The Molecular Basis of Sex: Linking Yeast to Human

Willie J. Swanson,[1] Jan E. Aagaard,[1] Victor D. Vacquier,[2] Magnus Monné,†,[3] Hamed Sadat Al Hosseini,[3] and Luca Jovine*,[3]

[1]Department of Genome Sciences, University of Washington

[2]Marine Biology Research Division, Scripps Institution of Oceanography, University of California San Diego

[3]Department of Biosciences and Nutrition and Center for Biosciences, Karolinska Institutet, Huddinge, Stockholm, Sweden

†Present address: Department of Chemistry, University of Basilicata, Potenza, Italy

***Corresponding author:** E-mail: luca.jovine@ki.se.

Associate editor: Michael Nachman

Abstract

Species-specific recognition between egg and sperm, a crucial event that marks the beginning of fertilization in multicellular organisms, mirrors the binding between haploid cells of opposite mating type in unicellular eukaryotes such as yeast. However, as implied by the lack of sequence similarity between sperm-binding regions of invertebrate and vertebrate egg coat proteins, these interactions are thought to rely on completely different molecular entities. Here, we argue that these recognition systems are, in fact, related: despite being separated by 0.6–1 billion years of evolution, functionally essential domains of a mollusc sperm receptor and a yeast mating protein adopt the same 3D fold as egg zona pellucida proteins mediating the binding between gametes in humans.

Key words: fertilization, egg–sperm interaction, egg coat, zona pellucida domain, yeast mating, protein structure.

Introduction

Like their counterparts in the vitelline (egg) envelope (VE) of other vertebrates as well as invertebrates such as the mollusc abalone (Aagaard et al. 2006), mammalian zona pellucida (ZP) subunits ZP1–4 assemble into the nascent egg coat using a common C-terminal "ZP domain" (Bork and Sander 1992; Jovine et al. 2002). This conserved polymerization module consists of two domains, ZP-N and ZP-C (Jovine et al. 2004; Wassarman and Litscher 2008) (fig. 1). Recent crystallographic studies of sperm receptor ZP3 revealed that the ZP-N domain defines a new subtype of the immunoglobulin (Ig) superfamily of proteins, characterized by two disulfide bonds with invariant 1–4, 2–3 connectivity, a unique E′ strand implicated in polymerization, and a conserved Tyr residue in strand F (Monné et al. 2008). Moreover, they showed that—despite having a very different sequence—ZP-C also adopts a β-sandwich fold with the same basic topology as ZP-N, suggesting that the two moieties of the ZP module might have originated by duplication of a single ancestral Ig-like domain (Han et al. 2010). Additional copies of ZP-N are found within the N-terminal region of some vertebrate ZP/VE components (Callebaut et al. 2007; Monné et al. 2008) (fig. 1), where—as in the case of mammalian ZP2—they can bind sperm (Tsubamoto et al. 1999) and regulate gamete recognition (Bleil et al. 1981; Gahlay et al. 2010). Notably, repeated sequences located within the N-terminal region of abalone VE subunits VERL and VEZP14 (fig. 1) are also thought to bind sperm (Swanson and Vacquier 1997; Aagaard et al. 2010), but because of very low-sequence similarity, no connection could be made between molluscan and mammalian repeats.

Molluscan Egg Coat Protein Repeats Adopt the ZP-N Fold of Mammalian ZP Proteins

Because rapid sequence divergence could mask potential relationships between reproductive proteins from evolutionary distant species (Swanson and Vacquier 2002), we performed a fold recognition analysis using sequence–structure comparison in FUGUE (Shi et al. 2001). Molluscan repeat sequences were threaded against a local copy of the HOMSTRAD database of structural profiles (de Bakker et al. 2001) that included an entry for the canonical ZP-N domain of VERL (Galindo et al. 2002), generated on the basis of the crystal structure of ZP3 ZP-N (Monné et al. 2008; Han et al. 2010). A high-confidence match was found between the sequence of VERL repeat 10 and the Ig-like fold variant specific to ZP-N (supplementary fig. S1, Supplementary Material online). An homology model of repeat 10 created on the basis of this sequence–structure alignment is structurally sound and exposes Asn side chains expected to be glycosylated (Swanson and Vacquier 1997) (fig. 2). Moreover, it can be readily extended to all other VERL repeats, as well as the VERL-like repeat of VEZP14 (Aagaard et al. 2010), because of significant sequence similarity (supplementary figs. S1 and S2*a*–*b*, Supplementary Material online). This suggests that all Cys within the repeat array of VERL are engaged in ZP-N-specific Cys_{1-4}, Cys_{2-3} disulfide bonds, with the exception of C201 and C294 (supplementary fig. S3*a*, Supplementary Material online). These additional Cys, located in repeat 2, may therefore be responsible for forming the intermolecular disulfides that have been shown to mediate homodimerization of VERL (Swanson and Vacquier 1997). This prediction was experimentally confirmed by loss of VERL dimerization upon

Letter

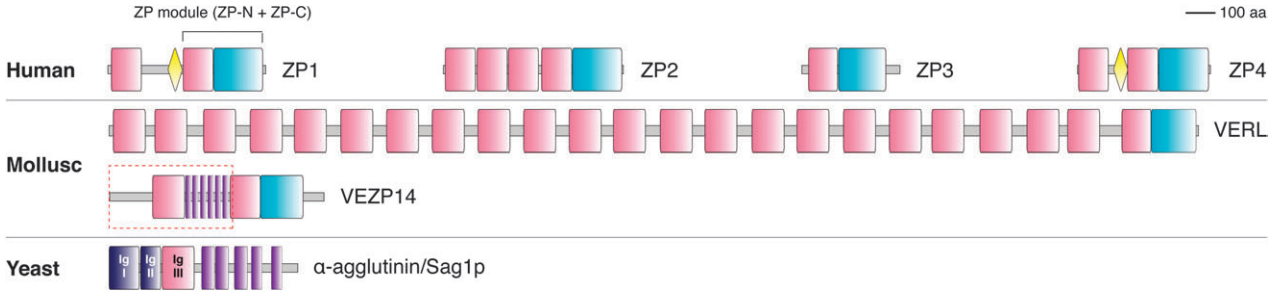


Fig. 1. Domain architecture of human ZP subunits, mollusc VERL and VEZP14, and yeast α-agglutinin/Sag1p. Pink: ZP-N domain; cyan: ZP-C domain; yellow: trefoil domain; violet: S/T-rich sequence repeat; dark blue: Sag1p Ig-like domains I, II; and dashed red box: SMART Pfam:Candida_ALS match in VEZP14.

introduction of C201D, C294S substitutions within a repeat 1–4 fragment secreted by insect cells (supplementary fig. S3*b*, Supplementary Material online). Considering that all other abalone VE subunits also contain a ZP module (Aagaard et al. 2010), this data collectively suggest that, as in mammals, the ZP-N domain accounts for the majority of the structure of the molluscan egg coat.

A Protein Domain Essential for Mating in Yeast Also Shares ZP-N-Specific Features

Domain analysis with SMART (Letunic et al. 2009) indicates that the N-terminal region of VEZP14, which contains the protein's VERL-like ZP-N repeat (Aagaard et al. 2010) (fig. 1), is in turn related to yeast agglutinin-like proteins (supplementary fig. S1, Supplementary Material online). These highly glycosylated adhesion molecules mediate extracellular interactions, such as mating in *Saccharomyces cerevisiae* and host invasion in *Candida albicans*, mainly using the last of three N-terminal Ig domains (Ig III; fig. 1) (Dranginis et al. 2007). Although Ig III was initially modeled on the basis of Ig Kol—the best template available at the time—(de Nobel et al. 1996), FUGUE threading of Ig III sequences against the current protein fold database identifies the ZP-N Ig subtype as the top hit for this domain (supplementary fig. S1, Supplementary Material online), a prediction supported by I-TASSER (Roy et al. 2010). Most importantly, our ZP-N-based model of *S. cerevisiae* mating protein α-agglutinin/Sag1p Ig III is not only physically realistic (supplementary fig. S4, Supplementary Material online) but also completely consistent with a large amount of available biochemical data (fig. 3). Specifically, the model agrees with circular dichroism spectroscopy studies of the N-terminal half of α-agglutinin (Chen et al. 1995); accounts for the experimentally determined disulfide bond between C202 and C300 (Chen et al. 1995), which corresponds to the canonical Cys_{1-4} disulfide of the ZP-N fold; predicts burial of C227 and C256 ($Cys_{2,3}$) (Chen et al. 1995); and is consistent with exposure of residues that were shown experimentally to be accessible to proteases (Chen et al. 1995), glycosylated (Chen et al. 1995), or involved in binding to a-agglutinin (Cappellaro et al. 1991; de Nobel et al. 1996). Moreover, Y270 of α-agglutinin is positioned in correspondence of the conserved F-strand Tyr that lies next to invariant Cys_4 within the E'-F-G extension of the ZP-N fold (Monné et al. 2008; Han et al. 2010). Taken together, these considerations suggest that this particular type of Ig-like domain may not be restricted to multicellular eukaryotes as previously thought but also

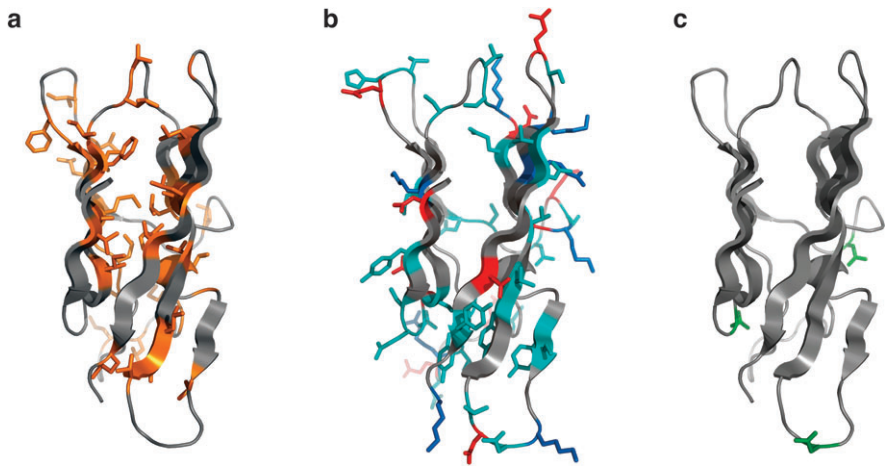


Fig. 2. Homology model of abalone VERL repeat 10 ZP-N domain, shown in side view using a cartoon representation with relevant residues depicted as sticks. The model is consistent with burial of hydrophobic residues (*a*; brown), exposure of positively charged, negatively charged, and polar side chains (*b*; blue, red, and cyan, respectively) and exposure of consensus sites for N-glycosylation (*c*; green).

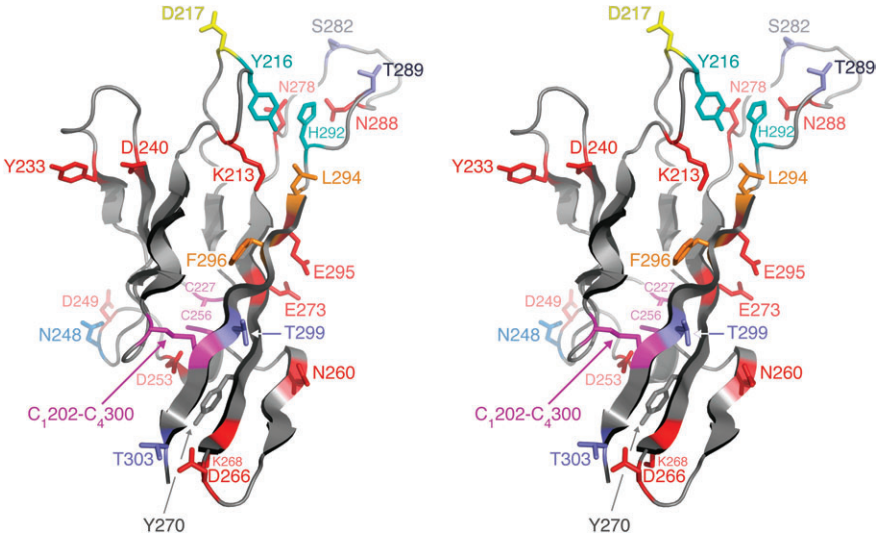


Fig. 3. Stereograph of the model of *Saccharomyces cerevisiae* α-agglutinin/Sag1p Ig III. Conserved Cys residues: magenta; protease-accessible residues: red; N-glycosylated residues: light blue; O-glycosylated residues: violet; residues that are essential, important, or play a minor role in binding to a-agglutinin: cyan, orange, and yellow, respectively; and Y270: dark gray.

exist in specialized extracellular proteins of yeast that play key roles in mating (*S. cerevisiae* Sag1p) or adhesion to human tissues and biofilm formation (*C. albicans* Als1p and Als3p).

Conclusions and Functional Implications

Although the rapid evolution of reproductive proteins makes the direct comparison of their sequences generally uninformative, relationships between these molecules could in principle be recognized by identifying suitably intermediate sequences that connect them (Park et al. 1997) or relying on conservation of common higher-order structural features. In this report, we have combined these approaches to detect unexpected structural similarities between reproductive proteins from both vertebrates and invertebrates, as well as yeast mating proteins. These findings suggest that some of the molecular features that regulate sexual interaction may be much more conserved during evolution than previously appreciated (fig. 1 and supplementary fig. S5, Supplementary Material Online). In this regard, it is particularly remarkable that α-agglutinin amino acids essential for interaction with *a*-agglutinin (de Nobel et al. 1996) (fig. 4*c*) are positioned so that they are exposed on the same face of the ZP-N fold as ZP2 and VERL residues implicated in sperm binding (fig. 4*a*–*b*). Moreover, specific adherence of *Candida* to human endothelial and epithelial cells requires Als1p Ig III amino acids centered around V285 (Fu et al. 1998; Loza et al. 2004; Sheppard et al. 2004; Dranginis et al. 2007), a residue that is also predicted to be exposed on the same region of the ZP-N domain (fig. 4*c*). Because threading per se simply estimates the likeness that a known 3D fold is adopted by a given

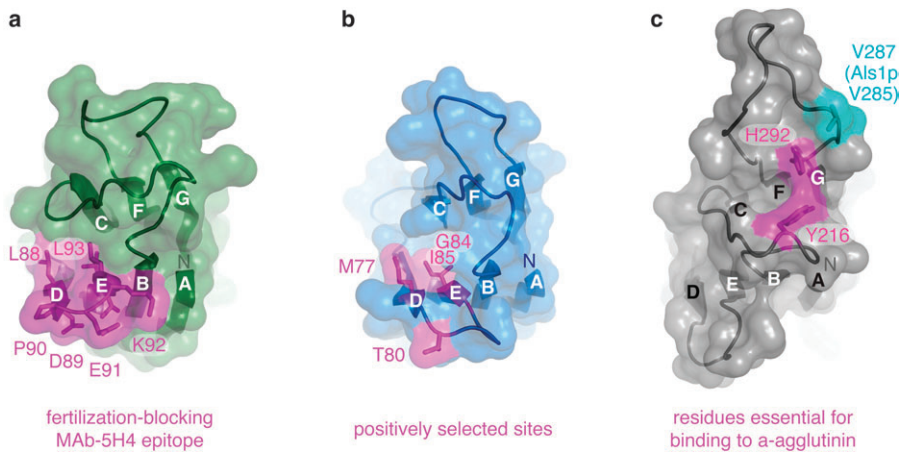


Fig. 4. Mapping of functionally important residues on the homology models of human ZP2 repeat 1 ZP-N (*a*), mollusc VERL repeat 1 ZP-N (*b*), and yeast α-agglutinin/Sag1p Ig III ZP-N (*c*). *Saccharomyces cerevisiae* Sag1p Ig III residue V287, corresponding to functionally crucial residue V285 of *Candida albicans* Als1p, is also indicated in (*c*). A top view of the proteins is shown, with N termini and Ig fold β-strands marked by uppercase letters.

sequence profile, it does not directly address the point of whether the corresponding proteins share common ancestry or just adopt a similar tertiary structure. Although this is currently unfeasible due to lack of abalone genome sequences and absence of significant conserved syntheny between yeast and human, future identification of related sequences from additional lineages may make it possible to assess whether the similarity that we have uncovered reflects direct homology or is instead the result of convergent evolution. Nevertheless, considering the widespread distribution of the Ig fold, it is striking that reproductive protein sequences from mollusc and yeast specifically match its ZP-N variant, repeats of which had previously only been detected in vertebrate egg coat proteins.

Supplementary Material

Supplementary methods and figures S1–S5 are available at *Molecular Biology and Evolution* online (http://mbe.oxfordjournals.org/).

Acknowledgments

This work was supported by National Institute of Health (NIH) Grants HD057974, HD042563, and HD 054631 (W.J.S.); NIH Grant HD12986 (V.D.V.); the Center for Biosciences, Swedish Research Council grant 2009-5193, an EMBO Young Investigator award, and the European Research Council under the European Union's Seventh Framework Program (FP7/2007–2013)/ERC grant agreement 260759 (L.J.). We thank Tsukasa Matsuda, Stevan Springer, and members of our laboratories for comments and discussions.

References

Aagaard JE, Vacquier VD, MacCoss MJ, Swanson WJ. 2010. ZP domain proteins in the abalone egg coat include a paralog of VERL under positive selection that binds lysin and 18-kDa sperm proteins. *Mol Biol Evol*. 27:193–203.

Aagaard JE, Yi X, MacCoss MJ, Swanson WJ. 2006. Rapidly evolving zona pellucida domain proteins are a major component of the vitelline envelope of abalone eggs. *Proc Natl Acad Sci U S A*. 103:17302–17307.

Bleil JD, Beall CF, Wassarman PM. 1981. Mammalian sperm-egg interaction: fertilization of mouse eggs triggers modification of the major zona pellucida glycoprotein, ZP2. *Dev Biol*. 86:189–197.

Bork P, Sander C. 1992. A large domain common to sperm receptors (Zp2 and Zp3) and TGF-β type III receptor. *FEBS Lett*. 300: 237–240.

Callebaut I, Mornon JP, Monget P. 2007. Isolated ZP-N domains constitute the N-terminal extensions of Zona Pellucida proteins. *Bioinformatics* 23:1871–1874.

Cappellaro C, Hauser K, Mrsa V, Watzele M, Watzele G, Gruber C, Tanner W. 1991. *Saccharomyces cerevisiae* a- and α-agglutinin: characterization of their molecular interaction. *EMBO J*. 10: 4081–4088.

Chen MH, Shen ZM, Bobin S, Kahn PC, Lipke PN. 1995. Structure of *Saccharomyces cerevisiae* α-agglutinin. Evidence for a yeast cell wall protein with multiple immunoglobulin-like domains with atypical disulfides. *J Biol Chem*. 270:26168–26177.

de Bakker PI, Bateman A, Burke DF, Miguel RN, Mizuguchi K, Shi J, Shirai H, Blundell TL. 2001. HOMSTRAD: adding sequence information to structure-based alignments of homologous protein families. *Bioinformatics* 17:748–749.

de Nobel H, Lipke PN, Kurjan J. 1996. Identification of a ligand-binding site in an immunoglobulin fold domain of the *Saccharomyces cerevisiae* adhesion protein α-agglutinin. *Mol Biol Cell*. 7:143–153.

Dranginis AM, Rauceo JM, Coronado JE, Lipke PN. 2007. A biochemical guide to yeast adhesins: glycoproteins for social and antisocial occasions. *Microbiol Mol Biol Rev*. 71:282–294.

Fu Y, Rieg G, Fonzi WA, Belanger PH, Edwards JEJ, Filler SG. 1998. Expression of the *Candida albicans* gene *ALS1* in *Saccharomyces cerevisiae* induces adherence to endothelial and epithelial cells. *Infect Immun*. 66:1783–1786.

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

Galindo BE, Moy GW, Swanson WJ, Vacquier VD. 2002. Full-length sequence of VERL, the egg vitelline envelope receptor for abalone sperm lysin. *Gene* 288:111–117.

Han L, Monné M, Okumura H, Schwend T, Cherry AL, Flot D, Matsuda T, Jovine L. 2010. Insights into egg coat assembly and egg-sperm interaction from the X-ray structure of full-length ZP3. *Cell* 143:404–415.

Jovine L, Qi H, Williams Z, Litscher E, Wassarman PM. 2002. The ZP domain is a conserved module for polymerization of extracellular proteins. *Nat Cell Biol*. 4:457–461.

Jovine L, Qi H, Williams Z, Litscher ES, Wassarman PM. 2004. A duplicated motif controls assembly of zona pellucida domain proteins. *Proc Natl Acad Sci U S A*. 101:5922–5927.

Letunic I, Doerks T, Bork P. 2009. SMART 6: recent updates and new developments. *Nucleic Acids Res*. 37:D229–D232.

Loza L, Fu Y, Ibrahim AS, Sheppard DC, Filler SG, Edwards JEJ. 2004. Functional analysis of the *Candida albicans ALS1* gene product. *Yeast* 21:473–482.

Monné M, Han L, Schwend T, Burendahl S, Jovine L. 2008. Crystal structure of the ZP-N domain of ZP3 reveals the core fold of animal egg coats. *Nature* 456:653–657.

Park J, Teichmann SA, Hubbard T, Chothia C. 1997. Intermediate sequences increase the detection of homology between sequences. *J Mol Biol*. 273:349–354.

Roy A, Kucukural A, Zhang Y. 2010. I-TASSER: a unified platform for automated protein structure and function prediction. *Nat Protoc*. 5:725–738.

Sheppard DC, Yeaman MR, Welch WH, Phan QT, Fu Y, Ibrahim AS, Filler SG, Zhang M, Waring AJ, Edwards JEJ. 2004. Functional and structural diversity in the Als protein family of *Candida albicans*. *J Biol Chem*. 279:30480–30489.

Shi J, Blundell TL, Mizuguchi K. 2001. FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. *J Mol Biol*. 310:243–257.

Swanson WJ, Vacquier VD. 1997. The abalone egg vitelline envelope receptor for sperm lysin is a giant multivalent molecule. *Proc Natl Acad Sci U S A*. 94:6724–6729.

Swanson WJ, Vacquier VD. 2002. The rapid evolution of reproductive proteins. *Nat Rev Genet*. 3:137–144.

Tsubamoto H, Hasegawa A, Nakata Y, Naito S, Yamasaki N, Koyama K. 1999. Expression of recombinant human zona pellucida protein 2 and its binding capacity to spermatozoa. *Biol Reprod*. 61:1649–1654.

Wassarman PM, Litscher ES. 2008. Mammalian fertilization: the egg's multifunctional zona pellucida. *Int J Dev Biol*. 52:665–676.

Supplementary Material for

The Molecular Basis of Sex: Linking Yeast to Human

Willie J. Swanson, Jan E. Aagaard, Victor D. Vacquier,

Magnus Monné, Hamed Sadat Al Hosseini and Luca Jovine

This file includes:

Supplementary Methods

Figures S1-S5

Supplementary References

Supplementary Methods

Bioinformatics Analysis

Homology models were generated by iterative alignment with MODELLER (Fiser and Sali, 2003), manual optimization in Coot (Emsley et al., 2010) and regularization with phenix.pdbtools in PHENIX (Adams et al., 2010) or molecular dynamics simulation in water with the knowledge-based YASARA2 force field of YASARA Structure (Krieger et al., 2009). They were validated using DOPE scores (Shen and Sali, 2006), Verify3D (Eisenberg et al., 1997) and MolProbity (Chen et al., 2010).

Sequence-structure threading was performed using FUGUE (Shi et al., 2001) and a local installation of the HOMSTRAD database (de Bakker et al., 2001) that consisted of 11848 folds and included a structure-based alignment for the canonical ZP-N domain of *H. rufescens* VERL (residues 3420-3516 of sequence database entry AAL50827; (Galindo et al., 2002). This belongs to a ZP module (AAL50827 residues 3420-3666) that has significant sequence similarity to that of well-characterized ZP domain proteins such as *Drosophila* Dusky (Roch et al., 2003), NompA (Chung et al., 2001), Piopio (Jazwinska et al., 2003) and Dumpy (Wilkin et al., 2000) (PSI-BLAST E-values at iteration 5: 5e-49, 1e-45, 9e-42 and 5e-35, respectively); and is identified as a bona-fide ZP domain by both InterProScan (E-value 0.0075) and FUGUE (Z-score relative to the full-length chicken ZP3 crystal structure (PDB ID 3NK3): 7.76). To exclude the possibility that adding an homology model-derived profile to HOMSTRAD would bias the outcome of FUGUE searches, we tested profiles based on incorrect alignments of the canonical VERL ZP-N sequence with either the N-terminal region of *E. coli* maltose-binding protein (residues 1-97 of PDB ID 1ANF; mixed α/β fold) or pig PSP-I (residues 4-101 of PDB ID 1SPP; all-β fold); none of these negative control profiles produced a significant match.

The results obtained from the FUGUE analysis of α-agglutinin/Sag1p Ig III were confirmed by submitting to the I-TASSER/Zhang server (Roy et al., 2010) the Ig III sequence by itself (run 1), or together with our alignment to ZP3 ZP-N and a distance constraint corresponding to the experimentally determined C202-C300 disulfide bond (Chen et al., 1995) (run 2). The Ig III model obtained from run 2 had a significantly higher scores (C-score -0.63, TM-score 0.63±0.13) than that the best model produced from run 1 (C-score -2.32, TM-score 0.44±0.14), indicating that aligning the Ig III sequence to ZP3 ZP-N as shown in fig. S5 results in a model that is more accurate than any other model that I-TASSER can produce on its own, based on all the structures within its database.

Domain searches were performed using SMART 6.0 (Letunic et al., 2009), which also scans input sequences using Pfam profile hidden Markov models (Finn et al., 2010). Repeat analysis was carried out with Radar (Heger and Holm, 2000). At each stage of the analysis outlined in fig. S1, several homologous sequences were evaluated in parallel for any given protein; the accession numbers reported in the figure refer to the sequences that gave the highest scores. Figs. 2, 3 and 4 were generated with PyMOL (DeLano, 2002), using secondary structure assignments produced by DSSPcont (Carter et al., 2003) and β-Spider (Parisien and Major, 2005). The structure-based multiple sequence alignment shown in fig. S5 was assembled using STRAP (Gille and Frömmel, 2001).

Analysis of VERL Conserved Cys Mutation

cDNA fragments encoding C-terminally histidine-tagged red abalone VERL repeats 1-4 (VERL_R01-04_6His), as well as double-mutant VERL_R01-04_6His C201D+C294S, were cloned into vector pIEx-5 (Novagen) and transiently expressed in Sf9 insect cells. After 72 hr of culture, cell medium was concentrated and used for immunoblotting experiments using

monoclonal anti-5His (QIAGEN; 1:1,000) and peroxidase-conjugate AffiniPure goat anti-mouse IgG (Jackson ImmunoResearch Lab, Inc.; 1:10,000).

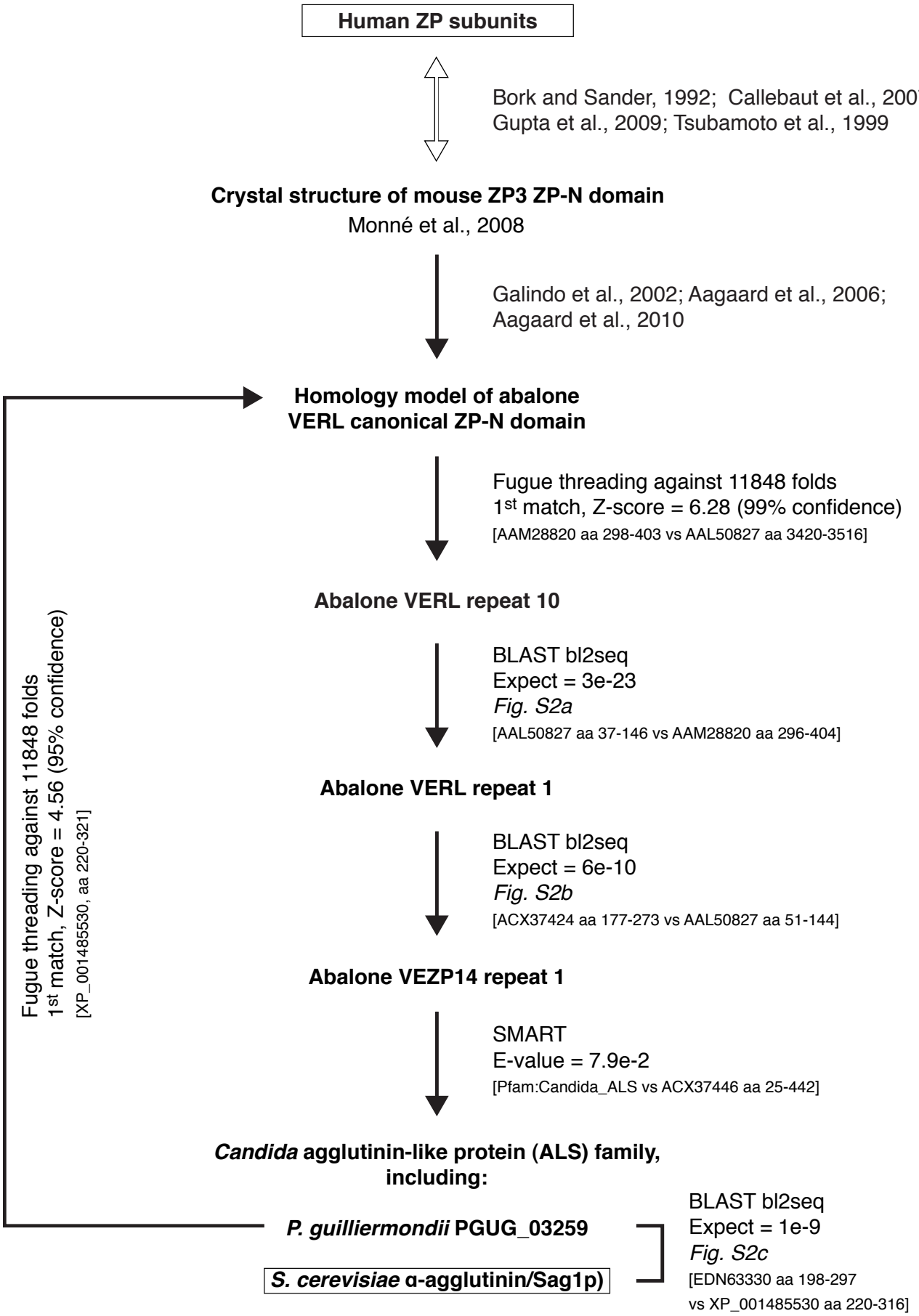


Fig. S1. Summary of bioinformatic analysis results linking abalone VERL, VEZP14 and yeast α-agglutinin/Sag1p to mammalian ZP proteins involved in sperm binding. Solid arrows indicate how results were serially built on one another, starting from the experimental 3D structure of ZP3 ZP-N.

a

```
Score = 88.2 bits (217),  Expect = 3e-23, Method: Compositional matrix adjust.
Identities = 46/110 (42%), Positives = 66/110 (60%), Gaps = 1/110 (0%)

AAM28820        296  IDWDVYCSQNESIPAKFISL-LTSKDQAVEKTEINCSNGLVPITQESGINMMLIQYTRND  354
                     +D  + CS ++S  A  IS  +T K   ++  +I C NG + +T+  GINM+ I Y +
AAL50827        37   LDLTLVCSDDKSKQATLISYPVTFKGHVIKDMQIFCKNGWMQMTRGRGINMIRIHYPQTY  96

AAM28820        355  LLDSPGMCVFWGPYSVPKNDTVVLYTVTARLKWSEGPPTNLSIECYMPKS  404
                        PG CVF GPYSVP ND++ +Y V+  L WS+G PT  S+EC + KS
AAL50827        97   TSVVPGACVFRGPYSVPTNDSIEMYNVSVALLWSDGTPTYESLECNVTKS  146
```

b

```
Score = 43.5 bits (101),  Expect = 6e-10, Method: Compositional matrix adjust.
Identities = 28/98 (29%), Positives = 47/98 (48%), Gaps = 5/98 (5%)

AAL50827        51   LTLVCSDDKSKQATLISYPVTFKGHVIKDMQIFCKNGWMQMT--RGRGINMIRIHYPQTY  96
                       + C+      A + S    ++   I   Q+ C N  + M         ++RI+ P
ACX37424        177  QIISCNKI---GAVVHSEKSVYRACAITYSQMICANSTVPMILPEKSTFYILRIYLPSNL  224

AAL50827        97   TSVVP--GACVFRGPYSVPTNDSIEMYNVSVALLWSDGTPTYESLECNVT  144
                      ++     AC+  GPY V    S++ YNV+V++LW+DG  T   + C V+
ACX37424        225  KNIKKDINACIMDGPYKVNGTSSLK-YNVTVSMLWNDGRATGAEVHCLVS  273
```

c

```
Score = 42.7 bits (99),  Expect = 1e-09, Method: Compositional matrix adjust.
Identities = 29/101 (29%), Positives = 45/101 (45%), Gaps = 5/101 (4%)

XP_001485530    220  LGTSCP---FETQSGTLGITFKNGGPPLQCSTLTASFSNQFNDWYFPQTASSLSYSVLC-  275
                     LG  CP   F   +  +     N    L CS++    SN FNDW+FPQ+ +  +  V C
EDN63330        198  LGMYCPNGYFLGGTEKIDYDSSNNNVDLDCSSVQVYSSNDFNDWWFPQSYNDTNADVTCF  257

XP_001485530    276  SLNLVMVTFAGVPAGSRPFFNIFASLLPVGRNSAVYTLKYR--CGNGP  316
                       NL +     +  G   + N   S LP   N+  + L+++  C +
EDN63330        258  GSNLWITLDEKLYDGEMLWVNALQS-LPANVNTIDHALEFQYTCLDTI  297
```

Fig. S2. bl2seq alignments. (**a**) Alignment of abalone VERL repeat 10 (database entry AAM28820) and VERL repeat 1 (AAL50827). (**b**) Alignment of abalone VERL repeat 1 (AAL50827) and VEZP14 repeat 1 (ACX37424). (**c**) Alignment of *P. guilliermondii* PGUG_03259 Ig III (XP_001485530) and *S. cerevisiae* α-agglutinin/Sag1p Ig III (EDN63330). Residues that were input to bl2seq but not retained in the program output are shown in light blue; invariant ZP-N domain $Cys_{1\text{-}4}$ residues and conserved F-strand Tyr residue are highlighted in magenta and grey, respectively. Alignment statistics and sequence numbers refer to residues retained in the bl2seq output (black).

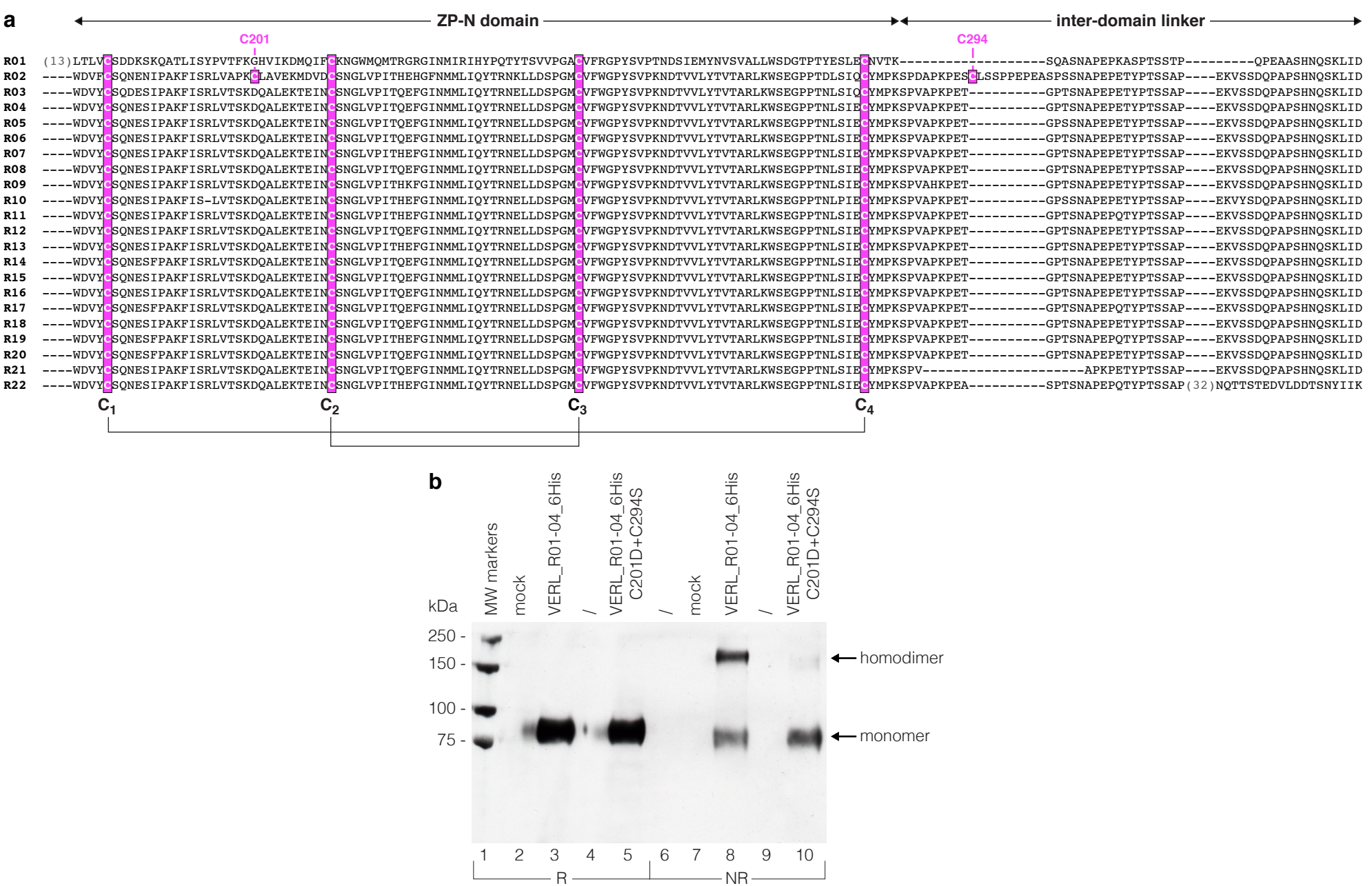
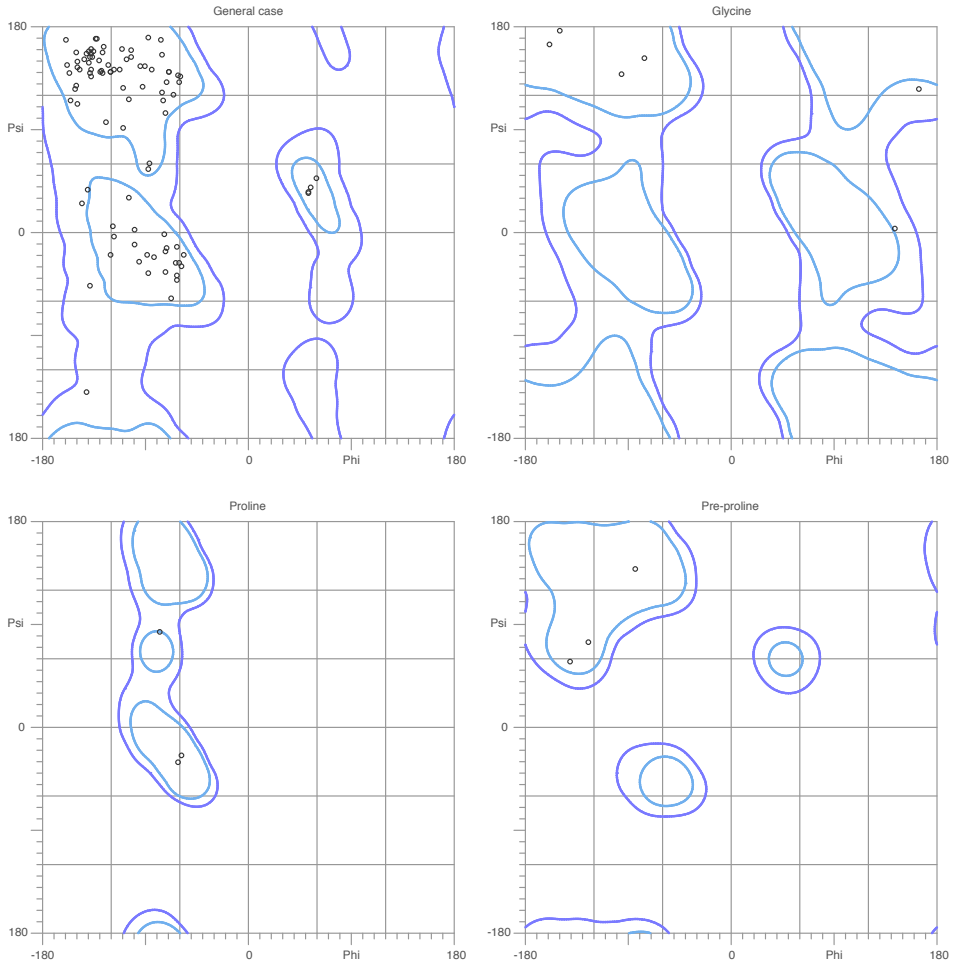


Fig. S3. (**a**) Sequence alignment of the 22 repeats of mature red abalone VERL (residues 39-3419), showing how C201 and C294 are the only Cys residues predicted not to be involved in canonical $Cys_{1\text{-}4}$ and $Cys_{2\text{-}3}$ ZP-N disulfides. The 13 residues preceding repeat 1 and the 32 additional residues within the linker of repeat 22 do not contain any Cys, whereas the 10 Cys found in the C-terminal part of the protein (residues 3420-3669, not shown) align with conserved ZP domain Cys residues involved in intramolecular disulfides. (**b**) Immunoblot analysis of recombinant VERL repeat 1-4 proteins with anti-5His antibody in reducing (R) and non-reducing (NR) conditions. Whereas a significant proportion of the wild-type protein forms disulfide-linked homodimers, a C201D+C294S double mutant is secreted in monomeric form.

a



93.3% (98/105) of all residues were in favored (98%) regions.
100.0% (105/105) of all residues were in allowed (>99.8%) regions.

There were no outliers.

b

All-Atom Contacts	Clashscore, all atoms:	0	100th percentile* (N=1784, all resolutions)
	Clashscore is the number of serious steric overlaps (> 0.4 Å) per 1000 atoms.		
Protein Geometry	Poor rotamers	1.03%	Goal: <1%
	Ramachandran outliers	0.00%	Goal: <0.2%
	Ramachandran favored	93.33%	Goal: >98%
	Cβ deviations >0.25Å	0	Goal: 0
	MolProbity score^	0.94	100th percentile* (N=27675, 0Å - 99Å)
	Residues with bad bonds:	0.00%	Goal: 0%
	Residues with bad angles:	0.00%	Goal: <0.1%

* 100th percentile is the best among structures of comparable resolution; 0th percentile is the worst.
^ MolProbity score is defined as the following: 0.42574*log(1+clashscore) + 0.32996*log(1+max(0,pctRotOut-1)) + 0.24979*log(1+max(0,100-pctRamaFavored-2)) + 0.5

c

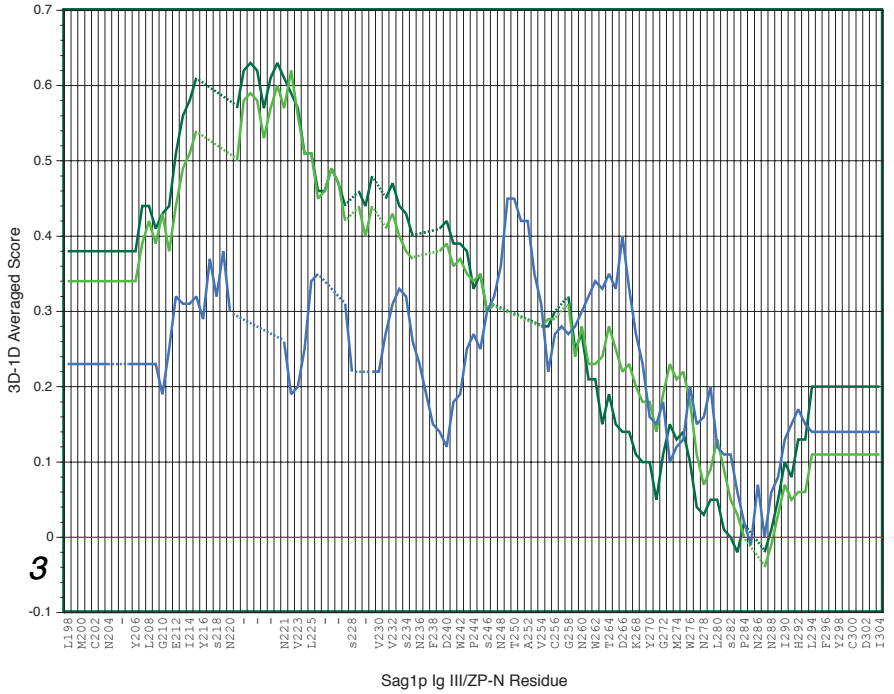


Fig. S4. Validation of the model of yeast α-agglutinin/Sag1p Ig III. (**a**,**b**) Ramachandran plot and all-atom contact/geometry analysis output by MolProbity (Chen et al., 2010). (**c**) 3D profile scores for the Sag1p Ig III model (blue) and the 2.3 Å resolution crystal structure of ZP3 ZP-N (Monné et al., 2008), both as deposited in PDB (ID 3D4G, residues 372-473 of chain A; dark green) and after MD refinement in YASARA (Krieger et al., 2009) (light green). Profiles were computed using Verify3D (Eisenberg et al., 1997) and are shown as a function of position along the Sag1p sequence, according to the alignment in fig. S5.

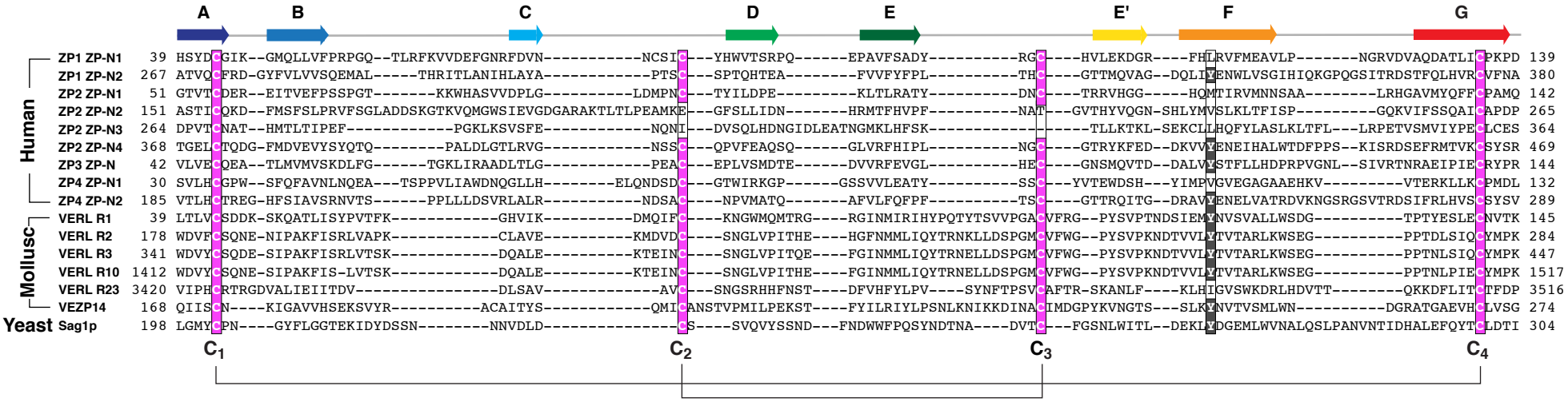


Fig. S5. Structure-based sequence alignment of ZP-N domains found in human ZP1-4, abalone VERL and VEZP14, and yeast α-agglutinin/Sag1p. VERL ZP-N repeats 4-9 and 11-22, which are almost identical to repeat 10, are not shown. ZP-N fold β-strands A-G are marked by arrows; invariant Cys_{1-4} and conserved F-strand Tyr are highlighted in magenta and grey, respectively.

Supplementary References

Aagaard, J. E., Vacquier, V. D., MacCoss, M. J., Swanson, W. J. 2010. ZP domain proteins in the abalone egg coat include a paralog of VERL under positive selection that binds lysin and 18-kDa sperm proteins. *Mol. Biol. Evol.* 27:193-203.

Aagaard, J. E., Yi, X., MacCoss, M. J., Swanson, W. J. 2006. Rapidly evolving zona pellucida domain proteins are a major component of the vitelline envelope of abalone eggs. *Proc. Natl. Acad. Sci. U. S. A.* 103:17302-17307.

Adams, P. D. et al. 2010. PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D Biol. Crystallogr.* 66:213-221.

Bork, P., Sander, C. 1992. A large domain common to sperm receptors (Zp2 and Zp3) and TGF-β type III receptor. *FEBS Lett.* 300:237-240.

Callebaut, I., Mornon, J. P., Monget, P. 2007. Isolated ZP-N domains constitute the N-terminal extensions of Zona Pellucida proteins. *Bioinformatics* 23:1871-1874.

Carter, P., Andersen, C. A., Rost, B. 2003. DSSPcont: Continuous secondary structure assignments for proteins. *Nucleic Acids Res.* 31:3293-3295.

Chen, M. H., Shen, Z. M., Bobin, S., Kahn, P. C., Lipke, P. N. 1995. Structure of *Saccharomyces cerevisiae* α-agglutinin. Evidence for a yeast cell wall protein with multiple immunoglobulin-like domains with atypical disulfides. *J. Biol. Chem.* 270:26168-26177.

Chen, V. B., Arendall, W. B. r., Headd, J. J., Keedy, D. A., Immormino, R. M., Kapral, G. J., Murray, L. W., Richardson, J. S., Richardson, D. C. 2010. MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr. D Biol. Crystallogr.* 66:12-21.

Chung, Y. D., Zhu, J., Han, Y., Kernan, M. J. 2001. *nompA* encodes a PNS-specific, ZP domain protein required to connect mechanosensory dendrites to sensory structures. *Neuron* 29:415-428.

de Bakker, P. I., Bateman, A., Burke, D. F., Miguel, R. N., Mizuguchi, K., Shi, J., Shirai, H., Blundell, T. L. 2001. HOMSTRAD: adding sequence information to structure-based alignments of homologous protein families. *Bioinformatics* 17:748-749.

DeLano, W. L. 2002. *The PyMOL Molecular Graphics System*.

Eisenberg, D., Luthy, R., Bowie, J. U. 1997. VERIFY3D: assessment of protein models with three-dimensional profiles. *Methods Enzymol.* 277:396-404.

Emsley, P., Lohkamp, B., Scott, W. G., Cowtan, C. 2010. Features and development of Coot. *Acta Crystallogr. D Biol. Crystallogr.* 66:486-501.

Finn, R. D. et al. 2010. The Pfam protein families database. *Nucleic Acids Res.* 38:D211-22.

Fiser, A., Sali, A. 2003. Modeller: generation and refinement of homology-based protein structure models. *Methods Enzymol.* 374:461-491.

Galindo, B. E., Moy, G. W., Swanson, W. J., Vacquier, V. D. 2002. Full-length sequence of VERL, the egg vitelline envelope receptor for abalone sperm lysin. *Gene* 288:111-117.

Gille, C., Frömmel, C. 2001. STRAP: editor for STRuctural Alignments of Proteins. *Bioinformatics* 17:377-378.

Gupta, S. K., Bansal, P., Ganguly, A., Bhandari, B., Chakrabarti, K. 2009. Human zona pellucida glycoproteins: functional relevance during fertilization. *J. Reprod. Immunol.* 83:50-55.

Heger, A., Holm, L. 2000. Rapid automatic detection and alignment of repeats in protein sequences. *Proteins* 41:224-237.

Jazwinska, A., Ribeiro, C., Affolter, M. 2003. Epithelial tube morphogenesis during *Drosophila* tracheal development requires Piopio, a luminal ZP protein. *Nat. Cell Biol.* 5:895-901.

Krieger, E., Joo, K., Lee, J., Lee, J., Raman, S., Thompson, J., Tyka, M., Baker, D., Karplus, K. 2009. Improving physical realism, stereochemistry, and side-chain accuracy in homology modeling: Four approaches that performed well in CASP8. *Proteins* 77 Suppl 9:114-122.

Letunic, I., Doerks, T., Bork, P. 2009. SMART 6: recent updates and new developments. *Nucleic Acids Res.* 37:D229-D232.

Monné, M., Han, L., Schwend, T., Burendahl, S., Jovine, L. 2008. Crystal structure of the ZP-N domain of ZP3 reveals the core fold of animal egg coats. *Nature* 456:653-657.

Parisien, M., Major, F. 2005. A new catalog of protein β-sheets. *Proteins* 61:545-558.

Roch, F., Alonso, C. R., Akam, M. 2003. *Drosophila miniature* and *dusky* encode ZP proteins required for cytoskeletal reorganisation during wing morphogenesis. *J. Cell Sci.* 116:1199-1207.

Roy, A., Kucukural, A., Zhang, Y. 2010. I-TASSER: a unified platform for automated protein structure and function prediction. *Nat. Protoc.* 5:725-738.

Shen, M. Y., Sali, A. 2006. Statistical potential for assessment and prediction of protein structures. *Protein Sci.* 15:2507-2524.

Shi, J., Blundell, T. L., Mizuguchi, K. 2001. FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. *J. Mol. Biol.* 310:243-257.

Tsubamoto, H., Hasegawa, A., Nakata, Y., Naito, S., Yamasaki, N., Koyama, K. 1999. Expression of recombinant human zona pellucida protein 2 and its binding capacity to spermatozoa. *Biol. Reprod.* 61:1649-1654.

Wilkin, M. B., Becker, M. N., Mulvey, D., Phan, I., Chao, A., Cooper, K., Chung, H. J., Campbell, I. D., Baron, M., MacIntyre, R. 2000. *Drosophila* Dumpy is a gigantic

extracellular protein required to maintain tension at epidermal-cuticle attachment sites. *Curr. Biol.* 10:559-567.