developing embryos [11-13]. In insects, Vg is mainly synthesized in fat body (a tissue that is functionally comparable to the vertebrate liver and white fat) before it is transported to the hemolymph (insect blood). From there, the protein is generally transferred to ovaries through a receptor-mediated process and deposited into eggs [14, 15]. In honey bees, Vg is found in the muscles, gut, and brain and in both queens and the functionally sterile female workers [14, 16-18]. These diverse locations point to Vg's functionality beyond the reproductive role. Vg has received much attention from honey bee researchers over the last two decades. Initially, researchers found that the protein influenced the division of labor between worker bees [19, 20]. Young bees that care for larvae (nurses) have a higher Vg titer compared to typically older foraging bees. The Vg levels shift according to bees' social tasks and affect their life expectancy. For example, if a foraging bee returns to nursing, the production of Vg increases, and life expectancy is enhanced [21]. Workers high in Vg also have higher titers of functional immune cells and better resistance to oxidative stress [22, 23]. Researchers have speculated that these latter associations rely on zinc [24, 25]. Zinc is a metal ion that is essential for development and important for thousands of proteins' structural shape and numerous animals' functional roles [26-28]. Vg is the main circulatory zinc-carrying protein in honey bees [29]. For example, studies have suggested that Vg donates zinc to help immune cells function properly [22, 23, 29]. The possible immune-related activity of Vg has been further studied, and it was found that Vg recognizes components of the cell walls on disease-causing bacteria and fungi (pathogens) and damaged or dying cells [30]. This recognition potential of Vg also exists in several species of fish [31, 32] as well as invertebrates other than bees [33, 34]. Most recently, researchers have identified that the immune function of honey bee Vg extends to trans-generational immune priming (TGIP). This process allows females to prime their offspring against the pathogens that they encounter and increases the likelihood that the offspring will survive. The study of honey bees has contributed to a better

understanding of TGIP by revealing that Vg can bind and carry fragments of pathogens into ovaries and to developing eggs [35]. Ongoing research on honey bee Vg reveals new functional roles, exemplified by this year's discovery that the protein can influence the transcription of genes via subdomain translocation and DNA binding [36].

Vg belongs to a protein family that arose early in animal evolution, called large lipid transfer proteins (LLTPs) [37]. The common structural feature of all members is the lipid binding cavity. During evolution, the superfamily members developed otherwise specific structural features, dividing them into subcategories. Vg is one subcategory with a large lipid binding cavity and a well-conserved N-terminal domain (ND) [38, 39]. The ND consists of two distinct structural folds that create two subdomains, the β-barrel and the α-helical. The remaining domains differ across species. Vg usually includes one or several domains of unknown function (DUF) and a von Willebrand factor (vWF) domain [14]. Studies have identified various features for Vg members, for example, the presence of an extended serine-rich region at different positions. Sequence analysis has shown that the so-called polyserine linker is between the ND subdomains in honey bee Vg [14]. Knowledge of the structural features of Vg primarily comes from a crystal structure of lamprey (*Ichthyomyzon unicuspis*) [40, 41], solved over two decades ago; this is still the only experimentally resolved structure of any Vg. Although lamprey and honey bees are distant relatives, homology-based modeling of the conserved ND was possible. The first subdomain reveals a β-barrel-like shape conserved in both species, but honey bee Vg includes additional structural regions only preserved in insects [42]. The cleavage of honey bee Vg at the polyserine linker was demonstrated; two fragments are created: one small 40 kDa fragment consisting of the β-barrel subdomain and one larger 150 kDa fragment. The β-barrel subdomain was shown to be phosphorylated and glycosylated [42, 43]. The large fragment consists of the remaining domains, including the α-helical subdomain of ND, one DUF (DUF1943), and a vWF domain [36], where modeling was only feasible for the second subdomain of ND. This was done soon after the first model and demonstrated a conserved structure and a missing insect-specific structural region [30]. The model did show an 18 α-helical repeated domain that included 34 positively charged residues on the surface side. This finding increased the understanding of the subdomain recognition potential to the negatively charged cell wall fragments of pathogens [30, 35]. Thus, the lamprey crystal structure has provided important structural insights. However, the low sequence similarity and different structural Vg domains between the lamprey and bee have left a restricted understanding of the remaining domains, including the lipid binding cavity and the vWF domain (Figure 2).

Thus, Vg is multifunctional and central for honey bee health and social behaviors. Its ancient protein family has given rise to proteins with central roles in lipid transport and immunity in species as diverse as fish and insects [37, 44]. Most egg-laying animals depend on Vg for reproduction [13]. In honey bees, Vg's additional abilities to recognize and transport

fragments from pathogens, participate in transcriptional regulation, behavioral regulation, and somatic maintenance highlight how this protein can contribute to understanding important biological processes. Thus, progress in understanding the activities of Vg can be fueled by an improved structural prediction for honey bee Vg.

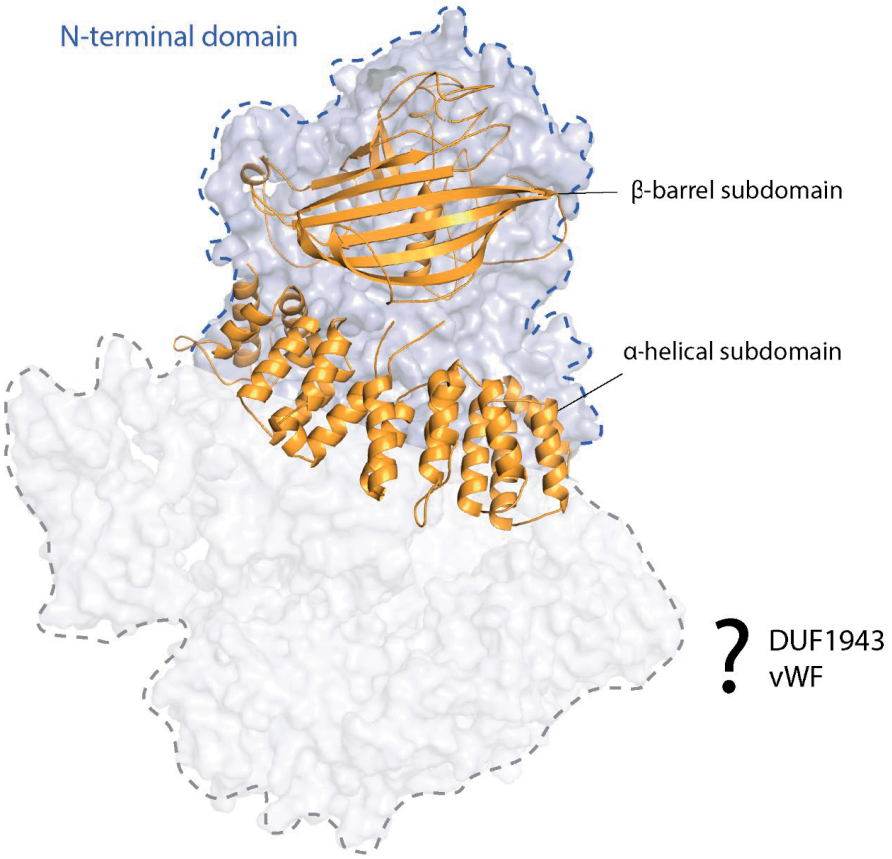


Figure 2: Illustration of the known structural features of honey bee Vg. The predicted structural folds of the β-barrel subdomain and α-helical subdomain (yellow) make up the ND (blue dotted line). The size of honey bee Vg is known, but the structural folds of the remaining domains are unknown (grey dotted line), for example, the DUF1943 and the vWF domain.

Protein structure prediction

When possible, laboratory methods are used to determine protein structures. Standard methods include X-ray crystallography and nuclear magnetic resonance (NMR) spectroscopy [3]. Both determine the relative positions of atoms in the protein but use different approaches. The former method requires the protein of interest to form a crystal, which can be time-consuming and sometimes fails. In theory, there is no size limit for determining protein structures in this way, and the method can potentially produce very precise models directly based on electron density calculated from the diffraction data [3]. NMR spectroscopy's application is limited to smaller proteins and usually requires relatively high concentrations of recombinant protein labeled with stable isotopes. However, NMR requires no crystals as the structure is solved while the protein is in solution or even inside a live cell [45]. The resulting model reflects the dynamic nature of proteins, which is often related to its function and can be further explored using NMR [46]. The relatively recent development of experimental approaches and new hardware and software technology have made it possible to use a third method for precise structure interpretation [47, 48], cryo-electron microscopy (cryo-EM). This method snap-freezes the protein before using advanced electron detectors to take high-quality images of the sample. Sophisticated programs are used to curate the collected images and can produce near-atomic representations of tertiary protein structures or protein complexes. The method only needs a small volume of the protein sample and usually provides the best results for sizeable and stable proteins [49]. The approaches have complementary strengths: cryo-EM can provide an overall shape, X-ray crystallography can give detailed information, and NMR can fill in the blanks about conformational changes and the dynamic nature of the protein.

Sometimes, it may be challenging to obtain the protein of interest from natural sources or in recombinant form, which restricts experimental progress. Computational modeling is then a good alternative. Currently, numerous approaches use different algorithms to predict the protein fold based on the amino acid sequence. Generally, the methods either use an experimentally solved structure as a template or attempt to predict how the amino acid sequences fold using the laws of physics [50]. The growing number of solved protein structures shared in public databases creates an increasingly stronger foundation for numerous computational resources [51]. The available software and algorithms are also becoming more powerful. For example, a neural network-based algorithm, AlphaFold, has reached groundbreaking accuracy for computational structure predictions [52].

Aims of the study

When I started this project, there was a general understanding of the ND structure for honey bee Vg, mainly provided through computational methods. However, information about several domains was lacking, including the essential lipid binding site and a precise anatomic representation of most regions. Previous work had met roadblocks, as the protein does not seem to crystallize (personal communications with supervisors). Furthermore, the large size (1,770 amino acids) disqualifies Vg for NMR and provides challenges for producing a synthetic construct. To obtain a full-length protein sample, Vg is purified from a natural source, which is time-consuming and produces a low yield. However, new developments in structural biology have created new opportunities. My project utilized these developments to arrive to a more detailed understanding of the protein structure of honey bee Vg.

Research has mostly described the functional roles of honey bee Vg at an individual level. To more fully understand the functional impact of the protein, detailed knowledge at the genetic, molecular, and anatomic levels is needed. My study focuses on the structure-function relationship of honey bee Vg using structural and genetic data. The project has four aims:

1. Constructing the first full-length structure prediction of honey bee Vg (Paper I)
2. Providing an in-depth analysis of a functional role of Vg using the available structural data (Paper II)
3. Mapping out the allelic diversity of the *vg*-gene on a global scale and investigating the structural effect. Outlining the functional consequences for the observed variation (Paper III)
4. Combining the results from aims 1-3 to investigate honey bee Vg mechanisms of action (Paper IV)

Methods

Bioinformatics

This research used two computational approaches to predict the tertiary structure of honey bee Vg: homology modeling and AlphaFold. Homology modeling is based on the principle that the protein structure is more conserved than the amino acid sequence. An experimentally resolved structure (template) is selected mainly based on the sequence identity to the protein of interest (target). The sequence identity should not fall below 25 % for the results to be reliable [53]. In addition, the template should ideally be from the same protein family as the target. Next, the target and template sequences are aligned. Multiple sequences should support the alignment to ensure that functional regions and secondary structural elements are correctly aligned [54]. Based on the alignment, the coordinates of all atoms in each amino acid from the template are copied to the target. Side chains and missing regions (for example, resulting from gaps in the alignment) are modeled. Finally, the tertiary structure of the target is refined using, for example, energy refinement and quality control, such as checking that the bond length and angles in the model. The approach was used to predict the subdomains in ND [30, 42] using the crystal structure of lamprey Vg as the template [40, 41]. The continuously increasing protein structure database also allowed me to resolve the vWF domain in this manner (Paper I). I used Swiss-PdbViewer [55] to perform homology modeling interactively.

The full-length structure of Vg was predicted using AlphaFold (v.2), a neural network developed and trained by DeepMind [52]. The network uses the inputted amino acid sequence to predict the distances between amino acids and the angles of their chemical bonds. In addition, related sequences are compiled into a multiple sequence alignment (MSA) that is fed to the network. AlphaFold calculates a confidence score for each amino acid in the prediction, which is used to interpret the model's reliability. Paper I presents the output of honey bee Vg and the confidence in the output, and all four papers use the full-length structure.

The only experimentally resolved representation of honey bee Vg is a low-resolution negative stain EM map. This experimental method fixates the protein using chemicals rather than cryo-temperatures; combined with different hardware, it results in lower resolution surface representations than cryo-EM. The EM map is presented in Paper I and was generated by Elías Herrero-Galán (co-author of Paper I), using *in vivo* samples of honey bee Vg collected by Heli Salmela (previous PhD student and postdoc in my research group). The EM map displays two cavities and was used to validate the AlphaFold structure. I used two methods to rigidly place the full-length structure into the EM map, meaning that no flexibility to either the EM map or the tertiary structure is allowed during the fitting process. Both methods, PowerFit [56, 57] and ADP_EM [58], calculate the correlation between the high-resolution structure and the low-resolution EM map at each point in a grid. The grid size is decided based on the inputted resolution of the EM map (27 Å). The automated correlation searches for all possible relative

rotations and translations at each grid point. The resulting correlation score and number of atoms protruding from the EM map density are used to judge the goodness of the fit. Paper I outlines the details and results.

Apiculture

At my University, I have access to five research hives that are part of a small apiary of mainly purebred, some freely mated, *Apis mellifera carnia*. Claus Kreibich, our research group beekeeper, maintains and cares for the hives. In connection to the apiary, there is a specialized honey bee laboratory. The lab is equipped with the necessary equipment for experiments and the safe handling of the bees. In 2009, the Animal Welfare Act in Norway was updated to include honey bees [59]. The insect's legal standing as livestock is well reflected in our standard operating procedures. Honey bees live in close contact with their nestmates inside a warm (35 °C), dark, and humid (50 %) hive. To obtain protein samples, hemolymph needs to be extracted from honey bees. Vg is the predominant protein circulating in the hemolymph [60, 61]. To take hemolymph samples, honey bees are removed from the hive using soft tweezers and placed in a small cage [62]. Between 20 to 50 honey bees are collected to keep the stress level to a minimum, and the cage is quickly placed in a heating cabinet with the optimal temperature, light, and humidity. The honey bees need to continuously pump the hemolymph to make the sampling possible. Before the procedure, honey bees are placed on ice to avoid stress and potential fleeing or stinging. The ice induces a "chill-coma," a reversible reduced neurological state [63]. The immobilized honey bees are placed under a microscope and pinned down on a waxed plate. As soon as the bees wake up, a thin needle is inserted between the second and third exoskeleton plate on the abdomen, making a small hole. With gentle pressure to the abdomen, tiny droplets of hemolymph weep out and are collected with a pipette (ca. 4 µL). The honey bees are quickly placed back on the ice for a more extended period (more than 4 hours), so the coma becomes irreversible. On average, ca. 1 µg/µL diluted samples of Vg were obtained per honey bee. The samples are pooled, and Vg is purified using ion-exchange chromatography (explained in Papers I and II).

The obtained purified samples of Vg were first used to evaluate the native state of the protein (using blue native polyacrylamide gel electrophoresis and size exchange chromatography, performed by Mateu Montserrat-Canals, co-author of Paper I). This was done to evaluate whether several polypeptides, or monomers, of Vg interact. Purified samples were also used in Paper II to measure the concentration of Zn^{2+}-ions bound to Vg (using inductively coupled plasma mass spectrometry or ICP-MS). Øyvind Enger (co-author of Paper II) performed the instrumental steps and analysis while I prepared the samples and calculated the molecular ratio. This method allows for a quick ionization of the protein sample and a typically precise detection of elements [64]. ICP-MS is also very sensitive and can measure several elements simultaneously. The sensitivity is beneficial for the Vg samples since a small sample volume at a low concentration is sufficient for detecting metal ions. However, zinc is a very common

metal element that increases the risk of sample contamination. Therefore, I included extra steps to ensure a minimum level of contamination. For example, I washed the tubes and containers with acid, avoided using glass or metal equipment, and included negative controls prepared in the same way as the Vg-containing samples.

I also used honey bees to extract gDNA for sequencing of the *vg*-gene (Paper III). My main supervisor and Bjørn Dahle at Norwegian Beekeepers Association helped me contact collaborators connected to different apiaries at several locations with diverse honey bee subspecies. The honey bee samples were collected and shipped by scientists at honey bee research labs or by managers of breeding associations across Europe and the USA. I created a collection kit to facilitate a sampling scheme that was as systematic and low-effort as possible. The kit included a step-by-step guide and the necessary equipment for collection and shipping (Figure 3). I successfully received samples from the 21 outlined apiaries. The gDNA was extracted from the flight muscle (thorax) in honey bees, a DNA-dense tissue [65]. To amplify the *vg*-gene, I used long-range PCR and successfully obtained a full-length gene amplicon (*vg* gene is 6,109 bp). The samples were barcoded with unique primer combinations, making it possible to trace the resulting allele sequences to individual honey bees. This protocol was developed and executed in collaboration with my co-supervisor, Jane Ludvigsen, and the team at Cigene (co-authors of Paper III). The use of Oxford Nanopore sequencing technology generated high-throughput results and allowed for strict error rates.

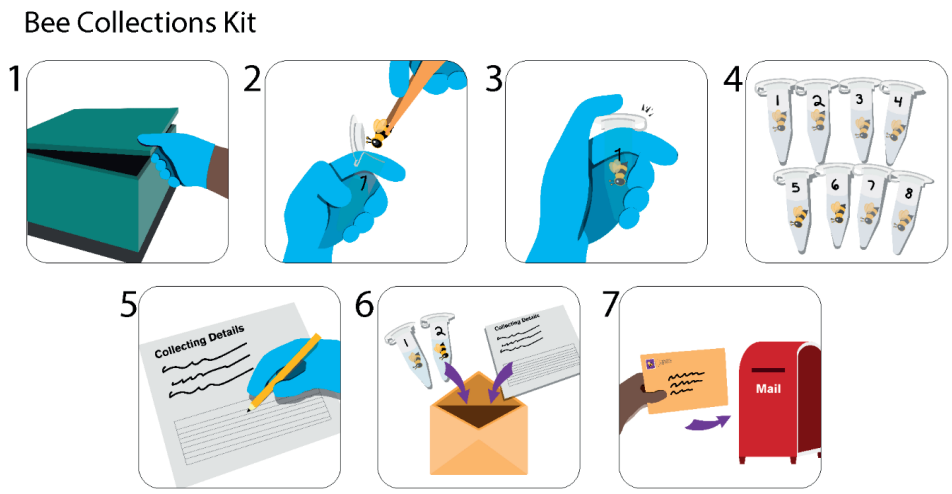


Figure 3: The collection kit included a step-by-step guide. We asked the beekeepers to choose 3-5 colonies and collect an equal number of workers from each, 30 bees in total. The instructions were: 1) Open the hive. Wear plastic gloves while handling to collection kit. 2) Collect 1 worker bee using tweezers, place bee in tube head first. 3) Close the lid until you hear a "click". 4) Open the next tube and repeat until you have collected the planned number of worker bees. 5) Add details to the sheet in the collection kit (such as hive ID and the number of collected worker bees per hive). 6) Place the collected samples and sheet in the return envelope. 7) Ship as soon as possible.

Results

Paper I:

Structure prediction of honey bee vitellogenin: a multi-domain protein important for insect immunity

This paper presents a structural prediction of honey bee Vg using computational approaches, including homology modeling and AlphaFold. A recently published crystallographic protein structure of the D'D3 assembly in human vWF protein demonstrates an appropriate sequence identity to the vWF domain in Vg. The template identification allows us to perform homology modeling of the vWF domain in honey bee Vg. We reveal a conserved Ca^{2+}-ion binding site in the domain that has not yet been described. Next, we used AlphaFold to generate a full-length structure prediction of honey bee Vg. The resulting prediction is estimated to have overall high confidence. AlphaFold was able to predict a structural fold of the C-terminal region. A long loop connects the C-terminal to the vWF domain, and the calculated properties of the loop might suggest a flexible region.

Comparing the structural folds predicted by the different computational methods shows good consistency and demonstrates that the results are reliable. We performed a rigid-body fitting using a low-resolution negative stain electron microscopy map to validate the predicted structural fold and domain assembly. The map is a rough surface representation of Vg and discloses two distinct cavities. We confirm the position of the lipid binding cavity within the protein. The vWF domain appears to be incorporated in the lipid cavity. We also place an additional cavity in the ND that had not yet been identified. The C-terminal region is outside the density barriers, while the full-length structure does not occupy some regions in the density map.

Thus, we present a high-quality structure prediction of honey bee Vg for the first time. Our results are an important contribution to understanding the functional roles of honey bee Vg. The presented structure builds a foundation for further work, and I use the structure in the papers presented below.

Paper II:

Where Honey Bee Vitellogenin may Bind Zn^{2+}-Ions

Here, we first confirm honey bee Vg to be a zinc-carrying protein. The element analysis shows that honey bee Vg can bind on average 3 Zn^{2+}-ions, which demonstrates a high binding capacity compared to earlier reports in other animals.

Zinc is a common structural, regulatory, or catalytic metal ion in many proteins. Its coordination environment and typical binding residues are well characterized. Using the protein structure from Paper I, I predicted seven potential binding sites of Zn^{2+}-ion in honey bee Vg. I also generated an MSA to identify conserved residues, which included a broad phylogenetic range of Vg sequences. The potential sites are presented as clusters that consist mainly of conserved histidine and cysteine residues, but I also looked for conserved aspartate, glutamate, and serine residues.

I identified seven potential clusters at several functional sites: two in the β-barrel subdomain, two in the α-helical subdomain, two in the lipid binding site, and one in the C-terminal region. We decided to look closer at the β-barrel subdomain and attempted to determine the number of Zn^{2+}-ions bound in here experimentally. However, the in vitro system did not provide a clear answer.

Overall, our findings show that honey bee Vg can bind 3 Zn^{2+}-ions on average. Identification of several potential sites suggests that zinc may be important for several activities in honey bee Vg. This paper discusses how Zn^{2+} could influence honey bee health.

Paper III:

Identification of 121 variants of honey bee Vitellogenin protein sequence with structural differences at functional sites

In this paper, we use the dataset of allelic sequence variation for the *vg*-gene to identify 121 Vg variants. The protein variants are identified based on non-synonymous single nucleotide polymorphisms (nsSNPs), which occur in different combinations in the variants. We first examined how the nsSNPs were distributed in Vg. We identify a clear difference between the subdomains and domains. To understand this pattern, we continued to explore the structural impact for each nsSNPs in the different subdomains or domains. The protein structure from Paper I was used for the structural analysis.

Our results showed that the β-barrel subdomain had relatively few changes. We found changes in the same region close to the identified N-terminal cavity from Paper I. In the α-helical subdomain, we identified three hotspots for amino acid substitutions. The first hotspot is close to the changes identified in the β-barrel subdomain. The changes identified here tended to introduce hydrophobic residues. The second hotspot in the α-helical subdomain is located in loop regions close to the lipid binding site, while the third hotspot is slightly buried in the subdomain. The identified substitutions at the second and third hotspot introduced variable amino acids. Similarly, interfacing sites from the lipid binding site were also diverse. The lipid binding site, in general, is a highly diverse region of the protein. The changes did not appear to alter the hydrophobic cavity. The vWF domain is also a highly diverse region, and most substitutions were identified at buried residues. Exposed changes in the vWF domain

occurred at the domain interface to the lipid binding site. The changes here were diverse. We also identified changes in the C-terminal region, which introduced polar residues.

Our findings confirm that the ND is, in general, well conserved. Our study also reflects earlier reports of high diversity in the lipid binding site. Interestingly, we observe a high diversity in the vWF domain. In this paper, we discuss the functional impact of the observed diversity pattern. Our observations point to honey bee Vg maintaining central functions, for example, protein-protein interactions or the proposed DNA binding, while at the same time selectively accommodating for functional regions that recognize pathogens and lipid molecules. Our sequencing approach provided insight into structural variants (such as deletions and insertions) on the *vg*-gene, which will be further explored in future work.

Paper IV:

How honey bee Vitellogenin holds its lipid cargo: A role of the C-terminal

In my final paper, I present a hypothesis concerning a possible mechanism of the C-terminal region in honey bee Vg. The predicted position of the region in the AlphaFold model is not coherent with the EM map density barriers shown in Paper I. The findings suggest possible flexibility in the loop region leading up to the domain that allows for a conformational shift. In Paper II, I identify two highly conserved disulfide bridges crossing each other in the C-terminal region. Formation of disulfide bonds during oxidative conditions or loss during reducing conditions could contribute to conformational change. In Paper III, I identify nsSNPs that often introduce serine residues in this region, increasing the polarity. In Paper IV, I present my previous findings and propose a hypothesis that the C-terminal region could fold over the opening to the lipid binding cavity and cover a large hydrophobic area. Complementary electrostatic surface charges at the C-terminal region and the lipid binding site supports the theory. We discuss the possibility that post-translational modifications, metal binding, and changes in conditions, such as the secretion from the fat body to hemolymph, could influence the proposed activity of the C-terminal. The proposed shielding mechanism could increase the solubility of the protein, which is beneficial for Vg during the uptake, transport, and delivery of lipid molecules. I present how the structural landscape of honey bee Vg has the potential to a large lipid cargo and compare structural features with homologous family members. The theory presented here demonstrates how knowledge of structural features could help better understand proteins' mechanisms and functional consequences.

Discussion

The continuous increase of available structural data and new solutions for predicting protein structures allowed me to present a full-length structure of honey bee Vg and thus complete aim 1. I used two structural templates that were resolved and published within the first year of my project (vWF factor D'D3 assembly, PDB-ID: 6N29 [66] and the LLTP member, microsomal triglyceride transfer protein, PDB-ID: 6I7S [67]). I was also granted the opportunity to include the negative stain EM map of honey bee Vg in my work. When AlphaFold [52, 68] was released, I generated a confident prediction and combined the data to present a detailed representation of honey bee Vg (Paper I). The combination of methods provided a unique insight into honey bee Vg; for example, I could identify the ND cavity and Ca^{2+}-binding site and display the coordinates for every amino acid in the protein. Reaching my first goal provided a solid foundation for the following papers (Papers II, III, and IV) and future work. The model is relevant for species beyond honey bees, as the protein belongs to a phylogenetically broad superfamily and represents any insect's first Vg protein structure.

Having generated a good representation of the structural region of honey bee Vg, I could continue with my next aim and investigate how Vg can carry out its many functional roles. Combining structure, sequence, and experimental data, I demonstrated that zinc could be important for several activities of the honey bee Vg (Paper II). The findings also indicate that zinc might bind or release Vg depending on the situation. If honey bees are exposed to damage or invading pathogens, circulating Vg in the hemolymph could release zinc from the α-helical domain or lipid binding site to promote the activity of immune-related cells. However, in the fat body, the smaller fragment of cleaved Vg could have adopted a zinc-specific fold that might be needed for the proposed DNA binding. Unfortunately, I could not produce experimental proof of zinc-binding to the β-barrel subdomain; nonetheless, the structural, sequence, and motif data are supported by typical activity for zinc-finger proteins and build a logical and encouraging hypothesis. Paper III further supports the seven identified zinc clusters, showing that the cluster-residues are conserved in 543 honey bees at both alleles. With this, I consider the second aim completed, as the study provides a new understanding of how zinc could be related to the activities of Vg.

High-throughput sequencing is also a field under rapid development, considering efficiency, cost, and availability are improving. Collaborating with beekeepers, honey bee researchers, and Nanopore specialists, I gained a unique insight into the *vg*-gene. The reasonably novel methodology enabled me to present the largest reported collection of honey bee Vg variants. I used the structural model from Paper I to present a detailed analysis of how diversity affects the protein structure (Paper III). The findings confirm earlier studies showing the lipid binding site of honey bee Vg to be diverse. Furthermore, the study finds that the vWF domain has a similar pattern, which confirms the Paper I results as an important structural element

in the cavity. The sizeable genetic dataset can reveal more information about the genetic diversity across geographical locations for several *Apis mellifera* subspecies, which completes, and even exceeds, my third aim.

The functional role for honey bee Vg to load, carry, and deliver lipid molecules is well documented on a cellular level, and earlier studies of proteins from the LLTP family indicate the structural features involved in the activity. I combed my findings from Papers I, II, and III to present a hypothesis for honey bee Vg. I confirm the structural features are present in the honey bee Vg model and provide evidence supporting the claim that the C-terminal region is an important functional area (conserved electrostatic surface charge, disulfide bridges, and a flexible loop region). The proposed flexibility and potential conformational changes indicate that Vg is highly dynamic and could have several active shapes. This hypothesis provides the completion of aim four. However, the mechanism that my hypothesis outlines is probably just the tip of the iceberg in terms of the activities of honey bee Vg.

To summarize, my thesis presents a detailed look at the structure-function relationship of honey bee Vg and provides a good foundation for future work (Figure 4). Due to the low yield of the purification protocol and the large size of honey bee Vg, Cryo-EM is a promising method for solving its protein structure and could confirm my findings. I have already started a collaboration with a research group at UiO (co-authors in Paper I), which is associated with the cryo-EM facility at the University of Aarhus, to attempt a structure prediction of Vg. The preliminary results indicate a potential high-resolution structure is in the making. In addition, the genetic dataset gives several possibilities for future studies, as we have information on structural variants, the non-coding regions of the gene, geographical location, and phylogenetic history. Future studies are planned and can give a new perspective on the genetic level while contributing to an understanding of honey bee Vg on a population and ecological level.

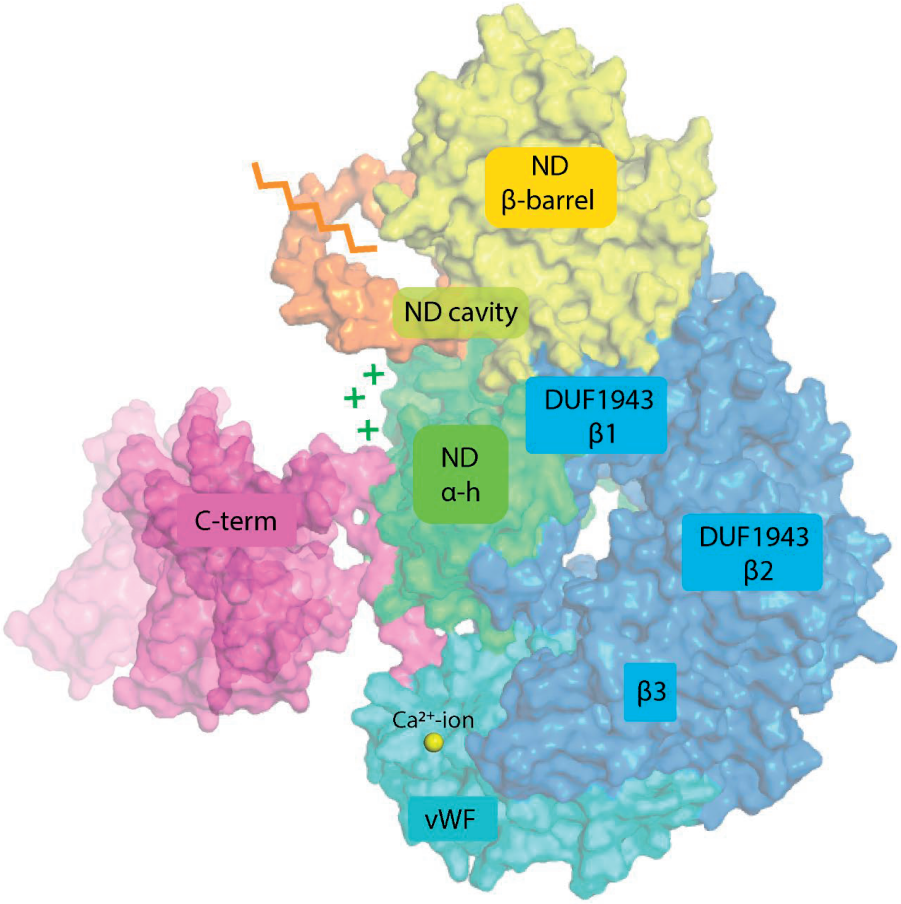


Figure 4: The surface representation of the full-length AlphaFold prediction of honey bee Vg with every subdomain and domain included. The β-barrel subdomain (yellow), the polyserine linker (orange), and the α-helical subdomain (green) make up the ND. The position of the ND cavity is ladled (yellow-green), and the zigzag line illustrates the proteolytic cleavage of the ND. The positively charged surface (green plus signs) on the α-helical subdomain is illustrated. The lipid binding cavity consists of four structural elements, the two β-sheets in the DUF1943 (blue), a third β-sheet (blue), and the vWF domain (cyan). The Ca^{v}-ion is shown in yellow. Finally, the C-terminal (magenta) is shown as a flexible region. This figure is adapted from Figure 1 in Paper III.

Concluding remarks

Studies have extensively examined Vg for several years in many species, but the protein keeps giving. I am proud to have enabled progress in understanding the multifunctional nature of Vg. I am now at the end of my project and feel like a kid in a candy store. For every turn I take in the massive structural landscape of this protein, a new aspect is revealed. My work represents an important step towards understanding the structure of Vg. However, the journey to understand this impressive protein's functions, molecular mechanisms, and properties has just started.

The illustration shows the AlphaFold prediction of honey bee Vg and is a photo taken by Cristofer Bang.

References

1. Marth JD. A unified vision of the building blocks of life. Nat Cell Biol. 2008;10(9):1015-6.
2. Alberts B, Johnson A, Lewis J, al. e. Molecular Biology of the Cell. 4th edition ed: New York: Garland Science; 2002.
3. Petsko GA, Ringe D. Protein Structure and Function. Middlesex House, 34-42 Cleveland Street, London W1P6LB, UK: New Science Press Ltd; 2004. 189 p.
4. Pál C, Papp B, Lercher MJ. An integrated view of protein evolution. Nature Reviews Genetics. 2006;7(5):337-48.
5. Seeley TD. Honeybee Democracy. 41 Wiliam Street, Princeton, New Jersey 08540: Princeton University Press; 2010. 273 p.
6. Weinstock GM, Robinson GE, Gibbs RA, Weinstock GM, Weinstock GM, Robinson GE, et al. Insights into social insects from the genome of the honeybee Apis mellifera. Nature. 2006;443(7114):931-49.
7. Mattila HR, Seeley TD. Genetic Diversity in Honey Bee Colonies Enhances Productivity and Fitness. Science. 2007;317(5836):362-4.
8. Tautz J. The Buzz about Bees: Biology of a Superorganism. 1 ed: Springer, Berlin, Heidelberg; 2008. XIV, 284 p.
9. Jarimi H, Tapia-Brito E, Riffat S. A Review on Thermoregulation Techniques in Honey Bees' (Apis Mellifera) Beehive Microclimate and Its Similarities to the Heating and Cooling Management in Buildings. Future Cities and Environment. 2020;6(1):7.
10. Cremer S. Social immunity in insects. Current Biology. 2019;29(11):R458-R63.
11. Pan ML, Bell WJ, Telfer WH. Vitellogenic Blood Protein Synthesis by Insect Fat Body. Science. 1969;165(3891):393.
12. Wallace RA, Selman K. Ultrastructural aspects of oogenesis and oocyte growth in fish and amphibians. Journal of electron microscopy technique. 1990;16(3):175-201.
13. Li H, Zhang S. Functions of Vitellogenin in Eggs. In: Kloc M, editor. Oocytes: Maternal Information and Functions. Cham: Springer International Publishing; 2017. p. 389-401.
14. Tufail M, Takeda M. Molecular characteristics of insect vitellogenins. Journal of insect physiology. 2008;54(12):1447-58.
15. Tseng DY, Chen YN, Kou GH, Lo CF, Kuo CM. Hepatopancreas is the extraovarian site of vitellogenin synthesis in black tiger shrimp, Penaeus monodon. Comparative biochemistry and physiology Part A, Molecular & integrative physiology. 2001;129(4):909-17.
16. Münch D, Ihle KE, Salmela H, Amdam GV. Vitellogenin in the honey bee brain: Atypical localization of a reproductive protein that promotes longevity. Experimental gerontology. 2015;71:103-8.
17. Corona M, Velarde RA, Remolina S, Moran-Lauter A, Wang Y, Hughes KA, et al. Vitellogenin, juvenile hormone, insulin signaling, and queen honey bee longevity. Proceedings of the National Academy of Sciences of the United States of America. 2007;104(17):7128-33.
18. Sappington TW, S. Raikhel A. Molecular characteristics of insect vitellogenins and vitellogenin receptors. Insect biochemistry and molecular biology. 1998;28(5):277-300.
19. Amdam GV, Norberg K, Hagen A, Omholt SW. Social exploitation of vitellogenin. Proceedings of the National Academy of Sciences of the United States of America. 2003;100(4):1799-802.
20. Amdam GV, Csondes A, Fondrk MK, Page RE, Jr. Complex social behaviour derived from maternal reproductive traits. Nature. 2006;439(7072):76-8.
21. Münch D, Amdam GV. The curious case of aging plasticity in honey bees. FEBS letters. 2010;584(12):2496-503.
22. Amdam GV, Simoes ZL, Hagen A, Norberg K, Schroder K, Mikkelsen O, et al. Hormonal control of the yolk precursor vitellogenin regulates immune function and longevity in honeybees. Experimental gerontology. 2004;39(5):767-73.

23. Seehuus SC, Norberg K, Gimsa U, Krekling T, Amdam GV. Reproductive protein protects functionally sterile honey bee workers from oxidative stress. Proceedings of the National Academy of Sciences of the United States of America. 2006;103(4):962-7.
24. Martin DJ, Rainbow PS. The kinetics of zinc and cadmium in the haemolymph of the shore crab Carcinus maenas (L.). Aquatic Toxicology. 1998;40(2):203-31.
25. Mocchegiani E, Muzzioli M, Giacconi R. Zinc, metallothioneins, immune responses, survival andageing. Biogerontology. 2000;1(2):133-43.
26. Falchuk KH. The molecular basis for the role of zinc in developmental biology. Molecular and cellular biochemistry. 1998;188(1-2):41-8.
27. Baltaci AK, Yuce K. Zinc Transporter Proteins. Neurochemical Research. 2018;43(3):517-30.
28. Andreini C, Banci L, Bertini I, Rosato A. Counting the Zinc-Proteins Encoded in the Human Genome. Journal of Proteome Research. 2006;5(1):196-201.
29. Amdam GV, Aase AL, Seehuus SC, Kim Fondrk M, Norberg K, Hartfelder K. Social reversal of immunosenescence in honey bee workers. Experimental gerontology. 2005;40(12):939-47.
30. Havukainen H, Munch D, Baumann A, Zhong S, Halskau O, Krogsgaard M, et al. Vitellogenin recognizes cell damage through membrane binding and shields living cells from reactive oxygen species. The Journal of biological chemistry. 2013;288(39):28369-81.
31. Zhang S, Dong Y, Cui P. Vitellogenin is an immunocompetent molecule for mother and offspring in fish. Fish & shellfish immunology. 2015;46(2):710-5.
32. Sun C, Zhang S. Immune-Relevant and Antioxidant Activities of Vitellogenin and Yolk Proteins in Fish. Nutrients. 2015;7(10):8818-29.
33. Du X, Wang X, Wang S, Zhou Y, Zhang Y, Zhang S. Functional characterization of Vitellogenin_N domain, domain of unknown function 1943, and von Willebrand factor type D domain in vitellogenin of the non-bilaterian coral Euphyllia ancora: Implications for emergence of immune activity of vitellogenin in basal metazoan. Developmental and comparative immunology. 2017;67:485-94.
34. Wu B, Liu Z, Zhou L, Ji G, Yang A. Molecular cloning, expression, purification and characterization of vitellogenin in scallop Patinopecten yessoensis with special emphasis on its antibacterial activity. Developmental and comparative immunology. 2015;49(2):249-58.
35. Salmela H, Amdam GV, Freitak D. Transfer of Immunity from Mother to Offspring Is Mediated via Egg-Yolk Protein Vitellogenin. PLoS pathogens. 2015;11(7):e1005015.
36. Salmela H, Harwood G, Münch D, Elsik C, Herrero-Galán E, Vartiainen MK, et al. Nuclear Translocation of Vitellogenin in the Honey Bee (Apis mellifera). bioRxiv. 2021:2021.08.18.456851.
37. Smolenaars MMW, Madsen O, Rodenburg KW, Van der Horst DJ. Molecular diversity and evolution of the large lipid transfer protein superfamily^{s⃞}. Journal of Lipid Research. 2007;48(3):489-502.
38. Roth Z, Weil S, Aflalo ED, Manor R, Sagi A, Khalaila I. Identification of Receptor-Interacting Regions of Vitellogenin within Evolutionarily Conserved β-Sheet Structures by Using a Peptide Array. ChemBioChem. 2013;14(9):1116-22.
39. Li A, Sadasivam M, Ding JL. Receptor-Ligand Interaction between Vitellogenin Receptor (VtgR) and Vitellogenin (Vtg), Implications on Low Density Lipoprotein Receptor and Apolipoprotein B/E: THE FIRST THREE LIGAND-BINDING REPEATS OF VTGR INTERACT WITH THE AMINO-TERMINAL REGION OF VTG *. Journal of Biological Chemistry. 2003;278(5):2799-806.
40. Thompson JR, Banaszak LJ. Lipid-protein interactions in lipovitellin. Biochemistry. 2002;41(30):9398-409.
41. Anderson TA, Levitt DG, Banaszak LJ. The structural basis of lipid interactions in lipovitellin, a soluble lipoprotein. Structure (London, England : 1993). 1998;6(7):895-909.
42. Havukainen H, Halskau O, Skjaerven L, Smedal B, Amdam GV. Deconstructing honeybee vitellogenin: novel 40 kDa fragment assigned to its N terminus. The Journal of experimental biology. 2011;214(Pt 4):582-92.
43. Havukainen H, Underhaug J, Wolschin F, Amdam G, Halskau O. A vitellogenin polyserine cleavage site: highly disordered conformation protected from proteolysis by phosphorylation. The Journal of experimental biology. 2012;215(Pt 11):1837-46.
44. Mahbubur Rahman M, Ma G, Roberts HLS, Schmidt O. Cell-free immune reactions in insects. Journal of insect physiology. 2006;52(7):754-62.

45. Sakakibara D, Sasaki A, Ikeya T, Hamatsu J, Hanashima T, Mishima M, et al. Protein structure determination in living cells by in-cell NMR spectroscopy. Nature. 2009;458(7234):102-5.
46. Kovermann M, Rogne P, Wolf-Watz M. Protein dynamics and function from solution state NMR spectroscopy. Quarterly reviews of biophysics. 2016;49:e6.
47. Callaway E. The revolution will not be crystallized: a new method sweeps through structural biology. Nature. 2015;525(7568):172-4.
48. Yip KM, Fischer N, Paknia E, Chari A, Stark H. Atomic-resolution protein structure determination by cryo-EM. Nature. 2020;587(7832):157-61.
49. Merk A, Bartesaghi A, Banerjee S, Falconieri V, Rao P, Davis MI, et al. Breaking Cryo-EM Resolution Barriers to Facilitate Drug Discovery. Cell. 2016;165(7):1698-707.
50. Kuhlman B, Bradley P. Advances in protein structure prediction and design. Nature Reviews Molecular Cell Biology. 2019;20(11):681-97.
51. Kleywegt GJ, Velankar S, Patwardhan A. Structural biology data archiving – where we are and what lies ahead. FEBS letters. 2018;592(12):2153-67.
52. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021;596(7873):583-9.
53. Venclovas C. Methods for sequence-structure alignment. Methods in molecular biology (Clifton, NJ). 2012;857:55-82.
54. Sali A, Potterton L, Yuan F, van Vlijmen H, Karplus M. Evaluation of comparative protein modeling by MODELLER. Proteins. 1995;23(3):318-26.
55. Guex N, Peitsch MC. SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. Electrophoresis. 1997;18(15):2714-23.
56. van Zundert GCP, Trellet M, Schaarschmidt J, Kurkcuoglu Z, David M, Verlato M, et al. The DisVis and PowerFit Web Servers: Explorative and Integrative Modeling of Biomolecular Complexes. Journal of Molecular Biology. 2017;429(3):399-407.
57. Zundert GCPv, Bonvin AMJJ. Fast and sensitive rigid-body fitting into cryo-EM density maps with PowerFit. 2015;2(2):73-87.
58. Garzón JI, Kovacs J, Abagyan R, Chacón P. ADP_EM: fast exhaustive multi-resolution docking for high-throughput coverage. Bioinformatics (Oxford, England). 2007;23(4):427-33.
59. Dyrevelferdsloven. Lov om dyrevelferd (dyrevelferdsloven). 2009.
60. Pinto LZ, Bitondi MM, Simões ZL. Inhibition of vitellogenin synthesis in Apis mellifera workers by a juvenile hormone analogue, pyriproxyfen. Journal of insect physiology. 2000;46(2):153-60.
61. Fluri P, Lüscher M, Wille H, Gerig L. Changes in weight of the pharyngeal gland and haemolymph titres of juvenile hormone, protein and vitellogenin in worker honey bees. Journal of insect physiology. 1982;28(1):61-8.
62. Huang SK, Csaki T, Doublet V, Dussaubat C, Evans JD, Gajda AM, et al. Evaluation of Cage Designs and Feeding Regimes for Honey Bee (Hymenoptera: Apidae) Laboratory Experiments. Journal of Economic Entomology. 2014;107(1):54-62.
63. Macmillan HA, Sinclair BJ. Mechanisms underlying insect chill-coma. Journal of insect physiology. 2011;57(1):12-20.
64. Wilschefski SC, Baxter MR. Inductively Coupled Plasma Mass Spectrometry: Introduction to Analytical Aspects. Clin Biochem Rev. 2019;40(3):115-33.
65. Bruusgaard JC, Liestøl K, Ekmark M, Kollstad K, Gundersen K. Number and spatial distribution of nuclei in the muscle fibres of normal mice studied in vivo. The Journal of physiology. 2003;551(Pt 2):467-78.
66. Dong X, Leksa NC, Chhabra ES, Arndt JW, Lu Q, Knockenhauer KE, et al. The von Willebrand factor D'D3 assembly and structural principles for factor VIII binding and concatemer biogenesis. Blood. 2019;133(14):1523-33.
67. Biterova EI, Isupov MN, Keegan RM, Lebedev AA, Sohail AA, Liaqat I, et al. The crystal structure of human microsomal triglyceride transfer protein. Proceedings of the National Academy of Sciences. 2019;116(35):17251-60.
68. Varadi M, Anyango S, Deshpande M, Nair S, Natassia C, Yordanova G, et al. AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. Nucleic Acids Res. 2021.

Paper I

FEBS openbio

FEBS PRESS
science publishing by scientists

Structure prediction of honey bee vitellogenin: a multi-domain protein important for insect immunity

Vilde Leipart[1], Mateu Montserrat-Canals[2], Eva S. Cunha[2], Hartmut Luecke[3], Elías Herrero-Galán[4,*], Øyvind Halskau[5] and Gro V. Amdam[1,6]

1 Faculty of Environmental Sciences and Natural Resource Management, Norwegian University of Life Sciences, Aas, Norway
2 Norwegian Center for Molecular Medicine, University of Oslo, Norway
3 Department of Physiology and Biophysics, University of California, Irvine, CA, USA
4 Department of Structure of Macromolecules, Centro Nacional de Biotecnología (CNB-CSIC), Madrid, Spain
5 Department of Biological Sciences, University of Bergen, Norway
6 School of Life Sciences, Arizona State University, Tempe, AZ, United States

Keywords
homology modeling; honey bee vitellogenin; rigid-body fitting; von Willebrand factor domain

Correspondence
V. Leipart, Faculty of Environmental Sciences and Natural Resource Management, Norwegian University of Life Sciences, Høgskoleveien 12, 1430 Ås, Norway
Tel: +47 99444807
E-mail: vilde.leipart@nmbu.no

Present address
E. Herrero-Galan, Molecular Mechanics of the Cardiovascular System Cell and Developmental Biology Area, Centro Nacional de Investigaciones Cardiovasculares (CNIC), Instituto de Salud Carlos III, C/ Melchor Fernández Almagro, Madrid, Spain

(Received 27 May 2021, revised 27 September 2021, accepted 18 October 2021)

doi:10.1002/2211-5463.13316

Edited by Cláudio Soares

Vitellogenin (Vg) has been implicated as a central protein in the immunity of egg-laying animals. Studies on a diverse set of species suggest that Vg supports health and longevity through binding to pathogens. Specific studies of honey bees (*Apis mellifera*) further indicate that the *vitellogenin* (*vg*) gene undergoes selection driven by local pathogen pressures. Determining the complete 3D structure of full-length Vg (flVg) protein will provide insights regarding the structure–function relationships underlying allelic variation. Honey bee Vg has been described in terms of function, and two subdomains have been structurally described, while information about the other domains is lacking. Here, we present a structure prediction, restrained by experimental data, of flVg from honey bees. To achieve this, we performed homology modeling and used AlphaFold before using a negative-stain electron microscopy map to restrict, orient, and validate our 3D model. Our approach identified a highly conserved Ca^{2+}-ion-binding site in a von Willebrand factor domain that might be central to Vg function. Thereafter, we used rigid-body fitting to predict the relative position of high-resolution domains in a flVg model. This mapping represents the first experimentally validated full-length protein model of a Vg protein and is thus relevant for understanding Vg in numerous species. Our results are also specifically relevant to honey bee health, which is a topic of global concern due to rapidly declining pollinator numbers.

Abbreviations
BN-PAGE, blue native polyacrylamide gel electrophoresis; CCS, cross-correlation score; DAMPs, damage-associated molecular patterns; DUF1943/1944, domain of unknown function 1943/1944; EM, electron microscopy; fbVg, fat body Vg; flVg, full-length Vg; LC, lower cavity; MSA, multiple sequence alignment; MTP, microsomal triglyceride transfer protein; ND, N-terminal domain; PAMPs, pathogen-associated molecular patterns; QMEAN, Qualitative Model Energy Analysis; SEC, size exclusion chromatography; SPDBV, Swiss-PdbViewer; UC, upper cavity; VADAR, volume, area, dihedral, angle reporter; Vg, vitellogenin; vWF, von Willebrand factor; Ω, gap region.

Vitellogenin (Vg) belongs to an ancient and phylogenetically broad protein family called large lipid transfer proteins [1]. In most egg-laying animals, Vg contributes to oogenesis by providing lipids. Over the last 20 years, studies of several species have demonstrated additional functions of this superfamily in health and behavior [2]. Many animals with one or more *vg* genes are commercially important, and this has incentivized analyses of reproductive and immune traits in which Vg is likely to play a role. Effects of Vg on host immunity have been studied in animals as diverse as bees and fishes [3,4]. For example, Vg recognizes gram-positive bacteria (i.e., *Staphylococcus aureus*, *Micrococcus luteus,* and *Bacillus subtilis*) and gram-negative bacteria (i.e., *Escherichia coli* and *Vibrio anguillarum*) in nonbilaterian coral (*Euphyllia ancora*) and zebrafish (*Danio rerio*) [5,6]. These studies also show that Vg recognizes general bacterial and fungal pathogen-associated molecular patterns (PAMPs). Antimicrobial activity was not detected in these studies, but the interaction promotes apoptosis. Zhang et al. [4] suggest that Vg in zebrafish functions as an inflammatory acute-phase protein leading to elimination of pathogens. This finding also applies to honey bees (*Apis mellifera*) where Vg appears to have similar immunological binding properties [7]. In addition, the Vg molecule of honey bees recognizes damage-associated molecular patterns (DAMPs) [3] and displays antioxidant activity [8–10].

The honey bee is one of the best studied species in terms of the diverse roles of Vg [8,11,12]. For example, this animal was used to show that via their eggs, females can protect their offspring against diseases using a Vg-mediated transfer mechanism: Fragments of bacterial cell walls (immune elicitors) are recognized by Vg and carried out to the honey bee eggs during oogenesis [7,13]. This phenomenon of transgenerational immune priming without the use of antibody-based (i.e., acquired) immunity was first detected a decade ago [14]. However, the underlying mechanisms were not understood before Vg was proposed as a causal element [7]. The availability of the genomic sequence and some functional genetic technologies in honey bees have also enabled studies of Vg's role in behavior [8,15], and such findings have been extended to ants, cockroaches, and mosquitos [16–18]. Honey bees are globally available due to apiculture and can be obtained in large numbers at low costs. Therefore, honey bees provide a practical and useful model for investigating the structure–function relationship of Vg.

In most egg-laying animals, Vg consists of three conserved domains: The N-terminal domain (ND), a domain of unknown function 1943 (DUF1943) and the von Willebrand factor (vWF) type D domain (Fig. S1). In honey bees, the ND is further subcategorized into two structural subdomains, the β-barrel and the α-helical domains, with a highly disordered polyserine region linking these two domains [19] (Fig. S1A). Circulating Vg in the hemolymph of honey bees has a molecular mass of approximately 180 kDa. Vg is cleaved into a 40 and a 150 kDa fragment in the abdominal fat body tissue, the main site for Vg synthesis and storage, and the polyserine linker has been identified as the cleavage site [19]. During investigation of pathogen recognition of Vg in honey bees, the full-length hemolymph Vg (flVg) and the 150 kDa fat body Vg (fbVg) subunit, together with a recombinant peptide of the α-helical domain, were shown to recognize dead and damaged cells [3]. The authors suggest that the heavily positively charged α-helical domain is the main contributor to pathogen recognition. The same study also includes a recombinant peptide of vWF, but this synthetic domain did not show similar binding activity. Studies in fishes and one coral species confirm that the ND can recognize PAMPs and DAMPs but also show that the DUF1943 and vWF can contribute to pathogen recognition [5,6]. Taken together, these findings indicate that Vg may have multiple pathogen-recognizing domains.

In vertebrates and invertebrates, the three main structural domains of Vg are highly conserved at the structural level [5] despite a low nucleic acid sequence similarity [1]. This conservation indicates that the main features of the Vg amino acid sequence are maintained by natural selection. At the level of nucleic acids, the β-barrel subdomain is the most conserved region of the honey bee *vg* gene, while the presumed lipid-binding region (α-helical domain and DUF1943) undergoes positive selection [20]. In a previous study, five residue positions were identified as candidates of functional polymorphisms (marked in Fig. S1A). Local pathogen pressure can be a significant selective force [21–23], and several studies suggest that Vg structure adapts to more efficiently recognize such local threats [7,12]. This hypothesis relies on structure–function relationships that are not fully understood. In fact, there is no complete and detailed structure of the full-length Vg (flVg) protein in any bee, insect, coral, or modern fish species. The only experimentally solved structure is that of lamprey (*Ichthyomyzon unicuspis*) Vg (PDB ID: 1LSH [24]), which consists only of the lipovitellin light and heavy chain (ca. 76% of the sequence is crystallized; Fig. S1B). Using this information as a resource, the conserved N-terminal subdomains (β-barrel and α-helical) in honey bees were

described using homology modeling [3,25] with lamprey Vg as a template. This approach has not been extended to the less conserved DUF1943 domain that is also present in lamprey. The vWF homologous domain, β-Component, is absent from the lamprey crystallographic structure, which eliminates lamprey as a possible template for homology modeling of the vWF domain in other species like honey bees.

Solving the structure of Vg in more species can increase our understanding of ligand interactions and provide important insights into structure–function relationships. However, even in otherwise well-studied species like honey bees, this centrally important information on the DUF1943 and vWF domain is lacking.

Fortunately, the number of experimentally solved protein structures is growing, and the computational modeling software is becoming more powerful. For example, a crystallographic protein structure of the D'D3 assembly in human vWF protein was resolved in 2019 [26], and the VWD3 domain in this assembly has a pairwise sequence identity slightly above 20% to the honey bee domain, which is sufficient to be used as a template during homology modeling.

In this study, we make progress in describing the structure and interpreting the function of the vWF domain in honey bees. In addition, we compile results from template-based, deep learning modeling methods, and the ground-breaking neural network-based algorithm, AlphaFold [27], to present, for the first time, a full-length model for an invertebrate Vg. We combine this new information with published data to begin to elucidate the domain assembly of flVg. Our findings suggest that vWF contributes to the structural organization and has a previously undescribed and valuable function in the protein. This study contributes to the understanding of a protein that is central to life in many animal species.

Materials and methods

Identification of templates

The full-length honey bee Vg sequence (UniProt ID: Q868N5) was inputted to the HHpred [28] server with default settings, which included 'PDB_mmCIF70_23_Jul' as the target database. HHpred returned 250 hits. Each hit was evaluated based on the sequence identity. For the vWF domain, the structural template was verified by performing a BLAST of honey bee Vg (UniProt ID: Q868N5) against the UniProtKB. The target database was restricted to only include UniProt sequences having a PDB ID. The query was run with default settings (*e*-threshold: 10, matrix: auto, filtering: none, gapped: yes, hits: 1000). This BLAST returned 26 hits, and hits from regions already satisfactorily modeled in earlier work were ignored. The remaining hits included the VWF_HUMAN (UniProt ID: P04275, *e*-value 7.2e-1, and 25.0% sequence identity). Residues 1453–1612 of the vWF domain in Vg were aligned to residues 864–1013 of vWF, *Homo sapiens*. These residues correspond to the WD3 domain in the D'D3 assembly in the human vWF protein.

Structural alignment and homology modeling of the von Willebrand factor domain

Both the target and template sequence are part of two larger assemblies, each comprising 4 and 12 domains, respectively. To identify the correct start and end points of the structural alignments, 16 alignments with different sequence lengths were performed. The highest sequence identity (26.3%) was obtained by aligning residues 1440–1634 (target) with residues 836–1031 (template) using the Emboss Needle pairwise alignment tool [29,30], with default settings (Table S1). To ensure that the functional and important regions were aligned correctly, the pairwise alignment was supplemented with a multiple sequence alignment (MSA). The MSA was executed using BLAST and representative Vg sequences from a wider selection of 16 species [3] (Table S2). To ensure a correct alignment of the full-length vWF *H. sapiens* in the MSA and not cause confusion among the four VWD modules in the protein, we referenced the alignment of the modules in the D assemblies from Dong et al. [26] (Fig. 2). The pairwise alignment was altered so that gaps were in the same positions as in the low-conserved regions of the MSA. The highly conserved residues were correctly aligned and were not altered. To avoid gaps in secondary structures or binding sites, the secondary structure annotations from template 6N29 were added to the alignment.

The homology model was interactively built using Swiss-PdbViewer [31] (SPDBV; v. 4.1.0), a recommended approach when building target models with low sequence identity to the template [32]. To initiate the modeling project, the raw sequence (Q868N5) was fitted onto the 3D coordinates of the template (PDB ID: 6N29). Backbone building was performed automatically after editing the alignment as described above. *Ab initio* loop building was performed to ligate breaks in the backbone caused by gaps in the alignment (insertions/deletions). The loop option with the lowest clash and energy scores was chosen in all cases. In this way, nine loops were inserted into the model (Table S3), leaving three unsolved regions (residues 1494–1504, 1515–1522, and 1537–1541) missing in the model. *Ab initio* and database loop building attempts failed to produce a reasonable output for these three 8–11 residue-long gaps. Side chain conformations of target residues aligned to residues with dissimilar characteristics in the template were

identified by detecting clashes and rearranged into the most optimal rotamer option. Rotamer libraries of the most observed orientations for side chains are included in the program. The entire model was energy minimized through a partial implementation of the GROMOS96 force field [33] integrated in the SPDBV software.

Quality control of the von Willebrand factor homology model

Quality control was performed on the model to determine whether the structural features are consistent with the physiochemical rules. Stereochemical consistency was evaluated residue-by-residue using PROCHECK [34]. Global and local quality estimates were performed using the Qualitative Model Energy Analysis (QMEAN) server [35], powered by SWISS-MODEL. The QMEAN output *Z*-score compares the query to similar values based on X-ray structures. VADAR (v. 1.8) [36] assesses the 3D profile, stereo/packing, accessible surface and residue volume. Based on these quality assessments, manual editing was applied to the residues listed in Table S4. The final model was deposited to ModelArchive and can be accessed at: https://modelarchive.org/doi/10.5452/ma-sfueo (access code: okHs98Pcl2).

The Ca^{2+}-ion was copied from the template to the target model, and the contacts to the binding residues were verified to be reasonable in PYMOL (v. 2.2.2) [37]. All illustrations of the model were made in PYMOL.

Full-length structure prediction of honey bee vitellogenin

The alignments from HHpred with the highest sequence identity were selected and forwarded to the implemented modeling software MODELLER [38]. Models 1–8 were built using the query sequences listed in Table 1. All models were built using default settings. A full-length prediction was also built using the RAPTORX web server [39] with the full-length honey bee Vg sequence (UniProt ID: Q868N5) as input, which generated a structure consisting of six domains, each built using one to five templates or template-free modeling (Table S7 and Fig. S7). The models were visualized with the program PYMOL and aligned, and the final model was assembled and built here.

To run AlphaFold v2.0 ([27], see Jumper et al. (2021) supplementary material for detailed description of the method), a P3.2xlagre instance was provisioned from AWS EC2, using the Deep Learning AMI (Ubuntu 18.04) Version 48.0 and a 300 GB disk. Additionally, a 4TB gp3 EBS volume, with 400 $MB{\cdot}s^{-1}$ of throughput and 3000 IOPS, was provisioned and mounted on the machine. The step-by-step guide (README.md, https://github.com/deepmind/alphafold) was followed for setting up and running AlphaFold using Docker. Dependencies that were not included in the AMI were installed manually using the apt package manager. The input sequence was UniProt ID: Q868N5, and AlphaFold was run with the full_dbs preset. Model parameters, downloaded databases, and the output files were stored on the 4TB EBS volume. The run resulted in five models, ranked by average plDDT (Fig. S8B,C). The PDB-file of the top ranked model is included in Appendix S2.

Rigid-body fitting into the electron microscopy map

The high-resolution full-length model and separate chains, in addition to two previously published homology models [3,25] and lamprey Vg (PDB ID: 1LSH) [24], were fitted into the low-resolution negative-stain electron microscopy (EM) map (Fig. S9, EMDB-22113, deposited) without

Table 1. Structure predictions generated by MODELLER and RAPTORX. The table presents all the models generated using MODELLER and RAPTORX (Figs S6 and S7) and lists the region of the amino acid sequence (aa seq.) that has been modeled and which domain it represents. The template used for the model (protein name, species, and PDB ID) and the sequence identity are listed. For Model 9, several templates have been used to generate the full-length model.

Model	Honey bee Vg aa seq.	Honey bee Vg domain	Template	Seq. iden. (%)
1	21–1059	ND and DUF1943	Lamprey Vg (PDB ID: 1LSH_A)	16
2	1190–1515	Undetermined and partly vWF	Lamprey Vg (PDB ID: 1LSH_B)	15
3	1442–1632	vWF	Human vWF (PDB ID: 6N29)	22
4	21–323	β-barrel	Lamprey Vg (PDB ID: 1LSH_A)	19
5	324–360	Polyserine linker	Honey bee Vg (PDB ID: 2ILC)	97
6	361–756	α-helical	Lamprey Vg (PDB ID: 1LSH_A)	19
7	760–1059	DUF1943	Human MTP (PDB ID: 6I7S)	13
8	760–1059	DUF1943	Lamprey Vg (PDB ID: 1LSH_A)	11
9	1–1770	Full-length Vg	PDB ID: 1LSH_A, 1LSH_B, 6RBF_A, 3WJB_A, 4YU8_A, 4JPH_A, 5BPA, 4NT5_A and 2KD3_A	12, 21, 8, 6, 5, 9, 10, 14 and 7

direct human intervention by using the PowerFit webserver [40,41] and the ADP_EM plugin in CHIMERA [42]. In both methods, the resolution was set to 27 Å based on the Fourier shell correlation curve (Fig. S9C), and for PowerFit, the rotational sampling interval parameter was set to 5.00. The PowerFit algorithm uses the cross-correlation between the EM map and the structure to be fitted to search for optimal fits. Output was provided as the structural model's orientation with a corresponding goodness of fit score. ADP_EM works similarly, but is optimized for low-resolution density maps. The fits were imported to the program UCSF CHIMERA (v. 1.14) [43] to optimize them using the volume data 'Fit-in-map' function. This function calculates a correlation score and an average map value both based on map grid points, but the former calculates overlap, while the latter only focuses on the atoms inside the map. In addition, the number of atoms outside the contour is shown. The setting was left as default, but the resolution of 27 Å was inputted. All resulting scores from both software systems are presented in Tables S5 and S6.

CHIMERA and PYMOL were also used to generate the figures of the fits and apply a hydrophobicity scale [44]. The final assembly was imported to PYMOL, where it was aligned to lamprey Vg (PDB ID: 1LSH). The generate symmetry function in PYMOL was used to produce the dimer formation presented by Anderson et al. [53] of lamprey Vg and aligned the final assembly to this structure to present the dimer of honey bee Vg (Fig. 4E). The conserved residues creating polar contacts in honey bee Vg were identified using the MSA produced by MODELLER (not shown). The distances of polar contacts were measured in PYMOL.

Purification of vitellogenin from honey bees

To obtain purified Vg, we collected 1–10 µL honey bee hemolymph in a 1 : 10 dilution in 0.5 M Tris/HCl pH 7.6, using BD needles (30 G) as described earlier [45]. The dilution was filtered using a 0.2 µm syringe filter. Vg was purified from honey bee hemolymph with ion-exchange chromatography using a HiTrap Q FF 1 mL column 0.5 M Tris/HCl as the sample buffer and 0.5 M Tris/HCl with 0.45 M NaCl as the elution buffer. 400–450 µL diluted hemolymph was manually injected and Vg eluted at a conductivity of 15–22 $mS \cdot cm^{-1}$. All fractions from this peak were collected, pooled and concentrated using an Amicon® Ultracel 100 kDa membrane centrifuge filter (Merck KGaA, Darmstadt, Germany). The fraction purity was verified by running SDS/PAGE, which contained only one band of the correct size (~ 180 kDa). The protein concentration was measured with Qubit.

Native gel and size exclusion chromatography

Blue native polyacrylamide gel electrophoresis (BN-PAGE) was performed at 4 °C in precast 3–12% acrylamide gels (Invitrogen, Waltham, MA, USA) for 2 h at a constant voltage of 150 V. The NativePAGE Novex Bis-Tris Gel System (Life Technologies, Carlsbad, CA, USA) protocol was used both for sample and buffer preparation, and Native-PAGE Running Buffer (1×) and the Dark Blue Cathode Buffer (0.4% Coomassie G-250) were used. Size exclusion chromatography (SEC) was performed of Vg in a Superose 6 Increase 3.2/300 column (GE Healthcare, Chicago, IL, USA) at 4 °C equilibrated with a buffer containing 50 mM Tris pH 7.6 and 225 mM NaCl. The SEC was run on an ÄKTA Pure 25 system (GE Healthcare) in micro configuration that allows the use of very small sample volumes. This modification prevents dilution of the sample by effectively reducing the internal volume since it bypasses the multicolumn valve and the pH flow cell and has a shorter path length between the injection valve and the UV monitor. We injected 50 µL of sample (0.26 $mg \cdot mL^{-1}$) and manually collected fractions directly from the outlet of the UV monitor.

Results

Template search

Increased insight into the tertiary structure of Vg's domains is beneficial to our understanding of how Vg contributes to honey bee immunity. To build a full-length structure prediction of honey bee Vg, we first identified potential templates using HHpred [28] (Fig. 1A) with the complete amino acid sequence as input. HHpred indicated that two templates are available for building the ND and DUF1943 domain, one for an undetermined region (residue 1190–1442), and three for the vWF domain. Except for Template 1 (PDB ID: 6N29_A), the sequence identities fall below 20%. By dividing the query sequence into known subdomains and domain boundaries and repeating the search, we generated more specific alignments. The top two ND subdomain templates increased their sequence identities to 19%. In contrast, the DUF1943 was demonstrated to be more distinct compared to human microsomal triglyceride transfer protein (MTP) and lamprey Vg, having sequence identities of only 13% and 11%, respectively.

Homology modeling of the von Willebrand factor domain

Among the three highly conserved domains, the vWF is a major unknown piece in the structural puzzle of Vg. Our initial search discovered a recently published and promising template for this domain, which we confirmed using BLAST [46]. The WD3 domain in the D'D3 assembly of the vWF protein of *H. sapiens* has

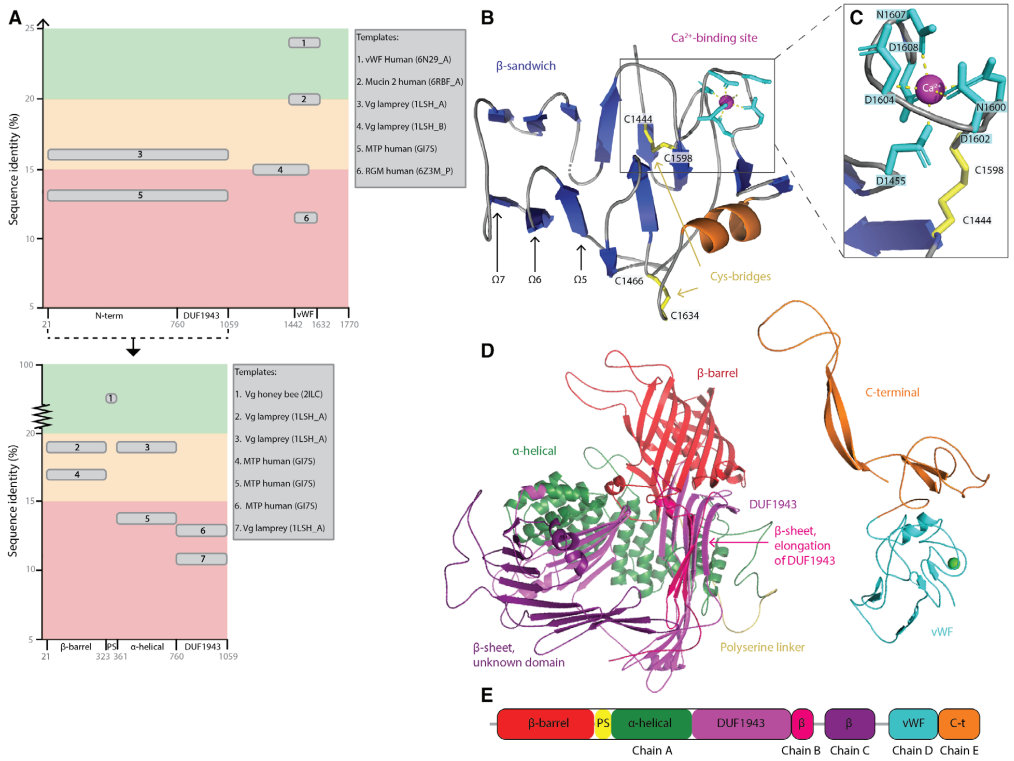


Fig. 1. Structure prediction of honey bee vitellogenin. (A) A graphical illustration of the identified templates using HHpred. On both graphs, the amino acid sequence of honey bee Vg is on the *x*-axis (with the subdomains and domains labeled), and the percentage of sequence identity to the templates is on the *y*-axis. The first graph displays all the templates (gray rounded edge boxes) identified when inputting the full-length sequence of honey bee Vg, while the second shows the templates identified when inputting only the sequence of the separate subdomains. The background colors on both graphs illustrate whether the sequence identity is below 15% (red), between 15% and 20% (orange) or above 20% (green). The templates are numbered according to the sequence identity (highest to lowest), and the protein name, species, and PDB ID are noted in the two large gray boxes. (B) Homology model of vWF: The β-sandwich is on the left side while the Ca^{2+}-segment is on the right side. The Cys-bridges connecting the two segments are shown as yellow sticks and arrows. The β-strands, α-helix and loops are colored blue, orange, and gray, respectively, and the positions of Ω5–7 are labeled with black arrows. The Ca^{2+}-binding residues are shown as cyan sticks, and the Ca^{2+}-ion is shown as a pink sphere. (C) Close-up of the Ca^{2+}-binding site. The coloring scheme is the same as in panel B. All Ca^{2+}-binding residues (D1455, N1600, D1602, D1604, N1607 and D1608) and one of the Cys-bridges (C1598 and C1444) are labeled, and this demonstrates how D1455 from the β-sandwich interacts with the Ca^{2+}-ion. (D) The full-length homology model compiled from several models with different templates. The subdomains and domains are colored as follows: the β-barrel subdomain (red), the polyserine linker (yellow), the α-helical subdomain (forest green), the DUF1943 domain (magenta), elongation of the DUF1943 domain (hot pink), the undetermined structural region (purple), the vWF domain (cyan), and the C-terminal region (orange). (E) A 2D illustration of the chains A to E, used when preforming rigid-body fitting of the homology model.

a sequence identity of the pairwise alignment of 24.1%, which is slightly below the suggested threshold (25%) for creating a reliable homology model [47]. In other words, a pairwise alignment may not be enough to identify gaps and robustly conserved amino acids. We therefore conducted a MSA to confirm gaps and alignment of conserved and domain-defining residues across 12 species, including representative insects, nematodes and mammals. The MSA and the final structural alignment are presented in Fig. S2.

A visual inspection of the structural alignment revealed some interesting aspects. In the almost 200 amino acid-long alignment, the first 40 residues and the last 80 residues are well conserved. In the less

conserved regions, four larger gap regions (Ω) have been introduced (Ω4–7). Ω4 is also missing in all species containing the vWF protein based on the MSA, while downstream Ω5 and Ω7 are conserved in most of the species containing the Vg protein (Fig. S2A). Ω6 seems to be included in all species but is missing in the VWD3, a cysteine-rich domain that forms four intrachain disulfide bridges and two interchain disulfide bridges. The interchain bridges stabilize dimerization of VWD domains in the human vWF protein as opposed to the intrachain bridges formed between cysteine residues inside a single VWD domain. The interchain bridging cysteine residues are not included in the target sequence, and based on the MSA, they are also not conserved in the template domain. However, the eight intrachain bridging cysteine residues are included in the template. Four of these are conserved in the target (C1444, C1466, C1598, and C1634; Fig. 1B). The VWD3 domain also contains a Ca^{2+}-binding site experimentally known from the structural template with key residues also present in the target sequence [26]. We recognize this as a class II calcium binding site because the coordinating residues, as well as the neighboring residues, make up two short regions [48] (r. 1453–1456 and r. 1596–1609; Fig. S2) that are well conserved among all species in the MSA. This indicates an essential site for function and/or stability of the domain. We conclude that the significant regions for domain function or stability, the intrachain disulfide bonds, as well as the Ca^{2+}-binding residues, are conserved and correctly aligned. We also conclude that the MSA was able to identify robustly conserved features of Vg, and we therefore proceeded with interactive homology modeling using the structural alignment provided by the MSA (Fig. S2B).

The amino acid sequence of the target was fitted onto the three-dimensional coordinates of the template using the structural alignment. Breaks in the backbone were ligated using loop building, and the side chains of nonconserved residues were rearranged to the most optimal rotamer orientation, reducing the number of steric clashes. Finally, we performed energy minimization to release local backbone strain and electron density clashes. The overall quality of the target model was validated using several software tools. To account for sequential errors, we also included the quality scores of the template (Figs S3 and S4). Based on the results, the backbone phi and psi angles of 14 residues, detected as outliers by Ramachandran analysis (Fig. S3C) [49], and rotamers of 19 residues, detected by PROCHECK, were manually edited (Table S4). The main limiting factor for the quality metrics of the model were the errors already listed as well as the presence of the longer gap regions. It was not possible to include Ω5–7 in the model because this creates a region with too many unfavorable interactions and torsion angles. However, these regions exhibit low conservation (Fig. S2). The local quality estimate by SWISS-MODEL (Fig. S3B) shows that the middle region is of lower quality relative to the first and last missing regions. The Ca^{2+}-binding residues and intrachain disulfide bonds are in higher-quality regions. The PROCHECK summary shows that the main difference between the target and template models originates from the calculated stereochemical parameters (geometry, bad contacts and bond length and angles; Fig. S3A). The residue-by-residue list produced by PROCHECK (Fig. S4E) identified residues deviating from the ideal values. However, these residues were altered during loop building, often resulting in an unfavorable orientation for the chosen residues [50]. We conclude that key structural features of the target are modeled correctly except for the low-quality middle region that contains residues with stereochemical parameters deviating from the ideal values. The homology modeling approach used has a proven track record of producing models of sufficient quality when facing similar challenges [51]. We demonstrated this by comparing our model to an automatically produced model by MODELLER. We find that in our model, the local quality is better in the regions of low conservation (Fig. S3B), and the global quality is higher (Fig. S4A–C). For the conserved region, our interactive modeling approach achieves a better result by including C1634, which creates an intrachain disulfide bond, two additional β-strands and a more appropriate rotamer option for the Ca^{2+}-binding residue N1607 (Fig. S5).

We are thus for the first time able to present a detailed structural model of the vWF domain of honey bee Vg. The structure can be understood as two segments: one consisting of 11 antiparallel β-strands organized into a β-sandwich while the other is comprised of the Ca^{2+}-binding site, a short α-helix, and three short β-strands (Fig. 1B,C). Connecting the two segments are the two intrachain disulfide bonds. The two segments are also connected through the Ca^{2+}-binding site via the interaction of residue D1455 (Fig. 1C). The Ca^{2+}-binding residues are in loop regions (i.e., normally flexible regions), but we suggest that binding of a Ca^{2+}-ion might confer stability to this region. The Ca^{2+}-binding segment of the domain exhibits higher quality than the antiparallel β-sandwich. Despite the lower quality, the residues in the secondary structure elements exhibit a higher local quality score compared to the residues in the loop regions (Fig. S3B). We conclude that the β-strands are organized in a sterically

reasonable manner, while the loop regions are most likely not described accurately.

Full-length structure prediction of honey bee vitellogenin

We performed template-based prediction of the remaining domains of honey bee Vg using the integrated MODELLER software in HHpred. We generated eight models using different sections of the honey bee Vg amino acid sequence as input (Table 1). By aligning the predicted models covering the same domains (Fig. S6), we observed that the general fold is the same except for models describing DUF1943 (Models 1, 7, and 8; Fig. S6B). Using human MTP as a template returned a straight β-sheet with fewer and longer β-strands. In addition, we also used the deep learning modeling method RAPTORX to generate a full-length and complete prediction (Fig. S7). The model is mostly based on nine different templates with sequence identity ranging from 5% to 21% but also includes regions resulting from deep learning predictions. The total model assembles all predicted domains like pearls on a string and cannot predict how they are organized relative to each other. However, the general fold of each model is consistent with the results from MODELLER (Fig. S6A–E). We built the final structure using Model 1 for residues 21–1059, Model 9 for residues 1060–1140, Model 2 for residues 1190–1408, the vWF homology model from Quality control of the von Willebrand factor homology model for residues 1440–1634 and Model 9 for residues 1635–1770. We selected these models based on whether their fold were consensus folds and removed the long, extending loop regions. The final model has 93.1% sequence coverage of honey bee Vg and includes the conserved domains (ND, DUF1943 and vWF) in addition to undetermined regions now structurally described for the first time for an invertebrate Vg (two β-sheets downstream of DUF1943 and the C-terminal region; Fig. 1D). Based on the compilation of models, the final prediction was divided into chains A (the ND), B (the β-sheet from Model 9), C (the β-sheet from Model 2), D (the vWF domain) and E (the C-terminal region) as presented in Fig. 1E.

The very recent publication and code availability for AlphaFold v2.0 [27] enabled us to produce a structure prediction of honey bee Vg. The first step of the pipeline is to produce an MSA, and the resulting number of hits can indicate the prediction accuracy. The developers observe a decrease in prediction accuracy when the alignment depth falls below 30 sequences and an increase of accuracy until 100 sequences, where they observe a threshold effect [27]. The honey bee Vg MSA have an average of 1988 hits per residue (Fig. S8A), suggesting a high prediction quality. The resulting AlphaFold models had an average predicted local distance difference test (plDDT) ranging from 81.7692 to 84.5747 (Fig. S8B), which is a per-residue estimate of confidence [27,52]. The highest-ranking model colored by the plDDT confidence scale (Fig. 2A) shows a generally confident backbone prediction of honey bee Vg. Some regions fall below 70, which the developers of AlphaFold state should be treated with caution, and these residues map to short loops in domains or longer flexible segments in-between domains (Fig. 2B). The developers state that plDDT residue scores below 50 strongly indicate disorder which in our case is consistent with our knowledge of the protein. The very low scoring residues 341–380 (average pIDDT: 33.1242) map to the polyserine linker, which is known to be flexible and disordered [19]. Similar disorder is predicted for the N-terminal signal peptide residues 1–17 (average plDDT: 47.8064) and the segments upstream and downstream of the vWF domain, residue 1425–1437 and 1674–1684 (average plDDT: 44.5930 and 42.9336), respectively. Aligning the top ranking AlphaFold predictions demonstrates a consistent fold for the confident regions and some inconsistency of the low-confidence regions (Fig. S8C). The predicted disorder of residues 1674–1684 results in a variable positioning of the downstream C-terminal region between the predictions, suggesting flexibility of the domain position.

The final homology model and the AlphaFold prediction agree on the fold of the stable domain (Fig. S8D). AlphaFold produces 3D coordinates for every atom in the protein, so the prediction takes up more space, compared to the homology model where there are missing atoms, particularly downstream of the DUF1943 domain (Fig. S8D). However, the overall consistency in both of our predictions confirms that our structural prediction is strong.

Using PowerFit, ADP_EM, and Chimera to determine the domain assembly of full-length vitellogenin

The full-length models of Vg indicate the general fold of each domain. However, the domain assembly in the final homology model is speculative and derived from lamprey Vg and the deep learning method along with strong biases. To reduce these biases and provide some validation of the structural assembly, we performed rigid-body fitting of our model to a low-resolution EM map (Fig. S9, EMDB-22113, deposited) of *in vivo*-

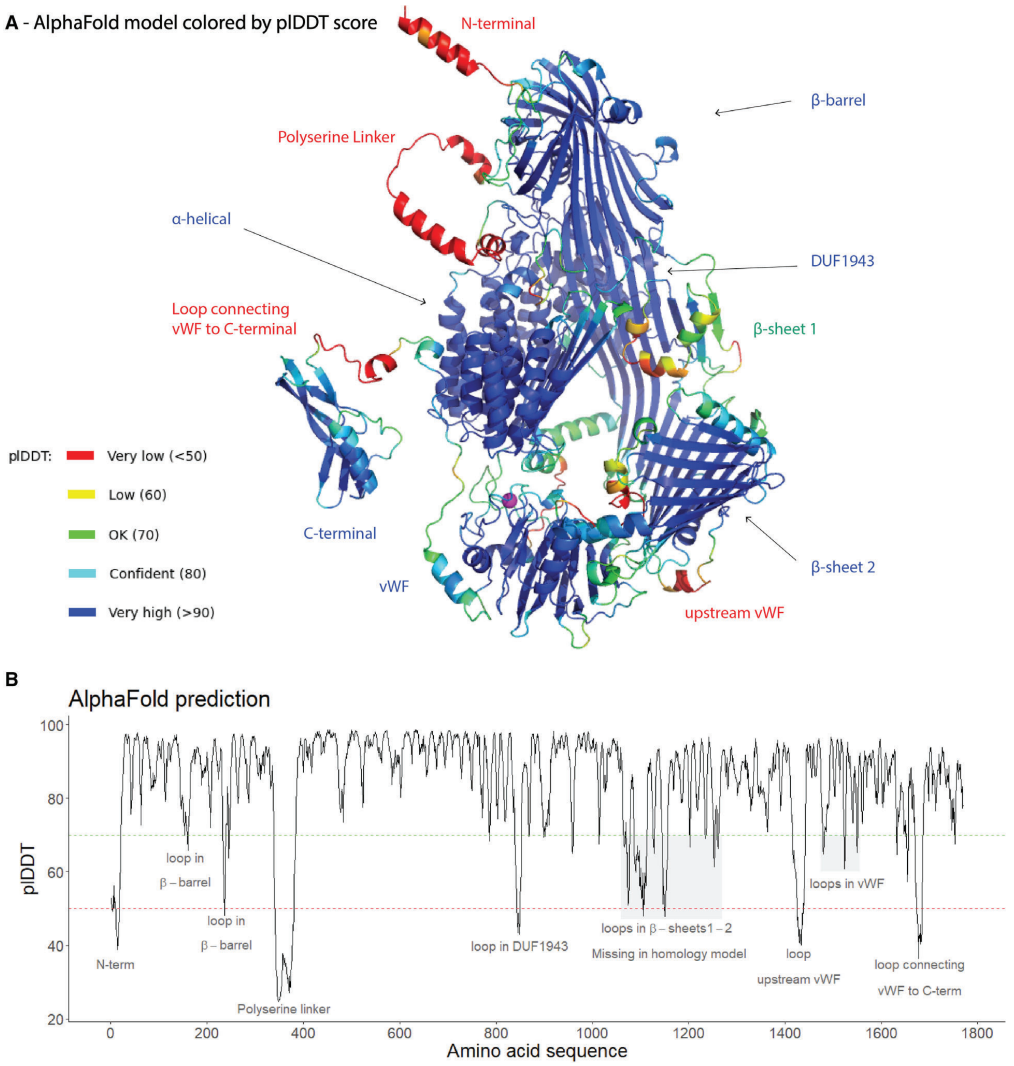


Fig. 2. AlphaFold prediction. (A) The top ranked AlphaFold model is shown as cartoon, colored by the plDDT scale. The high scoring domains (β-barrel, α-helical, DUF1943, β-sheet 2, vWF domain, and C-terminal domain) are labeled in blue, while the medium confident region (β-sheet 1) is labeled in green, and the low confident regions (N-terminal, polyserine linker, the segment upstream and downstream of vWF domain) are labeled in red. The Ca^{2+}-ion is shown as a magenta sphere. (B) The plDDT score is plotted per residue for the top ranked AlphaFold model. Each region that scores below 70 (green dotted line) is labeled. The very low plDDT (< 50) is indicated with a red dotted line.

obtained honey bee Vg. The EM map reveals a rough overview of the surface and two distinct cavities, hereafter named top, base, left and right, upper cavity (UC) and lower cavity (LC) in reference to this specific orientation (Fig. S10A). Fitting of the complete homology model placed chains D and E consistently outside the contour map, while chains A to C did not take up all the available space inside it (Fig. S10C).

This indicates incorrect domain assembly of chains D and E. Fitting of the RAPTORX structure gave similar results leaving chains C, D, and E outside the contour map, clearly demonstrating improper domain assembly (Fig. S10D). To avoid problems related to template-based assembly, we fitted the chains individually. Chains A and D occupy somewhat separate parts of the contour, but chain A overlaps with chain C and partly chain B and E (Fig. S10E,F). These individual domain fits support the assembly of chain A to C in the predicted model and further suggest improper assembly of chains D and E. Keeping chains A to C united but chains D and E separate resulted in two alternative orientations (Fig. S11B,C) leaving out chain E, which is not compatible with either alternative (Fig. S10F). The first 68 residues of chain E were built using a template-free method, while the last 58 residues were compiled from a multiple alignment of the last five templates (Table S7) ranging from 5% to 14% sequence identity. HHpred recognizes none of these templates. Faced with a speculative prediction and its incompatibility with the EM map, we removed the C-terminal domain from the domain assembly. The resulting fits from two independent rigid-body fitting methods (PowerFit [40,41] and ADP_EM [42]) was optimized using CHIMERA fit-in-map [43], producing correlation scores that could be compared directly (Figs S11A, S10B, and S12A). The highest scoring fit of chain A to C from ADP_EM is overlapping perfectly with the second-best fit from PowerFit (Fig. S11B1), while the highest scoring fit of the same chains from PowerFit is agreeing with the relative orientation of the domains. The best fit from Powerfit is not overlapping, however, with the second-best fit from ADP_EM (Fig. S11B2). The correlation score for the second ADP_EM fit is lower, and more atoms are outside the contour, compared to the other fits. Both alternatives are compatible with the ADP_EM and the PowerFit orientation of chain D (Fig. S11C). Secondary structure elements from the α-helical subdomain and DUF1943 are protruding outside the contour for both alternatives. For alternative 2, the DUF1943 and additionally the β-barrel subdomain are seemingly restricting access to both cavities (Fig. S11D).

To further investigate the two alternatives, we fitted previously generated homology models of the β-barrel and α-helical domains of honey bee Vg [3,25] and the X-ray structure of lamprey Vg (PDB ID: 1LSH [24]) to the EM map. The respective or homologous domains consistently fit in the two relative orientations and scored high values for both alternatives (Fig. S12). The β-barrel and α-helical domain supported alternative 1, while lamprey Vg favored alternative 2 according to the scores. The EM map is an *in vivo* representation of honey bee Vg, while the 1LSH structure is a distant homologue with 24% of the sequence missing in the crystal structure. The AlphaFold prediction with 100% sequence coverage serves as a far better representation of honey bee Vg. Fitting the top ranked AlphaFold prediction resulted in two different orientations by selecting the highest scoring fit from PowerFit and ADP_EM, respectively (Fig. 3A). The best fit from PowerFit has fewer atoms outside the contour and a higher correlation score, compared to the best fit from ADP_EM (Fig. 3B). The very low-confidence fold of the N-terminal signal peptide and the polyserine linker is protruding in both alternatives (Fig. 3C,D). In addition, smaller loops with a fold confidence ranging from low to intermediate are also protruding in both fits but these mismatches between model and contour map are more pronounced in the ADP_EM fit (Fig. 3D). The model cavities are restricted in the ADP_EM fit by the β-barrel and a long β-sheet which is the AlphaFold prediction of a more complete chain C, and these domains are confidently modeled. Both cavities in the PowerFit fit are also somewhat restricted by in-between domains segments, which have a lower confidence fold. Taken together, the orientation represented by PowerFit is the best fit of the AlphaFold prediction. This orientation also conforms to the best fits of individual domains: chain A to C (Fig. S11B2), chain D (Fig. S11C, PF1), β-barrel (Fig. S12B, ADP2), α-helical (Fig. S12C, PF2 and ADP1) and lamprey Vg (Fig. S12D, PF1 and ADP1). This further supports the PowerFit orientation of the AlphaFold prediction, but now with a more optimized fit. Using the full-length sequence representation results in a structure which fills more of the density space while keeping the percentage of protruding atoms low and the correlation score high. This suggests that the domain assembly in the AlphaFold prediction is an accurate representation of honey bee Vg.

The final model is presented in Fig. 4. The LC serves as the better-known lipid-binding site. It is easily accessible, while the hydrophobic core is buried in the EM map (Fig. 4A). The UC is partly built up by the β-barrel. The vWF domain is placed close to the LC bringing the Ca^{2+}-ion into close proximity to the cavity (Fig. 4B). This is supported by the results produced by the Volume, Area, Dihedral Angle Reporter (VADAR; Fig. S4D). The fractional accessible surface area report shows that the two short β-strands downstream of the Ca^{2+}-binding site are reported as exposed (r. 145–156 in plot 1, Fig. S4D).

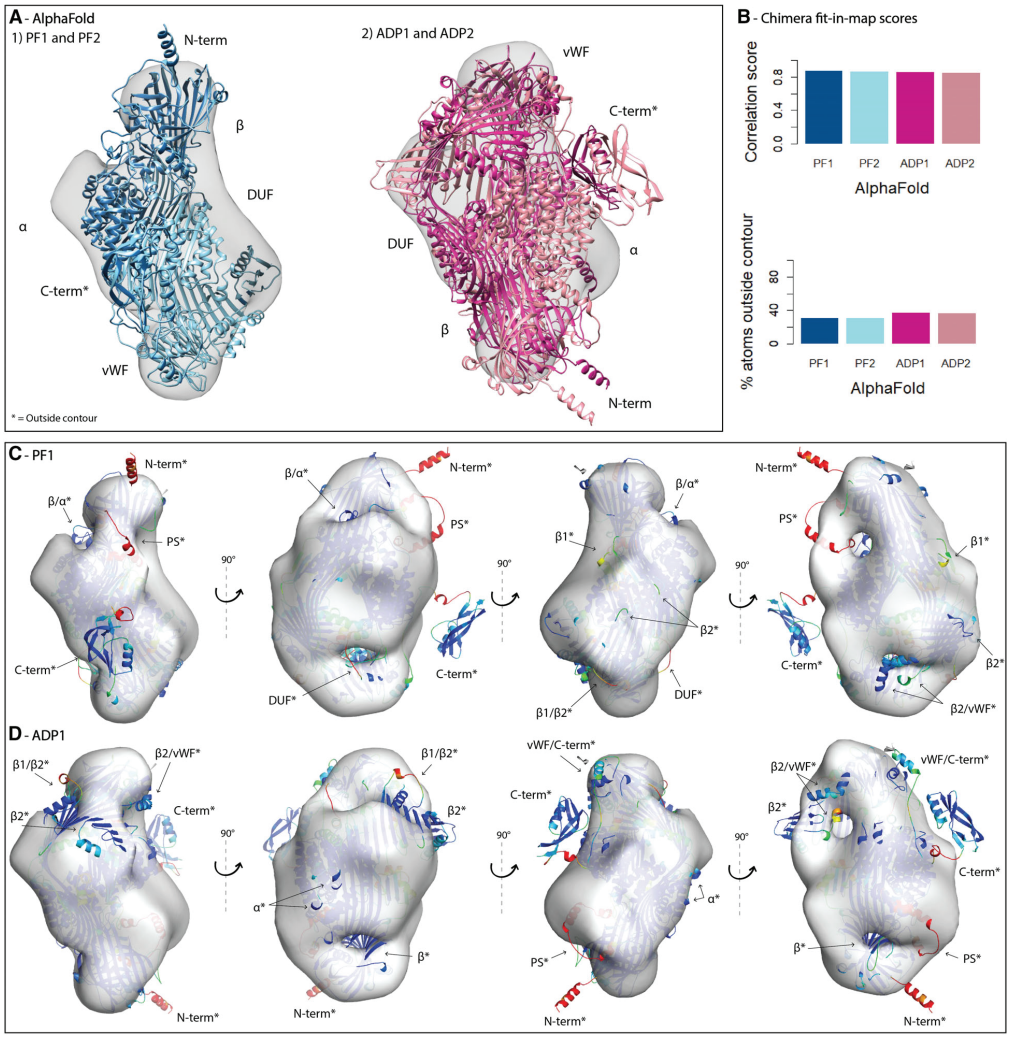


Fig. 3. Rigid-body fitting of AlphaFold. (A) The EM map are shown as a transparent surface, and the fits of AlphaFold from PowerFit (PF) and ADP_EM (ADP) are shown as cartoons and colored by method and scores (dark blue: PF1, light blue: PF2, dark pink: ADP1, light pink: ADP2). The N-terminal (N-term), β-barrel (β), α-helical (α), DUF1943 (DUF), vWF domain (vWF), and C-terminal (C-term) domains are labeled. (B) The correlation score and percent of atoms outside the contour calculated by CHIMERA were plotted for each fit from PowerFit (PF, blue) and ADP_EM (ADP, pink), and ranked according to the correlation score (dark color: highest score, light color: second highest score). (C) The EM map and the highest ranked PowerFit fit of AlphaFold is shown in at four different angels, colored by plDDT score. The label is marked with '*' if residues are outside the contour of the EM map and '/' between domain labels indicate that the pointed to segment is in-between domains. The polyserine linker and the two β-sheets downstream of the DUF1943 domain are labeled PS, β1, and β2, respectively. (D) The EM map and the highest ranked ADP_EM fit of AlphaFold. The same coloring and labeling are used as in panel C.

The fractional residue volume plot reports a potential cavity in the vicinity of the Ca^{2+}-binding site. In addition to the hydrophobic regions of Vg to be buried in the two cavities, the previously established hydrophilic and positively charged side of the α-helical domain [3] faces the surface in our model,

providing further support for a correct assembly. The polyserine region is also very exposed, favoring the reported dephosphorylation and cleavage events [19]. In the final model, we also mapped out residue positions of interest (Fig. 4C,D). The five functional polymorphisms are in association with a cavity (three in the lipid-binding site and two in the vWF domain). Anderson et al. [53] specified 12 polar interactions among nine residues on each monomer of lamprey Vg. Seven of these residues are conserved in honey bee Vg, and mapping these to the final model shows them to be accessible to solvent. Simulating the dimerization in PYMOL with the final model confirms dimerization to be a feasible oligomeric arrangement for honey bee Vg (Fig. 4E). However, re-fitting the Vg dimer in the EM map results in 33–39% of the atoms inside the contour (Tables S5 and S6). Taken together, this further supports the predicted assembly and demonstrates the EM map to be a representation of monomeric honey bee Vg.

Vitellogenin oligomerization state

While lamprey Vg forms a dimer with a modest 245 Å^2 hydrophobic interface in the crystal structure [24], mixed evidence exists for the oligomerization status of honey bee Vg. As described above, the negative-stain EM map with a resolution of 27 Å supports Vg to be monomeric since only one Vg molecule can be placed in the EM map, even at low contouring level. However, the sole known experimentally solved structure suggests that Vg can appear as a dimer [53], at least under some conditions. To further investigate this, we obtained purified Vg from honey bees and evaluated two different amounts using BN-PAGE (Fig. 5A). The lower molecular weight band (151 kDa) constitutes most of the material in the sample and is assumed to be monomeric Vg. The additional weaker band with higher molecular weight (345 kDa) is assumed to be a minor fraction of dimeric Vg. Contamination by other proteins in the sample seems

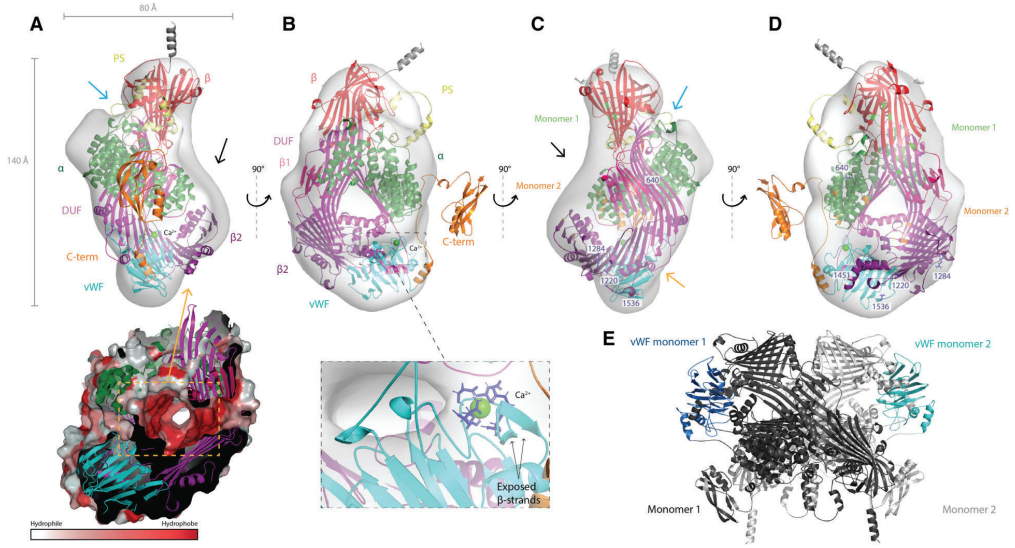


Fig. 4. Honey bee vitellogenin final assembly. The EM map is shown as a transparent surface from four different angles and have the AlphaFold model fitted inside. The polyserine linker (PS, yellow), β-barrel (β, red), α-helical (α, green), DUF1943 (DUF, magenta), β-sheet 1 (β1, hot pink), β-sheet 2 (β2, purple), vWF domain (vWF, cyan), and C-terminal (C-term, orange) domains are labeled, as well as the UC (blue arrow), LC (orange arrow), and empty density (black arrow). (A) The measurements of the EM map are shown along the x- and y-axis. The surface of the LC, colored by Eisenberg hydrophobicity scale [44], is shown inside the orange dashed box surrounded by the domains building up the cavity. (B) Here, we zoom in on the Ca^{2+}-binding sites, and show the two exposed β-strands (black arrows) and their proximity to the LC. (C, D) The five residue positions (640, 1220, 1284, 1451, and 1536) identified as candidates of functional polymorphisms are colored blue and labeled. The conserved residues in honey bee Vg that make polar contacts during dimerization are colored green (monomer 1) and orange (monomer 2). (E) The simulated Vg dimer is shown with monomer 1 (dark gray) and 2 (light gray). The vWF domain is colored in each monomer (monomer 1, dark blue and monomer 2, cyan).

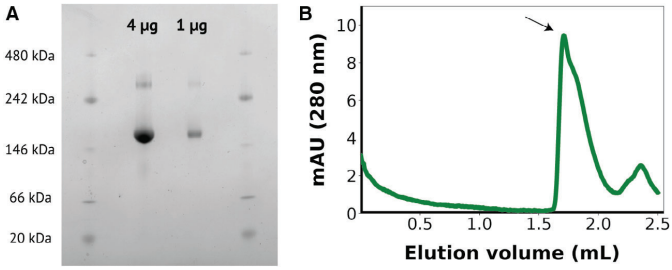


Fig. 5. *In vitro* oligomerization state analysis of vitellogenin. (A) BN-PAGE gel results. Both the bands corresponding to the monomer and the dimer can be observed for Vg loaded in different amounts. (B) SEC elution profile for purified Vg. The peak containing Vg is labeled with an arrow corresponding to an elution volume consistent with monomeric Vg.

unlikely since only one band for Vg can be observed from the sample in a denaturing PAGE (not shown). Next, we performed SEC (Fig. 5B), and the content of the concentrated fractions was analyzed with BN-PAGE (not shown). The main peak obtained corresponded to monomeric Vg, and its apparent molecular weight was estimated to be 178 kDa based on the elution volume. No peak corresponding to the dimeric form was obtained, although when the fraction from the main peak was concentrated, it showed on a native blue PAGE both as a monomer and a dimer in similar proportion to that observed in Fig. 5A. Together, these results suggest that Vg can dimerize at higher protein concentrations *in vitro*.

Discussion

With this study, we aimed to gain more insight into the structure of honey bee Vg and to attempt a full-length model of the protein. Our results reveal structural features that have not yet been described for Vg in invertebrates.

First, we presented a detailed structural prediction of the vWF domain. Through homology modeling, we identified a potential class II Ca^{2+}-binding site, which appears to be highly conserved across Vg and vWF-containing species. The Ca^{2+}-ion coordinates with 4 Asp and 2 Asn residues, through their OD1 or OD2 atoms, respectively, except for D1604, which coordinates through its main chain carbonyl O-atom. In the human WD3 domain, the residue corresponding to D1604 is I1002 (Fig. S2). The side chain of isoleucine is unable to interact meaningfully with calcium [54]. We speculate that the introduction of a sixth calcium-coordinating residue, aspartate, creates an additional bond to the Ca^{2+}-ion, increasing the interaction and strengthening the coordination. Identifying a total of six coordinating residues and a loop structure in the binding site enabled us to categorize this as a class II site [48].

We were able to present a full-length structure prediction of an invertebrate Vg. However, our concern about the remaining domains is that the use of distant homologues with low sequence identity can create predictions influenced by the template used. Studies show that general protein folds are well conserved across great phylogenetic distances despite low conservation of the amino acid sequence [55]. Focusing mainly on the general fold and creating several models by using different query sequence lengths, we increased our confidence in the prediction for each domain. The striking similarity between the AlphaFold prediction and the predicted homology model chains validates our modeling results. In addition, AlphaFold provides a confident domain fold of the C-terminal region, and predicts folds for loop regions missing in the homology model, enabling us to present a 100% complete structure representation of honey bee Vg, with considerable confidence within each domain. Using PowerFit, ADP_EM and CHIMERA, we were able to present a domain assembly of the full-length structure prediction. The negative-stain EM map has a low resolution (27 Å), which increases the margin of error. To limit the number of possible orientations, we fitted the homology models according to size, beginning with the largest. We also fitted the previously predicted domains, the crystal structure of lamprey Vg and the AlphaFold prediction to validate our modeled fold and its placement in the EM map. We evaluated each fit based on the scoring, protruding atoms and overlapping fits of separate domains. We concluded that the AlphaFold PowerFit orientation, with the DUF1943 domain, the two downstream β-sheets and vWF domain oriented around the LC and the β-barrel

and α-helical subdomain toward the UC (Fig. 3A1), was the most probable representation for honey bee Vg. The energetics for the full-length model and the separate domains (e.g., whether polar surfaces or hydrophobic surfaces were exposed to the solvent) are logical, as demonstrated for the lipid binding site (Fig. 4A). The final model does not occupy all available density while the C-terminal region is outside the contour, which represents about 4.6% of the atoms. The position of this domain is not clear as the Alphafold results indicate a flexibility in the connecting loop. The unassigned density in the low-resolution EM map above the UC could potentially be where the C-terminal region is positioned (Fig. 4A,C). Honey bee Vg is also found to be phosphorylated and glycosylated, [25] which is not represented in the protein structure and could explain the excess of density.

Both cavities identified in the EM map are compatible with the assembly, and the LC is identified as the lipid-binding site, which recognizes lipids, possible fragments of gram-negative and gram-positive bacteria [24]. The UC, built up partly by the β-barrel subdomain, has not been described earlier, and whether the UC has similar recognition potential, to the LC is not known. The *in vitro* mutagenesis experiments performed for the human vWF protein [56] illustrate the importance of the Ca^{2+}-binding site for recognition of factor VIII in a blood-clotting cascade. A study from 2013 shows fbVg to be membrane associated and speculates the receptor binding site to be in the 150 kDa subunit and not in the β-barrel domain as previously believed [3]. Insect Vg receptors belong to a subfamily of the low-density lipoprotein receptor family, and calcium interaction has been shown to be essential for ligand association [57,58]. Our findings support these results and suggest the vWF domain as the potential Vg receptor binding site. Additionally, the vWF domain has been implicated in having adhesive and lubricant properties [59,60] as seen for vWF and mucin proteins in humans. The structure of the WD3 domain, used as template here, was recently functionally compared to the MUC2 in humans. Since the two proteins shows high structural similarity, Javitt et al. [61] suggest that WD3 has a similar polymerization function and is essential for macromolecular assemblies in the epithelial mucosa and vasculature. Our study shows that the interchain disulfide bonds, essential for oligomerization in the human vWF [26,56], are not conserved in honey bees. In addition, residues in the β-barrel and α-helical domain are interacting in the Vg dimer, and not the vWF domains (Fig. 4E), thereby ruling out this kind of polymerization activity for the vWF domain in honey bees. However, the Ca^{2+}-binding site, the intrachain disulfide bonds and the β-sandwich are highly conserved, suggesting a similar function in mucosal immunity, as seen for mucins and vWF proteins in humans.

Insects, which have an open circulation system, have developed an efficient coagulation mechanism that is an essential part of their innate immune system [62]. When exposed to invading microbes, a clotting cascade is initiated, trapping and eventually killing the invaders [63]. The hemolymph clot was recently characterized in a Brazilian whiteknee tarantula, showing the main content to be proteins encompassing vWF-like domains. Sanggaard et al. [64] results also indicate that the clot functional and structural overlaps with such clots observed in insects. We propose that honey bee Vg can initiate or aid in this clotting mechanism, interacting through the vWF domain, and protect honey bees from pathogens and mechanical damage, like in zebrafish Vg [4]. Our identification of three residue positions exhibiting high genetic differentiation in the LC could be a result of adaption to binding substrates present in specific environments. Our results work well with this theory since we also identified the last two functional polymorphisms close to the LC. This suggests that the vWF domain recognizes environmental factors such as pathogens. Specifically, site 1451 (Fig. 4C,D) is in a small hydrophobic pocket close to the Ca^{2+}-binding site. Our MSA shows conservation of hydrophobicity in this position, which is often seen for binding sites. Based on our collected data, this speculation cannot be confirmed, but could form the basis of new experimental work in which this is explored.

Our results suggest that honey bee Vg is predominantly monomeric *in vitro*. First, only one copy of the Vg model could fit into the low-resolution EM map. Second, SEC analysis showed only one peak, and this corresponded to monomeric Vg. Third, native gel results also showed a higher tendency toward a monomeric state determined by the much weaker 345 kDa band (presumably a dimer). On the contrary, we demonstrated that the seven residues of each monomer that are creating polar contacts during dimerization in lamprey Vg are conserved in honey bee Vg, making it plausible that Vg dimers can form in honey bees in certain cases. We note that no reducing agent was present in the loading buffer or gel, making it possible that dimers are stabilized by disulfide bonds. Additionally, we cannot rule out that high salt concentration in the SEC prevented the formation of the Vg dimer. Taken together, it is difficult to determine whether dimerization occurs *in vivo* or is an artifact of the *in vitro* conditions, as dimerization occurs frequently in a high concentration sample containing just one

type of protein [65]. We speculate that dimerization can be dose-dependent and thus become more prevalent at elevated Vg concentration. The concentration of Vg in honey bee hemolymph has been reported as high as 100 $\mu g \cdot \mu L^{-1}$, illustrating that the protein is highly soluble [66]. More efforts are needed to conclude the oligomeric state of Vg in honey bees and to evaluate earlier evidence describing honey bee Vg to be monomeric [57,67].

To summarize, our study presents new evidence of the full-length protein and domain assembly for honey bee Vg. We are thus able to identify properties and describe the structural landscape of the large and versatile protein. Our results verify a second cavity of honey bee Vg in addition to the well described lipid-binding cavity and describe the structural units potentially forming this cavity. As a result, we are able to suggest the possibility that the vWF domain contributes to the immune system of honey bees, which is currently of global concern due to declining pollinator numbers. Efforts are being made to generate a higher resolution and up-to-date EM map, which could be used to preform molecular dynamic flexible fitting and enable studies of Vg protein–protein interactions and ligand binding. Our findings encourage future initiatives in investigating this domain together with the full-length protein to unravel some of the questions asked here.

Acknowledgements

We thank Eivind Fjeldstad for his valuable guidance for running AlphaFold. The authors acknowledge The Research Council of Norway grant number 262137 for funding toward running costs and positions. MM-C is supported by an H2020 MSCA International Training Network, ESC by an H2020 MSCA Individual Fellowship, HL by NCMM core funding. The FP7 WeNMR (project# 261572), H2020 West-Life (project# 675858), and the EOSC-hub (project# 777536) European e-Infrastructure projects are acknowledged for the use of their web portals, which make use of the EGI infrastructure with the dedicated support of CESNET-MetaCloud, INFN-PADOVA, NCG-INGRID-PT, TW-NCHC, SURFsara, and NIKHEF, and the additional support of the national GRID Initiatives of Belgium, France, Italy, Germany, the Netherlands, Poland, Portugal, Spain, UK, Taiwan, and the US Open Science Grid. Molecular graphics and analyses performed with UCSF CHIMERA, developed by the Resource for Biocomputing, Visualization, and Informatics at the University of California, San Francisco, with support from NIH P41-GM103311. The authors acknowledge BioCat (RCN grant number 249023) for travel grants and conferences support.

Conflict of interest

The authors declare no conflict of interest.

Author contributions

VL executed homology modeling, structure predictions, rigid-body fitting, and purification of honey bee Vg and ØH and GVA supervised the research. EH-G done the negative staining and generation of the EM map. MM-C and ESC performed native gel and SEC and HL supervised the research. VL wrote the manuscript with assistance from ØH and GVA. All authors contributed to the manuscript.

Data accessibility

The data that support the findings of this study are available in the supplementary material of this article (Tables S1–S7, Figs S1–S12, and the EM map validation report [Appendix S1]). The structural data from homology modeling of the vWF domain are openly available at ModelArchive https://modelarchive.org/doi/10.5452/ma-sfueo (access code: okHs98Pcl2), and the structural data from AlphaFold are available in the supplementary material of this article.

References

1 Hayward A, Takahashi T, Bendena WG, Tobe SS, Hui JH. Comparative genomic and phylogenetic analysis of vitellogenin and other large lipid transfer proteins in metazoans. *FEBS Lett*. 2010;**584**(6):1273–8.
2 Corona M, Velarde RA, Remolina S, Moran-Lauter A, Wang Y, Hughes KA, et al. Vitellogenin, juvenile hormone, insulin signaling, and queen honey bee longevity. *Proc Natl Acad Sci USA*. 2007;**104**(17):7128–33.
3 Havukainen H, Munch D, Baumann A, Zhong S, Halskau O, Krogsgaard M, et al. Vitellogenin recognizes cell damage through membrane binding and shields living cells from reactive oxygen species. *J Biol Chem*. 2013;**288**(39):28369–81.
4 Zhang S, Dong Y, Cui P. Vitellogenin is an immunocompetent molecule for mother and offspring in fish. *Fish Shellfish Immunol*. 2015;**46**(2):710–5.
5 Sun C, Hu L, Liu S, Gao Z, Zhang S. Functional analysis of domain of unknown function (DUF) 1943, DUF1944 and von Willebrand factor type D domain (VWD) in vitellogenin2 in zebrafish. *Dev Comp Immunol*. 2013;**41**(4):469–76.

6 Du X, Wang X, Wang S, Zhou Y, Zhang Y, Zhang S. Functional characterization of vitellogenin_n domain, domain of unknown function 1943, and von Willebrand factor type D domain in vitellogenin of the non-bilaterian coral *Euphyllia ancora*: implications for emergence of immune activity of vitellogenin in basal metazoan. *Dev Comp Immunol*. 2017;**67**:485–94.
7 Salmela H, Amdam GV, Freitak D. Transfer of immunity from mother to offspring is mediated via egg-yolk protein vitellogenin. *PLoS Pathog*. 2015;**11**(7): e1005015.
8 Seehuus SC, Norberg K, Gimsa U, Krekling T, Amdam GV. Reproductive protein protects functionally sterile honey bee workers from oxidative stress. *Proc Natl Acad Sci USA*. 2006;**103**(4):962–7.
9 Nakamura A, Yasuda K, Adachi H, Sakurai Y, Ishii N, Goto S. Vitellogenin-6 is a major carbonylated protein in aged nematode, *Caenorhabditis elegans*. *Biochem Biophys Res Comm*. 1999;**264**(2):580–3.
10 Ando S, Yanagida K. Susceptibility to oxidation of copper-induced plasma lipoproteins from Japanese eel: protective effect of vitellogenin on the oxidation of very low density lipoprotein. *Comp Biochem Physiol C Pharmacol Toxicol Endocrinol*. 1999;**123**(1):1–7.
11 Amdam GV, Norberg K, Hagen A, Omholt SW. Social exploitation of vitellogenin. *Proc Natl Acad Sci USA*. 2003;**100**(4):1799–802.
12 Havukainen H, Halskau O, Amdam GV. Social pleiotropy and the molecular evolution of honey bee vitellogenin. *Mol Ecol*. 2011;**20**(24):5111–3.
13 Hernandez Lopez J, Schuehly W, Crailsheim K, Riessberger-Galle U. Trans-generational immune priming in honeybees. *Proc Biol Sci*. 2014;**281** (1785):20140454.
14 Sadd BM, Kleinlogel Y, Schmid-Hempel R, Schmid-Hempel P. Trans-generational immune priming in a social insect. *Biol Lett*. 2005;**1**(4):386–8.
15 Nelson CM, Ihle KE, Fondrk MK, Page RE, Amdam GV. The gene vitellogenin has multiple coordinating effects on social organization. *PLoS Biol*. 2007;**5**(3):e62.
16 Kohlmeier P, Feldmeyer B, Foitzik S. Vitellogenin-like a-associated shifts in social cue responsiveness regulate behavioral task specialization in an ant. *PLoS Biol*. 2018;**16**(6):e2005747.
17 Suren-Castillo S, Abrisqueta M, Maestro JL. Foxo inhibits juvenile hormone biosynthesis and vitellogenin production in the German cockroach. *Insect Biochem Mol Biol*. 2012;**42**(7):491–8.
18 Dittmer J, Alafndi A, Gabrieli P. Fat body-specific vitellogenin expression regulates host-seeking behaviour in the mosquito *Aedes albopictus*. *PLoS Biol*. 2019;**17** (5):e3000238.
19 Havukainen H, Underhaug J, Wolschin F, Amdam G, Halskau O. A vitellogenin polyserine cleavage site: highly disordered conformation protected from proteolysis by phosphorylation. *J Exp Biol*. 2012;**215**(Pt 11):1837–46.
20 Kent CF, Issa A, Bunting AC, Zayed A. Adaptive evolution of a key gene affecting queen and worker traits in the honey bee, *Apis mellifera*. *Mol Ecol*. 2011;**20**(24):5226–35.
21 Pinto MA, Henriques D, Chávez-Galarza J, Kryger P, Garnery L, van der Zee R, et al. Genetic integrity of the dark European honey bee (*Apis mellifera mellifera*) from protected populations: a genome-wide assessment using SNPs and mtDNA sequence data. *J Apic Res*. 2014;**53**(2):269–78.
22 Munoz I, Henriques D, Jara L, Johnston JS, Chavez-Galarza J, De La Rua P, et al. SNPs selected by information content outperform randomly selected microsatellite loci for delineating genetic identification and introgression in the endangered dark European honeybee (*Apis mellifera mellifera*). *Mol Ecol Resour*. 2017;**17**(4):783–95.
23 Henriques D, Browne KA, Barnett MW, Parejo M, Kryger P, Freeman TC, et al. High sample throughput genotyping for estimating C-lineage introgression in the dark honeybee: an accurate and cost-effective SNP-based tool. *Sci Rep*. 2018;**8**(1):8552.
24 Thompson JR, Banaszak LJ. Lipid-protein interactions in lipovitellin. *Biochemistry*. 2002;**41**(30):9398–409.
25 Havukainen H, Halskau O, Skjaerven L, Smedal B, Amdam GV. Deconstructing honeybee vitellogenin: novel 40 kDa fragment assigned to its n terminus. *J Exp Biol*. 2011;**214**(Pt 4):582–92.
26 Dong X, Leksa NC, Chhabra ES, Arndt JW, Lu Q, Knockenhauer KE, et al. The von Willebrand factor D'D3 assembly and structural principles for factor VIII binding and concatemer biogenesis. *Blood*. 2019;**133** (14):1523–33.
27 Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. *Nature*. 2021;**596**(7873):583–9.
28 Zimmermann L, Stephens A, Nam S-Z, Rau D, Kübler J, Lozajic M, et al. A completely reimplemented MPI bioinformatics toolkit with a new HHpred server at its core. *J Mol Biol*. 2018;**430**(15):2237–43.
29 Rice P, Longden I, Bleasby A. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet*. 2000;**16**(6):276–7.
30 Needleman SB, Wunsch CD. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol*. 1970;**48** (3):443–53.
31 Guex N, Peitsch MC. SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. *Electrophoresis*. 1997;**18**(15):2714–23.
32 Guex N, Peitsch MC, Schwede T. Automated comparative protein structure modeling with SWISS-

MODEL and Swiss-PdbViewer: a historical perspective. *Electrophoresis*. 2009;**30**(S1):S162–73.
33 van Gunsteren WF. *Biomolecular simulations: the GROMOS96 manual and user guide*. Zürich: VDF Hochschulverlag AG an der ETH Zürich; 1996. p. 1–1042.
34 Laskowski RA, MacArthur MW, Moss DS, Thornton JM. Procheck: a program to check the stereochemical quality of protein structures. *J Appl Crystallogr*. 1993;**26**(2):283–91.
35 Benkert P, Biasini M, Schwede T. Toward the estimation of the absolute quality of individual protein structure models. *Bioinformatics*. 2011;**27**(3):343–50.
36 Willard L, Ranjan A, Zhang H, Monzavi H, Boyko RF, Sykes BD, et al. VADAR: a web server for quantitative evaluation of protein structure quality. *Nucleic Acids Res*. 2003;**31**(13):3316–9.
37 Schrodinger L. The pymol molecular graphics system, version 1.8; 2015.
38 Sali A, Potterton L, Yuan F, van Vlijmen H, Karplus M. Evaluation of comparative protein modeling by MODELLER. *Proteins*. 1995;**23**(3):318–26.
39 Xu J, Mcpartlon M, Li J. Improved protein structure prediction by deep learning irrespective of co-evolution information. *bioRxiv*. 2020. 2020.2010.2012.336859.
40 van Zundert GC, Trellet M, Schaarschmidt J, Kurkcuoglu Z, David M, Verlato M, et al. The DisVis and PowerFit web servers: explorative and integrative modeling of biomolecular complexes. *J Mol Biol*. 2017;**429**(3):399–407.
41 Zundert GCP, Bonvin AMJJ. Fast and sensitive rigid-body fitting into cryo-em density maps with powerfit. *AIMS Biophys*. 2015;**2**(2):73–87.
42 Garzón JI, Kovacs J, Abagyan R, Chacón P. ADP_EM: fast exhaustive multi-resolution docking for high-throughput coverage. *Bioinformatics*. 2007;**23**(4):427–33.
43 Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, et al. UCSF Chimera – a visualization system for exploratory research and analysis. *J Comput Chem*. 2004;**25**(13):1605–12.
44 Eisenberg D, Schwarz E, Komaromy M, Wall R. Analysis of membrane and surface protein sequences with the hydrophobic moment plot. *J Mol Biol*. 1984;**179**(1):125–42.
45 Aase ALTO, Amdam GV, Hagen A, Omholt SW. A new method for rearing genetically manipulated honey bee workers. *Apidologie*. 2005;**36**(3):293–9.
46 Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol*. 1990;**215**(3):403–10.
47 Venclovas C. Methods for sequence-structure alignment. *Methods Mol Biol*. 2012;**857**:55–82.
48 Pidcock E, Moore GR. Structural characteristics of protein binding sites for calcium and lanthanide ions. *J Biol Inorg Chem*. 2001;**6**(5–6):479–89.
49 Ramachandran GN, Ramakrishnan C, Sasisekharan V. Stereochemistry of polypeptide chain configurations. *J Mol Biol*. 1963;**7**:95–9.
50 Bordoli L, Schwede T. Automated protein structure modeling with SWISS-MODEL workspace and the protein model portal. *Methods Mol Biol*. 2012;**857**:107–36.
51 Dalton JAR, Jackson RM. An evaluation of automated homology modelling methods at low target–template sequence similarity. *Bioinformatics*. 2007;**23**(15):1901–8.
52 Mariani V, Biasini M, Barbato A, Schwede T. LDDT: a local superposition-free score for comparing protein structures and models using distance difference tests. *Bioinformatics*. 2013;**29**(21):2722–8.
53 Anderson TA, Levitt DG, Banaszak LJ. The structural basis of lipid interactions in lipovitellin, a soluble lipoprotein. *Structure*. 1998;**6**(7):895–909.
54 Lu C-H, Lin Y-F, Lin J-J, Yu C-S. Prediction of metal ion-binding sites in proteins using the fragment transformation method. *PLoS One*. 2012;**7**(6):e39252.
55 Friedberg I, Margalit H. Persistently conserved positions in structurally similar, sequence dissimilar proteins: roles in preserving protein fold and function. *Protein Sci*. 2002;**11**(2):350–60.
56 Springer TA. von Willebrand factor, Jedi knight of the bloodstream. *Blood*. 2014;**124**(9):1412–25.
57 Sappington TW, Raikhel AS. Molecular characteristics of insect vitellogenins and vitellogenin receptors. *Insect Biochem Mol Biol*. 1998;**28**(5):277–300.
58 Atkins AR, Brereton IM, Kroon PA, Lee HT, Smith R. Calcium is essential for the structural integrity of the cysteine-rich, ligand-binding repeat of the low-density lipoprotein receptor. *Biochemistry*. 1998;**37**(6):1662–70.
59 Faiz ZM, Mardhiyyah MP, Mohamad A, Hidir A, Nurul-Hidayah A, Wong L, et al. Identification and relative abundances of mRNA for a gene encoding the vWD domain and three Kazal-type domains in the ovary of giant freshwater prawns, *Macrobrachium rosenbergii*. *Anim Reprod Sci*. 2019;**209**:106143.
60 Finn RN. Vertebrate yolk complexes and the functional implications of phosvitins and other subdomains in vitellogenins. *Biol Reprod*. 2007;**76**(6):926–35.
61 Javitt G, Khmelnitsky L, Albert L, Bigman LS, Elad N, Morgenstern D, et al. Assembly mechanism of mucin and von Willebrand factor polymers. *Cell*. 2020;**183**(3):717–29.e716.
62 Loof TG, Schmidt O, Herwald H, Theopold U. Coagulation systems of invertebrates and vertebrates and their roles in innate immunity: the same side of two coins? *J Innate Immun*. 2011;**3**(1):34–40.

63 Eleftherianos I, Revenis C. Role and importance of phenoloxidase in insect hemostasis. *J Innate Immun*. 2011;**3**(1):28–33.
64 Sanggaard KW, Dyrlund TF, Bechsgaard JS, Scavenius C, Wang T, Bilde T, et al. The spider hemolymph clot proteome reveals high concentrations of hemocyanin and von Willebrand factor-like proteins. *Biochim Biophys Acta*. 2016;**1864**(2):233–41.
65 Wang W, Xu W-X, Levy Y, Trizac E, Wolynes PG. Confinement effects on the kinetics and thermodynamics of protein dimerization. *Proc Natl Acad Sci USA*. 2009;**106**(14):5517–22.
66 Amdam GV, Hartfelder K, Norberg K, Hagen A, Omholt SW. Altered physiology in worker honey bees (Hymenoptera: Apidae) infested with the mite Varroa destructor (Acari: Varroidae): a factor in colony loss during overwintering? *J Econ Entomol*. 2004;**97**(3):741–7.
67 Tufail M, Takeda M. Molecular characteristics of insect vitellogenins. *J Insect Physiol*. 2008;**54**(12):1447–58.

Supporting information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Fig. S1. Domain architecture of honey bee and lamprey vitellogenin. The N-term (green), DUF1943 (pink) and vWF (blue) domains are conserved in both species, as well as the two structural subdomains, β-barrel (red arrow) and α-helical domain (dark green curved line). A) Honey bee Vg contains a proteolytic cleavage site, polyserine region (yellow S) linking the two subdomains. The five residue-positions (640, 1220, 1284, 1451 and 1536) identified to be candidates of functional polymorphisms are marked (brown stars). B) Lamprey Vg contains an addition domain, DUF1943 (purple). The yolk protein organization of IuVg is shown as gray boxes; lipovitellin heavy chain (LvH), Phosvitin (Pv), lipovitellin light chain (LvL), β-Component (β-C) and C-terminal coding region (CT). The dotted lines indicate that these regions (Pv, β-C and CT) are missing from the crystallographic structure (PDB ID: 1LSH).

Fig. S2. Multiple sequence and structural alignment. The coloring for the conserved residues/regions, gaps and secondary structure annotations are explained in the green box. The conserved Ca2+-binding region are colored in two shades of pink, dark pink is more conserved compared to the lighter pink. A) Extraction of the MSA. The original residue numbering for honey bee Vg is included on top. B) The final structural alignment with the original residue numbering included above each sequence. The annotations are retrieved from the template (PDB ID: 6N29). Both figures are created in Geneious Prime (v. 2019.0.3) and Adobe illustrator (v. 24.0.02).

Fig. S3. ProCheck summary, local quality estimate and Ramachandran plots. A) The ProCheck quality evaluations summarized and categorized by calculation results. The ideal residue values and standard deviation for any given model are derived from Morris et al. 1992.[1] The max deviation, in residues properties, is calculated from the mean value of the residue-by-residue listing values (Fig. S4E) of the full-length structure. The number of bad contacts is defined as the non-bonded atoms at a distance of <= 2.6 Å. The bond length and angles are calculated in similar manner as the max deviation, but the ideal values are based on Engh and Huber 1991.[2] The Morris et al. (1992) class summarizes the three above stereochemical parameters by assigning a number between 1 (best) to 4 (worst), indicating the overall quality of the model. B) Local QMEAN results are presented. The first plot is analysis of the template (green), while the second is analysis of the target modeled interactively (cyan) and automatically (red). The Ca^{2+}-binding region (magenta Ca), the Cys residues forming the intra-chain disulfide bridges (orange, C) are in the higher quality region, while Ω5-7 (black) are in the lower quality region. The local score is calculated for each residue in the model and the average local score for the template is 0.93 ± 0.07, while the target average score is 0.40 ± 0.07 (cyan) and 0.44 ± 0.06 (red). C) The Ramachandran plot produced by ProCheck. The plot on the left is the template (PDB ID: 6N29), while the target (honey bee vWF domain) is on the right. Below each plot, the statistic is presented.

Fig. S4. Global quality estimate, VADAR plots and ProCheck residue listing. A-C) The plots of the global QMEAN have the QMEAN4 scores for a set PDB structures plotted (gray dots) with the QMEAN4 score along the x-axis and the number of residues in the structures as long the y-axis. The global scores value QMEAN4 range from 0 to 1, where 1 is good. A) Analysis of the template (red star) and the QMEAN4 value is written on the plot. B) Analysis of the interactively homology modeled (red star) structure and C) The automatically homology modeled (red star) structure from MODELLER. D) Four different analyses were performed by VADAR, presented in one plot each, with the template (gray) compared to the target (green). Plot 1: a low fractional ASA score indicates a buried residue, while a score above 0.5 (dotted black line) indicates an exposed residue. A score above 1.0 (red line) indicates a problem in the structure. Plot 2: When a protein structure is efficiently packed the score

should be around 1.0 ± 0.1. A score above 1.2 (blue line) or below 0.8 (red line) could indicate a poor refinement or identify cavities. Plot 3: Each residue is assigned a score between 0-3 (high is good quality) for three different measurements (torsion angle, omega angle and fractional volume). The total quality score for each residue can be from 0-9 and the threshold for a good quality is set to 6 (red line). Plot 4: Calculates the 3D quality of each residue based on its environment and gives a score between 0-9 (high is good quality), and the threshold for a good quality is set to 4 (red line). E) The Residue-by-Residue listing for ProCheck lists all residues in a structure and present all calculations for each. A short example is shown here for the first six residues in the target structure. Each value is compared to the ideal values which is noted on top. The deviating values are marked with * (one standard deviation) and + (half a standard deviation) sign. For example, the omega dihedral angle of residue S1443 is 16.9 standard deviation away from the ideal value, which is a result from the loop building of Ω1.

Fig. S5. Comparison of vWF homology models. A) The sort region around the Ca^{2+}-binding site (Ca^{2+}-ion, green) is shown from the interactively modeled (cyan) structure and the automatically modeled (gray) structure. The Cys-residues (C1444, C1466, C1598 and C1634) and Ca^{2+}-binding residues are shown as yellow/cyan (interactively) and orange/magenta (automatically) sticks. The missing C1634 and β-strands in the automatically modeled structure are shown (gray arrows). B) All the Ca^{2+}-binding residues are in the same orientation in both models (light blue: interactively and light pink: automatically), except N1607. The interactions to the Ca^{2+}-ion is shown as yellow dotted lines and measured (Å) for N1607.

Fig. S6. Comparison of homology models from MODELLER and RaptorX. A) The N-terminal domain: Model 1 (green) aligned with Model 4 (red), 5 (yellow) and 6 (forest green). B) The DUF1943 domain: Model 1 (magenta) aligned with Model 8 (cyan), Model 7 (orange) and Model 9 (blue). The identified curve in the longer β-sheet in Model 1, 8 and 9 and the missing curve in Model 7 is marked with arrows. C) The DUF1943 domain Model 1 (magenta), the downstream region residue 1060 to 1140 of Model 9 (hot pink) and the loop region (gray). D) The undetermined domain: Model 2 (purple) aligned with Model 9 (blue), with the long loop region (gray). E) The interactively homology model of vWF domain (cyan) with the C-terminal region from Model 9 (orange).

Fig. S7. RaptorX structural prediction of full-length honey bee vitellogenin. A) The β-barrel subdomain (red), the polyserine linker (yellow), the α-helical subdomain (forest green), the DUF1943 domain (magenta), elongation of the DUF1943 domain (hot pink arrow), the undetermined structural region (purple), the vWF domain (cyan) and the C-terminal region (orange) are generated as one full-length model. The two loop regions (gray arrows) are also predicted. B) Domain 1 to 6 from Table S7 are colored red, cyan, purple, blue, green and orange, respectively, and if templates was used, the PDB ID is written in parenthesis.

Fig. S8. AlphaFold output. A) The number of sequence hits in the MSA produced by AlphaFold, is plotted per residue. The average number of hits per residue (gray dotted line), and the threshold at 100 sequence per residue (red dotted line) is marked. B) The plDDT score for the five outputted models by AlphaFold is plotted per residue, and the average plDDT score per model is listed to the right, which produces the rank from 0 (best) to 4 (worst). C) The ranked models are aligned, colored by the same coloring scheme in panel B, and the consistently folded domains (β-barrel (β), α-helical (α), DUF1943 (DUF), β-sheet 1 (β1), β-sheet 2 (β2) and vWF domain (vWF)) are labeled in bold letters, while the more variable domains (N-terminal, polyserine linker (PS) and C-terminal) are labeled in grey letters. D) The final homology model domains (β-barrel (red), polyserine linker (yellow), α-helical (green), DUF1943 (magenta), β-sheet 1 (hotpink), β-sheet 2 (purple), vWF (cyan, Ca^{2+}-ion shown as green sphere) and C-terminal domain (orange) is aligned to their respective domains in the top ranked AlphaFold prediction (grey). The grey brackets to the lower right indicate the region where AlphaFold have predicted a fold for the main missing atoms in the homology model.

Fig. S9. EM map validation. A) Map visualization to allow visual inspection of the internal detail of the map and identification of artifacts. The primary map, central slices of the map and largest variance of the map is shown in three orthogonal directions. The 3D surface view of the primary map at recommended contour level 0.07. B) Statistical analysis of the map. In the first graph the map-value distributions is plotted in 128 intervals along the x-axis, and the y-axis is logarithmic. The spike around 0 indicate that the volume has been masked. The second graph shows how the enclosed volume varies with the contour level. The volume at the recommended contour (red line) is 289 nm3; this corresponds to an approximate mass of 261 kDa. C) The provided Fourier-Shell Correlation (blue) is plotted together with the reported resolution, (black line, *Reported resolution corresponds to spatial frequency of 0.037 $Å^{-1}$). A curve is displayed for the half-

bit criterion (dashed red), in addition to lines showing the 0.143 gold standard cut-off (dashed orange line) and 0.5 cut-off (green dotted line). All the graphs are assembled from the EmDataBank map validation report (copy included).

Fig. S10. Rigid-body fitting for honey bee vitellogenin homology models. A) The EM map is shown as a gray surface. The distinct cavity creases are marked with stars and arrows, upper cavity (blue) and lower cavity (yellow). The four curves in the surface are labeled (top, base, left and right). B) The correlation score and precent of atoms outside the contour calculated by Chimera was plotted for each fit from PowerFit (PF, blue) and ADP_EM (ADP, pink), and ranked according to the correlation score (dark color: highest score, light color: second highest score). C-E) The fits from the full-length homology model, RaptorX and chain A is presented inside the EM map, with the same coloring scheme as in panel B. The β-barrel (β), α-helical (α), DUF1943 (DUF), vWF and C-terminal (C-t) domains are labeled. If the domain is outside of the contour it is noted by a "*"-mark. F) The fits of chain B to E separately with the same coloring scheme as in panel B, but they are labeled according to chains and not domains.

Fig. S11. Rigid-body fitting of chain A to C and D. A) The correlation score and precent of atoms outside the contour calculated by Chimera was plotted for each fit from PowerFit (PF, blue) and ADP_EM (ADP, pink), and ranked according to the correlation score (dark color: highest score, light color: second highest score). B) The EM map are shown as a transparent surface, and the fits of chain A to C from PF and ADP are shown as cartoons and colored by method and scores (dark blue: PF1, light blue: PF2, dark pink: ADP1, light pink: ADP2). The β-barrel (β), α-helical (α) and DUF1943 (DUF) domains are labeled. C) The EM map and the fits of chain D is shown in same coloring scheme as in panel B. The label is marked with "*" if the fit is outside the contour of the EM map. D) The EM map are shown as a surface, less transparent than in panel B, with the fits of chain A to C (1: PF2 and ADP1, 2: PF1 and ADP2) in the same coloring scheme as in panel B. The EM map is shown at four different angels, and arrows points to secondary structure elements from β, α or DUF domain which are outside the contour of the EM map.

Fig. S12. Rigid-body fitting for previously published homology models and a distant homologue. A) The same plot as in Fig. S10 for the β-barrel and α-helical subdomains, and the crystal structure of lamprey Vg (1LSH). B-D) Same presentation and coloring scheme as in Fig. S10C-S10F.

Table S1. Alignment parameters.

Table S2. List of species used in the multiple sequence alignment.

Table S3. Loop building based on gaps in the structural alignment.

Table S4. Edited residues during quality control.

Table S5. Rigid-body fitting scores from PowerFit and Chimera.

Table S6. Rigid-body fitting scores from ADP_EM and Chimera.

Table S7. RaptorX structure prediction.

Appendix S1. wwPDB EM Validation Summary Report.

Appendix S2. Top ranked Vitellogenin model by AlphaFold.

Paper II

Where Honey Bee Vitellogenin may Bind Zn^{2+}-Ions

Vilde Leipart[1], Øyvind Enger[1], Diana Cornelia Turcu[2], Olena Dobrovolska[3], Finn Drabløs[4], Øyvind Halskau[2], Gro V. Amdam[1,5]

[1]Faculty of Environmental Sciences and Natural Resource Management, Norwegian University of Life Sciences, Aas, Norway

[2]Department of Biological Sciences, University of Bergen, Bergen, Norway

[3]Helse Bergen, Bergen, Norway

[4]Department of Clinical and Molecular Medicine, Faculty of Medicine and Health Sciences, NTNU – Norwegian University of Science and Technology, Trondheim, Norway

[5]School of Life Sciences, Arizona State University, Tempe, AZ, United States

Running title: Honey Bee Vitellogenin and Zn^{2+}-Binding

Abstract

The protein Vitellogenin (Vg) plays a central role in lipid transportation in most egg-laying animals. High Vg levels correlate with stress resistance and lifespan potential in honey bees (*Apis mellifera*). Vg is the primary circulating zinc-carrying protein in honey bees. Zinc is an essential metal ion in numerous biological processes, including the function and structure of many proteins. Measurements of Zn^{2+} suggest a variable number of ions per Vg molecule in different animal species, but the molecular implications of zinc-binding by this protein are not well understood. We used inductively coupled plasma mass spectrometry (ICP-MS) to determine that, on average, each honey bee Vg molecule binds 3 Zn^{2+}-ions. Our full-length protein structure and sequence analysis revealed seven potential zinc-binding sites. These are

located in the β-barrel and α-helical subdomains of the N-terminal domain, the lipid binding site, and the cysteine-rich C-terminal region of unknown function. Interestingly, two potential zinc-binding sites in the β-barrel can support a proposed role for this structure in DNA-binding. Overall, our findings illustrate the capacity of honey bee Vg to bind zinc at several functional regions, indicating that Zn^{2+}-ions are important for many of the activities of this protein. In addition to being potentially relevant for other egg-laying species, these insights provide a platform for studies of metal ions in bee health, which is of global interest due to recent declines in pollinator numbers.

Keywords: Honey bees, vitellogenin, zinc-binding, insect immunity, protein structure analysis

Abbreviations

Domain of unknown function (DUF1943), Inductively coupled plasma mass spectrometry (ICP-MS), multiple sequence alignment (MSA), Vitellogenin (Vg), von Willebrand factor (vWF)

Introduction

Zinc is necessary for living organisms to function properly (Sloup et al., 2017). The element is involved in basic life processes such as cell division and gene expression and it is an essential nutrient for growth and development (Falchuk, 1998, Baltaci and Yuce, 2018). Zinc is necessary for catalytic, structural, and regulatory functions for thousands of proteins (Andreini et al., 2006). Improving the understanding of zinc-carrying proteins is thus likely to reveal new information about many physiological processes across taxa.

An important zinc-carrying protein in egg-laying animals is the multi-domain glycolipophosphoprotein Vitellogenin (Vg) (Montorzi et al., 1994, Falchuk, 1998, Matozzo et al., 2008, Gupta et al., 2021). Vg provides nutrients to developing embryos by delivering lipids,

amino acids, and zinc (Pan et al., 1969). In some species, Vg is expressed in juveniles, as well as in males and in females that do not reproduce, hinting at roles beyond yolk formation (Sappington and S. Raikhel, 1998). Such roles have been most abundantly studied in honey bees (*Apis mellifera*), where Vg is recognized as a multi-functional protein impacting the behavior and health of workers (functionally sterile females). RNA-interference mediated gene knockdown reveals that honey bee Vg affects worker behavioral ontogeny, foraging choice, capacity to provide larval care, stress resistance, and longevity (Amdam et al., 2004, Guidugli et al., 2005). At least some of the physiological impact of honey bee Vg can be zinc-related, as low zinc levels may reduce the cell-based immune capacity of the workers (Amdam et al., 2005). However, apart from the finding that circulating zinc levels correlate strongly with the hemolymph level of Vg in honey bees, the molecular relationship between this ion and protein is largely unknown.

The capacity for zinc-binding is established for Vg proteins, but the number of ions per Vg molecule varies among species. For example, measurements of bound Zn^{2+}-ions in the hemolymph of shore crab (*Carcinus maenas*) (Martin and Rainbow, 1998) are higher than those of the American clawed frog (*Xenopus laevis*) (Montorzi et al., 1994) and domestic fowl (*Gallus gallus*) (Mitchell and Carlisle, 1991). The same studies also indicate that the number of Zn^{2+}-ions carried by Vg can vary with individual reproductive state and age. Such variation is likely biologically important, but to date is not well understood for Vg proteins.

A prerequisite for understanding zinc-related molecular mechanisms is finding the location and structural context of Zn^{2+}-binding sites within the protein of interest (Ataie et al., 2008, Daniel and Farrell, 2014). The coordinating environment for Zn^{2+}-ions in proteins is well characterized (Dudev and Lim, 2003, Pace and Weerapana, 2014), and binding sites are

usually sorted into two structurally distinct categories based on whether the Zn^{2+} has a catalytic or structural role. A catalytic binding site is often partially exposed to the solvent, and the Zn^{2+}-ion coordinates most often with histidine (H), cysteine (C), aspartate (D), glutamate (E), serine (S) residues, and/or water molecules (Jernigan et al., 1994, Ataie et al., 2008). A structural binding site is usually buried in the protein, surrounded by an intricate network of hydrogen bonds (Dudev and Lim, 2003), and the Zn^{2+}-coordinating residues are typically multiple H/C residues only. For example, the well-known transcription factor motif zinc fingers (C4 or C2H2) (Pace and Weerapana, 2014) represents a structural binding site.

The prevalence of coordinating residues differs between catalytic and structural binding sites. For catalytic sites, 4, 5, and 6 residues coordinate in 48, 44, and 6 % of cases, respectively. Correspondingly, the ratio is 79, 6, and 12 % for structural sites, respectively (Dudev and Lim, 2003, Ataie et al., 2008). These numbers imply that catalytic sites largely coordinate Zn^{2+} with 4 or 5 residues, while structural sites most commonly coordinate with 4 residues. In addition to these two categories, Zn^{2+}-binding is identified in regulatory Zn^{2+}-Cys complexes, called redox switches (Pace and Weerapana, 2014). Two common characteristics for all coordinating sites are: strong interaction between the residues and the Zn^{2+}-ion, and high hydrophobic contrast in the binding site (Dudev and Lim, 2003, Pace and Weerapana, 2014).

The zinc coordinating environment for honey bee Vg is not described, but speculations have been presented regarding lamprey (*Ichthyomyzon unicuspis*) and zebrafish (*Danio rerio*) Vg. Anderson *et al.* (1998) published the only experimentally solved protein structure of lamprey Vg (PDB-ID: 1LSH), which lacks Zn^{2+}-ions due to use of 1 mM EDTA during crystallization. In the absence of zinc, Anderson *et al.* (1998) proposed two potential binding sites (H312/H322 and H868/H887) based on the residues' locations in the crystal structure and the sequence

conservation. In comparison, Sullivan *et al.* (2018) suggest the phosphorylated serine-rich phosvitin domain is associated with Zn^{2+}-ions in zebrafish (Sullivan and Yilmaz, 2018). This domain is missing in some Vg proteins, including those of insects (Tufail and Takeda, 2008). For example, Honey bee Vg consists of an N-terminal domain, a lipid cavity, and a C-terminal region (Havukainen et al., 2011, Havukainen et al., 2012, Leipart et al., 2021). The N-terminal domain comprises the β-barrel subdomain followed by a flexible polyserine linker and the α-helical subdomain. The lipid cavity is built up by a domain of unknown function (DUF1943), a β-sheet, and a von Willebrand factor (vWF) domain. Honey bee Vg can be cleaved at the polyserine linker in the N-terminal domain. This cleavage creates a small fragment (40 kDa, the β-barrel subdomain) and a larger fragment (150 kDa, the α-helical subdomain, the lipid binding site, and the C-terminal region) (Havukainen et al., 2012). It is interesting to note that the smaller fragment of honey bee Vg may translocate into cell nuclei, bind DNA (potentially with co-factors), and influence gene expression (Salmela et al., 2021).

Honey bees provide a practical and useful research system as they are globally available as commercial pollinators and primary producers of honey, pollen, and wax. We recently published the first full-length Vg structure prediction for honey bees (Leipart et al., 2021) and use it here to provide insight into knowledge of where Vg binds zinc. First, we performed an element analysis of Vg protein obtained from worker bee hemolymph and found that, on average, it binds 3 Zn^{2+}-ions . Using structural data in combination with sequence data, we then conducted an in-depth analysis to predict the location(s) of potential zinc-binding sites. We identified areas in the β-barrel subdomain, α-helical subdomain, lipid binding site, and C-terminal region. We propose that a zinc-binding site in the β-barrel subdomain plays a role in DNA binding. In an attempt to characterize the zinc-binding site(s) of this subdomain, we expressed the β-barrel using bacterial recombinant expression systems in cultural medium

with various compositions of Zn^{2+} and/or Co^{2+} but this approach did not provide a clear answer. However, taken together, our results provide the first detailed insights into where honey bee Vg can bind Zn^{2+}.

Results

Identification of zinc in honey bee Vg using inductively coupled plasma mass spectrometry

We performed inductively coupled plasma mass spectrometry (ICP-MS) on Vg from worker bee hemolymph to confirm Vg as a zinc carrier and quantify the number of Zn^{2+}-ions per Vg molecule. We detected significant amounts of Zn^{2+} in Vg samples relative to the (zinc negative) controls (Kruskal-Wallis analysis: chi-squared = 7.81, df = 1, p-value = 0.00519, Figure 1A). Based on sample concentrations of Vg and Zn^{2+}, and using the theoretical molecular weight of Vg and Zn^{2+} (201147.7 g/mol and 65.30 g/mol, respectively), we calculated the molecular Zn:Vg ratio for each sample (Figure 1B). This analysis gave a range of 2.57–3.89 mol of Zn^{2+}-ions per Vg molecule, with an average ratio of 3 Zn^{2+}-ions per full-length Vg protein.

Identification of potential zinc coordinating residues in honey bee Vg

In this section, we analyze honey bee Vg to identify zinc-binding sites, referred to as clusters. The results are summarized in Table 1 and illustrated in Figure 2A. To achieve this, we took a comprehensive approach: using online bioinformatic tools developed for identification of zinc motifs in amino acid sequences, assessing suggested structural sites from studies on lamprey (Anderson et al., 1998) and zebrafish (Sullivan and Yilmaz, 2018), and finally, fully analyzing our recently published protein structure (Leipart et al., 2021). We used a multiple sequence alignment (MSA) of Vg sequences with a broad phylogenetic range to evaluate the findings.

Assessing potential zinc coordinating residues using online tools

The searches using online motif search algorithms (see Experimental procedures) resulted in only one hit. MotifScan (Pagni et al., 2007) identified a Zn^{2+}-binding motif at residue 926 to 936. The short region includes several conserved residues (F928, P929, G933, L934, P935, and F936, Figure S1). Regarding the MSA (Figure S1), the zinc-binding residue identified in the motif, H926, is only present in the *Apis mellifera* Vg sequence. The motif is located in a β-strand-turn-β-strand fold in the DUF1943, which extends into the β-barrel subdomain, close to cluster Duf.1 and βb.2, and the DNA binding motif (see Figure S2 for a structural overview of this region).

Assessing potential zinc coordinating residues from suggested zinc-binding sites

According to the MSA, the two sites proposed in lamprey Vg (H312/H322 and H868/H887) align to E427/V444 and N975/G994 in honey bee Vg, respectively (Figure S1B). Of the four residues in honey bee Vg, only the E residue is known to coordinate with Zn^{2+}. The low conservation of E427 and lack of other typically Zn^{2+}-coordinating residues (H/C/D/S) at both sites suggest that these locations do not take part in zinc-binding in the honey bee.

Sullivan and Yilmaz (2018) suggest that the phosvitin domain in zebrafish Vg has a Zn^{2+}-coordinating role. Honey bee Vg does not have this domain, but the polyserine linker in the N-terminal domain is a similar, serine-rich region. We did not identify any C, H, D, or E residues here that could support zinc-coordination, but the linker contains 14 S residues. One or two S residues can participate in zinc-coordination. However, this location is an unlikely candidate as it would be unprecedented to find Zn^{2+}-coordination by serines in a disordered (loop) region (Baglivo et al., 2009).

Assessing potential zinc coordinating residues based on protein structure modeling

Potential Zn^{2+} sites in the β-barrel subdomain of the N-terminal region

The N-terminal β-barrel subdomain contains a total of seven H residues (H20, H47, H113, H193, H210, H229, and H265) and two C residues (C178 and C222) (Figure 2B). Among these, H229 is not folded in the β-barrel subdomain (see cluster αh.1 in Figure 2C). The C residues are conserved in most of the Vg sequences included in the MSA (Figure S1A and Figure S3A). H20 and H113 are conserved in the Vg sequences in the MSA from insect species, while the remaining H residues are less conserved. The first cluster identified in the β-barrel subdomain, βb.1, contains H20 and H113 located in separate loop regions (Figure 2B). H265 is situated at the beginning of a β-strand close to βb.1. In addition, we found two conserved residues (D143 and E147) in a loop region and included them in cluster βb.1 (Table 1).

The second cluster in the β-barrel subdomain, βb.2, contains C178 and C222. The C residues form a disulfide bridge. We identified five conserved residues (E171, D172, S173, E179, and D223) close to the disulfide bridge and included them in cluster βb.2. All residues are in two neighboring β-strands, apart from C222 and D223 at the end of a loop region (Table 1, Figure 2B).

A short SRSSTSR sequence (residues 250 to 256) in the β-barrel subdomain is proposed to bind DNA (Salmela et al., 2021). The residues are located at a β-strand close to cluster βb.1 and βb.2 (Figure S2 and Figure S3A).

Potential Zn^{2+} sites in α-helical subdomain and lipid binding site

We identified four clusters of conserved H/C residues (Table 1), two in the α-helical subdomain (αh.1 and αh.2) and two in the DUF1943 (Duf.1 and Duf.2) (Figure S2C and Figure S1B). The first cluster in the α-helical subdomain, αh.1, contains two highly conserved H residues (H587 and H593) and two less conserved H residues (H577 and H602). All H residues

are in a well-conserved insect-specific loop in the α-helical subdomain (Figure S3B). The cluster also includes the well-conserved H229. The residue is part of a loop extending from the β-barrel subdomain, which positions H229 close to H587 and H593. The second cluster in the α-helical subdomain, αh.2, is located at the beginning of the well-conserved 15th α-helix in the subdomain (Figure S3B) containing H697, C701, and E698 (Figure 2C). The cavity-facing side of the α-helical subdomain is close to a β-sheet region of DUF1943. More specifically, the α-helices are near two loops connecting the β-strands on DUF1943. We identified three well-conserved residues in those loops (S775, S800, and D802) and included them in cluster αh.1.

The first cluster in the DUF1943, Duf.1, contains three conserved H residues (H988, H990, and H1045). They are packed together in two loop regions at the end of two adjacent β-strands (Figure 2C). Two conserved residues (E987 and D1046) were identified in the same loops and included in Duf.1. The second cluster in DUF194, Duf.2, contains two conserved H residues (H1000 and H1035) positioned on the same β-sheet as Duf.1 (Figure S1B and Figure S3C). We also found two conserved residues (D996 and S1037) in the β-sheet region and two residues in the α-helical subdomain (H445 and E449 in the second α-helix of the subdomain) and included those in Duf.2.

Potential Zn^{2+} sites in lipid binding site β-sheet and the C-terminal region

Following the DUF1943, we identified four conserved C residues (C1242, C1279, C1310, and C1324) in a β-sheet in the lipid binding site (Figure S1C). C1242 and C1279 are highly conserved and create a disulfide bridge in the β-sheet region. C1310 and C1324 also create a disulfide bridge connecting two α-helices. Neither of the disulfide bridges has any conserved D, E, or S residues in proximity, so the bridges are not identified as potential zinc clusters here.

After this β-sheet in the lipid binding site, the following domain is the vWF domain, which does not coordinate Zn^{2+} (Leipart et al., 2021).

The final region in honey bee Vg is the C-terminal (residue 1635 to 1770). This region contains seven C and two H residues that are conserved in most Vg sequences in the MSA (Figure S1D and Figure S3D). Six C residues create three separate disulfide bridges (Figure 2C). Two of these bridges cross each other (C1687, C1711, C1715, and C1768) and were identified as cluster Ct (Table 1). The remaining disulfide bridge and conserved C and two H residues were not nearby and therefore not considered a potential zinc-binding site.

Assessing functional roles of specific zinc coordinating residues

Motif comparison in the β-barrel subdomain

The functional roles of zinc-binding sites are not well defined for any Vg. However, a possible exception is made by the short SRSSTSR sequence close to clusters βb.1 and βb.2 in the β-barrel subdomain that may bind DNA in honey bees (Figure 2B and S2). A recent study presents a proposed DNA-binding motif that the subdomain recognizes (see Figure 5 in Salmela *et al.* 2021). Motif A is similar to the transcription factor CTCF motifs in *Drosophila melanogaster* (see Figure S5A and S5B for motif comparison). CTCF contains eleven C2H2 zinc finger factor motifs (Maksimenko et al., 2021) that bind one Zn^{2+}-ion each, shown to stabilize the structural fold in CTCF required to bind DNA (Maksimenko et al., 2021). We aligned the C2H2 motifs to the Vg sequences in our MSA (Figure S5C) and found five predicted zinc-binding residues in the β-barrel subdomain aligned to C and H residues in the C2H2 motifs. D143 and E147 from βb.1 align to H and D residues in the C2H2 motifs, and C178 and C222 from βb.2 align to C residues in the C2H2 motifs. In addition, H229 from αh.1 aligns to H residues in the C2H2 motifs. This configuration of zinc-binding residues in honey bee Vg

supports a functional role of the β-barrel in DNA binding and, more specifically, suggests that Vg needs at least one Zn^{2+}-ion to stabilize the binding to DNA.

Attempting detection of zinc in the β-barrel subdomain

To begin validating the role of zinc in the honey bee Vg β-barrel subdomain, we initially attempted to obtain a sample through dephosphorylation and proteolysis of native Vg at the polyserine linker. However, this was unsuccessful. A weak 40kDa band was obtained under certain conditions, but we could not validate its identity (see Figure S6 for SDS-PAGE gel). Therefore, the subdomain was expressed in *E. coli* with a solubility tag (SUMO) by Genscript Biotech, and ICP-MS was repeated. The Zn^{2+} concentration for the tagged β-barrel subdomain samples was significantly higher than the negative controls of buffer samples (Mann-Whitney U test: w = 0, p-value = 0.0106) (see Figure S7 for ICP-MS results), but not significantly different from samples of the free SUMO-tag, including all samples (Mann-Whitney U test: w = 4, p-value = 0.0937). Excluding one outlier in the SUMO-tag sample set, however, yielded significant results (Mann-Whitney U test: w = 0, p-value = 0.0195). This outcome seems encouraging, but the tagged β-barrel subdomain was exposed to Zn^{2+} during expression in culture, while the free SUMO-tag was synthetically produced without similar opportunity for zinc-binding. We attempted to develop systems to control for this confounding factor (see supplementary.docx and Figure S7 and S8), but without sufficient success.

Discussion

Our study validates that honey bee Vg binds Zn^{2+}, as suggested previously (Amdam et al., 2004). The Zn:Vg ratio calculated by us, 3:1, is higher than the 1 or possible 2 zinc ions reported for each monomer of lamprey Vg (Anderson et al., 1998), which was a calculation based on measurements from the American clawed frog Vg (Montorzi et al., 1994, Auld et al.,

1996). After the initial validation, we performed a sequence and structural analysis to understand the structural basis and possible functional outcomes of the zinc-binding capacity of honey bee Vg. We identified seven zinc clusters located at different subdomains and domains.

We found two clusters in the β-barrel subdomain. The β-barrel subdomain is proposed to function as a transcription factor (Salmela et al., 2021). A classical feature in the DNA binding domain of transcription factor is the coordination of Zn^{2+}, called zinc finger domains (Cassandri et al., 2017). The Zn^{2+}-binding assists the protein in folding, creating the structural form that can recognize DNA (Chang et al., 2010). We show a similarity between a known zinc finger protein and honey bee Vg in the DNA binding motif. The CTCF transcription factor requires coordination of Zn^{2+} at the C2H2 motifs to adopt the correct fold to bind DNA (Maksimenko et al., 2021). Similar to CTCF, the β-barrel subdomain might also bind zinc, and through this process build a different fold than seen in our model. The MSA (Figure S5C) shows two C residues in cluster βb.2 and H229 in the β-barrel subdomain, aligned to C and H zinc-binding residues in CTCF. The three residues are at distant positions in the subdomain (Figure S2). The β-barrel subdomain is presumably cleaved from honey bee Vg when translocating to the nucleus (Salmela et al., 2021). Proteolytic cleavage makes H229 available (no longer associated in the αh.1 cluster). The loop region of H229 allows for flexibility and possible association with the two C residues in cluster βb.2, creating a C2H1 site. Cluster βb.1 consists of three H residues in loop regions, which could fold similarly, resulting in a C2H2 zinc site. Taken together, we demonstrated that the β-barrel subdomain can potentially create a C2H2 zinc site if flexibility in the loop regions is allowed for. Zinc-binding-related flexibility is documented for several C2H2 zinc-binding factors, in addition to CTCF, in insects (Jauch et al., 2003, Stubbs L. et al., 2011, Maksimenko et al., 2021). Such flexibility could be interrupted by

the bound SUMO-tag in our expression system. In addition, the CTCF transcription factor consists of several C2H2 motifs linked together when binding DNA (Maksimenko et al., 2021, Jauch et al., 2003). Similarly, the β-barrel subdomain might require co-factors or proteins to bind DNA (Salmela et al., 2021).

Cluster αh.1 is in a histidine-rich loop in the α-helical subdomain, a loop identified earlier as flexible and insect-specific in honey bee Vg (Havukainen et al., 2013). A structural zinc-binding site could increase the stability of the α-helical subdomain by structuring the loop and stabilizing the N-terminal domain through interaction with H229 (Figure 2C). However, proteolytic cleavage would disengage H229 from the site. We suggest that the remaining four H located in the loop (Figure 2C) could create a new structural zinc site in such situations, similar to some zinc transporter proteins (Fukada and Kambe, 2011, Tanaka et al., 2013, Zhang et al., 2019). We suggest that a conformational change induced by zinc is feasible for the flexible loop and could stabilize the α-helical subdomain and, in turn, the lipid binding site. Cluster αh.1 is at the surface of honey bee Vg. We propose that the αh.1 cluster could sense the cellular environment more efficiently than the clusters inside the lipid cavity. Zinc transporters with similar histidine-loop coordination sites regulate zinc homeostasis (Fukada and Kambe, 2011), supporting our findings.

Electrostatic and hydrophobic interactions between the α-helices in the α-helical subdomain and two β-sheets in the DUF1943 stabilize the Vg lipid cavity (Babin et al., 1999, Smolenaars et al., 2007, Biterova et al., 2019). We suggest that structural zinc-binding sites in the cavity could support stability. Clusters αh.2, Duf.1, and Duf.2 have residues at the α-helical subdomain and the DUF1943. The clusters are in conserved (Figure S3B and S3C) and hydrophobic regions (see Figure S4 for a structural overview of the hydrophobic areas), and

Duf.1 and Duf.2 consist of three H residues, typical at a structural zinc site (Dudev and Lim, 2003). Cluster αh.2 has only one H residue but has an additional C residue, a rare arrangement for a structural site, but identified in a ubiquitin-binding protein ((Lim et al., 2019) PDB-ID: 6H3A), indicating that just two H/C residues could coordinate zinc in cluster αh.2. The structural zinc-binding sites in zinc transporters can have an additional D residue (Fukada and Kambe, 2011), which suggests a possibility for the structural coordination event in cluster αh.2 to include a D residue (D802).

We speculated whether an interaction between H926 at the DUF1943 and the two C residues in cluster βb.2 (Figure S2) is possible. However, the residues are too distant (~30 Å) for interaction (normally ~2.0–2.4 Å (Dudev and Lim, 2003)) in our model. The generally rigid β-sheets (Perczel et al., 2005) at both positions make a conformational change induced by zinc unlikely. H926 and Duf.1 are at the same β-sheet. However, a similar rigidness would also make interaction unlikely. Therefore, H926 is less likely to coordinate zinc, while cluster αh.2 and Duf.1, Duf.2 are feasible structural zinc-binding sites in the lipid cavity and can coordinate zinc (e.g., during transport). We suggest an optimal solution for honey bee Vg would be to carry lipid molecules and zinc in the same location so it could be released together upon delivery.

The well-conserved disulfide bridge on the C-terminal region presumably contributes to a stable structural fold. Such bridges are usually stable oxidative conditions(Sevier and Kaiser, 2002). Two of the disulfide bridges create an interesting arrangement and suggest the possibility for a ZnC4 coordination site, which would probably maintain the stable fold when Vg is experiencing reducing conditions. ZnC4 is a typical coordination site for redox switches (Ilbert et al., 2006, Pace and Weerapana, 2014). Redox switches can sense oxidative stress,

which could generate a response of the protein to change cellular location or release zinc (Ilbert et al., 2006). We propose a similar process that can subside at the Ct cluster in honey bee Vg. This idea has some support in the observation that Vg levels in worker honey bees are positively correlated with oxidative stress resilience (Seehuus et al 2006). It is possible that the zinc released from Vg can explain this phenomenon via binding to protective enzymes or cell membranes (Marreiro et al., 2017, Seehuus et al., 2006).

Regarding the caveats of this study, we assert that *in silico* analysis of the β-barrel subdomain is not fully reflected in our experimental results, despite using several methodologies to approach this problem. Proteolytic cleavage of the SUMO-tag resulted in low yields, and separation during purification of the tag-free subdomain from the SUMO-tag did not work, despite optimization. Changing the solubility-tag to maltose-binding protein (MBP) improved expression yields slightly. However, we faced the same challenges as earlier during purification. Adding two affinity column purification steps followed by a size exclusion chromatography did not successfully separate the SUMO-tag from the subdomain. The tagged subdomain, therefore, became the best option for element analysis. Another drawback of *in silico* prediction is that the orientation for some side chains can be imprecise, even when located in a confident backbone fold (Jumper et al., 2021). The residues in cluster αh.2, H593, H229, and H587 look to have an optimal side chain arrangement to coordinate zinc, creating a small triangle (Figure 2C). However, the side chains might not always be in such an optimal arrangement as seen, for example, in Duf.1. Specifically, the side chain orientation can be inaccurately predicted or could adopt another side chain orientation when zinc is present (Kluska et al., 2018). Due to these caveats, we identified the residues as potential clusters. We also assumed this for cluster βb.1, βb.2, αh.2, and Duf.2 (Figure 2B-C).

Our analysis relied on our recently published full-length protein structure that was predicted using AlphaFold (Leipart et al., 2021). AlphaFold calculates a confidence score that evaluates the predicted structure's stereochemical integrity (Mariani et al., 2013, Jumper et al., 2021). The honey bee Vg prediction has a confidently folded backbone, with the exception of a few short loops and regions that make up approximately 13% of the Vg residues (see Figure 2 in (Leipart et al., 2021)). The zinc clusters that we identified here are fully embedded in confidently folded regions. However, loop regions are flexible structures (Barozet et al., 2021) and residues located in such flexible regions could potentially change the backbone fold when zinc is present, as seen in zinc regulators and zinc transporter proteins (Liu et al., 2021, Tanaka et al., 2013, De Angelis et al., 2010). We assume this could be true in our model, and this insight was applied in clusters βb.1, αh.1, αh.2, and Duf.1 (Figure 2B-C).

While the localization of one or more zinc-binding sites to the recombinantly expressed β-barrel remains uncertain, the presence of, on average, 3 Zn^{2+} cations per honey bee Vg molecule was determined with confidence using ICP-MS (Figure 1). Our structural analysis illustrates where Vg can bind zinc ions, presumably one at each position. The seven sites might provide unknown flexibilities based on Vg having Zn^{2+}-ions bound at different combinations of sites depending on the protein's situation.

Experimental procedures

Collection and purification of honey bee Vg

To obtain Vg, we collected 1-10 µl honey bee hemolymph in a 1:10 dilution in 0.5 M Tris HCl, using BD needles (30G), as described earlier (Aase et al., 2005). The dilution was filtered using an 0.2 µm syringe filter. Vg was purified from honey bee hemolymph with ion-exchange

chromatography using a HiTrap Q FF 1mL column. The sample buffer (0.5 M Tris HCl) and elution buffer (0.5 M Tris HCl with 0.45 M NaCl) was prepared with ion-free water and acid-treated (10% HNO_3 ON) plastic bottles to eliminate ion contamination. Then 400–450 µl diluted hemolymph was manually injected and Vg eluted at the conductivity of 15–22 mS/cm. All fractions from the peak were collected, pooled, and up-concentrated using an Amicon® Ultracel 100 kDa membrane centrifuge filter. We verified the fraction purity by running SDS-PAGE, which contained one band only of the correct size (~180 kDa). The purification protocol was repeated to produce five samples with concentrations between 1.2 and 2.8 µg/µl in 65 µl. Five blank samples were created using the same protocol; here, only sample buffer was injected. The protein concentration, measured with Qubit, confirmed the samples contained no protein. All samples were collected in 1.5 mL Eppendorf tubes pretreated with 10% HNO_3 for 24 hours and dried at 65 °C.

Identification of zinc in honey bee Vg using Inductively Coupled Plasma mass spectrometry

ICP-MS was performed to detect metal ions associated with purified Vg. For the ICP-QQQ-MS of full-length Vg extracted from bees, 32 µL of concentrated ultra-pure(up) nitric acid was added to the 65µL samples (purified Vg and blanks as described above). The samples were placed in a heating cabinet at 90°C for 3 hours and subsequently put into an ultrasound bath for 60 seconds to dissolve any remaining particles before analysis. Then the samples were diluted to 325 µL by adding a solution of 1% (V/V) HNO_3 and 28.5 µg/L Rhodium (Rh, used as an internal Zn-standard), an x5 total volume dilution. This gives a Rh concentration of 20 µg/L in the final sample. For ICP-QQQ-MS analysis on the recombinantly expressed and purified β-barrel, 30 µL of concentrated ultra-pure(up) nitric acid was added to 60µL samples. The same conditions for heat and ultrasonic bath as described above were applied. Then the samples

were diluted to 300 µL by adding a solution of 1% (V/V) HNO_3 and 30 µg/L Rh for an x5 total volume dilution. This gives a Rh concentration of 21 µg/L in the final sample.

Both sample series were analyzed using an Agilent Technologies 8800 ICP-QQQ-MS (Table S1 for instrumental parameters). The ICP-MS was fitted with a micro nebulizer with a flow of 50 µl/min to accommodate small sample volumes. The sample introduction was a high throughput setup. External standards containing Zn were used to calibrate the ICP-MS. Rh was added to the standards in the same concentration as the samples. Zn was analyzed in ammonia mode to remove any interferences on Zn, which was measured at 64 and 66 amu. Mass 64 has an isobaric overlap with ^{64}Nickel (Ni); therefore, an inter-element correction for ^{64}Ni interference on ^{64}Zn was performed. The internal standard Rh was measured at 103 amu.

The method limit of detection and limit of quantification was calculated as 3 times the standard deviation of the buffer blank samples and 10 times the standard deviation of the buffer blank samples.

Multiple sequence alignment

The sequences were aligned using Clustal Omega (McWilliam et al., 2013). The protein sequences used for the multiple sequence alignment were (UniProt ID): *Apis mellifera* (Q868N5), *Athalia rosae* (Q17083), *Pimpla nipponica* (O17428), *Pteromalus puparum* (B2BD67), *Encarsia formosa* (Q698K6), *Bombus ignites* (B9VUV6), *Bombus hypocrite* (C7F9J8), *Solenopsis invicta* (Q7Z1M0 and Q2VQM6), *Riptortus clavatus* (O02024), *Anthonomus grandis* (Q05808), *Lethocerus deyrollei* (B1B5Z4), *Aedes aegypti* (Q16927), *Nilaparvata lugens* (A7BK94), *Graptopsaltria nigrofuscata* (Q9U5F1), *Antheraea pernyi* (Q9GUX5), *Saturnia japonica* (Q59IU3), *Periplaneta Americana* (Q9U8M0), *Blattella germanica* (O76823), *Rhyparobia maderae* (Q5TLA5), *Homalodisca vitripennis* (Q0ZUC7), *Ichthyomyzon unicuspis*

(Q91062), *Acipenser transmontanus* (Q90243), *Oreochromis aureus* (Q9YGK0), *Oncorhynchus mykiss* (Q92093), *Fundulus heteroclitus* (Q90508), *Xenopus laevis* (P18709), *Gallus gallus* (P87498), *Homo sapiens* (P55157) and *Anolis carolinensis* (Q9PUB1), based on Havukainen *et al.* (2011). The protein sequences used in the MSA with the CTCF protein were the same as listed above, but only including the β-barrel subdomain (residue 1 to 326 in honey bee Vg). The CTCF protein sequences used were: *Drosophila melanogaster* (Q9VS55), *Anopheles gambiae* (Q4G266), *Bombyx mori* (H9IXV8), *Tribolium castaneum* (D6WGY1), *Apis mellifera* (A0A7M7MWC7 and A0A7M7MWF4), and *Nasonia vitripennis* (A0A7M7H6S9). The CTCF proteins were identified through BLAST (Altschul et al., 1990), and the species represent the evolutionary relationship between *A. mellifera* and *D. melanogaster* (Honeybee Genome Sequencing, 2006). See Figure S5C for MSA.

Identification of clusters

We used the full-length AlphaFold prediction of honey bee Vg for our structural analysis (Leipart et al., 2021). The first step was to identify conserved H and C residues (See Figure S1 for MSA). Second, we inspected their positions in 3D space to determine the distance and positions in relation to each other. Zinc usually coordinates with a minimum of four residues, so if our initial search did not identify at least four H/C in proximity, we went back to the MSA to identify adjacent (in sequence or 3D space) conserved S/D/E residues or regions of conserved residues (potential coordination through backbone association). Since the structural model lacks water molecules, we could not account for possible coordination involving hydrogen bonds from water molecules. The potential zinc-binding sites found in this way are called clusters. The MSA was uploaded to ConSurf (Ashkenazy et al., 2016) to create a PyMol script coloring atoms based on the degree of conservation. The results are presented in Figure S3. Finally, we evaluated the hydrophobic contrast of each cluster using Dudev and

Lim's (2003) formalism and using the PyMol command based on the Eisenberg hydrophobicity scale (Eisenberg et al., 1984). The results are presented in Figure S4. The *Apis mellifera* (Q868N5) sequence was input to ZincBind (Ireland and Martin, 2019), MotifScan (Pagni et al., 2007), and MOTIF Search (GenomeNet, Kyoto University Bioinformatics Center). Only MotifScan resulted in a positive hit.

Limited proteolysis

A limited proteolysis analysis was performed to obtain the 40 kDa fragment of the β-barrel subdomain. Purified Vg was dephosphorylated with Lambda Protein Phosphatase (New England BioLabs, MA, USA). The samples were incubated with 1 µL Lambda Protein Phosphatase, 5 µL 10x NEBuffer for Protein MetalloPhosphatases and 5 µL of 10 mM $MnCl_2$, for 30 minutes at 30°C. Then 6.5 ng of dephosphorylated Vg was digested with 5 and 10 units of caspase-1 (Sigma-Aldrich) for 2 hours at 37°C, and with 0.01, 0.1, and 1 unit of chymotrypsin (Sigma-Aldrich) for 30 minutes at 25°C. Two standards (ThermoFisher PageRuler™ unstained and pre-stained High Range Protein Ladder), unphosphorylated and undigested full-length Vg, and the digested samples were run on 4–20 % SDS-PAGE gel (Bio-Rad, CA, USA) under reducing conditions.

Recombinant protein and preparation for element analysis

The β-barrel (amino acid 21 to 323) subdomain was produced by Genscript Biotech. The DNA was subcloned into a pET30a vector, with an N-terminal His tag and a SUMO solubility tag. The construct was expressed in E. coli Arctic Express (D3). A single colony was inoculated into an LB medium containing kanamycin, including 100 uM Zn^{2+}. The bacteria were grown and harvested using standard approaches, and the target protein was resolubilized and purified using Ni-affinity purification. The tagged β-barrel subdomain was used in ICP-MS. We made

three sets of blank samples: buffer blank and two sets with SUMO-1 (human, His-tag, Enzo Life Sciences). One collection of SUMO-1 samples were incubated with 25uM $ZnCl_2$ at 4 °C for 1 hour. The buffers in all four sample sets were exchanged to 0.5 M Tris HCl with 0.225 M NaCl. The ICP-MS protocol is explained above. ICP-MS revealed minimal Zn^{2+}-coordination in all samples. The expression system was re-ordered; the LB medium included 42 uM Zn^{2+} or 50 uM Co^{2+}, and one included both 42 uM Zn^{2+} and 50 uM Co^{2+}. A SUMO-tag only expression system with the same three conditions as above was also ordered to avoid false positive results. We used the samples to perform UV-Vis spectroscopy, NMR spectroscopy and intrinsic tryptophan fluorescence spectroscopy for results and protocols (see supplementary.docx and Figure S8).

Acknowledgment

The authors acknowledge The Research Council of Norway grant number 262137 for running costs and positions and BioCat (RCN grant number 249023) for travel grants and conference support.

Data availability statement

All data presented here are included in the main article or the supplementary material (Supplementary.docx). The structural data for honey bee Vg was published earlier and supplemented there (Leipart et al., 2021).

References

ALTSCHUL, S. F., GISH, W., MILLER, W., MYERS, E. W. & LIPMAN, D. J. 1990. Basic local alignment search tool. *J Mol Biol,* 215**,** 403-10.

AMDAM, G. V., SIMOES, Z. L., HAGEN, A., NORBERG, K., SCHRODER, K., MIKKELSEN, O., KIRKWOOD, T. B. & OMHOLT, S. W. 2004. Hormonal control of the yolk precursor vitellogenin regulates immune function and longevity in honeybees. *Exp Gerontol,* 39**,** 767-73.

AMDAM, G. V., AASE, A. L., SEEHUUS, S. C., KIM FONDRK, M., NORBERG, K. & HARTFELDER, K. 2005. Social reversal of immunosenescence in honey bee workers. *Exp Gerontol,* 40**,** 939-47.

ANDERSON, T. A., LEVITT, D. G. & BANASZAK, L. J. 1998. The structural basis of lipid interactions in lipovitellin, a soluble lipoprotein. *Structure,* 6**,** 895-909.

ANDREINI, C., BANCI, L., BERTINI, I. & ROSATO, A. 2006. Counting the Zinc-Proteins Encoded in the Human Genome. *Journal of Proteome Research,* 5**,** 196-201.

ASHKENAZY, H., ABADI, S., MARTZ, E., CHAY, O., MAYROSE, I., PUPKO, T. & BEN-TAL, N. 2016. ConSurf 2016: an improved methodology to estimate and visualize evolutionary conservation in macromolecules. *Nucleic Acids Research,* 44**,** W344-W350.

ATAIE, N. J., HOANG, Q. Q., ZAHNISER, M. P. D., TU, Y., MILNE, A., PETSKO, G. A. & RINGE, D. 2008. Zinc coordination geometry and ligand binding affinity: the structural and kinetic analysis of the second-shell serine 228 residue and the methionine 180 residue of the aminopeptidase from Vibrio proteolyticus. *Biochemistry,* 47**,** 7673-7683.

AULD, D. S., FALCHUK, K. H., ZHANG, K., MONTORZI, M. & VALLEE, B. L. 1996. X-ray absorption fine structure as a monitor of zinc coordination sites during oogenesis of Xenopus laevis. *Proceedings of the National Academy of Sciences,* 93**,** 3227-3231.

BABIN, P. J., BOGERD, J., KOOIMAN, F. P., VAN MARREWIJK, W. J. A. & VAN DER HORST, D. J. 1999. Apolipophorin II/I, Apolipoprotein B, Vitellogenin, and Microsomal Triglyceride Transfer Protein Genes Are Derived from a Common Ancestor. *Journal of Molecular Evolution,* 49**,** 150-160.

BAGLIVO, I., RUSSO, L., ESPOSITO, S., MALGIERI, G., RENDA, M., SALLUZZO, A., DI BLASIO, B., ISERNIA, C., FATTORUSSO, R. & PEDONE, P. V. 2009. The structural role of the zinc ion can be dispensable in prokaryotic zinc-finger domains. *Proceedings of the National Academy of Sciences of the United States of America,* 106**,** 6933-6938.

BALTACI, A. K. & YUCE, K. 2018. Zinc Transporter Proteins. *Neurochemical Research,* 43**,** 517-530.

BAROZET, A., CHACÓN, P. & CORTÉS, J. 2021. Current approaches to flexible loop modeling. *Current research in structural biology,* 3**,** 187-191.

BITEROVA, E. I., ISUPOV, M. N., KEEGAN, R. M., LEBEDEV, A. A., SOHAIL, A. A., LIAQAT, I., ALANEN, H. I. & RUDDOCK, L. W. 2019. The crystal structure of human microsomal triglyceride transfer protein. *Proceedings of the National Academy of Sciences,* 116**,** 17251-17260.

CASSANDRI, M., SMIRNOV, A., NOVELLI, F., PITOLLI, C., AGOSTINI, M., MALEWICZ, M., MELINO, G. & RASCHELLÀ, G. 2017. Zinc-finger proteins in health and disease. *Cell Death Discovery,* 3**,** 17071.

CHANG, S., JIAO, X., HU, J.-P., CHEN, Y. & TIAN, X.-H. 2010. Stability and folding behavior analysis of zinc-finger using simple models. *International journal of molecular sciences,* 11**,** 4014-4034.

DANIEL, A. G. & FARRELL, N. P. 2014. The dynamics of zinc sites in proteins: electronic basis for coordination sphere expansion at structural sites. *Metallomics,* 6**,** 2230-2241.

DE ANGELIS, F., LEE, J. K., CONNELL, J. D., MIERCKE, L. J. W., VERSCHUEREN, K. H., SRINIVASAN, V., BAUVOIS, C., GOVAERTS, C., ROBBINS, R. A., RUYSSCHAERT, J.-M., STROUD, R. M. & VANDENBUSSCHE, G. 2010. Metal-induced conformational changes in ZneB suggest an active role of membrane fusion proteins in efflux resistance systems. *Proceedings of the National Academy of Sciences,* 107**,** 11038.

DUDEV, T. & LIM, C. 2003. Principles Governing Mg, Ca, and Zn Binding and Selectivity in Proteins. *Chemical Reviews,* 103**,** 773-788.

EISENBERG, D., SCHWARZ, E., KOMAROMY, M. & WALL, R. 1984. Analysis of membrane and surface protein sequences with the hydrophobic moment plot. *J Mol Biol,* 179**,** 125-42.

FALCHUK, K. H. 1998. The molecular basis for the role of zinc in developmental biology. *Mol Cell Biochem,* 188**,** 41-8.

FUKADA, T. & KAMBE, T. 2011. Molecular and genetic features of zinc transporters in physiology and pathogenesis†. *Metallomics,* 3**,** 662-674.

GUIDUGLI, K. R., NASCIMENTO, A. M., AMDAM, G. V., BARCHUK, A. R., OMHOLT, S., SIMOES, Z. L. & HARTFELDER, K. 2005. Vitellogenin regulates hormonal dynamics in the worker caste of a eusocial insect. *FEBS Lett,* 579**,** 4961-5.

GUPTA, G., SRIVASTAVA, P. P., GANGWAR, M., VARGHESE, T., CHANU, T. I., GUPTA, S., ANDE, M. P., KRISHNA, G. & JANA, P. 2021. Extra-Fortification of Zinc Upsets Vitellogenin Gene Expression and Antioxidant Status in Female of Clarias magur brooders. *Biological Trace Element Research*.

HAVUKAINEN, H., HALSKAU, O., SKJAERVEN, L., SMEDAL, B. & AMDAM, G. V. 2011. Deconstructing honeybee vitellogenin: novel 40 kDa fragment assigned to its N terminus. *J Exp Biol,* 214**,** 582-92.

HAVUKAINEN, H., MUNCH, D., BAUMANN, A., ZHONG, S., HALSKAU, O., KROGSGAARD, M. & AMDAM, G. V. 2013. Vitellogenin recognizes cell damage through membrane binding and shields living cells from reactive oxygen species. *J Biol Chem,* 288**,** 28369-81.

HAVUKAINEN, H., UNDERHAUG, J., WOLSCHIN, F., AMDAM, G. & HALSKAU, O. 2012. A vitellogenin polyserine cleavage site: highly disordered conformation protected from proteolysis by phosphorylation. *J Exp Biol,* 215**,** 1837-46.

HONEYBEE GENOME SEQUENCING, C. 2006. Insights into social insects from the genome of the honeybee Apis mellifera. *Nature,* 443**,** 931-949.

ILBERT, M., GRAF, P. C. & JAKOB, U. 2006. Zinc center as redox switch--new function for an old motif. *Antioxid Redox Signal,* 8**,** 835-46.

IRELAND, S. M. & MARTIN, A. C. R. 2019. ZincBind—the database of zinc binding sites. *Database,* 2019.

JAUCH, R., BOURENKOV, G. P., CHUNG, H.-R., URLAUB, H., REIDT, U., JÄCKLE, H. & WAHL, M. C. 2003. The Zinc Finger-Associated Domain of the Drosophila Transcription Factor Grauzone Is a Novel Zinc-Coordinating Protein-Protein Interaction Module. *Structure,* 11**,** 1393-1402.

JERNIGAN, R., RAGHUNATHAN, G. & BAHAR, I. 1994. Characterization of interactions and metal ion binding sites in proteins. *Current Opinion in Structural Biology,* 4**,** 256-263.

JUMPER, J., EVANS, R., PRITZEL, A., GREEN, T., FIGURNOV, M., RONNEBERGER, O., TUNYASUVUNAKOOL, K., BATES, R., ŽÍDEK, A., POTAPENKO, A., BRIDGLAND, A., MEYER, C., KOHL, S. A. A., BALLARD, A. J., COWIE, A., ROMERA-PAREDES, B., NIKOLOV, S., JAIN, R., ADLER, J., BACK, T., PETERSEN, S., REIMAN, D., CLANCY, E., ZIELINSKI, M., STEINEGGER, M., PACHOLSKA, M., BERGHAMMER, T., BODENSTEIN, S., SILVER, D., VINYALS, O., SENIOR, A. W., KAVUKCUOGLU, K., KOHLI, P. & HASSABIS, D. 2021. Highly accurate protein structure prediction with AlphaFold. *Nature,* 596**,** 583-589.

KLUSKA, K., ADAMCZYK, J. & KRĘŻEL, A. 2018. Metal binding properties, stability and reactivity of zinc fingers. *Coordination Chemistry Reviews,* 367**,** 18-64.

LEIPART, V., MONTSERRAT-CANALS, M., CUNHA, E. S., LUECKE, H., HERRERO-GALÁN, E., HALSKAU, Ø. & AMDAM, G. V. 2021. Structure prediction of honey bee vitellogenin: a multi-domain protein important for insect immunity. *FEBS Open Bio,* In press.

LIM, M., NEWMAN, J. A., WILLIAMS, H. L., MASINO, L., AITKENHEAD, H., GRAVARD, A. E., GILEADI, O. & SVEJSTRUP, J. Q. 2019. A Ubiquitin-Binding Domain that Binds a Structural Fold Distinct from that of Ubiquitin. *Structure,* 27**,** 1316-1325.e6.

LIU, F., SU, Z., CHEN, P., TIAN, X., WU, L., TANG, D.-J., LI, P., DENG, H., DING, P., FU, Q., TANG, J.-L. & MING, Z. 2021. Structural basis for zinc-induced activation of a zinc uptake transcriptional regulator. *Nucleic Acids Research,* 49**,** 6511-6528.

MAKSIMENKO, O. G., FURSENKO, D. V., BELOVA, E. V. & GEORGIEV, P. G. 2021. CTCF As an Example of DNA-Binding Transcription Factors Containing Clusters of C2H2-Type Zinc Fingers. *Acta Naturae,* 13**,** 31-46.

MARIANI, V., BIASINI, M., BARBATO, A. & SCHWEDE, T. 2013. lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests. *Bioinformatics (Oxford, England),* 29**,** 2722-2728.

MARREIRO, D. D. N., CRUZ, K. J. C., MORAIS, J. B. S., BESERRA, J. B., SEVERO, J. S. & DE OLIVEIRA, A. R. S. 2017. Zinc and Oxidative Stress: Current Mechanisms. *Antioxidants (Basel, Switzerland),* 6**,** 24.

MARTIN, D. J. & RAINBOW, P. S. 1998. The kinetics of zinc and cadmium in the haemolymph of the shore crab Carcinus maenas (L.). *Aquatic Toxicology,* 40**,** 203-231.

MATOZZO, V., GAGNÉ, F., MARIN, M. G., RICCIARDI, F. & BLAISE, C. 2008. Vitellogenin as a biomarker of exposure to estrogenic compounds in aquatic invertebrates: A review. *Environment International,* 34**,** 531-545.

MCWILLIAM, H., LI, W., ULUDAG, M., SQUIZZATO, S., PARK, Y. M., BUSO, N., COWLEY, A. P. & LOPEZ, R. 2013. Analysis Tool Web Services from the EMBL-EBI. *Nucleic Acids Research,* 41**,** W597-W600.

MITCHELL, M. A. & CARLISLE, A. J. 1991. Plasma zinc as an index of vitellogenin production and reproductive status in the domestic fowl. *Comparative Biochemistry and Physiology Part A: Physiology,* 100**,** 719-724.

MONTORZI, M., FALCHUK, K. H. & VALLEE, B. L. 1994. Xenopus laevis vitellogenin is a zinc protein. *Biochemical and biophysical research communications,* 200**,** 1407-1413.

PACE, N. J. & WEERAPANA, E. 2014. Zinc-binding cysteines: diverse functions and structural motifs. *Biomolecules,* 4**,** 419-34.

PAGNI, M., IOANNIDIS, V., CERUTTI, L., ZAHN-ZABAL, M., JONGENEEL, C. V., HAU, J., MARTIN, O., KUZNETSOV, D. & FALQUET, L. 2007. MyHits: improvements to an interactive resource for analyzing protein sequences. *Nucleic Acids Res,* 35**,** W433-7.

PAN, M. L., BELL, W. J. & TELFER, W. H. 1969. Vitellogenic Blood Protein Synthesis by Insect Fat Body. *Science,* 165**,** 393.

PERCZEL, A., GÁSPÁRI, Z. & CSIZMADIA, I. G. 2005. Structure and stability of β-pleated sheets*. *Journal of Computational Chemistry,* 26**,** 1155-1168.

SALMELA, H., HARWOOD, G., MÜNCH, D., ELSIK, C., HERRERO-GALÁN, E., VARTIAINEN, M. K. & AMDAM, G. 2021. Nuclear Translocation of Vitellogenin in the Honey Bee (Apis mellifera). *bioRxiv***,** 2021.08.18.456851.

SAPPINGTON, T. W. & S. RAIKHEL, A. 1998. Molecular characteristics of insect vitellogenins and vitellogenin receptors. *Insect Biochemistry and Molecular Biology,* 28**,** 277-300.

SEEHUUS, S. C., NORBERG, K., GIMSA, U., KREKLING, T. & AMDAM, G. V. 2006. Reproductive protein protects functionally sterile honey bee workers from oxidative stress. *Proc Natl Acad Sci U S A,* 103**,** 962-7.

SEVIER, C. S. & KAISER, C. A. 2002. Formation and transfer of disulphide bonds in living cells. *Nature Reviews Molecular Cell Biology,* 3**,** 836-847.

SLOUP, V., JANKOVSKÁ, I., NECHYBOVÁ, S., PEŘINKOVÁ, P. & LANGROVÁ, I. 2017. Zinc in the Animal Organism: A Review. *Scientia Agriculturae Bohemica,* 48**,** 13-21.

SMOLENAARS, M. M. W., MADSEN, O., RODENBURG, K. W. & VAN DER HORST, D. J. 2007. Molecular diversity and evolution of the large lipid transfer protein superfamily^{s⃞}. *Journal of Lipid Research,* 48**,** 489-502.

STUBBS L., SUN Y. & CAETANO-ANOLLES D. 2011. Function and Evolution of C2H2 Zinc Finger Arrays. . *In:* T., H. (ed.) *A Handbook of Transcription Factors. Subcellular Biochemistry.* Springer, Dordrecht.

SULLIVAN, C. V. & YILMAZ, O. 2018. Vitellogenesis and Yolk Proteins, Fish. *In:* SKINNER, M. K. (ed.) *Encyclopedia of Reproduction (Second Edition).* Oxford: Academic Press.

TANAKA, N., KAWACHI, M., FUJIWARA, T. & MAESHIMA, M. 2013. Zinc-binding and structural properties of the histidine-rich loop of Arabidopsis thaliana vacuolar membrane zinc transporter MTP1. *FEBS Open Bio,* 3**,** 218-224.

TUFAIL, M. & TAKEDA, M. 2008. Molecular characteristics of insect vitellogenins. *Journal of Insect Physiology,* 54**,** 1447-1458.

ZHANG, T., KULIYEV, E., SUI, D. & HU, J. 2019. The histidine-rich loop in the extracellular domain of ZIP4 binds zinc and plays a role in zinc transport. *Biochem J,* 476**,** 1791-1803.

AASE, A. L. T. O., AMDAM, G. V., HAGEN, A. & OMHOLT, S. W. 2005. A new method for rearing genetically manipulated honey bee workers. *Apidologie,* 36**,** 293-299.

Tables

Table 1: Identified zinc clusters.

Cluster	Vg domain	Residues
βb.1	β-barrel subdomain	H20, H113, D143, E147, H265
βb.2	β-barrel subdomain	E171, D172, S173, C178, E179, C222, D223
αh.1	α-helical subdomain	H229, H577, H587, H593, H602
αh.2	α-helical subdomain	H697, E698, C701, S775, S800, D802
Duf.1	DUF1943	E987, H988, H990, H1045, D1046
Duf.2	DUF1943	H445, E449, D996, H1000, H1035, S1037
Ct	C-terminal	C1687, C1711, C1715, C1768

The clusters are named (column 1) based on their location in honey bee Vg (column 2). The residues included in each cluster are listed in column 3.

Figure legends

Figure 1. ICP-MS. A) The concentration measured with ICP-MS for the 5 blank and full-length Vg samples is plotted. The mean and the standard deviation of the mean are indicated for each group (lines). **B)** The calculated molecular Zn:Vg ratio for each sample of Vg is provided, and the calculated mean value for all samples is included as a separate bar.

Figure 2. Identified clusters: A) 2D representation of the honey bee Vg with conserved domain, subdomains, and regions (gray boxes). The amino acid sequence was divided into two regions based on the naturally cleaved fragments, the α-helical subdomain, the lipid binding site, including vWF and C-terminal region (blue, 150 kDa), and the β-barrel subdomain (green, 40 kDa). The identified clusters are marked below the gray boxes (αh.1: green, αh.2: red, Duf.1: blue, Duf.2: orange, Ct: yellow, βb.1: pink and βb.2: cyan). The two zinc-binding

locations identified by studies of lamprey are marked above the gray boxes with gray dots (L). The DNA binding motif (pink box, DNA) and the zinc-binding motif identified by MotifScan (yellow box, web) are also marked above the gray boxes. Finally, the two conserved disulfide bridges not included in a cluster are included here (smaller brown dots). **B)** Identified clusters in the 40 kDa fragment of honey bee Vg. The residues of cluster βb.1 (magenta) and βb.2 (cyan) are in the β-barrel subdomain (green) and close to the DNA binding motif (pink β-sheet). The disulfide bridge in cluster βb.2 is shown as a yellow stick. **C)** Identified clusters in the 150 kDa fragment of honey bee Vg. Cluster αh.1: The insect-specific loop region in the α-helical subdomain (purple) and a short β-strand from the β-barrel subdomain (green) contains the residues of cluster αh.1 (green sticks). Cluster αh.2: The α-helical subdomain (purple) and the β-sheet region in the DUF1943 (light purple) are shown as a cartoon. The identified cluster-residues are shown as red sticks. Cluster Duf.1: Two loops from the DUF1943 (light purple) contain the identified residues, shown as blue sticks. The zinc-binding motif identified by MotifScan (yellow) is located in the neighboring β-strands in the DUF1943. Cluster Duf.2: Cluster Duf.2 (orange sticks) is found further down on the same β-sheet in DUF1943 (light purple). Two of the residues are located in the α-helical subdomain (purple). Cluster Ct: Three disulfide bridges (yellow sticks) are found in the C-terminal (light pink). The dotted box zoom on two of the disulfide bridges shows them from another direction. The two disulfide bridges can create a tetrahedral geometry and are identified as cluster Ct.

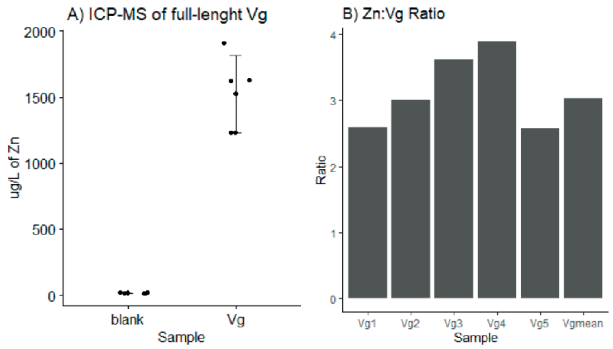


Figure 1. ICP-MS. A) The concentration measured with ICP-MS for the 5 samples of blank and full-length Vg, are plotted. The mean and the standard deviation of the mean is indicated for each group (lines). B) The calculated molecular Zn:Vg ratio for each sample of Vg is shown, and the calculated mean value for all samples is included as a separate bar.

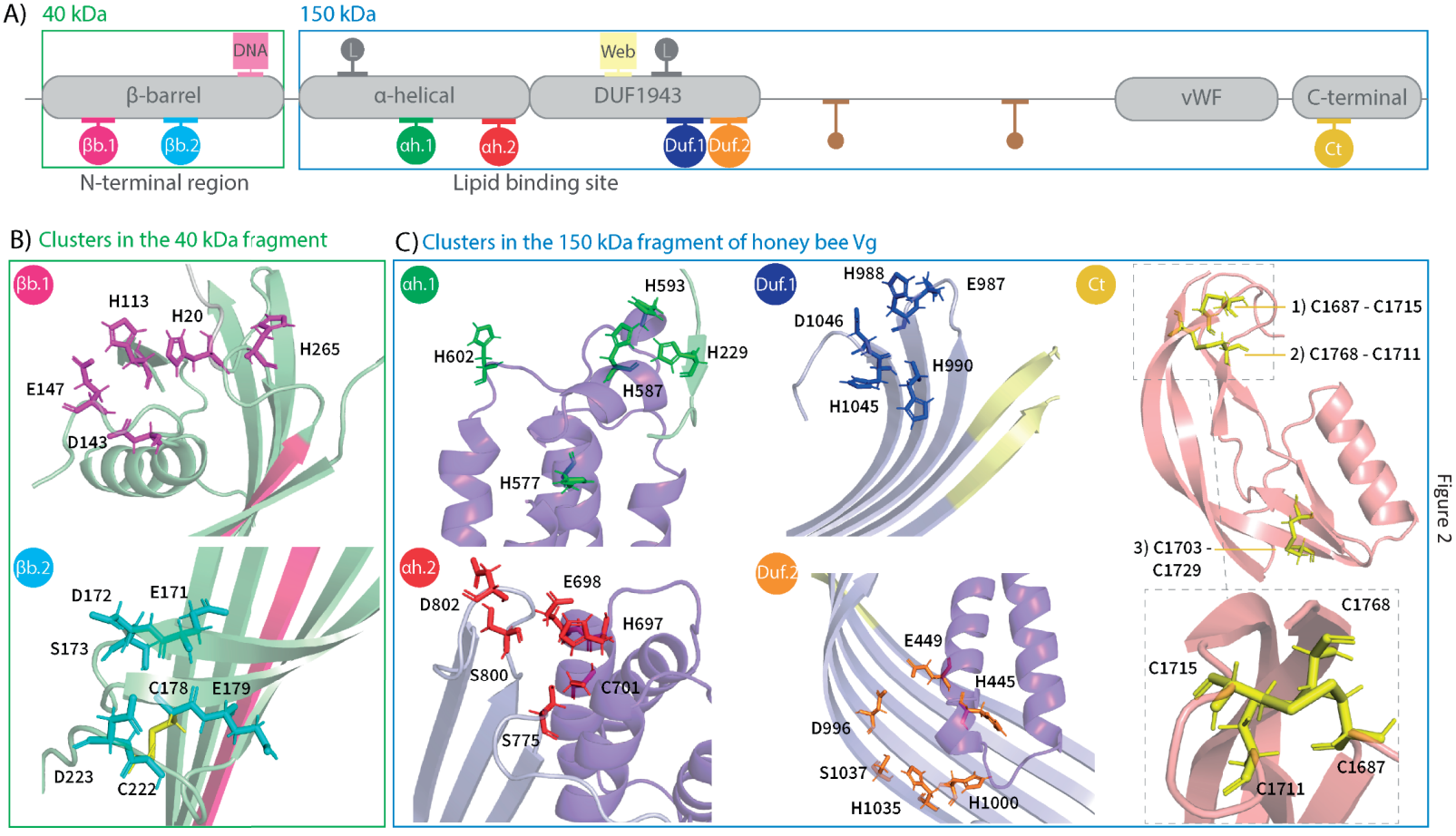


Figure 2. Identified clusters: A) 2D representation of the honey bee Vg with conserved domain, subdomains and regions (gray boxes). The amino acid sequence was divided into two regions based on the naturally cleaved fragments, the lipid binding site, including vWF and C-terminal region (blue, 150 kDa) and the β-barrel subdomain (green, 40 kDa). The identified clusters are marked below the gray boxes (αh.1: green, αh.2: red, Duf.1: blue, Duf.2: orange, Ct: yellow, βb.1: pink and βb.2: cyan). The two zinc binding locations identified by studies of lamprey is marked above the gray boxes with gray dots (L). The DNA binding motif (pink box, DNA) and the zinc binding motif identified by MotifScan (yellow box, Web) are also marked above the gray boxes. Lastly, the two locations of conserved disulfide bridges, not included in a cluster is included here (brown smaller dots). B) Identified clusters in the 40 kDa fragment of honey bee Vg. The residues of cluster βb.1 (magenta) and βb.2 (cyan) is in the β-barrel domain (green) and is close to the DNA binding motif (pink β-sheet). The disulfide bridge in cluster βb.2 is show in as a yellow stick. C) Identified clusters in the 150 kDa fragment of honey bee Vg. Cluster αh.1: The insect specific loop region in the α-helical domain (purple) and a short β-strand from the β-barrel subdomain (green) contains the residues of cluster αh.1 (green sticks). Cluster αh.2: The α-helical domain (purple) and the β-sheet region in the DUF1943 domain (light purple) is shown as cartoon. The identified cluster-residues are shown as red sticks. Cluster Duf.1: Two loops from the DUF1943 domain (light purple) contains the identified residues, shown as blue sticks, and the zinc binding motif identified by MotifScan (yellow) is located in the neighboring β-strands in the DUF1943 domain. Cluster Duf.2: Further down on the same β-sheet in DUF1943 domain (light purple) we find cluster Duf.2 (orange sticks). Two of the residues are located in the α-helical domain (purple). Cluster Ct: Three disulfide bridges (yellow sticks) are found in the C-terminal (light pink). The dotted box zoom on two of the disulfide bridges, showing them from another direction. The two disulfide bridges can create a tetrahedral geometry and is identified as cluster Ct.

Supplementary Material

Table S1. Instrumental parameters used for Agilent 8800 ICP-MS

ICP-MS	**Settings[1]**	**Settings[2]**
RF power	1600 W	1600 W
Plasma gas	15 L/min	15 L/min
Auxiliary gas	0,9 L/min	0,9 L/min
Nebulizer gas	0,90 L/min	0,95 L/min
Makeup gas	0,30 l/min	0,20 l/min
Nebulizer pump	0,32 rpm	0,32 rpm
Sampler/skimmer	Ni	Ni
Sample Depth	8,0 mm	8,0 mm
Data registration for ICP-MS		
Rinse time	20 s	20 s
Flush time	50 s	50 s
Read delay	15 s	15 s
Scanning mode	peak hop	peak hop
Points/spectral peak	1	1
Sweeps/reading	10	10
Replicates	5	5
P/A detector (puls/analog)	on	on
Temperature spray chamber	12 °C	2 °C

1) First analysis series (full-length Vg)
2) Second analysis series (β-barrel subdomain)

Supplementary methods for detection of zinc in the β-barrel subdomain

ICP-MS continued

We attempted to control for Zn^{2+} interacting with the SUMO-tag by incubating the SUMO-tag with Zn^{2+} before ICP-MS. Incubation resulted in significantly higher Zn^{2+} levels in SUMO-tag samples compared to the non-incubated tagged β-barrel subdomain (Mann-Whitney U test: w = 0, p-value = 0.0114). The difference in concentration indicated that the incubation caused the association of Zn^{2+} with the SUMO-tag. However, we could not rule out more specific Zn^{2+}-binding to SUMO. The net determined Zn^{2+} concentration for the β-barrel subdomain was much lower compared to the concentration for native full-length Vg. Figure 1A shows a mean of 1524.0000 µg/L (SD ± 261.6562), while Figure S7 shows the mean as 64.2000 µg/L (SD ± 13.4350), and the significance suggested 0.01 bound Zn^{2+} per β-barrel molecule. We attempted to examine this result with an independent approach and expressed the β-barrel subdomain in the presence of Co^{2+}, attempting to displace Zn^{2+} with this cation. Co^{2+} is considered a good structural and functional model for studying Zn^{2+}-binding sites as the coordination is similar and exchange for Zn^{2+} to Co^{2+} occurs in nature (Lane and Morel, 2000, Shumilina et al., 2014). In contrast to Zn^{2+}, Co^{2+} coordination causes a readily detected change in the protein's UV-Vis spectrum (Bertini and Luchinat, 1984, Shumilina et al., 2014, Sivo et al., 2017). The conducted experiments are presented below.

UV-Vis Spectroscopy

To identify the presence of Zn^{2+}-binding sites by Co^{2+} substitution, SUMO fusion proteins were expressed and purified as described in the main manuscript (i.e., expressed in the presence of Zn^{2+} [42 uM], Zn^{2+} and Co^{2+} [42 uM and 50 uM, respectively], and Co2+ [50 uM]).

Initially, in a storage buffer consisting of PBS at pH 7.4, 10% glycerol, and 0.5 M L-arginine, were centrifuged at 17000xg for 10 min. Supernatants were concentrated using Amicon filters with 10 kDa cut-off at 3250xg for 30 min, producing approximately 4 mg/mL protein samples. These were transferred to the UV-VIS cuvette (path length 10 mm, 500 uL capacity, Hellma item number 108-002-10-40). A UV-Vis spectra in the range of 200–800nm at room temperature was acquired. If Co^{2+} successfully replaced Zn^{2+}, we would have expected to see a distinct Co^{2+}-specific peak pattern near the 500–750 nm range (Sivo et al., 2017, Lane and Morel, 2000). However, we did not observe this (negative results are presented in Figure S8A). The samples were then used for NMR and intrinsic tryptophan fluorescence spectroscopy (see below).

NMR Spectroscopy

Co^{2+} can create a paramagnetic shift in protein NMR spectra (Lane and Morel, 2000). Therefore, we transferred stocks of SUMO-fusion proteins exposed to Zn^{2+}, Zn^{2+} Co^{2+}, and Co^{2+} (as described above) to NMR tubes. D_2O up to 5% v/v was added. We then acquired 1D proton spectra (25°C, water suppression using Watergate, 512 scans, processed using exponential multiplication with a line broadening of 0.3 Hz), and inspected the spectra without finding any significant differences between the samples (Figure S8B).

Intrinsic Tryptophan Fluorescence Spectroscopy

We also looked for fold changes in response to divalent cation(s) present during expression. To do this, we transferred 5 uL of the NMR samples (described above) into 300 PBS buffer at pH 6.5 in a quartz cuvette (path length 5 mm) and conducted an emission scan (excitation wavelength 295 nm, slit widths 5 nm, 310–500 nm). The spectra, which primarily provide

information on the microenvironment of the tryptophans in the protein (Knappskog and Haavik, 1995, Takita et al., 2003) did not show any significant difference.

Figure legends

Figure S1: Multiple sequence alignment. Snapshots of the multiple sequence alignment. All the included species are noted with their UniProt ID. Residues included in clusters are in bold colors (αh.1: green, αh.2: red, Duf.1: dark blue, Duf.2: orange, Ct: yellow, βb.1: pink and βb.2: cyan). Some alignment regions are excluded (noted with "…") since they are not relevant to this study or have significant gaps. **A)** The residues from βb.1, βb.2, and one residue from cluster αh.1 are in the β-barrel subdomain. In addition, the DNA binding motif (pink box) is part of this subdomain. The conserved residues are colored (bold black). **B)** Cluster αh.1, αh.2, Duf.1 and Duf.2 are in the lipid binding site (DUF1943). In addition, the suggested zinc coordinating residues from studies in Lamprey (gray bold) and the MotifScan zinc-binding site (yellow box) are found in this region. H926 is colored (black bold). **C)** No cluster was identified in the region before the vWF domain, but the conserved disulfide bridges residues are colored (brown bold). **D)** Cluster Ct is in the C-terminal region, where all the conserved C and H residues are marked.

Figure S2: MotifScan. The zinc-binding motif (yellow) identified by MotifScan is located in the DUF1943 domain (purple) but extends into the cavity of the β-barrel (green). The predicted zinc-binding H residue (H926) is shown as a stick. Cluster Duf.1 (blue spheres), βb.1 (magenta spheres), βb.2 (cyan spheres), H229 from αh.1 (green stick), and the DNA binding motif (pink) is in close proximation to this predicted zinc-binding motif.

Figure S3: Conservation. Residues are colored using ConSurf, based on the MSA. The low to highly conserved residues are colored from light blue to dark pink (scale presented in the lower-right corner). Both the secondary structures and the spheres (representing the clusters) are colored according to this scale. **A)** The buried residues in the β-barrel subdomain are well

conserved, including the residues in cluster βb.1, βb.2, and the DNA binding motif β-sheet. The regions closer to the surface are less conserved. **B)** The α-helices in the α-helical subdomain are well conserved. The residues in Cluster αh.1 and αh.2 are also conserved. **C)** One of the β-sheet in the DUF1943 domain includes cluster Duf.1, Duf.2, and the zinc motif identified by MotifScan. The conservation of the residues in the β-sheet are variable, but the clusters and zinc motif are conserved. As shown in the MSA, residue H926 is not conserved. **D)** The C-terminal is generally not conserved, except the four residues presented as cluster Ct and the third disulfide bridge (labeled).

Figure S4: Hydrophobicity. The surface and secondary structure are colored using the Eisenberg hydrophobicity scale (scale presented in the lower-left corner). In both panels, the clusters are shown as spheres and marked with a blue dotted circle. Their respective domains are presented as surface and cartoon. **A)** Cluster αh.2 is the position between the highly hydrophobic core of the lipid binding site (β-sheet) and the polar surface of the α-helical subdomain. **B)** Cluster Duf.1 and Duf.2 are positioned on the same β-sheet that make up one side of the lipid binding site (highly hydrophobic), while the other side is facing the surface and is more polar.

Figure S5: Logo representation of DNA binding site motifs and sequence analysis of CTCF. **A)** The most significant motif (motif A) found by Salmela *et al.* (2021) for Vg-DNA binding sites, as shown in Figure 5. **B)** The motif for CTCF in Drosophila melanogaster from the JASPAR database (Castro-Mondragon et al., 2021) (matrix MA0531.1). The logo representation was made with WebLogo3 (Schneider and Stephens, 1990, Crooks et al., 2004). **C)** Residue 140 to 233 in the β-barrel subdomain, using the same species as in the full-length MSA (Figure S1),

aligned to CTCF proteins. The conserved residues from the β-barrel subdomain, identified in the CTCF proteins, are in bold.

Figure S6. Proteolysis of honey bee Vg by caspase-1 and chymotrypsin. The black arrow emphasizes the probable 40 kDa cleavage products of the full-length Vg (flVg) with 5 and 10 units of caspase-1 in lanes 3 and 4, respectively. We did not identify a clear 150 kDa band. The smaller bands outside the range of the standard could be the lambda protein phosphates (25 kDa) or caspase-1 (30 kDa). The chymotrypsin (lane 5 to 7) cleaves Vg completely into small fragments, and no 40 kDa band was identified. The smaller bands outside the range of the standard could be lambda protein phosphates (25 kDa) or chymotrypsin (25 kDa).

Figure S7. ICP-MS results for the β-barrel subdomain. The concentration was measured with ICP-MS for the x5 samples of SUMO tagged β-barrel subdomain (bb), sample buffer (blk), non-incubated SUMO tag (Sblk), and SUMO-tag incubated with Zn^{2+} (SZnblk). The mean and the standard deviation of the mean are indicated for each group.

Figure S8. Spectroscopic analyses of SUMO-fusion proteins expressed in Zn^{2+}, Zn^{2+} and Co^{2+}, and Co^{2+} medium. A) UV-Vis spectra of protein expressed in medium enriched with Zn^{2+} (green traces), Zn^{2+} and Co^{2+}(blue traces), and Co^{2+} (red traces). Expressions of the Sumo tag only are represented by dashed lines, and the SUMO beta-barrel fusion protein is represented by whole lines. **B)** Amide region from ^{1}H NMR spectra of SUMO beta-barrel fusion protein expressed in medium enriched with Zn^{2+} (green trace), Zn^{2+} and Co^{2+}(blue trace), and Co^{2+} (red trace). **C)** Intrinsic tryptophan fluorescence spectra of SUMO beta-barrel fusion protein expressed in medium enriched with Zn^{2+} (green trace), Zn^{2+} and Co^{2+}(blue traces), and Co^{2+} (red trace).

References

BERTINI, I. & LUCHINAT, C. 1984. High spin cobalt(II) as a probe for the investigation of metalloproteins. *Adv Inorg Biochem,* 6**,** 71-111.

CASTRO-MONDRAGON, J. A., RIUDAVETS-PUIG, R., RAULUSEVICIUTE, I., BERHANU LEMMA, R., TURCHI, L., BLANC-MATHIEU, R., LUCAS, J., BODDIE, P., KHAN, A., MANOSALVA PÉREZ, N., FORNES, O., LEUNG, TIFFANY Y., AGUIRRE, A., HAMMAL, F., SCHMELTER, D., BARANASIC, D., BALLESTER, B., SANDELIN, A., LENHARD, B., VANDEPOELE, K., WASSERMAN, W. W., PARCY, F. & MATHELIER, A. 2021. JASPAR 2022: the 9th release of the open-access database of transcription factor binding profiles. *Nucleic Acids Research*.

CROOKS, G. E., HON, G., CHANDONIA, J. M. & BRENNER, S. E. 2004. WebLogo: a sequence logo generator. *Genome Res,* 14**,** 1188-90.

KNAPPSKOG, P. M. & HAAVIK, J. 1995. Tryptophan Fluorescence of Human Phenylalanine Hydroxylase Produced in Escherichia coli. *Biochemistry,* 34**,** 11790-11799.

LANE, T. W. & MOREL, F. M. 2000. Regulation of carbonic anhydrase expression by zinc, cobalt, and carbon dioxide in the marine diatom Thalassiosira weissflogii. *Plant physiology,* 123**,** 345-352.

SCHNEIDER, T. D. & STEPHENS, R. M. 1990. Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res,* 18**,** 6097-100.

SHUMILINA, E., DOBROVOLSKA, O., DEL CONTE, R., HOLEN, H. W. & DIKIY, A. 2014. Competitive cobalt for zinc substitution in mammalian methionine sulfoxide reductase B1 overexpressed in E. coli: structural and functional insight. *Journal of biological inorganic chemistry : JBIC : a publication of the Society of Biological Inorganic Chemistry,* 19**,** 85-95.

SIVO, V., D'ABROSCA, G., RUSSO, L., IACOVINO, R., PEDONE, P. V., FATTORUSSO, R., ISERNIA, C. & MALGIERI, G. 2017. Co(II) Coordination in Prokaryotic Zinc Finger Domains as Revealed by UV-Vis Spectroscopy. *Bioinorg Chem Appl,* 2017**,** 1527247.

TAKITA, T., NAKAGOSHI, M., INOUYE, K. & TONOMURA, B. I. 2003. Lysyl-tRNA Synthetase from Bacillus stearothermophilus: The Trp314 Residue is Shielded in a Non-polar Environment and is Responsible for the Fluorescence Changes Observed in the Amino Acid Activation Reaction. *Journal of Molecular Biology,* 325**,** 677-695.

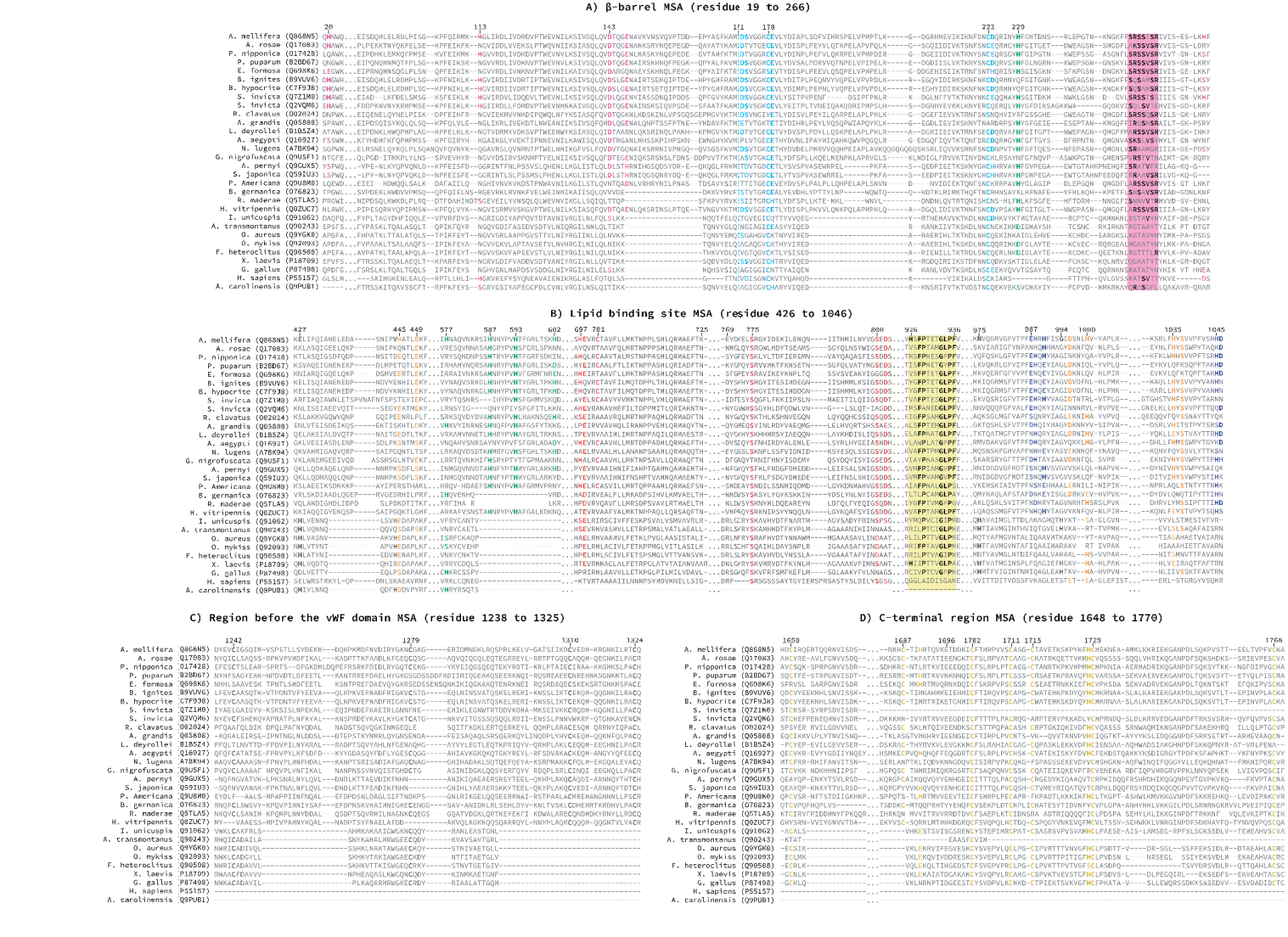

A) β-barrel MSA (residue 19 to 266)
B) Lipid binding site MSA (residue 426 to 1046)
C) Region before the vWF domain MSA (residue 1238 to 1325)
D) C-terminal region MSA (residue 1648 to 1770)

Figure S1

Figure S2

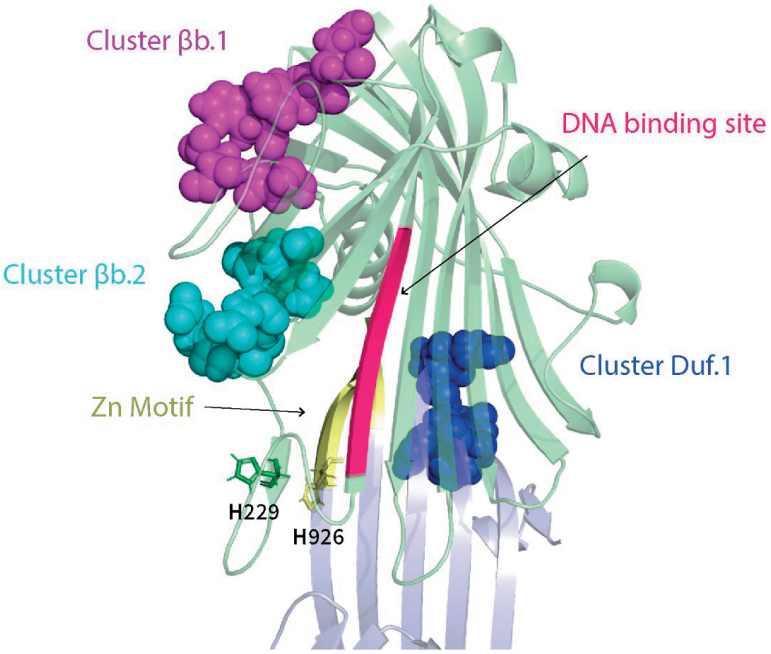
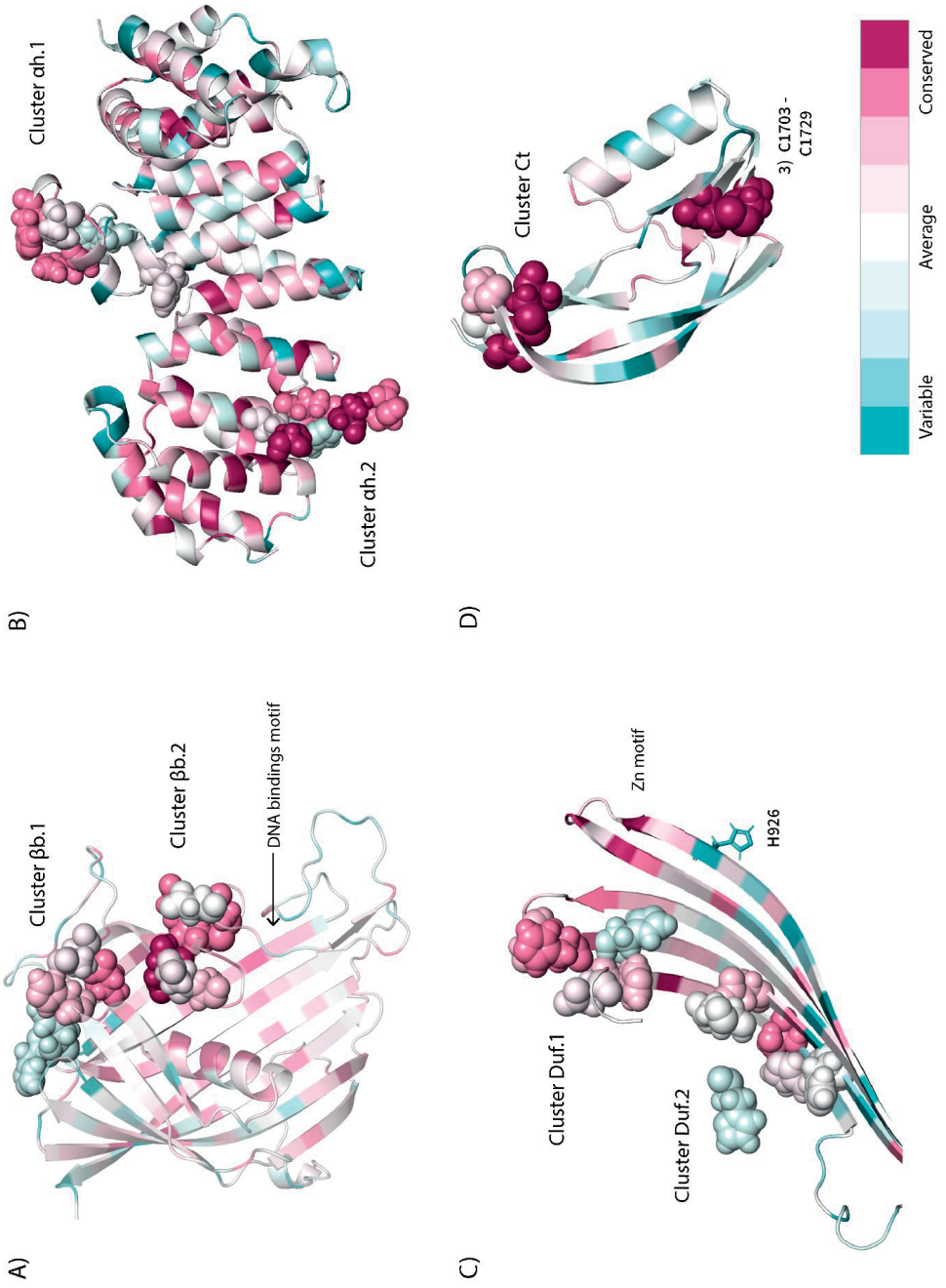

Cluster βb.1
DNA binding site
Cluster βb.2
Cluster Duf.1
Zn Motif
H229
H926

Figure S3


A)
Cluster βb.1
Cluster βb.2
DNA bindings motif
B)
Cluster αh.1
Cluster αh.2
C)
Cluster Duf.1
Cluster Duf.2
Zn motif
H926
D)
Cluster Ct
3) C1703 - C1729
Variable
Average
Conserved

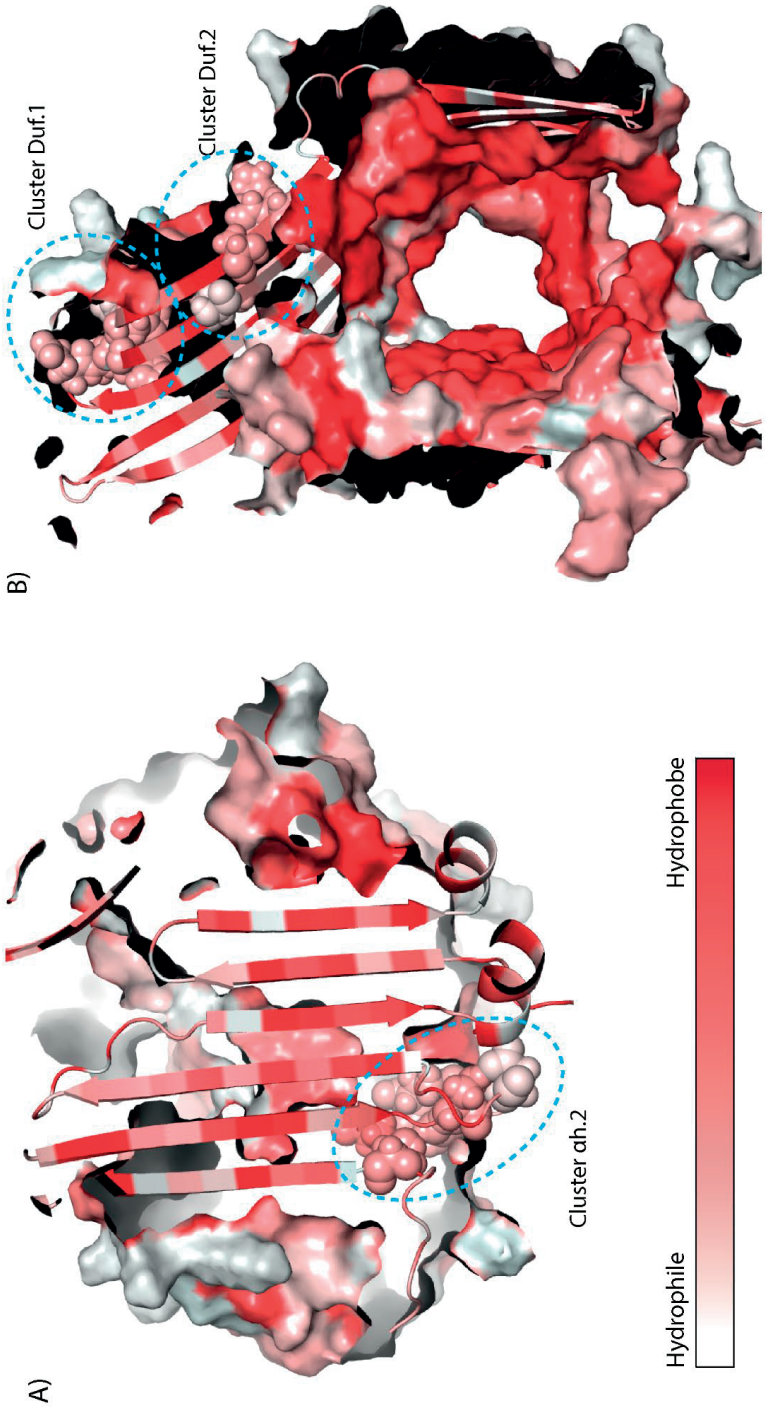

B)
Cluster Duf.1
Cluster Duf.2
A)
Cluster dh.2
Hydrophobe
Hydrophile

Figure S5

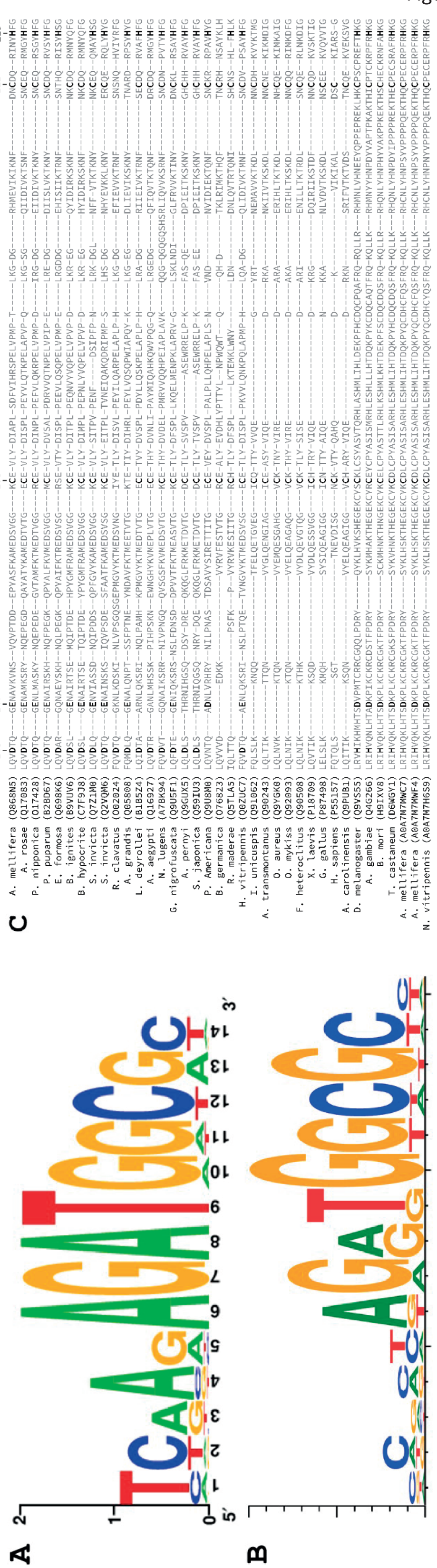
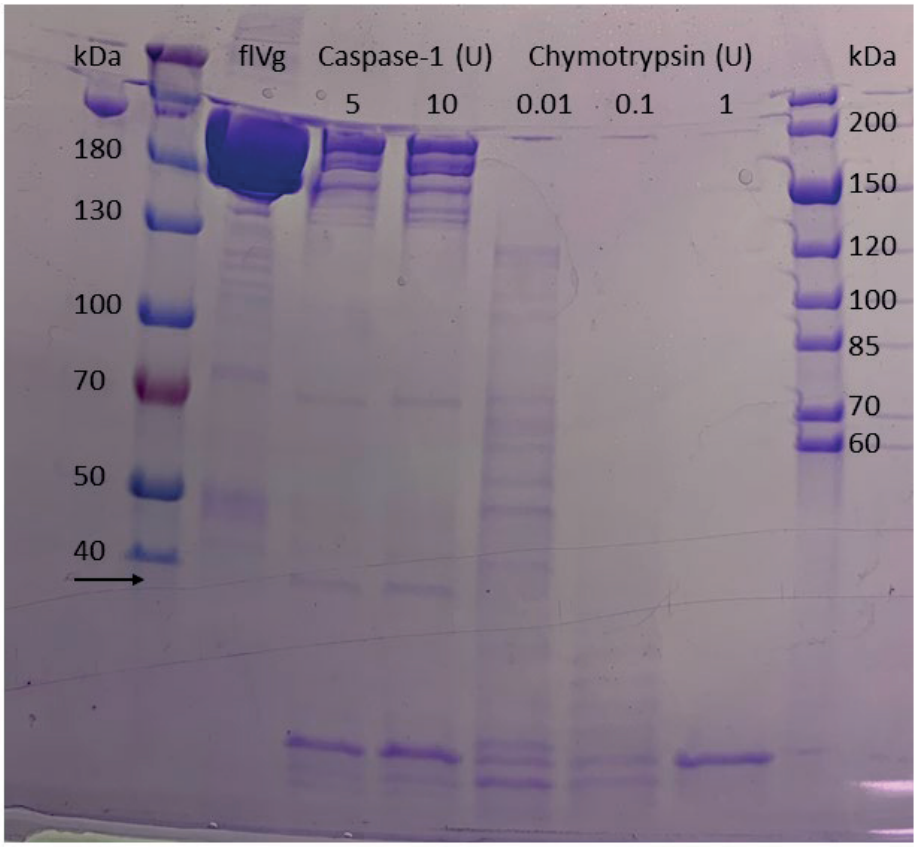


Figure S6


kDa
flVg
Caspase-1 (U)
5
10
Chymotrypsin (U)
0.01
0.1
1
kDa
180
130
100
70
50
40
200
150
120
100
85
70
60

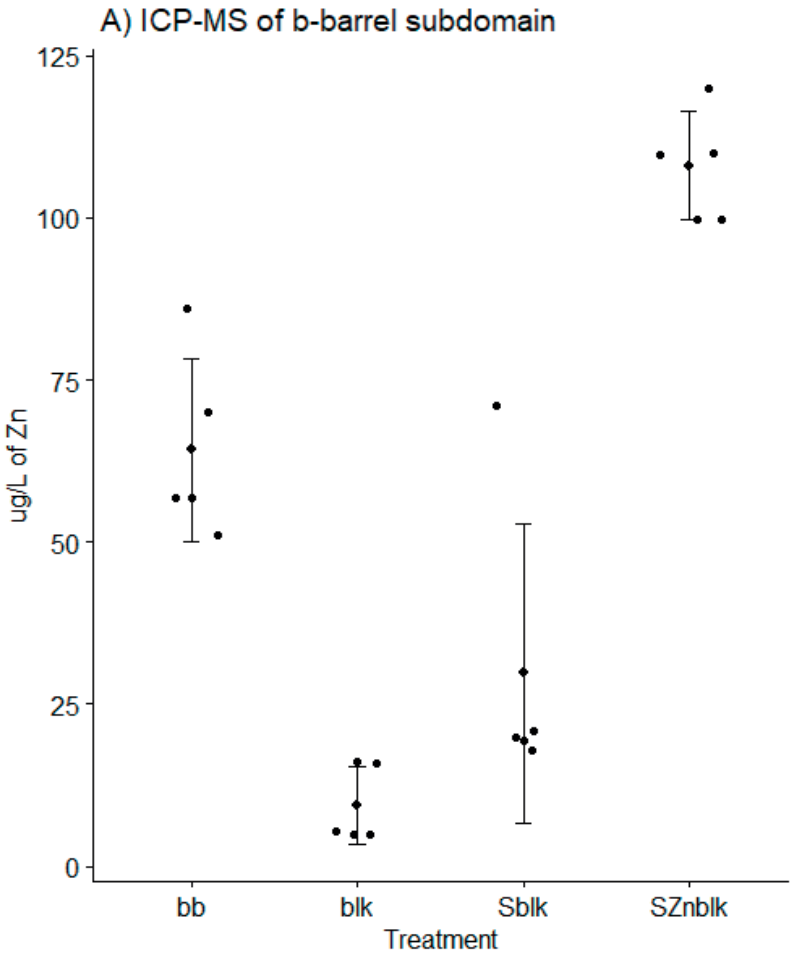
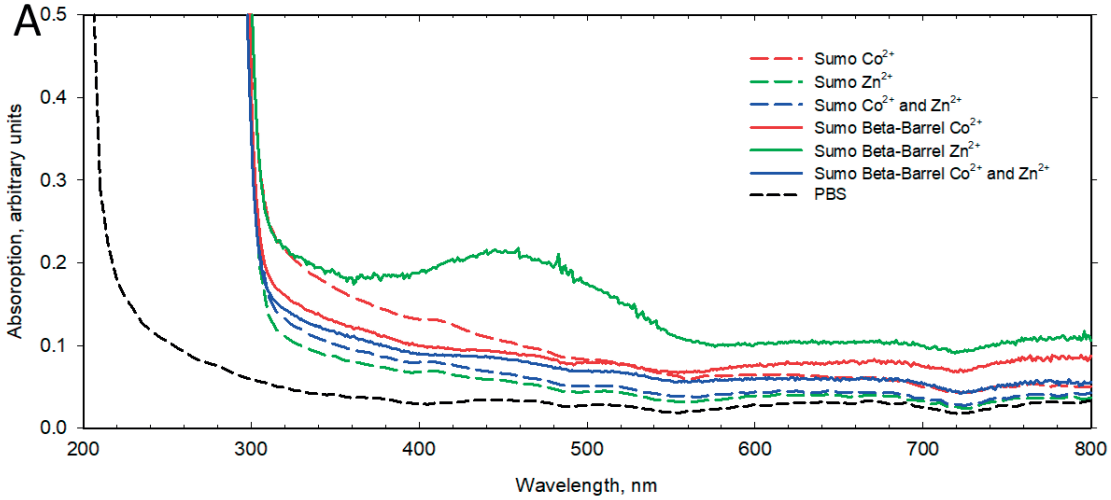
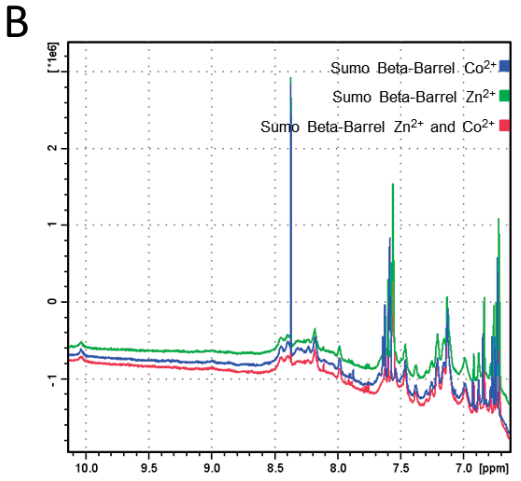
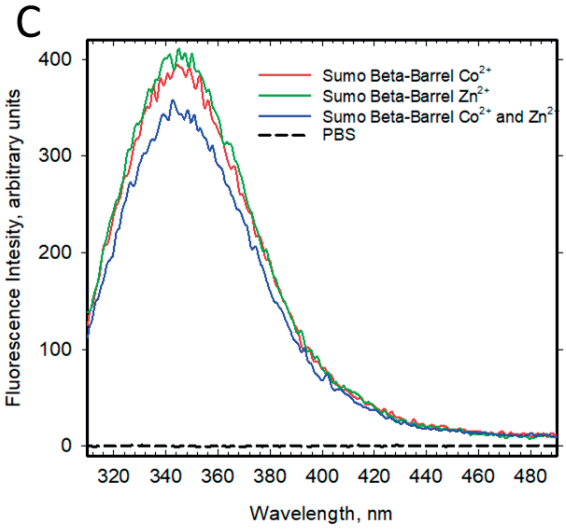

A) ICP-MS of b-barrel subdomain
125
100
75
50
25
0
ug/L of Zn
bb
blk
Sblk
SZnblk
Treatment

Figure S8


A
Absorption, arbitrary units
Wavelength, nm
Sumo Co2+
Sumo Zn2+
Sumo Co2+ and Zn2+
Sumo Beta-Barrel Co2+
Sumo Beta-Barrel Zn2+
Sumo Beta-Barrel Co2+ and Zn2+
PBS


B
[*1e6]
Sumo Beta-Barrel Co2+
Sumo Beta-Barrel Zn2+
Sumo Beta-Barrel Zn2+ and Co2+
[ppm]


C
Fluorescence Intesity, arbitrary units
Wavelength, nm
Sumo Beta-Barrel Co2+
Sumo Beta-Barrel Zn2+
Sumo Beta-Barrel Co2+ and Zn2+
PBS

Paper III

Identification of 121 variants of honey bee Vitellogenin protein sequences with structural differences at functional sites

Short title: Identification of 121 honey bee Vitellogenin variants

Vilde Leipart[1], Jane Ludvigsen[1,4], Matthew Kent[2], Simen Sandve[2], Thu-Hien To[2], Mariann Árnyasi[2], Claus D Kreibich[1], Bjørn Dahle[1], Gro V. Amdam[1,3]

1: Faculty of Environmental Sciences and Natural Resource Management, Norwegian University of Life Sciences, Ås, Norway

2: Centre for Integrative Genetics (CIGENE), Department of Animal and Aquacultural Sciences, Norwegian University of Life Sciences, Ås NO-1432, Norway

3: School of Life Sciences, Arizona State University, Tempe, AZ, United States

4: Fürst medisinsk laboratorium, Søren Bulls vei 25, 1051 Oslo

Corresponding Author: Vilde Leipart, vilde.leipart@nmbu.no

Author Contributions

Conceptualization: Vilde Leipart, Jane Ludvigsen, Gro V. Amdam

Data Curation: Vilde Leipart, Thu-Hien To

Formal Analysis: Vilde Leipart

Funding Acquisition: Gro V. Amdam

Investigation: Vilde Leipart, Jane Ludvigsen, Matthew Kent, Mariann Árnyasi, Thu-Hien To

Methodology: Vilde Leipart, Jane Ludvigsen, Matthew Kent, Simen Sandve, Mariann Árnyasi, Thu-Hien To, Claus D Kreibich, Gro V. Amdam

Project Administration: Vilde Leipart, Jane Ludvigsen, Gro V. Amdam

Resources: Gro V. Adam, Matthew Kent, Simen Sandve, Bjørn Dahle

Software: Thu-Hien To

Supervision: Jane Ludvigsen, Gro V. Amdam

Validation: Vilde Leipart, Jane Ludvigsen, Matthew Kent, Simen Sandve, Mariann Árnyasi, Thu-Hien To, Gro V. Amdam

Visualization: Vilde Leipart, Thu-Hien To

Writing – Original Draft Preparation: Vilde Leipart, Jane Ludvigsen

Writing – Review and Editing: All Authors

Abstract

Proteins are under selection to maintain central functions and to accommodate needs that arise in ever-changing environments. The positive selection and neutral drift that preserve functions result in a diversity of protein variants. The amount of diversity differs between proteins: multifunctional or disease-related proteins tend to have fewer variants than proteins involved in some aspects of immunity. Our work focuses on the extensively studied protein Vitellogenin (Vg), which in honey bees (*Apis mellifera*) is multifunctional and highly expressed and plays roles in immunity. Yet, almost nothing is known about the natural variation in the coding sequences of this protein or how amino acid-altering variants might impact structure–function relationships. Here, we map out allelic variation in honey bee Vg using biological samples from 15 countries. The successful barcoded amplicon Nanopore sequencing of 543 bees revealed 121 protein variants, indicating a high level of diversity in Vg. We find that the distribution of non-synonymous single nucleotide polymorphisms (nsSNPs) differs between protein regions with different functions; domains involved in DNA and protein–protein interactions contain fewer snSNPs than the protein's lipid binding cavities. We outline how the central functions of the protein can be maintained in different variants and how the variation pattern may inform about selection from pathogens and nutrition.

Abbreviations

The domain of unknown function 1943 (DUF1943), hotspot (H), large lipid transfer protein (LLTP), N-terminal domain (ND), non-synonymous single nucleotide polymorphisms (nsSNPs), pathogen-associated molecular patterns (PAMP), relative solvent accessible surface area (rASA), Vitellogenin (Vg), von Willebrand factor (vWF) domain

Introduction

Protein function relies on the protein's structural shape, which is dictated by the amino acid sequence that determines the biophysical properties of the molecule. Mutations resulting in non-synonymous single nucleotide polymorphisms (nsSNPs) alter the amino acid sequence and provide an opportunity for new protein variants to enter populations. New variants can be detrimental, neutral, or beneficial in terms of the protein's impact on phenotype, and these different selective contexts create specific patterns of diversity [1, 2]. For example, multifunctional proteins or proteins at high titers or expressed in several tissues tend to be under strong purifying selection pressure, which results in low diversity [3-5]. An increase in the number of protein-protein interactions is also negatively associated with diversity [6, 7], as is enzyme-function in essential metabolic pathways where changes in the proteins' active site are unlikely to be beneficial [8]. Conversely, proteins that accommodate diverse or rapidly evolving interaction partners [1, 9]; as exemplified by the histocompatibility complex [10, 11] that recognizes antigens and as observed for membrane- or surface-exposed proteins involved in host-specificity of bacteria [12]. More diversity is also seen in proteins with high designability (i.e., several amino acid configurations accommodate the same fold) [13]. Finally, specific structural features are associated with diversity patterns, such as when exposed structures show more diversity than buried structures [14, 15], or when flexible structures show more diversity than stable β-sheets or α-helices [16].

Vitellogenin (Vg) is a large glycolipo-protein broadly distributed phylogenetically and well known for its role in egg yolk formation. In several species of fish, Vg has immunological functions [17, 18], and in honey bees (*Apis mellifera*), the protein is further recognized for pleiotropic effects on complex behavior [19, 20]. Honey bees are important ecologically and economically as pollinators of native plants and cash crops, and they are key producers of honey, wax, and propolis worldwide [21]. In addition, they represent a flagship species in social insect research [22]. Largely due to these features, Vg has been more intensely studied in honey bees than in most other invertebrates [23]. The protein is found at high titers in hemolymph (insect blood) [24] and localizes to multiple honey bee tissues, including muscle, fat body (functionally analogous to liver and white adipose tissue), gut epithelial cells, and glial cells in the brain [25, 26]. Structurally, the protein has a subdomain of 18 amphipathic α-helices that, together with a β-barrel subdomain and a flexible polyserine linker, form a highly conserved N-terminal domain (ND) [27]. The ND is positioned around a large lipid binding site consisting of a domain of unknown function 1943 (DUF1943) and one β-sheet, followed by a von Willebrand factor (vWF) domain (Figure 1). The final C-terminal region comprises a small structure connected to the vWF domain through a presumed flexible linker [28].

Specifically, the ND likely represents the receptor-binding region of all Vg proteins [29-31]. The ND is also a surface-to-surface contact site in Vg homodimerization, as seen in lamprey (*Ichthyomyzon unicuspis*) [32, 33]. Dimerization at this site is supported in honey bees [28], although Vg appears to be monomeric under most conditions in this insect [34, 35]. Moreover, in honey bees, the β-barrel subdomain of the ND can be proteolytically cleaved at the polyserine linker [36]. The β-barrel appears to subsequently translocate to the nucleus and bind DNA (potentially with co-factors) to influence gene expression [37]. The honey bee ND has a cavity of unknown function in the cleft between the β-barrel and α-helical subdomain [28], while the positively charged α-helical subdomain can account for some of the proteins' binding to honey bee pathogens [38, 39]. Zooming out, the three structural elements of the large lipid binding cavity create a network of β-sheets with an extensive hydrophobic interior. The hydrophobic core of this site is crucial for the transport and storage of lipids [32], and its structural fold and polarity are conserved across the large lipid transfer protein (LLTP) superfamily to which the Vg proteins belong [40]. The DUF1943 and vWF domains are, in addition, important for innate and mucosal immunity in several species [17, 41, 42]. In contrast, no specific function has been assigned to the C-terminal region of Vg to date [28].

The multifunctionality of honey bee Vg, as well as its high expression, expression in many tissues, and the protein's interaction with a receptor and dynamics of dimerization may indicate that few Vg variants are found in the bee. The protein's functions in immunity, in contrast, can suggest that many variants are found. Some support for the latter is provided by previous research [20, 43]. Motivated by these questions, our study seeks a deeper understanding of patterns of variation in honey bee Vg. We examine sequence variation from 15 countries, identify domains under different selective pressures, and characterize the putative functional impact of amino acid changing variants. We reveal 121 unique Vg variants, including 81 nsSNPs that are non-uniformly distributed across the domains and subdomains of the protein. Our analysis illustrates how the structural elements of honey bee Vg experience differing degree of selection pressures.

Results

Identification of Vg variants, frequency, and distribution of nsSNPs

Successful amplicon sequencing and variant-calling from 543 individual worker honey bees (diploid females) generated 1,086 full-length *vg* allele sequences, corresponding to 340 unique haplotypes (see Figure S1 for an overview of workflow and Materials and Methods for further details). These haplotypes include different combinations of 81 nsSNPs (see Table S1 for information on the nsSNPs' properties) resulting in 121 protein variants of honey bee Vg (Table S2; see Figure S2 for an overview of the geographical location of these variants).

In all domains and subdomains of the Vg gene, nsSNPs were identified, with a mean total number of nsSNPs per Vg variant of 5.56 (SD = 1.76). Some nsSNPs occurred more frequently than others: specifically, 15 of the 81 nsSNPs were identified in ≥5 % of the Vg variants. Except for one (p.Arg1292Ser) (6 %), these common nsSNPs caused subtle changes in residue type (Figure 2A). Variants with only common nsSNPs carry the same (one) change in the α-helical subdomain of the ND, and the β-barrel subdomain of the ND and in the C-terminal region typically has few nsSNPs (see Figure 2C for examples). The specific number of nsSNPs and their combinations vary more for the lipid binding site, which thus becomes unique for each Vg variant. In contrast to the common nsSNPs, 21 of 42 (50 %) of the nsSNPs observed only once (i.e., rare nsSNPs) conferred major changes in amino acid characteristics (Figure 2B). For a look at rare nsSNPs, we present a set of Vg variants that includes several rare changes (Figure 2D). In these examples, as seen with the rare Vg nsSNPs overall, we find some in the α-helical subdomain (see variant nr. 34, Figure 2D), and only one change in the β-barrel subdomain, in contrast to several changes in the lipid binding site, including the vWF domain.

Taken together, the distribution of the rare nsSNPs across protein domains mirrors that of the common nsSNPs. The α-helical subdomain of the ND tends to carry similar changes between variants. In contrast, the lipid binding sites, including the vWF domain, tend to carry more diverse sets of nsSNPs between variants (Figure 3).

To examine the distribution of 81 nsSNPs in the domains and subdomains, we calculated the frequency of nsSNPs per domain and subdomain site (aa; Figure 3A) and found nsSNP frequency to be lower in the β-barrel subdomain than in the remainder of the domains and subdomains. We subsequently separated the Vg variants into domains and subdomains and counted the number of unique combinations of nsSNPs. This number is higher for the lipid binding site than for the remainder of the domains and subdomains (Figure 3B). The number of amino acids comprising each domain and subdomain varies, which results in a different number of available sites for substitutions at the domains and subdomains. To calculate a ratio to control for this difference, we divided the number of unique Vg variants by the number of sites (aa) per subdomain and domain (Figure 3C). This represents a ratio of unique Vg variant per subdomain and domain. The ratio is higher for the lipid binding site and vWF domain than the remainder of the domains and subdomains (Figure 3C).

We classified the nsSNPs identified in the domains and subdomains into three categories. First, we used the common and rare categories described above and included the remaining nsSNPs (other). Then, we considered if the changes were modest or drastic and calculated whether the substituted residues were at buried or exposed sites in the protein structure. Figure 3D shows the resulting plot

for each subdomain and domain. The plot reveals considerable differences between the structural elements of Vg. We assessed whether this variability in distribution and classification of nsSNPs justified a domain- or subdomain-specific approach in the next-step analyses.

Implications of β-barrel subdomain variants

Only 7 of the 81 nsSNPs were identified in the β-barrel subdomain (Figure 3D), which is less than for other domains (Figure 3A). Except for p.Gly146Ser, all of the nsSNPs cluster at one side of the structure (Figure 4A). Gly146 is buried in the subdomain, close to a set of predicted Zn^{2+}-coordinating residues and a proposed DNA binding region (Leipart *et al*. in manuscript)[37]. The remaining nsSNPs increase the polarity of buried residues or increase the hydrophobicity at the surface, except for p.Ile132Met, which maintains the hydrophobic core (Figures 4B and 4C). Overall, the 121 Vg variants identified in this study either contain none or one nsSNP in the β-barrel subdomain, except for Vg variant nr. 5, which carries two common nsSNPs (Figure 4C).

Implications of α-helical subdomain variants

We identified 17 nsSNPs in the α-helical subdomain (Figure 3D). By mapping the nsSNPs onto the structure, we identified three hotspots of amino acid substitutions (H1, H2, H3; see Figure 5A). The same classification outlined in Figure 3D was repeated here for the nsSNPs in the identified hotspots (Figure 5A). The only common nsSNP (p.Ile489Val, green in the plots in Figures 3D and 5A) is a modest substitution identified in H2. All hotspots contain rare nsSNPs. Out of the 121 Vg variants identified here, 119 variants include one nsSNP in H2, and 15 variants have at least one nsSNP in H1 and/or H3, as shown for Vg variant nr. 24, 41, and 45 (Figure 5B).

Looking at the H1 in more detail, we find that it represents moderate substitutions at three buried and two exposed residues (Figure 5A). The polarity is maintained by these nsSNPs, except for the rare p.Thr594Met, which decreases the polarity of the buried region of the hotspot (see variant nr. 41, Figure 5B). H2 encompasses residues frequently substituted in the short loop regions connecting the α-helices, close to the lipid binding site. These substitutions are modest, except the exposed p.Thr522Ile (see variant nr. 24, Figure 5B). The effects of the nsSNPs on the polarity and electrostatic potential of the structure vary as hydrophobic and hydrophilic residues are introduced. One nsSNP provides a positive charge (p.Asn560His), while another nsSNP removes a negative charge (p.Asp626Asn). The same variability for electrostatic potential is seen in H3, which is buried between two of the subdomain α-helices and the first β-sheet of DUF1943: one nsSNP maintains a negative charge (p.Asp608Glu), while another flips the charge from negative to positive (p.Glu642Lys; see variant nr. 24 and 45, Figure 5B). The remaining three nsSNPs in H3 maintain hydrophobicity at their specific sites.

Implications of lipid binding site variants

We identified 37 nsSNPs at the lipid binding site (Figure 3D). The nsSNPs were found in 56 combinations (Table S2 and Figure 3B) without discernable clustering into hotspots. The 56 combinations represent a high ratio relative to domain size (aa; Figure 3C). Only three out of the 121 Vg variants lack nsSNPs in the lipid binding site, confirming that it represents a highly diverse protein region. Underlining this level of diversity is the identification of 10 different nsSNPs in just two Vg variants (see variant nr. 1 and 49, Figure 6B).

Specifically, drastic substitutions at the lipid binding site were identified at exposed residues, altering the polarity and electrostatic charge of the surface (Figure 3D). This dynamicity of surface residues is a common finding [16], as are moderate substitutions at buried residues [15]. We observed that moderate substitutions do not appear to alter the hydrophobic core or the two charged centers of the Vg lipid binding cavity (Figures 6C and 6D). In addition, however, we find three rare and drastic substitutions at buried residues. Two of these nsSNPs increase the hydrophobicity at the end of the long β-sheet spanning the ND (Figures 6A and 6B), while the third nsSNP increases the polarity of a buried loop, folded away from the domain core.

Implications of vWF domain variants

We found 14 nsSNPs in the domain (Figure 3D). They were identified in 17 unique combinations, which represents a high ratio relative to domain size (aa; Figure 3C). Overall, the changes are diverse and distributed without discernable hotspots, as we observed for the lipid binding site that interfaces with the vWF domain (Figure 6A).

Interestingly, the vWF domain shows a total of 5 drastic (but rare) substitutions at buried residues. This is the highest number of drastic, buried nsSNPs, compared to the other Vg domains or subdomains (Figure 3D). Three of these nsSNPs either maintain or introduce a polarity, while the other two increase hydrophobicity. Among the 14 nsSNPs in the vWF domain, Ser1587 is the only substituted residue directly exposed to the lipid cavity. This nsSNP introduces a large aromatic residue to the cavity (see variant nr. 103, Figure 2D). Additionally, we find three common nsSNPs that maintain hydrophobicity at buried or exposed sites. These three occur together in Vg variant nr. 40 (Figure 6E). The remaining 5 nsSNPs are modest substitutions. Three are buried and maintain hydrophobicity, while two are exposed and maintain polarity.

Implications of C-terminal variants

We identified 6 nsSNPs in the C-terminal of Vg (Figure 3D). Four out of the 6 nsSNPs in the exposed structure introduce a serine residue (Figure 7A). These are positioned at the presumed flexible linker or an exposed loop extending from the folded structure, which increases the polarity of the C-terminal.

Two serine-introducing nsSNPs occur together in Vg variant nr. 11 (Figure 7B). The two remaining nsSNPs, not introducing serine, are rare and drastic substitutions (Figure 3D), one increasing the hydrophobicity of the buried structural elements, and the other introducing a large aromatic residue close to a predicted Zn^{2+}-binding site (Leipart *et al.* in manuscript). The positive surface charge of the C-terminal is not altered by any of the 6 nsSNPs (Figure 7C).

Implications of nsSNPs at three domain or subdomain interfaces

Viewing the patterns of nsSNPs in the light of domain or subdomain interfaces, we find that the most variable region of the β-barrel subdomain is adjacent to H1 on the α-helical subdomain. Together, these structures create a hydrophobic and slightly negatively charged cavity (Figures 8A and 8B). A positively charged β-sheet in the DUF1943 domain extends into the cavity, forming an intriguing subdomain interface (Figure 8B). The interface carries 10 nsSNPs: seven introduce a methionine, while the remaining three introduce a tyrosine, leucine, or alanine. One nsSNP decreases the positive charge (p.His412Tyr), while the remaining changes do not influence negative charges of buried or exposed residues, and hydrophobic characteristics are maintained. The conservative nature of these variations is in part explained by the 10 nsSNPs being mostly rare (Figure 8C, for classification of the nsSNPs) and thus unlikely to occur together on one Vg variant (seen in 26 of 121 variants).

Moving on, we find that H2 localizes to an opening where the α-helical subdomain of the ND interfaces with the lipid binding site (Figure 8D). The subdomain interface has a positive charge close to H2, while the edge of the opening (i.e., at the lipid binding site) is hydrophobic (Figure 8D, for classification of the nsSNPs). The nsSNPs in this region are rare and modest substitutions, except for the common p.Ile489Val and the drastic p.Thr522Ile in H2, which maintain and increase hydrophobicity, respectively (Figures 8C and 8D). Two other nsSNPs (p.Asn560His and p.Glu906Lys) slightly increase the positively charged surface, while the hydrophobic region remains undisturbed. The majority of Vg variants identified in this study have only one nsSNP at this subdomain–domain interface (seen in 119 out of 121 variants, including the common p.Ile489Val; excluding this, it is seen in 13 of 121 variants).

Next, we observe that the vWF domain is adjacent to an additional opening into the lipid binding site. At this interface, we find 13 nsSNPs (Figure 8E) that do not appear to introduce a consistent type of change. The domain interface is mainly hydrophilic, which is maintained by two common nsSNPs (p.Ser803Asn and p.Arg1174Lys). Other nsSNPs introduce polar and hydrophobic residues: a positive and a negative charge are lost at two different positions (p.Lys1171Asn and p.Asp1491Asn), mirrored by the introduction of a positive and a negative charge at two other positions (p.Gly1016Asp and p.Thr1567Lys). Both buried and exposed residues are modestly or drastically substituted, but these nsSNPs are generally rare (Figure 8C, for classification of the nsSNPs). Adding to the region's diversity,

there are two aa positions with alternative substitutions (p.Gly1016Asp/Thr and p.Thr1567Met/Tyr, see Figure 8E). As observed for the previous domain interface, the Vg variants tend to carry only one nsSNP at the vWF-lipid binding site interface (seen in 36 of 121 variants).

Implications for the full-length protein structure

When mapping all of the nsSNPs on the surface of the full-length structure of Vg (colored red in Figure 9), we find that the three domain or subdomain interfaces (described above) are located on the same surface side, referred to here as side A (Figure 9). Interestingly, all but one surface-exposed nsSNPs in the ND are located either around the ND cavity where the β-barrel subdomain is interfacing with H1 in the α-helical subdomain or where H2 in the α-helical subdomain interfaces with the DUF1943 domain around an opening to the lipid binding cavity. The one exception is a nsSNP from H3 in the α-helical domain (Figure 9, side A). For the lipid binding site, including the vWF domain, the exposed nsSNPs on side A are found concentrated around a small opening into the lipid cavity, except for two exposed nsSNPs on the vWF domain (Figure 9). Moreover, we find no surface-exposed nsSNPs in the ND when we rotate Vg 180° about the y-axis (as seen in Figure 9, side B). On side B, the exposed nsSNPs are distributed in the lipid binding site, including vWF, making no specific pattern on the surface and not seeming to cluster around the wide opening into the lipid cavity (shaded area in Figure 9, side B). Taken together, our findings demonstrate that honey bee Vg has surface-exposed nsSNPs in every domain and subdomain on side A, while the exposed nsSNPs on side B are only located in the lipid binding site, including vWF.

Discussion

This study presents new information on the diversity pattern of Vg. Our geographically broad sampling strategy resulted in over 100 full length Vg protein sequence variants, which is the largest collection of Vg protein variants in any species. Our data confirm the conserved nature of the ND: no changes were observed at aa positions in functional sites for DNA interaction [37] in the β-barrel subdomain, nor at positions suggested for homodimerization at either ND subdomain [28, 33] (see Figures 4A and 5A). The oligomerization state in native honey bee Vg is uncertain [28], but a requirement to protect surface properties involved in homodimerization is supported by our data. Simulating a homodimerization event exposes side A, while ND becomes inaccessible on side B (Figure S3). We find additional support for the conservation of the β-barrel subdomain, since none of the nsSNPs appear to introduce instabilities to the β-barrel fold (Figures 3D and 4A). Similarly, we find evidence supporting studies on varying selection pressures on honey bee Vg. These studies pinpoint the lipid binding site as the primary region of diversity [20, 43], as do our data (see, e.g., Figures 3B and 3D). Yet, in addition to these expected findings, our data reveal information that, combined with the first full-length protein structure for honey bee Vg, contributes to a new understanding of the diversity pattern.

New insights involving the ND

Given the conserved nature of the ND, our finding of 7 nsSNPs in this region might come as a surprise. As shown, 6 of the 7 nsSNPs cluster at the interface to the α-helical subdomain, adjacent to H1 (Figure 8A). We found that these nsSNPs are rare and tend to introduce hydrophobic residues, particularly methionine. These observations support the idea that selection acts to maintain the characteristics of this structure. Specifically, the region of the 6 clustering nsSNPs is part of a cavity [28], and the conservation of hydrophobic residues is typical for a binding site [44, 45].

The functionality of binding cavities is defined by the residue types, shapes, and locations in the protein [46]. At the β-barrel/α-helical subdomain interface, the β-barrel residues create a hydrophobic and slightly negatively charged region, which meets a positive interior. This structure resembles the large lipid cavity further downstream in the aa sequence. However, the overall shape of ND differs, since the cavity is closer to the protein surface and smaller. We interpret this difference to indicate that the two cavities of honey bee Vg are not functionally equivalent. A distant homolog found in lamprey supports this interpretation, since no phospholipids were observed at the location of the ND cavity [32]. The more conserved nature of the ND cavity compared to the Vg lipid binding site lends further support (Figures 3A–D): the more conserved ND cavity could have a consistent binding partner,

while the lipid binding site might interact with various groups of lipids. We suggest that the compatible binding partner of the ND cavity is the Vg receptor. In support of this suggestion, it is assumed that the ND provides the receptor-binding site of the Vg proteins [29-31], and we observe that all but one of the nsSNPs (p.His412Tyr, seen in three Vg variants) introduce no or little change in the electrostatic potential of the ND cavity. Such electrostatic potential is generally important for receptor binding [30, 47]. It has previously been demonstrated that β-sheets in the β-barrel subdomain, as well as α-helices in the α-helical subdomain, have affinity and/or enhanced affinity to the Vg receptor [29-31]. Still, no specific residues in the ND had been specified to participate in this interaction before our work.

The second subdomain in the ND, the α-helical subdomain, has an immune-related function in honey bees that involves the transport of immune elicitors (fragments of bacterial cell wall, i.e., lipopolysaccharides or peptidoglycans) [39] and the recognition of pathogen-associated molecular patterns (PAMP) [38]. The PAMP recognition by Vg is demonstrated for several species of fish [48-50]. High levels of diversity are found in at least some proteins involved in immune defense mechanisms, such as pattern recognition receptors that bind to bacteria via PAMP. In these receptors, the recognition domain is characterized by a leucine-rich repeat that carries nsSNPs, modulating the ability to identify various pathogens [51-53]. Based on this mechanism and the *in vitro* detection of PAMP binding by the α-helical subdomain of honey bee Vg [38], we expected to find a level of diversity in one or more regions of the subdomain. Indeed, we find three nsSNP hotspots: the first (H1) is part of the ND cavity discussed above. The second and third interface with the lipid binding site (H2) or are buried in the subdomain (H3), respectively (see Figures 5A and 8D). In assessing their potential for binding PAMP, we find a high level of diversity at exposed residues in H2. This diversity represents substitutions with a lack of consistency for the introduced residue types that is similarly observed in leucine-rich repeats of protein recognition receptors [51-53]. The buried nature of H3 makes it a less attractive candidate for a direct role in PAMP binding. Instead, nsSNPs could influence subdomain stability and functionality [54-59]. Thus, we speculate that H2 has the potential to be involved in binding specificity with PAMP, while H3 has the potential for being indirectly involved by influencing subdomain functionality for recognition.

Currently, no specific description exists of a molecular mechanism of pathogen binding by the α-helical subdomain of honey bee Vg. Yet, we find 34 positively charged exposed residues (arginine and lysine) that could have an affinity to negatively charged pathogen membrane surfaces [38]. Similar positive surface charge can create high host affinity (but low specificity) for pathogen recognition [60].

Interestingly, the 34 positively charged residues in the α-helical subdomain are conserved in all of the 121 Vg variants identified by our study. We identify no nsSNPs on any exposed 34 arginine or lysine

residues (Figure S4). Taken together, the combination of a variable hotspot possibly involved in binding specificity (i.e., H2) and a conserved surface area involved in pathogen affinity (i.e., the 34 positively charged residues) could help provide a molecular understanding of how the α-helical subdomain of Vg contributes to honey bee immunity. At the same time, this insight helps explain why the α-subdomain, overall, may have lower diversity than expected for immune-related activity.

New insights involving the lipid binding site and vWF domain

The lipid binding site interfaces with H2 and the vWF domain. At both interfaces, we identify nsSNPs that introduce a positively charged residue and a high diversity. Also, when folded, the surface-exposed domain interfaces between the lipid binding site and vWF domain are near the α-helical subdomain (Figure 9A). This structural constellation could imply that the pathogen recognition region of Vg expands beyond the α-helical domain – a proposition supported by previous observation: full-length honey bee Vg binds PAMP better than the α-subdomain alone [38]. Several members of the LLTP family have similar recognition potential through the α-helical subdomain [40] but have additional protective roles as lipid presenting proteins. For example, the microsomal triglyceride transfer protein (MTP) has an important role in loading endogenous and exogenous lipids onto antigen-presenting cells in the human immune system [61, 62]. Similarly, apolipophorin III and I/II in insects can recognize pathogens [63, 64]. Studies of apolipophorin III show that additional immunological function, such as the ability to regulate and activate hemocytes (immune cells) or stimulate cellular encapsulation, is gained in a lipid-associated state [64]. This conditional functionality is explained by a conformational change when the protein binds lipids [65]. We speculate that the recognition surface presented by honey bee Vg could increase in response to lipid binding; thereby, maintaining the stability of the lipid binding cavity is important for immunological function.

Pathogen membrane surfaces are large relative to a protein [66, 67], so presenting several regions on the protein for affinity and/or specificity is certainly feasible. In this context, we note that the vWF domain of Vg can recognize pathogens in coral (*Euphyllia ancora*) [41] and zebrafish (*Danio rerio*) [17]. Interestingly, we identify a high level of diversity in the honey bee vWF domain (Figure 3C), yet these substitutions mostly occur at buried residues (Figure 3D). The vWF domain is predicted to be an important β-sheet structural region in the lipid binding cavity [28] (Figure 1), and the β-sheet structure is central to the stability of this cavity [40, 57]. Substitutions at buried regions, like those seen for the vWF domain, can affect stability and consequently regulate the size of the lipid load in Vg.

Interestingly, we find exposed residues undergoing changes inside the lipid binding cavity. The lipid cavity interior of Vg is not hypothesized to partake in immune-related activities directly. Instead, the region is recognized for a role in the transport and storage of nutritional phospholipids. Studies of

proteins in the LLTP superfamily show that maintaining the large hydrophobic core of the cavity facilitates a high affinity but low specificity for lipid molecules [32, 68]. Our data confirm that the hydrophobicity is conserved in honey bee Vg and suggest that the exposed nsSNPs inside the lipid cavity might influence lipid specificity. Phospholipids usually occupy the positively charged center, as shown in the lipid cavity for a distant homolog [32]. We find diversity at regions close to this charged center, suggesting that phospholipids might enter the cavity here (side A, Figure 9A). These diverse regions might also influence specificity for lipid molecules as well as pathogen specificity, as discussed above. Thus, overall, an evolutionary arms race with changing pathogens that further vary at different geographies could be a possible explanation for the pattern we observe, as suggested in previous research [69-71].

New insights involving the C-terminal region

We confirm the C-terminal region on honey bee Vg to be soluble and find 4 nsSNPs introducing polar residues (Figure 7A, seen in 35 Vg variants). This finding supports our previous study showing the region is exposed and connected to a presumed flexible linker [28]. We additionally provide new evidence showing a conserved positively charged surface (Figure 7C). A positively charged C-terminal region in other proteins has been linked to signaling for recruitment and translocation [72], protein assembly [73], and sensing changes in the extracellular environment [74]. Honey bee Vg has been demonstrated to sense oxidative stress [75] and suggested protecting honey bees from reactive oxidative species. Our earlier study shows that two disulfide bridges are conserved in the C-terminal region, which is proposed to coordinate Zn^{2+} (Leipart et al. 2021 *in manuscript*) (Figure 7A). Proteins with a positive surface charge and disulfide bridges on neighboring residues, sometimes including Zn^{2+}, are shown to protect against oxidative stress [76, 77]. Our findings support a conserved polarity and positive charged region; thus, we speculate that the C-terminal has a similar functional role.

Concluding remarks

None of the nsSNPs identified here are detrimental for honey bee Vg. The structural fold in the ND is highly conserved, and the drastic changes in the remaining domains are either exposed at the surface or buried at non-structural loop regions, except for the p.Thr939Met shown in Figure 6A. These nsSNPs increase the hydrophobicity at the protein core, which is unlikely to reduce structural stability. All of these observations are expected for a protein that is essential for fitness in its yolk-precursor role. At the same time, we observe new variability patterns that are likely associated with aspects of lipid

binding. In assessing these nsSNPs, we provide new insights on the possible interface between Vg, its lipid cargo, and honey bee pathogens. We believe these suggestive findings are thought-provoking and warrant further study. Additionally, it is worth mentioning that the long-read sequencing technology used here creates an opportunity to identify and characterize genomic structural variants that are difficult or impossible to detect with alternative approaches [78, 79]. Such variants can significantly impact protein structure and should be receiving increasing attention in studies seeking to link genotype to phenotypic variation. Correspondingly, a preliminary examination of our data suggests the presence of larger structural variants (deletions) that will be fully explored in a future manuscript.

Materials and Methods

Bee sampling

452 samples of *Apis mellifera* were collected from Europe. Nine protected *Apis mellifera mellifera* apiaries were selected and sampled based on earlier introgression studies [69, 71]: Norway (Flekkefjord, N=30; Rena, N=32), Sweden (Jämtland, N=30), Denmark (Læsø, N=32), Scotland (Isle of Colonsay, N=30), Ireland (Connemara, N=30), Poland (Augustów Primeval Forest, N=30), the Netherlands (Texel, N=30), and France (Les Belleville, N=30). Samples from six European subspecies, from separate apiaries, were chosen for comparison: Slovenia (*A. m. carnica*, N=25), Italy (*A. m. ligustica*, N=30), Portugal (*A. m. iberiensis*, N=30), Macedonia (*A. m. macedonica*, N=33), Malta (*A. m. ruttneri*, N=30), and Turkey (*A. m. anatolica*, N=30). The samples from Europe were provided by researchers and managers of breeding associations working with each subspecies to ensure that samples were obtained from purebred populations. In addition, we collected 186 samples from the USA, used as one control group, from 6 different apiaries covering the north, west, south, northeast, east, and central regions: Minnesota (N=33), California (N=30), Arizona (N=30), Maryland (N=30), North-Carolina (N=33), and Illinois (N=30), respectively. To ensure genetic variation among the samples, the collectors in Europe and the USA sampled 25 to 33 bees from three to six separate hives in their apiaries. The specimens were collected and shipped in 2 ml Eppendorf tubes filled with 1.9 ml 96 % ethanol and stored at −20 °C.

gDNA extraction

Genomic DNA (gDNA) was extracted from the thorax of each bee. The head, wings, legs, and abdomen were removed, before the thorax was washed in PBS for 5 minutes. The equipment used for dissection was washed in 10 % chlorine and 96 % ethanol between every bee. After washing, the thorax was cut in half vertically and weighed, with weights ranging from 18 to 30 mg. Half of each thorax was used in the DNA extraction protocol. The thorax piece was placed in a tube filled with 200 µl ATL buffer (1:2 ratio) and three sterile ceramic beads (2.8 mm). The samples were ground in Retsch® mixer mill MM 400 (Retsch GmbH, Germany) at 15/s for 20 seconds, before 20 µl Proteinase K and 2 µl Rnase A were added and mixed by vortexing, and the samples were incubated at 56 °C overnight while mixing. The remaining steps followed the QIAGEN® DNeasy® Blood & Tissue Kit standard protocol (QIAGEN, Redwood City, CA). The eluate was eluted twice with a final volume of 100 µl. The concentration was measured on Qubit® 2.0 Fluorometer using the Qubit™ dsDNA HS Assay kit standard protocol (ThermoFisher Scientific, Waltham, MA). The extracted gDNA was run on 0.4 % TAE Agarose gels containing TAE buffer containing StainIN™ GREEN Nucleic Acid Stain (highQu, Germany), at 40V for 1h and 50 min, with the Thermo Scientific™ GeneRuler™ High Range DNA ladder to determine the size and quality of gDNA. Eluted gDNA was stored at −20 °C for 1–2 days, then at −80 °C.

PCR, pooling, and clean-up

To enable the simultaneous sequencing of amplicons from 543 bee samples, a two-tier barcoding strategy was used, whereby barcodes were included in both the PCR primers and the sequencing adapters. PCR primers were developed to amplify the full-length *vg* gene (including introns) from position 5,029,433 to 5,035,683 in NC_037641.1 [80] (see Table S3 for the primer sequences). In addition to the *vg*-specific sequence, unique barcodes from the PCR Barcoding Expansion 1-96 kit (EXP-PBC096; Oxford Nanopore Technologies, see Table S3 for barcode sequences) were incorporated into the 5' ends of the forward (n=8) and reverse (n=12) primers, which enabled 96 different barcode combinations. PCR was performed in 96-well plates, wherein each PCR reaction contained 10 ng gDNA, a unique combination of forward and reverse primers (0.5 uM each), 0.5U Q5® High-Fidelity DNA Polymerase (New England BioLabs, MA, USA), 1X Q5 Reaction Buffer, 200 µM dNTP, and Nuclease-free water, to a final volume of 25 µl. Cycling conditions were as follows: 98 °C for 1 min, 30 cycles of 98 °C for 10 s, 58 °C for 30 s, 72 °C for 5 min, and then 72 °C for 7 min and a hold at 4 °C. One positive control sample and one negative control (PCR water) were included for each of the 6 PCR plates that were run. After PCR, the concentration of each amplicon was measured in a plate reader using PicoGreen (ThermoFisher Scientific, Waltham, MA). The positive and negative controls were checked on a 1 % TAE agarose gel to verify amplification and the lack of contamination (see Figure S5 for agarose gel). From each of the 94 samples within each plate, 16 ng was pooled, creating six plate pools (see Table S3 for a plate set up used for each pool). The six plate pools had concentrations ranging from 5.4 to 10.8 ng/µl (Qubit® 2.0, dsDNA BR Assay) and volumes ranging from 392.4 to 731.4 µl. Each pool was concentrated and purified using 0.75X AMPure XP beads (Beckman Coulter, Brea, CA) before being eluted in 60 µl nuclease-free water pre-heated to 50 °C. The concentration of each pool was measured (Qubit® 2.0, dsDNA BR Assay) and found to range from 10.9 to 22.8 ng/µl. Three of the pools with concentrations lower than 15 ng/µl were up-concentrated using a vacuum centrifuge to be able to start with a minimum input of 620 ng amplicons from each pool.

Library preparation and Nanopore sequencing

For Nanopore sequencing, the library was prepared using the Ligation Sequencing kit 1D (SQK-LSK109) and the Native Barcode Expansion kit (EXP-NBD104), following the "Native barcoding amplicons" Nanopore protocol. The workflow is illustrated in Figure S1A. Briefly, 620 to 850 ng amplicons from each plate pool were used as input to prepare the DNA ends for barcode attachments; native barcodes NB01-NB06 were then ligated to the end-prepared amplicons. After measuring the concentration of the six native barcoded sample plate pools, equal amounts from each pool were combined, and a total of 800 ng mix was taken to adapter ligation. After flow cell priming, 200 ng (equal to 50fmol) final prepared library was loaded into a PromethION flow cell (v9.4.1). MinKNOW v20.06.18 was used for

operating sequencing. Base-calling and filtering were performed with Guppy v4.0.11 using the "High-accuracy sequencing" base called model, and the minimum qscore for read filtering was 7. Oxford Nanopore Technologies sequence data were base called real-time using the MinKNOW Fast base calling model from Fast5 into FastQ file format. Raw reads were classed as passed by MinKNOW based on the average read quality score >7.

Bioinformatic pipeline

The bioinformatic pipeline is illustrated in Figure S1B. About 18 million raw reads were downloaded from the PromethION sever and demultiplexed each native and inner barcodes into separate samples using cutadapt v. >=2.10 [81]. The error rate for the inner barcodes was set to 0.17, and the minimum and maximum length of reads after trimming the inner barcodes was set to 6,000 and 7,000, respectively, reducing the number of raw reads to 6,193,310. Each read was written into a separate folder, and the native and inner barcodes and primer sequences were removed from the reads. The medaka tool (v. 1.0.3 https://nanoporetech.github.io/medaka/index.html, source code, and analysis scripts (available at https://github.com/nanoporetech/medaka) were used to create consensus sequences and variant calling. A consensus sequence for each demultiplexed sample was generated using medaka_consensus based on reference sequence NC_037641.1 [80]. To create haplotype consensus sequences, the phased alignments of the medaka_variant pipeline were first applied and separated the reads into haplotypes for each sample. The medaka_consensus was then re-used, with the same reference sequence as above, to generate a consensus sequence for each haplotype. The variant calling pipeline of medaka was also used for SNP calling for each haplotype using the same reference sequence. The pipeline was implemented using snakemake v.>=5.6.0 (available at https://gitlab.com/cigene/computational/bee_amplicon). We illustrate the pipeline in Figure S1B. The downstream analysis was done on the allele sequences generated from a minimum of 100 raw reads (31 samples had fewer than 100 reads and were not included in the downstream protocol). This resulted in 1,086 allele sequences, generated from an average of 6,497.34 (SD=5,328.55) raw reads per allele sequence.

Identifying Vitellogenin variants

The raw allele sequences were uploaded to Geneious Prime v.2019.0.03, where we created FASTA files starting at first to the last codon for the *vg* gene (6109 bp, including introns, NP_001011578.1). DNA Sequence Polymorphism v.6.12.03 [82] was used to identify 340 haplotypes and the 81 nsSNPs (See table S1 for an overview of the nsSNPs properties). The nsSNPs are written using the Human Genome Variation Society [83]. Haplotypes with identical nsSNPs combinations were identified as identical Vg variants. The Vg variants are presented in Table S2. The AlphaFold prediction of full-length

honey bee Vg was generated from UniProt ID Q868N5, and we used this sequence as a reference for nsSNP analysis.

Structural analysis

The structural analysis was performed in PyMol v.2.4.1 [84] using AlphaFold Vg structure [28]. We considered nsSNPs identified in more than 5 Vg variants as common and identified only one as rare. Other nsSNPs identified in 5 to 2 Vg variants were also considered and classified as "other." The relative solvent accessible surface area (rASA) was calculated in PyMol, and residues scoring <20 % were deemed buried [85]; otherwise, they were classified as exposed, although thresholds from 5–25% have been used in literature. The rASA calculation indicates how exposed the residue is at the specific position in the protein structure [86]. The similarity between amino acids was classified for each substitution using a substitution matrix [87]. A negative score indicates that the physiochemical properties are not preserved. Negative scores in the BLOSUM62 matrix were considered drastic; otherwise, they were considered modest. We illustrate these three characteristics for each nsSNP in Figures 3D, 5A, and 8C. The Eisenberg hydrophobicity scale [88] was used to analyze hydrophobicity. The APBS electrostatic plugin in PyMol was used to identify charged regions, and the illustrations were made in PyMol.

Acknowledgments

We extend our greatest gratitude to the researchers and managers of breeding associations who sampled, handled, and shipped the bee samples collected for our research herein: Anja Laupstad Vatland (Managing Director at Molti AS, Norway), Tor Erik Rødsdalen (Leader of Norsk brunbielag), Ingvard Arvidsson (Adviser at Nordbiföreningen, Sweden), Flemming Vejsnæs (Adviser at Danish Beekeepers Association), Andrew Abrahams (Manager of Colonsay Black Bee Reserve, Scotland), Gerard Coyne (Vice Chairperson at The Native Irish Honey Bee Society and Regional Director of Connacht), Małgorzata Bieńkowska (Lab head at the Research Institute of Horticulture in Skierniewice, Poland), Romée van der Zee (Dutch Center for Bee Research), Klébert Silvestre (President of the Center for Technical Apicultural Studies of Savoie, France), Peter Kozmus (Professional Leader of Breeding Program for Carniolan Honeybee for the Slovenian Beekeepers' Associations), Cecilia Costa (Researcher at Council for Agriculture Research and Agricultural Economy Analysis, Bologna, Italy), Maria Alice de Silva Pinto (Coordinator Professor at Instituto Politécnico de Bragança, Portugal), Aleksandar Uzunov (Associate Professor at Faculty of Agricultural Sciences and Food, Skopje, North Macedonia), Thomas Galea (committee member of the Malta Beekeepers Association), Irfan Kandemir (Professor at Department of Biology, Ankara University, Turkey), Adam G. Dolezal (Assistant Professor – Entomology at School of Integrative Biology, University of Illinois), Olav Rueppell (Florence Schaeffer Distinguished Professor of Science at Department of Biology, University of North Carolina at Greensboro), Jay Evans (Research Entomologist at Bee Research Laboratory, United States Department of Agriculture, Maryland), Tim Kenney (Beekeeper and manager of Red Mountain Cattle Company, Arizona), Randy Oliver (Manager of Scientific Beekeeping, California), Marla Spivak (Professor in Entomology, University of Minnesota), and Mike Goblirsch (Post-doc at the Spivak Honey Bee Lab, University of Minnesota). We thank you all for your cooperation. The authors acknowledge The Research Council of Norway grant number 262137 for funding toward running costs and positions and BioCat (RCN grant number 249023) for travel grants and conference support.

The authors declare no conflicts of interest.

Main Figure Captions

Figure 1. Illustration of the honey bee Vg structure. Vg consists of the N-terminal domain (ND) comprised of two subdomains, β-barrel (yellow) and α-helical (α-h, green), and a lipid binding site (blue), the vWF domain (vWF, cyan), and a C-terminal (C-term, magenta). The orange zig-zag line shows the proteolytic cleavage site on the polyserine linker in ND. The green plus-signs next to the α-helical subdomain illustrate the net positive surface charge. Three β-sheets (β1, β2, and β3) build up the lipid binding site. DUF1943 is defined by β-sheets 1 and 2, while the third sheet is considered

part of the lipid binding site; we refer to this structural region as the lipid binding site throughout the article. The C-terminal has been demonstrated to be flexible, as illustrated here. We show the interacting or binding units recognized by honey bee Vg to the right, colored according to the interacting domain or subdomain. We use this coloring scheme throughout the article.

Figure 2. A, B) The common and rare nsSNPs (determined by the number of occurrence in the Vg variants, more than 5 Vg variants are common, while only observed once is rare) are divided by whether they introduce a drastic or modest change in residue type. The drastic substitutions are defined determined by a change of physicochemical properties. For a complete overview of the nsSNPs properties, see Table S1. C) The surface view of five Vg variants (same coloring scheme as in Figure 1) with only common nsSNPs (red spheres). An orange asterisk (*) marks the drastic nsSNPs. D) Vg variants with several rare nsSNPs (labeled in pink) and drastic, labeled as in panel C.

Figure 3. A) The frequency of nsSNPs (used same colors as in Figure 1 for the domains and subdomains) per amino acid (y-axis) presented for the Vg domains and subdomains (x-axis). B) The Vg variants have nsSNPs in different combinations. We divided the Vg variants into domains and subdomains and found the number of unique combinations for the domains and subdomains. These are plotted here (same colors as in panel A). C) The number of unique domains and subdomains used to find the ratio to the size of the domain and subdomain sites (aa). The ratio is plotted here (same colors as in panel A). D) The nsSNPs are colored by how often they were identified on the Vg variants. We considered nsSNPs common when identified on more than 5 Vg variants (green), while the nsSNPs only identified once are considered rare (pink). The nsSNPs identified in 5 to 2 Vg variants were also considered and classified as “other” (light pink). The nsSNPs were divided into the same subdomains or domains used in panels A, B, and C and plotted according to the nsSNPs’ properties (see Table S1 for a complete overview). We calculated the relative solvent accessible surface area (rASA) for each substituted residue, determining how exposed the site is in the protein structure. We considered nsSNPs with a value of 20 % or less as buried; otherwise, they were classified as exposed. The effect of each substitution was determined using a substitution matrix (BLOSUM62) since it shows whether the physicochemical properties are preserved. The nsSNPs with a negative score were considered drastic; otherwise, they were considered modest. We plotted the nsSNPs according to the following classifications: buried or exposed and drastic or modest.

Figure 4. The identified nsSNPs in the β-barrel subdomain (yellow cartoon) are plotted together on the structure, even though the nsSNPs are not identified on the same Vg variant. The spheres represent nsSNPs (red), proposed Zn2+-binding residues (purple) and homodimerization active residues (orange). The DNA binding β-sheet is colored in pink. B) The hydrophobic core adjacent to

p.Ile132met is circled, and we show the polar surface for the subdomain. C) β-barrel variant nr. 5 and 26 are shown with the identified nsSNPs (*drastic nsSNPs).

Figure 5. A) The identified nsSNPs in the α-helical subdomain (green cartoon) are plotted together on the structure, even though the nsSNPs are not identified on the same Vg variant. The spheres represent nsSNPs (red) and homodimerization active residues (orange). The identified hotspots H1 (blue), H2 (dark pink), and H3 (yellow) are circled. The nsSNPs are also plotted according to properties per hotspot in the same way as in Figure 3D. B) We show the α-helical variant nr. 24, 41, and 45 with the identified nsSNPs labeled according to the colors of the hotpots used in panel A (*drastic nsSNPs).

Figure 6 . A) The identified nsSNPs in the lipid binding site (blue cartoon) and vWF domain (cyan) are plotted together on the structure, even though the nsSNPs are not identified on the same Vg variant. Spheres represent nsSNPs (red), and the two rare nsSNPs are labeled (pink=rare; *drastic nsSNPs). The three β-sheets shown in Figure 1 are labeled. B) Lipid binding site variants nr. 1 and 49 are shown with the identified nsSNPs (pink=rare; *drastic nsSNPs). C) The lipid cavity is very hydrophobic. D) The two charged centers are shown (black arrows). E) We show vWF variant nr. 40 with the three common nsSNPs. The Ca2+-ion is a blue sphere.

Figure 7. A) The identified nsSNPs in C-terminal (magenta cartoon) are plotted together on the structure, even though they are not identified on the same Vg variant. The spheres represent nsSNPs (red) and proposed Zn2+-binding residues (purple). NsSNPs are labeled (pink=rare; *drastic nsSNPs). B) C-terminal variant nr. 11 is shown with the two serine-introducing nsSNPs. C) The net positive exposed surface is not affected by the nsSNPs.

Figure 8. A) The identified nsSNPs in domain or subdomain interface of β-barrel subdomain (yellow), α-helical subdomain (green), and DUF1943 (blue) are plotted together on the structure, even though they are not identified on the same Vg variant. Spheres represent nsSNPs (red) and are labeled (pink=rare; *drastic nsSNPs). B) The hydrophobic core and the electrostatic charges is shown in the dashed boxes, is the same region shown in panel A. C) The same categorization of the nsSNPs identified at three domain or subdomain interfaces, as in Figure 3D. D) The identified nsSNPs in domain interface of H2 in the α-helical subdomain (green) and DUF1943 domain (blue) are plotted together on the structure, even though they are not identified on the same Vg variant. Spheres represent nsSNPs (red) and are labeled (pink=rare; *drastic nsSNPs). We show the positively charged and hydrophobic patches to the right for the same region. E) The identified nsSNPs in the domain interface of DUF1943 (blue) and vWF domain (cyan) are plotted together on the structure, even

though they are not identified on the same Vg variant. Spheres represent nsSNPs (red) and are labeled (pink=rare; *drastic nsSNPs). We show the neutral surface to the right for the same region.

Figure 9. The full-length Vg structure. The colors of domains and subdomains are the same as in Figure 1. Side A: The nsSNPs are colored red on the surface, and the gray lines indicate which domain or subdomain interface the nsSNPs belong to the ND cavity, the α-helical H2 subdomain to the DUF1943, or the DU1943 to the vWF domain. The two smaller cavities (shaded area) leading into the lipid binding site have black arrows pointing to them. Three exposed nsSNPs, not part of a domain or subdomain interface, are marked, one from H3 in the α-helical subdomain and two at the vWF domain. Side B: Rotating 180° about the y-axis reveals a large opening to the lipid binding site (shaded area and black arrow). The surface-exposed nsSNPs are colored red.

Supporting Information

Figure S1. A) Illustration of the workflow. The thorax of the sampled bees was used to extract gDNA. The *vg* gene (6109 bp) was amplified using barcoded primers (see Table S3 for the primer and barcode sequences). We used long-range PCR to amplify the full gene. A native barcode and adapter were ligated to each amplicon sequence before being loaded onto PromethION for Nanopore sequencing. B) Overview of the bioinformatic pipeline. The tools used at each step are written in blue letters above the arrow. First, the raw sequences (first box, NativeBC) were sorted using the native and inner barcodes (see Table S3), resulting in a long list of raw sequences for each sample (second box, sample). For each sample set, three tools were used to generate 1) a consensus sequence for each sample, 2) variant calling files, and 3) haplotype sequences. The blue asterisk (*) at these three steps illustrates that the reference sequence (NC_037641.1 [80]) was applied. (TIFF)

Figure S2. The identified Vg variants (yellow circles, numerated from Table S2) are plotted according to their geographical location. The sampled populations are shown as black dots (more information about the sampled locations is provided in the first section in “Materials and Methods”). (TIFF)

Figure S3. Simulation of the proposed homodimerization event. The full-length Vg is colored the same way as in Figure 9 and represents monomer 1. The second monomer is colored orange. In the event of dimerization, side A will still be exposed, while the ND could be shielded by monomer 2 on side B. (TIFF)

Figure S4. The α-helical subdomain is green with the nsSNPs (red spheres) and the 34 positively charged residues as blue sticks. No changes are introduced at the positively charged residues. (TIFF)

Figure S5. The plate setup (see Table S3) was repeated six times. Positive and negative control was included on each plate (6 plates in total). The controls were run on a 1 % TAE agarose gel shown

here. We had successful amplification of correct size (6296 bp) in the positive controls (gel 1 and 2 (+): lane 2, 4, and 6) and no significant background contamination in the negative controls (gel 1 and 2 (-): lane 3, 5, and 7). GeneRuler 1kb ladder (ThermoFisher Scientific, Waltham, MA) is in lanes 1 and 8 (1kb) in gel 1 and lane 1 (1kb) in gel 2. (TIFF)

Table S1. This table provides the identified 81 nsSNPs (listed using the recommended format from the Human Genome Variation Society [83]) and details about their properties. The number of occurrences in the Vg variants are listed in column C. The scores from the substitution matrix (BLOSUM62) are listed for each nsSNP (Column D). Negative scores indicate that the physicochemical properties are not preserved. The calculated relative solvent accessible surface area (rASA) for each substituted residue position (column E) shows the percentage of the residue exposed to the solvent. Below 20% was considered buried. (XLSX)

Table S2. This table provides the identified 121 Vg variants. The Vg variants are numerate, and the table includes the identified nsSNPs per Vg variant. The nsSNPs are written using the same format as in Table S1. (XLSX)

Table S3. This table provides the PCR primers used for amplification of the *vg*-gene. Oligo sequence, melting temperature, and oligo size are provided for the forward and reverse primers. The forward and reverse primers were barcoded, creating 8 forward primers (F1 to F8) and 12 reverse primers (R1 to R12). Here we list the full oligo sequence, where the barcodes are written in red. The oligo size includes the primers, barcodes, and the Vg gene. We also include the plate setup, which shows the barcoded forward and reverse primer combinations used in each well. We repeated this setup for all 6 plates (see “Materials and Methods” for more details on the PCR protocol and Figure S1 for the complete workflow). (XLSX)

References

1. Pál C, Papp B, Lercher MJ. An integrated view of protein evolution. Nature Reviews Genetics. 2006;7(5):337-48.
2. Camps M, Herman A, Loh E, Loeb LA. Genetic constraints on protein evolution. Crit Rev Biochem Mol Biol. 2007;42(5):313-26.
3. Subramanian S, Kumar S. Gene Expression Intensity Shapes Evolutionary Rates of the Proteins Encoded by the Vertebrate Genome. Genetics. 2004;168(1):373-81.
4. Duret L, Mouchiroud D. Determinants of substitution rates in mammalian genes: expression pattern affects selection intensity but not mutation rate. Molecular biology and evolution. 2000;17(1):68-74.
5. Zhang J, Yang J-R. Determinants of the rate of protein sequence evolution. Nat Rev Genet. 2015;16(7):409-20.
6. Podder S, Mukhopadhyay P, Ghosh TC. Multifunctionality dominantly determines the rate of human housekeeping and tissue specific interacting protein evolution. Gene. 2009;439(1):11-6.
7. Salathé M, Ackermann M, Bonhoeffer S. The Effect of Multifunctionality on the Rate of Evolution in Yeast. Molecular biology and evolution. 2005;23(4):721-2.
8. Peregrin-Alvarez JM, Tsoka S, Ouzounis CA. The phylogenetic extent of metabolic enzymes and pathways. Genome Res. 2003;13(3):422-7.
9. De S, Lopez-Bigas N, Teichmann SA. Patterns of evolutionary constraints on genes in humans. BMC Evolutionary Biology. 2008;8(1):275.
10. Langefors Å, Von Schantz T, Widegren B. Allelic variation of Mhc class II in Atlantic salmon; a population genetic analysis. Heredity. 1998;80(5):568-75.
11. de Bakker PIW, McVean G, Sabeti PC, Miretti MM, Green T, Marchini J, et al. A high-resolution HLA and SNP haplotype map for disease association studies in the extended human MHC. Nature genetics. 2006;38(10):1166-72.
12. Yue M, Han X, Masi LD, Zhu C, Ma X, Zhang J, et al. Allelic variation contributes to bacterial host specificity. Nature Communications. 2015;6(1):8754.
13. Helling R, Li H, Mélin R, Miller J, Wingreen N, Zeng C, et al. The designability of protein structures. Journal of molecular graphics & modelling. 2001;19(1):157-67.
14. Friedberg I, Margalit H. Persistently conserved positions in structurally similar, sequence dissimilar proteins: roles in preserving protein fold and function. Protein science : a publication of the Protein Society. 2002;11(2):350-60.
15. Franzosa EA, Xia Y. Structural determinants of protein evolution are context-sensitive at the residue level. Molecular biology and evolution. 2009;26(10):2387-95.
16. Liu J, Zhang Y, Lei X, Zhang Z. Natural selection of protein structural and functional properties: a single nucleotide polymorphism perspective. Genome Biology. 2008;9(4):R69.
17. Sun C, Hu L, Liu S, Gao Z, Zhang S. Functional analysis of domain of unknown function (DUF) 1943, DUF1944 and von Willebrand factor type D domain (VWD) in vitellogenin2 in zebrafish. Developmental and comparative immunology. 2013;41(4):469-76.
18. Zhang S, Dong Y, Cui P. Vitellogenin is an immunocompetent molecule for mother and offspring in fish. Fish & shellfish immunology. 2015;46(2):710-5.
19. Havukainen H, Halskau O, Amdam GV. Social pleiotropy and the molecular evolution of honey bee vitellogenin. Molecular ecology. 2011;20(24):5111-3.
20. Kent CF, Issa A, Bunting AC, Zayed A. Adaptive evolution of a key gene affecting queen and worker traits in the honey bee, Apis mellifera. Molecular ecology. 2011;20(24):5226-35.
21. vanEngelsdorp D, Meixner MD. A historical review of managed honey bee populations in Europe and the United States and the factors that may affect them. Journal of Invertebrate Pathology. 2010;103:S80-S95.
22. Weinstock GM, Robinson GE, Gibbs RA, Weinstock GM, Weinstock GM, Robinson GE, et al. Insights into social insects from the genome of the honeybee Apis mellifera. Nature. 2006;443(7114):931-49.

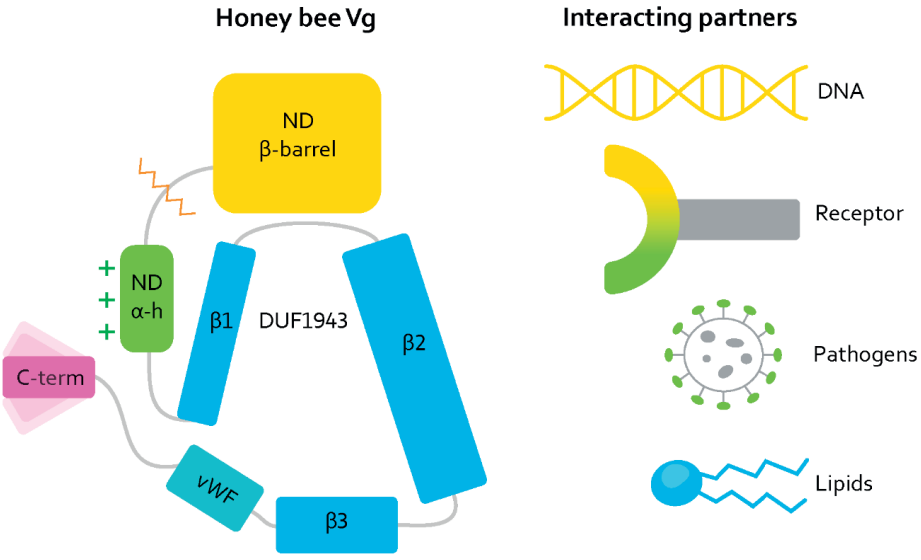
23. Menzel R, Leboulle G, Eisenhardt D. Small brains, bright minds. Cell. 2006;124(2):237-9.
24. Amdam GV, Simoes ZL, Hagen A, Norberg K, Schroder K, Mikkelsen O, et al. Hormonal control of the yolk precursor vitellogenin regulates immune function and longevity in honeybees. Experimental gerontology. 2004;39(5):767-73.
25. Münch D, Ihle KE, Salmela H, Amdam GV. Vitellogenin in the honey bee brain: Atypical localization of a reproductive protein that promotes longevity. Experimental gerontology. 2015;71:103-8.
26. Corona M, Velarde RA, Remolina S, Moran-Lauter A, Wang Y, Hughes KA, et al. Vitellogenin, juvenile hormone, insulin signaling, and queen honey bee longevity. Proceedings of the National Academy of Sciences of the United States of America. 2007;104(17):7128-33.
27. Havukainen H, Halskau O, Skjaerven L, Smedal B, Amdam GV. Deconstructing honeybee vitellogenin: novel 40 kDa fragment assigned to its N terminus. The Journal of experimental biology. 2011;214(Pt 4):582-92.
28. Leipart V, Montserrat-Canals M, Cunha ES, Luecke H, Herrero-Galán E, Halskau Ø, et al. Structure prediction of honey bee vitellogenin: a multi-domain protein important for insect immunity. FEBS Open Bio. 2021;In press(In press).
29. Li A, Sadasivam M, Ding JL. Receptor-Ligand Interaction between Vitellogenin Receptor (VtgR) and Vitellogenin (Vtg), Implications on Low Density Lipoprotein Receptor and Apolipoprotein B/E: THE FIRST THREE LIGAND-BINDING REPEATS OF VTGR INTERACT WITH THE AMINO-TERMINAL REGION OF VTG *. Journal of Biological Chemistry. 2003;278(5):2799-806.
30. Roth Z, Weil S, Aflalo ED, Manor R, Sagi A, Khalaila I. Identification of Receptor-Interacting Regions of Vitellogenin within Evolutionarily Conserved β-Sheet Structures by Using a Peptide Array. ChemBioChem. 2013;14(9):1116-22.
31. Upadhyay SK, Singh H, Dixit S, Mendu V, Verma PC. Molecular Characterization of Vitellogenin and Vitellogenin Receptor of Bemisia tabaci. PloS one. 2016;11(5):e0155306-e.
32. Thompson JR, Banaszak LJ. Lipid-protein interactions in lipovitellin. Biochemistry. 2002;41(30):9398-409.
33. Anderson TA, Levitt DG, Banaszak LJ. The structural basis of lipid interactions in lipovitellin, a soluble lipoprotein. Structure (London, England : 1993). 1998;6(7):895-909.
34. Tufail M, Takeda M. Molecular characteristics of insect vitellogenins. Journal of insect physiology. 2008;54(12):1447-58.
35. Sappington TW, S. Raikhel A. Molecular characteristics of insect vitellogenins and vitellogenin receptors. Insect biochemistry and molecular biology. 1998;28(5):277-300.
36. Havukainen H, Underhaug J, Wolschin F, Amdam G, Halskau O. A vitellogenin polyserine cleavage site: highly disordered conformation protected from proteolysis by phosphorylation. The Journal of experimental biology. 2012;215(Pt 11):1837-46.
37. Salmela H, Harwood G, Münch D, Elsik C, Herrero-Galán E, Vartiainen MK, et al. Nuclear Translocation of Vitellogenin in the Honey Bee (Apis mellifera). bioRxiv. 2021:2021.08.18.456851.
38. Havukainen H, Munch D, Baumann A, Zhong S, Halskau O, Krogsgaard M, et al. Vitellogenin recognizes cell damage through membrane binding and shields living cells from reactive oxygen species. The Journal of biological chemistry. 2013;288(39):28369-81.
39. Salmela H, Amdam GV, Freitak D. Transfer of Immunity from Mother to Offspring Is Mediated via Egg-Yolk Protein Vitellogenin. PLoS pathogens. 2015;11(7):e1005015.
40. Smolenaars MMW, Madsen O, Rodenburg KW, Van der Horst DJ. Molecular diversity and evolution of the large lipid transfer protein superfamily^{s⃞}. Journal of Lipid Research. 2007;48(3):489-502.
41. Du X, Wang X, Wang S, Zhou Y, Zhang Y, Zhang S. Functional characterization of Vitellogenin_N domain, domain of unknown function 1943, and von Willebrand factor type D domain in vitellogenin of the non-bilaterian coral Euphyllia ancora: Implications for emergence of

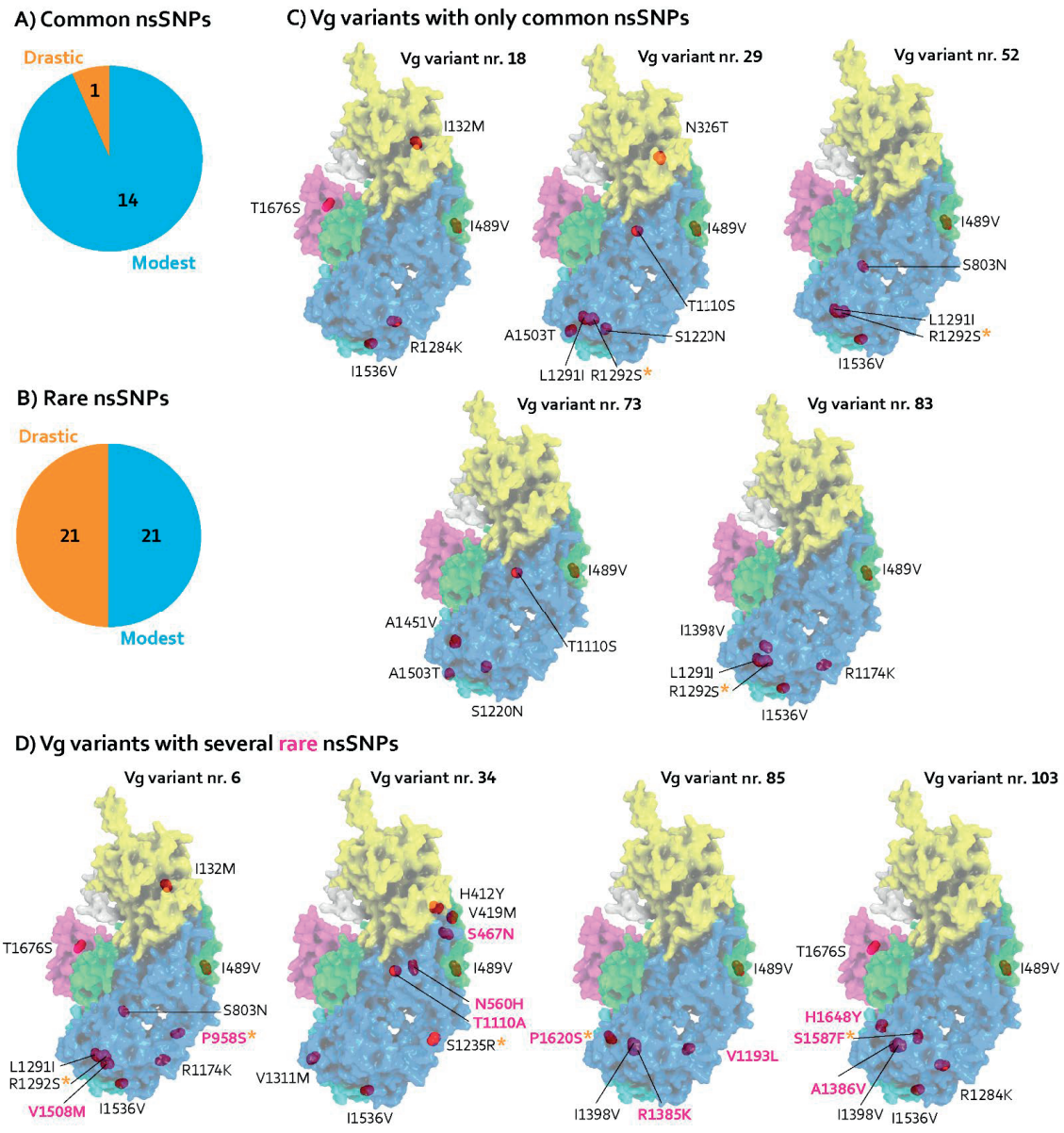
immune activity of vitellogenin in basal metazoan. Developmental and comparative immunology. 2017;67:485-94.
42. Qiao K, Jiang C, Xu M, Chen B, Qiu W, Su Y, et al. Molecular Characterization of the Von Willebrand Factor Type D Domain of Vitellogenin from Takifugu flavidus. Marine Drugs. 2021;19(4):181.
43. Ilyasov RA, Poskryakov AV, Nikolenko AG. [New SNP markers of the honeybee vitellogenin gene (Vg) used for identification of subspecies Apis mellifera mellifera L]. Genetika. 2015;51(2):194-9.
44. Morita M, Katta AM, Ahmad S, Mori T, Sugita Y, Mizuguchi K. Lipid recognition propensities of amino acids in membrane proteins from atomic resolution data. BMC biophysics. 2011;4:21.
45. Ma B, Elkayam T, Wolfson H, Nussinov R. Protein-protein interactions: structurally conserved residues distinguish between binding sites and exposed protein surfaces. Proceedings of the National Academy of Sciences of the United States of America. 2003;100(10):5772-7.
46. Stank A, Kokh DB, Fuller JC, Wade RC. Protein Binding Pocket Dynamics. Accounts of Chemical Research. 2016;49(5):809-15.
47. Li A, Sadasivam M, Ding JL. Receptor-ligand interaction between vitellogenin receptor (VtgR) and vitellogenin (Vtg), implications on low density lipoprotein receptor and apolipoprotein B/E. The first three ligand-binding repeats of VtgR interact with the amino-terminal region of Vtg. The Journal of biological chemistry. 2003;278(5):2799-806.
48. Li Z, Zhang S, Liu Q. Vitellogenin functions as a multivalent pattern recognition receptor with an opsonic activity. PloS one. 2008;3(4):e1940-e.
49. Sun C, Zhang S. Immune-Relevant and Antioxidant Activities of Vitellogenin and Yolk Proteins in Fish. Nutrients. 2015;7(10):8818-29.
50. Liu Q-H, Zhang S-C, Li Z-J, Gao C-R. Characterization of a pattern recognition molecule vitellogenin from carp (Cyprinus carpio). Immunobiology. 2009;214(4):257-67.
51. Seabury CM, Womack JE. Analysis of sequence variability and protein domain architectures for bovine peptidoglycan recognition protein 1 and Toll-like receptors 2 and 6. Genomics. 2008;92(4):235-45.
52. Haunshi S, Burramsetty AK, Ramasamy K, Chatterjee RN. Polymorphisms in pattern recognition receptor genes of indigenous and White Leghorn breeds of chicken. Arch Anim Breed. 2018;61(4):441-9.
53. Seabury CM, Seabury PM, Decker JE, Schnabel RD, Taylor JF, Womack JE. Diversity and evolution of 11 innate immune genes in Bos taurus taurus and Bos taurus indicus cattle. Proceedings of the National Academy of Sciences of the United States of America. 2010;107(1):151-6.
54. Bhaskara RM, Srinivasan N. Stability of domain structures in multi-domain proteins. Scientific reports. 2011;1:40-.
55. Ittisoponpisan S, Islam SA, Khanna T, Alhuzimi E, David A, Sternberg MJE. Can Predicted Protein 3D Structures Provide Reliable Insights into whether Missense Variants Are Disease Associated? Journal of molecular biology. 2019;431(11):2197-212.
56. Tokuriki N, Stricher F, Schymkowitz J, Serrano L, Tawfik DS. The Stability Effects of Protein Mutations Appear to be Universally Distributed. Journal of Molecular Biology. 2007;369(5):1318-32.
57. Wang L, Walsh MT, Small DM. Apolipoprotein B is conformationally flexible but anchored at a triolein/water interface: A possible model for lipoprotein surfaces. Proceedings of the National Academy of Sciences. 2006;103(18):6871-6.
58. Lai J-S, Cheng C-W, Lo A, Sung T-Y, Hsu W-L. Lipid exposure prediction enhances the inference of rotational angles of transmembrane helices. BMC Bioinformatics. 2013;14(1):304.
59. Zhang Z, Miteva MA, Wang L, Alexov E. Analyzing effects of naturally occurring missense mutations. Comput Math Methods Med. 2012;2012:805827-.
60. Peacock TP, Sealy JE, Harvey WT, Benton DJ, Reeve R, Iqbal M, et al. Genetic Determinants of Receptor-Binding Preference and Zoonotic Potential of H9N2 Avian Influenza Viruses. Journal of Virology. 2021;95(5):e01651-20.

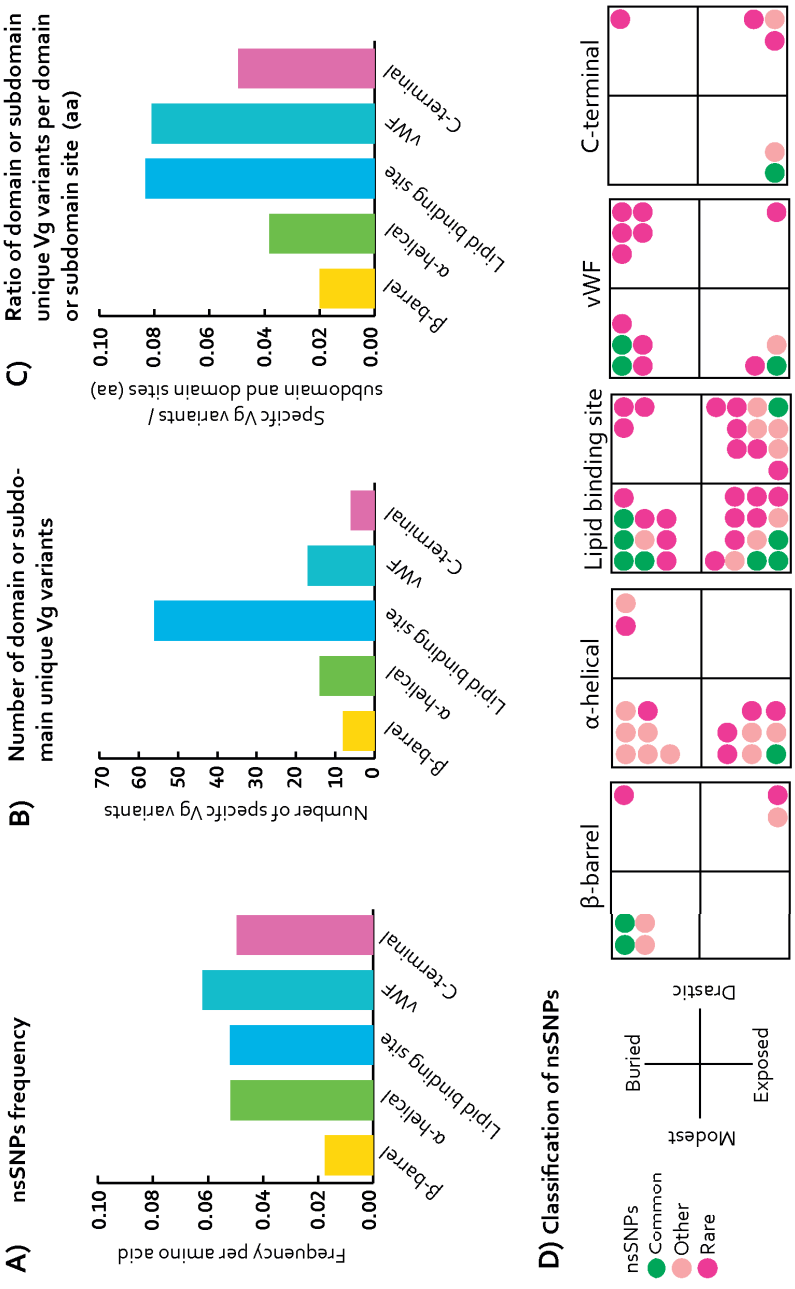
61. Dougan SK, Salas A, Rava P, Agyemang A, Kaser A, Morrison J, et al. Microsomal triglyceride transfer protein lipidation and control of CD1d on antigen-presenting cells. Journal of Experimental Medicine. 2005;202(4):529-39.
62. Rakhshandehroo M, Gijzel SMW, Siersbæk R, Broekema MF, de Haar C, Schipper HS, et al. CD1d-mediated Presentation of Endogenous Lipid Antigens by Adipocytes Requires Microsomal Triglyceride Transfer Protein *. Journal of Biological Chemistry. 2014;289(32):22128-39.
63. Mahbubur Rahman M, Ma G, Roberts HLS, Schmidt O. Cell-free immune reactions in insects. Journal of insect physiology. 2006;52(7):754-62.
64. Whitten MMA, Tew IF, Lee BL, Ratcliffe NA. A Novel Role for an Insect Apolipoprotein (Apolipophorin III) in β-1,3-Glucan Pattern Recognition and Cellular Encapsulation Reactions. The Journal of Immunology. 2004;172(4):2177-85.
65. Niere M, Dettloff M, Maier T, Ziegler M, Wiesner A. Insect Immune Activation by Apolipophorin III Is Correlated with the Lipid-Binding Properties of This Protein. Biochemistry. 2001;40(38):11502-8.
66. Spurny R, Přidal A, Pálková L, Kiem HKT, Miranda JRd, Plevka P, et al. Virion Structure of Black Queen Cell Virus, a Common Honeybee Pathogen. Journal of Virology. 2017;91(6):e02100-16.
67. Škubník K, Nováček J, Füzik T, Přidal A, Paxton RJ, Plevka P. Structure of deformed wing virus, a major honey bee pathogen. Proceedings of the National Academy of Sciences. 2017;114(12):3210-5.
68. Biterova EI, Isupov MN, Keegan RM, Lebedev AA, Sohail AA, Liaqat I, et al. The crystal structure of human microsomal triglyceride transfer protein. Proceedings of the National Academy of Sciences. 2019;116(35):17251-60.
69. Henriques D, Browne KA, Barnett MW, Parejo M, Kryger P, Freeman TC, et al. High sample throughput genotyping for estimating C-lineage introgression in the dark honeybee: an accurate and cost-effective SNP-based tool. Scientific reports. 2018;8(1):8552-.
70. Munoz I, Henriques D, Jara L, Johnston JS, Chavez-Galarza J, De La Rua P, et al. SNPs selected by information content outperform randomly selected microsatellite loci for delineating genetic identification and introgression in the endangered dark European honeybee (Apis mellifera mellifera). Molecular ecology resources. 2017;17(4):783-95.
71. Pinto MA, Henriques D, Chávez-Galarza J, Kryger P, Garnery L, van der Zee R, et al. Genetic integrity of the Dark European honey bee (Apis mellifera mellifera) from protected populations: a genome-wide assessment using SNPs and mtDNA sequence data. Journal of Apicultural Research. 2014;53(2):269-78.
72. Vergunst AC, van Lier MCM, den Dulk-Ras A, Grosse Stüve TA, Ouwehand A, Hooykaas PJJ. Positive charge is an important feature of the C-terminal transport signal of the VirB/D4-translocated proteins of Agrobacterium. Proceedings of the National Academy of Sciences of the United States of America. 2005;102(3):832-7.
73. Papakonstantinou T, Galanis M, Nagley P, Devenish RJ. Each of three positively-charged amino acids in the C-terminal region of yeast mitochondrial ATP synthase subunit 8 is required for assembly. Biochimica et biophysica acta. 1993;1144(1):22-32.
74. Waclawska I, Ziegler C. Regulatory role of charged clusters in the N-terminal domain of BetP from Corynebacterium glutamicum. Biological Chemistry. 2015;396(9-10):1117-26.
75. Seehuus SC, Norberg K, Gimsa U, Krekling T, Amdam GV. Reproductive protein protects functionally sterile honey bee workers from oxidative stress. Proceedings of the National Academy of Sciences of the United States of America. 2006;103(4):962-7.
76. Finkel T. Redox-dependent signal transduction. FEBS letters. 2000;476(1-2):52-4.
77. Cremers CM, Jakob U. Oxidant sensing by reversible disulfide bond formation. The Journal of biological chemistry. 2013;288(37):26489-96.
78. Beyter D, Ingimundardottir H, Oddsson A, Eggertsson HP, Bjornsson E, Jonsson H, et al. Long-read sequencing of 3,622 Icelanders provides insight into the role of structural variants in human diseases and other traits. Nature genetics. 2021;53(6):779-86.

79. Sakamoto Y, Zaha S, Suzuki Y, Seki M, Suzuki A. Application of long-read sequencing to the detection of structural variants in human cancer genomes. Computational and Structural Biotechnology Journal. 2021;19:4207-16.
80. Wallberg A, Bunikis I, Pettersson OV, Mosbech MB, Childers AK, Evans JD, et al. A hybrid de novo genome assembly of the honeybee, Apis mellifera, with chromosome-length scaffolds. BMC genomics. 2019;20(1):275.
81. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. 2011. 2011;17(1):3.
82. Rozas J, Ferrer-Mata A, Sánchez-DelBarrio JC, Guirao-Rico S, Librado P, Ramos-Onsins SE, et al. DnaSP 6: DNA Sequence Polymorphism Analysis of Large Data Sets. Molecular biology and evolution. 2017;34(12):3299-302.
83. den Dunnen JT, Dalgleish R, Maglott DR, Hart RK, Greenblatt MS, McGowan-Jordan J, et al. HGVS Recommendations for the Description of Sequence Variants: 2016 Update. Human mutation. 2016;37(6):564-9.
84. Schrodinger L. The PyMOL Molecular Graphics System, Version 1.8. 2015.
85. Savojardo C, Manfredi M, Martelli PL, Casadio R. Solvent Accessibility of Residues Undergoing Pathogenic Variations in Humans: From Protein Structures to Protein Sequences. Frontiers in Molecular Biosciences. 2021;7(460).
86. Tien MZ, Meyer AG, Sydykova DK, Spielman SJ, Wilke CO. Maximum allowed solvent accessibilites of residues in proteins. PloS one. 2013;8(11):e80635-e.
87. Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. Proceedings of the National Academy of Sciences of the United States of America. 1992;89(22):10915-9.
88. Eisenberg D, Schwarz E, Komaromy M, Wall R. Analysis of membrane and surface protein sequences with the hydrophobic moment plot. J Mol Biol. 1984;179(1):125-42.

Figure 1


Honey bee Vg
ND
β-barrel
ND
α-h
β1
DUF1943
β2
C-term
vWF
β3
Interacting partners
DNA
Receptor
Pathogens
Lipids


A) Common nsSNPs
Drastic
1
14
Modest
C) Vg variants with only common nsSNPs
Vg variant nr. 18
I132M
T1676S
I489V
R1284K
I1536V
Vg variant nr. 29
N326T
I489V
T1110S
A1503T
S1220N
L1291I
R1292S*
Vg variant nr. 52
I489V
S803N
L1291I
R1292S*
I1536V
B) Rare nsSNPs
Drastic
21
21
Modest
Vg variant nr. 73
I489V
A1451V
T1110S
A1503T
S1220N
Vg variant nr. 83
I489V
I1398V
L1291I
R1292S*
R1174K
I1536V
D) Vg variants with several rare nsSNPs
Vg variant nr. 6
I132M
T1676S
I489V
S803N
P958S*
L1291I
R1292S*
R1174K
V1508M
I1536V
Vg variant nr. 34
H412Y
V419M
S467N
I489V
N560H
T1110A
S1235R*
V1311M
I1536V
Vg variant nr. 85
I489V
P1620S*
V1193L
I1398V
R1385K
Vg variant nr. 103
T1676S
I489V
H1648Y
S1587F*
A1386V
R1284K
I1398V
I1536V


A)
nsSNPs frequency
Frequency per amino acid
0.10
0.08
0.06
0.04
0.02
0.00
β-barrel
α-helical
Lipid binding site
vWF
C-terminal
B)
Number of domain or subdo-
main unique Vg variants
Number of specific Vg variants
70
60
50
40
30
20
10
0
β-barrel
α-helical
Lipid binding site
vWF
C-terminal
C)
Ratio of domain or subdomain
unique Vg variants per domain
or subdomain site (aa)
Specific Vg variants /
subdomain and domain sites (aa)
0.10
0.08
0.06
0.04
0.02
0.00
β-barrel
α-helical
Lipid binding site
vWF
C-terminal
D) Classification of nsSNPs
β-barrel
α-helical
Lipid binding site
vWF
C-terminal
Drastic
Modest
Buried
Exposed
nsSNPs
Common
Other
Rare

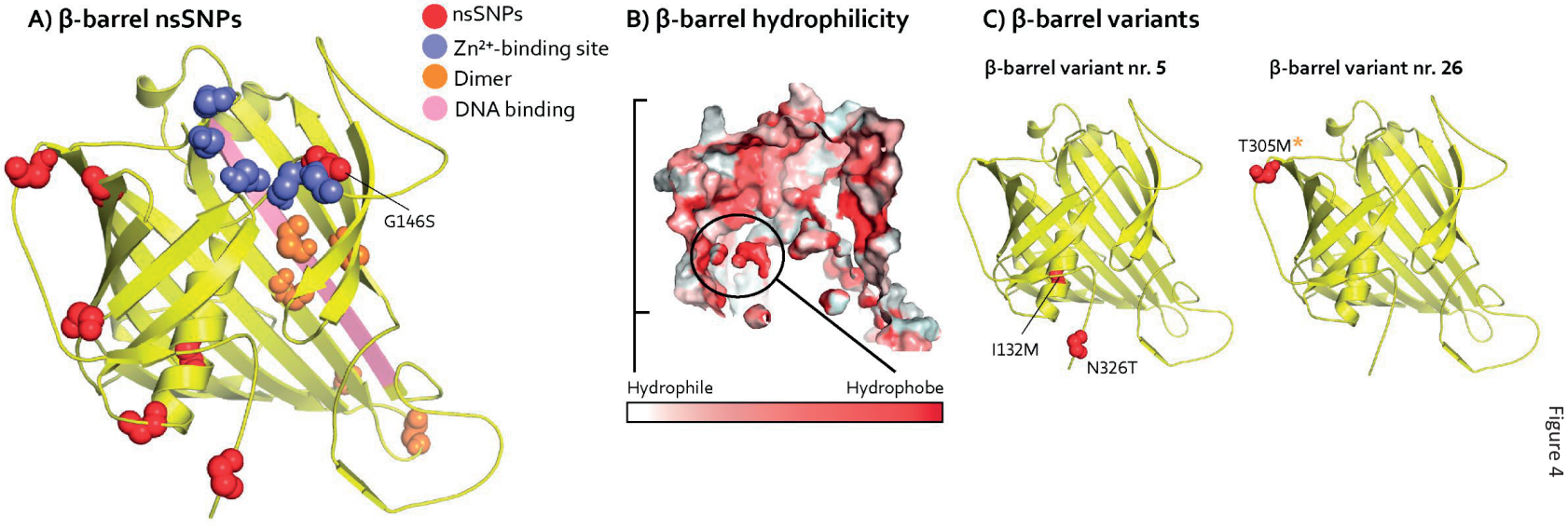

A) β-barrel nsSNPs
nsSNPs
Zn²⁺-binding site
Dimer
DNA binding
G146S
B) β-barrel hydrophilicity
Hydrophile
Hydrophobe
C) β-barrel variants
β-barrel variant nr. 5
β-barrel variant nr. 26
T305M*
I132M
N326T

Figure 4

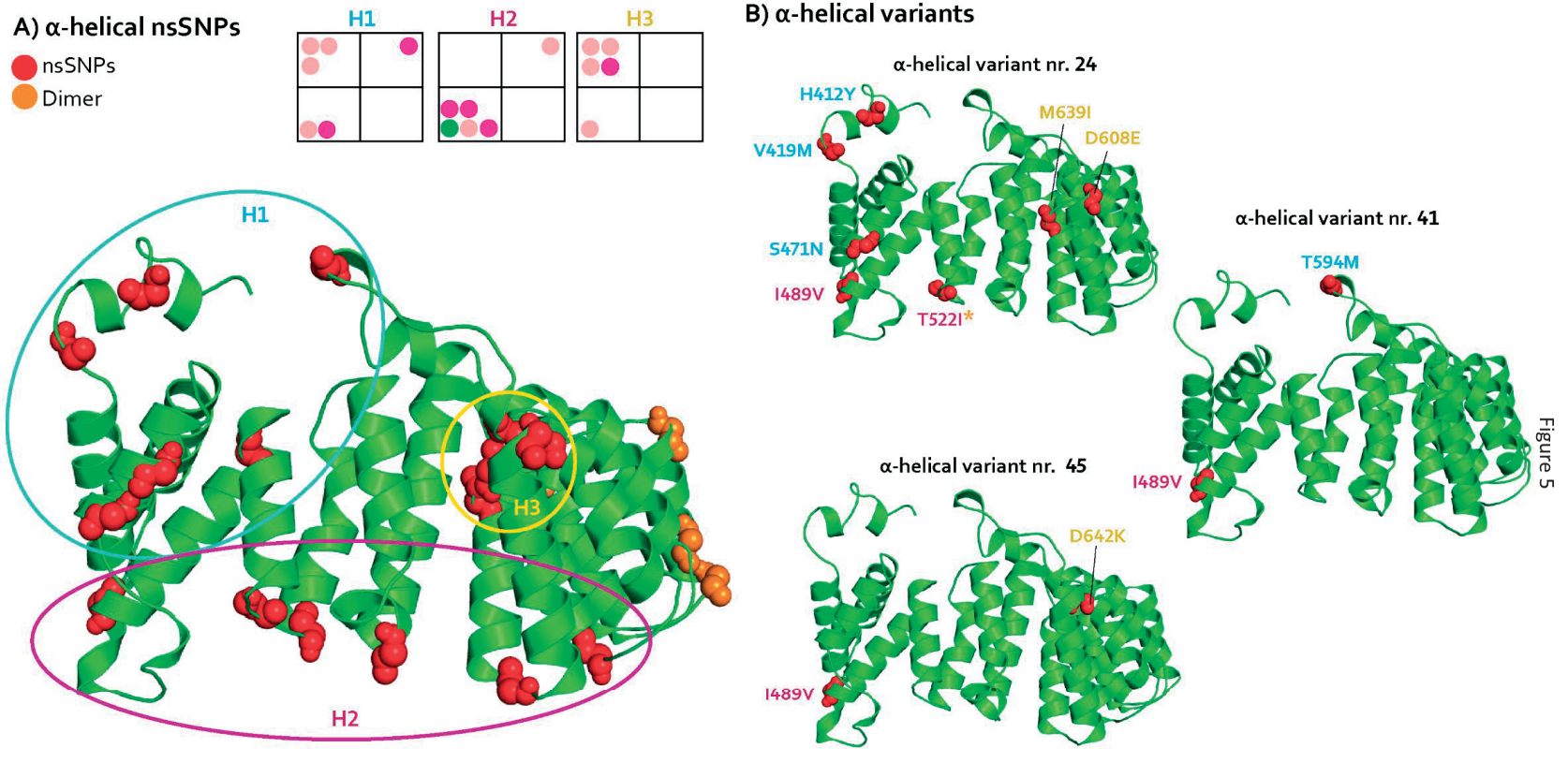
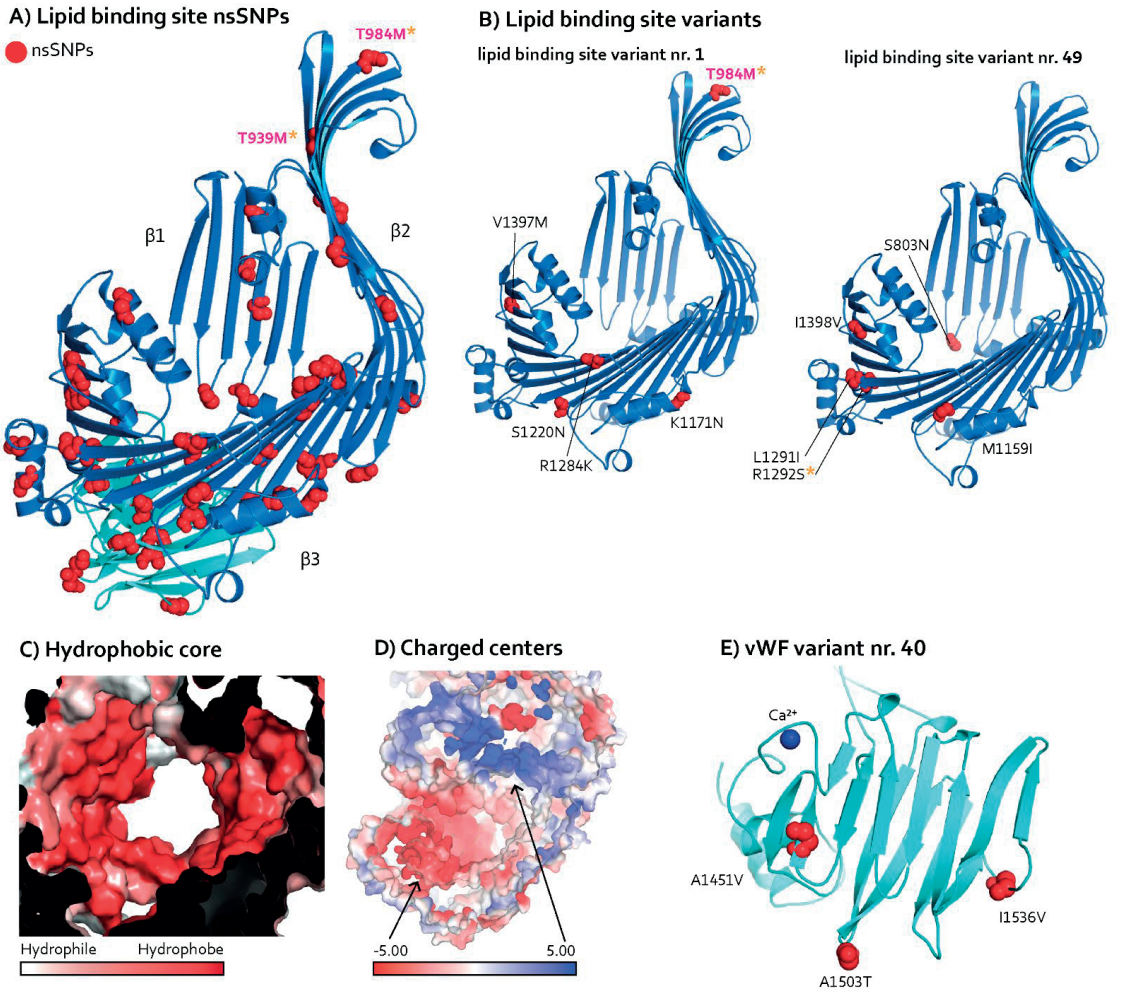
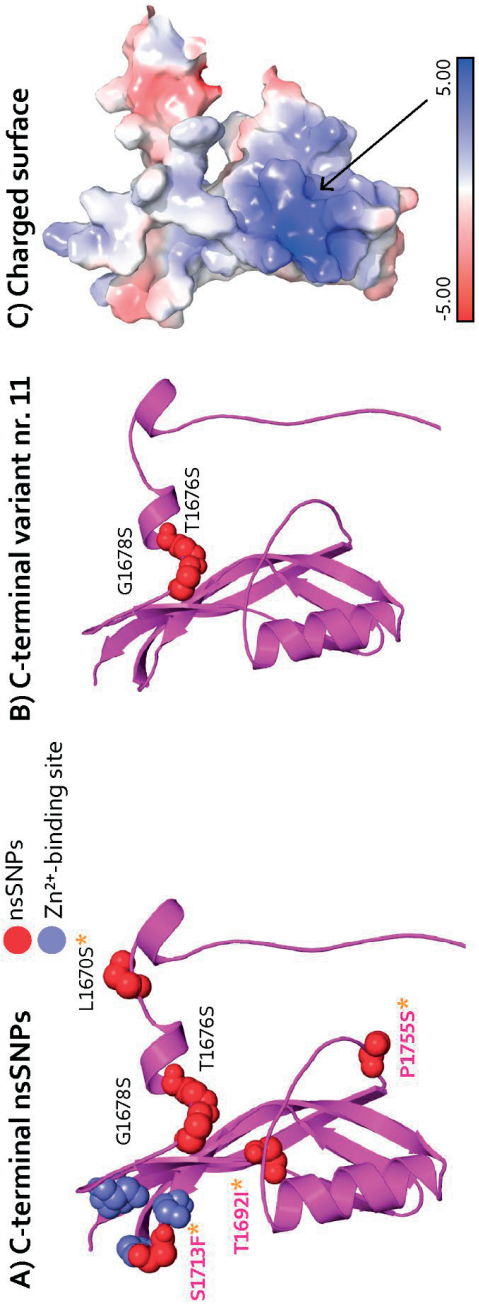
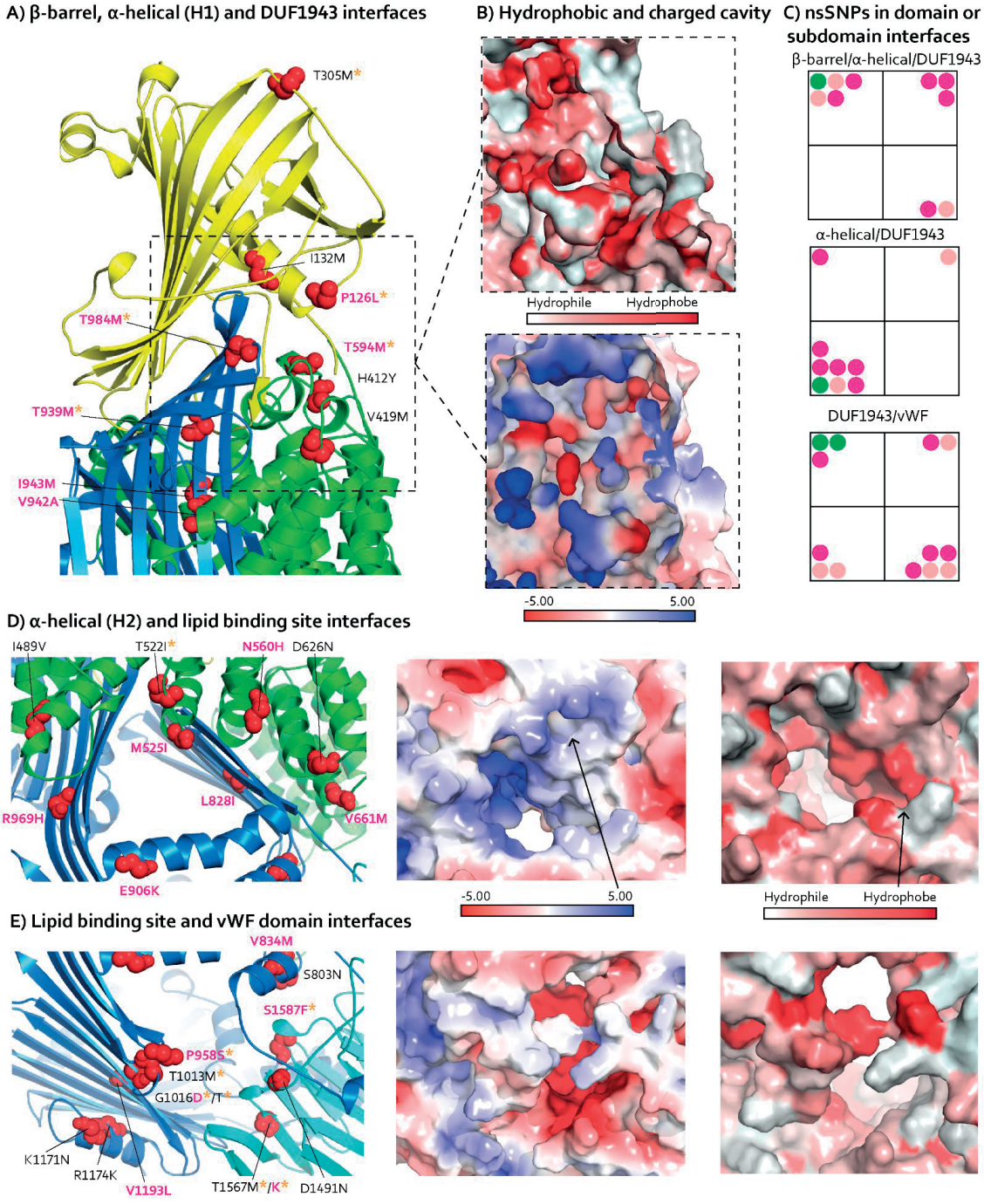
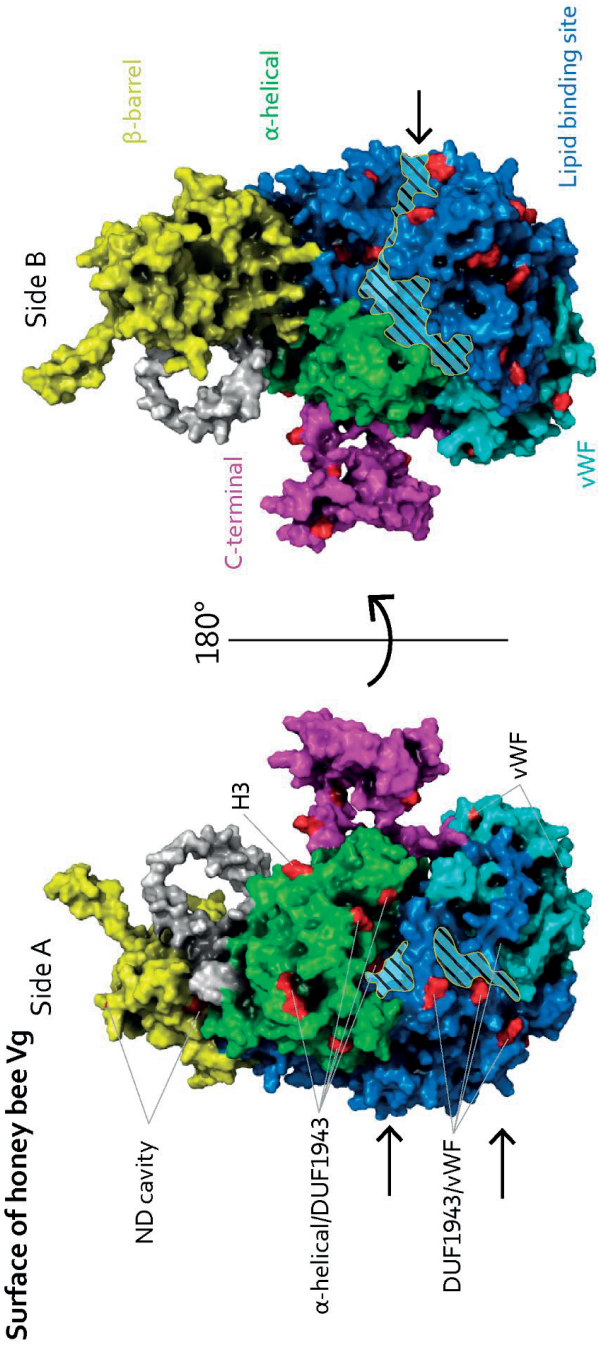

A) α-helical nsSNPs
nsSNPs
Dimer
H1
H2
H3
H1
H3
H2
B) α-helical variants
α-helical variant nr. 24
H412Y
V419M
M639I
D608E
S471N
I489V
T522I*
α-helical variant nr. 41
T594M
I489V
α-helical variant nr. 45
D642K
I489V

Figure 5


A) Lipid binding site nsSNPs
nsSNPs
T984M*
T939M*
β1
β2
β3
B) Lipid binding site variants
lipid binding site variant nr. 1
T984M*
V1397M
S1220N
K1171N
R1284K
lipid binding site variant nr. 49
S803N
I1398V
L1291I
R1292S*
M1159I
C) Hydrophobic core
Hydrophile
Hydrophobe
D) Charged centers
-5.00
5.00
E) vWF variant nr. 40
Ca²⁺
A1451V
I1536V
A1503T


A) C-terminal nsSNPs
nsSNPs
Zn²⁺-binding site
L1670S*
G1678S
T1676S
S1713F*
T1692I*
P1755S*
B) C-terminal variant nr. 11
G1678S
T1676S
C) Charged surface
5.00
-5.00


A) β-barrel, α-helical (H1) and DUF1943 interfaces
T305M*
I132M
P126L*
T984M*
T594M*
H412Y
T939M*
V419M
I943M
V942A
B) Hydrophobic and charged cavity
Hydrophile
Hydrophobe
-5.00
5.00
C) nsSNPs in domain or subdomain interfaces
β-barrel/α-helical/DUF1943
α-helical/DUF1943
DUF1943/vWF
D) α-helical (H2) and lipid binding site interfaces
I489V
T522I*
N560H
D626N
M525I
L828I
R969H
V661M
E906K
-5.00
5.00
Hydrophile
Hydrophobe
E) Lipid binding site and vWF domain interfaces
V834M
S803N
S1587F*
P958S*
T1013M*
G1016D*/T*
K1171N
R1174K
V1193L
T1567M*/K*
D1491N


Surface of honey bee Vg
Side A
ND cavity
H3
α-helical/DUF1943
DUF1943/vWF
vWF
180°
Side B
β-barrel
α-helical
C-terminal
Lipid binding site
vWF

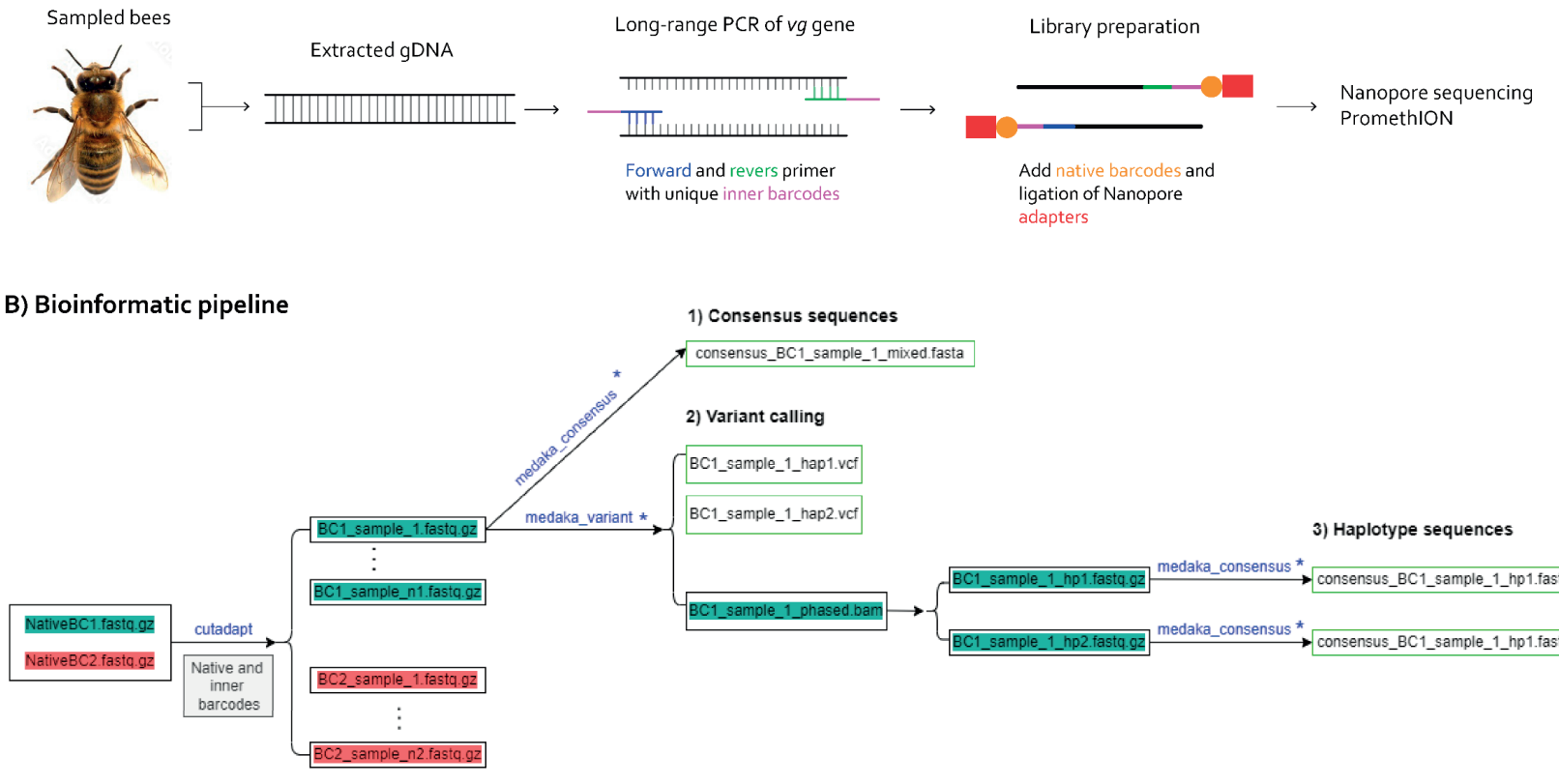

A) Workflow
Sampled bees
Extracted gDNA
Long-range PCR of vg gene
Forward and revers primer with unique inner barcodes
Library preparation
Add native barcodes and ligation of Nanopore adapters
Nanopore sequencing PromethION
B) Bioinformatic pipeline
1) Consensus sequences
consensus_BC1_sample_1_mixed.fasta
medaka_consensus *
2) Variant calling
BC1_sample_1_hap1.vcf
BC1_sample_1_hap2.vcf
medaka_variant *
3) Haplotype sequences
BC1_sample_1_hp1.fastq.gz
BC1_sample_1_hp2.fastq.gz
medaka_consensus *
consensus_BC1_sample_1_hp1.fasta
consensus_BC1_sample_1_hp1.fasta
BC1_sample_1_phased.bam
BC1_sample_1.fastq.gz
BC1_sample_n1.fastq.gz
BC2_sample_1.fastq.gz
BC2_sample_n2.fastq.gz
NativeBC1.fastq.gz
NativeBC2.fastq.gz
cutadapt
Native and inner barcodes

Figure S1

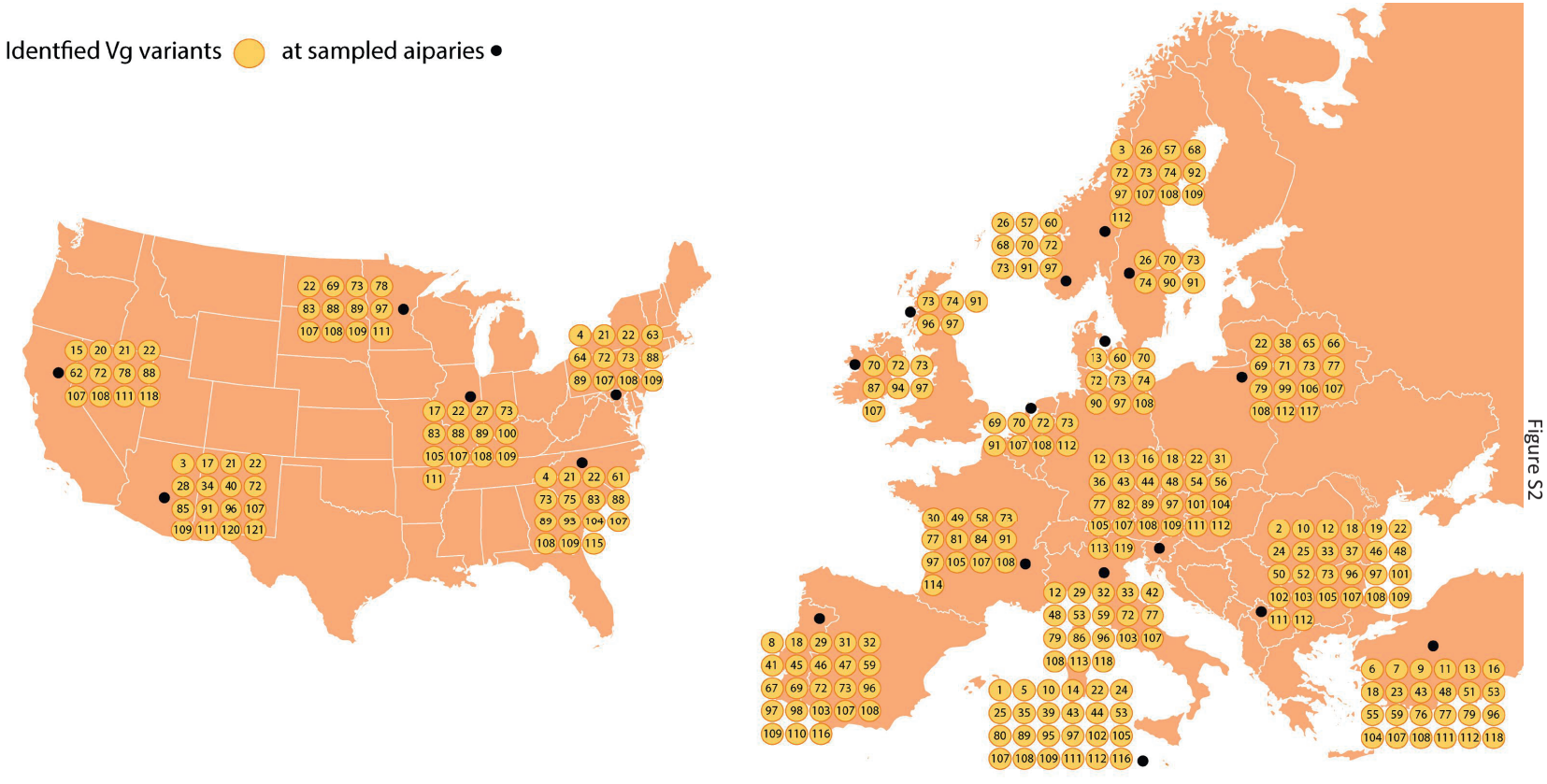


Figure S2

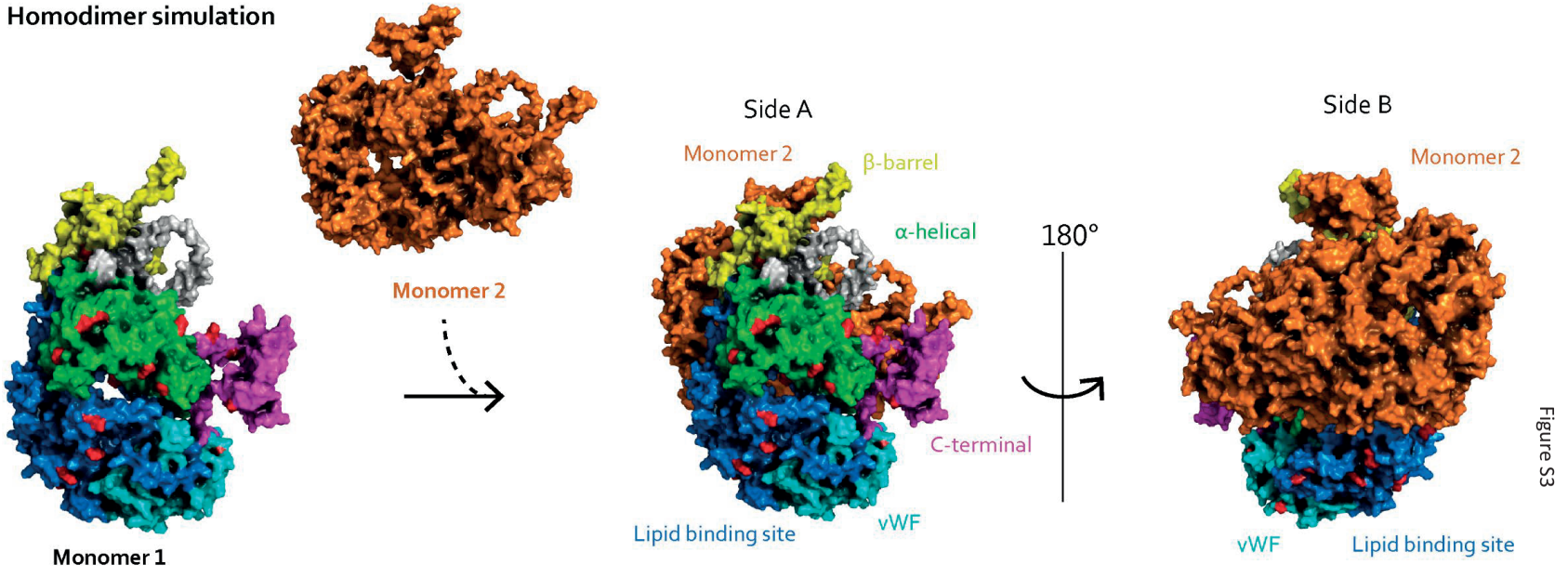

Homodimer simulation
Monomer 2
Monomer 1
Side A
Monomer 2
β-barrel
α-helical
C-terminal
vWF
Lipid binding site
180°
Side B
Monomer 2
vWF
Lipid binding site

Figure S3

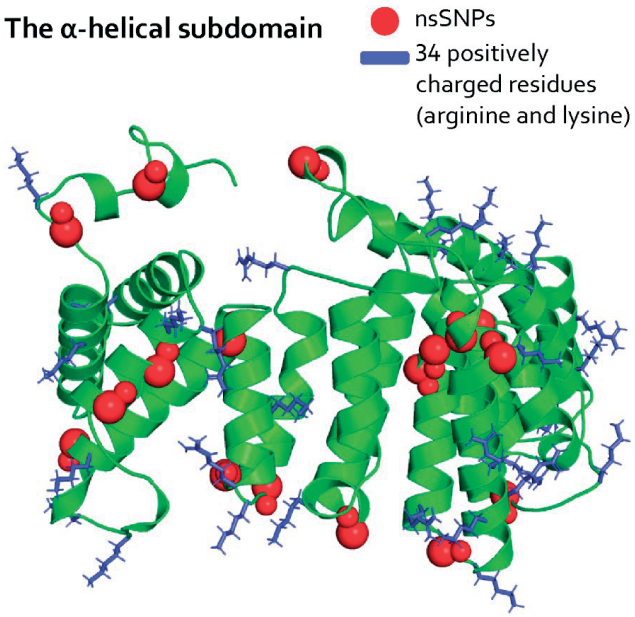

The α-helical subdomain
nsSNPs
34 positively charged residues (arginine and lysine)

Figure S5

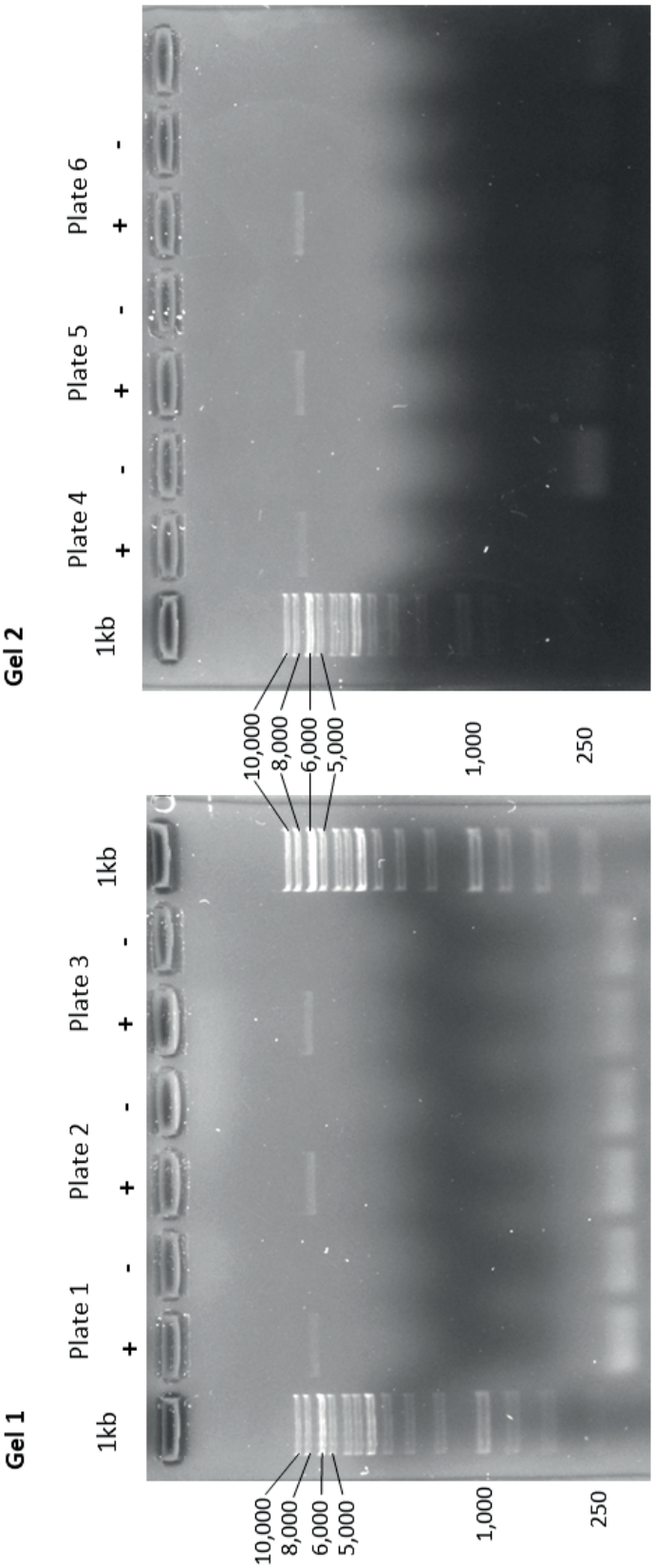

Gel 1
1kb
Plate 1
+
-
Plate 2
+
-
Plate 3
+
-
1kb
10,000
8,000
6,000
5,000
1,000
250
Gel 2
1kb
Plate 4
+
-
Plate 5
+
-
Plate 6
+
-
10,000
8,000
6,000
5,000
1,000
250

Table S1. The identifed nsSNPs

Number	nsSNPs	Number of occurences on Vg variants	Modest (positive values) or drastic (negative values)	Exposed (>20 %), otherwise buried
1	p.Ala60Thr	2	0	2
2	p.Pro106Ser	1	-1	15
3	p.Pro126Leu	1	-3	28
4	p.Ile132Met	16	1	1
5	p.Gly146Ser	3	0	16
6	p.Thr305Met	3	-1	22
7	p.Asn326Thr	8	0	3
8	p.His412Tyr	3	2	3
9	p.Val419Met	3	1	5
10	p.Ser467Asn	1	1	34
11	p.Ser471Asn	3	1	59
12	p.Ile489Val	119	3	36
13	p.Ala509Thr	4	0	2
14	p.Thr522Ile	5	-1	1
15	p.Met525Ile	1	1	61
16	p.Asn560His	1	1	77
17	p.Thr594Met	1	-1	2
18	p.Leu606Phe	1	0	4
19	p.Asp608Glu	4	2	68
20	p.Asp626Asn	2	1	27
21	p.Met639Ile	4	1	6
22	p.Ile640Val	2	3	0
23	p.Glu642Lys	2	1	9
24	p.Val661Met	1	1	26
25	p.Ser803Asn	12	1	16
26	p.Leu828Ile	1	2	7
27	p.Val834Met	1	1	1
28	p.Pro866Ser	1	-1	18
29	p.Glu906Lys	1	1	49
30	p.Thr939Met	1	-1	18
31	p.Val942Ala	1	0	3
32	p.Ile943Met	1	1	0
33	p.Pro958Ser	1	-1	91
34	p.Arg969His	1	0	27
35	p.Thr984Met	1	-1	0
36	p.Thr1013Met	2	-1	46
37	p.Gly1016Asp	1	-1	29
38	p.Gly1016Thr	2	-2	29
39	p.Leu1072Phe	1	0	31
40	p.Asp1103Tyr	1	-3	65
41	p.Thr1110Ala	1	0	23
42	p.Thr1110Ser	14	1	23

43	p.Met1159Ile	5	1	60
44	p.Lys1171Asn	4	0	61
45	p.Arg1174Lys	12	2	4
46	p.Val1193Leu	1	1	31
47	p.Val1199Ile	1	3	33
48	p.Thr1207Ile	1	-1	23
49	p.Ser1220Asn	43	1	3
50	p.Ser1235Arg	2	-1	58
51	p.Ala1237Leu	1	-1	35
52	p.Arg1284Lys	54	2	39
53	p.Leu1291Ile	22	2	17
54	p.Arg1292Ser	25	-1	49
55	p.Gly1302Glu	2	-2	35
56	p.Val1311Met	2	1	24
57	p.Phe1357Val	1	-1	43
58	p.Arg1385Lys	1	2	37
59	p.Ala1386Val	1	0	16
60	p.Val1397Met	2	1	19
61	p.Ile1398Val	64	3	27
62	p.Ala1451Val	31	0	0
63	p.Asp1491Asn	3	1	43
64	p.Ala1503Thr	28	0	64
65	p.Gly1504Arg	1	-2	11
66	p.Val1508Met	1	1	0
67	p.Ile1536Val	83	3	7
68	p.Met1559Ile	1	1	3
69	p.Gly1565Ser	1	0	0
70	p.Thr1567Met	1	-1	3
71	p.Thr1567Lys	1	-1	3
72	p.Ser1587Phe	1	-1	32
73	p.Ser1605Leu	1	-2	13
74	p.Pro1620Ser	1	-1	19
75	p.His1648Tyr	1	2	42
76	p.Leu1670Ser	3	-2	88
77	p.Thr1676Ser	31	1	76
78	p.Gly1678Ser	5	0	64
79	p.Thr1692Ile	1	-1	2
80	p.Ser1713Phe	1	-2	94
81	p.Pro1755Ser	1	-1	72

Table S2. The 121 Vg variants

Vg variants	nsSNP											
1	p.Ala60Thr	p.Ile489Val	p.Thr522Ile	p.Thr984Met	p.Lys1171Asn	p.Ser1220Asn	p.Arg1284Lys	p.Val1397Met	p.Ala1451Val	p.Ile1536Val		
2	p.Ala60Thr	p.Ile489Val	p.Lys1171Asn	p.Ser1220Asn	p.Ser1235Arg	p.Val1397Met	p.Ile1536Val					
3	p.Pro106Ser	p.Ile489Val	p.Arg1284Lys	p.Ile1398Val	p.Ile1536Val	p.Thr1676Ser						
4	p.Pro126Leu	p.Ile489Val	p.Ser1220Asn	p.Ala1451Val								
5	p.Ile132Met	p.Asn326Thr	p.Ile489Val	p.Thr522Ile	p.Asp608Glu	p.Met639Ile	p.Lys1171Asn	p.Ser1220Asn	p.Arg1284Lys	p.Ile1398Val	p.Ala1451Val	p.Ile1536Val
6	p.Ile132Met	p.Ile489Val	p.Ser803Asn	p.Pro958Ser	p.Arg1174Lys	p.Leu1291Ile	p.Arg1292Ser	p.Val1508Met	p.Ile1536Val	p.Thr1676Ser		
7	p.Ile132Met	p.Ile489Val	p.Ser803Asn	p.Arg1174Lys	p.Arg1284Lys	p.Ile1398Val	p.Ile1536Val					
8	p.Ile132Met	p.Ile489Val	p.Ser803Asn	p.Leu1291Ile	p.Arg1292Ser	p.Ile1398Val	p.Ile1536Val					
9	p.Ile132Met	p.Ile489Val	p.Ser803Asn	p.Leu1291Ile	p.Arg1292Ser	p.Ile1536Val						
10	p.Ile132Met	p.Ile489Val	p.Val834Met	p.Arg1284Lys	p.Ile1398Val	p.Ile1536Val	p.Thr1676Ser					
11	p.Ile132Met	p.Ile489Val	p.Gly1016Asp	p.Ser1220Asn	p.Arg1284Lys	p.Ile1398Val	p.Ile1536Val	p.Thr1676Ser	p.Gly1678Ser			
12	p.Ile132Met	p.Ile489Val	p.Met1159Ile	p.Leu1291Ile	p.Arg1292Ser	p.Ile1398Val	p.Ile1536Val					
13	p.Ile132Met	p.Ile489Val	p.Arg1174Lys	p.Arg1284Lys	p.Ile1398Val	p.Ile1536Val						
14	p.Ile132Met	p.Ile489Val	p.Arg1174Lys	p.Arg1284Lys	p.Ile1536Val							
15	p.Ile132Met	p.Ile489Val	p.Arg1174Lys	p.Leu1291Ile	p.Arg1292Ser	p.Ile1398Val	p.Ile1536Val					
16	p.Ile132Met	p.Ile489Val	p.Arg1284Lys	p.Ile1398Val	p.Ile1536Val							
17	p.Ile132Met	p.Ile489Val	p.Arg1284Lys	p.Ile1398Val	p.Ile1536Val	p.Thr1567Lys	p.Thr1676Ser	p.Gly1678Ser				
18	p.Ile132Met	p.Ile489Val	p.Arg1284Lys	p.Ile1536Val	p.Thr1676Ser							
19	p.Ile132Met	p.Ile489Val	p.Ile1398Val	p.Asp1491Asn	p.Ile1536Val							
20	p.Ile132Met	p.Ile489Val	p.Ile1536Val									
21	p.Gly146Ser	p.Ile489Val	p.Thr1110Ser	p.Ser1220Asn	p.Ala1451Val	p.Ala1503Thr						
22	p.Gly146Ser	p.Ile489Val	p.Val1199Ile	p.Arg1284Lys	p.Ile1398Val	p.Ile1536Val						
23	p.Gly146Ser	p.Ile489Val	p.Arg1284Lys	p.Ala1451Val	p.Ala1503Thr							
24	p.Thr305Met	p.His412Tyr	p.Val419Met	p.Ser471Asn	p.Ile489Val	p.Thr522Ile	p.Asp608Glu	p.Met639Ile	p.Lys1171Asn	p.Ser1220Asn	p.Ala1451Val	p.Ala1503Thr
25	p.Thr305Met	p.His412Tyr	p.Val419Met	p.Ser471Asn	p.Ile489Val	p.Thr522Ile	p.Asp608Glu	p.Met639Ile	p.Arg1284Lys	p.Ile1398Val	p.Ile1536Val	
26	p.Thr305Met	p.Ile489Val	p.Thr1110Ser	p.Ser1220Asn	p.Leu1291Ile	p.Arg1292Ser	p.Ala1451Val	p.Ala1503Thr				
27	p.Asn326Thr	p.Ile489Val	p.Ile640Val	p.Gly1504Arg	p.Leu1670Ser							
28	p.Asn326Thr	p.Ile489Val	p.Ile640Val	p.Leu1670Ser								
29	p.Asn326Thr	p.Ile489Val	p.Thr1110Ser	p.Ser1220Asn	p.Leu1291Ile	p.Arg1292Ser	p.Ala1503Thr					
30	p.Asn326Thr	p.Ile489Val	p.Ser1220Asn									
31	p.Asn326Thr	p.Ile489Val	p.Ser1220Asn	p.Ala1451Val								
32	p.Asn326Thr	p.Ile489Val	p.Ser1220Asn	p.Ala1451Val	p.Ala1503Thr							
33	p.Asn326Thr	p.Ile489Val	p.Ser1220Asn	p.Ala1451Val	p.Ser1605Leu							
34	p.His412Tyr	p.Val419Met	p.Ser467Asn	p.Ile489Val	p.Asn560His	p.Thr1110Ala	p.Ser1235Arg	p.Val1311Met	p.Ile1536Val			
35	p.Ser471Asn	p.Ile489Val	p.Thr522Ile	p.Asp608Glu	p.Met639Ile	p.Arg1284Lys	p.Ile1398Val	p.Ile1536Val				
36	p.Ile489Val	p.Ala509Thr	p.Arg1284Lys	p.Gly1302Glu	p.Ile1398Val	p.Ile1536Val						
37	p.Ile489Val	p.Ala509Thr	p.Arg1284Lys	p.Ile1398Val	p.Ile1536Val							
38	p.Ile489Val	p.Ala509Thr	p.Arg1284Lys	p.Ile1398Val	p.Ile1536Val	p.Thr1676Ser						
39	p.Ile489Val	p.Ala509Thr	p.Arg1284Lys	p.Ile1536Val	p.Thr1676Ser	p.Ser1713Phe						
40	p.Ile489Val	p.Met525Ile	p.Ser1220Asn	p.Ala1451Val	p.Ala1503Thr	p.Ile1536Val						
41	p.Ile489Val	p.Thr594Met	p.Ser1220Asn									
42	p.Ile489Val	p.Leu606Phe	p.Arg1284Lys	p.Ile1398Val	p.Ile1536Val	p.Thr1676Ser						
43	p.Ile489Val	p.Asp626Asn	p.Arg1284Lys	p.Ile1398Val	p.Ile1536Val							
44	p.Ile489Val	p.Asp626Asn	p.Arg1284Lys	p.Ile1398Val	p.Ile1536Val	p.Thr1676Ser						
45	p.Ile489Val	p.Glu642Lys	p.Arg1284Lys	p.Ile1398Val	p.Ile1536Val	p.Thr1676Ser						
46	p.Ile489Val	p.Glu642Lys	p.Arg1284Lys	p.Ala1451Val	p.Ala1503Thr							
47	p.Ile489Val	p.Val661Met	p.Ser1220Asn	p.Ala1451Val	p.Ala1503Thr							
48	p.Ile489Val	p.Ser803Asn	p.Met1159Ile	p.Arg1284Lys	p.Ile1398Val	p.Asp1491Asn	p.Ile1536Val	p.Thr1676Ser				
49	p.Ile489Val	p.Ser803Asn	p.Met1159Ile	p.Leu1291Ile	p.Arg1292Ser	p.Ile1398Val	p.Ile1536Val					
50	p.Ile489Val	p.Ser803Asn	p.Arg1284Lys	p.Ile1398Val	p.Ala1503Thr	p.Ile1536Val						
51	p.Ile489Val	p.Ser803Asn	p.Arg1284Lys	p.Ile1398Val	p.Ile1536Val	p.Thr1676Ser						
52	p.Ile489Val	p.Ser803Asn	p.Leu1291Ile	p.Arg1292Ser	p.Ile1536Val							
53	p.Ile489Val	p.Ser803Asn	p.Ile1398Val	p.Asp1491Asn	p.Ile1536Val							
54	p.Ile489Val	p.Ser803Asn	p.Ile1398Val	p.Ile1536Val	p.Thr1676Ser	p.Gly1678Ser						
55	p.Ile489Val	p.Ser803Asn	p.Ser1220Asn	p.Arg1284Lys	p.Ile1398Val	p.Ile1536Val	p.Pro1755Ser					
56	p.Ile489Val	p.Leu828Ile	p.Arg1284Lys	p.Ile1398Val	p.Ile1536Val	p.Thr1676Ser	p.Gly1678Ser					
57	p.Ile489Val	p.Pro866Ser	p.Thr1110Ser	p.Ser1220Asn	p.Ala1451Val	p.Ala1503Thr						
58	p.Ile489Val	p.Glu906Lys	p.Ser1220Asn	p.Ala1451Val	p.Ala1503Thr							
59	p.Ile489Val	p.Thr939Met	p.Arg1284Lys	p.Ile1398Val	p.Ile1536Val	p.Thr1676Ser						
60	p.Ile489Val	p.Val942Ala	p.Ser1220Asn									
61	p.Ile489Val	p.Ile943Met	p.Arg1284Lys	p.Ile1398Val	p.Ile1536Val							
62	p.Ile489Val	p.Arg969His	p.Arg1284Lys	p.Ile1398Val	p.Ile1536Val	p.Thr1676Ser						
63	p.Ile489Val	p.Thr1013Met	p.Arg1284Lys	p.Ile1398Val	p.Ile1536Val							
64	p.Ile489Val	p.Thr1013Met	p.Arg1284Lys	p.Ile1398Val	p.Ile1536Val	p.Thr1676Ser						
65	p.Ile489Val	p.Gly1016Thr	p.Arg1174Lys	p.Leu1291Ile	p.Arg1292Ser	p.Ile1398Val	p.Ile1536Val					
66	p.Ile489Val	p.Gly1016Thr	p.Arg1284Lys	p.Ile1398Val	p.Ile1536Val							
67	p.Ile489Val	p.Leu1072Phe	p.Thr1110Ser	p.Ser1220Asn	p.Ala1451Val	p.Ala1503Thr						
68	p.Ile489Val	p.Asp1103Tyr	p.Thr1110Ser	p.Ser1220Asn	p.Ala1451Val	p.Ala1503Thr						
69	p.Ile489Val	p.Thr1110Ser	p.Ser1220Asn									
70	p.Ile489Val	p.Thr1110Ser	p.Ser1220Asn	p.Leu1291Ile	p.Arg1292Ser	p.Ala1451Val	p.Ala1503Thr					
71	p.Ile489Val	p.Thr1110Ser	p.Ser1220Asn	p.Ile1398Val	p.Ile1536Val							
72	p.Ile489Val	p.Thr1110Ser	p.Ser1220Asn	p.Ala1451Val								
73	p.Ile489Val	p.Thr1110Ser	p.Ser1220Asn	p.Ala1451Val	p.Ala1503Thr							
74	p.Ile489Val	p.Thr1110Ser	p.Ser1220Asn	p.Ala1503Thr								
75	p.Ile489Val	p.Thr1110Ser	p.Ser1220Asn	p.Ile1536Val	p.Thr1692Ile							
76	p.Ile489Val	p.Thr1110Ser	p.Arg1284Lys	p.Ile1398Val	p.Ala1503Thr	p.Ile1536Val						
77	p.Ile489Val	p.Met1159Ile	p.Leu1291Ile	p.Arg1292Ser	p.Ile1398Val	p.Ile1536Val						
78	p.Ile489Val	p.Met1159Ile	p.Ile1398Val	p.Ile1536Val								
79	p.Ile489Val	p.Arg1174Lys	p.Arg1284Lys	p.Ile1398Val	p.Ile1536Val							
80	p.Ile489Val	p.Arg1174Lys	p.Arg1284Lys	p.Ile1398Val	p.Ile1536Val	p.Thr1676Ser						
81	p.Ile489Val	p.Arg1174Lys	p.Arg1284Lys	p.Ile1536Val								
82	p.Ile489Val	p.Arg1174Lys	p.Arg1284Lys	p.Ile1536Val	p.Thr1676Ser							
83	p.Ile489Val	p.Arg1174Lys	p.Leu1291Ile	p.Arg1292Ser	p.Ile1398Val	p.Ile1536Val						
84	p.Ile489Val	p.Arg1174Lys	p.Leu1291Ile	p.Arg1292Ser	p.Ile1398Val	p.Ile1536Val	p.Met1559Ile					
85	p.Ile489Val	p.Val1193Leu	p.Arg1385Lys	p.Ile1398Val	p.Pro1620Ser							
86	p.Ile489Val	p.Thr1207Ile	p.Arg1284Lys	p.Ile1398Val	p.Ile1536Val	p.Thr1676Ser						
87	p.Ile489Val	p.Ser1220Asn	p.Ala1237Leu	p.Ala1451Val	p.Ala1503Thr							
88	p.Ile489Val	p.Ser1220Asn	p.Arg1284Lys	p.Ile1398Val	p.Ile1536Val							
89	p.Ile489Val	p.Ser1220Asn	p.Leu1291Ile	p.Arg1292Ser								

90	p.Ile489Val	p.Ser1220Asn	p.Leu1291Ile	p.Arg1292Ser	p.Ala1451Val				
91	p.Ile489Val	p.Ser1220Asn	p.Leu1291Ile	p.Arg1292Ser	p.Ala1451Val	p.Ala1503Thr			
92	p.Ile489Val	p.Ser1220Asn	p.Leu1291Ile	p.Arg1292Ser	p.Ala1451Val	p.Ala1503Thr	p.Thr1567Met		
93	p.Ile489Val	p.Ser1220Asn	p.Arg1292Ser						
94	p.Ile489Val	p.Ser1220Asn	p.Arg1292Ser	p.Ile1536Val					
95	p.Ile489Val	p.Ser1220Asn	p.Ile1398Val	p.Ile1536Val	p.Gly1565Ser	p.Thr1676Ser			
96	p.Ile489Val	p.Ser1220Asn	p.Ala1451Val						
97	p.Ile489Val	p.Ser1220Asn	p.Ala1451Val	p.Ala1503Thr					
98	p.Ile489Val	p.Ser1220Asn	p.Ala1451Val	p.Ala1503Thr	p.Ile1536Val				
99	p.Ile489Val	p.Ser1220Asn	p.Ala1451Val	p.Ala1503Thr	p.Thr1676Ser				
100	p.Ile489Val	p.Ser1220Asn	p.Ala1451Val	p.Ile1536Val					
101	p.Ile489Val	p.Arg1284Lys	p.Gly1302Glu	p.Phe1357Val	p.Ile1398Val	p.Ile1536Val			
102	p.Ile489Val	p.Arg1284Lys	p.Val1311Met	p.Ile1398Val	p.Ile1536Val	p.Thr1676Ser			
103	p.Ile489Val	p.Arg1284Lys	p.Ala1386Val	p.Ile1398Val	p.Ile1536Val	p.Ser1587Phe	p.His1648Tyr	p.Thr1676Ser	
104	p.Ile489Val	p.Arg1284Lys	p.Ile1398Val						
105	p.Ile489Val	p.Arg1284Lys	p.Ile1398Val	p.Ala1503Thr	p.Ile1536Val				
106	p.Ile489Val	p.Arg1284Lys	p.Ile1398Val	p.Ala1503Thr	p.Ile1536Val	p.Thr1676Ser			
107	p.Ile489Val	p.Arg1284Lys	p.Ile1398Val	p.Ile1536Val					
108	p.Ile489Val	p.Arg1284Lys	p.Ile1398Val	p.Ile1536Val	p.Thr1676Ser				
109	p.Ile489Val	p.Arg1284Lys	p.Ile1398Val	p.Ile1536Val	p.Thr1676Ser	p.Gly1678Ser			
110	p.Ile489Val	p.Arg1284Lys	p.Ala1451Val	p.Ala1503Thr					
111	p.Ile489Val	p.Arg1284Lys	p.Ile1536Val						
112	p.Ile489Val	p.Arg1284Lys	p.Ile1536Val	p.Thr1676Ser					
113	p.Ile489Val	p.Leu1291Ile	p.Arg1292Ser	p.Ile1398Val	p.Ile1536Val				
114	p.Ile489Val	p.Leu1291Ile	p.Arg1292Ser	p.Ile1398Val	p.Ile1536Val	p.Thr1676Ser			
115	p.Ile489Val	p.Leu1291Ile	p.Arg1292Ser	p.Ile1536Val					
116	p.Ile489Val	p.Leu1291Ile	p.Arg1292Ser	p.Ile1536Val	p.Thr1676Ser				
117	p.Ile489Val	p.Arg1292Ser	p.Ile1398Val	p.Ile1536Val					
118	p.Ile489Val	p.Ile1398Val	p.Ile1536Val						
119	p.Ile489Val	p.Ala1451Val	p.Ala1503Thr						
120	p.Ile1398Val	p.Ile1536Val							
121	p.Ile1398Val	p.Ile1536Val	p.Leu1670Ser						

Table S3. PCR primers, barcodes and PCR plate setup

Primers for *vg* gene	Oligo sequence (5' to 3')	Tm	Oligo size
Forward	AGCCGAATCAAATGCATCGT	58.6	20
Reverse	ACGAAAGAAAGGATTATTGAAAACA	56	25

Primer and Barcodes	Oligo sequence (5' to 3')	Oligo size	Full-length fragment size
F1	AAGAAAGTTGTCGGTGTCTTTGTGAGCCGAATCAAATGCATCGT	44	6296 bp
F2	TCGATTCCGTTTGTAGTCGTCTGTAGCCGAATCAAATGCATCGT	44	6296 bp
F3	GAGTCTTGTGTCCCAGTTACCAGGAGCCGAATCAAATGCATCGT	44	6296 bp
F4	TTCGGATTCTATCGTGTTTCCCTAAGCCGAATCAAATGCATCGT	44	6296 bp
F5	CTTGTCCAGGGTTTGTGTAACCTTAGCCGAATCAAATGCATCGT	44	6296 bp
F6	TTCTCGCAAAGGCAGAAAGTAGTCAGCCGAATCAAATGCATCGT	44	6296 bp
F7	GTGTTACCGTGGGAATGAATCCTTAGCCGAATCAAATGCATCGT	44	6296 bp
F8	TTCAGGGAACAAACCAAGTTACGTAGCCGAATCAAATGCATCGT	44	6296 bp
R1	AGAACGACTTCCATACTCGTGTGAACGAAAGAAAGGATTATTGAAAACA	49	6296 bp
R2	AACGAGTCTCTTGGGACCCATAGAACGAAAGAAAGGATTATTGAAAACA	49	6296 bp
R3	AGGTCTACCTCGCTAACACCACTGACGAAAGAAAGGATTATTGAAAACA	49	6296 bp
R4	CGTCAACTGACAGTGGTTCGTACTACGAAAGAAAGGATTATTGAAAACA	49	6296 bp
R5	ACCCTCCAGGAAAGTACCTCTGATACGAAAGAAAGGATTATTGAAAACA	49	6296 bp
R6	CCAAACCCAACAACCTAGATAGGCACGAAAGAAAGGATTATTGAAAACA	49	6296 bp
R7	GTTCCTCGTGCAGTGTCAAGAGATACGAAAGAAAGGATTATTGAAAACA	49	6296 bp
R8	TTGCGTCCTGTTACGAGAACTCATACGAAAGAAAGGATTATTGAAAACA	49	6296 bp
R9	GAGCCTCTCATTGTCCGTTCTCTAACGAAAGAAAGGATTATTGAAAACA	49	6296 bp
R10	ACCACTGCCATGTATCAAAGTACGACGAAAGAAAGGATTATTGAAAACA	49	6296 bp
R11	CTTACTACCCAGTGAACCTCCTCGACGAAAGAAAGGATTATTGAAAACA	49	6296 bp
R12	GCATAGTTCTGCATGATGGGTTAGACGAAAGAAAGGATTATTGAAAACA	49	6296 bp

Plate setup

	1	2	3	4	5	6	7	8	9	10	11	12
A	F1R1	F1R2	F1R3	F1R4	F1R5	F1R6	F1R7	F1R8	F1R9	F1R10	F1R11	F1R12
B	F2R1	F2R2	F2R3	F2R4	F2R5	F2R6	F2R7	F2R8	F2R9	F2R10	F2R11	F2R12
C	F3R1	F3R2	F3R3	F3R4	F3R5	F3R6	F3R7	F3R8	F3R9	F3R10	F3R11	F3R12
D	F4R1	F4R2	F4R3	F4R4	F4R5	F4R6	F4R7	F4R8	F4R9	F4R10	F4R11	F4R12
E	F5R1	F5R2	F5R3	F5R4	F5R5	F5R6	F5R7	F5R8	F5R9	F5R10	F5R11	F5R12
F	F6R1	F6R2	F6R3	F6R4	F6R5	F6R6	F6R7	F6R8	F6R9	F6R10	F6R11	F6R12
G	F7R1	F7R2	F7R3	F7R4	F7R5	F7R6	F7R7	F7R8	F7R9	F7R10	F7R11	F7R12
H	F8R1	F8R2	F8R3	F8R4	F8R5	F8R6	F8R7	F8R8	F8R9	F8R10	F8R11	F8R12

Paper IV

How honey bee Vitellogenin holds lipid cargo: A role for the C-terminal

Vilde Leipart[1*], Øyvind Halskau[2], Gro V. Amdam[1,3]

[1]Faculty of Environmental Sciences and Natural Resource Management, Norwegian University of Life Sciences, Ås, Norway

[2]Department of Biological Sciences, University of Bergen, Bergen, Norway

[3]School of Life Sciences, Arizona State University, Tempe, AZ, United States

*** Correspondence:**
Vilde Leipart
vilde.leipart@nmbu.no

Keywords: Honey bee Vitellogenin$_1$, AlphaFold$_2$, Lipid binding$_3$, Oxidative stress$_4$, C-terminal region$_5$

Word count: 2208, Figures: 4

Running title: How Vitellogenin holds lipid cargo

Article type: Hypothesis and theory

Abstract

Vitellogenin (Vg) is a phylogenetically broad glycolipophosphoprotein. A major function of this protein is holding lipid cargo for storage and transportation. Vg has been extensively studied in honey bees (*Apis mellifera*) due to additional functions in social traits. Using AlphaFold and EM contour mapping, we recently described the protein structure of honey bee Vg. The full-length protein structure reveals a large hydrophobic lipid binding site and a well-defined fold at the C-terminal region. Now, we outline a shielding mechanism that allows the C-terminal region of Vg to cover a large hydrophobic area exposed in the all-atom model. We propose that this C-terminal movement influences lipid molecules' uptake, transport, and delivery. The mechanism requires elasticity in the Vg lipid core as described for homologous proteins in the large lipid transfer protein (LLTP) superfamily to which Vg belongs. Honey bee Vg has, additionally, several structural arrangements that we interpret as beneficial for the functional flexibility of the C-terminal region. The mechanism proposed here may be relevant for the Vg molecules of many species.

1 Introduction

Vitellogenin (Vg) is the most ancient yolk precursor protein in animals [1]. It is well-known for transporting lipids and other nutrients to developing embryos but is recognized for additional roles in innate immunity and complex behavior [2-4]. As is true of all large lipid transfer protein (LLTP) superfamily members, Vg contains a hydrophobic lipid binding domain that defines a cavity structure. Superfamily members have similar structural landscapes in their binding cavities despite variations in amino acid sequence [5, 6]. Specifically, LLTPs across taxa have amphipathic α-helical

repeats surrounding several amphipathic β-sheets [7, 8]. The β-sheets provide a hydrophobic lining to the interior cavity. The connective loops and the flexibility of the α-helices and β-sheets provide elasticity to expand or compress the cavity during uptake or delivery of the lipid cargo. Beyond the α-helical and β-sheet structures and features, the characteristics and functions of LLTPs differ [1]. In terms of the Vg proteins, they all contain a well-conserved N-terminal domain, composed of a β-barrel and α-helical subdomain [9, 10]. The remaining domains are otherwise variable but usually include one or several domains of unknown function (DUF) and may include a von Willebrand factor (vWF) domain.

For most animals, the detailed structural composition of Vg remains undescribed. The majority of insight into its structure is derived through template-based modeling from the only experimentally-solved Vg structure, lamprey (*Ichthyomyzon unicuspis*) [11]. The determination of this crystal structure was a central contribution to understanding Vg proteins, but its partial sequence coverage and distant homology toward many other Vg molecules limit its usefulness for *in silico* structural predictions. Also, few template options for isolated Vg domains or subdomains exist. The massive size and the complex domain organization of Vg proteins, which includes their sizeable hydrophobic (lipid binding) cavity and extensive post-translational modifications [12], have probably contributed to the relative lack of detailed structural insights.

In parallel with these challenges, intriguing data have emerged on the non-reproductive roles of Vg [13, 14]. To date, these have been most studied in honey bees (*Apis mellifera*) [15], in which the protein influences social behavior, oxidative stress resilience, and cell-based and trans-generational immunity, in addition to its traditional role in yolk formation [4, 16-18]. Recent progress made possible by DeepMind's AlphaFold, a neural network for structure prediction [19], allowed us to generate a full-length structure prediction of honey bee Vg with high confidence [20]. This structure prediction reveals the N-terminal domain folding around the lipid binding cavity, as expected for a Vg protein. Surprisingly, four structural units build up the cavity (Figure 1, reproduced here from Leipart *et al.* 2021a, see ref list [21]). Two β-sheets (β1 and β2, also referred to as C- and A-sheet, respectively) comprise the so-called DUF1943 domain. A third β-sheet (β3) and the vWF domain follow the DUF1943, completing the circular or funnel-like shape of the lipid cavity. The domains and subdomains are interconnected. For example, the longer β2 sheet extends toward the N-terminal domain, and the α-helical subdomain covers and scaffolds the DUF1943 structural elements, which reduce the lipid cavity's exposure to the solvent. The C-terminal constitutes a small structural fold connected to the vWF domain through a presumably flexible linker. The folded C-terminal region does not appear to be in direct contact with the lipid binding site but instead appears at the flank of the large Vg structure (see [20] for more details).

Our previous study fitted the AlphaFold prediction into a low-resolution EM map [20]. However, the C-terminal position was not compatible with EM density barriers. Therefore, we proposed an alternative position of the C-terminal above the lipid binding site, as the fitting revealed available space at the opening of the lipid cavity (marked in Figure 2, see panel C). But what is the C-terminal region of Vg possibly doing there? In the current article, we assess this structural organization's feasibility and possible functional relevance.

2 C-terminal flexibility

To our knowledge, the tertiary structure of the C-terminal region of Vg proteins has not been solved. The structural fold of the C-terminal region is composed of four short β-strands, an α-helix, and two longer β-strands. Three disulfide bridges connect the short and longer β-strands (Figure 2A). The

AlphaFold database (https://alphafold.ebi.ac.uk/), a collaboration between DeepMind and EMBL's European Bioinformatics Institute, contains 21 predicted proteomes [22], including that of *Caenorhabditis elegans*, which has six Vg-encoding sequences (*vit* gene 1 to 6) [23]. Superimposing the C-terminal region (amino acid 1530 to 1613) in *C. elegans* Vg-2 with our prediction of the C-terminal in honey bee Vg (amino acid 1688 to 1770) shows an almost identical fold (RMSD = 1.035 [24], Figure 2B), although Vg-2 has only two disulfide bridges. These predictions have low confidence for the loop connecting the C-terminal to the vWF domain, indicating a disordered region for both animals [19, 25]. This disorder suggests flexibility of the loop region linking the C-terminal to the entrance of the lipid binding cavity. The position of the C-terminal region differs between the bee and worm predictions in reference to the lipid binding site (Figure 2C). However, AlphaFold states that predicting positions for extended linkers or isolated structural elements may be less reliable due to the frequent lack of inter-residue contacts [19].

The vWF and C-terminal regions are often described as single C-terminal domains in Vg proteins. Our prediction of honey bee Vg, in contrast, describes two separate and distinct structural folds. The vWF domain is packed tightly in the lipid binding site, while the C-terminal is a separate solvent-exposed region (Figure 2C and 2E). We proposed a possible zinc-coordination site that resides between the two adjacent disulfide bridges in the C-terminal region (Leipart *et al.* (2021b) in manuscript, see ref list [26]). Similar coordination sites of four cysteine residues are often found in redox switches [27, 28] that can cause conformational changes: During oxidative stress, zinc is released, resulting in oxidative folding and creation of disulfide bridges [29]. A similar mechanism could be relevant for folding at the C-terminal region of honey bee Vg.

Taken together, structural analysis suggests that the C-terminal of honey bee Vg can be flexible and take part in conformational changes such as domain repositioning.

3 Exposed lipid binding site

The proposed C-terminal repositioning is in line with complementary electrostatic forces on the C-terminal region and lipid binding cavity. Insect Vg proteins have conserved positively charged residues at the C-terminal [12]. These reside at the α-helix, creating a net positive surface charge (Figure 2D). The lipid binding site has a negatively charged center on β3 and the vWF domain (Figure 2E). Additionally, the wide opening of the lipid binding site, exposing the hydrophobic cavity to the solvent, is costly in terms of entropy. Shielding the opening would aid the stability and solubility of Vg, particularly during transport or storage of large lipid cargo. We propose that the C-terminal region provides this shielding. As illustrated in Figure 3, the "closed" position resembles the contour of the EM map [20], while the AlphaFold prediction would represent the "open" (flanking) position. Similar conformational shifts, including an "open" and "closed" state, have previously been reported for LLTPs [1, 8]. The precise position of the C-terminal region in our "open" state, however, is uncertain due to reduced inter-residue contacts, as noted above (Figure 2C). A more likely scenario is perhaps a position closer to the Vg structure.

4 Expansion and compression

LLTPs can bind up to hundreds of lipid molecules [30, 31]. Their packing requires interior stability to withstand differences in pressure on the lipid cavity lining and support the elasticity of the lipid core to handle the changing lipid loads. For honey bee Vg, the β-sheet network and the identification of five disulfide bridges distributed between β3 and the vWF domain may contribute to a stable interior (Figure 4A). The lipid binding site of microsomal triglyceride transfer protein (MTP),

another LLTP member, has a narrower lipid binding cavity compared to lamprey Vg [32]. MTP has a flexible junction to accommodate lipid binding, despite lower lipid binding capacity. We believe that honey bee Vg might require greater flexibility than MTP due to the larger cavity volume.

Interestingly, MTP and lamprey Vg have conserved disulfide bridges in their respective α-helical subdomains, which are reported to create stability for the subdomain and the encompassing β1 and β2 [11, 32]. The disulfide bridges are not conserved in honey bee Vg, suggesting lower fold stability and a greater potential for flexibility. In addition, honey bee Vg has an insect-specific loop region between α-helices 9 and 10 [16], which is consistent with an elastic subdomain arrangement. The insect-specific loop aligns with the opening between β1 and β2 (Figure 4B). The β-sheets interact with the α-helical subdomain through hydrophobic and electrostatic interactions; this is important for maintaining a stable fold. Moreover, β1 and β2 are connected through a long α-helix, creating a triangle-like shape of the lipid binding cavity (Figure 4B). The connecting α-helix is reported as stabilizing for the tertiary structure for other LLTPs [11, 32]. In honey bee Vg, the α-helix is longer (19 amino acids) than in MTP (11 amino acids), lamprey Vg (13 amino acids), and *C. elegans* Vg-2 (11 amino acids followed by a loop region and an additional α-helix of nine amino acids, folded in parallel with β2). The long and continuous α-helix in honey bee Vg creates a larger lipid binding triangle.

To summarize: the α-helical subdomain surrounding the lipid binding cavity in honey bee Vg contains regions that can provide elasticity during expansion and compression (Figure 4C). The subdomain and the connecting α-helix also support an ability to carry very large lipid loads. We note, however, that the expansion of the hydrophobic core of a lipid binding cavity can result in a less soluble surface [32, 33]. In this context, we propose that the C-terminal region provides a cover that increases the solubility of Vg, possibly shifting deeper into the cavity in response to increasing loads. Similar shielding has been reported for MTP, which interacts with its β-subunit, protein disulfide isomerase (PDI): PDI binds to the α-helical subdomain and shields the lipid binding site opening [32]. Interestingly, MTP lacks a C-terminal region homologous to that of honey bee Vg. In contrast, honey bee Vg is a single subunit protein that does not pair with a PDI homolog.

5 Post-translational modifications (PTMs)

Extensive protein modifications, such as ubiquitinoylation and sumoylation, are not observed for honey bee Vg [34, 35], but the protein is known to be phosphorylated and glycosylated [34] (extent and exact positions of these PTMs are unknown). There are well-documented examples of phosphorylation and glycosylation providing increased solubility [36-39], resistance to disordered elements against proteases [34, 35, 40], modulation of the conformational propensities of flexible elements [41-43], and steric hindrances or complementarity for ligand binding or domain reorganization [44, 45]. Methylation or acetylation of Vg could also conceivably be involved. These modifications are found on lysine and arginine and are associated with epigenetic control but tend to decrease solubility [46]. We acknowledge that PTMs at the folded part of the C-terminal, at its flexible linker, or in the putative binding site could affect conformational propensities of the whole region, regulate or support the correct and timely insertion of the folded element, and protect hydrophobic surfaces from the solvent [36]. However, further discussion of these possibilities requires more actual and accurate data on Vg PTMs.

6 Concluding remarks

At this point, we arrive at an explanation for how the lipid binding site of honey bee Vg may be optimized for large lipid cargo. This optimization includes a large lipid binding triangle, an ability to flex and compress to load and unload the lipid cargo, and utilization of the C-terminal to shield the exposed hydrophobic surface. Our model includes predictions about an "open" vs. "closed" protein configuration. This model sets the stage for performing molecular simulation, protein docking, and experimental dynamical studies to test our speculations. We further note that thorough mapping of potential PTM sites using both *in silico* and experimental approaches is required for a complete understanding of the molecular mechanisms. Taken together, these considerations provide a roadmap for future studies of how honey bee Vg holds its lipid cargo. We also hope they are inspirational and relevant for research on the Vg molecules of other species.

7 Conflict of Interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

8 Author Contributions

VL wrote the paper and made the figures. ØH and GVA contributed to and edited the manuscript.

9 Acknowledgments

The authors acknowledge The Research Council of Norway grant number 262137 for funding toward running costs and positions and BioCat (RCN grant number 249023) for travel grants and conferences support.

10 Data Availability Statement

The datasets analyzed for this study can be found in the PDB at https://www.rcsb.org/ (PDB-ID: 1LSH and 6I7S) and in AlphaFold database https://alphafold.ebi.ac.uk/ (UniProt ID: P05690). The AlphaFold prediction of honey bee Vg is supplemented in our recent publication [20].

Contribution to the field statement (max 200 words)

Honey bee Vitellogenin transports and stores lipids, which is an important function for developing and surviving long winter seasons. This protein has also been linked to providing immune responses during oxidative stress and for long-lived workers. We have limited detailed knowledge of how Vitellogenin behaves at the molecular level to support these functional roles but made progress when describing the tertiary protein structure using AlphaFold and fitted the predicted model into an EM density map. The protein structure reveals a structural fold at the C-terminal, which, we propose, could potentially function during lipid uptake, storage, and delivery. This mechanism could also be related to insect immunity. Our theory is relevant for a range of egg-laying species and is consistent with known functions of similar lipid transfer proteins.

11 Figures

Figure 1 (Reproduced here from Leipart *et al.* 2021a [21])**.** Illustration of the honey bee Vg structure. Vg consists of the N-terminal domain (ND) comprised of two subdomains: β-barrel (yellow) and α-

helical (α-h, green), and a lipid binding site (blue), the vWF domain (vWF, cyan) and a C-terminal (C-term, magenta). The orange zigzag line shows the proteolytic cleavage site on the polyserine linker in ND. The green plus signs next to the α-helical subdomain illustrate the net positive surface charge. Three β-sheets (β1, β2, and β3) build up the lipid binding site. DUF1943 is defined by β-sheets 1 and 2, while the third sheet is considered part of the lipid binding site. We refer to this structural region as the lipid binding site throughout the article. The C-terminal has been demonstrated to be flexible, as illustrated here. We show the interacting or binding units recognized by honey bee Vg to the right, colored according to the interacting domain or subdomain. We use this coloring scheme throughout the article.

Figure 2. A) The C-terminal region of honey bee Vg is composed of an α-helix (cyan) and four short and two longer β-strands (magenta), connected by three disulfide bridges (yellow sticks, black arrows). B) The C-terminal region from *C. elegans* Vg-2 (orange) superimposed honey bee Vg (the same colors as in panel A) and has two disulfide bridges (red sticks), while the third, as seen in honey bee Vg, is missing (black arrow). C) Protein surface representation of full-length honey bee Vg with the N-terminal domain colored gray, the lipid binding site including the vWF domain colored blue, and the C-terminal colored magenta. We show the EM map density barriers as a grid representation around the surface. The black arrow points to the available density above the lipid binding site. Honey bee Vg is turned 90° about the y-axis compared to the presented *C. elegans* Vg-2, shown from the same angle and colored the same, except the C-terminal is colored orange. D) The positively charged residues (blue sticks) on the α-helix (cyan) contribute to a positively charged surface region. E) The negatively charged surface of β3 and the vWF domain are shown. In panels D and E, the electrostatic charges are calculated using the APBS plugin in PyMol.

Figure 3. We suggest that the flexibility loop (red) connecting the C-terminal region (magenta) to the vWF domain (cyan), in addition to the electrostatic forces (shown in Figure 2D), contributes to a conformational change in Vg. In an "open" conformation, the C-terminal region is flanking on the side of the lipid binding site (blue) and α-helical subdomain (green). However, when a shielding of the lipid binding site is necessary, for example, during storage or transport of lipid molecules, the loop region is flexible so that the C-terminal region can be positioned over the lipid binding site. The position is likely to contribute to a more soluble protein.

Figure 4. A) The β3 (blue) contains two disulfide bridges (black arrows), while the vWF domain (cyan) contains three (black arrows). The disulfide bridges contribute stability to the binding site. B) The α-helical subdomain (green) wraps around β1 and β2 (blue). The insect-specific loop (orange) aligns with an opening between the β-sheets. β1 and β2 are connected through a long α-helix (light pink), creating a triangle-like shape of the cavity. C) When Vg unloads lipid molecules, the lipid binding site is compressed. Vg requires elasticity to expand the lipid binding cavity during the loading and storage of many lipid molecules. We speculate that the insect-specific loop (orange) and the long α-helix (light pink) add flexibility, as illustrated here (double-pointed arrows).

12 References

1. Smolenaars MMW, Madsen O, Rodenburg KW, Van der Horst DJ. Molecular diversity and evolution of the large lipid transfer protein superfamily^{s⃞}. Journal of Lipid Research. 2007;48(3):489-502.

2. Sun C, Zhang S. Immune-Relevant and Antioxidant Activities of Vitellogenin and Yolk Proteins in Fish. Nutrients. 2015;7(10):8818-29.

3. Amdam GV, Csondes A, Fondrk MK, Page RE, Jr. Complex social behaviour derived from maternal reproductive traits. Nature. 2006;439(7072):76-8.

4. Salmela H, Amdam GV, Freitak D. Transfer of Immunity from Mother to Offspring Is Mediated via Egg-Yolk Protein Vitellogenin. PLoS pathogens. 2015;11(7):e1005015.

5. Shelness GS, Ledford AS. Evolution and mechanism of apolipoprotein B-containing lipoprotein assembly. Current Opinion in Lipidology. 2005;16(3).

6. Babin PJ, Bogerd J, Kooiman FP, Van Marrewijk WJA, Van der Horst DJ. Apolipophorin II/I, Apolipoprotein B, Vitellogenin, and Microsomal Triglyceride Transfer Protein Genes Are Derived from a Common Ancestor. Journal of Molecular Evolution. 1999;49(1):150-60.

7. Van der Horst DJ, Van Hoof D, Van Marrewijk WJA, Rodenburg KW. Alternative lipid mobilization: The insect shuttle system. Molecular and cellular biochemistry. 2002;239(1):113-9.

8. Wang L, Walsh MT, Small DM. Apolipoprotein B is conformationally flexible but anchored at a triolein/water interface: A possible model for lipoprotein surfaces. Proceedings of the National Academy of Sciences. 2006;103(18):6871-6.

9. Li A, Sadasivam M, Ding JL. Receptor-ligand interaction between vitellogenin receptor (VtgR) and vitellogenin (Vtg), implications on low density lipoprotein receptor and apolipoprotein B/E. The first three ligand-binding repeats of VtgR interact with the amino-terminal region of Vtg. The Journal of biological chemistry. 2003;278(5):2799-806.

10. Roth Z, Weil S, Aflalo ED, Manor R, Sagi A, Khalaila I. Identification of Receptor-Interacting Regions of Vitellogenin within Evolutionarily Conserved β-Sheet Structures by Using a Peptide Array. ChemBioChem. 2013;14(9):1116-22.

11. Thompson JR, Banaszak LJ. Lipid-protein interactions in lipovitellin. Biochemistry. 2002;41(30):9398-409.

12. Tufail M, Takeda M. Molecular characteristics of insect vitellogenins. Journal of insect physiology. 2008;54(12):1447-58.

13. Corona M, Libbrecht R, Wurm Y, Riba-Grognuz O, Studer RA, Keller L. Vitellogenin underwent subfunctionalization to acquire caste and behavioral specific expression in the harvester ant Pogonomyrmex barbatus. PLoS genetics. 2013;9(8):e1003730.

14. Kohlmeier P, Alleman AR, Libbrecht R, Foitzik S, Feldmeyer B. Gene expression is more strongly associated with behavioural specialization than with age or fertility in ant workers. Molecular ecology. 2019;28(3):658-70.

15. Pan ML, Bell WJ, Telfer WH. Vitellogenic Blood Protein Synthesis by Insect Fat Body. Science. 1969;165(3891):393.

16. Havukainen H, Munch D, Baumann A, Zhong S, Halskau O, Krogsgaard M, et al. Vitellogenin recognizes cell damage through membrane binding and shields living cells from reactive oxygen species. The Journal of biological chemistry. 2013;288(39):28369-81.

17. Seehuus SC, Norberg K, Gimsa U, Krekling T, Amdam GV. Reproductive protein protects functionally sterile honey bee workers from oxidative stress. Proceedings of the National Academy of Sciences of the United States of America. 2006;103(4):962-7.

18. Amdam GV, Simoes ZL, Hagen A, Norberg K, Schroder K, Mikkelsen O, et al. Hormonal control of the yolk precursor vitellogenin regulates immune function and longevity in honeybees. Experimental gerontology. 2004;39(5):767-73.

19. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021;596(7873):583-9.

20. Leipart V, Montserrat-Canals M, Cunha ES, Luecke H, Herrero-Galán E, Halskau Ø, et al. Structure prediction of honey bee vitellogenin: a multi-domain protein important for insect immunity. FEBS Open Bio. 2021;In press(In press).

21. Leipart V, Ludvigsen J, Kent M, Sandve S, To T-H, Árnyasi M, et al. Identification of 121 variants of honey bee Vitellogenin protein sequence with structural differences at functional sites. 2021a.

22. Varadi M, Anyango S, Deshpande M, Nair S, Natassia C, Yordanova G, et al. AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. Nucleic Acids Res. 2021.

23. Perez MF, Lehner B. Vitellogenins - Yolk Gene Function and Regulation in Caenorhabditis elegans. Frontiers in Physiology. 2019;10(1067).

24. Schrodinger L. The PyMOL Molecular Graphics System, Version 1.8. 2015.

25. Mariani V, Biasini M, Barbato A, Schwede T. lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests. Bioinformatics (Oxford, England). 2013;29(21):2722-8.

26. Leipart V, Enger Ø, Turcu DC, Dobrovolska O, Drabløs F, Halskau Ø, et al. Where Honey Bee Vitellogenin may Bind Zn2+-Ions. 2021b.

27. Pace NJ, Weerapana E. Zinc-binding cysteines: diverse functions and structural motifs. Biomolecules. 2014;4(2):419-34.

28. Ilbert M, Graf PC, Jakob U. Zinc center as redox switch--new function for an old motif. Antioxidants & redox signaling. 2006;8(5-6):835-46.

29. Ruddock LW, Klappa P. Oxidative stress: Protein folding with a novel redox switch. Current Biology. 1999;9(11):R400-R2.

30. Segrest JP, Jones MK, De Loof H, Dashti N. Structure of apolipoprotein B-100 in low density lipoproteins. J Lipid Res. 2001;42(9):1346-67.

31. Hevonoja T, Pentikäinen MO, Hyvönen MT, Kovanen PT, Ala-Korpela M. Structure of low density lipoprotein (LDL) particles: basis for understanding molecular changes in modified LDL. Biochimica et biophysica acta. 2000;1488(3):189-210.

32. Biterova EI, Isupov MN, Keegan RM, Lebedev AA, Sohail AA, Liaqat I, et al. The crystal structure of human microsomal triglyceride transfer protein. Proceedings of the National Academy of Sciences. 2019;116(35):17251-60.

33. Ptak-Kaczor M, Banach M, Stapor K, Fabian P, Konieczny L, Roterman I. Solubility and Aggregation of Selected Proteins Interpreted on the Basis of Hydrophobicity Distribution. Int J Mol Sci. 2021;22(9):5002.

34. Havukainen H, Halskau O, Skjaerven L, Smedal B, Amdam GV. Deconstructing honeybee vitellogenin: novel 40 kDa fragment assigned to its N terminus. The Journal of experimental biology. 2011;214(Pt 4):582-92.

35. Havukainen H, Underhaug J, Wolschin F, Amdam G, Halskau O. A vitellogenin polyserine cleavage site: highly disordered conformation protected from proteolysis by phosphorylation. The Journal of experimental biology. 2012;215(Pt 11):1837-46.

36. Srikanth B, Vaidya MM, Kalraiya RD. O-GlcNAcylation determines the solubility, filament organization, and stability of keratins 8 and 18. The Journal of biological chemistry. 2010;285(44):34062-71.

37. Broyard C, Gaucheron F. Modifications of structures and functions of caseins: a scientific and technological challenge. Dairy Science & Technology. 2015;95(6):831-62.

38. Darewicz M, Dziuba J, Mioduszewska H. Some physico-chemical properties and structural changes of bovine β-casein upon glycation. Food / Nahrung. 1998;42(03-04):213-4.

39. Darewicz M, Dziuba J, Mioduszewska H, Minkiewicz P. Modulation of physico-chemical properties of bovine b-casein by nonenzymatic glycation associated with enzymatic dephosphorylation. Acta Alimentaria. 1999;28(4):339-54.

40. Niu C, Luo H, Shi P, Huang H, Wang Y, Yang P, et al. N-Glycosylation Improves the Pepsin Resistance of Histidine Acid Phosphatase Phytases by Enhancing Their Stability at Acidic pHs and Reducing Pepsin's Accessibility to Its Cleavage Sites. Appl Environ Microbiol. 2015;82(4):1004-14.

41. Fraser JA, Vojtesek B, Hupp TR. A Novel p53 Phosphorylation Site within the MDM2 Ubiquitination Signal I. PHOSPHORYLATION AT SER269 IN VIVO IS LINKED TO INACTIVATION OF p53 FUNCTION. Journal of Biological Chemistry. 2010;285(48):37762-72.

42. He EB, Yan GH, Zhang J, Wang J, Li WF. Effects of phosphorylation on the intrinsic propensity of backbone conformations of serine/threonine. Journal of Biological Physics. 2016;42(2):247-58.

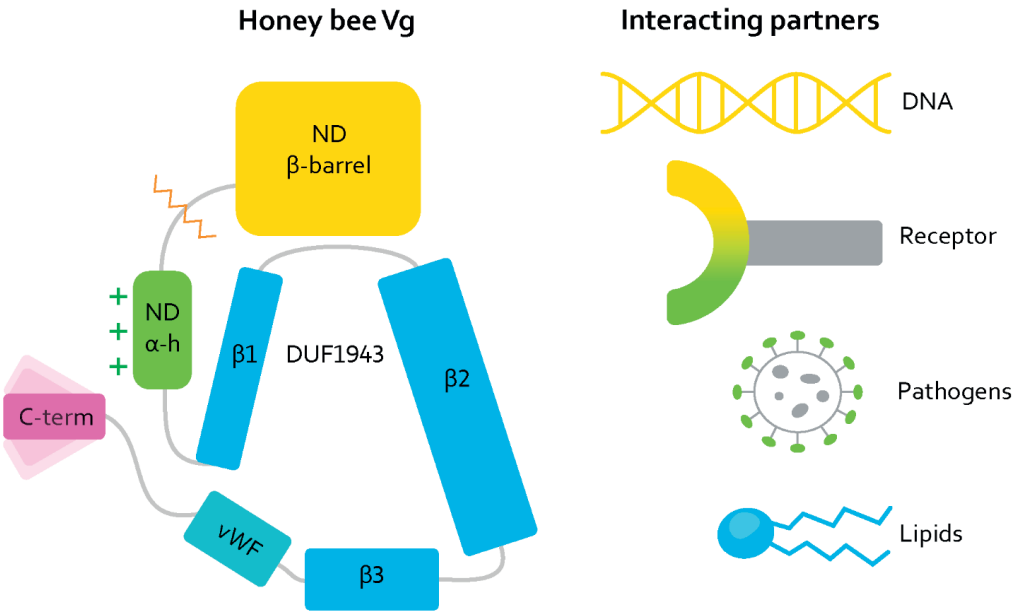
43. Kurotani A, Sakurai T. In Silico Analysis of Correlations between Protein Disorder and Post-Translational Modifications in Algae. Int J Mol Sci. 2015;16(8):19812-35.

44. Shipley JM, Grubb JH, Sly WS. The role of glycosylation and phosphorylation in the expression of active human beta-glucuronidase. The Journal of biological chemistry. 1993;268(16):12193-8.

45. Dean AM, Koshland DE, Jr. Electrostatic and steric contributions to regulation at the active site of isocitrate dehydrogenase. Science. 1990;249(4972):1044-6.

46. Ritchie TJ, Macdonald SJF, Pickett SD. Insights into the impact of N- and O-methylation on aqueous solubility and lipophilicity using matched molecular pair analysis. MedChemComm. 2015;6(10):1787-97.

Figure 1


Honey bee Vg
ND
β-barrel
ND
α-h
β1
DUF1943
β2
C-term
vWF
β3
Interacting partners
DNA
Receptor
Pathogens
Lipids

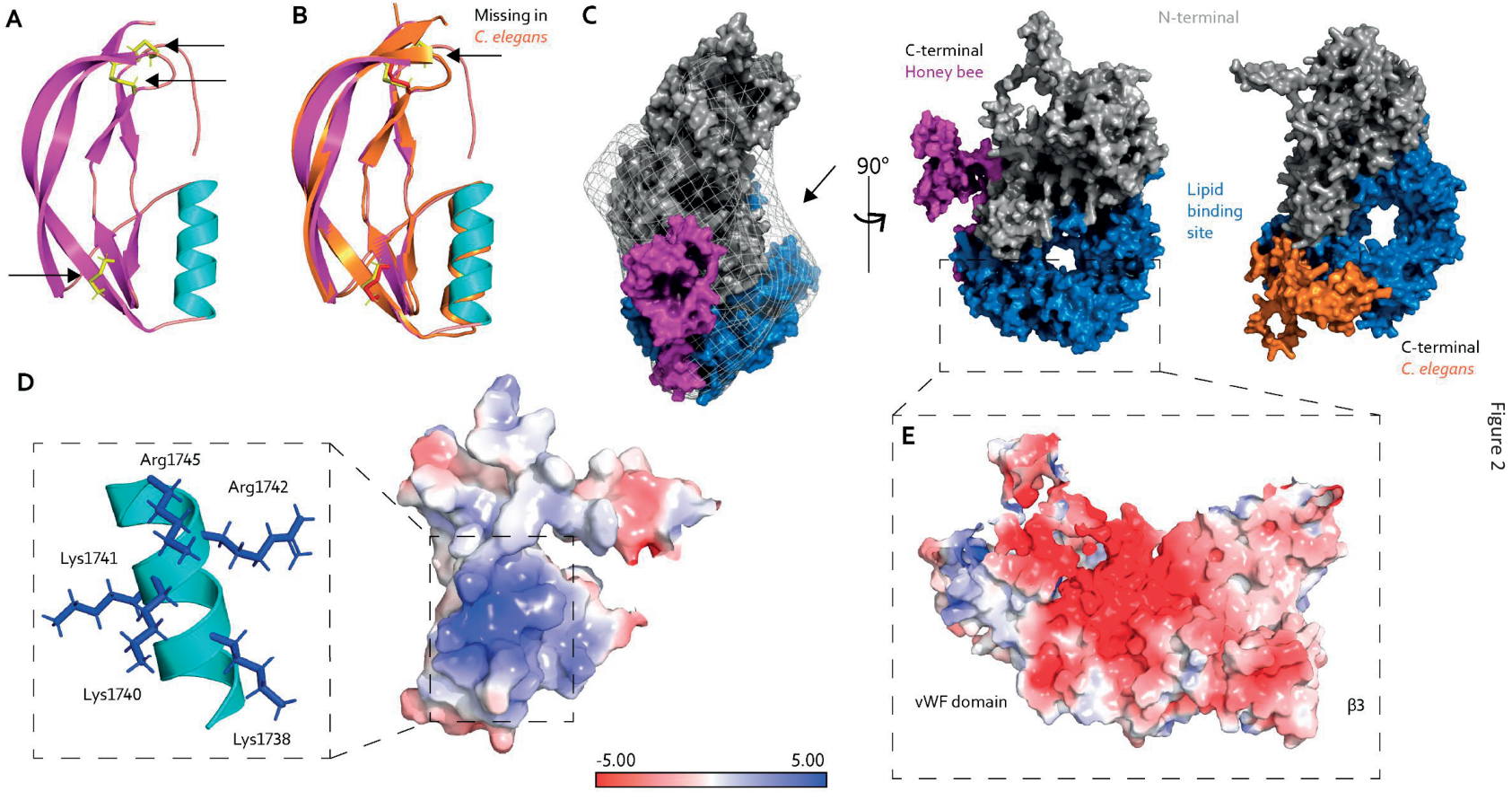
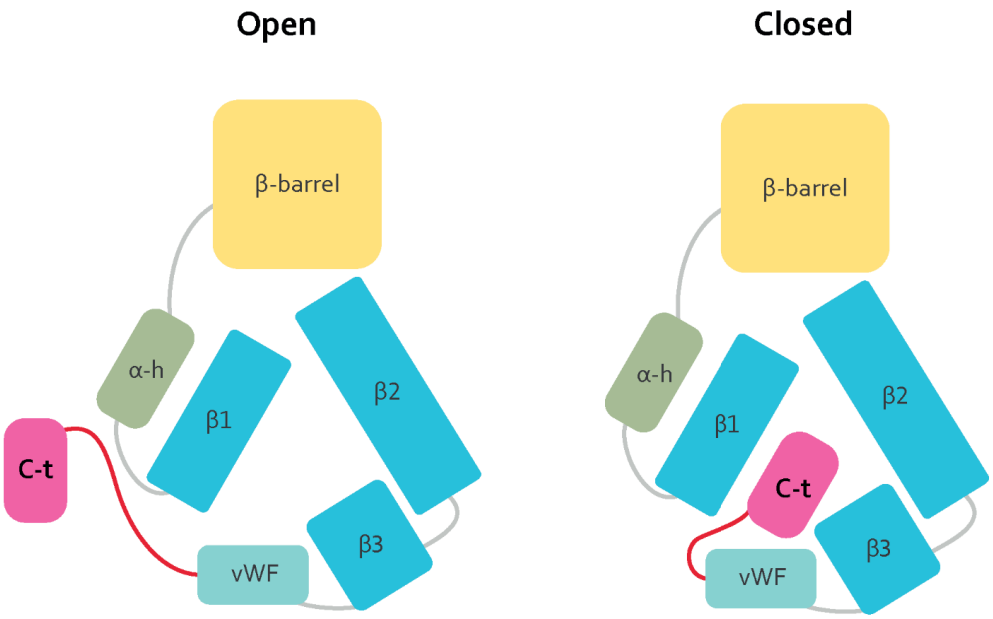
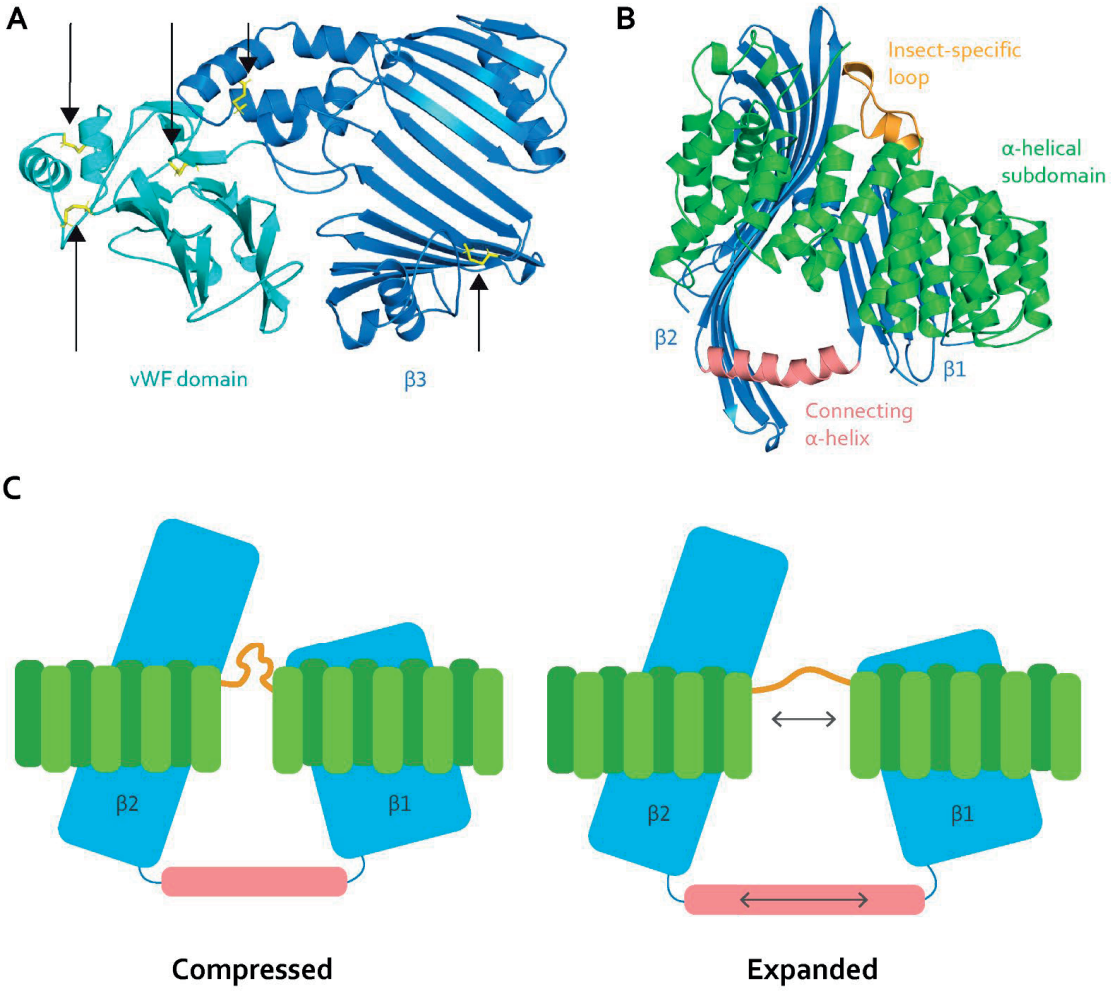

A
B
Missing in
C. elegans
C
90°
N-terminal
C-terminal
Honey bee
Lipid
binding
site
C-terminal
C. elegans
D
Arg1745
Arg1742
Lys1741
Lys1740
Lys1738
-5.00
5.00
E
vWF domain
β3

Figure 2

Figure 3


Open
β-barrel
α-h
β1
β2
C-t
β3
vWF
Closed
β-barrel
α-h
β1
β2
C-t
β3
vWF


A
vWF domain
β3
B
Insect-specific loop
α-helical subdomain
β2
β1
Connecting α-helix
C
β2
β1
β2
β1
Compressed
Expanded

ISBN: 978-82-575-1885-1
ISSN: 1894-6402

Norwegian University
of Life Sciences

Postboks 5003
NO-1432 Ås, Norway
+47 67 23 00 00
www.nmbu.no

112715 / ANDVORDGRAFISK.NO