Acta Societatis Botanicorum Poloniae

INVITED ORIGINAL RESEARCH PAPER Acta Soc Bot Pol 83(4):337–347 DOI: 10.5586/asbp.2014.034
Received: 2014-08-20 Accepted: 2014-10-13 Published electronically: 2014-12-31

The number of cell types, information content, and the evolution of complex multicellularity

Karl J. Niklas[1]*, Edward D. Cobb[1], A. Keith Dunker[2]

[1] Section of Plant Biology, School of Integrative Plant Science, Cornell University, Ithaca, NY 14853, USA

[2] Center for Computational Biology and Bioinformatics, School of Medicine, Indiana University, Indianapolis, IN 46202, USA

Abstract

The number of different cell types (NCT) characterizing an organism is often used to quantify organismic complexity. This method results in the tautology that more complex organisms have a larger number of different kinds of cells, and that organisms with more different kinds of cells are more complex. This circular reasoning can be avoided (and simultaneously tested) when NCT is plotted against different measures of organismic information content (e.g., genome or proteome size). This approach is illustrated by plotting the NCT of representative diatoms, green and brown algae, land plants, invertebrates, and vertebrates against data for genome size (number of base-pairs), proteome size (number of amino acids), and proteome functional versatility (number of intrinsically disordered protein domains or residues). Statistical analyses of these data indicate that increases in NCT fail to keep pace with increases in genome size, but exceed a one-to-one scaling relationship with increasing proteome size and with increasing numbers of intrinsically disordered protein residues. We interpret these trends to indicate that comparatively small increases in proteome (and not genome size) are associated with disproportionate increases in NCT, and that proteins with intrinsically disordered domains enhance cell type diversity and thus contribute to the evolution of complex multicellularity.

Keywords: algae; alternative splicing; embryophytes; genome size; G paradox; intrinsically disordered proteins; land plants

Introduction

The number of different cell types characterizing an organism has been used to measure organismic complexity because it quantifies the extent to which cellular structure is differentiated into phenotypically different entities (e.g., [1–5]), because this approach is indifferent to whether an organism is a fungus, plant, or animal, and because it is insensitive to most methods of categorizing grade or clade levels of organization. It can even be used to quantify the complexity of unicellular organisms because the vast majority of species with this body plan achieve different cell functionalities and morphologies at different stages in their life cycles (e.g., resting cysts versus actively motile cells). Nevertheless, the use of the number of cell types as a measurement of complexity has some drawbacks. For example, some workers may disagree about whether two cells differ sufficiently to be called different cell types (e.g., xylary versus extra-xylary fibers, and parenchyma versus prosenchyma cells). Gauging complexity by counting the number of different cell types an organism can make also becomes problematic when dealing with coenocytic organisms that can achieve considerable morphological complexity without benefit of cellularization (e.g., the green alga *Caulerpa*), or when dealing with comparatively simple morphologies that nevertheless have highly organized patterns of nuclear or cellular placement (e.g., angiosperm megagametophytes).

There are two other important concerns. The first is that using the number of cell types to quantify complexity results in an untested tautology – "complexity increases as the number of different types of cells increases, and this number increases as an organism's complexity increases" – which requires collateral evidence to show that this reasoning is more than a self-sustaining description of a phenotypic property. The second concern is that cellular simplicity need not reflect an absence of complexity. There is no reason a priori to argue that an organism capable of producing two different types of cells is more complex than an organism capable of producing only one. And, if there is justification, is an organism composed of two cell types twice as complex as an organism composed of one cell type? Further, the supposition that cellular simplicity denotes low

* Corresponding author. Email: kjn2@cornell.edu

Handling Editor: Beata Zagorska-Marek

This is an Open Access article distributed under the terms of the Creative Commons Attribution 3.0 License (creativecommons.org/licenses/by/3.0/), which permits redistribution, commercial and non-commercial, provided that the article is properly cited. © The Author(s) 2014 Published by Polish Botanical Society

levels of organismic complexity hinges on the notion that "complexity" must be defined at the level of the diversity of cell phenotypes, a notion that is challenged when considering the physiological (as opposed to the structural) complexity of parasitic organisms that often undergo morphological and anatomical reduction (e.g., the Indian Pipe *Monotropa* and the tapeworm *Taenia*).

Nevertheless, although the concept of biological complexity is philosophically slippery, intuition encourages the notion that more "information" of some sort is required to produce or sustain more different kinds of cells, and that "information content" is a measure of "complexity". If we accept this premise, the number of different kinds of cells is not a direct measure of complexity per se but rather an indirect reflection of information content and thus complexity. This logic fosters a research agenda requiring the quantification of "information context" and a statistical assessment of its correlation with the number of cell types an organism produces. Only in this way can the tautology be tested.

A number of candidates for biological "information" present themselves in this capacity. For example, genome size as measured by the number of base-pairs, or proteome size as measured by the number of amino acids arguably provide reasonable measures of an organism's information content, particularly if attention is paid to the effects of polyploidy on genome size. Another measure of information content is the number of intrinsically disordered protein domains or residues in an organism's proteome, which gauge a proteome's functional versatility. This measure is particularly interesting because intrinsically disordered domains lack an equilibrium 3D structure under normal physiological conditions. Consequently, each domain can assume multiple conformations (and thus multiple developmental regulatory functions) within the same cell without an unnecessary and inefficient expansion of genome or proteome size [6–8]. In this way, proteins with intrinsically disordered domains contain information that is not encoded directly in the genome, a feature that helps in part to explain the G paradox. Finally, two of the advantages of using these three measurements of organismic information content is that each can be applied to unicellular as well as to multicellular organisms, and each can be used to assess whether complexity has increased over any lineage's evolutionary history.

The goal of this paper is to examine whether genome size, proteome size, and, in particular, the number of intrinsically disordered proteins correlate with the number of different cell types diverse organisms produce, and, if they do, to explore the implications of these correlations to understanding the evolution of multicellularity. Toward this goal, we present and analyze data drawn from animal as well as algal and land plant species because there is no reason a priori to expect plant and animal evolution to produce vastly different relationships between information content and cell type diversity, and because, should different patterns emerge among different kinds of organisms, they might shed light on how multicellularity emerged in different lineages. Finally, although it is not conventional, for the purposes of this paper, we define "plants" as photosynthetic eukaryotes so as to include the polyphyletic algae as well as the monophyletic embryophytes because excluding the algae would preclude an examination of unicellular and colonial organisms that are conventionally used to assemble ancestor-descendant transformation series for the evolution of multicellular body plans [9–13].

Cell type numbers do not increase in all lineages

Despite the tautology that results from the supposition that the number of different cell types increases as organismic complexity increases, it is informative to examine whether any trends of this sort are evident in successively evolutionarily divergent taxa within well-documented clades, particularly if these trends flout the conventional wisdom that "plants are invariably simpler than animals". There are however a number of considerations and limitations to this enterprise. For example, the level of taxonomic resolution and the availability of sufficiently resolved phylogenies can present problems. A too finely resolved phylogeny may reveal no pattern because very closely related taxa may share the same or very similar cellular structure and thus the same or very similar number of different cell types. Conversely, too coarsely a resolved phylogeny may fail to reveal a pattern because distantly related taxa may have adapted to very different niches, or undergone anatomical reduction. Two other limitations exist: phylogenies resolved at the optimal level may not be available, and authoritative tabulations of the number of different cell types for species in sufficiently resolved phylogenies may not exist.

These concerns are illustrated by mapping estimates of the number of different cell types reported for representative taxa onto published phylogenies based on molecular data. Specifically, we use the maximum number of cell types (NCT) reported in the primary literature as tabulated by Bell and Mooers [2] and map NCT onto the different plant and animal lineages for which there are published phylogenies using molecular data. The data set assembled by Bell and Mooers [2] undoubtedly reflects divergent philosophies about how to distinguish and classify animal and plant cell types (i.e., the data reflect the opinions of "splitters" and "lumpers"), and thus may be biased in unforeseen ways. However, this data set has the advantage of reflecting the opinions of authorities in diverse fields of research. It also consolidates a large literature scattered in diverse journals.

The first example of this approach employs a multi-locus, time calibrated phylogeny for the brown algae (Phaeophyceae) [14] and a broader phylogeny for the heterokont algae [15]. Mapping maximum NCT onto these phylogenies reveals no clearly discernable pattern within the phylogeny of the Phaeophyceae or the heterokont algae (Fig. 1). Within the Phaeophyceae, the early divergent Sphacelariales, Syringodermatales, and Dictyotales are represented by species with a maximum of four to seven different cell types, whereas the late divergent Fucales and Nemodermatales share similar and small cell type numbers (Fig. 1a). The exception is the Laminariales with a maximum of 14 different types of cells. This order is notable because it is characterized by species with large (up to 60 m in length), highly differentiated sporophytes with a holdfast-stipe-blade construction (e.g., *Nereocystis luetkaena* and *Pelagophycus porra*) [16],

and because many species have highly specialized tissues (e.g., "trumpet cells" in *Macrocystis pyrifera*) that provide symplastic conduits for nutrient transport [17–19]. These features are associated with the ecological expansion of the kelps into intertidal zones, which provides an ecological explanation for the morphological and anatomical complexity seen in the kelps.

When the Phaeophyceae is embedded within the larger phylogeny of the heterokont algae based on a concatenated analysis of SSU rRNA and *rbcL* reported by Andersen [15], we see that this group as a whole contains species with larger cell type numbers compared to groups preceding the divergence of the Phaeophyceae within the clade and that these groups contain predominantly unicellular or colonial organisms with two to four cell types. For example, the species in the sister group of the Phaeophyceae, the raphidophytes, characteristically have two cell types (Fig. 1b). Unfortunately, comparisons of the various molecular phylogenies for the heterokont algae identify few consensus relationships among many lineages, with the exception perhaps of the diatoms and bolidophytes [14]. Consequently, until future studies accurately ascertain the evolutionary relationships within this large clade, little more can be said about the evolution of NCT within this important group of organisms.

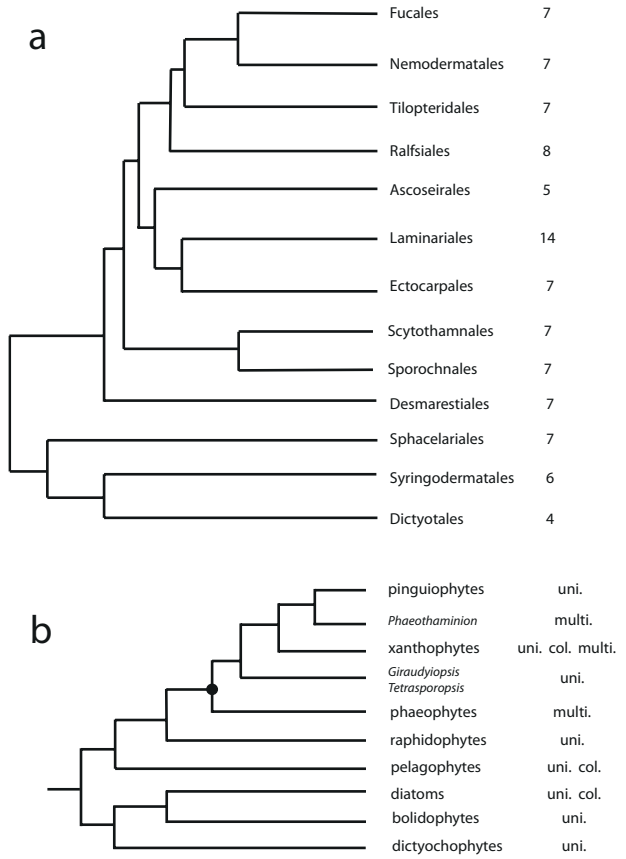

Fig. 1 Redacted phylogenetic relationships among (**a**) Phaeophyceae orders and the maximum number of cell types reported for any species within each order, and (**b**) redacted phylogenetic relationships among the heterokont algae showing the characteristic body plans reported for each group (uni. – unicellular; col. – colonial; multi. – multicellular). Data for number of cell types taken from Bell and Mooers [2]; Phaeophyceae phylogeny based on Silberfield et al. [14]; heterokont phylogeny based on Andersen [15].

In contrast, the consensus phylogeny of the viridiplantae (the chlorophycean and charophycean algae, and the embryophytes) indicates that NCT has increased, albeit not monotonically within this important clade (Fig. 2a). The early divergent chlorophytes are unicellular and have two cell types (e.g., the prasinophyte *Ostreococcus*), whereas later divergent green algae have a maximum of five different types of cells (e.g., *Fritschiella*). Among the streptophytes (the charophycean algae and the embryophytes), the early divergent charophycean alga *Mesostigma* has two cell types, whereas the more derived charophycean alga *Chara* has ten. Curiously, a phylogenetic analysis using 360 plastid genes [20] identifies the Zygnematophyceae as the closest related group to the land plants. This relationship contrasts with the hypothesis that the charophycean algae, such as *Chara* and *Coleochaete*, are more closely related to the land plants (e.g., [15]). This hypothesis is consistent with a large number of what have been traditionally considered shared derived

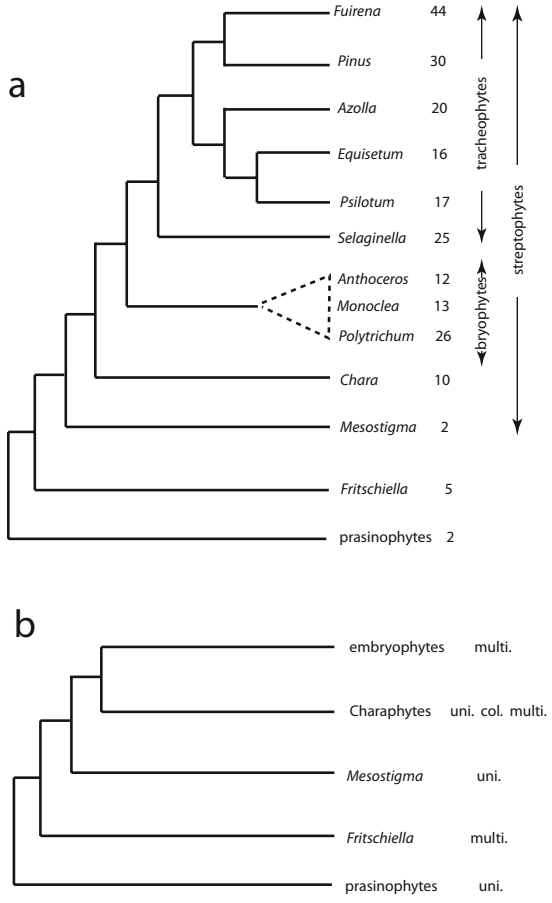

Fig. 2 Redacted phylogenetic relationships among representative taxa within the viridiplantae, the maximum number of cell types reported for species within each taxon (**a**), and the body plans represented within each taxon (uni. – unicellular; col. – colonial; multi. – multicellular; **b**). With the exception of the early divergent chlorophytes (the prasinophytes), each major taxon is represented by the genus reported to have the greatest diversity of cell types in its lineage (e.g., *Chara* and *Polytrichum* are reported to have the largest number of cell types among the charophycean algae and the mosses, respectively). The uncertainty in the phylogenetic relationships among the nonvascular land plants (the bryophytes) is depicted by dashed lines. Data for number of cell types taken from Bell and Mooers [2]. Phylogenetic relationships based in part on Ruhfel [20].

features such as oogamy and plasmodesmata [21], and IAA polar transport [22]. The perplexing placement of the Zygnematophyceae in the streptophytes based on molecular data is explicable if considerable evolutionary reduction and loss of numerous synapomorphies (such as oogamy and flagellated sperm cells) occurred. For example, the zygnematacean algae do not produce flagella and reproduce sexually by conjugation. They also produce a maximum number of only three or four cell types, which is significantly less than that observed for *Chara*, *Nitella*, or *Coleochaete*. Perhaps, like the complex Laminariales, which adapted to a structurally demanding intertidal environment, the comparatively simple Zygnematophyceae, may have radiated into ecologically less demanding environments.

Turning to the land plants, we see that the maximum number of cell types among the nonvascular embryophytes is reported for the mosses (i.e., *Polytrichum*; NCT = 26), which is approximately twice that reported for the hornworts and liverworts (e.g., *Anthoceros* and *Monoclea*; NCT = 12 and 13, respectively). Among the lycopods, the maximum NCT is 25 for *Selaginella*, whereas among the ferns and horsetails, the number of different cell types ranges between 17 and 20 (*Psilotum* and *Azolla*, respectively). Finally, *Pinus monophylla* is reported to have the maximum number of different cell types among conifers, whereas *Fuirena ciliaris* is described as having 44 different kinds of cells, which is the maximum reported for flowering plants ([2]; see [23] and [24]).

Turning to metazoans, it is worth noting that the maximum number of different cell types reported for the flowering plants exceeds that reported for Placozoa (NCT = 4) and early divergent animal groups such as the cnidarians and ctenophorans, and is comparable to those reported for the protostomes (i.e., ecdysozoans and lophotrochozoans; Fig. 3, Fig. 4). Although the resolution of the phylogeny of animals is not complete [25], the available data indicate that maximum cell type numbers range between 16 and 22 for the early divergent poriferans and cniderians, and between 9 and 69 for the mollusks and arthropods. The greatest number of different cell types among vertebrates is 240, which is 5.5 times that of the maximum number of angiosperm cell types (Fig. 4). The diversity of cell types across plant and animals is perhaps best illustrated when we compare the mean number of cell types reported for plants and metazoans (Fig. 3).

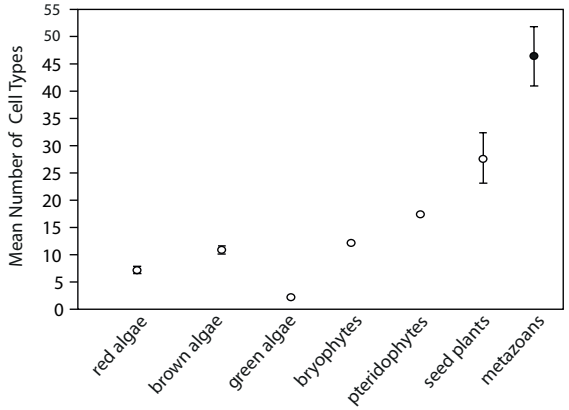

Fig. 3 Mean number of cell types (±*SE*) reported for extant red, brown, and green algae, bryophytes, pteridophytes, seed plants, and metazoans. Data taken from Bell and Mooers [2].

Among the plants, the red and brown algae produce a larger number of different kinds of cells compared to the green algae (chlorophytes and charophycean algae), whereas the NCT of brown algae is statistically indistinguishable from the NCT of the bryophytes. Within the viridiplantae, mean NCT increases progressively across the bryophytes, pteridophytes, and vascular plants (tracheophytes). The metazoans, on average, manifest the greatest diversity of cell types and the greatest range among all of the multicellular lineages.

The preceding provides very limited insights into how the numbers of cell types are distributed in the major plant lineages. It does, however, illustrate that the ability to detect patterns in NCT depends on the degree to which a lineage or clade is phylogenetically resolved. Nevertheless, it is clear that the number of different cell types does not invariably increase over the course of evolutionary history. Consequently, if the generalization that "cell type numbers gauge complexity" is accepted as true, it follows that evolution has not always resulted in more complex organisms in all lineages or clades (even in the absence of the evolution of adopting a parasitic life style). They also dispel the myth that plants are invariably less complex than animals, since many land plants have as many or more different cell types as annelids or arthropods (see Fig. 2, Fig. 4).

Cell type numbers increase with increasing proteome size

If the maximum number of cell types does not invariably increase over the course of a lineage's evolutionary history, does it correlate across taxa with organismic information content as gauged by genome or proteome size?

The expectation is that statistically significant and positive correlations will exist between NCT and proteome size based on a consideration of how alternative splicing and intrinsically disordered protein domains (which occur widely in all eukaryotic lineages) amplify protein functionalities without inefficiently increasing genome size. These molecular processes result in "moonlighting" proteins (i.e., proteins that

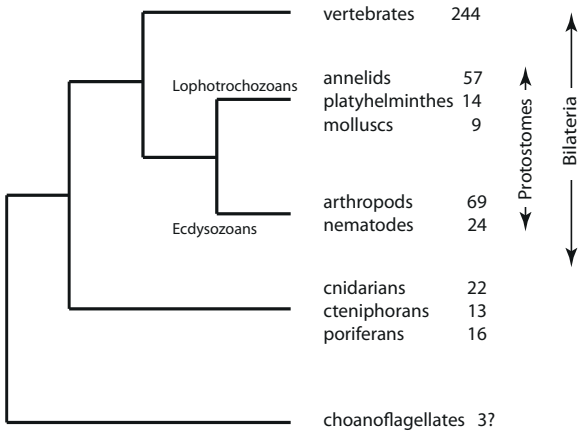

Fig. 4 Redacted phylogenetic relationships among the major groups of metazoans and the maximum number of cell types reported for each group. Data for number of cell types (NCT) taken from Bell and Mooers [2]. Phylogenetic relationships based on Adoutte et al. [25].

can take on a variety of functions, some of which increase the ability to diversify cell types and thus increase morphological and anatomical complexity).

Alternative splicing is an adaptive and highly conserved mechanism because it increases the number of proteins encoded by pre-mRNA in a context-dependent manner in both plants and animals. For example, Chang et al. [26] report a conserved alternative splicing pattern for heat shock transcription factors in the moss *Physcomitrella patens* and the flowering plant *Arabidopsis thaliana*, and show that the alternative splicing mechanism for heat regulation among the land plants is an ancestral condition. Using mRNA sequence data, Pan et al. report that transcripts from ≈95% of human multi-exon genes undergo alternative splicing and that ≈100 000 intermediate to high abundance alternative splicing events occur in a tissue-specific manner [27]. These results are similar to those reported by Johnson et al. using microarray analyses of human tissues [28]. In turn, alternative splicing produces a disproportionate number of intrinsically disordered proteins that lack an equilibrium 3D structure and thus can take on multiple functionalities under normal physiological conditions [6–8]. The majority of transcription factors, which are key players in cell fate specification, have intrinsically disordered protein domains affected by alternative splicing. This enables functional and regulatory diversity while avoiding structural complications [29]. It is reasonable to surmise therefore that the diversity of cell types will increase as a function of the size of a proteome and (in particular) as the number of intrinsically disordered domains increases.

This speculation is consistent with an examination of the data gathered from the primary literature reporting NCT, genome and proteome size, and the number of intrinsically disordered protein domains as gauged by the number of intrinsically disordered residues (IDResidues). Across 19 species of algae ($n = 8$), nonvascular and vascular land plants ($n = 6$), and invertebrates and vertebrates ($n = 5$) whose genomes have been completely sequenced, we find that NCT correlates positively and significantly ($P < 0.0001$) with each of the three metrics of information content (Fig. 5) as gauged by ordinary least squares protocols (which stipulate Y and X as the dependent and independent variables, respectively) using $\log_{10}$-transformed data (Tab. 1). Equally informative is the numerical values of the slopes of each of the bivariate regression curves, because these slopes are scaling exponents (denoted here by α) that indicate numerically the extent to which NCT proportionally increases or decreases with respect to increasing values of each of the other variables on a log scale [30]. Inspection of these slopes indicates that NCT scales roughly as the 2.18 power of the number of intrinsically disordered residues and as the 2.36 power of proteome size. Thus, the degree of cellular specialization increases dramatically as each of these two measures of information content increases. Although the number of intrinsically disordered residues scales nearly one-to-one (isometrically) with respect to proteome size (i.e., $\alpha = 0.97$), NCT fails to increase one-to-one with increasing genome size (i.e., $\alpha = 0.88$), which is a reflection of the so-called G paradox (for a discussion in the context of alternative splicing, see [31]). Collectively, these scaling relationships indicate that the number of intrinsically disordered residues is a highly significant measure of information content if we accept the hypothesis that NCT reflects organismic information content) and that the information contained in these residues helps to explain where some of the "missing" genetic information exists.

Doubts and admonitions

A number of caveats with regard to the preceding analyses exist. The first is that correlations among variables of interest do not constitute proof for cause and effect relationships. This can be illustrated by noting that each of the organisms used in the preceeding analyses is determinate in its growth in overall size. Using data for the maximum body length of the species examined (reported by Lang et al. [32]), regression of the $\log_{10}$-transformed data obtains a statistically significant and positive correlation between maximum body size and NCT ($r^2 = 0.861$, $P < 0.0001$) and a numerically large scaling exponent (i.e., $\alpha = 2.84$; Fig. 6). However, there is no biologically valid reason to assume that body size is dependent on the number of different cell types an organism can produce. The vascular plants produce a maximum of 44 different cell types [24], yet plants are the largest organisms on Earth. In contrast, *Homo sapiens* produce approximately 240 different cell types, yet humans reach a maximum body length of approximately 2.5 m. This example shows that strong correlations can be spurious, if not irrelevant.

Another concern is the nearly flat-line relationship between the number of different cell types and the information content of unicellular plants. For example, once again using $\log_{10}$-transformed data, regression of NCT against the number of intrinsically disordered residues in the six unicellular plants yields a scaling exponent of 0.41 ($r^2 = 0.282$, $P = 0.278$), as opposed to a scaling exponent of 1.84 for the 13 multicellular plants and animals ($r^2 = 0.536$, $P = 0.004$). Taken at face value, these relationships indicate that increases in the information content of unicellular organisms have less effect on NCT compared to multicellular organisms. A third concern is that we know comparatively little about the ancestors of many of the species used in these analyses. Drawing inferences about ancestral morphologies or life styles based on the anatomy or ecology of extant species is highly problematic because we cannot discount the possibility that a seemingly simple species is the descendant of a morphologically or reproductively more complex species.

Clearly, much more research is required to draw definitive conclusions. All that can be said with any certainty is that NCT increases with proteome size and with the number of intrinsically disordered protein residues for the species used in these analyses.

Mechanisms for cell fate specification existed before the evolution of multicellularity

Much has been written about the evolution of multicellularity from the perspective of the mechanisms giving rise to it and the selective advantages, if any, it confers. The traditional

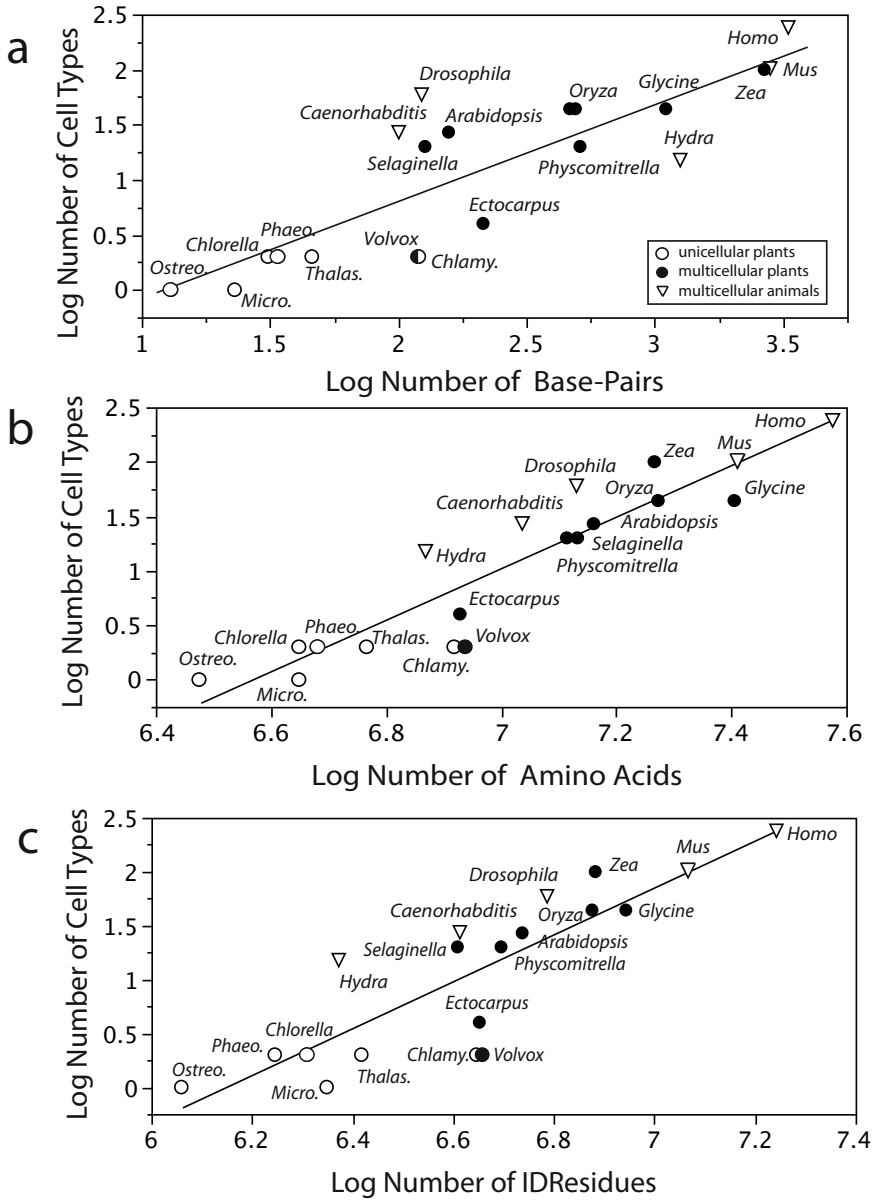

Fig. 5 Log_{10}-transformed data for different cell types (NCT) plotted against log_{10}-transformed data for genome size (number of base-pairs; **a**), proteome size (number of amino acids; **b**), and protein functional diversity (number of intrinsically disordered protein residues, IDResidues; **c**) reported for unicellular algae, multicellular algae and land plants, invertebrates, and vertebrates. Straight lines are ordinary least squares regression curves. *Chlamy.* – *Chlamydomonas reinhardii*; *Micro.* – *Micromonas pusilla* NOUM 17; *Ostreo.* – *Ostreococcus tauri*; *Phaeo.* – *Phaeodactylum tricornutum*; *Thalas.* – *Thalassiosira pseudonana*. For statistical regression parameters, see Tab. 1. Data for NCT taken from Bell and Mooers [2]. Data for IDResidues and AA taken from Oates et al. [46]. Data for IDResidues are based in predictions made by PONDR® VLS2b (see [45]), which ranked as the best for predicting segments of disorder and gave the highest per-residue accuracy for long regions of disorder [52]. Data for genome size taken from Internet sources.

and widely accepted scenario for the evolution of multicellularity postulates a transformation series of body plans starting with a unicellular ancestor, passing through an intermediate organism with a colonial body plan, and culminating with the evolution of simple multicellularity, i.e., a unicellular → colonial → (simple) multicellular body plan transformation series (e.g., [9,10]). This scenario complies reasonably well with phylogenetic trends in body plans within plant and animal groups (see Fig. 2b, Fig. 4), although instances of coenocytic → (simple) multicellular transformation series are known (see [33]). In some lineages, the unicellular → colonial → (simple) multicellular transformation series led to the acquisition of complex multicellularity (for reviews, see [12,33]) and, in some cases, the segregation of germ and soma cell lines [34–36]. Simple multicellularity occurs when every cell in a body plan makes direct contact with the external environment (e.g., filaments of cells), whereas complex multicellularity occurs when some cells make contact only with neighboring cells (e.g., solid blastula-like cellular structures) [37] (see also [38]). The internalization of cells is often associated with the presence of super-cellular transport systems (e.g., vascular plant tissues and mammalian

Tab. 1 Bivariate regression parameters (see Fig. 5) for $\log_{10}$–transformed data of the number of different cell types (NCT), genome size (G, in mbp), number of amino acids (AA), the number of intrinsically disordered residues (IDResidues), and maximum body length (BL) reported for 19 species of algae, land plants, and animal species[a].

Log Y vs. Log X	slope (α–value)	r^2	P	F
NCT vs. IDResidues	2.18	0.721	<0.0001	44
NCT vs. AA	2.36	0.894	<0.0001	110
IDResidues vs. AA	0.97	0.94	<0.0001	46.7
NCT vs. G	0.88	0.709	<0.0001	43.9
BL vs. NCT	2.84	0.861	<0.0001	112

[a] Species composition: brown algae (n = 1 species; *Ectocarpus siliculosus*), green algae (n = 3 species: *Chlorella* sp., *Chlamydomonas reinhardtii*, and *Volvox carteri*), diatoms (n = 4 species: *Micromonas pusilla* NOUM 17, *Ostreococcus tauri*, *Phaeodactylum tricornutum*, and *Thalassiosira pseudonana*), mosses (n = 1: *Physcomitrella patens*), lycophytes (n = 1 species: *Selaginella moellendorfii*), flowering plants (n = 4 species, *Arabidopsis thaliana*, *Glycine max*, *Oryza sativa*, and *Zea mays*), and metazoans (n = 5 species: *Hydra attenuate*, *Caenorhabditis elegans*, *Drosophila melanogaster*, *Mus musculus*, and *Homo sapiens*). Data for NCT taken from Bell and Mooers [2]. Data for IDResidues and AA taken from Oates et al. [46]. IDResidues values based on PONDR® VLS2b (see [45]) predictions, which is ranked as the best protocol for predicting segments of disorder and gives the highest per-residue accuracy for long regions of disorder [52]. Data for G taken from Internet sources.

pulmonary system), because it makes reliance on passive diffusion for the acquisition of nutrients insufficient to keep pace with metabolic demands (see [34]).

However, it is apparent that the ability to produce more than one cell type is not confined to complex or even simple multicellular life forms. Numerous unicellular, colonial, and filamentous organisms produce different cell types, e.g., the unicellular cyanobacterium *Nostoc* and the colonial volvocine alga *Pleodorina* produce two cell types, the filamentous cyanobacterium *Rivularia* produces three cell types, and the filamentous brown alga *Ectocarpus* produces four cell types. The ability of unicellular organisms to produce two or more different cell types reflects the existence of alternative stable states of gene expression patterns. It is reasonable therefore to suppose that the mechanism responsible for cell fate specification is ancient and prokaryotic in origin, and that its initial adaptive value had nothing to do with multicellularity per se but rather with adapting to different internal or external conditions.

The traditionally accepted mechanism for cell fate specification relies on the inherent multi-stability of gene regulatory networks (e.g., [39–41]). Each stable state provides the conditions that specify a cell type that can be "fixed" in a developmental repertoire ad hoc by natural selection as an organism adapts to its different environmental conditions. Mathematical models show that cellular differentiation can emerge among genetically identical cells as a response to poor compatibility among competing physiological processes [42], whereas, in more derived lineages, an alignment of fitness among adhering cells in a colony might compensate for conflicts of interest among cellular components such that a division of cellular labor becomes possible and even necessary (for a review, see [12]).

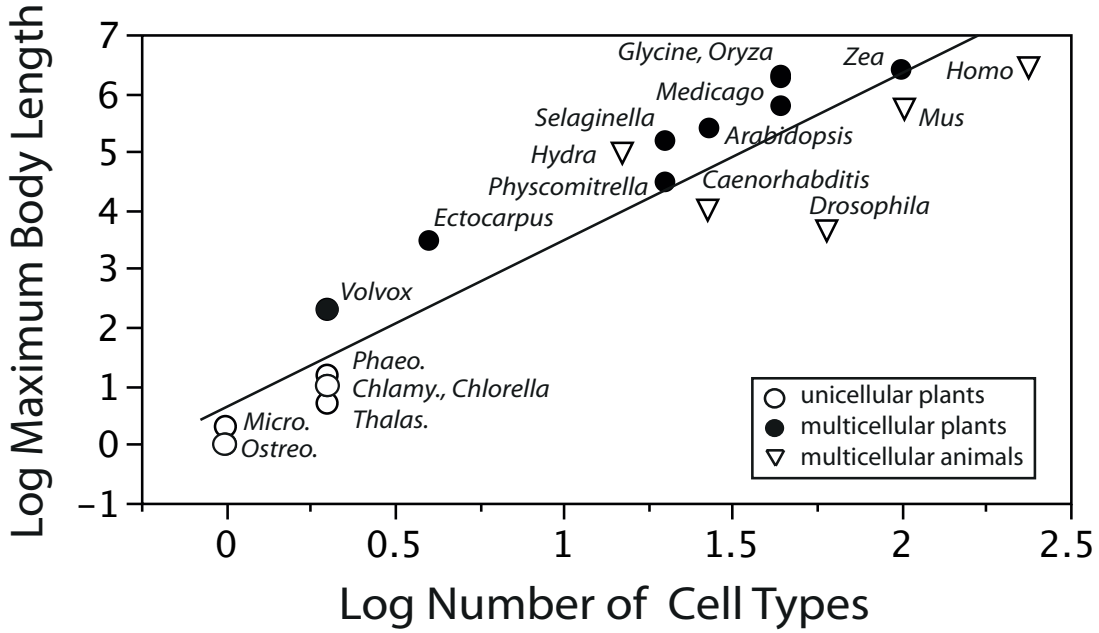

Fig. 6 Log_{10}-transformed data for maximum body length (original units in μm) plotted against $\log_{10}$-transformed data for number of different cell types (NCT) reported for unicellular and multicellular algae, land plants, invertebrates, and vertebrates. Straight lines are ordinary least squares log-log regression curves. *Chlamy.* – *Chlamydomonas reinhardii*; *Micro.* – *Micromonas pusilla* NOUM 17; *Ostreo.* – *Ostreococcus tauri*; *Phaeo.* – *Phaeodactylum tricornutum*; *Thalas.* – *Thalassiosira pseudonana*. For regression parameters, see Tab. 1. Data for NCT taken from Bell and Mooers [2]. Data for body length taken from Lang et al. [32] (supplementary Tab. 7 therein) and supplemented with data reported on the Internet (e.g., tallest human).

A critical assumption shared by most attempts to explain cell fate specification is that the multi-stable states achieved by gene regulatory networks are largely deterministic, i.e., cell fate specification is a direct and by and large invariant result of each state of multi-stable gene regulatory networks. However, there is sufficient evidence that the joint effects of alternative splicing (AS) and intrinsically disordered protein (IDP) domains on cell type specification and differentiation, and post-translational modification (PTM) are critical to the operations of gene regulatory networks. As noted, AS provides a mechanism that increases protein functional diversity without increasing proteome or genome size, whereas IDP domains further amplify protein functionalities. Further, both AS and IDP are dependent on intra- and extracellular conditions, which makes their operation context-dependent and therefore potentially highly adaptive. These properties engender the hypothesis that AS and IDP operate as a very ancient motif that functions in an interactive and context-dependent manner, and that this motif has been elaborated evolutionarily to yield more diverse cell types in successively divergent eukaryotic lineages.

This hypothesis is consistent with the data presented here and by other workers. For example, Schad et al. [5] observe a significant and positive correlation between organismic complexity (as gauged by the number of different cell types) and proteome size (as gauged by the total number of amino acids), and show that the fraction of intrinsically disordered proteins (IDPs) increases significantly between prokaryotes and eukaryotes. However, these authors could find no evidence for further increases in the fraction of IDPs "over the course of evolution" and concluded that "complexity is … determined by interaction potential, alternative splicing capacity, tissue-specific protein disorder and, above all, proteome size" [5]. These trends are consistent with most, but not all of those reported in Tab. 1, because the number of different cell types produced by the 19 different organisms examined here increases in a log-log monotonic manner as the number of amino acids and the number of intrinsically disordered residues increases. Further, as observed by Schad et al. [5], these trends have scaling exponents that exceed unity, which indicates a disproportionate increase in cell types with increasing proteome size. In contrast to the findings of Schad et al. [5], the trends shown in Fig. 5 indicate that the fraction of IDPs has likely increased "over the course of evolution" from unicellular diatoms and green algae (e.g., *Thalassiosira* and *Chlorella*) to simple multicellular organisms (e.g., *Volvox*, and *Hydra*), and culminating in highly derived plant and animal multicellular life forms (e.g., *Zea* and *Homo*).

The manner by which IDP within the AS–IDP motif can expand cell type diversity is illustrated by a model based on the appearance of five isoforms of a hypothetical regulatory (transcription factor) IDP and a simple Turing reaction-diffusion system (Fig. 7a). Beginning with an undifferentiated (meristematic or stem) cell, each successive cycle of cell division is accompanied by the context-dependent appearance of new isoforms. Cells acquire their "context", which in this case is their position with respect to neighboring cells, by means of the reaction-diffusion system, i.e., cells to the left of each neighbor produce an even-numbered isoform, whereas cells to the right produce an odd-numbered isoform. This scenario does not invoke alternative splicing (which theoretically would expand the repertoire of protein functionality well beyond what is illustrated by this scenario) and depends on only two parameters: (***i***) a chronometer (the cell cycle) that defines when different IDP isoforms are produced, and (***ii***) a spatial discriminator ("left" vs. "right") that defines which different IDP isoforms are produced. Yet, theoretically, even in the absence of AS, within three cell division cycles, eight different cell types are produced as defined by their IDP isoform compositions. Extension of this scenario to eight cell division cycles (and a total of 8 IDP isoforms) results in 258 different cell types, which exceeds the maximum number of cell types reported for any metazoan.

Clearly, this scenario is a naive contrivance if for no other reason than that it presupposes cell type specification depends on the presence or absence of a few protein isoforms, rather than on determinant gene regulatory dynamics. However, it can be expanded easily at the level of an individual cell by considering the synergistic effects of alternative splicing (AS), proteins with intrinsically disordered domains (IDP), and post-translational modifications (PTMs) as shown in Fig. 7b. In this more elaborate model for cell fate specification, a single pre-mRNA undergoes alternative splicing to produce three protein isoforms, each of which has an intrinsically disordered domain that can take on one of three functional variations. Alternatively, each of the three isoforms with disordered domains can be post-translationally modified in one of three ways to yield nine proteins, each of which can assume one of three functional domains. Regardless of the exact numbers, it is clear that a single pre-mRNA can be modified by an AS–IDP–PTMs motif to produce a large number of functionally different proteins.

Concluding remarks

We have tested the supposition that the number of different cell types is a measure of organismic complexity, and we conclude that it is a useful gauge for this purpose, albeit not invariably so. We have also made two assertions based on the relationships between the number of different cell types and proteome size: (***i***) the mechanism for cell fate specification depends on a context-dependent motif consisting of alternative splicing (AS) and intrinsically disordered residues (IDResidues) as well as post-translational modifications (PTMs), and (***ii***) this motif seriously challenges the notion that complex gene regulatory networks regulate development in a deterministic fashion.

Evidence for the operation of the AS–IDR–PTM motif in both plants and metazoans is extensive. As in animals, the pre-mRNAs of model plant organisms, such as *Arabidopsis thaliana*, undergo extensive AS, contain substantial fractions of IDR, and undergo PTM, typically in the form of phosphorylation (e.g., [4,43,44]). Plants and animals use small molecules for inter-cellular signaling that are transduced by binding to disorder-containing receptor proteins, and PONDR VSL2b [45] and other disorder predictors indicate that the various plant homeodomain-containing transcription factors contain significant amounts of disorder, and

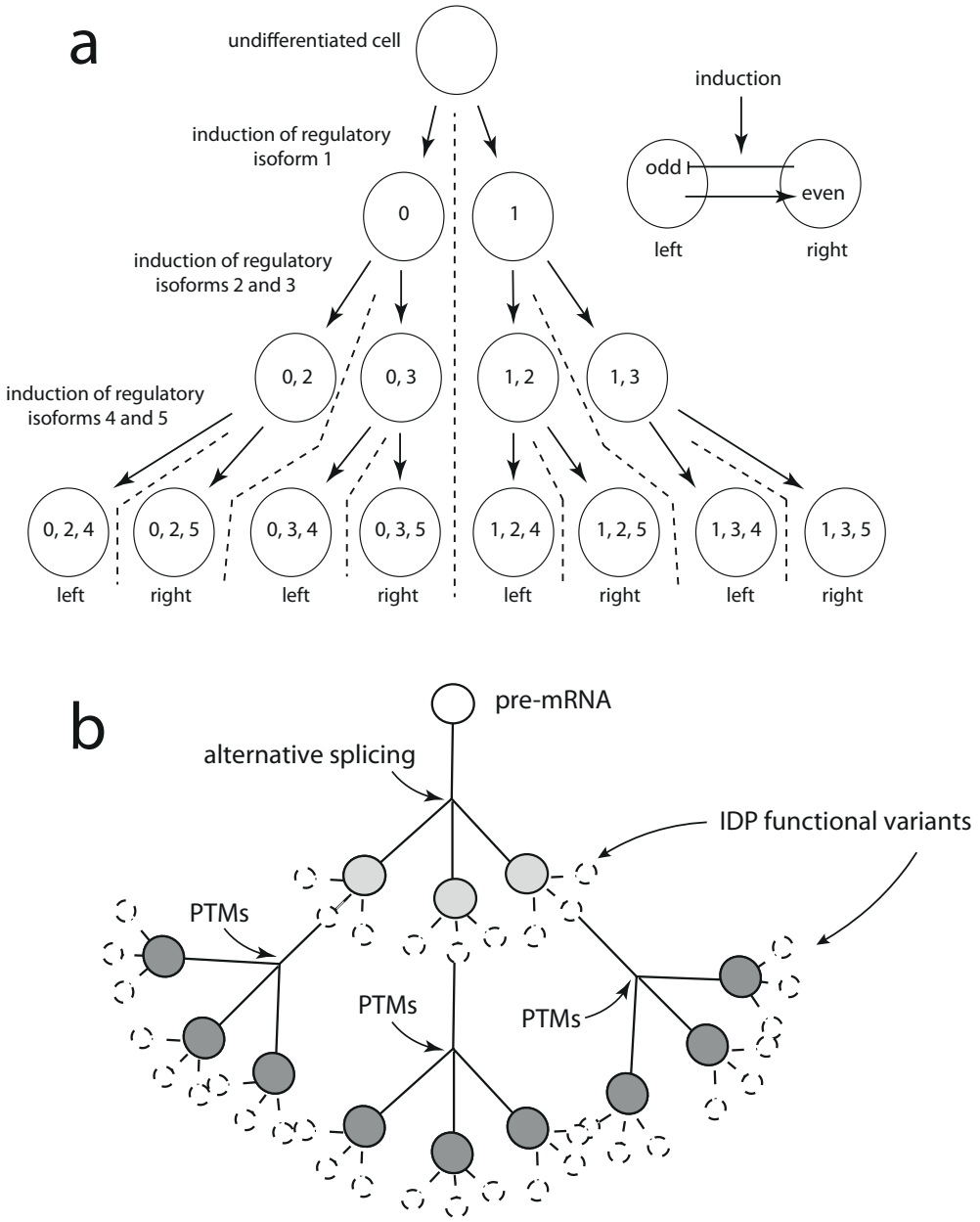

Fig. 7 Scenarios for cell fate specification involving the effects of alternative splicing, intrinsically disordered proteins (IDP), and post-translational modification (PTM). **a** A two-parameter scenario for the specification of eight different cell types using five isoforms (1–5) of an intrinsically disordered regulatory protein (transcription factor). The two parameters are (***i***) a chronometer (using successive cell division cycles) and (***ii***) a reaction-diffusion induction system (upper right diagram) in which paired derivative cells to the right of their neighbor inhibit the formation of odd numbered isoforms and cells to the left of their derivative neighbor activate the formation of even numbered isoforms. Each of three successive divisions (top to bottom) produces successively different isoform compositions, each of which specifies a different cell fate. **b** A single precursor mRNA (pre-mRNA) undergoes alternative splicing to yield three protein isoforms (shown in light gray), each of which has an intrinsically disordered domain that can take on three functions (circles with dashed outlines). Alternatively, each isoform is subjected to PTMs in one of three ways to yield nine proteins (shown in darker gray), each of which has an intrinsically disordered domain that can take on three functions (circles with dashed outlines), to yield a total of 27 IDP functional variants.

they undergo PTM [46]. For example, rice KNOX undergoes tissue-specific AS [47], whereas DELLA proteins are phosphorylated for degradation [48].

The operation of this motif casts doubt on the notion that cell fate specification is genetically deterministic. Across diverse unicellular and multicellular plants and animals, the number of different cell types does not increase one-to-one with increasing genome size (the G paradox), but it does increase (and in a log-log disproportionate manner) with increasing proteome size, particularly with increases in the faction of intrinsically disordered residues, which is a reflection of protein functional versatility. We are sensitive to the fact that the number of species examined in this study is small, and that even very significant and positive correlations between variables of interest do not provide proof of cause and effect. Yet, the majority of transcription

factors contain intrinsically disordered protein domains whose functionalities are context-dependent [29,49], and post-translational modification further diversifies protein functionalities. These features, taken even in isolation, challenge the conventional perspective on the deterministic control of gene regulatory networks. They also help us to understand how evolutionarily conserved transcription factors can evolve new functionalities because the AS–IDR–PTM motif can buffer and even isolate a genome from the immediate consequences of initially maladaptive gene mutations and provide a genome sufficient time to evolve adaptively as transcription factors assume new roles. The origin of cellular differentiation may reside therefore in the ability of the AS–IDP–PTM motif to permit somatic or reproductive functional roles for different cell types to become established ad hoc by natural selection. This speculation is consistent with the hypothesis that the evolution of complex genomic architecture was driven by non-adaptive stochastic events, rather than by adaptive evolution by means of natural selection [50], and it is consistent with the observation that metazoan embryonic cell lineages are significantly simpler than would be expected by chance, which suggests that selection for decreased complexity plays a significant role in shaping cell lineages [51]. Under any circumstances, our perspective indicates that sequence homologies do not invariably indicate protein functional homologies, which makes BLAST analyses potentially misleading.

Acknowledgments

The authors thank Dr. Beata Zagorska-Marek for inviting this submission, and Drs. Dominick Paolillo Jr. and Randy Wayne (Cornell University), and two anonymous reviewers for constructive suggestions. Funding from the College of Agriculture and Life Science is also gratefully acknowledged.

Authors' contributions

The following declarations about authors' contributions to the research have been made: conceived of and wrote the paper: KJN, EDC, AKD; prepared the figures: KJN.

Competing interests

No competing interests have been declared.

References

1. Valentine JW, Collins AG, Meyer CP. Morphological complexity increase in metazoans. Paleobiology. 1994;20(2):131–142.
2. Bell G. Size and complexity among multicellular organisms. Biol J Linn Soc. 1997;60(3):345–363. http://dx.doi.org/10.1006/bijl.1996.0108
3. Erwin DH. Early origin of the bilaterian developmental toolkit. Phil Trans R Soc B. 2009;364(1527):2253–2261. http://dx.doi.org/10.1098/rstb.2009.0038
4. Chen L, Bush SJ, Tovar-Corona JM, Castillo-Morales A, Urrutia AO. Correcting for differential transcript coverage reveals a strong relationship between alternative splicing and organism complexity. Mol Biol Evol. 2014;31:1402–1413. http://dx.doi.org/10.1093/molbev/msu083
5. Schad E, Tompa P, Hegyi H. The relationship between proteome size, structural disorder and organism complexity. Genome Biol. 2011;12:R120. http://dx.doi.org/10.1186/gb-2011-12-12-r120
6. Liu J, Perumal NB, Oldfield CJ, Su EW, Uversky VN, Dunker AK. Intrinsic disorder in transcription factors. Biochemistry (Mosc). 2006;45(22):6873–6888. http://dx.doi.org/10.1021/bi0602718
7. Oldfield CJ, Meng J, Yang JY, Yang MQ, Uversky VN, Dunker AK. Flexible nets: disorder and induced fit in the associations of p53 and 14-3-3 with their partners. BMC Genomics. 2008;9(1 suppl):S1. http://dx.doi.org/10.1186/1471-2164-9-S1-S1
8. Tompa P, Fuxreiter M. Fuzzy complexes: polymorphism and structural disorder in protein-protein interactions. Trends Biochem Sci. 2008;33(1):2–8. http://dx.doi.org/10.1016/j.tibs.2007.10.003
9. Bonner JT. First signals: the evolution of multicellular development. Princeton, NJ: Princeton University Press; 2000.
10. Kirk DL. A twelve-step program for evolving multicellularity and a division of labor. Bioessays. 2005;27(3):299–310. http://dx.doi.org/10.1002/bies.20197
11. Herron MD, Michod RE. Evolution of complexity in the volvocine algae: transitions in individuality through Darwin's eye. Evolution. 2008;62(2):436–451. http://dx.doi.org/10.1111/j.1558-5646.2007.00304.x
12. Folse HJ, Roughgarden J. What is an individual organism? A multilevel selection perspective. Q Rev Biol. 2010;85(4):447–472.
13. Niklas KJ. The evolutionary-developmental origins of multicellularity. Am J Bot. 2014;101(1):6–25. http://dx.doi.org/10.3732/ajb.1300314
14. Silberfeld T, Leigh JW, Verbruggen H, Cruaud C, de Reviers B, Rousseau F. A multi-locus time-calibrated phylogeny of the brown algae (Heterokonta, Ochrophyta, Phaeophyceae): investigating the evolutionary nature of the "brown algal crown radiation". Mol Phylogenet Evol. 2010;56(2):659–674. http://dx.doi.org/10.1016/j.ympev.2010.04.020
15. Andersen RA. Biology and systematics of heterokont and haptophyte algae. Am J Bot. 2004;91(10):1508–1522. http://dx.doi.org/10.3732/ajb.91.10.1508
16. Graham LE. Algae. 2nd ed. San Francisco, CA: Pearson/Benjamin Cummings; 2009.
17. Parker BC. Translocation in the giant kelp *Macrocystis*. I. Rates, direction, quantity of C^{14}-labeled products and fluorescein. J Phycol. 1965;1(2):41–46. http://dx.doi.org/10.1111/j.1529-8817.1965.tb04554.x
18. Parker BC. Translocation in *Macrocystis*. III. Composition of sieve tube exudate and identification of the major C^{14}-labeled products. J Phycol. 1966;2(1):38–41. http://dx.doi.org/10.1111/j.1529-8817.1966.tb04590.x
19. Buggeln RG, Fensom DS, Emerson CJ. Translocation of ^{14}C-photoassimilate in the blade of *Macrocystis pyrifera* (Phaeophyceae). J Phycol. 1985;21(1):35–40. http://dx.doi.org/10.1111/j.0022-3646.1985.00035.x
20. Ruhfel BR, Gitzendanner MA, Soltis PS, Soltis DE, Burleigh JG. From algae to angiosperms-inferring the phylogeny of green plants (*Viridiplantae*) from 360 plastid genomes. BMC Evol Biol. 2014;14:23. http://dx.doi.org/10.1186/1471-2148-14-23
21. Cook M, Graham L, Botha C, Lavin C. Comparative ultrastructure of plasmodesmata of *Chara* and selected bryophytes: toward an elucidation of the evolutionary origin of plant plasmodesmata. Am J Bot. 1997;84(9):1169–1178.
22. Boot KJM, Libbenga KR, Hille SC, Offringa R, van Duijn B. Polar auxin transport: an early invention. J Exp Bot. 2012;63(11):4213–4218. http://dx.doi.org/10.1093/jxb/ers106
23. Foster AS. Comparative morphology of vascular plants. 2nd ed. San Francisco, CA: W.H. Freeman; 1974.
24. Govindarajalu E. The systematic anatomy of south Indian Cyperaceae. Bot J Linn Soc. 1969;62:27–40.
25. Adoutte A, Balavoine G, Lartillot N, Lespinet O, Prud'homme B, de Rosa R. The new animal phylogeny: reliability and implications. Proc Natl Acad Sci USA. 2000;97(9):4453–4456. http://dx.doi.org/10.1073/pnas.97.9.4453
26. Chang CY, Lin WD, Tu SL. Genome-wide analysis of heat-sensitive alternative splicing in *Physcomitrella patens*. Plant Physiol. 2014;165:826–840. http://dx.doi.org/10.1104/pp.113.230540
27. Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nat Genet. 2008;40(12):1413–1415. http://dx.doi.org/10.1038/ng.259
28. Johnson JM, Castle J, Garrett-Engele P, Kan Z, Loerch PM, Armour CD, et al. Genome-wide survey of human alternative pre-mRNA splicing

with exon junction microarrays. Science. 2003;302(5653):2141–2144. http://dx.doi.org/10.1126/science.1090100

29. Liu J, Perumal NB, Oldfield CJ, Su EW, Uversky VN, Dunker AK. Intrinsic disorder in transcription factors. Biochemistry (Mosc). 2006;45(22):6873–6888. http://dx.doi.org/10.1021/bi0602718
30. Niklas KJ. Plant allometry: the scaling of form and process. Chicago, IL: University of Chicago Press; 1994.
31. Hahn MW, Wray GA. The g-value paradox. Evol Dev. 2002;4(2):73–75. http://dx.doi.org/10.1046/j.1525-142X.2002.01069.x
32. Lang D, Weiche B, Timmerhaus G, Richardt S, Riaño-Pachón DM, Corrêa LGG, et al. Genome-wide phylogenetic comparative analysis of plant transcriptional regulation: a timeline of loss, gain, expansion, and correlation with complexity. Genome Biol Evol. 2010;2:488–503. http://dx.doi.org/10.1093/gbe/evq032
33. Niklas KJ, Cobb ED, Crawford DR. The evo-devo of multinucleate cells, tissues, and organisms, and an alternative route to multicellularity. Evol Dev. 2013;15(6):466–474. http://dx.doi.org/10.1111/ede.12055
34. Niklas KJ, Newman SA. The origins of multicellular organisms. Evol Dev. 2013;15(1):41–52. http://dx.doi.org/10.1111/ede.12013
35. Buss LW. The evolution of individuality. Princeton, NJ: Princeton University Press; 1987.
36. Michod RE. Evolution of the individual. Am Nat. 1997;150 suppl 1:S5–S21. http://dx.doi.org/10.1086/286047
37. Schlichting CD. Origins of differentiation via phenotypic plasticity. Evol Dev. 2003;5(1):98–105.
38. Knoll AH. The multiple origins of complex multicellularity. Annu Rev Earth Planet Sci. 2011;39(1):217–239. http://dx.doi.org/10.1146/annurev.earth.031208.100209
39. Britten RJ, Davidson EH. Gene regulation for higher cells: a theory. Science. 1969;165(3891):349–357.
40. Laurent M, Kellershohn N. Multistability: a major means of differentiation and evolution in biological systems. Trends Biochem Sci. 1999;24(11):418–422.
41. Jaeger J, Monk N. Bioattractors: dynamical systems theory and the evolution of regulatory processes. J Physiol. 2014;592(11):2267–2281. http://dx.doi.org/10.1113/jphysiol.2014.272385
42. Ispolatov I, Ackermann M, Doebeli M. Division of labour and the evolution of multicellularity. Proc R Soc B. 2012;274:1768–1776. http://dx.doi.org/10.1098/rspb.2011.1999
43. Yao Q, Gao J, Bollinger C, Thelen JJ, Xu D. Predicting and analyzing protein phosphorylation sites in plants using musite. Front Plant Sci. 2012;3:186. http://dx.doi.org/10.3389/fpls.2012.00186
44. Yruela I, Contreras-Moreira B. Protein disorder in plants: a view from the chloroplast. BMC Plant Biol. 2012;12(1):165. http://dx.doi.org/10.1186/1471-2229-12-165
45. Peng K, Radivojac P, Vucetic S, Dunker AK, Obradovic Z. Length-dependent prediction of protein intrinsic disorder. BMC Bioinformatics. 2006;7(1):208. http://dx.doi.org/10.1186/1471-2105-7-208
46. Oates ME, Romero P, Ishida T, Ghalwash M, Mizianty MJ, Xue B, et al. D2P2: database of disordered protein predictions. Nucl Acids Res. 2012;41:D508–516. http://dx.doi.org/10.1093/nar/gks1226
47. Ito Y, Hirochika H, Kurata N. Organ-specific alternative transcripts of KNOX family class 2 homeobox genes of rice. Gene. 2002;288(1–2):41–47.
48. Qin Q, Wang W, Guo X, Yue J, Huang Y, Xu X, et al. *Arabidopsis* DELLA protein degradation is controlled by a type-one protein phosphatase, TOPP4. PLoS Genet. 2014;10(7):e1004464. http://dx.doi.org/10.1371/journal.pgen.1004464
49. Minezaki Y, Homma K, Nishikawa K. Genome-wide survey of transcription factors in prokaryotes reveals many bacteria-specific families not found in archaea. DNA Res. 2005;12(5):269–280. http://dx.doi.org/10.1093/dnares/dsi016
50. Lynch M. The origins of eukaryotic gene structure. Mol Biol Evol. 2006;23(2):450–468. http://dx.doi.org/10.1093/molbev/msj050
51. Azevedo RBR, Lohaus R, Braun V, Gumbel M, Umamaheshwar M, Agapow PM, et al. The simplicity of metazoan cell lineages. Nature. 2005;433(7022):152–156. http://dx.doi.org/10.1038/nature03178
52. Peng ZL, Kurgan L. Comprehensive comparative assessment of in-silico predictors of disordered regions. Curr Protein Pept Sci. 2012;13(1):6–18.