



HAL
open science

New genomic data and analyses challenge the traditional vision of animal epithelium evolution

Hassiba Belahbib, Emmanuelle Renard, Sébastien Santini, Cyril Jourda,
Jean-Michel Claverie, Carole Borchellini, André Le Bivic

► **To cite this version:**

Hassiba Belahbib, Emmanuelle Renard, Sébastien Santini, Cyril Jourda, Jean-Michel Claverie, et al..
New genomic data and analyses challenge the traditional vision of animal epithelium evolution. *BMC Genomics*, 2018, 19 (1), pp.393. 10.1186/s12864-018-4715-9 . hal-01951941

HAL Id: hal-01951941

<https://amu.hal.science/hal-01951941>

Submitted on 11 Dec 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

New genomic data and analyses challenge the traditional vision of animal epithelium evolution

Hassiba Belahbib¹, Emmanuelle Renard², Sébastien Santini¹, Cyril Jourda¹, Jean-Michel Claverie^{1*}, Carole Borchiellini^{2*} and André Le Bivic^{3*} 

Abstract

Background: The emergence of epithelia was the foundation of metazoan expansion. Epithelial tissues are a hallmark of metazoans deeply rooted in the evolution of their complex developmental morphogenesis processes. However, studies on the epithelial features of non-bilaterians are still sparse and it remains unclear whether the last common metazoan ancestor possessed a fully functional epithelial toolkit or if it was acquired later during metazoan evolution.

Results: To investigate the early evolution of animal epithelia, we sequenced the genome and transcriptomes of two new sponge species to characterize epithelial markers such as the E-cadherin complex and the polarity complexes for all classes (Calcarea, Demospongiae, Hexactinellida, Homoscleromorpha) of sponges (phylum Porifera) and compare them with their homologues in Placozoa and in Ctenophora. We found that Placozoa and most sponges possess orthologues of all essential genes encoding proteins characteristic of bilaterian epithelial cells, as well as their conserved interaction domains. In stark contrast, we found that ctenophores lack several major polarity complex components such as the Crumbs complex and Scribble. Furthermore, the E-cadherin ctenophore orthologue exhibits a divergent cytoplasmic domain making it unlikely to interact with its canonical cytoplasmic partners.

Conclusions: These unexpected findings challenge the current evolutionary paradigm on the emergence of epithelia. Altogether, our results raise doubt on the homology of protein complexes and structures involved in cell polarity and adhesive-type junctions between Ctenophora and Bilateria epithelia.

Keywords: Epithelium evolution, Non-bilaterian animals, Cell polarity, Cell-cell junctions

Background

Multicellular organisms evolved from unicellular ancestors several times during the evolution of life [1, 2] resulting in an extensive morphological diversity. For metazoans, this major transition is linked with the emergence of a new type of cellular organization, the epithelium [3–6]. Historically, epithelia were defined in

bilaterians by the presence of three major features: apico-basal cell polarity, cell-cell junctions between the apical and the lateral domains and the presence of a basement membrane. These central features delineate key epithelial functions: the regulation of vectorial transport and morphogenesis [6]. By analogy, this typical bilaterian epithelium organization was extended to all eumetazoan i.e. including Cnidaria and Ctenophora [7–9], yet a lack of molecular evidence prevents evolutionary interpretations of epithelial structures [4, 9, 10].

From a morphological point of view, non bilaterian animals display a variety of cell sheet organizations. For example, the basal lamina is absent from all but one sponge classes [11, 12], in placozoa [13] and in several ctenophoran species [14]. From a functional point of

* Correspondence: jean-michel.claverie@univ-amu.fr; carole.borchiellini@imbe.fr; andre.le-bivic@univ-amu.fr

¹Structural and Genomic Information Laboratory, Aix-Marseille Université & CNRS UMR 7256, Mediterranean Institute of Microbiology (IMM FR 3479), Marseille, France

²Aix Marseille Univ, Univ Avignon, CNRS, IRD, UMR 7263, Mediterranean Institute of Marine and Continental Biodiversity and Ecology (IMBE), Station Marine d'Endoume, Marseille, France

³Aix-Marseille University, CNRS, UMR 7288, Developmental Biology Institute of Marseille Luminy (IBDM), Marseille, France

view, these epithelial-like cell layers show selective transport differences with bilaterian ones [4, 15–18]. It is now essential to determine the identity of the genes and proteins involved in these basal metazoan tissues – and consequently their homology across animals [5].

Despite the diversity of epithelial structures among animals, apico-basal cell polarity and AJs are believed to be present in all extant animal phyla [3, 5, 6, 8, 15, 19]. We thus chose to characterize molecularly these two bilaterian epithelial hallmarks among non bilaterian phyla. Former studies performed on Placozoa [5, 13, 20–22] and sponges [23–25] initiated the study of candidate epithelial genes in the different lineages. The conservation of critical functions was not assessed, however, due to the lack of detailed analyses of key protein interaction domains and residues. On the other hand, studies on the epithelial organization of Ctenophora were neglected in favor of studies focused on the mesoderm and nervous system [26–32] due to the previously unquestioned position of this phylum among eumetazoans [33–35].

In the present study, we first sequenced the genomes of two additional sponge species, *O. lobularis* (belonging to the Homoscleromorpha class) and *O. minuta* (the first Hexactinellida), and used RNA-seq data to help with the annotation procedure. This new data was then combined with information available from public databases to carefully identify and analyze homologues of genes coding for proteins known to compose polarity complexes and adherens junctions for all classes of Porifera (Calcarea, Demospongiae, Hexactinellida, Homoscleromorpha), several genera of Ctenophora with contrasted features [14] and Placozoa. Classical cadherins (E- type) and catenins [5, 36], Par, Crumbs (Crb) apical polarity complexes and Scribble (SCRIB) lateral polarity complex were identified and analyzed [8, 37–40]. We hypothesize that sponge species exhibit highly contrasted tissue features related to molecular divergence of some of their polarity complex proteins. Finally, we revealed an unexpected lack of conservation of the epithelial toolkit in Ctenophores asking for a profound revision of our understanding of Ctenophore biology. Altogether, our results raise a doubt on the homology of protein complexes and structures involved in cell polarity and adhesive type junctions between Ctenophora and Bilateria epithelia.

Results

New genomic and transcriptomic data from *Oopsacas minuta* (Hexactinellida) and *Oscarella lobularis* (Homoscleromorpha)

We used two platforms (Pacific Bioscience and Illumina) and a combination of paired end and mate pair sequencing approaches (see Materials and Methods) to generate and assemble the data. Concerning *Oopsacas minuta*,

the assembly yielded a total of 61.46 Mb of unique haploid genome sequence distributed in 365 contigs longer than 1 kb (N_{50} length = 676,369 bp, L_{50} number = 31, mean coverage = 381). Following the mapping of 207,529,788 RNAseq reads from a polyA+ cDNA library (mean Open Reading Frame (ORF) coverage = 1443), we predicted the presence of 17,043 protein-coding genes. The small final number of contigs and the above coverage values suggest that our delineation of the (protein-coding) gene content is very close to 100% completion.

Concerning *Oscarella lobularis*, we generated and assembled a total of 52.34 Mb of unique haploid genome sequence distributed in 2658 contigs longer than 1 kb (N_{50} length = 265,395 bp, L_{50} number = 58, mean coverage = 98). Following the mapping of 231,475,388 RNAseq reads from a polyA+ cDNA library (mean ORF coverage = 710), we predicted the presence of 17,885 protein-coding genes. The large, albeit unavoidable, proportion (> 50%) of sequence data from bacterial and archaean origin, as well as unfavorable (repeated or variable) genome structures caused the final number of contigs to remain significantly larger than for *O. minuta*. However, the above coverage values remain large enough to correspond to a near 100% complete delineation of the (protein-coding) gene content. The quality of our transcriptomes and genome drafts enables us to be confident on the completion of the predicted proteins and the number of copies found for each candidate gene.

Porifera common ancestor most likely possessed functional adherens junctions

Classical cadherins contain extracellular repeat domains that mediate trans-interactions with the extracellular domain of cadherins on opposing cells, and a cytoplasmic domain that binds p120 and β -catenin [5, 41–43]. β -catenin binds to α -catenin thereby forming the core cytoplasmic protein complex of the classical cadherin/ β -catenin/ α -catenin complex (CCC). In this complex, α -catenin is the key protein that links the CCC complex to the underlying actin cytoskeleton. In turn, p120 is the critical actor for the surface stability of cadherin-catenin cell-cell adhesion by controlling cytoskeletal dynamics and regulating cadherin endocytosis. Cadherin and catenin families are present outside of metazoans, but the C-terminal catenin-binding motifs that define classical E-cadherins are a metazoan novelty [5, 36, 44]. Consistent with earlier reports, our analyses, combining homology searches, phylogenetic reconstructions and domain predictions, confirm that Placozoa and all Porifera possess homologues of classical E-cadherins [5, 13, 20–25, 36, 42]. All characteristic domains were identified with high confidence (Fig. 1a):

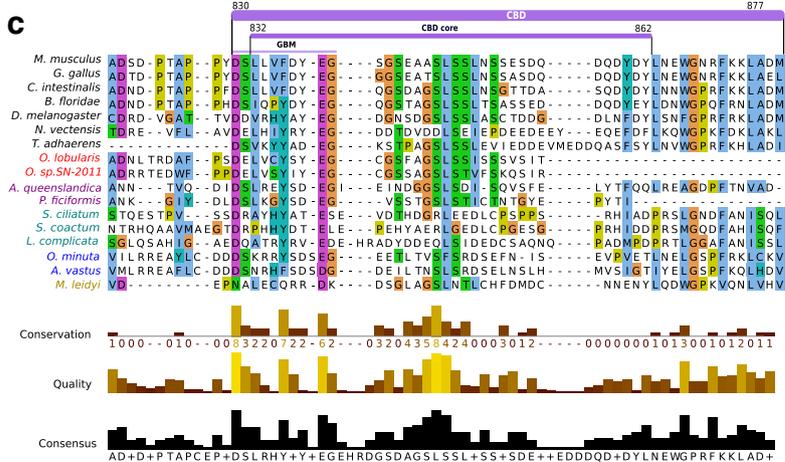
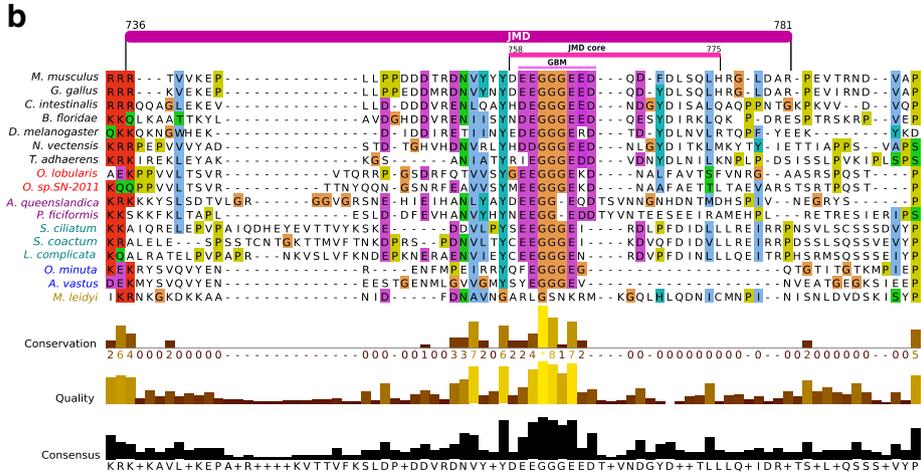
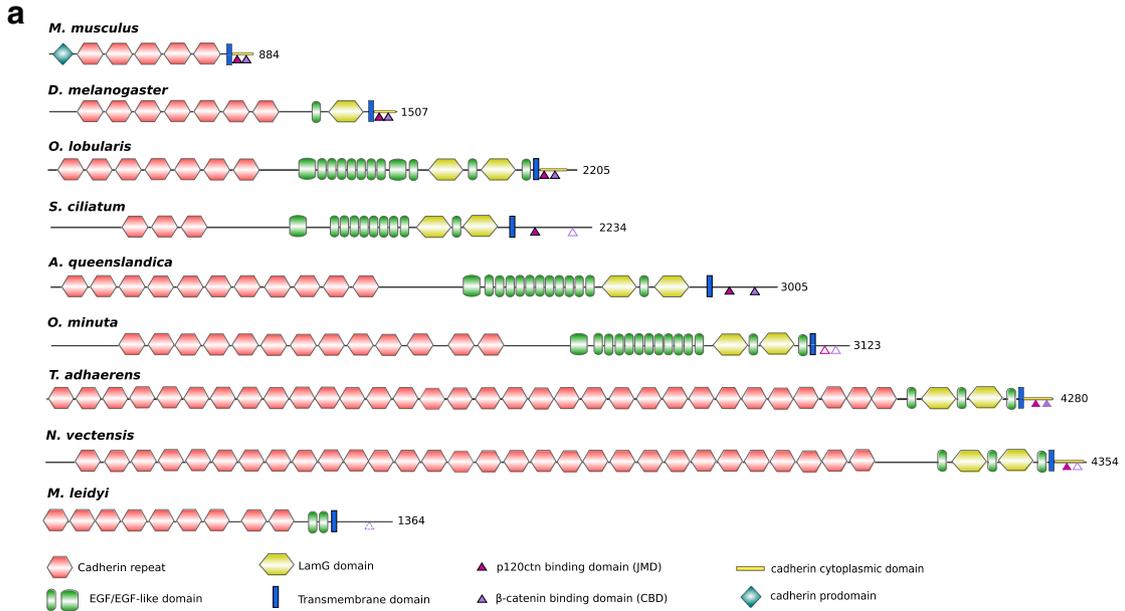


Fig. 1 (See legend on next page.)

(See figure on previous page.)

Fig. 1 Comparison of E-cadherin domains and motifs between metazoans. Porifera: Homoscleromorpha in red (*Oscarella lobularis*, *Oscarella. sp.*), Demospongiae in magenta (*Amphimedon queenslandica*, *Petrosia ficiformis*), Calcarea in green (*Sycon ciliatum*, *Sycon coactum*, *Leucosolenia complicata*), Hexactinellida in blue (*Oopsacas minuta*, *Aphrocallistes vastus*). Other represented clades are Placozoa (*Trichoplax adhaerens*); Ctenophora (*Mnemiopsis leidyi*) in yellow, Cnidaria (*Nematostella vectensis*), Bilateria (Deuterostomia: *Mus musculus*; Protostomia: *Drosophila melanogaster*). Sequences were aligned with MAFFT v7 web server and visualized with Jalview. **a** Representative cadherin proteins depicted with their domains. *Mus musculus* and *Drosophila melanogaster* E-cadherins are taken as reference. *Oscarella lobularis* has the sole poriferan cadherin the cytoplasmic-specific domain of which is detected by Pfam (E-value = 2.10^{-11}) and InterProScan as in the mouse and fruitfly E-cadherin (depicted in yellow at the C-terminal part). Degrees of conservation of p120 and β -catenin binding domains are indicated by full, dashed or open triangles. **b** Alignment of the cytoplasmic cadherin p120 binding domain (Juxtamembrane domain, JMD). The JMD consists of 50 residues immediately following the transmembrane domain (in Mouse E-cadherin). The JMD core consists of 20 residues. The groove-binding motif (GBM) required for binding p120 is well conserved in metazoans. **c** Alignment of the cytoplasmic cadherin β -catenin binding domain (CBD). The CBD consists of approximately 50 residues. The groove-binding motif (GMB) consists of 10 residues

- The extracellular cadherin (EC) repeated domains (ranging from 3 in *Sycon ciliatum* to 32 in *Trichoplax adhaerens* units) that mediate trans-interactions with the extracellular domain of cadherins on opposing cells;
- The transmembrane region (TM) and the cytoplasmic tail, which contains the conserved specific binding domain for p120-catenin in the juxta-membrane domain (JMD) and the β -catenin-binding domain (CBD);
- The epidermal growth factor (EGF) domains and Laminin G (Lam-G) domains in a membrane-proximal position considered typical of non-vertebrate classical cadherins [19, 42].

The alignment of E-cad JMD that mediates binding to p120 catenin (Fig. 1b) shows that the Groove-Binding Motif (GBM motif) (XX [ED] GGGEXX) is highly conserved in placozoan and in three classes of sponges. In contrast a G residue is missing in the two demosponges studied, which may modify the interactions with p120-catenin, since the three consecutive glycine residues are thought to anchor the region in a small hydrophobic pocket in the armadillo (ARM) repeats of p120-catenin [42, 45]. p120-catenin consists of central ARM domain repeats involved in E-cad JMD interactions flanked by an N-terminal regulatory region (NTR) and a C-terminal tail region (CTR). Among the key residues of p120 involved in E-cadherin-binding (Additional file 1: Figure S1A), the 13 essential residues involved in electrostatic interactions (Q391 to K574) with the N-terminal acidic region of the JMD core (residues 758–766, [45] (Fig. 1b) are highly conserved in sponges and placozoans with minor exceptions (H392 -> Q and K574-> Q residues) in glass (Hexactinellid) sponges. In contrast, the eight amino acids in the N-terminus of p120 (R364 to Y389, Additional file 1: Figure S1A) known to be involved in hydrophobic interactions with the C-terminal anchor region of the JMD core (residues 767–775) are more variable. This region (Fig. 1b) of E-cadherin appears also less constrained suggesting

that electrostatic interactions dominate the p120-catenin-JMD core interaction.

We detected a striking exception in the ctenophore *Mnemiopsis leidyi*, in which the E-cadherin GBM motif is not conserved (Fig. 1b) possibly forbidding interaction with p120-catenin. In contrast, *M. leidyi* p120-catenin residues, essential for electrostatic binding with E-cadherin classical cytoplasmic domain, remain highly conserved (Additional file 1: Figure S1A), thus excluding a compensatory co-evolution process [44] that may have preserved the interaction. In the cadherin domain binding to β -catenin (CBD, Fig. 1c), the interaction was shown to require a GBM of about 10 amino acids (DXXXX ϕ XXEG where ϕ is an aromatic residue) [42, 45, 46]. As for the p120-catenin-binding motif, this motif is conserved in *Trichoplax* and all sponges except in calcareous sponges where it slightly diverges at the end (Fig. 1c). Whether such a change in the CBD results in a weakening (or loss) of the interaction with β -catenin in calcareous sponges has yet to be investigated.

A single β -catenin gene copy was identified in each studied species except for calcareous sponges that exhibit a specific duplication (Additional file 1: Figure S1B). All β -catenin proteins identified in sponges, placozoans and ctenophores harbor the same ARM repeats as described in bilaterians. In sponges and placozoans, β -catenin residues that were identified as essential for E-cadherin interaction [47] are highly conserved except for the R386 and N387 residues (respectively replaced by L and T) in two hexactinellids and a more anecdotic change: A656 -> S in placozoans.

In *M. leidyi* again the E-cadherin CBD motif diverged from that of other metazoans. The D, an aromatic residue, and the G were replaced in the DXXXX ϕ XXEG sequence (Fig. 1c), which might impair interactions with the classical β -catenin respective interaction motif. Conversely *M. leidyi* β -catenin amino acids involved in the interaction with E-cadherin (Additional file 1: Figure S1B) are modified (Y331V, K335Y, D390N and R582C) either suggesting a loss of interaction with the E-cadherin CBD intracellular domain or its maintenance

through co-evolution. The α -catenin (member of the vinculin family) links E-cadherin to the actin cytoskeleton by interacting with β -catenin [36]. All species studied, including *T. adhaerens* and *M. leidy*, have at least two vinculin family members: one orthologous to α -catenin and one orthologous to vinculin. Interestingly, the α -catenin/vinculin N-terminal region, known to interact with β -catenin, is conserved (Additional file 1: Figure S1C). In addition, in the β -catenin of sponges, placozoans and ctenophores, most of the crucial α -catenin-binding residues [47, 48] are conserved (Additional file 1: Figure S1B) suggesting that such an interaction was already present in the common ancestor of metazoans. Our analyses of two additional sponge species (*Oscarella lobularis*, class Homoscleromorpha; *Oop-sacas minuta*, class Hexactinellida) confirm the presence of bona fide E-cadherin complexes in the four Porifera classes. Moreover, the motifs governing the interactions between the members of this CCC complex essential for the establishment of adherens junctions appear very conserved in Placozoa, Calcispongiae and Homoscleromorpha. Even if a few substitutions were identified in demosponges and glass sponges that may modulate these interactions and explain the absence of AJs in these two lineages, we can nevertheless infer that the last ancestor of Porifera already possessed all the component needed to build functional adherens junctions similar to those of bilaterians. However, this is probably not true of the ctenophores as their E-cadherin cytoplasmic domain lacks most bona fide E-cadherin cytoplasmic domain binding sites.

A par apical polarity complex inherited from Urmetazoa

Next we investigated whether placozoans, ctenophores, and sponges of all classes, harbor the polarity protein complexes that are necessary for epithelium formation and morphogenesis [37]. There are at least three types of polarity complexes:

- The Par complex made of atypical Protein Kinase C (aPKC), Partition defective 3 (Par3) and Partition defective 6 (Par6);
- The Scribble complex made of Scribble (Scrib/Src), lethal giant larvae (Lgl) and Disc large (Dlg);
- The Crumbs complex made of Crumbs (Crb), stardust (Sdt, or MPP5 (Membrane Palmitoylated Protein 5) in mammals also known as Pals1 (protein associated with Lin-7 1) and Pals1-associated tight junction protein (Patj).

First, we looked for the Par complex, considered to be a metazoan innovation [23]. Par6 contains an N-terminal Phox and Bem1 (PB1) domain, a C-terminal Postsynaptic-density-95/Disc-large/Zona-occludens1

(PDZ) domain and a semi-CRIB (Cdc42/Rac interactive binding domain) motif immediately preceding the PDZ domain. The PB1 domain of Par6 forms a heterodimer with the PB1 domain of aPKC. Par3 is associated with Par6/aPKC complex via the PDZ-PDZ domain interaction. The activity of the PAR complex is dynamically regulated by phosphorylation of PAR3 and its association with the stable PAR6-aPKC complex. Highly conserved sequences for all members of this complex are present in all available genomes of sponges, placozoans and ctenophores. All characteristic domains as well as the residues essential for their interactions within the complex were also identified (Figs. 2 and 3, Additional file 1: Figure S2). For example, Par6 (Fig. 2) interacts with aPKC through its PB1 domain and, in mouse Par6, lysine K19 is essential for this interaction [49]. This Lysine residue is strictly conserved in all species studied here, which strongly suggests that the interaction between aPKC and Par6 is conserved throughout metazoan evolution. In addition, there are increasing evidences that the formation of this complex is regulated by phosphorylation, mainly on serine S980 in the aPKC-binding region of Par3 [50]. This phosphorylated site (S/T) is conserved in all studied species (Additional file 1: Figure S2). All these data strongly suggest that aPKC, Par3 and Par6 have co-evolved from a functional metazoan ancestral complex.

The scribble lateral polarity complex is present in all non bilaterians except ctenophores

Next, we investigated the presence of the Scribble polarity complex composed of Scrib, Dlg and Lgl members. Members of this complex contain multiple protein-protein interaction domains, in particular PDZ, Src homology 3 (SH3) domain and guanylate kinase (GUK) domains capable of recruiting a complex network of proteins.

We identified with high confidence Dlg orthologous genes containing all specific domains (Lin2 and Lin7 binding domain (L27), GUK, SH3 and three PDZ domains) (Fig. 4) in all sponges and in *T. adhaerens* (even though the L27 domain is lacking). In ctenophores, Dlg orthologous genes were found without GUK domain. Since Dlg predates the emergence of metazoans (Dlg homologues have been reported in Choanoflagellata, Filasterea and Ichthyosporea) [23, 36, 51], the absence of this key domain is probably due to secondary loss. Lgl, characterized by short WD40 repeats and specific phosphorylation sites, is not present in Choanozoa [24] but was identified in all Porifera and Placozoa in agreement with previous studies [23, 24] and in Ctenophora (Additional file 1: Table S1), suggesting that it appeared in the last common metazoan ancestor.

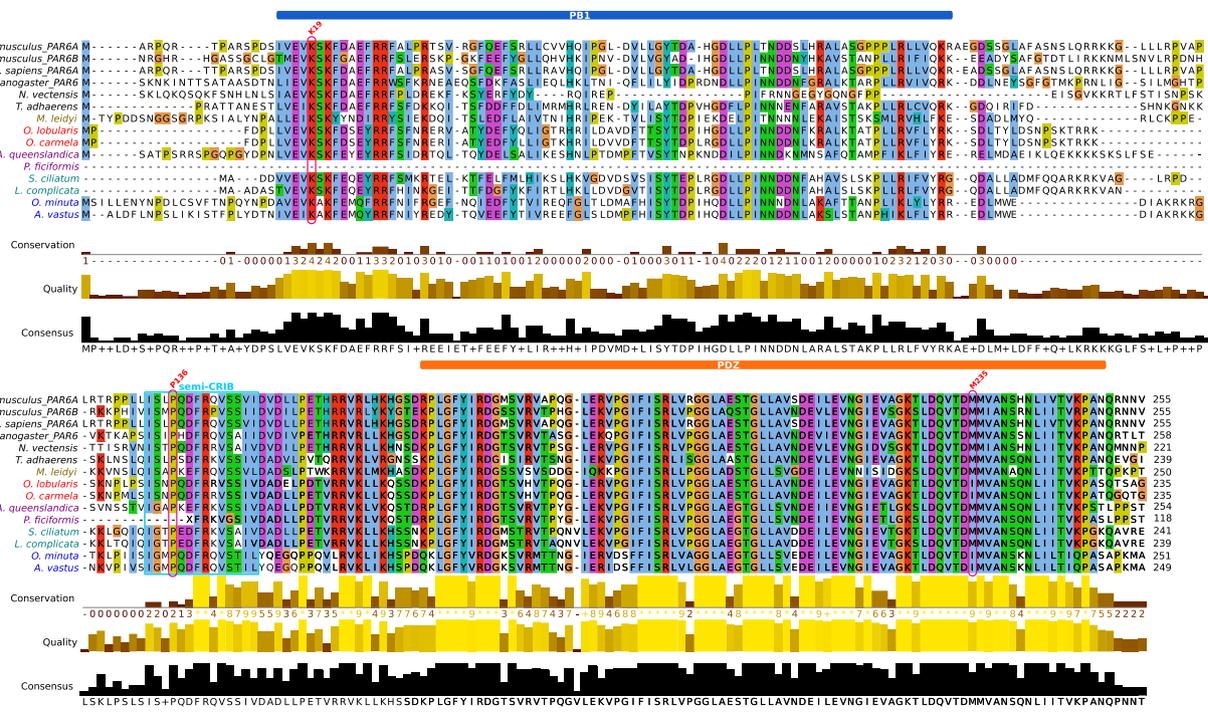
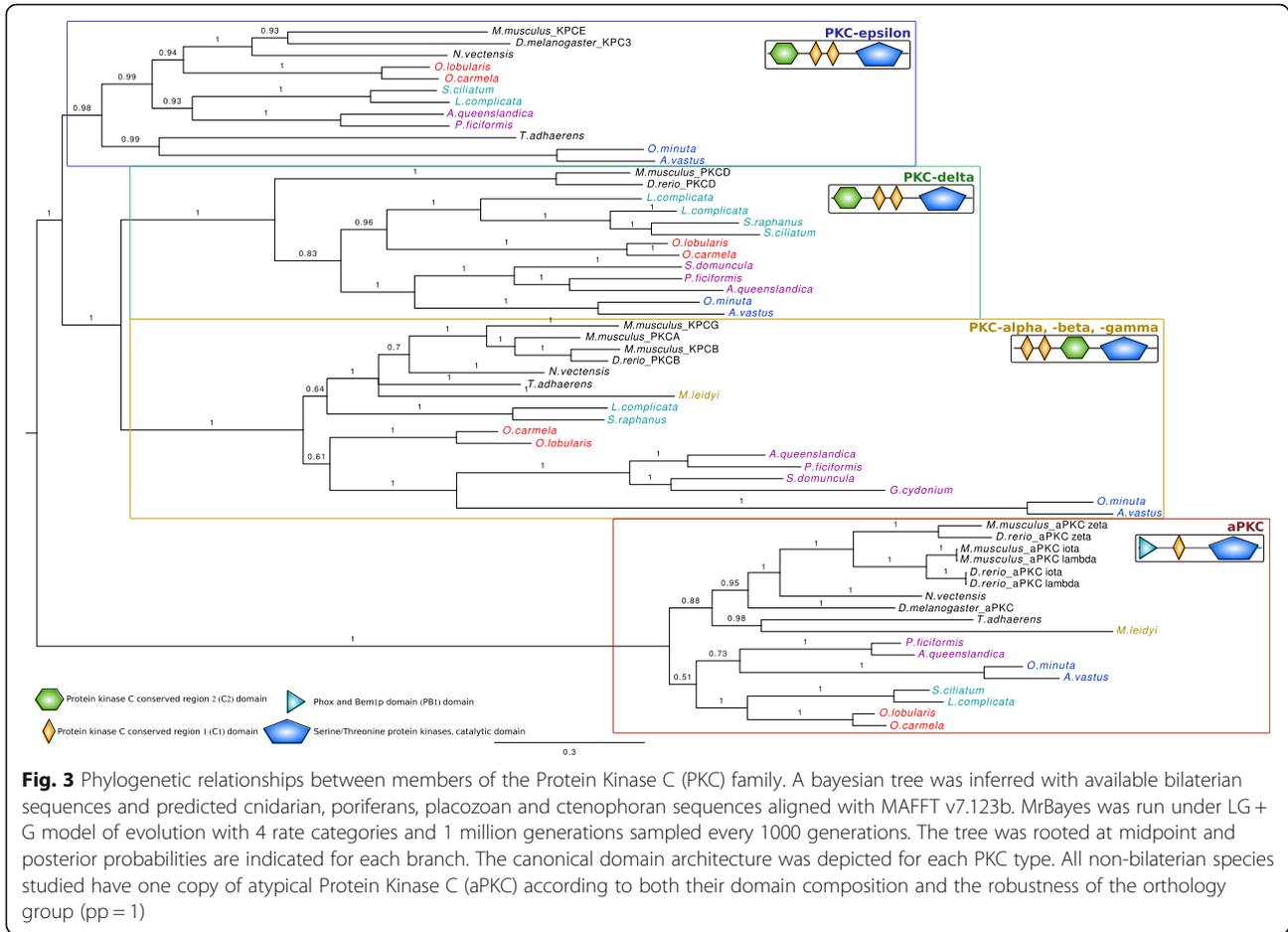


Fig. 2 Comparison of the sequences of the diagnostic domains of Par6: The N-terminal Phox and Bem1 (PB1) domains required for interaction with the PB1 domain of atypical Protein Kinase C (aPKC); the semi-CRIB (Cdc42/Rac interactive binding) domain and the C-terminal Postsynaptic-density-95/Disc-large/Zona-occludens1 (PDZ) domain required for the interaction with Par3. Sequences were aligned with MAFFT v.7 and visualized with Jalview 2.9. Critical residues are labelled in red: Lysine K19 in PB1 domain is essential for the interaction with aPKC; Proline 136 (P136) in the semi-CRIB motif is necessary to bind cdc42; Methionine 235 (M235) in the PDZ domain binds the LGL protein

As previously reported, Scrib homologues were identified in all sponge classes [24] and placozoans [36] (Additional file 1: Table S2). Scribble is a LAP [LRR (leucine-rich repeats) and PDZ (PSD-95/Disc-large/ZO-1) protein containing 16 LRRs and either one or four PDZ domains [40, 52]. In striking contrast, we could not detect a protein associating a PDZ domain and LRRs in ctenophores. To discard the hypothesis that divergent evolution led to a specific loss of Scribble in *M. leidyi*, we investigated its presence in two other genera (*Pleurobrachia* and *Beroe*) and we confirmed the absence of Scribble homologue (the only LRRs domains we identified belong to other classes). The LRR domain is critical for Scribble function, since in *Drosophila* Scribble proteins mutated in the LRR domains mimic the complete loss of Scribble protein [40]. The PDZ domain of Scribble was also shown to be important for its recruitment to the junctional complex and plasma membrane [53] and for the correct localization of Dlg [40] in *Drosophila*. The absence of a bona fide Scribble homologue in ctenophores might indicate a change in Dlg/Lgl localization or function.

The crumbs apical polarity complex is divergent in syncytial glass sponges and absent in ctenophores
 We then investigated the conservation of the Crumbs complex in metazoans. Crumbs is a central regulator of epithelial apical actin cytoskeleton organization and adherens junction formation in bilaterians and was proposed to be a metazoan innovation [8, 23]. The formation of the Crumbs complex is ensured by physical interactions between different core components. The central component, Stardust/MPP5, organizes a plasma membrane-associated protein scaffold via an interaction between its PDZ domain and the C-terminal ERL1 motif of Crb. The two L27 domains of Sdt bind to the L27 domains of PATJ and Lin-7. Crumbs transmembrane proteins, consist of extracellular EGF, laminin-like repeats and a short cytoplasmic domain (less than 40 amino acids) with two essential localization motifs [37] (Fig. 5a). These motifs are the signature of Crumbs proteins and are essential for their morphogenetic function [54, 55]. The membrane proximal motif RxxxGxYxPS or FERM-binding motif (FBM) is required for the interaction with proteins of the ERM (Ezrin-Radixin-Moesin) family that associates with the actin cytoskeleton [56] (Fig. 5b). The second motif consists of the last 4 amino acids, ERL1, at the C-terminus



(Fig. 5b). It is a class II PDZ-binding motif (PBM), which interacts with stardust (MPP5) [57] and Par6 [58]. There is a strong conservation of the class II PDZ-binding site with conservative variations (E/D-R/K-L/I-I/L) in three of the four sponge classes and in *Trichoplax*, most likely under evolutionary pressures maintaining the interaction with PDZ containing proteins (Sdt or Par6). A unique exception was found in hexactinellids where the ERLI motif in replaced by ETLI (Fig. 5b). This change allows the binding of class I PDZ domains instead of class II. Analysis of FBM in *Trichoplax* and Porifera reveals that hexactinellids exhibit the most divergent sequences with only two conserved residues (XxxxXxYxPX) while *O. lobularis* has a conserved FBM (RxxxGxYxPT) (Fig. 5b) suggesting that homoscleromorphs have a truly functional Crumbs complex while it might be defective in hexactinellids. Therefore, there might be a relationship between the loss of some protein interactions and the syncytial organization characteristic of this sponge lineage. In Calcarea and in Demospongiae, 4 and 3 of the 5 FBM residues are conserved. Placozoans also exhibit a conserved FBM with RxxxGxFxPS in one of their two Crumbs homologues (Fig. 5b). Another feature of the FBM in bilaterians

is the presence of two phosphorylation sites recognized by aPKC (TxGTYx), which regulates the binding to Moesin, an ERM protein [59]. These two phosphorylation sites are absent from all sponge species and from *Trichoplax*, suggesting that this regulation by aPKC is an innovation shared by cnidarians and bilaterians (except for *Caenorhabditis elegans*) since at least one phosphorylation site is present in a Crumbs isoform of cnidarians (*Nematostella*) (Fig. 5b). Thus, even though previous studies identified Crumbs-like proteins in all sponge classes, these studies did not verify the conservation of key functional motifs [24] required for their functional interactions. Here, we show that the high divergence of key Crumbs residues in glass sponges is hardly compatible with the formation of a fully conserved complex (Fig. 5b).

Finally, the most unexpected result was the absence of any Crumbs-like gene or transcript in *M. leidy* (and in other ctenophore transcriptomes available in databases: *Pleurobrachia bachei*, *Beroe abyssicola* and *Beroe sp.*) exhibiting a significant similarity with the conserved cytoplasmic domain, while we identified homologues of transmembrane proteins with extracellular domains made of EGF and laminin-like repeats. This suggests that

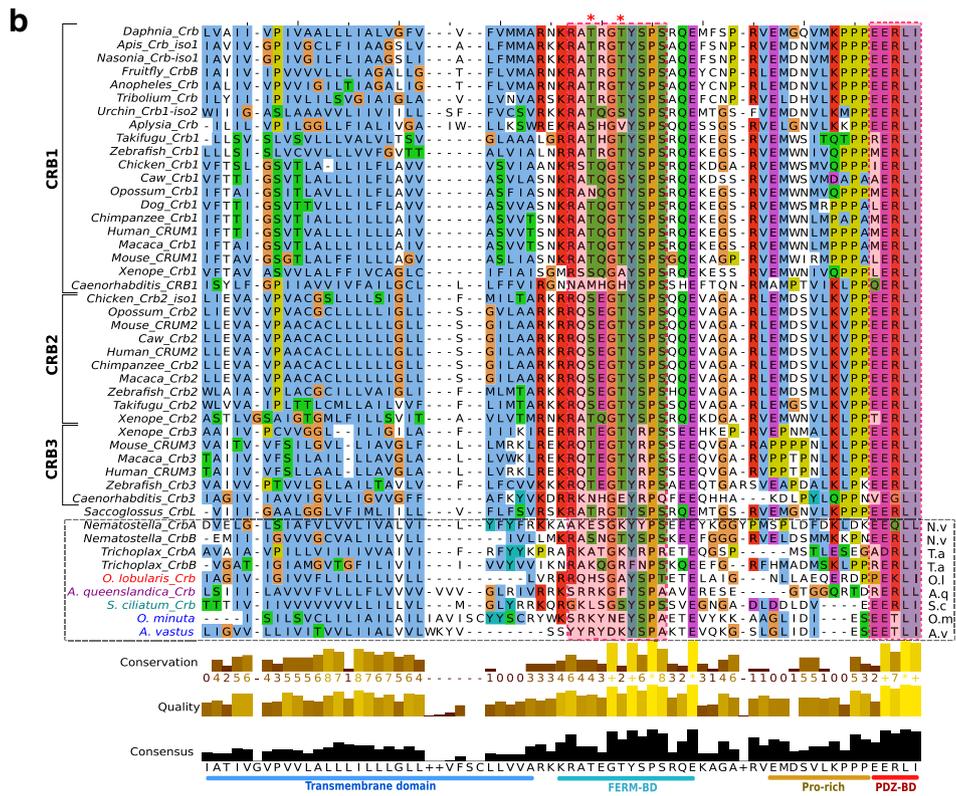
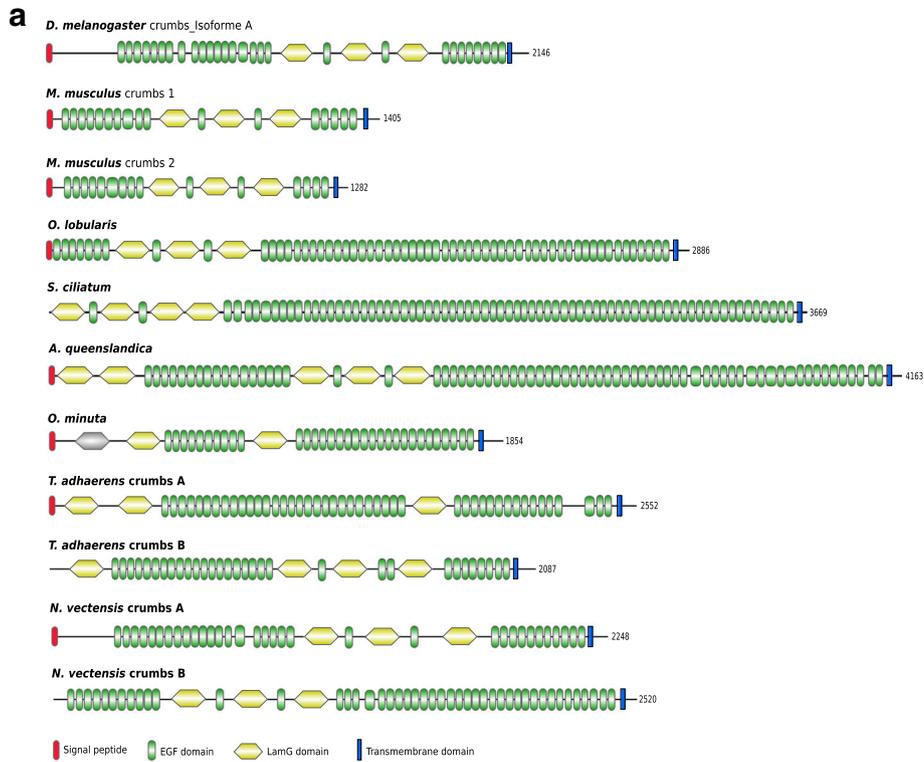


Fig. 5 (See legend on next page.)

(See figure on previous page.)

Fig. 5 a Domain composition of Crumbs proteins of *D. melanogaster* (Dcrb), *M. musculus* (CRB1, CRB2), *O. lobularis*, *S. ciliatum*, *A. queenslandica*, *O. minuta*, *N. vectensis* and *T. adhaerens*. Domains are shown as detected by SMART and Pfam and scaled with IBS software. Crumbs transmembrane proteins consist of extracellular epidermal growth factor (EGF), laminin-like (LAM) repeats and a short cytoplasmic domain. No crumbs was detected in Ctenophora. In contrast to other animals, sponges have only one copy of Crumbs. **b** Alignment of the cytoplasmic domain of Crumbs that binds PALS1 and PAR6. Transmembrane and intracellular domains were aligned with MAFFT v7.123b and displayed with JalView. The transmembrane domain, the FERM binding domains (FBM) containing the RxxxGxYxPS motif needed for the interaction with the Ezrin-Radixin-Moesin (ERM) protein family, the Proline rich domain and the PDZ binding domain (PDZ-BD) are depicted at the bottom. The FBM presents a different conservation depending on the sponge class considered (from 2 to 5 conserved residues). Asterisks indicate the position of the two phosphorylation sites recognized by aPKC. These sites are absent in placozoans and poriferans. The PDZ-BD domain (interacting with MPP5) is well-conserved in non bilaterians except in glass sponges

homoscleromorph sponge *Oscarella lobularis* and to a lesser extent in calcareans and demosponges. As no PatJ homologue was found in ctenophores, we can safely conclude that the whole Crumbs/Sdt/Patj complex is entirely absent in this phylum.

Discussion

By investigating the presence of genes and proteins involved in epithelial polarity and adherens junctions, we found that Placozoa and Porifera (despite some divergence observed in hexactinellids) possess all polarity

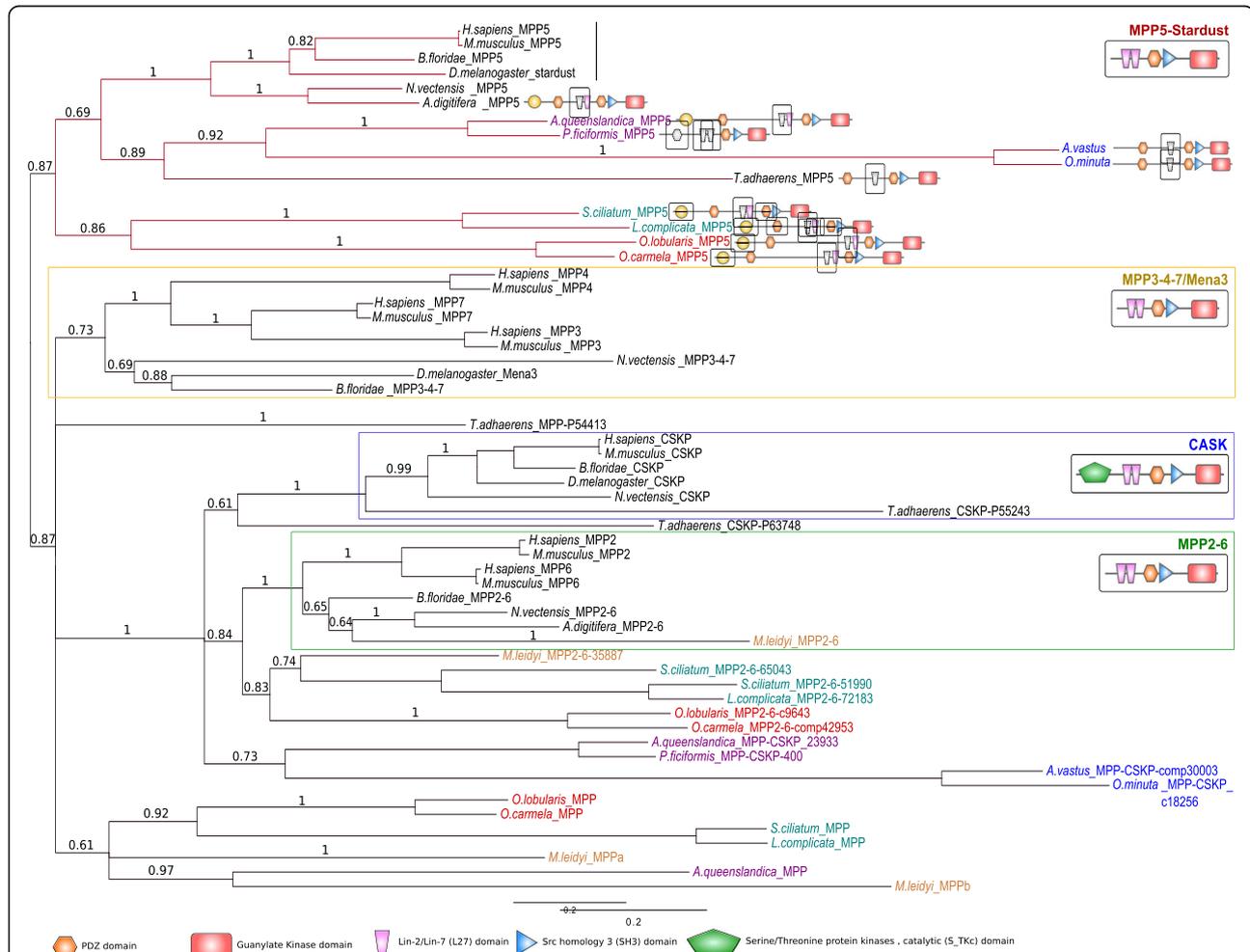


Fig. 6 Phylogenetic tree of MAGUK proteins based on their shared MPP PDZ + SH3 + GUK domains. A bayesian tree was inferred with available bilaterian sequences and predicted cnidarian, poriferan, placozoan and ctenophoran sequences aligned with MAFFT v7.123b. MrBayes was run under LG + G model of evolution with 4 rate categories and 1 million generations sampled every 1000 generations. The tree was rooted at midpoint and posterior probabilities are indicated for each branch. The canonical domain architecture was depicted for each main MPP class: MPP5-stardust, CASK, MPP2-6, MPP3-4-7. Whereas Ctenophora lack a MPP5/Sdt orthologue, all 4 sponge classes and Placozoa have one copy of the corresponding gene. Nevertheless, the first Lin2/Lin7 (L27) domain involved in interaction with the L27 domain of PatJ (Pals1-associated tight junction protein) is missing in glass sponges. Domains in grey are predicted but divergent

complex members and adherens junction components. In contrast, ctenophores lack the Crumbs complex and Scribble as homologues of the corresponding genes were not found in currently available species. In addition, *M. leidy* possesses an E-cadherin-like cytoplasmic sequence divergent enough from canonical E-cadherin to raise doubt on its ability to interact with p120 catenin and with β -catenin. These unexpected findings are shedding a new light on the ongoing controversy about the morpho-anatomy of the last metazoan ancestor based on various phylogenetic reconstructions [27, 34, 35, 60–62]. One hypothesis favors ctenophores as a sister group of all other metazoans [35, 62], proposing that complex traits such as neurons and muscles might have been acquired independently in Ctenophora and Parahoxozoa [60]. The alternative hypothesis favors Porifera as the sister group of other metazoans [34, 61] in agreement with more traditional interpretations. According to our results, sponges (in particular Homoscleromorphs) now appear to have an epithelial toolkit (collagen IV, polarity complexes, E-cadherin complex) that is more complete (and expected to be functional according to motif conservation) than that of ctenophores. This finding is all the more unexpected since the epithelial organization of ctenophores appears fully accepted [51], while the presence of “true” epithelia in sponges remains debated.

In the context of the ctenophore-first evolutionary scenario, the compromised interactions between catenins and cadherins and the absence of two typical polarity complexes in ctenophora epithelia can be interpreted in two ways. Either it is the result of secondary losses, or it was inherited from an ancestral state. In this later case, it implies that additional components such as the Crumbs complex and p120 binding for E-cadherin were later acquired during the course of evolution in the last common ancestor of sponges and parahoxozoa.

Interestingly, the epithelial features of Ctenophora are very different [14, 26, 27, 63, 64] leading to incongruent interpretations of authors concerning the presence or not of bona fide AJs. According to our survey, if there are adhesive-type junctions in ctenophores, they cannot be considered as bilaterian AJ homologues. Our results also challenge the notion that there is a straightforward relationship between a genetic toolkit and morphological features. Sponges from different classes possess very different tissue organizations; homoscleromorphs have epithelial-like layers with adherens junction while hexactinellids exhibit a syncytial organization without AJ-like junctions. Up to now, however, the small molecular variations observed in these different species are not sufficient to explain the huge differences seen in body plans. Similarly, for ctenophores, *Beroe*, *Pleurobrachia* and *Mnemiopsis* exhibit different epithelial features [14] despite their very similar gene contents. Consequently, gene

inventories alone are not sufficient to explain tissue and structure diversity.

Conclusions

Altogether, our results raise a doubt on the homology of protein complexes and structures involved in cell polarity and adhesive type junctions between Ctenophora and Bilateria epithelia.

Our study strongly advocates for more functional studies of the epithelium-like tissue of all non bilaterian animals. It is also an incentive to develop sponges and ctenophores as new experimental models for cellular biology to elucidate how epithelial cell layers that are key to the rise of animal diversity emerged throughout evolution.

Methods

Genome sequencing and assembly

Oscarella lobularis (Schmidt 1862) and *Oopsacas minuta* Topsent, 1927 were collected by SCUBA diving in the north-western Mediterranean Sea (Marseille Bay). All new sequences are available on NCBI website: accession numbers and links are provided in Additional file 1: Tables S5 and S6.

Oscarella lobularis genome sequencing was performed using Illumina technology with DNA-seq paired-end and Nextera mate pair protocols on a HiSeq2500 sequencer. Adapter sequences were removed and low-quality bases were trimmed using Cutadapt [65]. Remaining reads were assembled using a pipeline including IDBA-UD ([66], Platanus [67], GapFiller [68] and cap3 [69]. A transcriptomic dataset was mapped with Tophat [70] to all contigs longer than 1 kb to identify potential Eukaryotic sequences. The result was passed to Braker [71] to predict genes. All predicted protein sequences were submitted to BLAST ([72] to refine taxonomic assignment and automatically assess genes function. Seventeen thousand eight hundred eighty-five protein-coding were predicted from a total of 2658 contigs (Additional file 1: Table S3). Following the mapping of the individual reads to the final genome and transcriptome assembly using Bowtie2 [73], the coverage values were found to be 98, ensuring a near 100% completeness of the predicted gene content.

Oopsacas minuta genome sequencing was first performed in the same conditions as *Oscarella lobularis*. In addition, a second genome sequencing step was performed using PacBio technology on an isolated sponge fragment to limit bacterial contaminations. These long reads were filtered based on their length and quality with Pacific Biosciences (PacBio) tools (SMRT Portal) then self-corrected with canu [74]. All Illumina reads were mapped on the corrected PacBio reads with Bowtie2. Mapped Illumina reads and corrected PacBio reads were then assembled together with SPAdes [75]. The

number of contigs longer than 1 kb was low enough to rapidly identify Eukaryotic sequences using MetaGeneMark [76] and BLAST through a homebrew web server. Finally, a super scaffolding and polishing step was achieved using Sspace [77], Pilon [78] and GapFiller. Seventeen thousand forty-three genes were predicted with Braker from the 365 remaining contigs (see Additional file 1: Table S3 for metrics). The same method as for *Oscarella lobularis* genome was applied to predict proteins and their functions. The mapping of the individual read to the Genome and transcriptome assembly resulted into an estimated coverage of 381, again ensuring a near 100% completeness of the gene content.

Sequence annotation and structure prediction

Epithelial hallmarks were investigated using data from various sources: *O. lobularis* and *O. minuta* de novo assembled genomes and transcriptomes (this work), and on available poriferan, placozoan, cnidarian and ctenophoran genomes and/or transcriptomes retrieved from the sources listed in Additional file 1: Table S4. *D. melanogaster*, *M. musculus* and *A. queenslandica* Epithelial cadherin, Crumbs, PAR and Scribble complexes retrieved from NCBI database were used to perform reciprocal best-hits with BLAST 2.3.0 run locally using an E-value cutoff of 10^{-5} .

Ab initio protein-coding gene prediction was performed on the best candidate genomic and/or transcriptomic contigs with GeneMark.hmm eukaryotic web server (<http://exon.gatech.edu/gmhmm.cgi>) [79], GenScan (<http://genes.mit.edu/GENSCAN.html>) [80], Augustus v3.0.3 (<http://bioinf.uni-greifswald.de/augustus/>) [81] run locally and FgeneSH web server (<http://www.softberry.com/>) [82]. Protein domains were predicted and checked with Pfam v28.0 (<http://pfam.xfam.org/>) run locally [83], InterProScan v52 (<http://www.ebi.ac.uk/interpro/>) [84] and SMART (<http://smart.embl-heidelberg.de/>) [85].

The newly identified early branching metazoan proteins were added to the previous database sequences for iterative BLAST searches to identify more potential homologues. Proteins containing repeated domains generate false positive and best-hits characteristic of protein motifs and /or domains were used to enhance the detection of real homologues. These motifs and domains were aligned with MAFFT.7 [86] (<http://mafft.cbrc.jp/alignment/server/>) and/or MUSCLE v3.8.31 [87] (implemented in Seaview 4.5.2 [88]) depending on the level of conservation of the proteins. HMM profiles were built with HMMER 3.1b1 (<http://hmmerr.org/>) using aligned sequences with an E-value cutoff of 10^{-5} . Retrieved motifs of potential homologues to the protein queries were used as new baits to recompute HMM profiles and perform further HMM searches. C-terminal

characteristic parts were used to build HMM profiles of Crumbs and E-cadherin proteins.

Since Lin2/Lin7 (L27) (N)-terminal domain is important in the mediation of MAGUK family protein interactions with other proteins, a particular attention was given to this domain. To identify all potential MAGUK proteins, local BLAST searches (BLASTp and tBLASTn) were performed against complete or on-going genome projects (Additional file 1: Table S4) using L27 domains retrieved from the Pfam database (PF02828). To improve the detection, L27 domain HMM profile was rebuilt by iteratively adding the best scoring L27 domains identified in sponges, placozoans, cnidarians and ctenophorans.

L27 domain HMM searches led also to the identification of PATJ homologues but for some proteins containing multiple PDZ domains (MPDZ) no L27 domain was detected. For MPDZ lacking the L27 domain, a BLASTp was run locally using the PDZ domains retrieved from the Pfam database (PF00595) as reference. This step led to the identification of *Petrosia ficiformis* MPDZ which has (N)- and (C)- terminal regions predicted on two different contigs, including L27 and PDZ domains on the first contig and the remaining PDZ domains on a second contig merged with Emboss merger webtool (<http://www.bioinformatics.nl/cgi-bin/emboss/merger>).

In addition to protein domain analyses, conservation of critical residues was investigated on aligned sequences according to previous publications. Alignments were visualized with JalView 2.9 using Clustalx amino acids color display [89].

Phylogenetic analyses

To confirm the annotation of identified sponge genes as well as other early branching metazoans, a maximum likelihood (ML) analysis with 100 bootstraps was conducted using PhyML v3.1 [90] (implemented in Seaview 4.5.2). Bayesian analyses were conducted using MrBayes v3.2.5 [91]. Both ML and Bayesian analyses were run under the appropriate model recommended by ProtTest v3.4 [92].

For PALS1/MPP5/Stardust phylogeny, a first step consisted on inferring an ML and a Bayesian tree from SH3 + GUK domains of the “core MAGUK” of all proteins containing GUK domains retrieved from NCBI or predicted sequences of available non-bilaterian genome and/or transcriptomes (data not shown). This phylogeny includes MAGUK, Dlg, LRR and GUK domain, Zonula occludens (ZO), Caspase recruitment family (CARMA), Membrane-associated Guanylate kinase Inverted (MAGI) and Calcium channel β -subunit (CACNB) classes were also included even MAGI do not have a SH3 domain, their GUK domain is truncated, and Calcium Voltage-gated Channel auxiliary subunit beta (CACNB) are divergent.

Once all proteins of MPP classes were clearly identified, another phylogenetic analysis was performed including the PDZ domains besides the SH3 + GUK domains previously used.

The PATJ/MUPP-1/DLG/LIN phylogeny was performed on sequences identified using the L27 domains and the two following PDZ domains (except for the LIN family that exhibits a single PDZ domain).

All sequences annotated and analyzed are listed in Additional file 1: Tables S5 and S6.

Additional file

Additional file 1: Figure S1A. Comparison of p120 sequences. Residues involved in interaction with E-cadherin are boxed in red. Most of them are conserved. **Figure S1B.** Comparison of β -catenin sequences. A single β -catenin gene copy was identified in every studied species except for calcareous sponges that exhibit a duplication. All residues essential for E-cadherin interaction are boxed in pink and are highly conserved except for the R386 and N387 residues (replaced by L and T, respectively) in two hexactinellids and a more anecdotal change from A656 to S in placozoans. Residues boxed in blue are involved in α -catenin binding and in orange for the DTDL PDZ binding motif. **Figure S1C.** Analyses of α -catenins and vinculin sequences. Sequences of α -catenins and vinculin were aligned based on the structural domains helix0 to helix5 in *Mus musculus* α -catenin and vinculin. Helices are boxed and the numbers at the end of each sequence indicate the range encompassed in the alignment. Secondary structure prediction by JNet (Jalview option) identified six helices in all sponge α -catenin sequences except for *A. queenslandica* (missing the 4 first helices) and *A. vastus* (missing helix0). All species analyzed in this study have one copy of α -catenin and one copy of vinculin well-separated in Bayesian tree with high support ($pp = 1$) (bottom). **Figure S2.** Structure of Par3 proteins in metazoans. Par3 exhibits a conserved N-terminal domain (CR1), three central PDZ domains, and a C-terminal region containing multiple protein binding sites including the aPKC-binding motif. **Figure S5.** Domain composition of PatJ (*D. melanogaster*), INADL and MUPP1 (*M. musculus*) and Multiple PDZ containing protein (MPDZ) (*O. lobularis*, *S. ciliatum*, *A. queenslandica* and *O. minuta*). Note that only *O. lobularis* exhibits an MPDZ with a well-detected L27 domain (Evalue = $8.5 \cdot 10^{-4}$) as bilaterians. *A. queenslandica* and *S. ciliatum* MPDZ have a low-scoring L27 domain (shaded in grey) according to the HMM profile search. There is no recognizable similarity to the L27 domain in the N-terminal region of *O. minuta* MPDZ. **Tables S1.** and **S2.** Information on the domain structures of the Lethal giant larvae (LGL) and Scribble (Src) proteins in various metazoans. Spreadsheet containing **Tables S3.** and **S4.** The spreadsheet contains information on the characteristics of the new private databases and public databases (nature = genome/transcriptome) and links for new sequences used in this study. Spreadsheet containing **Tables S5.** and **S6.** The spreadsheet contains information on the accession numbers or contig/scaffold references where candidate genes were identified. In bold accession numbers of sequences annotated from our new transcriptomic and genomic sponge datasets. Links for new sequences used in this study are provided. (PDF 2610 kb)

Acknowledgments

We thank Chris Toret for critical reading and English editing of this manuscript, the Observatoire des Sciences de l'Univers (Pythéas institute) and Christian Marshall (IMBE) for collecting *O. lobularis* and *O. minuta* samples and the Service Commun de Biologie Moléculaire (SCBM) of IMBE for providing facilities needed to prepare samples for sequencing.

Funding

This work was supported by a grant from A*MIDEX (n° ANR-11-IDEX-0001-02) funded by the "investissements d'Avenir" French Government program, managed by the French National Research Agency (ANR). We acknowledge the use of the PACA-Bioinfo Platform, supported by IBISA and France-Génomique (ANR-10-INBS-0009). The Le Bivic group is an "Equipe labellisée 2008 de La Ligue

Nationale contre le Cancer" and is supported by the labex INFORM (grant ANR-11-LABX-0054), CNRS and Aix-Marseille University. All funding bodies indicated had no role in the design or collection, analysis or interpretation of the data or in the writing of the manuscript.

Availability of data and materials

All new sequences are available on NCBI website: accession numbers and links are provided in Additional file 1: Tables S5 and S6.

Alignments and trees were deposited in Treebase: <http://purl.org/phylo/treebase/phylovs/study/TB2:S22408>

Authors' contributions

HB sequences retrieval and analysis and manuscript writing. ER project design, data analysis and manuscript writing. SS genome and transcriptome analysis, manuscript editing. CJ genome and transcriptome analysis, manuscript editing. JMC project design, data analysis, and manuscript writing. CB project design, data analysis and manuscript writing. ALB project design, data analysis and manuscript writing. All authors read and approved the final manuscript.

Ethics approval and consent to participate

Sponges were collected by Christian Marschal (IMBE) for the purpose of the study with the permission provided (permission n°894 of December, 18th, 2017 from the Provence-Alpes-Côte d'Azur prefecture).

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Received: 11 December 2017 Accepted: 23 April 2018

Published online: 24 May 2018

References

1. Niklas KJ. The evolutionary-developmental origins of multicellularity. *Am J Bot.* 2014;101:6–25.
2. Parfrey LW, Lahr DJG. Multicellularity arose several times in the evolution of eukaryotes (response to DOI 10.1002/bies.201100187). *BioEssays.* 2013;35:339–47.
3. Abedin M, King N. Diverse evolutionary paths to cell adhesion. *Trends Cell Biol.* 2010;20:734–42.
4. Leys SP, Riesgo A. Epithelia, an evolutionary novelty of metazoans. *J Exp Zool B Mol Dev Evol.* 2012;318:438–47.
5. Miller PW, Clarke DN, Weis WI, Lowe CJ, Nelson WJ. The evolutionary origin of epithelial cell-cell adhesion mechanisms. *Curr Top Membr.* 2013;72:267–311.
6. Tyler S. Epithelium—the primary building block for metazoan Complexity1. *Integr Comp Biol.* 2003;43:55–63.
7. Adamska M. Sponges as models to study emergence of complex animals. *Curr Opin Genet Dev.* 2016;39:21–8.
8. Le Bivic A. Evolution and cell physiology. 4. Why invent yet another protein complex to build junctions in epithelial cells? *Am J Physiol - Cell Physiol.* 2013;305:C1193–201.
9. Lanna E. Evo-devo of non-bilaterian animals. *Genet Mol Biol.* 2015;38:284–300.
10. Jenner RA, Wills MA. The choice of model organisms in evo-devo. *Nat Rev Genet.* 2007;8:311–4.
11. Boute N, et al. Type IV collagen in sponges, the missing link in basement membrane ubiquity. *Biol Cell.* 1996;88:37–44.
12. Ereskovsky AV, et al. The Homoscleromorph sponge *Oscarellalobularis*, a promising sponge model in evolutionary and developmental biology. *BioEssays.* 2009;31:89–97.
13. Ringrose JH, et al. Deep proteome profiling of *Trichoplax adhaerens* reveals remarkable features at the origin of metazoan multicellularity. *Nat Commun.* 2013;4:1408.
14. Fidler AL, et al. Collagen IV and basement membrane at the evolutionary dawn of metazoan tissues. *elife.* 2017;6. <https://doi.org/10.7554/eLife.24176>.
15. Adams EDM, Goss GG, Leys SP. Freshwater sponges have functional, sealing epithelia with high Transepithelial resistance and negative Transepithelial potential. *PLoS One.* 2010;5:e15040.

16. Leys SP, Hill A. The physiology and molecular biology of sponge tissues. *Adv Mar Biol.* 2012;62:1-56. <https://doi.org/10.1016/B978-0-12-394283-8.00001-1>.
17. Leys SP, Nichols SA, Adams EDM. Epithelia and integration in sponges. *Integr Comp Biol.* 2009;49:167-77.
18. Smith CL, Reese TS. Adherens junctions modulate diffusion between epithelial cells in *Trichoplax adhaerens*. *Biol Bull.* 2016;231:216-24.
19. Oda H, Takeichi M. Structural and functional diversity of cadherin at the adherens junction. *J Cell Biol.* 2011;193:1137-46.
20. Smith CL, et al. Novel cell types, neurosecretory cells and body plan of the early-diverging metazoan, *Trichoplax adhaerens*. *Curr. Biol. CB.* 2014;24:1565-72.
21. Srivastava M, et al. The *Trichoplax* genome and the nature of placozoans. *Nature.* 2008;454:955-60.
22. Hulpiau P, van Roy F. New insights into the evolution of metazoan Cadherins. *Mol Biol Evol.* 2011;28:647-57.
23. Fahey B, Degnan BM. Origin of animal epithelia: insights from the sponge genome: evolution of epithelia. *Evol Dev.* 2010;12:601-17.
24. Riesgo A, Farrar N, Windsor PJ, Giribet G, Leys SP. The analysis of eight transcriptomes from all Poriferan classes reveals surprising genetic complexity in sponges. *Mol Biol Evol.* 2014;31:1102-20.
25. Srivastava M, et al. The Amphimedon queenslandica genome and the evolution of animal complexity. *Nat.* 2010;466:720-6.
26. Dunn CW, Leys SP, Haddock SHD. The hidden biology of sponges and ctenophores. *Trends Ecol Evol.* 2015;30:282-91.
27. Jager M, Manuel M. Ctenophores: an evolutionary-developmental perspective. *Curr Opin Genet Dev.* 2016;39:85-92.
28. Leys SP. Elements of a 'nervous system' in sponges. *J Exp Biol.* 2015; 218:581-91.
29. Moroz LL. Convergent evolution of neural systems in ctenophores. *J Exp Biol.* 2015;218:598-611.
30. Moroz LL, Kohn AB. Independent origins of neurons and synapses: insights from ctenophores. *Philos Trans R Soc B Biol Sci.* 2016;371:20150041.
31. O'Malley MA, Wideman JG, Ruiz-Trillo I. Losing complexity: the role of simplification in macroevolution. *Trends Ecol Evol.* 2016;31:608-21.
32. Ryan JF, Chiodin M. Where is my mind? How sponges and placozoans may have lost neural cell types. *Philos Trans R Soc B Biol Sci.* 2015;370. <https://doi.org/10.1098/rstb.2015.0059>.
33. Halanych KM, Whelan NV, Kocot KM, Kohn AB, Moroz LL. Miscues misplace sponges. *Proc Natl Acad Sci U S A.* 2016;113:E946-7.
34. Pisani D, et al. Genomic data do not support comb jellies as the sister group to all other animals. *Proc Natl Acad Sci U S A.* 2015;112:15402-7.
35. Telford MJ, Moroz LL, Halanych KM. Evolution: a sisterly dispute. *Nat.* 2016; 529:286-7.
36. Murray PS, Zaidel-Bar R. Pre-metazoan origins and evolution of the cadherin adhesome. *Biol Open.* 2014;3:1183-95.
37. Assémat E, Bazellieres E, Pallesi-Pocachard E, Le Bivic A, Massey-Harroche D. Polarity complex proteins. *Biochim Biophys Acta BBA - Biomembr.* 2008; 1778:614-30.
38. Bazellieres E, Assémat E, Arsanto JP, Le Bivic A, Massey-Harroche D. Crumbs proteins in epithelial morphogenesis. *Front Biosci.* 2009;14:2149-69.
39. Chen J, Zhang M. The Par3/Par6/aPKC complex and epithelial cell polarity. *Exp Cell Res.* 2013;319:1357-64.
40. Elsum I, Yates L, Humbert PO, Richardson HE. The scribble-Dlg-Lgl polarity module in development and cancer: from flies to man. *Essays Biochem.* 2012;53:141-68.
41. Boggon TJ, et al. C-cadherin Ectodomain structure and implications for cell adhesion mechanisms. *Sci.* 2002;296:1308-13.
42. Nichols SA, Roberts BW, Richter DJ, Fairclough SR, King N. Origin of metazoan cadherin diversity and the antiquity of the classical cadherin/ β -catenin complex. *Proc Natl Acad Sci U S A.* 2012;109:13046-51.
43. Shapiro L, Weis WI. Structure and biochemistry of Cadherins and catenins. *Cold Spring Harb Perspect Biol.* 2009;1:a003053.
44. Clarke DN, Miller PW, Lowe CJ, Weis WI, Nelson WJ. Characterization of the cadherin/Catenin complex of the sea Anemone *Nematostella vectensis* and implications for the evolution of metazoan cell-cell adhesion. *Mol Biol Evol.* 2016;33:2016-29.
45. Ishiyama N, et al. Dynamic and static interactions between p120 catenin and E-cadherin regulate the stability of cell-cell adhesion. *Cell.* 2010;141:117-28.
46. Huber AH, Weis WI. The structure of the β -catenin/E-cadherin complex and the molecular basis of diverse ligand recognition by β -catenin. *Cell.* 2001; 105:391-402.
47. Bao R, Fischer T, Bolognesi R, Brown SJ, Friedrich M. Parallel duplication and partial subfunctionalization of γ -catenin/Armadillo during insect evolution. *Mol Biol Evol.* 2012;29:647-62.
48. Pai L-M, et al. Drosophila α -catenin and E-cadherin bind to distinct regions of Drosophila Armadillo. *J Biol Chem.* 1996;271:32411-20.
49. Noda Y, et al. Molecular recognition in dimerization between PB1 domains. *J Biol Chem.* 2003;278:43516-24.
50. Horikoshi Y, et al. Interaction between PAR-3 and the aPKC-PAR-6 complex is indispensable for apical domain development of epithelial cells. *J Cell Sci.* 2009;122:1595-606.
51. Ganot P, et al. Structural molecular components of septate junctions in cnidarians point to the origin of epithelial junctions in eukaryotes. *Mol Biol Evol.* 2015;32:44-62.
52. Su W-H, Mruk DD, Wong EWP, Lui W-Y, Cheng CY. Polarity protein complex scribble/Lgl/Dlg and epithelial cell barriers. *Adv Exp Med Biol.* 2012;763:149-70.
53. Albertson R, Chabu C, Sheehan A, Doe CQ. Scribble protein domain mapping reveals a multistep localization mechanism and domains necessary for establishing cortical polarity. *J Cell Sci.* 2004;117:6061-70.
54. Bulgakova NA, Knust E. The crumbs complex: from epithelial-cell polarity to retinal degeneration. *J Cell Sci.* 2009;122:2587-96.
55. Pocha SM, Knust E. Complexities of crumbs function and regulation in tissue morphogenesis. *Curr Biol CB.* 2013;23:R289-93.
56. Baines AJ, Lu H-C, Bennett PM. The protein 4.1 family: hub proteins in animals for organizing membrane proteins. *Biochim Biophys Acta.* 2014; 1838:605-19.
57. Bachmann A, Schneider M, Theilenberg E, Grawe F, Knust E. Drosophila stardust is a partner of crumbs in the control of epithelial cell polarity. *Nat.* 2001;414:638-43.
58. Lemmers C, et al. CRB3 binds directly to Par6 and regulates the morphogenesis of the tight junctions in mammalian epithelial cells. *Mol Biol Cell.* 2004;15:1324-33.
59. Wei Z, Li Y, Ye F, Zhang M. Structural basis for the phosphorylation-regulated interaction between the cytoplasmic tail of cell polarity protein crumbs and the actin-binding protein Moesin. *J Biol Chem.* 2015;290:11384-92.
60. Borowiec ML, Lee EK, Chiu JC, Plachetzki DC. Extracting phylogenetic signal and accounting for bias in whole-genome data sets supports the Ctenophora as sister to remaining Metazoa. *BMC Genomics.* 2015;16:987.
61. Simion P, et al. A large and consistent Phylogenomic dataset supports sponges as the sister group to all other animals. *Curr Biol CB.* 2017;27: 958-67.
62. Whelan NV, Kocot KM, Halanych KM. Employing Phylogenomics to resolve the relationships among cnidarians, ctenophores, sponges, Placozoans, and Bilaterians. *Integr Comp Biol.* 2015;55:1084-95.
63. Magie CR, Martindale MQ. Cell-cell adhesion in the Cnidaria: insights into the evolution of tissue morphogenesis. *Biol Bull.* 2008;214:218-32.
64. Nielsen C. Animal evolution: interrelationships of the living Phyla. Oxford: OUP; 2012.
65. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnetJ.* 2011;17:10.
66. Peng Y, Leung HCM, Yiu SM, Chin FYL. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinforma Oxf Engl.* 2012;28:1420-8.
67. Kajitani R, et al. Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads. *Genome Res.* 2014;24:1384-95.
68. Boetzer M, Pirovano W. Toward almost closed genomes with GapFiller. *Genome Biol.* 2012;13:R56.
69. Huang X, Madan A. CAP3: a DNA sequence assembly program. *Genome Res.* 1999;9:868-77.
70. Kim D, et al. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* 2013; 14:R36.
71. Hoff KJ, Lange S, Lomsadze A, Borodovsky M, Stanke M. BRAKER1: unsupervised RNA-Seq-based genome annotation with GeneMark-ET and AUGUSTUS. *Bioinforma Oxf Engl.* 2016;32:767-9.
72. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol.* 1990;215:403-10.
73. Langmead B, Salzberg SL. Fast gapped-read alignment with bowtie 2. *Nat Methods.* 2012;9:357-9.
74. Koren S, et al. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res.* 2017;27:722-36.

75. Bankevich A, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol. J. Comput. Mol. Cell Biol.* 2012;19:455–77.
76. Zhu W, Lomsadze A, Borodovsky M. Ab initio gene identification in metagenomic sequences. *Nucleic Acids Res.* 2010;38:e132.
77. Boetzer M, Henkel CV, Jansen HJ, Butler D, Pirovano W. Scaffolding pre-assembled contigs using SSPACE. *Bioinforma Oxf Engl.* 2011;27:578–9.
78. Walker BJ, et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One.* 2014; 9:e112963.
79. Lukashin AV, Borodovsky M. GeneMark.Hmm: new solutions for gene finding. *Nucleic Acids Res.* 1998;26:1107–15.
80. Burge C, Karlin S. Prediction of complete gene structures in human genomic DNA. *J Mol Biol.* 1997;268:78–94.
81. Stanke M, Morgenstern B. AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucleic Acids Res.* 2005;33: W465–7.
82. Solovyev V, Kosarev P, Seledsov I, Vorobyev D. Automatic annotation of eukaryotic genes, pseudogenes and promoters. *Genome Biol.* 2006;7(Suppl 1):S10.1–12.
83. Finn RD, et al. The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.* 2016;44:D279–85.
84. Jones P, et al. InterProScan 5: genome-scale protein function classification. *Bioinforma Oxf Engl.* 2014;30:1236–40.
85. Schultz J, Milpetz F, Bork P, Ponting CP. SMART, a simple modular architecture research tool: identification of signaling domains. *Proc Natl Acad Sci U S A.* 1998;95:5857–64.
86. Katoh K, Misawa K, Kuma K, Miyata T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* 2002;30:3059–66.
87. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 2004;32:1792–7.
88. Gouy M, Guindon S, Gascuel O. SeaView version 4: a multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Mol Biol Evol.* 2010;27:221–4.
89. Waterhouse AM, Procter JB, Martin DMA, Clamp M, Barton GJ. Jalview version 2—a multiple sequence alignment editor and analysis workbench. *Bioinforma Oxf Engl.* 2009;25:1189–91.
90. Guindon S, et al. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol.* 2010;59:307–21.
91. Ronquist F, Huelsenbeck JP. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinforma Oxf Engl.* 2003;19:1572–4.
92. Abascal F, Zardoya R, Posada D. ProtTest: selection of best-fit models of protein evolution. *Bioinforma Oxf Engl.* 2005;21:2104–5.