In Silico Resurrection of the Major Vault Protein Suggests It Is Ancestral in Modern Eukaryotes

Toni K. Daly*, Andrew J. Sutherland-Smith, and David Penny

Institute of Fundamental Sciences, Massey University, Palmerston North, New Zealand

*Corresponding author: E-mail t.daly1@massey.ac.nz; tonidaly@mac.com.

Accepted: July 17, 2013

Abstract

Vaults are very large oligomeric ribonucleoproteins conserved among a variety of species. The rat vault 3D structure shows an ovoid oligomeric particle, consisting of 78 major vault protein monomers, each of approximately 861 amino acids. Vaults are probably the largest ribonucleoprotein structures in eukaryote cells, being approximately 70 nm in length with a diameter of 40 nm—the size of three ribosomes and with a lumen capacity of 50 million Å^3. We use both protein sequences and inferred ancestral sequences for in silico virtual resurrection of tertiary and quaternary structures to search for vaults in a wide variety of eukaryotes. We find that the vault's phylogenetic distribution is widespread in eukaryotes, but is apparently absent in some notable model organisms. Our conclusion from the distribution of vaults is that they were present in the last eukaryote common ancestor but they have apparently been lost from a number of groups including fungi, insects, and probably plants. Our approach of inferring ancestral 3D and quaternary structures is expected to be useful generally.

Key words: vault ribonucleoprotein, ancestral reconstruction (ASR), BLAST, I-TASSER, RosettaDock, last eukaryotic common ancestor.

Introduction

Phylogenetic reconstruction of the last eukaryotic common ancestor (LECA) and ultimately the path of life itself is a goal of evolutionary biologists. Molecular phylogenetics has sped up this search and has shown that LECA had many more properties than simply a nucleus and mitochondria. For example, LECA had linear genetic material, essential for meiosis and the advantages that sex and recombination bring (Ishikawa and Naito 1999), but it does lead to the issue of terminal erosion of chromosomes. Although there are a number of fixes, the telomerase complex is the standard caretaker of eukaryote telomeres (Nosek et al. 2006) and is also ancestral. LECA already had introns and a complex spliceosome to process them (Collins and Penny 2005). LECA could synthesize sterols, essential for phagocytosis and cell signaling (Desmond and Gribaldo 2009). If the vault particle were also in LECA, what possible role could it have?

Our interest has been in using in silico methods for inferring 3D structure of proteins (Daly et al. 2013) from tertiary structures determined by standard X-ray crystallography methods. These do not require strict adherence to known 3D structures, they do allow variation, but still based on known structures. In addition, we have used quaternary structural information, estimating the extent that the tertiary models will assemble into the expected quaternary vault structure. Our approach here is to combine the three methods: searching for proteins that are widespread in eukaryotes by BLAST searches; using I-TASSER to test that the sequences found by BLAST searches (or their inferred ancestral sequences) will really fold into the expected tertiary structures; and using RosettaDock to infer quaternary structure.

In addition, the search for proteins widespread in eukaryotes has recently been extended to allow for some losses in specific lineages (Tabach et al. 2013). These authors reported a loss of some proteins, particularly a loss of homologs to the ciliated sensory ending component (BBS-1) in plants and fungi. However, this loss did not affect the conclusion that BBS-1 proteins were ancestral in eukaryotes. Our primary contribution here is to consider tertiary and quaternary structure in the search for ancestral eukaryote proteins, especially the vaults. Allowing loss of a few ancestral proteins from specific groups is an important advance.

Currently, we do not know the full details of deeper eukaryote phylogeny, probably because the ability of

Markov models to reconstruct sequences (or the tree) falls off exponentially at deeper times (Mossel and Steel 2005). This means that the definite relationship of the main groups of eukaryotes is not yet known. There had been hints that the root could be within the excavates in 2007 (Rodríguez-Ezpeleta et al. 2007). More recently, Cavalier-Smith (2010) proposed that the root of the eukaryotic ancestor lay between euglenozoa and the rest of the eukaryotes, that is, within the excavates, breaking euglenozoa away from excavata. However, with equal confidence, he had previously favored the root between the opisthokonts (animals plus fungi = fungamals) and all other eukaryotes (Stechmann and Cavalier-Smith 2003).

Regardless of the placement of the root, the accepted approach is to find features that are in all the major groups of eukaryotes, for example, those defined by Keeling et al. (2005). These are considered to be 1) Opisthokonts (fungi and animals) and the amoebozoa (arguably a supergroup of their own); 2) Plants (Plantae); 3) Excavates (such as *Naegleria*, *Trichomonas*, *Giardia*, and also including Euglenoids); 4) Stramenopiles and Alveolates, together known as chromalveolates; and 5) Rhizaria, these include Radiolarians and Foraminfera. More recently, the chromalveolates have been grouped with Rhizaria forming a supergroup known as SAR (Stramenopile, Alevolate, Rhizaria) (Burki et al. 2007; Elias et al. 2009). Our strategy is to identify vaults in as many of these groups as possible, especially because vaults appear to have been lost in several significant groups of eukaryotes (see later).

With the publication of the *Naegleria gruberi* genome (Fritz-Laylin et al. 2010), the number of genes in the putative LECA increased by 700 additional genes to over 4,000, summarized by Koonin (2010). This is likely to be a conservative estimate, as it does not account for gene loss in a few lineages. Using our in silico protocol of searching for remote sequence homologs of the major vault protein (MVP) monomer, and assessing their putative tertiary and quaternary structure, we found that there were at least two plausible candidate genes in *N. gruberi* encoding proteins predicted to fold as MVP (UniProtKB:D2V5B9 and D2W0Z9) and we predicted that they would ultimately form a vault particle (Daly et al. 2013).

If we are to propose that a particle such as the vault were present in LECA, we would anticipate that many signals from sequence homology would have been largely erased by successive substitutions. We expect that the tertiary and quaternary structural information would persist longer even where other primary sequence becomes randomized (Mossel and Steel 2005). The distribution of amino acid substitution is not random because some residues are essential for tertiary and quaternary structure, so consequently we expect them to be more highly conserved than residues of lesser structural import.

The primary (initial) function of vaults is not known. Extant vaults are associated with signaling pathways (Berger et al. 2009) they are known to be upregulated in treatment resistant cancers (Herlevsen et al. 2007) and epilepsy (Liu et al. 2011), but these clearly were not their original role in protists. Vaults are enriched in tissue types that are involved with scavenging such as in macrophages (Chugani et al. 1991) and in lipid rafts (Kowalski et al. 2007). They have been observed containing cargo (suggested to be mRNA [Paspalas et al. 2009]), and sea urchin vaults appear to contain many proteins (Stewart et al. 2005). Researchers are now using vaults to deliver cargo such as vaccines (Champion et al. 2009) and drugs (Buehler et al. 2011) to targeted cells. Again, carrying vaccines is clearly not their original function, nevertheless, everything we can learn about their distribution and functions will help understand their original role.

As mentioned earlier, we use a 3-fold protocol for inferring structure using BLASTp, I-TASSER, and RosettaDock. First, a BLASTp search of the UniProt and NCBI protein sequence databases is used to find potential MVP homologs. Second, the tertiary structure of these sequences is predicted by I-TASSER (iterative threading assembly refinement server) (Zhang 2008). This program uses a combination of protein structure prediction techniques to produce potential models of secondary and tertiary structures for a given sequence, based on a structural template. If there is structural similarity to an MVP monomer, we anticipate that I-TASSER will predict the greatest similarity to the rat MVP, as it is the only 3D crystallographic structure in the Protein Data Bank that is almost full-length. For this reason, we have used the rat sequence (UniProtKB:Q62667) as the standard against which others are measured. Third, the output structures from I-TASSER are then submitted to RosettaDock (Lyskov and Gray 2008) to determine whether the predicted monomeric MVP structures are expected to assemble into vaults. Because of this potential for some groups to lose vaults, it is important to test the predicted tertiary and quaternary structure of vault sequences to see that they really do fold and dock in the expected manner (Daly et al. 2013).

Sequences that meet the three criteria of being more or less complete at the primary sequence level, structurally creditable as MVP monomers similar to the rat structure (Tanaka et al. 2009; the only complete crystal structure in the protein data bank), and are predicted by RosettaDock to dock laterally, were grouped phylogenetically and used to infer ancestral MVP sequences using PAML4 (Phylogenetic Analysis by Maximum Likelihood) (Yang 2007) and FastML (Ashkenazy et al. 2012). When we reconstructed the putative ancestor of all eukaryotes, we added Mega5 (Tamura et al. 2011) to try to limit bias shown by both PAML4 and FastML described later. There are more sequences potentially available than have ultimately been used because it is often difficult to decide whether a sequence annotated as complete really is, when there are apparent gaps. There is also the issue of orphan sequences, where there is a cDNA or mRNA sequence that is not ascribed to a gene. In some cases, these are placed on a branch on a phylogenetic tree consistent with established taxonomy increasing the likelihood of being correct. In other instances, the position on a tree seems so unlikely that RNA contamination must be suspected, instances of these problems are described later.

Because some groups of eukaryotes may lack vaults, we make an additional test and infer ancestral sequences for a broader group, for example, invertebrates—(though insects appear to lack vaults). We then undertake the same tertiary and quaternary structure predictions of the inferred ancestral sequences. Ancestral reconstruction (ASR) takes a multiple sequence alignment (MSA) (nucleotides or proteins), together with a tree representing the sequences in the MSA and calculates the most likely ancestral sequence at each node of the tree. For example, ASR has been used to calculate a putative ancestral RNAse P sequence to submit as a BLAST query in the search for evolutionarily distant protein homologs (Collins et al. 2003). Usually, the ASRs retrieved known sequences for proteins associated with RNAse P with a higher *E* value than by BLASTing with known sequences; in one instance a protein homolog was found in *Giardia lamblia* using the reconstruction that could not be retrieved using any of the known sequences. Ancestral proteins have also been experimentally resurrected (that is, synthesized) from sequences determined by ASR (Chang et al. 2002; Gullberg et al. 2010). However, here we only resurrect MVP in silico, with a combination of two protocols, first by reconstructing the ancestral sequence for each group (using PAML4 and FastML), and second by inferring the structures using I-TASSER. Explicitly, our test is—will the inferred ancestral MVP sequences be as capable of forming vault particles using our modeling protocol, as the extant MVP?

ASR can use a variety of methods, perhaps the most reliable uses posterior probabilities from known trees in their reconstructions (maximum likelihood [ML] and empirical Bayes). Empirical Bayes may overlook the best guess in terms of most likely substitution resulting in slightly less accurate sequence reconstruction but may better preserve structural and functional properties (Williams et al. 2006). An additional form of ASR involving topological empirical Bayes, which weights the trees differently to other methods, has not been found to alter the resultant sequence (Hanson-Smith et al. 2010). We have found that a combination of two ASR algorithms (PAML4 and FastML) combined with human intervention results in a suitable ancestral sequence to put forward for ancestral protein structural prediction, or in silico resurrection.

Materials and Methods

Full-length MVP sequences were found by BLASTp and PSI BLASTing of the NCBI and UniProtKB databases using the rat MVP sequence (UniProtKB:Q62667) as the query. Using our established protocol of tertiary and quaternary structural modeling (Daly et al. 2013), we retrieved 116 eukaryote and 10 bacterial protein sequences that fulfilled the criteria of structural homology with sufficient lateral docking capacity for vault particle formation. Much of the available sequence data had not been subject to detailed analysis, with only few sequences ascribed to a chromosomal position even if they are from genomic DNA. Some MVP sequences are derived from mRNA, for example, cat (*Felis catus*; UniProtKB:Q18PA2), diamondback rattlesnake (*Crotalus adamanteus*; UniProtKB:J3RZY3), and barley (*Hordeum vulgare*; UniProtKB:F2E078). A tree constructed from all protein sequences used in the ancestor of all eukaryotes is available online (supplementary material S1b, Supplementary Material online).

Each MSA used for ASR was generated by MUSCLE (Edgar 2004). Trees were also calculated for each MSA using MrBayes (Huelsenbeck and Ronquist 2001) with both algorithms run via the Geneious platform (Geneious Pro 5.5.7 Biomatters available from http://www.geneious.com/, last accessed August 2, 2013). Most ASR algorithms require a tree formed from the MSA under scrutiny. FastML will calculate a tree from the MSA but we found that MrBayes trees produced the most plausible and reliable trees for submission to both PAML and FastML, although computationally the most expensive. At least four sequences are required for MrBayes tree, which means that any groups with less than four representative sequences could not be used for ASR. Although we do not expect that a tree built from a single gene would necessarily be the same as a tree built using combined gene sequences (Philippe et al. 2011), we have used the method that produces the same tree for MVP for a given species set each time it is calculated to limit systematic error. We also tried calculating the tree using different roots to be certain that the ASR algorithms were being provided with the best initial data. Sequences are continually being added to the databases and the ancestral MVP sequences can be refined. Both PAML4 and FastML ASR methods use ML analysis (empirical Bayes) to estimate the ancestral sequence. However, there are unfortunately significant differences between methods in their handling of sites with missing data. PAML deletes sites in the ancestral sequence where any one sequence contains a gap in the MSA, whereas FastML uses a binary matrix to reconstruct indels and adds them back into regions of more highly conserved sequence. This means that the PAML sequences are shorter, and the FastML sequences much longer. One standard fix with PAML is to use "X" in the position of gaps; this stops the automatic deletion of sites with gaps (PAML FAQ; http://abacus.gene.ucl.ac.uk/software/paml.html, last accessed July 2, 2013).

Limitations of MVP Ancestral Sequence Reconstruction

PAML and FastML generally give identical ancestral sequence for the same input MSA and where there are no gaps. Because gaps result in ambiguity, sequences were checked carefully for completeness to limit the inclusion of sequences that would adversely affect ASR. For instance, the *Naegleria* MVP sequences are significantly short (~530 residues rather than ~880), but do appear to be complete (Daly et al. 2013). In other instances, the sequences are within the anticipated range of length but have region(s) missing. The pika MVP sequence has 23 residues missing from exon 9 (residues 523–545 compared with the rat sequence), these residues were also missing in the 2012 version of the chimp MVP sequence. Because this region is essential for correct tertiary structure, is highly conserved, and because rhesus macaques are known to have vault particles (Paspalas et al. 2009), this missing sequence region seemed likely to be an artifact in the database. Subsequently, an updated chimp MVP sequence has been deposited in the UniProt database (January 9, 2013) and is full length. Pika MVP was included in the ASR because this region of exon 9 was the only missing sequence and is more likely to be due to sequencing problems rather than the absence of this conserved region. However, this issue highlighted a limitation of using PAML; figure 1 shows how PAML and FastML (mis)treat the missing sequence region.

FastML includes insertions in the ancestral sequence even if it is representative of only one species, and so the FastML ancestral sequences are longer. Deleting insertions in the MSA where they are represented in only a single species solved this ambiguity. It is more parsimonious that an insertion has occurred in a single species, than the alternative where deletions have occurred in all other species.

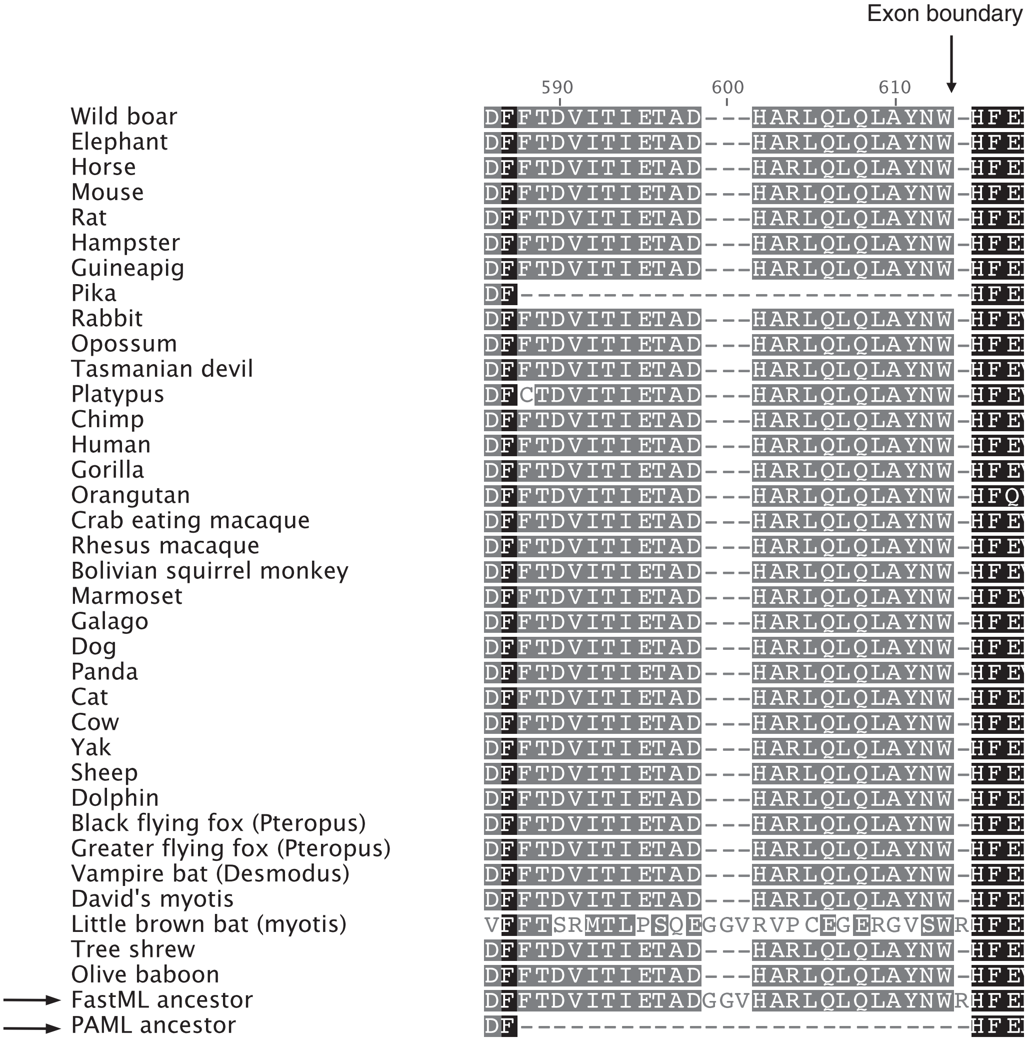

Exon boundary
590
600
610
Wild boar DFFTDVITIETAD---HARLQLQLAYNW-HFE
Elephant DFFTDVITIETAD---HARLQLQLAYNW-HFE
Horse DFFTDVITIETAD---HARLQLQLAYNW-HFE
Mouse DFFTDVITIETAD---HARLQLQLAYNW-HFE
Rat DFFTDVITIETAD---HARLQLQLAYNW-HFE
Hampster DFFTDVITIETAD---HARLQLQLAYNW-HFE
Guineapig DFFTDVITIETAD---HARLQLQLAYNW-HFE
Pika DF---------------------------HFE
Rabbit DFFTDVITIETAD---HARLQLQLAYNW-HFE
Opossum DFFTDVITIETAD---HARLQLQLAYNW-HFE
Tasmanian devil DFFTDVITIETAD---HARLQLQLAYNW-HFE
Platypus DFCTDVITIETAD---HARLQLQLAYNW-HFE
Chimp DFFTDVITIETAD---HARLQLQLAYNW-HFE
Human DFFTDVITIETAD---HARLQLQLAYNW-HFE
Gorilla DFFTDVITIETAD---HARLQLQLAYNW-HFE
Orangutan DFFTDVITIETAD---HARLQLQLAYNW-HFQ
Crab eating macaque DFFTDVITIETAD---HARLQLQLAYNW-HFE
Rhesus macaque DFFTDVITIETAD---HARLQLQLAYNW-HFE
Bolivian squirrel monkey DFFTDVITIETAD---HARLQLQLAYNW-HFE
Marmoset DFFTDVITIETAD---HARLQLQLAYNW-HFE
Galago DFFTDVITIETAD---HARLQLQLAYNW-HFE
Dog DFFTDVITIETAD---HARLQLQLAYNW-HFE
Panda DFFTDVITIETAD---HARLQLQLAYNW-HFE
Cat DFFTDVITIETAD---HARLQLQLAYNW-HFE
Cow DFFTDVITIETAD---HARLQLQLAYNW-HFE
Yak DFFTDVITIETAD---HARLQLQLAYNW-HFE
Sheep DFFTDVITIETAD---HARLQLQLAYNW-HFE
Dolphin DFFTDVITIETAD---HARLQLQLAYNW-HFE
Black flying fox (Pteropus) DFFTDVITIETAD---HARLQLQLAYNW-HFE
Greater flying fox (Pteropus) DFFTDVITIETAD---HARLQLQLAYNW-HFE
Vampire bat (Desmodus) DFFTDVITIETAD---HARLQLQLAYNW-HFE
David's myotis DFFTDVITIETAD---HARLQLQLAYNW-HFE
Little brown bat (myotis) VFFTSRMTLPSQEGGVRVPCEGERGVSWRHFE
Tree shrew DFFTDVITIETAD---HARLQLQLAYNW-HFE
Olive baboon DFFTDVITIETAD---HARLQLQLAYNW-HFE
FastML ancestor DFFTDVITIETADGGVHARLQLQLAYNWRHFE
PAML ancestor DF---------------------------HFE

FIG. 1.—Problems with FastML and PAML. MUSCLE alignment of mammal MVP sequences at the 3′ exon 9 boundary, a region which is responsible for shoulder domain formation (shaded by similarity). The FastML algorithm has included the little brown bat sequence to its calculated ancestral sequence even though it is an outlier with respect to the consensus MSA. PAML4 omitted the 23 extremely well-conserved residues from the ancestral sequence unless "X" was put in place of the gaps in the pika sequence.

The PAML mammal MVP ancestral sequence, derived from an alignment of 35 mammal MVP sequences, was 740 residues in length while the FastML ancestor from the same MSA had 965 residues. In contrast, the average length of extant MVP sequences is approximately 880 residues. However, replacing missing residues with XXs does not work with large numbers of MVP sequences, or for more diverse sequences such as for invertebrates, as there are gaps in the ASR due to sequence divergence, rather than there being a single sequence region missing.

We took this observation to the extreme in testing how different the PAML and FastML ancestors would be, using 119 MVP sequences to calculate an MVP ancestor of everything. The resulting PAML sequence is 185 residues long in contrast to the FastML ancestral sequence of 1,382 residues in length! It is unrealistic to think that ancestral sequences were generally either significantly shorter or longer. A simpler explanation is that the PAML algorithm has a bias toward reconstructing shorter sequences and FastML to reconstructing longer sequences the more ancient the ancestor becomes. At this point, we added Mega5 (Tamura et al. 2011) to our repertoire of ASR algorithms. Mega5 allows control over the percentage of residues from the MSA that must be considered. When set at use all sites—Mega5 resulted sequences shorter than the total length of the MSA where FastML would have given a results that was as long as the total of the MSA, that is, it was selecting which residue was most likely and not necessarily including residues when they appeared to be restricted to just a few sequences, or deleting all residues where there was a gap in one or few sequences as PAML4 would. This meant that the resultant sequence was a more realistic length. However, it did have a bias of its own, in that it removed some highly conserved sequence that was not present in what could be argued to be more ancient species (discussed later).

FastML additionally has the option of marginal reconstruction (where the residue replaced is based on the posterior probability of the next step at that position from the tree), or joint reconstruction (where the probability is the product of the next two steps, that is, the next most likely substitution is based on two steps rather than one). There were very minor differences at the local level, that is, mammal, invertebrate, and so forth, but the sequence of the ancestor of all eukaryotes then had 17% sequence difference depending on whether marginal or joint reconstruction was used. In practice, the method of reconstruction made little difference to either I-TASSER or RosettaDock. Ultimately, PAML and FastML sequences were combined to generate the final ancestral sequence for each group (described later) for submission to I-TASSER. Inserts unique to just one genus were removed when reconstructing the ancestor of all eukaryotic MVPs—resulting in a more realistic range of ancestor of 678 (PAML)—892 (FastML) residues. A comparison was made between various methods for reconstructing the overall ancestor (discussed later).

Determination of Vault Particle Formation

The completed ASR MVP sequences were analyzed by I-TASSER without constraint (described in Daly et al. 2013) to test that they would be predicted to fold similarly to the rat MVP. Briefly, I-TASSER uses a suite of threading programs known collectively as LOMETS (Wu and Zhang 2007) and outputs up to five structural predictions scored by the confidence in the topology of the model; known as the C score (range is from −5 to +2). We have used a C score cut off of greater than −1.5, which is indicative of a correct fold (Roy et al. 2010). There is an additional score calculated by I-TASSER the template modeling (TM) score that quantifies structural similarity between two superimposed protein structures analogous to the traditional root mean squared difference. A TM score of greater than 0.5 indicates high confidence that the topology of two models, in this case predicted and native (rat) MVP are the same. We have therefore additionally used a TM score of 0.5 or higher as a cut off for inclusion in our analysis.

Then two identical copies of the shoulder and coil domains of each I-TASSER output of ancestral MVP models were submitted to RosettaDock (Gray et al. 2003). For oligomeric vault formation, the crystal structure shows that MVP monomers dock laterally along the length of both sides to make the distinctive barrel shape (Tanaka et al. 2009). RosettaDock uses a low resolution Monte Carlo search and backbone optimization algorithm to optimally position a submitted monomer pair, followed by a refinement to relax the backbone and accommodate the side chains (Gray et al. 2003). Bona fide vault monomers would dock along their entire length with a negative RosettaDock energy score.

Docking a pair of full-length MVP monomers cannot be done via the RosettaDock web server because of a 600 residue limit. It has previously been demonstrated that the coiled coil region is essential for vault formation (van Zon et al. 2002), but we have found that improved in silico docking usually includes the MVP shoulder region as well (Daly et al. 2013). Our test requires that MVP shoulder and coiled coil (known as the cap-helix) would be predicted to dock laterally with an identical monomer, indicating vault particle formation (fig. 2).

Explicitly then, our determination of an MVP monomer is that it will form a vault particle by meeting both our two I-TASSER cut off criteria and then docking with both a negative RosettaDock energy score and the majority of the 1,000 models produced by RosettaDock clustering around this structure. In general, we would expect that such a docked pair of monomers would be in the lowest 10 energy models. All sequences used for the MSAs fulfilled this criteria, as did the ancestors produced by the ASR algorithms (table 1).

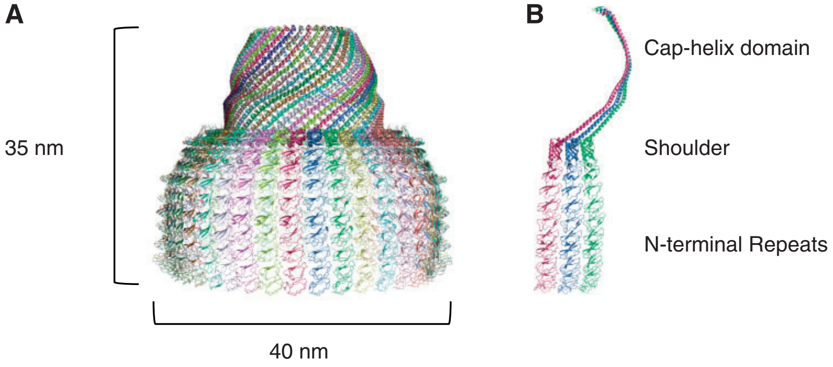


FIG. 2.—Vault ribonucleoprotein structure. (*A*) Rat MVP quaternary structure showing half a vault colored by monomer (PDB: 2ZUO, 2ZU4, and 2ZV5). A full vault will have an opposing copy of the upper half vault associated at the N terminii. (*B*) Three rat MVP monomers (PDB 2ZUO stripped down to three monomers). This figure highlights the extensive lateral association required to dock into the vault quaternary structure. All ribbon diagrams are rendered in PyMol version 1.3.

Results

Reconstructing Eukaryote Ancestral MVP Sequences

In the cases of metazoa (63 sequences), amoebozoa (9 sequences), and kinetoplast MVP (29 sequences), complete protein sequences with high homology to the rat crystal structure could be found by simple BLASTp searches using default parameters. There are additionally many more sequences that are fragments of MVP. But in the case of the stramenopiles (e.g., diatoms and oomycetes), there were only just enough sequences from different species to create an ancestor (5 sequences), and although there were five alveolate sequences they came from just two ciliate species: *Paramecium tetraurelia* (3 sequences) and *Oxytricha trifallax* (2 sequences). Sequences that fulfilled the I-TASSER criteria for inclusion (I-TASSER C score of >-1.5 and TM score of >0.5) and with a negative RosettaDock energy score (lower is more favorable), but had insufficient representation for ASR, were used as individual sequences. The inferred 3D structures of all of the ancestors are shown later in figure 8 and in the supplementary material S3a (Supplementary Material online).

Metazoa

MSAs were calculated to optimize various phylogenetic groupings. Eventually metazoa were split into; mammals, other sarcopterygii (coelacanth, *Xenopus laevis, X. tropicalis,* the Carolina anole, diamondback rattlesnake, chicken, and turkey), fish, and invertebrates (supplementary material S2a–d [Supplementary Material online] for these four trees). In the invertebrates, we have representative sequences from sponges (where there are 20 sequences though few are complete), cnidarians, bivalves, annelids, nematodes, and echinoderms. However, we have not found sequences from any arthropods. Although there are a few sequences with limited homology that will fold to resemble parts of MVP, we suggest that the whole vault particle has been lost and the *mvp* gene degraded beyond recognition in this group. Because the lancelet (an isolated lineage) was difficult to place, it was initially omitted from all of the groups and only added to the final tree of all opisthokonts (supplementary material S1a, Supplementary Material online).

Other Opisthokonts

Opisthokonts comprise all metazoa, fungi, plus choanoflagellates, and capsaspora (the latter two are neither animal nor fungi but are closely related and share many gene homologs) (Sebe-Pedros et al. 2011). There were too few MVP sequences to calculate an ancestral sequence for capsaspora or the choanoflagellata. *Capsaspora owczarzaki* is a single cell eukaryote that is neither an animal nor a choanoflagellate but closely related to both and is a symbiont of the freshwater snail *Biomphlaria glabrata*. There are three putative *Cap. owczarzaki* MVP homologs—an insufficient number to infer an ancestor. The choanoflagellates are represented by two species: *Salpingoeca rosetta* and *Monosiga brevicollis*. A single capsaspora MVP sequence and the two choanoflagellate sequences were added to metazoan MVP sequences in the reconstruction of the opisthokont ancestor (supplementary material S1a, Supplementary Material online). The tree of opisthokonts placed the capsaspora and choanoflagellates within the invertebrates. This is probably a reflection of the increased evolutionary rate of the sponge, *Amphimedon queenslandica,* the parasitic nematodes; *Clonorchis sinensis* and *Schistosoma mansoni* (Tsai et al. 2013), and the tunicate *Oikopleura dioica* (Denoeud et al. 2010). The placement of *Cap. owczarzaki* with hydra is more surprising.

Included within the grouping opisthokont are the fungi. There are proteins that will fold similarly to MVP but these require the constraint of the rat structure when submitted to I-TASSER and do not score within our cut off criteria. Some of these models have been submitted to RosettaDock and show that they will dock although the score is poor in comparison with metazoa. Additionally, vaults have not been found in fungi and are generally described as missing (Suprenant 2002). The few sequences that we have retrieved are unlikely to form vault particles and we speculate that they may have derived originally from the *mvp* gene and are now significantly diverged, they are annotated as uncharacterized proteins.

Amoebozoa

An amoebozoan MVP MSA was constructed for MVPα and MVPβ separately to reconstruct the ancestor of each MVP form. The dictyostelids form chimeric vaults with proteins from both α and β genes (Vasu and Rome 1995). Both MVP sequences from *Polysphondylium pallidum* (UniProtKB: P34118 and D3BM96) are annotated MVPβ; this is clearly a mis-annotation because P34118 is phylogenetically positioned within the MVPα sequences (fig. 3; supplementary material S2e, Supplementary Material online). An ancestral MVP sequence was also reconstructed from an MSA of MVP α and β sequences combined and it is interesting to note that this ancestor docked readily in RosettaDock, indicating that the product from a single original gene could have made a

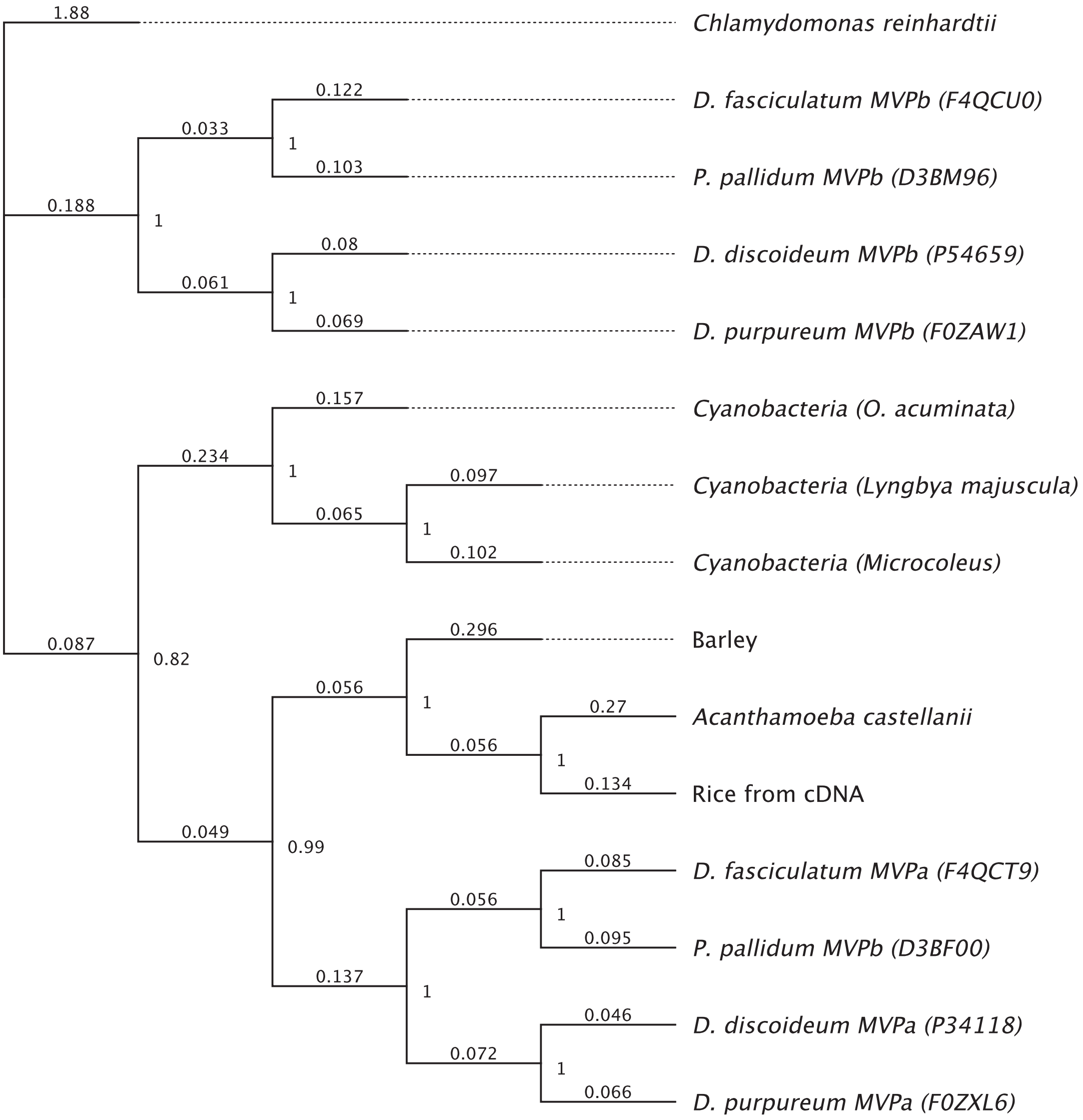

1.88
Chlamydomonas reinhardtii
0.122
D. fasciculatum MVPb (F4QCU0)
0.033
1
0.103
P. pallidum MVPb (D3BM96)
0.188
1
0.08
D. discoideum MVPb (P54659)
0.061
1
0.069
D. purpureum MVPb (F0ZAW1)
0.157
Cyanobacteria (O. acuminata)
0.234
1
0.097
Cyanobacteria (Lyngbya majuscula)
0.065
1
0.102
Cyanobacteria (Microcoleus)
0.296
Barley
0.087
0.82
0.056
1
0.27
Acanthamoeba castellanii
0.056
1
0.134
Rice from cDNA
0.049
0.99
0.085
D. fasciculatum MVPa (F4QCT9)
0.056
1
0.095
P. pallidum MVPb (D3BF00)
0.137
1
0.046
D. discoideum MVPa (P34118)
0.072
1
0.066
D. purpureum MVPa (F0ZXL6)

Fig. 3.—MrBayes tree showing the unlikely position of the barley, rice, and, cyanobacteria MVP sequences grouped within the cellular amoebozoa. Because the plant sequences have yet to be attributed to genomic DNA, it seems more likely that they are cDNAs derived from contaminants. The tree shown is rooted by *Chlamydomonas reinhardtii* but other root choices produce the same results. The number by the node represents posterior probability, the number on the branches represents the number of replacements per residue—of course the same residue may have been replaced multiple times. Note that *Polysphondylium pallidum* D3BF00 has been mis-annotated, both *Pol. pallidum* sequences are designated as MVPβ but D3BF00 has greater sequence similarity with the MVPα sequences of the other amoebozoa.

vault in this group (table 1). In *Dictyostelium discoideum*, knocking out expression of either of MVPα or MVPβ interferes with vault structure in that the vaults are abnormally ovoid though vaults still form (Vasu and Rome 1995).

Excavates

There are currently in excess of 30 gene sequences with homology to MVP within the leishmania and trypanosomes (which are grouped within euglenozoa) that fulfill our criteria for vault particle formation and 29 of them are complete. However, neither gene nor protein sequences have been found in any other excavates with the exception of the heterolobosea *N. gruberi*. The situation with excavates has become more complicated in that Cavalier-Smith (2010) has redefined them to be on both sides of his proposed rooting of modern eukaryotes. Our general approach has been to search for MVPs in all major groups of eukaryotes. Nevertheless, we have considered both options, the root being within the traditional excavates, or not. Consequently one of the *Naegleria* sequences has been included in the final tree as an individual but has not contributed to an ancestor other than the all eukaryotic MVP ancestor.

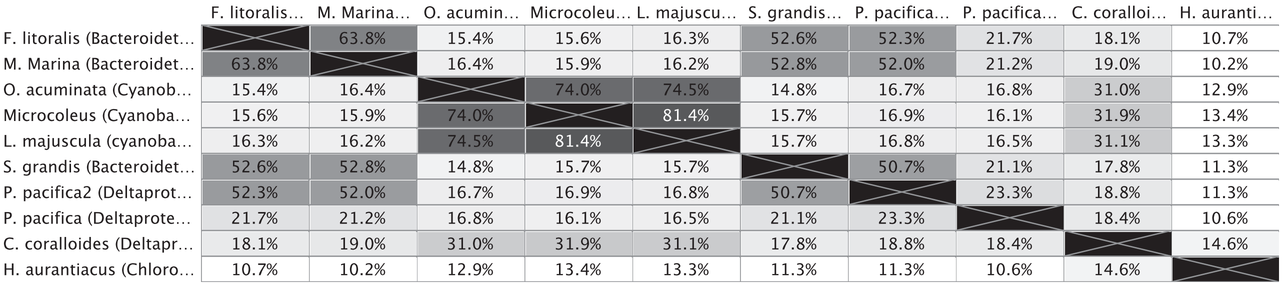
Kinetoplasts fall within euglenozoa and were treated either as two groups, (leishmania and trypanosomes), or as one MSA (and tree) to reconstruct a general kinetoplast ancestor. This was because there is a complex history of gene duplication and the sequences between groups clearly indicate a greater relationship across species in tiers rather than within any individual species (supplementary material S2f, Supplementary Material online). If the root of the eukaryotic ancestor requires that euglenozoa are removed from the excavates (Cavalier-Smith 2010), the main point is that the vault particle still appears on both sides of the proposed root, even though it may have been lost in a number of lineages.

Plants

The first land plant sequence identified as an MVP homolog was deposited in databases in March 2011 (Matsumoto et al. 2011) from mRNA of *H. vulgare* (domesticated barley; GenBank:BAK00750 UniProt:F2E078). This was unexpected because up to then no plant had then been shown to have either whole vaults or MVP monomers. Furthermore, the barley MVP sequence has a surprisingly high level of homology (55% identical residues) with rat (UniProtKB:Q62667). A BLAST search of the NCBI plant genomic database using the barley cDNA sequence (NIASHv2093C22) (GenBank: AK369549) resulted in a match to a cDNA sequence in the rice database (*Oryza sativa*) (GenBank:CT836653) with 60% homology to barley, a value considerably lower than expected since both taxa are members of the grass family (Gramineae). However, further BLASTing failed to retrieve either rice genomic DNA or protein sequence. A BLAST of the recently released barley genome (Klaus 2012), using stretches of cDNA described as MVP from the original find (NIASHv2093C22), has also failed to retrieve any hits.

This could be the result of incomplete genomic sequencing (though to affect the same protein in two species—barley and rice—seems unlikely), or that in both cases the mRNA annotated as MVP, was a contaminant. For example, a DNA extract made from an Antarctic moss using RAPDs (random amplified polymorphic DNA) appeared to be very diverse (Skotnicki et al. 2004); however, it turned out that the DNA of the moss extract came from a mixture of three sources (moss, fungi, and protozoa), and so contamination had occurred from animals living in the clumps of moss (Stevens et al. 2007). Reciprocal BLASTp using the barley MVP protein sequence as the query identifies MVP homologs in UniProt and NCBI with greatest similarity to the cellular slime mold *Pol. pallidum* (UniProtKB:D3BF00), and *D. discoideum* MVPα (UniProtKB:P34118); 60% identical residues shared with each. The translated rice MVP cDNA sequence showed homology to *D. discoideum* MVPα with 68% identical amino acids. I-TASSER predicts that the barley and rice sequences fold into the canonical MVP structure with a greater confidence score than even that of the rat sequence (UniProtKB:Q62667). Additionally, RosettaDock docks identical monomers with a superior (lower) energy score to rat (table 1). So if the barley and rice sequences are truly expressed from the plant genomes, then they are compelling MVP sequences, via linear sequence homology, as well as structure and docking analysis and a functional vault is predicted in both species. However, if they are genuine grass (Graminae) MVPs, their phylogenetic placement within amoebozoa seems unlikely.

Within the amoebozoan tree (fig. 3), three species of Cyanobacteria: *Lyngbya majuscula* (UniProtKB:F4Y3B4), *Oscillatoria acuminata* (UniProtKB: K9TKX8), and Microcoleus sp. PCC 7113 (UniProtKB: K9WB38), have homologous sequences described as colicin uptake transmembrane protein that are predicted to fold as MVP. The *Lyngbya* sequence F4Y3B4 is annotated as MVP by InterPro (a membership of 11 protein family databases) (Hunter et al. 2012). The cyanobacterium MVP homologs are more similar in sequence homology to each other (~74%) than to the cellular slime molds (~56%). However, it would not be anticipated that plant or bacterial sequences would group with the amoebozoa. When trees are made of all the individual sequences used in the study, the position of the plant and cyanobacteria sequences remain with the amoebozoa grouping with *Acanthamoeba castelinnii* (UniProtKB:L8GQU5), a free living soil protozoa and occasional human pathogen (supplementary material S1b, Supplementary Material online). In the case of the cyanobacteria, it is possible that horizontal gene transfer (HGT) could be responsible, possibly once from a eukaryote, and then shared between cyanobacteria. Although HGT from eukaryote to prokaryote is rare, there are incidences that have been described (Desmond and Gribaldo 2009; Schönknecht



	F. litoralis…	M. Marina…	O. acumin…	Microcoleu…	L. majuscu…	S. grandis…	P. pacifica…	P. pacifica…	C. coralloi…	H. auranti…
F. litoralis (Bacteroidet…	☒	63.8%	15.4%	15.6%	16.3%	52.6%	52.3%	21.7%	18.1%	10.7%
M. Marina (Bacteroidet…	63.8%	☒	16.4%	15.9%	16.2%	52.8%	52.0%	21.2%	19.0%	10.2%
O. acuminata (Cyanob…	15.4%	16.4%	☒	74.0%	74.5%	14.8%	16.7%	16.8%	31.0%	12.9%
Microcoleus (Cyanoba…	15.6%	15.9%	74.0%	☒	81.4%	15.7%	16.9%	16.1%	31.9%	13.4%
L. majuscula (cyanoba…	16.3%	16.2%	74.5%	81.4%	☒	15.7%	16.8%	16.5%	31.1%	13.3%
S. grandis (Bacteroidet…	52.6%	52.8%	14.8%	15.7%	15.7%	☒	50.7%	21.1%	17.8%	11.3%
P. pacifica2 (Deltaprot…	52.3%	52.0%	16.7%	16.9%	16.8%	50.7%	☒	23.3%	18.8%	11.3%
P. pacifica (Deltaprote…	21.7%	21.2%	16.8%	16.1%	16.5%	21.1%	23.3%	☒	18.4%	10.6%
C. coralloides (Deltapr…	18.1%	19.0%	31.0%	31.9%	31.1%	17.8%	18.8%	18.4%	☒	14.6%
H. aurantiacus (Chloro…	10.7%	10.2%	12.9%	13.4%	13.3%	11.3%	11.3%	10.6%	14.6%	☒

Fig. 4.—A table produced from an alignment of putative bacterial MVP homologs shaded by the distances as a percentage of identical residues between each pair.

et al. 2013). Clearly, it is necessary to be very careful when attributing a total mRNA extract to just a single species. We suspect that the barley and rice homologs are contaminating sequences from unsequenced amoebozoa. This is our hypothesis until further notice.

We cannot totally discount MVP in land plants, even though the barley and rice sequences appear unlikely. Our prediction is that as additional amoebozoa are sequenced, we will find that one of them has an MVP that is closer to, for example, the barley or the rice sequence. There are a few remote candidate plant MVP sequences: *Petunia integrifolia* (UniProtKB:A9XLF3), *Arabidopsis lyrata* (UniProtKB:D7MVK4), *Zea mays* (UniProtKB:B8A0P4), but all fall far short of our criteria for inclusion as MVP. The only MVP sequence from the super group Plantae that falls within our criteria of folding without constraint in I-TASSER with a C score greater than −1.5 and a TM score greater than 0.5, other than rice and barley is a sequence from the single-celled green algae *Chlamydomonas reinhardtii* (UniProtKB:A8JEL9). Owing to the uncertainty around, and low number of, plant MVP sequences barley, rice, and *Chl. reinhardtii* sequences have been included as individual sequences, but not assigned particularly to plants and were used only in the reconstruction ancestor of all eukaryotic MVP (supplementary material S1b, Supplementary Material online).

Although the validity of the cyanobacterium MVP sequences are uncertain, there are a number of putative MVP homologs found in a variety of bacteria with approximately 16% sequence identity with the cyanobacteria putative MVP, but approximately 25% with all other eukaryotic MVP. These bacterial sequences include the following: *Corallococcus coralloides* (UniProtKB:H8MNI3), *Plesiocystis pacifica* SIR-1 (UniProtKB:A6FXM1), and (UniProtKB:A6FXE2), *Microscilla marina* ATCC 23134 (UniProtKB:A1ZGE7), *Saprospira grandis* (UniProtKB:H6L4P8), *Flexibacter litoralis* (UniProtKB:I4AHY9), and *Herpetosiphon aurantiacus* (UniProtKB:A9AUD4). All sequences are predicted to fold into the shape of MVP according to our I-TASSER criteria and are able to dock in accordance with our RosettaDock criteria. Additionally, *Sap. grandis* has been provisionally annotated as MVP and *Fle. litoralis* as MVP shoulder domain containing. The matrix (fig. 4) shows the relationship between the bacterial sequences.

Plesiocystis pacifica is a fruiting gliding bacterium that has a sterol synthesis pathway related to eukaryotes and the genes are likely to have been acquired by HGT (Desmond and Gribaldo 2009). We now find that it also has two copies of a putative MVP homolog. It is a member of the deltaproteobacteria suggested to be a symbiont with a methanogenic archaea, at the root of eukaryotes (López-García and Moreira 1999) suggesting a possible source of ancestral MVP, though the MVP could also have been acquired from a eukaryote by HGT. We did not include bacterial sequences other than the three cyanobacteria in any ancestral MVP sequence reconstruction.

Alveolates

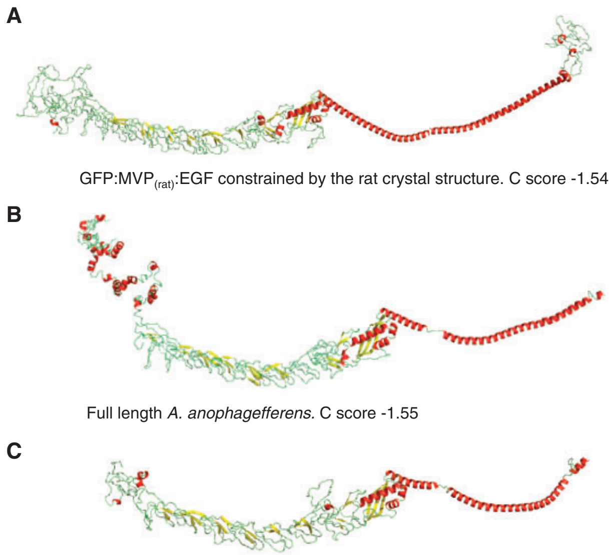
Alveolates fall within the super-group of chromalveolates, or SAR, a group reasoned to be the result of a single endosymbiosis process between a bikont (a protist with two flagella) and a red alga containing a plastid (bestowing the capability of photosynthesis) (Keeling 2004). Although there have been many challenges to the membership of this group; the alveolates and stramenopiles remain core members even though many of the alveolates can no longer photosynthesize (Keeling 2009). So are vault particles also found within the alveolates? *Paramecium tetraurelia* is a well-researched alveolate ciliate that feeds on bacteria, algae, and yeast and has a protein sequence (UniProtKB:A0CI16) containing a domain annotated as MVP shoulder, and is predicted by I-TASSER to adopt the MVP fold with a very high C score of +1.13, RosettaDock confirms that it is likely to oligomerize, with an energy score of −402 (both scores are more favorable than that of the rat). There are three homologous sequences in *P. tetraurelia*, insufficient to calculate an ancestral sequence; however, sequences added October 31, 2012, from the ciliate *O. trifallax* (UniProtKB:J9IML7 and UniProtKB: J9HVS2) are also predicted to fold as MVP. An ancestral MVP sequence was reconstructed from two *O. trifallax* and three paramecium sequences. Two of the paramecium sequences appear to be fairly recent duplications (UniProtKB:A0CI16 and A0DWW7) with 95% amino acid identity but the third (UniProtKB: A0EGV2) shows only 42% identity with the other two (supplementary material S2g, Supplementary Material online).

Stramenopiles

The other main group of the chromoalveolates (SAR) are the stramenopiles, with their ancestral sequence reconstructed from five sequences, four from oomycetes: *Phytophthora infestans* (potato blight—annotated as MVP; UniProtKB: D0N745)*, Phytophthora sorgae* (soybean stem and root rot; UniProtKB:G4Z1M3), and *Phytophthora ramorum* (sudden oak death; UniProtKB:H3G9I8), *Pythium ultimum* (a plant pathogen of many food crops and grasses; UniProtKB:K3X224), and *Aureococcus anophagefferens* (UniProtKB:F0YA32), a harmful algal bloom (supplementary material S2g, Supplementary Material online). The *Aur. anophagefferens* sequence is clearly different to the other stramenopiles, it has approximately 16% similarity with the other stramenopile sequences; however, the fold predicted by I-TASSER is very similar.

The highest scoring I-TASSER model for the complete *Aur. anophagefferens* sequence had a C score of −1.55, that is, just outside our cut off score of −1.5. However, one of the reasons for a lower than anticipated C score is the extension of the sequence either at the C or N terminal beyond the rat MVP crystal structure template. Extensions to the core MVP sequence do not necessarily prevent vault formation because vault particles form with Green Fluorescent Protein (GFP) fused to the N terminus of MVP (van Zon et al. 2003), and tags, for example, epidermal growth factor (EGF), added to the C terminal to direct the particle to particular cells (Kickhoefer et al. 2009). Indeed, when the sequence of such an engineered protein (GFP:MVP$_{(rat)}$:EGF) was submitted to I-TASSER the C score was much lower than our cut off score of −1.5, and the highest C score model predicted did not look convincingly like an MVP. Even when constrained by the rat crystal structure the C score was only −1.54, still low compared with 0.42 for the rat sequence (Q62667) alone (fig. 5*A*). When the *Aur. anophagefferens* sequence was resubmitted with the N terminal non MVP-like domain truncated it resulted in a score of −0.70, well above our threshold score (fig. 5*B* and *C*).

There are three points to be made here: first, that the C score is affected by the extended sequence presumably because it lacks a template for modeling. Second, the extra sequence may interfere with in silico docking if not predicted to fold correctly. We have confined docking between two monomers to the shoulder and coiled coil regions, as the coil was found to be critical to the vault formation in a yeast two-hybrid system (van Zon et al. 2002) and our previous work shows that pairs of the repeat domain region, in general, will readily dock along their length (Daly et al. 2013). Finally, the *Aur. anophagefferens* sequence has greater homology with the green algal MVP sequence used to root the alveolate and stramenopile tree (supplementary material S2g, Supplementary Material online) (26% full length, 35% with the N terminal removed) than with any of the stramenopile or alveolate sequences, full length or truncated. However, structurally I-TASSER predicts *Aur. anophagefferens* MVP to fold more similarly to the other chromalveolate MVPs. The stramenopile ancestral MVP structure is unaffected by the inclusion of the algal bloom sequence with minimal primary sequence homology (fig. 6).



GFP:MVP$_{(rat)}$:EGF constrained by the rat crystal structure. C score -1.54

Full length *A. anophagefferens.* C score -1.55

A. anophagefferens with 250 N terminal residues removed. C score -0.70

Fig. 5.—(*A*) GFP:MVP$_{(rat)}$:EGF (1,152 residues and constrained by the rat crystal structure 2ZUO*b). The C score was −1.54 low compared with +0.42 for the rat sequence (UniProtKB:Q62667) alone. (*B*) *Aureococcus anophagefferens* (UniProtKB:F0YA32) complete sequence 962 residues submitted to I-TASSER without constraint and resulted with a C score of −1.55. (*C*) The same sequence with the N terminal 250 residues removed – resulted with a C score of −0.70.

The inclusion of rhizaria as part of the super-group with chromalveolates (known as SAR) is becoming more compelling (Burki et al. 2010; Parfrey et al. 2010). Rhizaria are difficult to culture and consequently underrepresented in sequence databases (Sierra et al. 2013), BLASTing the few genomes sequenced thus far has not retrieved any sequences that resemble MVP.

Finally, a tree was made from both the ancestors (where possible) and the individuals that represented poorly covered families (supplementary material S3a, Supplementary Material online). Initially, the ancestor was comprised of all our eukaryote data set plus the three cyanobacteria—which we had identified as either contamination or gained from eukaryote via HGT. This fulfilled all of our criteria but it could be argued that the number of kinetoplast sequences that were included influenced the resultant ancestral sequence. We therefore limited the number of sequences to one per species. Additionally—because of the issues that affected the output

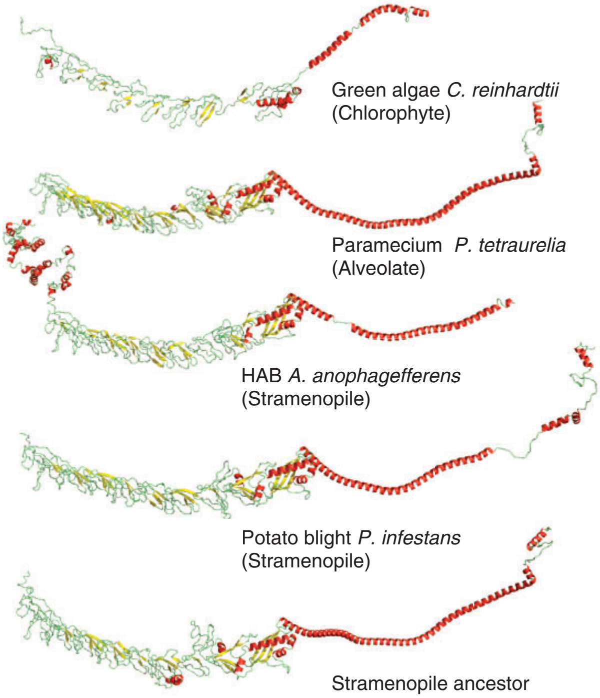


Fig. 6.—Although *Aureococcus anophagefferens* has greater sequence similarity with the green algae, *Chlamydomonas*, I-TASSER predicts that structurally it is more similar to either the ciliate paramecium or to the oomycete *Phytophthora infestans.* The stramenopile ancestor is unaffected structurally by the inclusion of *A. anophagefferens* even though it has very low primary sequence homology with the oomycetes sequences.

from each of the ASR algorithms that we used, making them either unrealistically long (FastML) or unrealistically short (PAML4)—we made ancestors by removing from MSAs, inserts that were present in only one ancestor (columns 2–5), or by deleting inserts represented by just one genus in the MSA of the individuals (columns 6–9). This resulted in sequences that were of a more likely length (number of residues shown) because this removed most of the gaps that the various algorithms dealt with in different ways.

1. Sequences derived from an MSA of the ancestors for the five major groups; amoebozoa, opisthokonts, kinetoplasts, alveolates, and stramenopiles—FastML joint (856 residues) columns 2 and FastML marginal (856 residues) column 3 in figure 7*A*, Mega5 (853 residues) column 4, PAML (819 residues) column 5. Rat is included in column 1 for comparison.
2. Ancestral sequences reconstructed from individual species using an MSA limited to one sequence per species, with inserts unique to a single genus removed—resulting in FastML joint (892 residues) and marginal (892 residues) sequences columns 6 and 7, Mega5 (770 residues) column 8, and PAML (679 residues) column 9.

Our main point is that regardless of how the ancestor is reconstructed, whether from ancestral sequences from each major group or from sequences from individual species used all together to make an ancestor, the resultant protein sequence folds and docks within the constraints of our original criteria.

Although the sequences had reduced overall similarity, there were blocks of highly conserved sequence (alignment supplementary material S3b, Supplementary Material online). Particularly highly conserved is a sequence region close to the C terminus (fig. 7*C*). The crystal structure for this region has not been resolved but ab initio modeling by MODELLER (Sali and Blundell 1993), part of the I-TASSER suite of programs, consistently predicts the structure depicted in figure 7*D*. This fold was also found by Phyre^2 (Kelley and Sternberg 2009), which also retrieved known MVP structures.

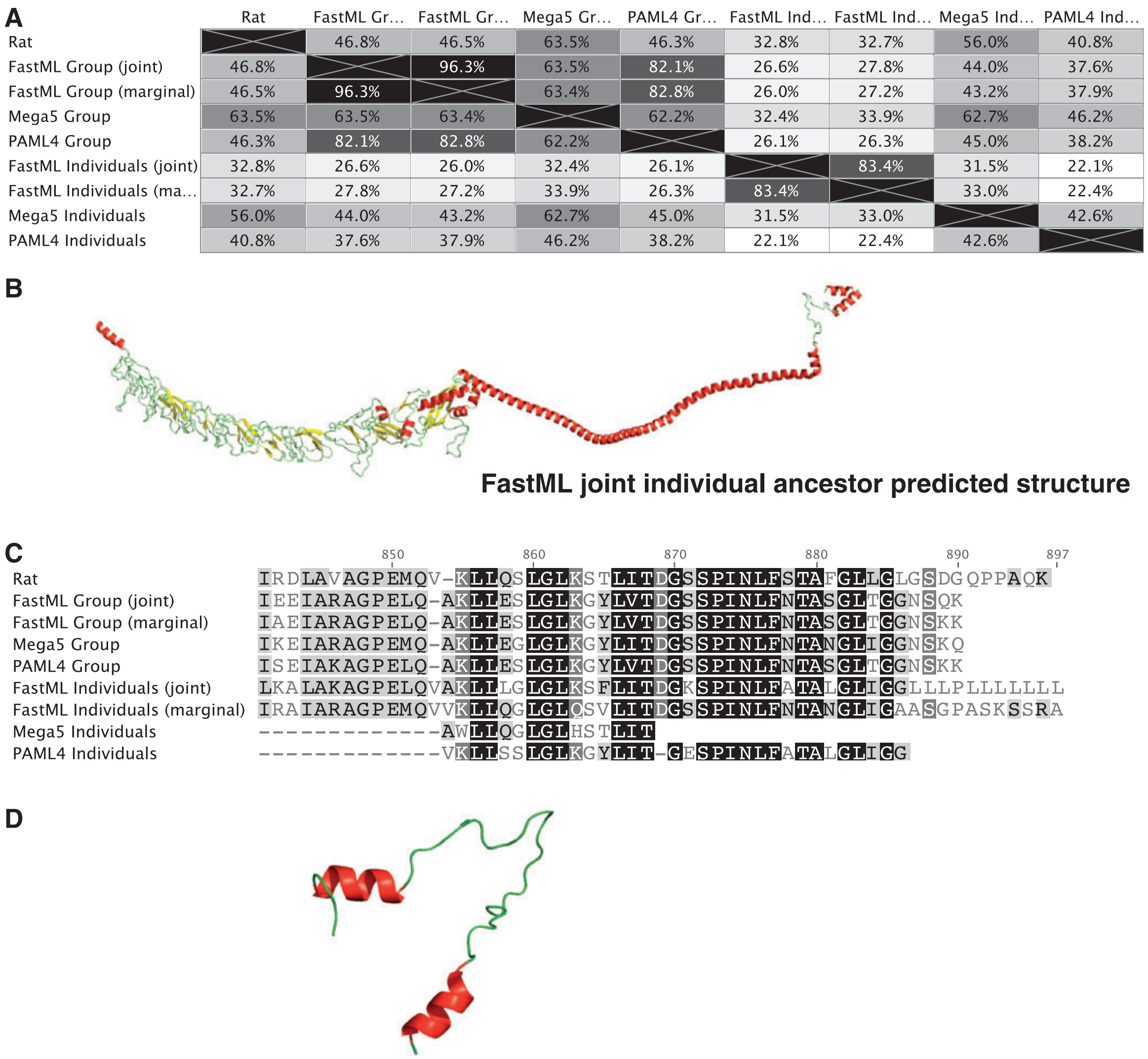
A BLASTp using just these conserved sequences resulted in hits only from known MVP sequences no matter how loose the parameters were. A structural search with the Dali server (Holm and Rosenström 2010) also failed to retrieve any other models with similar folds. This indicates that this sequence is found only in MVP. Could this define MVP? It could be essential for sealing the cap, or to hook the vaults onto cellular structures, vaults have been shown to bind to microtubules via their caps (Eichenmüller et al. 2003). This addition could have expanded the function of the vault from sequestration of ions or molecules to transportation.

This gives us an interesting dilemma. Even though the PAML4 individual ancestor is the shortest, it is Mega5 that has left out the very most highly conserved region of the alignment when asked to include 100% of the MSA. This is because the sequences from the branches (*Chl. reinhardtii, N. gruberi,* and *Aur. anophagefferens*) do not have the C terminal folds that appear to be specific to vaults. This could be correct and the extra sequence has arisen more recently. However, this is at odds with Cavalier–Smiths' latest version of the root of eukaryotes because extant vault MVP, from both sides of the proposed root, has this structure. It is unlikely to have arisen twice because the primary sequence is so highly conserved across all domains.

A summary of our results is shown in figure 8 based on a eukaryote tree (Keeling et al. 2005).

The Consurf representation (fig. 9*A* and *B*) (Ashkenazy et al. 2010) shows the nonconserved residues (blue), and the highly conserved residues (red) from an MSA of MVP sequences from all species discussed. The nonconserved residues are generally either solvent exposed on the exterior surface of the vault or line the interior, but are not those involved with inter molecular contacts docking monomers for vault formation. The conserved residues cluster around the shoulder of the vault, and also along the length of each monomer within the lateral contacts.

In general, the docking is relatively poorer amongst the ancestors (table 1) than docking in the individual sequences that made up the original ASR input. This is to be anticipated; a core MVP fold is conserved with sequence variation existing



	Rat	FastML Gr...	FastML Gr...	Mega5 Gr...	PAML4 Gr...	FastML Ind...	FastML Ind...	Mega5 Ind...	PAML4 Ind...
Rat		46.8%	46.5%	63.5%	46.3%	32.8%	32.7%	56.0%	40.8%
FastML Group (joint)	46.8%		96.3%	63.5%	82.1%	26.6%	27.8%	44.0%	37.6%
FastML Group (marginal)	46.5%	96.3%		63.4%	82.8%	26.0%	27.2%	43.2%	37.9%
Mega5 Group	63.5%	63.5%	63.4%		62.2%	32.4%	33.9%	62.7%	46.2%
PAML4 Group	46.3%	82.1%	82.8%	62.2%		26.1%	26.3%	45.0%	38.2%
FastML Individuals (joint)	32.8%	26.6%	26.0%	32.4%	26.1%		83.4%	31.5%	22.1%
FastML Individuals (ma...	32.7%	27.8%	27.2%	33.9%	26.3%	83.4%		33.0%	22.4%
Mega5 Individuals	56.0%	44.0%	43.2%	62.7%	45.0%	31.5%	33.0%		42.6%
PAML4 Individuals	40.8%	37.6%	37.9%	46.2%	38.2%	22.1%	22.4%	42.6%	

```
                                         850       860       870       880       890      897
Rat                          IRDLAVAGPEMQV-KLLQSLGLKSTLITDGSSPINLFSTAFGLLGLGSDGQPPAQK
FastML Group (joint)         IEEIARAGPELQ-AKLLESLGLKGYLVTDGSSPINLFNTASGLTGGNSQK
FastML Group (marginal)      IAEIARAGPELQ-AKLLESLGLKGYLVTDGSSPINLFNTASGLTGGNSKK
Mega5 Group                  IKEIARAGPEMQ-AKLLEGLGLKGYLITDGSSPINLFNTANGLIGGNSKQ
PAML4 Group                  ISEIAKAGPELQ-AKLLESLGLKGYLVTDGSSPINLFNTASGLTGGNSKK
FastML Individuals (joint)   LKALAKAGPELQVAKLLLGLGLKSFLITDGKSPINLFATALGLIGGLLLPLLLLLLL
FastML Individuals (marginal) IRAIARAGPEMQVVKLLQGLGLQSVLITDGSSPINLFNTANGLIGAASGPASKSSRA
Mega5 Individuals            --------------AWLLQGLGLHSTLIT
PAML4 Individuals            --------------VKLLSSLGLKGYLIT-GESPINLFATALGLIGG
```

Fig. 7.—(*A*) Table produced from MSA of ancestral sequences of the super-groups identified in the text and from the alignment of the ancestors made by individuals (one per species). Rat has been included for comparison. (*B*) The I-TASSER structural prediction for the reconstruction of the ancestor from 89 individual sequences, this cartoon depicts the FastML reconstruction that bears least sequence similarity with either the rat or with the other ancestors. (*C*) The MSA close to the C terminal identifying an area of very high conservation. (*D*) Cartoon diagram of this region modeled by I-TASSER utilizing the ab initio modeling capacity of MODELLER as this area has not been resolved in the crystal structure.

amongst groups that still allows interface docking, possibly through covariance of sites between monomers within particular species. Our RosettaDock analysis with the known rat structure (2ZUO monomers) indicates redundancy in docking possibilities, that is, the docked rat monomer pairs were not all utilizing the same residues as those found in the solved crystal structure (Daly et al. 2013). If, as expected, the mutation rate was equivalent at all positions within the MVP sequence positions, then the docking of one MVP monomer for another would quickly deteriorate so there must be selective pressure to maintain the residues important for docking even if the vault structure was ancestrally rather simpler.

Discussion

From our general approach of reconstructing tertiary structures, and inferring quaternary formations, the MVP gene appears to be ancestral to eukaryotes and it is likely that vault particles were present in LECA. MVP is retained in most eukaryote super-groups; opisthokonts (fungi plus animals), amoebozoa, chromalveolates (though we are not so sure about rhizaria that have been latterly included with the chromalveolates in the SAR supergroup), and excavates distributed in groups both sides of the proposed initial divergence of the last common eukaryote ancestor (Cavalier-Smith 2010). Plantae is rather more controversial, although *Chl. reinhardtii*

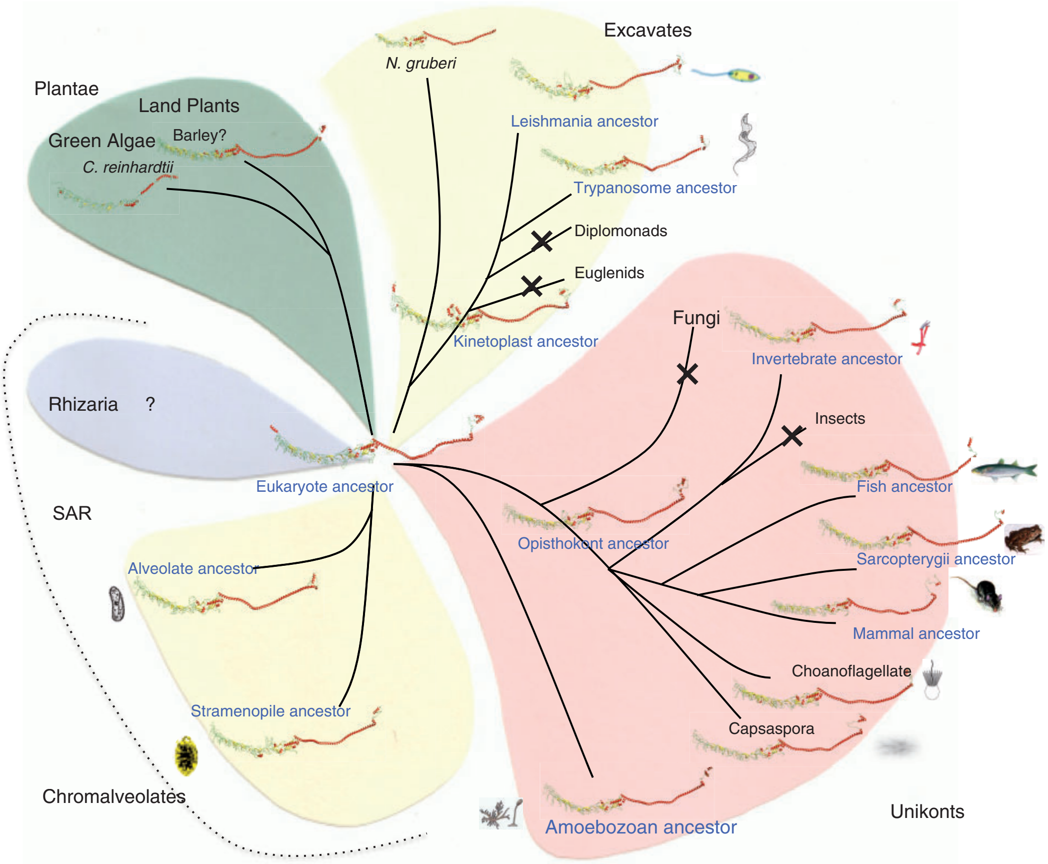


Fig. 8.—Structural diagrams of I-TASSER predictions from individual extant sequences (black type face) and from reconstructed sequences, derived from a combination of PAML4 and FastML ASR (blue type face).

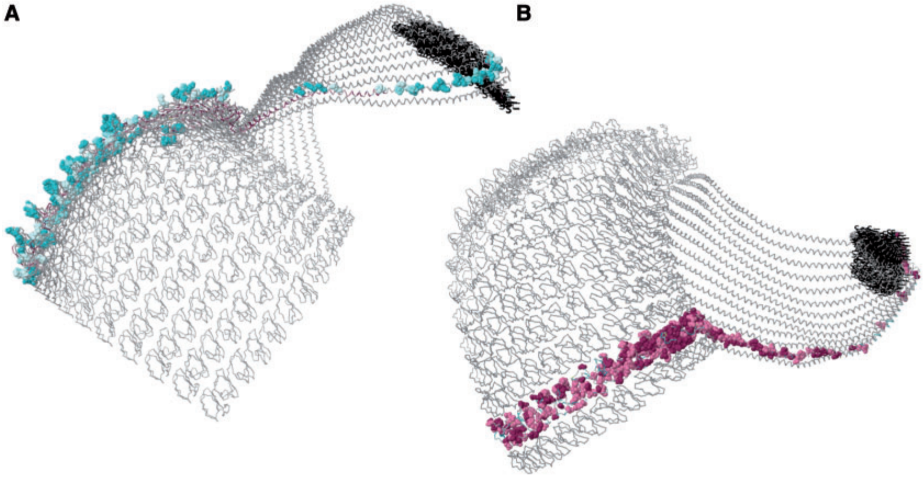


Fig. 9.—Consurf diagrams of the structural back bone of oligomerized rat MVP (PDB:2ZUO) showing one chain with spheres depicting the similarity score of the MSA for all sequences. Thirteen monomers (of a total 39) of a half rat vault are shown for clarity. (*A*) Nonconserved residues are shown as blue spheres. These nonconserved residues stick out from the surface of the vault (outward or inward), but are not involved in docking. (*B*) Shows completely conserved residues (red) and highly conserved residues (pink). The most conserved residues are found in the shoulder region and along the length of the monomer within the lateral contacts. The extreme C terminal is also highly conserved but unresolved in the crystal structure (2ZUO), those residues show as black points.

Table 1

I-TASSER and RosettaDock Results for Individuals from Poorly Represented Groups and from ASRs

Accession Number	Organism	Length	% Residues Identical to Rat	I-Tasser TM Score,[a] Max Is 1.0	I-Tasser C Score[b] (Range −5 +2)	RosettaDock Energy Score[c]
Q62667[d]	*R. norvegicus* (Rat)	861	100	0.77 ± 010	0.42	−254
Extant sequences used individually						
A9V809	*Monosiga brevicollis* (Choanoflagellate)	861	59	0.90 ± 0.06	1.34	−515
F2UN76	Salpingocea (Choanoflagellate)	853	59	0.86 ± 0.07	1.07	−474
E9CE06	*Capsaspora owczarzaki* (Capsaspora)	860	63	0.82 ± 0.09	0.77	−199
F2E078	*Hordeum vulgare* (Barley)	843	55	0.86 ± 0.07	1.06	−496
CT836653	*Oryza sativa* from cDNA (Rice)	831	60	0.90 ± 0.06	1.37	−511
F4Y3B4	*Lyngbya majuscula* (Cyanobacteria)	879	54	0.77 ± 0.10	0.43	−440
D2V5B9	*N. gruberi* (Heterolobosea)	559	17	0.62 ± 0.14	−0.74	−441
A8JEL9	*Chlamydomonas reinhardtii* (Chlorophyte)	529	17	0.62 ± 0.14	−0.86	−17
Ancestors created from combined PAML and FastML ASR						
ASR	All Eukaryotes[e]	892	32	0.79 ± 0.09	0.56	−156
ASR	Stramenopiles	912	45	0.74 ± 0.11	0.2	−208
ASR	Alveolate	871	45	0.88 ± 0.0	1.18	−161
ASR	Leishmania	995	34	0.58 ± 0.14	−1.06	−180
ASR	Trypanosome	916	38	0.81 ± 0.09	0.69	−205
ASR	Kinetoplast	1,025	34	0.53 ± 0.15	−1.46	−148
ASR	Amoebozoa	859	55	0.80 ± 0.09	0.67	−248
ASR	Opisthokont	913	65	0.96 ± 4.6	−0.38	−174
ASR	Invertebrate	853	67	0.90 ± 0.06	1.34	−261
ASR	Fish	887	67	0.88 ± 0.07	1.24	−519
ASR	Sarcopterygii	857	72	0.87 ± 0.07	1.11	−226
ASR	Mammal	945	79	0.78 ± 0.10	0.51	−238

Note.—The rat is given at the beginning for comparison; it is the only vault for which its 3D structure is known.

[a] I-TASSER TM score higher is better—cut off is −1.5.

[b] I-TASSER C score higher is better—cut off is 0.5.

[c] RosettaDock energy score lower is better.

[d] Q62667 is included for comparison. Rat MVP is the only complete MVP 3D X-ray crystallographic structure in the Protein Data Bank.

[e] All Eukaryotes. These figures refer to the joint ancestor of all individuals (one sequence per species), all final ancestors, scored within our criteria.

seems to be a bona fide inclusion, the grasses look more like contamination. Additionally, *Chl. reinhardtii*, *Aur. anophagefferens,* and *N. gruberi* do not have the highly conserved helices and loop at the C terminal. Although we have concentrated, for obvious reasons, on the MVPs, we see prediction of 3D structure as a general approach that could be used much more frequently for understanding earlier phases of molecular evolution.

MVP is under selective pressure to maintain structure and appropriate residues for docking. In general, the docking scores for the putative ancestors are lower than for the extant sequences though our prediction would still be that these sequences would be capable of self-assembling into a vault particle as they are known to do in metazoa.

Is there any evidence that could support this assertion? If we consider the proteins that are known to associate with vaults there are differences even within extant species. Vault poly ADP-ribose polymerase from the PARP protein family (VPARP) is found only in metazoa and amoebozoa, and therefore seems like a more recent adaptation (Citarelli et al. 2010). These authors found six clades of PARP protein and suggest that LECA already had at least proteins from clades 1 and 6. The only MVP sequence that we have used that comes from a species where no PARP family member has been found is *Aur. anophagefferens;* although vaults can form without PARP (Stephen et al. 2001).

Similarly, TEP1 (telomerase-associated protein-1) is a component of vault particles in metazoa and amoebozoa. TEP1 contains a TROVE domain (Telomerase, Ro, and Vault), that binds vault associated RNA (vtRNA) (Poderycki et al. 2005). TEP1 is ubiquitous but vaults form without it and without vtRNA. In metazoa, the only characterized group, approximately 80% of the vtRNA is found outside of the vault (Kickhoefer et al. 1998). The sequence homology of vtRNA—even within metazoa, is slim (Stadler et al. 2009). It is a pol III transcribed RNA and outside of the A and B box regions, structural homology would be the best search method to find it in other groups.

If vault particles were formed in LECA—with or without any other association—what functional role could they have played? Extant vault particles open at low pH (Goldsmith et al. 2009) and anions can enter, possibly attracted by positively charged amino acids facing the vault interior (Ng et al. 2008). Vaults have been associated with detoxification processes (Suprenant et al. 2007), though they have never been proven to be vital (Herlevsen et al. 2007). Some kind of early encapsulation of substances toxic to the cell would be a desirable trait. One possible function could be protection from harmful bacteria that are engulfed by eukaryotes.

Vault particles are probably missing from plants, have not been found in insects and although traces of MVP monomer sequences appear in some fungi, they fall short of having vault forming capability using our criteria (not shown). The loss of vault particles in plants and fungi might be explained because they do not normally consume bacteria? Again, their loss in insects might be explained by their hosting of complex communities of bacterial, fungal and viral symbionts when feeding on plants hosting pathogens and producing toxic chemicals (Frago et al. 2012) that would be protective without the need for vault particles.

Protein compartments that encapsulate and compartmentalize contents are ubiquitous, although a variety of designs are utilized. In many ways vaults are reminiscent of virus particles; they are large assemblies that have a protein shell composed of multiple copies of a single protein and have a large central cavity. However, the geometry of viruses can be classified as; icosahedral, helical or complex (the classification given to the pox virus), but none have the radially symmetrical halves joined together like the vault particle.

Prokaryotes also form compartments, both larger and smaller than the vault (Heinhorst and Cannon 2008) that concentrate linked functional mechanisms; however, these are mostly icosahedral, for example, carboxysomes. Although vault particles were originally thought to be absent from prokaryotes, there are a number of convincing homologs which could have been acquired by HGT from eukaryotes. However, there are other proteins, ubiquitous in prokaryotes that have sequence and structural similarities with MVP in whole or in part. BLASTs with the rat MVP sequence repeatedly result in TolA and band 7 protein homologs being identified within default parameters. In fact, the cyanobacterial MVP homologs have been annotated colicin uptake transmembrane protein, which is a pathway that utilizes TolA. The mechanism for colicin uptake has been mostly studied in *Escherichia coli* and comprises the Tol/Pal system. The function of TolA is not fully understood, it is involved in the structural integrity of *E. coli* and related bacterial cell membranes. It is also involved in active transport across the membrane but can be parasitized by colicins produced by other *E. coli* resulting in the death of the cell (Li et al. 2012). The Tol system also allows uptake of phage DNA, although generally deleterious imported DNA may contain genes that could give the cell an advantage. It seems unlikely that the Tol/Pal system would be retained specifically for the uptake of pathogens, but conservation of an active, if promiscuous transport system, might have been essential to the early eukaryote.

One of the bacterial sequences, *Her. aurantiacus* (UniProtKB:A9AUD4), that folds as MVP within our criteria is annotated as a band 7 protein. These band 7 sequences are ubiquitous proteins that include stomatins, prohibitin, flotillin HlfK/C, and podicin, known collectively as SPFH domain-containing proteins. Tanaka et al. (2009) identified the shoulder domain of MVP as homologous to the stomatin core from *Pyroccoccus horikoshii.* Band 7 proteins have been found to form ring-like oligomeric structures, for example, membrane-bound prohibitin rings in mitochondria (Tatsuta et al. 2005), and free ring structures in cyanobacteria (Boehm et al. 2009). SPFH domain proteins are often linked with lipid rafts (Browman et al. 2007). Extant vault particles are also found in association with lipid rafts (Kowalski et al. 2007). Vault particles are capable of detoxification of anions (Suprenant et al. 2007), and are linked with multi drug resistance in both cancer and epilepsy (Herlevsen et al. 2007; Liu, Mao, et al. 2011). The capacity for sequestration or even ejection of toxins from the early eukaryote would be a reason for the high level of conservation.

Conclusion

MVP has been identified by our Ancestral Sequence Reconstruction methods in; opisthokonts, amoebozoa, excavates (including euglenids), chromalveolates, bacteria, and possibly plants. We additionally predict that these MVP monomers could dock to form complex oligomeric vaults as they are known to do in opisthokonts and amoebozoa. We propose that vaults in LECA could have functioned in membrane transport, the sequestering of cell toxins, or provide protection from engulfing pathogenic bacteria, but have now diversified into the multitude of roles seen today, to the point where they are being harnessed and utilized for drug and vaccine delivery and possible future bioremediation.

Supplementary Material

Supplementary materials S1–S3 are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

T.K.D. performed the analyses and wrote the drafts; A.J.S.-S. assisted with the analysis programs; and D.P. designed the original research project. All authors regularly discussed the results, and contributed to the final manuscript. This work was supported by the Institute of Fundamental Sciences, Massey University.

Literature Cited

Ashkenazy H, Erez E, Martz E, Pupko T, Ben-Tal N. 2010. ConSurf 2010: calculating evolutionary conservation in sequence and structure of proteins and nucleic acids. Nucleic Acids Res. 38:W529–W533.

Ashkenazy H, et al. 2012. FastML: a web server for probabilistic reconstruction of ancestral sequences. Nucleic Acids Res. 40:W580–W584.

Berger W, Steiner E, Grusch M, Elbling L, Micksche M. 2009. Vaults and the major vault protein: novel roles in signal pathway regulation and immunity. Cell Mol Life Sci. 66:43–61.

Boehm M, et al. 2009. Structural and mutational analysis of band 7 proteins in the cyanobacterium *Synechocystis* sp. strain PCC 6803. J Bacteriol. 191:6425–6435.

Browman DT, Hoegg MB, Robbins SM. 2007. The SPFH domain-containing proteins: more than lipid raft markers. Trends Cell Biol. 17: 394–402.

Buehler DC, Toso DB, Kickhoefer VA, Zhou ZH, Rome LH. 2011. Vaults engineered for hydrophobic drug delivery. Small 7:1432–1439.

Burki F, et al. 2007. Phylogenomics reshuffles the eukaryotic supergroups. PLoS One 2(8):e790.

Burki F, et al. 2010. Evolution of Rhizaria: new insights from phylogenomic analysis of uncultivated protists. BMC Evol Biol. 10:377.

Cavalier-Smith T. 2010. Kingdoms Protozoa and Chromista and the eozoan root of the eukaryotic tree. Biol Lett. 6:342–345.

Champion CI, et al. 2009. A vault nanoparticle vaccine induces protective mucosal immunity. PLoS One 4(4):e5409.

Chang BSW, Jönsson K, Kazmi MA, Donoghue MJ, Sakmar TP. 2002. Recreating a functional ancestral archosaur visual pigment. Mol Biol Evol. 19:1483–1489.

Chugani DC, Kedersha NL, Rome LH. 1991. Vault immunofluorescence in the brain: new insights regarding the origin of microglia. J Neurosci. 11:256–268.

Citarelli M, Teotia S, Lamb RS. 2010. Evolutionary history of the poly(ADP-ribose) polymerase gene family in eukaryotes. BMC Evol Biol. 10:308.

Collins L, Penny D. 2005. Complex spliceosomal organization ancestral to extant eukaryotes. Mol Biol Evol. 22:1053–1066.

Collins LJ, Poole AM, Penny D. 2003. Using ancestral sequences to uncover potential gene homologues. Appl Bioinformatics. 2:S85–S95.

Daly TK, Sutherland-Smith AJ, Penny D. 2013. Beyond BLASTing: tertiary and quaternary structure analysis helps identify major vault proteins. Genome Biol Evol. 5:217–232.

Denoeud F, et al. 2010. Plasticity of animal genome architecture unmasked by rapid evolution of a pelagic tunicate. Science 330: 1381–1385.

Desmond E, Gribaldo S. 2009. Phylogenomics of sterol synthesis: insights into the origin, evolution, and diversity of a key eukaryotic feature. Genome Biol Evol. 1:364–381.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32:1792–1797.

Eichenmüller B, et al. 2003. Vaults bind directly to microtubules via their caps and not their barrels. Cell Motil Cytoskeleton. 56:225–236.

Elias M, Patron NJ, Keeling PJ. 2009. The RAB family GTPase Rab1A from *Plasmodium falciparum* defines a unique paralog shared by chromalveolates and rhizaria. J Eukaryot Microbiol. 56:348–356.

Frago E, Dicke M, Godfray HCJ. 2012. Insect symbionts as hidden players in insect-plant interactions. Trends Ecol Evol. 27:705–711.

Fritz-Laylin LK, et al. 2010. The genome of *Naegleria gruberi* illuminates early eukaryotic versatility. Cell 140:631–642.

Goldsmith LE, Pupols M, Kickhoefer VA, Rome LH, Monbouquette HG. 2009. Utilization of a protein "shuttle" to load vault nanocapsules with gold probes and proteins. ACS Nano. 3:3175–3183.

Gray JJ, et al. 2003. Protein-protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations. J Mol Biol. 331:281–299.

Gullberg M, et al. 2010. Characterization of a putative ancestor of coxsackievirus B5. J Virol. 84:9695–9708.

Hanson-Smith V, Kolaczkowski B, Thornton JW. 2010. Robustness of ancestral sequence reconstruction to phylogenetic uncertainty. Mol Biol Evol. 27:1988–1999.

Heinhorst S, Cannon GC. 2008. A new, leaner and meaner bacterial organelle. Nat Struct Mol Biol. 15:897–898.

Herlevsen M, Oxford G, Owens CR, Conaway M, Theodorescu D. 2007. Depletion of major vault protein increases doxorubicin sensitivity and nuclear accumulation and disrupts its sequestration in lysosomes. Mol Cancer Ther. 6:1804–1813.

Holm L, Rosenström P. 2010. Dali server: conservation mapping in 3D. Nucleic Acids Res. 38:W545–W549.

Huelsenbeck JP, Ronquist F. 2001. MRBAYES: Bayesian inference of phylogeny. Bioinformatics 17:754–755.

Hunter S, et al. 2012. InterPro in 2011: new developments in the family and domain prediction database. Nucleic Acids Res. 40: D306–D312.

Ishikawa F, Naito T. 1999. Why do we have linear chromosomes? A matter of Adam and Eve. Mutat Res. 434:99–107.

Keeling PJ. 2004. Diversity and evolutionary history of plastids and their hosts. Am J Bot. 91:1481–1493.

Keeling PJ. 2009. Chromalveolates and the evolution of plastids by secondary endosymbiosis. J Eukaryot Microbiol. 56:1–8.

Keeling PJ, et al. 2005. The tree of eukaryotes. Trends Ecol Evol. 20: 670–676.

Kelley LA, Sternberg MJ. 2009. Protein structure prediction on the Web: a case study using the Phyre server. Nat Protoc. 4:363–371.

Kickhoefer VA, et al. 1998. Vaults are up-regulated in multidrug-resistant cancer cell lines. J Biol Chem. 273:8971–8974.

Kickhoefer VA, et al. 2009. Targeting vault nanoparticles to specific cell surface receptors. ACS Nano. 3:27–36.

Klaus FX. 2012. A physical, genetic and functional sequence assembly of the barley genome. Nature 491:711–716.

Koonin EV. 2010. The incredible expanding ancestor of eukaryotes. Cell 140:606–608.

Kowalski MP, et al. 2007. Host resistance to lung infection mediated by major vault protein in epithelial cells. Science 317: 130–132.

Li C, et al. 2012. Structural evidence that colicin A protein binds to a novel binding site of TolA protein in *Escherichia coli* periplasm. J Biol Chem. 287:19048–19057.

Liu JL, Mao ZY, Gallick GE, Yung WKA. 2011. AMPK/TSC2/mTOR-signaling intermediates are not necessary for LKB1-mediated nuclear retention of PTEN tumor suppressor. Neuro Oncol. 13:184–194.

Liu B, et al. 2011. Up-regulation of major vault protein in the frontal cortex of patients with intractable frontal lobe epilepsy. J Neurol Sci. 308: 88–93.

López-García P, Moreira D. 1999. Metabolic symbiosis at the origin of eukaryotes. Trends Biochem Sci. 24:88–93.

Lyskov S, Gray JJ. 2008. The RosettaDock server for local proteinprotein docking. Nucleic Acids Res. 36:233–238.

Matsumoto T, et al. 2011. Comprehensive sequence analysis of 24,783 Barley full-length cDNAs derived from 12 clone libraries. Plant Physiol. 156:20–28.

Mossel E, Steel M. 2005. How much can evolved characters tell us about the tree that generated them? Oxford: Oxford University Press.

Ng BC, et al. 2008. Encapsulation of semiconducting polymers in vault protein cages. Nano Lett. 8:3503–3509.

Nosek J, Kosa P, Tomaska L. 2006. On the origin of telomeres: a glimpse at the pre-telomerase world. BioEssays 28:182–190.

Parfrey LW, et al. 2010. Broadly sampled multigene analyses yield a well-resolved eukaryotic tree of life. Syst Biol. 59:518–533.

Paspalas CD, et al. 2009. Major vault protein is expressed along the nucleus-neurite axis and associates with mRNAs in cortical neurons. Cereb Cortex. 19:1666–1677.

Philippe H, et al. 2011. Resolving difficult phylogenetic questions: why more sequences are not enough. PLoS Biol. 9(3):1–10.

Poderycki MJ, Rome LH, Harrington L, Kickhoefer VA. 2005. The p80 homology region of TEP1 is sufficient for its association with the telomerase and vault RNAs, and the vault particle. Nucleic Acids Res. 33: 893–902.

Rodríguez-Ezpeleta N, et al. 2007. Toward resolving the eukaryotic tree: the phylogenetic positions of Jakobids and Cercozoans. Curr Biol. 17: 1420–1425.

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

Sali A, Blundell TL. 1993. Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol. 234:779–815.

Schönknecht G, et al. 2013. Gene transfer from bacteria and archaea facilitated evolution of an extremophilic eukaryote. Science 339: 1207–1210.

Sebe-Pedros A, de Mendoza A, Lang BF, Degnan BM, Ruiz-Trillo I. 2011. Unexpected repertoire of metazoan transcription factors in the unicellular holozoan *Capsaspora owczarzaki*. Mol Biol Evol. 28:1241–1254.

Sierra R, et al. 2013. Deep relationships of Rhizaria revealed by phylogenomics: a farewell to Haeckel's Radiolaria. Mol Phylogenet Evol. 67: 53–59.

Skotnicki ML, Mackenzie AM, Ninham JA, Selkirk PM. 2004. High levels of genetic variability in the moss *Ceratodon purpureus* from continental Antarctica, subantarctic Heard and Macquarie Islands, and Australasia. Polar Biol. 27:687–698.

Stadler PF, et al. 2009. Evolution of vault RNAs. Mol Biol Evol. 26: 1975–1991.

Stechmann A, Cavalier-Smith T. 2003. The root of the eukaryote tree pinpointed. Curr Biol. 13:R665–R666.

Stephen AG, et al. 2001. Assembly of vault-like particles in insect cells expressing only the major vault protein. J Biol Chem. 276:23217–23220.

Stevens MI, Hunger SA, Hills SFK, Gemmill CEC. 2007. Phantom hitchhikers mislead estimates of genetic variation in Antarctic mosses. Plant Syst Evol. 263:191–201.

Stewart PL, et al. 2005. Sea urchin vault structure, composition, and differential localization during development. BMC Dev Biol. 5:3.

Suprenant KA. 2002. Vault ribonucleoprotein particles: Sarcophagi, gondolas, or safety deposit boxes? Biochemistry 41:14447–14454.

Suprenant KA, Bloom N, Fang JW, Lushington G. 2007. The major vault protein is related to the toxic anion resistance protein (TelA) family. J Exp Biol. 210:946–955.

Tabach Y, et al. 2013. Identification of small RNA pathway genes using patterns of phylogenetic conservation and divergence. Nature 493: 694–698.

Tamura K, et al. 2011. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol Biol Evol. 28:2731–2739.

Tanaka H, et al. 2009. The structure of rat liver vault at 3.5 Angstrom resolution. Science 323:384–388.

Tatsuta T, Model K, Langer T. 2005. Formation of membrane-bound ring complexes by prohibitins in mitochondria. Mol Biol Cell. 16: 248–259.

Tsai IJ, et al. 2013. The genomes of four tapeworm species reveal adaptations to parasitism. Nature 496:57–63.

van Zon A, et al. 2002. Structural domains of vault proteins: a role for the coiled coil domain in vault assembly. Biochem Biophys Res Commun. 291:535–541.

van Zon A, et al. 2003. The formation of vault-tubes: a dynamic interaction between vaults and vault PARP. J Cell Sci. 116:4391–4400.

Vasu SK, Rome LH. 1995. *Dictyostelium* vaults: disruption of the major proteins reveals growth and morphological defects and uncovers a new associated protein. J Biol Chem. 270:16588–16594.

Williams PD, Pollock DD, Blackburne BP, Goldstein RA. 2006. Assessing the accuracy of ancestral protein reconstruction methods. PLoS Comput Biol. 2:0598–0605.

Wu S, Zhang Y. 2007. LOMETS: a local meta-threading-server for protein structure prediction. Nucleic Acids Res. 35:3375–3382.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol. 24:1586–1591.

Zhang Y. 2008. I-TASSER server for protein 3D structure prediction. BMC Bioinformatics 9:40.

Associate editor: Dan Graur