Engineering 57 (2026) 87–99

Contents lists available at ScienceDirect

Engineering

journal homepage: www.elsevier.com/locate/eng

Research
Glycomedicine—Article

Contrasting Macroevolutionary Patterns in the Human N-Glycosylation Pathway

Domagoj Kifer [a], Nina Čorak [b], Mirjana Domazet-Lošo [c], Niko Kasalo [b], Gordan Lauc [a,d], Göran Klobučar [e,*], Tomislav Domazet-Lošo [b,f,*]

[a] Faculty of Pharmacy and Biochemistry, University of Zagreb, Zagreb 10000, Croatia
[b] Laboratory of Evolutionary Genetics, Division of Molecular Biology, Ruđer Bošković Institute, Zagreb 10000, Croatia
[c] Department of Applied Computing, Faculty of Electrical Engineering and Computing, University of Zagreb, Zagreb 10000, Croatia
[d] Genos Glycoscience Research Laboratory, Zagreb 10000, Croatia
[e] Department of Biology, Faculty of Science, Division of Zoology, University of Zagreb, Zagreb 10000, Croatia
[f] School of Medicine, Catholic University of Croatia, Zagreb 10000, Croatia

ARTICLE INFO

Article history:
Received 1 November 2024
Revised 23 May 2025
Accepted 29 June 2025
Available online 15 July 2025

Keywords:
N-glycosylation
Phylostratigraphy
Evolution
Endoplasmic reticulum
Golgi

ABSTRACT

Building on coding mutations and splicing variants, post-translational modifications add a final layer to protein diversity that operates at developmental and physiological timescales. Although protein glycosylation is one of the most common post-translational modifications, its evolutionary origin remains largely unexplored. Here, we performed a phylostratigraphic tracking of glycosylation machinery (GM) genes and their targets—glycoproteins (GPs)—in a broad phylogenetic context. Our results show that the vast majority of human GM genes trace back to two evolutionary periods: the origin of all cellular organisms and the origin of all eukaryotes. This indicates that protein glycosylation is an ancient process likely common to all life, further elaborated in early eukaryotes. In contrast, human glycoproteins exhibited prominent enrichment signals in more recent evolutionary periods, suggesting an important role in the transition from metazoans to vertebrates. Focusing specifically on the N-glycosylation (NG) pathway, we noted that the majority of NG genes acting on the cytoplasmic side of the endoplasmic reticulum (ER) trace back to the origin of cellular organisms. This sharply contrasts with the rest of the NG pathway, which is oriented toward the ER lumen, where genes of eukaryotic origin predominate. In the Golgi, we also identified an analogous binary evolutionary origin of GM genes. We discuss these findings in the context of the evolutionary emergence of the eukaryotic endomembrane system and propose that the ER evolved through the invagination of a prokaryotic cell membrane containing an NG pathway.

1. Introduction

Glycosylation is the enzymatic process that transfers an oligosaccharide (glycan) to an aglycon; that is, protein, lipid, or RNA [1]. Interestingly, a recent work also considers the possibility of DNA glycosylation [2]. In *Homo sapiens* (*H. sapiens*), glycans are synthesized and attached to aglycons through one of the 16 glycosylation pathways consisting of about 700 genes encoding enzymes transporters, and other proteins found in cellular glycosylation machinery (GM) [3–6]. Protein glycosylation is a complex post-translational modification occurring almost completely in the endoplasmic reticulum (ER) and the Golgi [7], where glycans are attached to proteins in a series of reactions catalyzed by glycotransferases and glycoside hydrolases [8]. It has long been known that glycans play major metabolic, structural, and physiological roles in biological systems, and that more than half of all human proteins are glycosylated [9–11].

Protein glycosylation can be found in all three domains of life [12,13]. Compared to bacteria and archaea, eukaryotes use a narrower spectrum of monosaccharide building blocks in glycan synthesis [14,15]. However, eukaryotic glycans, despite the reduced panel of monosaccharide units, show highly complex linear and branching structural diversity [10,15–17]. Protein glycosylation can be viewed as a polygenic trait, where glycosylation phenotypes

* Corresponding authors.
E-mail addresses: goran.klobucar@biol.pmf.hr (G. Klobučar), tdomazet@irb.hr (T. Domazet-Lošo).

https://doi.org/10.1016/j.eng.2025.06.039

result from the actions of many genes along the glycosylation pathways, partly modulated by intracellular and extracellular environments [16,18]. The studies of protein-bound glycan diversity in different evolutionary lineages showed remarkably discontinuous patterns that are shaped by various evolutionary forces, including coevolutionary interactions within a specific holobiont [16–21]. However, despite the biological importance and ubiquitous presence of glycosylation on the tree of life [12,14,19], our understanding of the evolutionary dynamics of GM genes is still cursory [22–27].

For instance, N-glycosylation (NG) biosynthesis is the most frequent and the most studied type of protein glycosylation in eukaryotes [9,23]. It starts at the cytoplasmic side of the ER membrane where monosaccharides are sequentially added by GM proteins to the activated lipid carrier until a glycan with $Man_5GlcNAc_2$ structure is formed [8]. This glycan structure is then translocated by a flippase to the luminal side of the ER membrane, where it is further elongated by the luminal GM to the $Glc_3Man_9GlcNAc_2$ structure, which is common to most eukaryotes [19,28]. Further glycan editing continues in the Golgi where, under the control of Golgi-specific GM proteins (glycosyltransferases and glycosidases), their final structure is formed [8]. In eukaryotes, this final glycan structure varies within populations and between species [12,14,17,19,27].

Eukaryotic, bacterial, and archaeal NG pathways share a common topology where the first part of stepwise oligosaccharide synthesis unfolds on the cytosolic side of a membrane. At some point, this process includes flipping of the nascent oligosaccharide across the membrane [12,23,29,30]. In addition, NG pathways show chemical resemblance that include the use of polyisoprenol lipid carriers and phosphate-linked substrates [22,29,31]. These topological and chemical similarities suggest that the last universal common ancestor (LUCA) possessed an NG like pathway [22]. The remarkable difference, however, is that in bacteria and archaea, the NG pathway is embedded in the plasma membrane, whereas in eukaryotes, it is part of the ER membrane [12,19,23,29,30,32].

On the other hand, the evolutionary origin of individual GM genes acting in the eukaryotic NG pathway is not yet fully resolved [23]. Because of the symbiotic origin of eukaryotes with the contribution of both archaeal and bacterial ancestors, it is not fully clear which prokaryotic group is ancestral to the eukaryotic NG pathway [33–35]. Interestingly, recent phylogenomic analysis suggests a mixed origin of eukaryotic NG with the contribution of both archaea and bacteria [23]. Nevertheless, the evolutionary origin of many eukaryotic NG genes remains unresolved and some of them seem to be a eukaryotic innovation [23].

Another facet of the glycosylation process involves proteins that are targeted by the GM. At least in humans, it is clear that many proteins are glycosylated, however evolutionary dynamics of these glycoproteins (GPs) have not been addressed so far. To obtain a coherent view of the evolutionary origin of GM and GP genes, and consequently to gain a better understanding of the role of glycosylation in the evolution of metazoans, particularly humans, we applied here a phylostratigraphic approach [36–48]. Although the phylostratigraphic approach has proven to be a powerful tool for addressing various macroevolutionary questions—including those related to human diseases [36,42,47,48]—it has never been applied to problems related to glycobiology.

Our results showed that GM and GP genes have divergent macroevolutionary dynamics and that the intracellular localization of NG proteins on the ER membrane follows an evolutionarily polarized pattern. This finding led us to propose that the ER evolved through the invagination of the prokaryotic plasma membrane. Together, our study reveals that eukaryotic glycosylation is built upon a fundamental core that is largely homologous to prokaryotic glycosylation genes, along with a set of eukaryota-specific genes that are unique to the glycosylation process in eukaryotic cells.

2. Methods

2.1. GM genes and GPs

We compiled a list of GM genes from several sources: genes listed in the study from Moremen et al. [4] (refer to Data S1 in Appendix A), Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways related to glycosylation in *H. sapiens* (hsa00510, hsa00511, hsa00512, hsa00514, hsa00515, hsa00531, hsa00532, hsa00533, hsa00534, hsa00563, hsa00601, hsa00603, and hsa00604) [49], *H. sapiens* entries in Carbohydrate Active enZYmes (CAZY) database [3], genes listed in the study from Varki et al. [6] and from Schjoldager et al. [5] (Data S1). We also retrieved *H. sapiens* orthologues of mouse glycosylation genes listed in Moremen et al. [4]. The obtained dataset was manually reviewed for duplicates and pseudogenes which finally yielded a total of 673 GM genes with unique ensemble gene IDs (Data S1).

In addition, we collected a list of GPs from UniProtKB/Swiss-Prot database [50] using the following query: "annotation:(type:carbohyd) AND reviewed:yes AND organism: "Homo sapiens (Human) [9606]"." In this way, we obtained all the reviewed *H. sapiens* proteins that show some evidence to possess a posttranslational modification of glycosylation type. The final GP dataset included a total of 4565 genes with unique ensemble gene IDs (Data S1).

2.2. Consensus phylogeny and reference genomes

We constructed a consensus tree of 503 organisms representing the diversity of major lineages that lead from the origin of cellular organisms to *H. sapiens*, resulting in 29 phylogenetic levels (phylostrata (ps), Fig. 1, Data S2 and S3 in Appendix A). For phylogenetic relationships between taxa, we followed the relevant literature [51–58]. Reference genomes (that is, corresponding amino acid sequences (proteomes) of all organisms included in the phylogeny) were downloaded mostly from the Ensembl and National Center for Biotechnology Information (NCBI) databases. Proteomes of choanoflagellates were retrieved from Richter et al. [59]. Prior to the phylostratigraphic analysis, the proteomes were processed to keep only the longest splicing variant of each gene. Other details on the construction of a phylostratigraphic database are described in Domazet-Lošo et al. [37].

2.3. Phylostratigraphic analysis

The theoretical background and phylostratigraphic procedure have been described previously [37,39,40,47]. Briefly, the longest splicing variant of every *H. sapiens* gene was compared to the reference database using the blastp algorithm with the *e*-value cutoff equal to 10^{-3}, which has repeatedly been shown to be optimal in phylostratigraphic analysis [37,39,43,60–63]. The obtained sequence similarity search results, along with the consensus phylogeny, were used to estimate the evolutionary origin (phylostratum; ps) of all human genes. We further tested the robustness of this cutoff by conducting a sliding *e*-value test, which confirmed the stability of the recovered patterns [37,43]. It is important to note that phylostrata represent key evolutionary events along the lineage leading to the focal species, reflecting the tree topology and thereby providing a broad overview of gene dynamics over geological time—that is, a macroevolutionary perspective [37,48].

The frequency distributions of studied genes (GM and GP) were compared to the frequency distribution of all human genes across

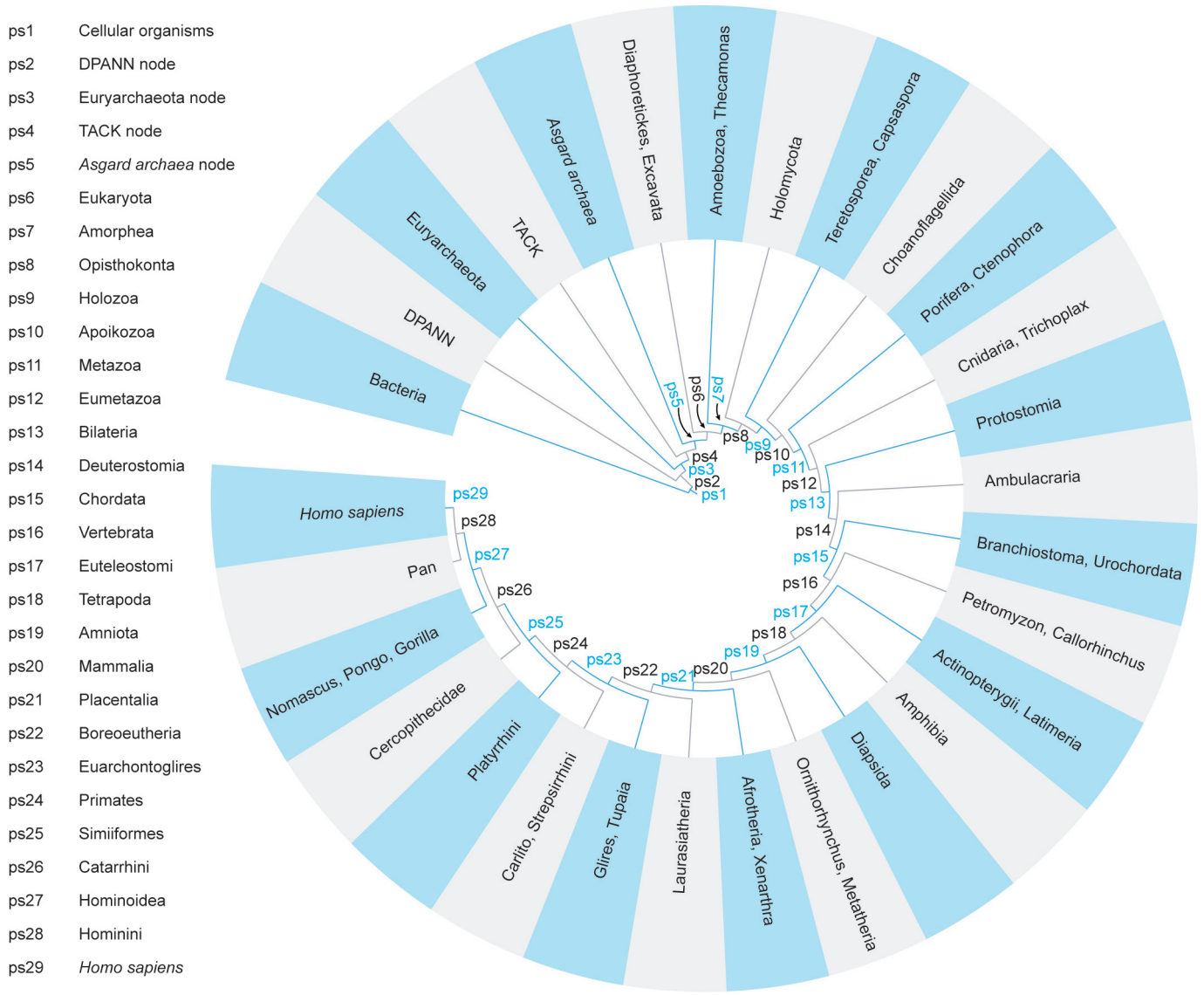

Fig. 1. The consensus phylogeny used in the phylostratigraphic analysis. The condensed consensus tree covers divergence from the last common ancestor of cellular organisms to *H. sapiens* as a focal organism. For a fully resolved tree see Data S2 in Appendix A. The tree is constructed by considering the importance of evolutionary transitions and availability of reference genomes. The internodes (29 ps) that lead from the root of the tree to the focal species (*H. sapiens*) are marked by ps1–ps29. The numbers of *H. sapiens* genes traced to each phylostratum and corresponding percentages (in parentheses) are given after phylostrata names. We mapped in total 23 237 *H. sapiens* genes. DPANN: phylum of archaea named by first five groups discovered—Diapherotrites, Parvarchaeota, Aenigmarchaeota, Nanoarchaeota, and Nanohaloarchaeota; TACK: phylum of proteoarchaeota named by first four groups discovered Thaumarchaeota (now Nitrososphaerota), Aigarchaeota, Crenarchaeota (now Thermoproteota), and Korarchaeota (now Thermoproteota).

phylostrata (expected distribution). In all graphical representations the ratio of these frequencies is shown as odds ratios. Deviations from the expected frequencies at each phylostratum were tested using a two-tailed hypergeometric test. The obtained *p*-values were corrected for multiple testing using the Benjamini–Hochberg method. To account for a potential sequence similarity search bias, we applied a sliding *e*-value protocol [43] where we repeated sequence similarity searches and all downstream calculations using a range of *e*-value cut-offs (1, 10^{-1}, 10^{-2}, 10^{-5}, 10^{-10}, 10^{-15}, and 10^{-20}).

2.4. Statistical analyses

To test whether there is an association between the evolutionary origin of genes coding for enzymes of the NG pathway acting on the ER and membrane topology, we generated a contingency table with these categories and tested observed biases using Fisher's exact test. We excluded requiring fifty three 1 homolog (RFT1) from the statistical analyses due to its ambiguous topology on the ER membrane (lumen vs cytosol). We used the same approach in the Golgi to test the association between the evolutionary origin and enzyme type.

To assess if there is an association between the order of enzymes of the NG pathway in the ER and their evolutionary origin, we devised the homogeneity index (HI). For an ordered list of *N* genes with assigned evolutionary origin (phylostratum), HI is calculated by iterating from the 2nd to the *N*th element of the list and comparing every element with the one immediately preceding it. If the evolutionary origin of the two elements is not equal, the index is increased by 1. In essence, a high HI means that the sequence of enzymes is disordered with respect to their evolutionary origin, while a sequence in which the enzymes of the same evolutionary origin are clustered together will result in a low HI.

To test whether the observed biased grouping of enzymes of the NG pathway in the ER is statistically significant, we generated all possible permutations of evolutionary origins and calculated the HI for each permutation, analogous to the procedure utilized in earlier publications [64,65]. We arranged the calculated HI values into bins containing identical values. We divided the number of elements in each bin with the total number of permutations, resulting in the empirical mass density function of HI values. Finally, we calculated the *p*-value by summing the probabilities of bins containing the HI values equal to or lower than the observed one. All statistical analyses were performed using Python and R [66]. Figures were created using R packages "ggplot2" and "cowplot" [67,68].

3. Results

To reconstruct the evolutionary origin of *H. sapiens* GM and GP genes, we first mapped all human protein sequences (23 237) on the consensus phylogeny consisting of 29 phylogenetic levels (ps, Fig. 1, Data S2 and S3) [37–39,41,43].

To get an overview of macroevolutionary patterns related to glycobiology, we made a compilation of 673 human GM genes. The majority of these genes (56%) map to the origin of all cellular organisms (ps1, Fig. 2), which is significantly more than expected by chance (odds ratio = 2.71, $p = 1.29 \times 10^{-37}$). The ps1 (cellular organisms) corresponds to the LUCA, thus all genes/proteins shared between bacteria and archaea are classified in this phylostratum. It is important to note that the focal linage (*H. sapiens*) determines which phylostrata are on the evolutionary path (trajectory) to the root of the tree. In our case, given that *H. sapiens* is a focal species, all bacteria are integral part of ps1 (cellular organisms).

The origin of cellular organisms (ps1, Fig. 2) is the only evolutionary period in which we detected the enrichment of GM genes indicating that glycosylation is an ancient process common to all life. However, we detected also a substantial number of GM genes at the origin of eukaryota (24%; ps6, Fig. 2) suggesting that in this evolutionary transition glycosylation pathways were further elaborated. Another 17% of GM genes we traced back to the period between the origin of Amorphea and Bilateria (ps7–ps13, Fig. 2). These numbers reveal that the origin of animal multicellularity as well as the body plan assembly of eumetazoans and bilaterians required new additions to glycosylation pathways. Notably, after the origin of amniotes (ps19, Fig. 2) no new GM genes were mapped on our phylostratigraphic map. Taken together, our phylostratigraphic profile of human GM genes suggests that glycosylation is an evolutionarily ancient process, fully established before the radiation of mammals (ps20, Fig. 2).

The reported results are based on the blastp *e*-value cutoff of 10^{-3} which has repeatedly been shown to be optimal in phylostratigraphic analysis [37,39,43,60–62]. However, to test the stability of the observed phylostratigraphic patterns we recently introduced a test, where the analysis is repeated for a broad range of *e*-value cutoffs; for example, between 1 and 10^{-20} [43]. This sliding *e*-value protocol intentionally inflates false positive (*e*-values closer to 1) and false negative rates (*e*-values closer to 10^{-20}) and

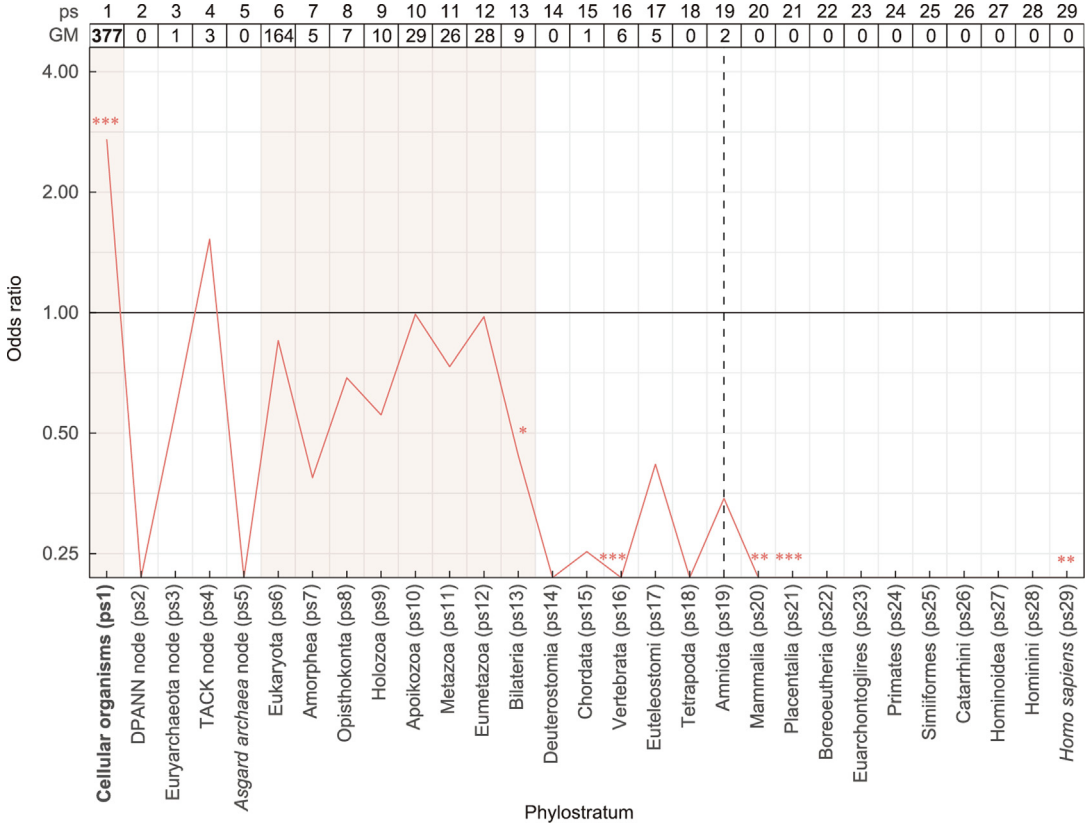

Fig. 2. Phylostratigraphic analysis of *H. sapiens* GM genes. The *x*-axis shows 29 phylogenetic levels (ps) of the consensus phylogeny with *H. sapiens* as a focal species (Fig. 1). The table at top shows distributions of 673 human GM genes along phylostrata. The numbers in bold mark phylostrata with a substantial number of GM genes. The odds ratio, shown on a log-scaled *y*-axis, indicates deviation from the expected frequency. We tested the significance of GM gene enrichments and depletions using a two-tailed hypergeometric test (*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$). GM genes are only significantly enriched in ps1. However, the evolutionary origin of most GM genes could be traced to cellular organisms (ps1), eukaryota (ps6), and the period between Amorphea and Bilateria (ps7–ps13). These periods are shaded in red. The dashed vertical line marks the origin of amniotes (ps19). After this phylostratum, no new GM genes emerged (ps20–ps29).

thus tests the robustness of the initially observed phylostratigraphic pattern at 10^{-3} *e*-value cutoff [43]. Our sliding *e*-value analysis confirmed the stability of the signal, which we initially found in ps1, in the full range of tested cutoff values (Fig. S1 in Appendix A). This result reassured us that the observed pattern at the 10^{-3} *e*-value reflects a genuine evolutionary imprint, rather than noise from the error rates of the blastp algorithm.

To get a more detailed insight into the evolutionary patterns of GM genes, we divided them into seven subgroups, which reflect their specific roles in glycobiology [50,69]. We then analyzed these subgroups independently using phylostratigraphic procedure (Fig. 3). The monosaccharide metabolism (MM) genes that contribute to glycosylation process are prevailingly mapped to the evolutionary oldest phylostrata (ps1) where they contribute to a strong enrichment signal (Fig. 3). This pattern suggests that the basic sugar metabolism necessary for glycosylation (i.e., monosaccharide activation) was present at the origin of cellular organisms. Comparably, genes that contribute to the NG and O-glycosylation (OG) of proteins show a bimodal pattern with significant enrichments at the origin of cellular organisms (ps1, Fig. 3) and eukaryota (ps6, Fig. 3). This enrichment profile suggests that protein glycosylation was present at the origin of cellular organisms (ps1), but also points to further innovations linked to these pathways at the origin of eukaryota (ps6).

Genes involved in the synthesis of glycosaminoglycans (GAGs) and glycan-binding proteins (GBPs) produced very similar phylostratigraphic patterns (Fig. 3). Both groups of GM genes showed enrichment signals at the origin of cellular organisms (ps1, Fig. 3) and at the origin of eumetazoans (ps12, Fig. 3). The first signal at the origin of cellular organisms (ps1) suggests that the synthesis of GAGs and the production of GBPs have deep roots in evolutionary history. However, the second signal at the origin of eumetazoans (ps12) points that novel GAGs and GBPs played an important role in the emergence of the first metazoans. This is not surprising because GAGs are the main component of the extracellular matrix and endothelial glycocalyx layer—the crucial elements of true tissues that determine their physical characteristics [70,71]. Similarly, GBPs play an important role in

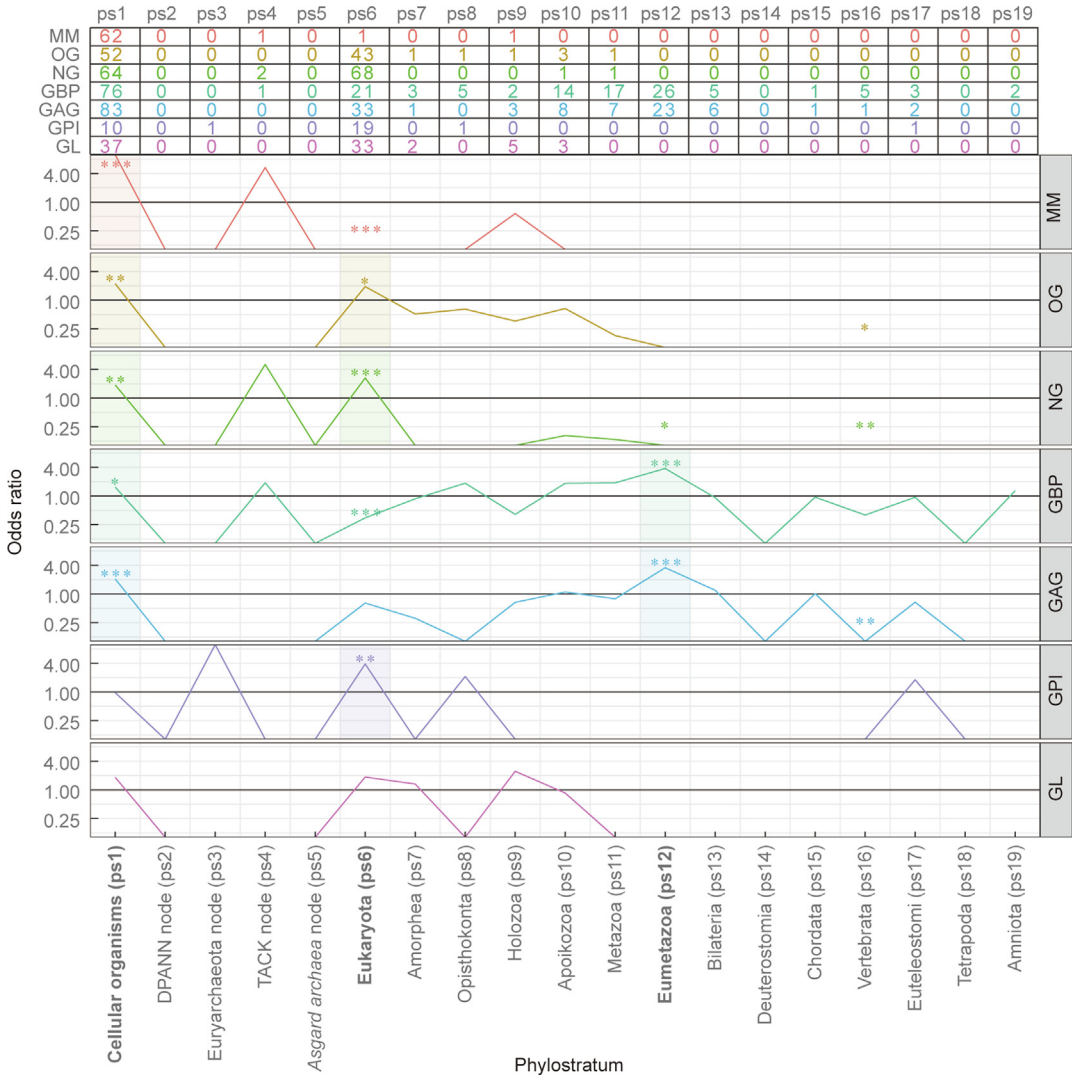

Fig. 3. Phylostratigraphic maps of GM genes per functional groups. The table at the top shows the distributions along 19 ps of *H. sapiens* GM genes divided into seven specific glycosylation roles. In contrast to Fig. 2, we here depicted only ps1 to ps19 because there are no new GM genes after the origin of amniotes (ps19, Fig. 2). The odds ratio, shown on a log-scaled *y*-axis, indicates deviation from the expected frequency for every subgroup independently. We tested the significance of the enrichments and depletions using a two-tailed hypergeometric test (*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$). The shaded areas mark phylostrata with significant enrichments. MM: monosaccharide metabolism; OG: O-glycosylation; GBP: glycan-binding protein; GAG: glycosaminoglycan; GPI: glycosylphosphatidylinositol; GL: glycolipid.

cell adhesion, cell–cell interactions, cell–matrix interactions, and immune processes via self-nonself recognition; all of which are important functions of animal lifestyle [72].

The glycosylphosphatidylinositol (GPI) anchor is a glycan-based posttranslational modification that allows modified proteins to be attached to the outer surface of the plasma membrane [73–75]. The GM involved in the formation of GPI anchored proteins showed a strong enrichment signal at the origin of eukaryota (ps6, Fig. 3). This signal, in line with the observation that GPI modifications are found only in eukaryotes [76,77], suggests that GPI anchor is a eukaryogenesis-linked innovation that facilitates protein trafficking within highly compartmentalized eukaryotic cells. In contrast, GM genes involved in glycolipid (GL) formation did not show any statistically significant enrichment (Fig. 3). Regardless of this, the vast majority of GL genes, similar to the other subgroups of GM genes, map to the origin of cellular organisms (ps1) or eukaryota (ps6, Fig. 3). Finally, all enrichment profiles that we recovered from the subgroups of GM genes showed stability in sliding *e*-value analyses (Fig. S2 in Appendix A).

To compare evolutionary patterns of GM genes with their target proteins we mapped 4565 human genes coding for GPs onto the consensus phylogeny (Fig. 4). Around 38% of GPs could be traced back to the ancestor of cellular organisms (ps1), which is significantly more than expected by chance (odds ratio = 1.30, $p < 0.001$). Apart from that, we detected significant enrichment signals in the evolutionary period that spans the origin of animals (Apoikozoa, ps10, Fig. 4) and bony vertebrates (Euteleostomi, ps17, Fig. 4). Approximately 37% of GPs have their origin in this period with statistically significant odds ratio in the range between 1.26 and 2.32. These results corroborate the notion that glycosylation is an ancient process common to all life, because we found the overlapping enrichment of both types of genes, that is, GM genes as well as GP genes, at the origin of cellular organisms (ps1, Figs. 2 and 4). In addition, the strong enrichment signals of GPs in the span between the unicellular ancestors of animals (ps10, Fig. 4) to bony vertebrates (ps17, Fig. 4) suggests that the increased use of GPs played an important role in the origin of the first animals and in their further radiation along the Cambrian and Ordovician geological periods [21], which is likely connected with the immense energetic shift at the origin of animals [64]. The sliding *e*-value analyses confirmed the robustness of these results (Fig. S3 in Appendix A).

The vast majority of human GPs (97%) in our dataset carry N-linked glycans. This sharply contrasts OGs which are much less frequent (7%) post-translational modification. By assuming that these large differences reflect the relative biological importance of these two post-translational modification types, and by considering that NG is the best-studied protein glycosylation process, we focused our further analysis on N-GM. We first depicted the part of the NG pathway that acts in the ER (Fig. 5). This subset includes 40 proteins that act directly in the biosynthetic pathway on the cytoplasmatic and luminal side of the ER membrane as well as those that are indirectly involved by providing activated monosaccharide blocks or are acting as membrane transporters (Fig. 5(a)). The NG pathway starts with dolichol-phosphate biosynthesis or recovery at the cytosolic side of ER (Fig. 5(a)). We traced the evolutionary origin of three genes involved in these processes (dolichyldiphosphatase 1 (DOLPP1), steroid 5α-reductase (SRD5A3), and dolichol

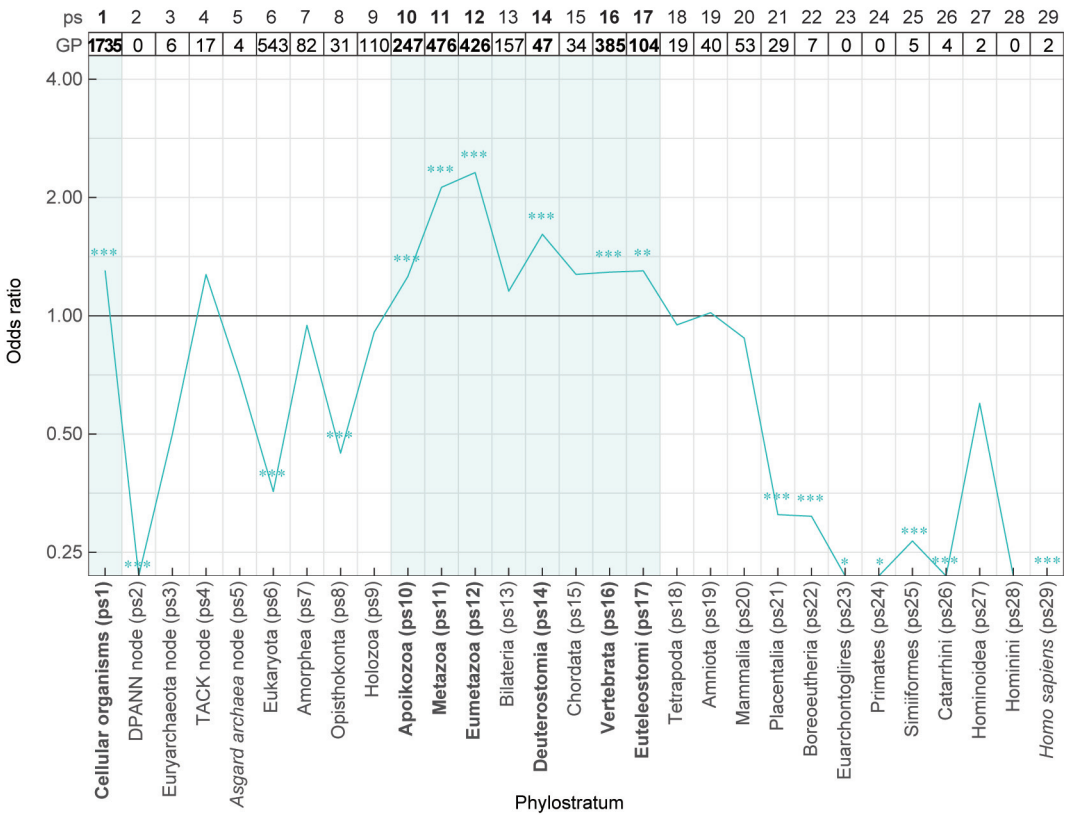

Fig. 4. Phylostratigraphic analysis of *H. sapiens* genes coding for GPs. The *x*-axis shows 29 phylogenetic levels (ps) of the consensus phylogeny with *H. sapiens* as a focal species (Fig. 1). The table at top shows distributions of 4565 genes which code for GPs genes along phylostrata. The odds ratio, shown on a log-scaled *y*-axis, indicates deviation from the expected frequency. We tested significance of the enrichments and depletions using two-tailed hypergeometric test (*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$). The evolutionary origin of most GP genes could be traced to cellular organisms (ps1) and eukaryota (ps6). GP genes are strongly enriched in cellular organisms (ps1) and in the evolutionary period that spans the origin of animals and bony vertebrates (Apoikozoa, ps10 to Euteleostomi, ps17). The shaded areas mark phylostrata with significant enrichments.

kinase (DOLK)) to eukaryota (ps6) which suggests that they appeared for the first time in the last eukaryotic common ancestor (LECA); in line with the results of a previous study [22].

The next step in the NG process is the addition of the first glycan on the dolichol-phosphate carrier catalyzed by a glycosyltransferase dolichyl-phosphate *N*-acetylglucosamine phosphotransferase 1 (DPAGT1) on the cytosolic side of the ER membrane (Fig. 5(a)). After this modification, a series of glycosyltransferases (ALG13, ALG14, ALG1, ALG2, and ALG11) sequentially add monosaccharides to yield the $Man_5GlcNAc_2$ structure. All these genes are mapped to ps1, suggesting their evolutionary occurrence in the last common ancestor of all cellular organisms (Fig. 5(a)). This newly synthesized glycan structure, which is attached to the dolichol carrier, is then flipped by RFT1 into the lumen of the ER (Fig. 5(a)). The NG process proceeds further on the luminal side of the ER membrane where a series of glycosyltransferases (ALG3, ALG9, ALG12, ALG6, ALG8, and ALG10) add additional glycans to reach the final $Glc_3Man_9GlcNAc_2$ structure common to all eukaryotes. All these proteins that act on the luminal side of the ER membrane are mapped to the origin of eukaryota (ps6) (Fig. 5(a)).

In the next step, the oligosaccharyltransferase (OST) protein complex transfers the glycan from the dolichol carrier to the polypeptide (Fig. 5(a)). This protein complex is assembled from five constant and three variable subunits. We traced all these subunits to the origin of eukaryota (ps6), with the exception of staurosporine and temperature sensitive oligosaccharyltransferase complex catalytic subunit (SST3), which we mapped at the origin of cellular organisms (ps1). This phylostratigraphic pattern suggests that the OST complex evolved during eukaryogenesis by the addition of new regulative subunits to the ancient core catalytic protein (SST3). Once an oligosaccharide is attached to a protein, the processing of the glycan begins with the removal of the glucose residues with two glucosidases (mannosyl-oligosaccharide glucosidase (MOGS) and GANAB). We traced MOGS to the origin of TACK *Asgard archaea* (ps4) and GANAB to the origin of cellular organisms (ps1). When the GP is finally properly folded, α-mannosidase I (MAN1B1), which we traced to cellular organisms (ps1, Fig. 5(a)), removes one mannose from the central glycan arm which signals that the GP is ready for transport outside the ER, usually to the Golgi apparatus [4].

By looking at the phylogenetic distribution of NG ER proteins (Fig. 5(a)), we found that essentially all of them come from only two evolutionary periods: cellular organisms (ps1) or eukaryota (ps6). Besides this binary evolutionary origin of NG ER proteins, which follows the global evolutionary pattern of NG genes (Fig. 3), we noticed that the proteins of the same evolutionary origin tend to have adjacent positions along the biosynthetic pathway (Fig. 5(a)). For instance, there is a block of seven consecutive biochemical steps (DPAGT1 to ALG11) where all proteins have evolutionary roots at the origin of cellular organisms (ps1, Fig. 5(a)). This block is followed by an evolutionary younger one (RFT1–ALG10) that contains eight biochemical steps where all proteins have phylogenetic roots at the origin of eukaryota (ps6, Fig. 5(a)).

To assess this phenomenon quantitatively, we devised a positional HI which estimates the clustering strength of identical phylostrata along a biosynthetic pathway. Low HI values reflect an extensive positional grouping of the identical phylostrata, whereas high HI values point to a scattered arrangement of phylostrata along the biosynthetic pathway. By applying this measure to the sequence of all NG reactions in the ER, we found a very low HI value that signals the biased grouping of phylostrata (HI = 5, Fig. 5(b)). To test if this biased grouping is statistically significant, we compared the observed HI value to the distribution of all possible HI values in the ER part of the NG pathway. We found that is extremely unlikely that such strong positional clustering of identical phylostrata occurred by chance ($p = 5.68 \times 10^{-3}$, Fig. 5(b)). We obtained similar results if we focused only on the core segment of the ER NG pathway from DPAGT1 to OST, which starts with the first attachment of a monosaccharide to the lipid carrier and ends with the transfer of the glycan to the polypeptide chain (HI = 2, $p = 3.86 \times 10^{-3}$, Fig. 5(b)).

Besides these longitudinal clustering biases along the ER NG pathway, we noticed that phylostrata were not randomly distributed with respect to the position of N-GM on the two sides of the ER membrane. The reactions that occur at the cytosol side of the ER membrane tend to use N-GM proteins that we traced to the origin of cellular organisms (ps1, Figs. 5(a) and (c)). Conversely, the reactions that unfold at the luminal side of the ER membrane preferentially rely on the N-GM proteins that we traced to the origin of eukaryota (ps6, Figs. 5(a) and (c)). This biased arrangement is significant when all genes are analyzed (Fisher's exact test, $p = 0.0104$) as well as when only the core part of the NG pathway from DPAGT1 to OST is considered (Fisher's exact test, $p = 2.77 \times 10^{-4}$). These results are quite stable in a broad range of *e*-value cutoffs (Fig. S4 in Appendix A).

After reaching Golgi, *N*-glycans go through further processing that involves coordinated action of glycosidases and glycosyltransferases that gives rise to three main classes of glycans: oligomannosidic, hybrid, and complex glycans. However, some glycans can completely skip these modifications [78]. The Golgi apparatus is organized into discrete cisternae, each containing a distinct subset of N-GM proteins which include glycosidases, glycosyltransferases, and nucleotide sugar transporters that provide glycosyltransferases with nucleotide sugars as glycosyl donors [78–80] (Fig. 6(a)). Similar to the ER, we found that NG proteins in Golgi are also assigned only to two phylostrata: cellular organisms (ps1) and eukaryota (ps6, Fig. 6(a)).

We traced seven nucleotide sugar membrane transporters acting in Golgi (solute carrier 35 (*SLC35*) gene family) to the origin of eukaryota (ps6, Fig. 6(a)). In contrast, we tracked all five glycosidases (mannose trimming enzymes), located in the *cis*- (mannosidase alpha class 1A members (MAN1A1, MAN1A2), and mannosidase alpha class 1C member 1 (MAN1C1)) and *medial*-Golgi (mannosidase alpha class 2A members (MAN2A1 and MAN2A2)), to the origin of cellular organisms (ps1, Fig. 6). On the other hand, 22 glycosyltransferases have binary phylogenetic origin. The majority of them (16) could be traced to eukaryota (ps6) and the rest (6) to the origin of cellular organisms (ps1, Fig. 6). In contrast to the ER, the Golgi NG pathway is not linear and lacks membrane polarity, as all glycosidase or glycosyltransferase reactions occur within the Golgi lumen. These properties prevented us from applying heterogeneity index or testing location biases similar to those applied in the ER (Fig. 5). However, we noticed an obvious bias in the phylogenetic origin of proteins acting in Golgi depending on their functional roles, where all glycosidases could be traced back to the origin of cellular organisms (ps1) and the majority of glycosyltransferases to the origin of eukaryota (ps6). There is a low probability of observing such a distribution by chance, (Fisher's exact test, $p = 5.72 \times 10^{-3}$, Fig. 6(b)), and the pattern is stable for *e*-values above 10^{-3} (Fig. S5 in Appendix A).

4. Discussion

Our global phylostratigraphic analysis of GM genes in humans revealed that glycosylation is an evolutionary old process most likely common to all life. However, GM pathways relevant to humans were established along protracted macroevolutionary time via three important periods. The first one is the origin of cellular organisms (ps1, Figs. 2 and 3) where we traced most of the human GM. The second is the origin of eukaryota (ps6, Figs. 2

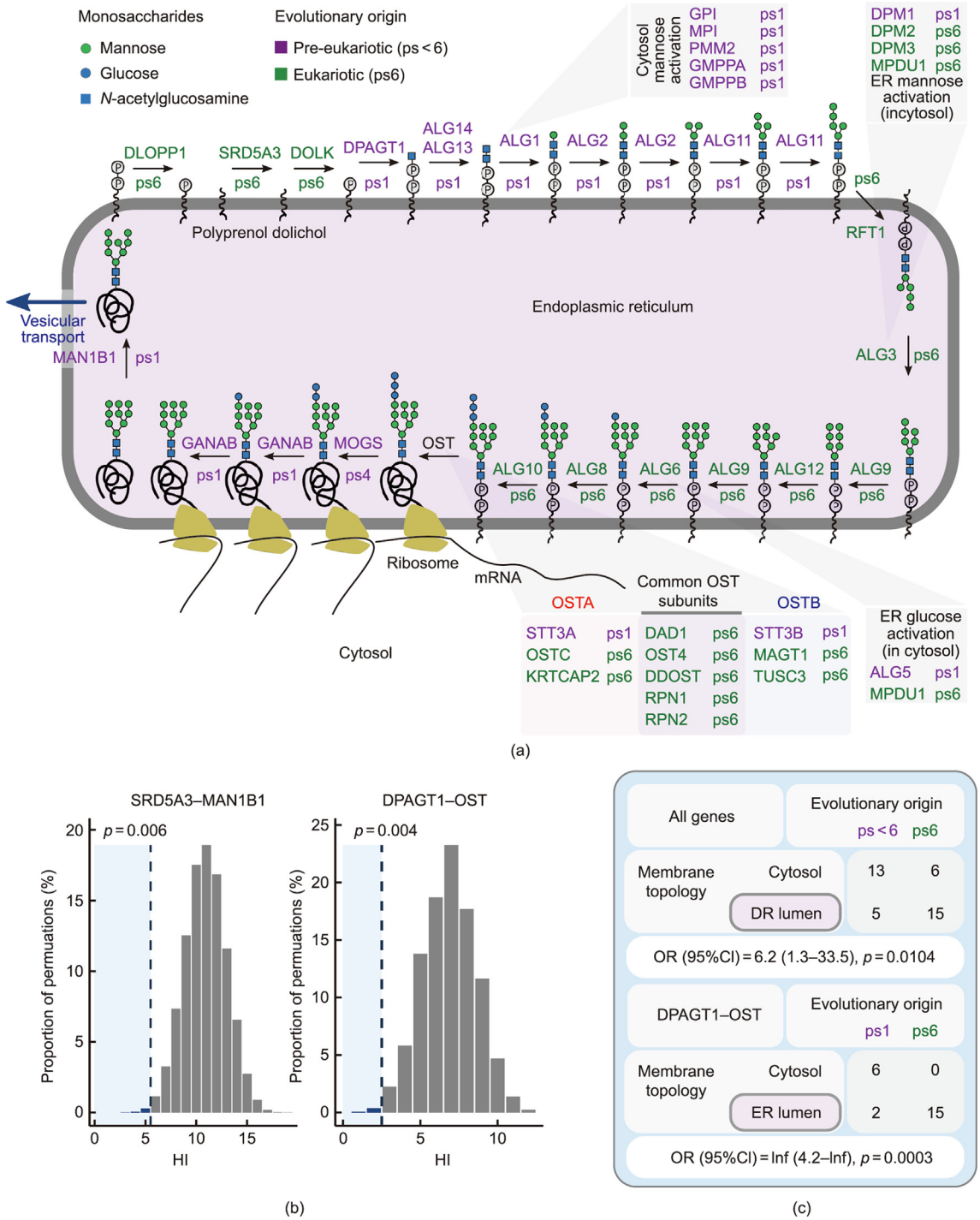

Fig. 5. Biased evolutionary origin of N-GM genes in the ER. (a) The part of the human NG biosynthetic pathway which is linked to the ER. The N-GM proteins (40 genes) are represented with Human Genome Organisation (HUGO) Gene Nomenclature Committee (HGNC) gene symbols, which are followed by corresponding phylostrata of their evolutionary origin (ps). Genes (proteins) that emerged at origin of cellular organisms (ps1), plus mannosyl-oligosaccharide glucosidase (*MOGS*) gene that we traced to TACK Archeae (ps4), are in violet. Genes (proteins) that emerged at the origin of eukaryota (ps6) are in green. (b) Empirical probability distribution of HI calculated by positional permutation of phylostrata. Dashed line represents the observed HI value, while *p*-value is probability of obtaining exactly that or a smaller HI value. (c) Contingency tables show the distribution of genes across two factors: evolutionary origin and membrane topology. We tested nonrandom associations between evolutionary origin and membrane topology using Fisher's exact test. Odds ratio (OR) and 95% confidence intervals (95%CI) are shown. RFT1 was excluded from the statistical analyses due to its ambiguous topology on the ER membrane (lumen vs cytosol). ALG: asparagine-linked glycosylation homologues; DAD1: defender against cell death 1; DDOST: dolichyl-diphosphooligosaccharide-protein glycosyltransferase non-catalytic subunit; DPAGT1: dolichyl-phosphate *N*-acetylglucosamine phosphotransferase 1; DPM: dolichyl-phosphate mannosyltransferase; GANAB: glucosidase II alpha subunit; GMPPA/B: GDP-mannose pyrophosphorylase A/B; Inf: infinity; KRTCAP2: keratinocyte associated protein 2; MAGT1: magnesium transporter 1; MAN1B1: α-mannosidase I; MPDU1: mannose-P-dolichol utilization defect 1; MPI: mannose phosphate isomerase; OST: oligosaccharyltransferase; OST4/A/C: oligosaccharyltransferase complex subunit 4/A/C; PMM2: phosphomannomutase 2; RPN: ribophorin; STT3A: STT3 oligosaccharyl-transferase complex catalytic subunit A; TUSC3: tumor suppressor candidate 3.

and 3) and the third one covers eukaryotic radiation and early metazoan diversification (ps7–ps13, Figs. 2 and 3). Finally, the complete GM, on which extant humans depend, was most likely fully established before the origin of mammals (ps20, Figs. 2 and

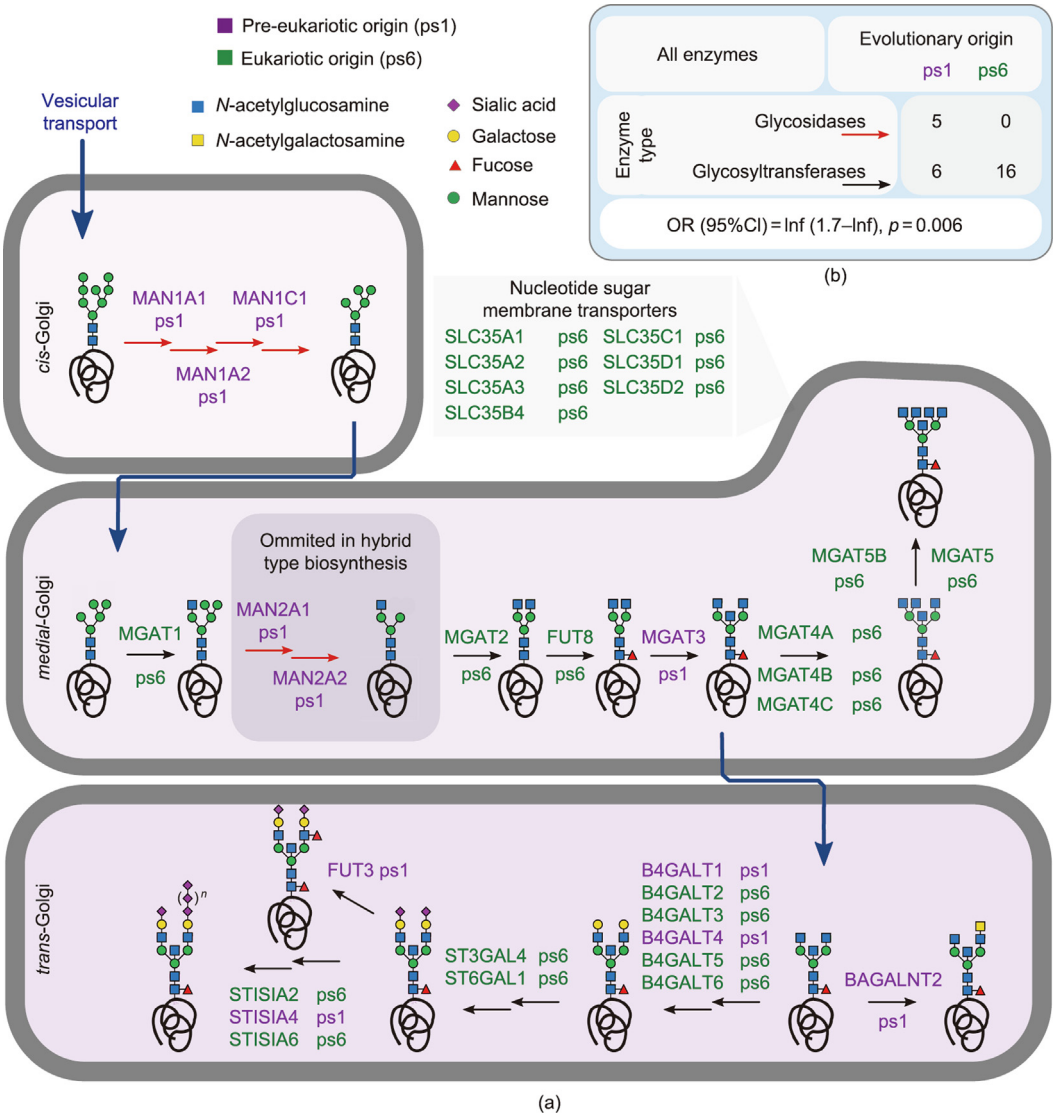

Fig. 6. Biased evolutionary origin of NG enzyme types in the Golgi apparatus. (a) The part of the human NG biosynthetic pathway which is linked to the Golgi apparatus. N-GM proteins (27) are represented with HGNC gene symbols, which are followed by corresponding phylostrata of their evolutionary origin (ps). Genes (proteins) that emerged at origin of cellular organisms (ps1) are in violet. Genes (proteins) that emerged at the origin of eukaryota (ps6) are in green. (b) The contingency table shows the distribution of genes across two factors: evolutionary origin and enzyme type. We tested nonrandom associations between evolutionary origin and enzyme type using Fisher's exact test. The OR, *p*-value, and 95%CI are shown. B4GALNT2: beta-1,4-*N*-acetyl-galactosaminyltransferase 2; B4GALT1–6: beta-1,4-galactosyltransferase 1–6; FUT3/8: fucosyltransferase 3/8; MGAT1: alpha-1,3-mannosyl-glycoprotein 2-beta-*N*-acetylglucosaminyltransferase; MGAT2: alpha-1,6-mannosyl-glycoprotein 2-beta-*N*-acetylglucosaminyltransferase; MGAT3: beta-1,4-mannosyl-glycoprotein 4-beta-*N*-acetylglucosaminyltransferase; MGAT4A/B/C: alpha-1,3-mannosyl-glycoprotein 4-beta-*N*-acetylglucosaminyltransferase A/B/C; MGAT5(B): alpha-1,6-mannosylglycoprotein 6-beta-*N*-acetylglucosaminyltransferase (B); ST3GAL1/4: ST3 beta-galactoside alpha-2,3-sialyltransferase 1/4; ST8SIA2/4/6: ST8 alpha-*N*-acetyl-neuraminide alpha-2,8-sialyltransferase 2/4/6; SLC35A1–A3/B4/C1/D1/D2: solute carrier family 35 member A1–A3/B4/C1/D1/D2.

3). However, it is interesting that we traced the comparably small number of GM genes to the archaea lineage (ps2–ps5, Figs. 2 and 3). This pattern suggests that glycosylation in humans predominantly relies on the GM that was already present at the origin of cellular organisms (ps1) and not on some archaea-specific innovations (ps2–ps5). However, this sharply contrasts the subsequent evolutionary period, the onset of eukaryotes, where we detected the burst of new GM genes (ps6, Figs. 2 and 3). This pattern creates a polar distribution in the analysis of the NG pathway where GM genes are either common to all cellular life (ps1) or specific to eukaryota (ps6, Figs. 5 and 6).

For instance, the structural and functional divergence between prokaryotic and eukaryotic OSTs reflects fundamental differences in the biological context of N-linked glycosylation. In prokaryotes, the OST is composed of a single catalytic subunit, such as undecaprenyl-diphosphooligosaccharide-protein glycotransferase (PglB) in *Campylobacter jejuni* or dolichyl-phosphooligosaccharide-protein glycotransferase (AglB) in archaea, which catalyzes the *en bloc* transfer of an oligosaccharide to asparagine residues in target proteins post-translationally, often after protein folding has occurred [22,23,81]. In contrast, eukaryotic OST is a multi-subunit complex embedded in the ER membrane, composed of the STT3 oligosaccharyltransferase complex catalytic subunit A or B (STT3A or STT3B) and multiple non-catalytic accessory proteins, each with specialized roles [82]. This complex organization supports the co-translational nature of glycosylation in eukaryotes, enabling synchronized interaction with the protein translocation machinery, as well as with quality control systems that monitor folding and post-translational modifications.

Subunits such as magnesium transporter 1 (MAGT1)/tumor suppressor candidate 3 (TUSC3) confer oxidoreductase activity necessary for glycosylation near disulfide-forming cysteines, while

others like oligosaccharyltransferase complex subunit 4 (OST4), dolichyl-diphosphooligosaccharide-protein glycosyltransferase non-catalytic subunit (DDOST), ribophorin 1 (RPN1), and RPN2 contribute to complex stability, lipid-linked oligosaccharide recruitment, and substrate positioning within the catalytic site [82]. The evolutionary transition from a single-subunit to a multi-subunit OST likely reflects the increased complexity of GP biosynthesis and trafficking in compartmentalized eukaryotic cells. Indeed, sequence and structural analyses have shown that the eukaryotic STT3 subunit shares homology with prokaryotic PglB/AglB, suggesting that the eukaryotic OST evolved via subunit accretion around an ancient catalytic core [81,82]. Thus, the presence of multiple non-catalytic subunits in eukaryotic OST is not merely a reflection of redundancy but an adaptation that enables precise spatial and temporal control of NG, essential for the biosynthesis of complex multi-domain GPs characteristic for eukaryotic organisms.

This binary evolutionary origin of N-GM genes (ps1 vs ps6) is a useful marker which allowed us to look for potential biases in the distribution of N-GM genes on the endoplasmic membrane. Surprisingly, we found a non-random distribution where NG genes specific for cellular organisms (ps1) are predominantly located on the cytosolic side and those specific for eukaryota (ps6) on the luminal side of the ER membrane (Fig. 5). This biased distribution has important evolutionary implications because it suggests that the lumen of the ER is a eukaryotic innovation which was devised by the invagination of the prokaryotic plasma membrane (Fig. 7). Bacterial and archaeal N-GM is always positioned at the innermost cell membrane where glycan synthesis occurs on the cytoplasmic side of the membrane. When completed, glycan is flipped to the extracellular or periplasmatic space (Fig. 7(a)). If we assume that this cross-membrane directionality of NG is preserved in eukaryotes, then the most likely explanation for the fact that in eukaryotes the flipping of the $Man_5GlcNAc_2$ structure occurs towards the ER lumen is that ER emerged by the invagination of the prokaryotic innermost membrane that contained N-GM (Fig. 7). Our finding that NG genes located on the cytoplasmic side of the ER membrane tend to be common to all cellular organisms (ps1), and that those at the luminal side are preferentially of eukaryotic origin (ps6), further corroborate this view (Fig. 5).

However, the question arises regarding the capability of extant and ancestral prokaryotes to form intracellular membrane vesicles (IMVs). In contrast to extracellular membrane vesicles (EMVs) which have been recently extensively studied in prokaryotes [83,84], IMVs are a much less explored feature. Nevertheless, it is clear that bacterial cells could also form IMVs [85–90]. For instance, it was shown that cell wall-deficient bacteria can encapsulate extracellular material by invagination of the cytoplasmic membrane by an endocytosis-like process [88]. This is a particularly suggestive finding, as wall-deficient bacteria are considered models of early cellular life that existed before the evolution of the cell wall [86,88,91].

Another important question relates to the evolutionary origin of the NG genes that in our analysis map to the cellular organisms (ps1). In principle, there are two contexts when these genes are placed at ps1. In the first one, these genes could have a significant match to bacteria only, while in the second one, these genes could have significant matches to both bacteria and archaea. A situation where most of the NG genes that are placed at ps1 have significant matches only to bacteria would indicate that the evolutionary oldest part of the eukaryotic NG pathway stems from the bacterial ancestor. In turn, this would imply that bacteria and archaea evolved NG independently. Conversely, a situation where most of the NG genes that are placed at ps1 have significant matches to both bacteria and archaea would suggest that already LUCA possessed basic N-GM that was then inherited in all subsequent lineages.

To test for these possibilities, we created a heatmap of the best matches per phylostratum for NG genes (Fig. S6 in Appendix A).

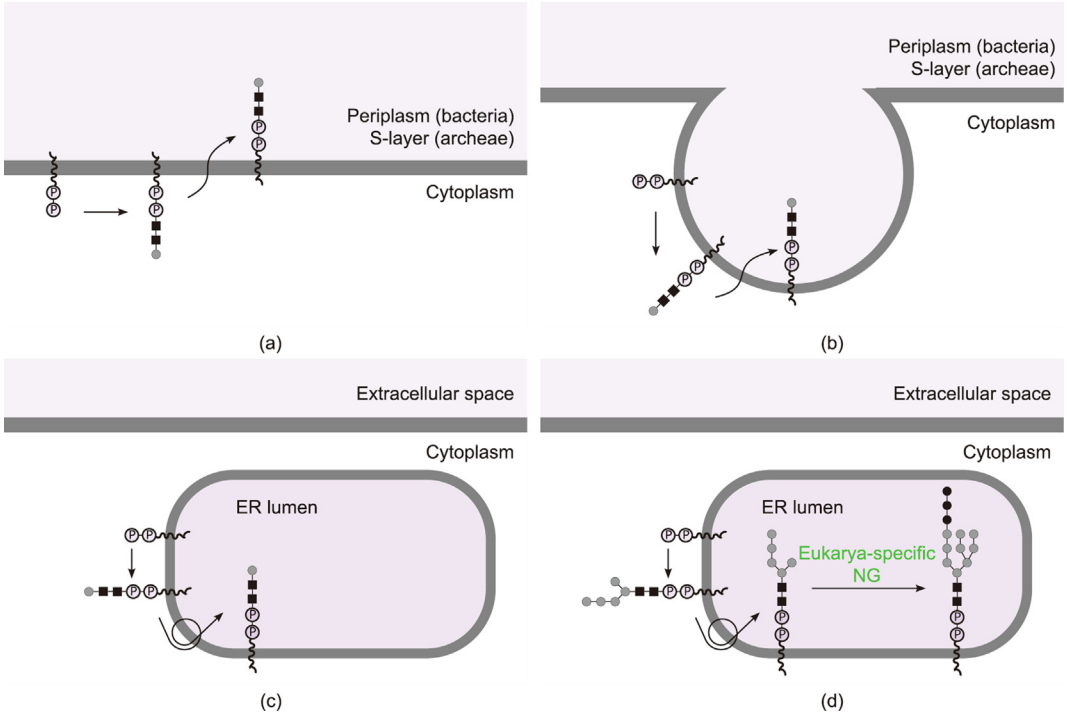

Fig. 7. A plausible evolutionary scenario for the placement of N-GM in the ER. (a) Schematic representation of the NG pathway in bacteria and archaea. NG occurs on the cytoplasmatic side of the innermost cell membrane. After the glycan synthesis is completed on the membrane-bound lipid carrier the resulting glycan is enzymatically flipped to the periplasm or S-layer of bacteria and archaea. (b) The plausible invagination of the prokaryotic innermost membrane that contained N-GM. (c) If this forming vesicle pinched off the cell membrane, then the residing N-GM gains the spatial position identical to that present in the eukaryotic ER. (d) During eukaryogenesis new genes were added to the NG pathway which further process glycans inside the lumen of ER.

This analysis revealed that most of NG genes mapped to ps1, which act in ER, have significant matches in both bacteria and archaea, suggesting that the basic N-GM was already present in LUCA and that all diverging lineages were built on this ancient core. A notable exception is ALG1, mannose phosphate isomerase (MPI), and phosphomannomutase 2 (PMM2) which seem to have significant matches only in bacteria. Similarly, the majority of genes that map to ps1 in Golgi have significant matches exclusively in bacteria (Fig. S6). Together, this gives some credence to the idea that eukaryotic NG genes that map to ps1 were actually inherited from bacterial ancestor, in line with the notion that eukaryotic membranes are of bacterial origin [35]. One approach to better resolve this issue in the future would be to conduct a detailed phylogenetic analysis of individual NG ps1 genes.

Our heatmap analysis also reveals that N-GM is largely preserved in eukaryotic side-branches along human lineage (*x*-axis, ps6–ps29, Fig. S6). However, there are also some exceptions. The most striking example is the loss of Golgi glycosyltransferase genes in fungi (*x*-axis, ps8, Fig. S6). This finding agrees with the observed lack of galactose and *N*-acetylglucosamine residues in glycan structures found in fungi [12].

It is important to establish links between our evolutionary findings and current biomedical research in glyco-therapeutics, glycan-based diagnostics, and the functional glycomics in human disease [92–94]. A good example is a recent finding that increased levels of α2,6-sialylation, catalyzed via α2,6-sialyltransferase-I (ST6Gal-I), have a significant role in development and progression of Alzheimer's disease [93]. In this analysis, we found that this sialyltransferase originated at the origin of eukaryota (ps6, Fig. S6). This information can be useful in tracing other glycosylation genes implicated in Alzheimer's disease, as it was previously shown that genes with similar functional significance correlate in terms of their phylostratigraphic origin [46]. Yet another possible application of here recovered evolutionary information is to add the phylostratigraphic origin of glycosylation genes, as a parameter, to platforms for glycosylation-omics analysis [92].

Here presented evolutionary findings on protein glycosylation can significantly inform and inspire current biomedical research. Since aberrant glycosylation is involved in numerous human diseases, such as cancer, autoimmune disorders, and infections, in depth understanding of the glycan biosynthesis machinery is obligatory for elucidating disease mechanisms and identifying new therapeutic targets [95]. Glycosylation plays a critical role in cancer immunotherapy as many tumor-associated glycans on the tumor glycocalyx help tumors evade immune surveillance and trigger immunosuppressive signaling via glycan-binding receptors [96]. For example, the sialylated glycan epitopes interact with the lectin receptors (GBP), leading to immune suppression [95]. The binary evolutionary origin of the GM in the ER and Golgi apparatus, with prokaryotic origins on the cytoplasmic side and eukaryotic origins in the lumen, offers a novel framework for designing glyco-therapeutics. These may include strategies such as microRNA-based targeting of either conserved, prokaryotic-like glycosylation processes or more recently evolved, eukaryotic-specific processes. This evolutionary perspective also supports the discovery and development of novel carbohydrate biomarkers for cancer stratification and improved clinical outcomes.

Taken together, our study showed that glycosylation in *H. sapiens* is an ancient feature, with a biphasic origin that comprises the origin of cellular organisms and the origin of eukaryotes. Despite this ancient history of GM, the usage of protein glycosylation intensified with the development of complex multicellular organisms, probably linked to selective pressures related to self-nonself recognition and improved coordination between the differentiated cells. Finally, using NG pathway as an evolutonary marker, we provide some support for the idea that ER evolved through invagination of the prokaryotic cell membrane that possessed the NG pathway.

CRediT authorship contribution statement

Domagoj Kifer: Writing – review & editing, Writing – original draft, Visualization, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Nina Čorak:** Writing – review & editing, Visualization, Methodology, Formal analysis, Data curation. **Mirjana Domazet-Lošo:** Writing – review & editing, Software, Methodology, Investigation, Formal analysis, Data curation, Funding acquisition. **Niko Kasalo:** Writing – review & editing, Validation, Resources, Investigation, Formal analysis. **Gordan Lauc:** Writing – review & editing, Validation, Supervision, Project administration, Funding acquisition, Conceptualization. **Göran Klobučar:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Methodology, Investigation, Formal analysis, Conceptualization. **Tomislav Domazet-Lošo:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We thank M. Futo, A. Tušar, S. Koska, and D. Franjević for discussions. This work was supported by the Croatian Science Foundation (IP-2016-06-5924), the City of Zagreb, and the Adris Foundation to Tomislav Domazet-Lošo; and the European Regional Development Fund (KK.01.1.1.01.0009 DATACROSS) to Mirjana Domazet-Lošo and Tomislav Domazet-Lošo. We used the computational resources of the University Computing Center—SRCE (Padobran) and the Institute Ruđer Bošković.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.eng.2025.06.039.

References

[1] Flynn RA, Pedram K, Malaker SA, Batista PJ, Smith BAH, Johnson AG, et al. Small RNAs are modified with *N*-glycans and displayed on the surface of living cells. Cell 2021;184(12):3109–24.e22.
[2] Wang W. Can DNA be glycosylated? Engineering. In press.
[3] Cantarel BL, Coutinho PM, Rancurel C, Bernard T, Lombard V, Henrissat B. The Carbohydrate-Active EnZymes database (CAZy): an expert resource for glycogenomics. Nucleic Acids Res 2009;37(Database):D233–8.
[4] Moremen KW, Tiemeyer M, Nairn AV. Vertebrate protein glycosylation: diversity, synthesis and function. Nat Rev Mol Cell Biol 2012;13(7):448–62.
[5] Schjoldager KT, Narimatsu Y, Joshi HJ, Clausen H. Global view of human protein glycosylation pathways and functions. Nat Rev Mol Cell Biol 2020;21(12):729–49.
[6] Varki A, Cummings RD, Esko JD, Stanley P, Hart GW, Aebi M, editors. Essentials of glycobiology. Cold Spring Harbor (NY): Cold Spring Harbor Laboratory Press; 2015.
[7] Colley KJ, Varki A, Kinoshita T. Cellular organization of glycosylation. In: Varki A, Cummings RD, Esko JD, Stanley P, Hart GW, Aebi M, editors. Essentials of glycobiology. Cold Spring Harbor (NY): Cold Spring Harbor Laboratory Press; 2017.
[8] Stanley P, Taniguchi N, Aebi M. *N*-glycans. In: Varki A, Cummings RD, Esko JD, Stanley P, Hart GW, Aebi M, editors. Essentials of glycobiology. Cold Spring Harbor (NY): Cold Spring Harbor Laboratory Press; 2017.
[9] Apweiler R. On the frequency of protein glycosylation, as deduced from analysis of the SWISS-PROT database. BBA-GS 1999;1473(1):4–8.

[10] Lauc G, Krištić J, Zoldoš V. Glycans are the third revolution in evolution. Front Genet 2014;5:00145.
[11] Varki A. Biological roles of glycans. Glycobiology 2017;27(1):3–49.
[12] Chung CY, Majewska NI, Wang Q, Paul JT, Betenbaugh MJ. SnapShot: N-glycosylation processing pathways across kingdoms. Cell 2017;171(1):258–e1.
[13] Joshi HJ, Narimatsu Y, Schjoldager KT, Tytgat HLP, Aebi M, Clausen H, et al. SnapShot: O-glycosylation pathways across kingdoms. Cell 2018;172(3):632–e2.
[14] Gagneux P, Aebi M, Varki A. Evolution of glycan diversity. In: Varki A, Cummings RD, Esko JD, Stanley P, Hart GW, Aebi M, editors. Essentials of glycobiology. Cold Spring Harbor (NY): Cold Spring Harbor Laboratory Press; 2017.
[15] Springer SA, Gagneux P. Glycomics: revealing the dynamic ecology and evolution of sugar molecules. J Proteomics 2016;135:90–100.
[16] Bishop JR, Gagneux P. Evolution of carbohydrate antigens—microbial forces shaping host glycomes? Glycobiology 2007;17(5):23R–34R.
[17] West CM, Malzl D, Hykollari A, Wilson IBH. Glycomics, glycoproteomics, and glycogenomics: an inter-taxa evolutionary perspective. Mol Cell Proteomics 2021;20:100024.
[18] Suzuki N. Glycan diversity in the course of vertebrate evolution. Glycobiology 2019;29(9):625–44.
[19] Corfield AP, Berry M. Glycan variation and evolution in the eukaryotes. Trends Biochem Sci 2015;40(7):351–9.
[20] Moran AP, Gupta A, Joshi L. Sweet-talk: role of host glycosylation in bacterial pathogenesis of the gastrointestinal tract. Gut 2011;60(10):1412–25.
[21] Schröder K, Bosch TCG. The origin of mucosal immunity: lessons from the holobiont Hydra. MBio 2016;7(6):e01184–216.
[22] Lombard J. Early evolution of polyisoprenol biosynthesis and the origin of cell walls. PeerJ 2016;4:e2626.
[23] Lombard J. The multiple evolutionary origins of the eukaryotic N-glycosylation pathway. Biol Direct 2016;11(1):36.
[24] Nikolayev S, Cohen-Rosenzweig C, Eichler J. Evolutionary considerations of the oligosaccharyltransferase AglB and other aspects of N-glycosylation across Archaea. Mol Phylogenet Evol 2020;153:106951.
[25] Petit D, Teppa E, Cenci U, Ball S, Harduin-Lepers A. Reconstruction of the sialylation pathway in the ancestor of eukaryotes. Sci Rep 2018;8(1):2946.
[26] Tomono T, Kojima H, Fukuchi S, Tohsato Y, Ito M. Investigation of glycan evolution based on a comprehensive analysis of glycosyltransferases using phylogenetic profiling. Biophysics 2015;12:57–68.
[27] Wang P, Wang H, Gai J, Tian X, Zhang X, Lv Y, et al. Evolution of protein N-glycosylation process in Golgi apparatus which shapes diversity of protein *N*-glycan structures in plants, animals and fungi. Sci Rep 2017;7(1):40301.
[28] Chen S, Pei CX, Xu S, Li H, Liu YS, Wang Y, et al. Rft1 catalyzes lipid-linked oligosaccharide translocation across the ER membrane. Nat Commun 2024;15(1):5157.
[29] Eichler J, Imperiali B. Biogenesis of asparagine-linked glycoproteins across domains of life—similarities and differences. ACS Chem Biol 2018;13(4):833–7.
[30] Nothaft H, Szymanski CM. Protein glycosylation in bacteria: sweeter than ever. Nat Rev Microbiol 2010;8(11):765–78.
[31] Jones MB, Rosenberg JN, Betenbaugh MJ, Krag SS. Structure and synthesis of polyisoprenoids used in N-glycosylation across the three domains of life. BBA-GS 2009;1790(6):485–94.
[32] Li H, Debowski AW, Liao T, Tang H, Nilsson HO, Marshall BJ, et al. Understanding protein glycosylation pathways in bacteria. Future Microbiol 2017;12(1):59–72.
[33] Baum DA, Baum B. An inside-out origin for the eukaryotic cell. BMC Biol 2014;12(1):76.
[34] Gould SB, Garg SG, Martin WF. Bacterial vesicle secretion and the evolutionary origin of the eukaryotic endomembrane system. Trends Microbiol 2016;24(7):525–34.
[35] López-García P, Moreira D. The Syntrophy hypothesis for the origin of eukaryotes revisited. Nat Microbiol 2020;5(5):655–67.
[36] Čorak N, Anniko S, Daschkin-Steinborn C, Krey V, Koska S, Futo M, et al. Pleomorphic variants of *Borreliella* (syn. *Borrelia*) *burgdorferi* express evolutionary distinct transcriptomes. IJMS 2023;24(6):5594.
[37] Domazet-Lošo M, Široki T, Šimičević K, Domazet-Lošo T. Macroevolutionary dynamics of gene family gain and loss along multicellular eukaryotic lineages. Nat Commun 2024;15(1):2663.
[38] Domazet-Lošo T, Brajković J, Tautz D. A phylostratigraphy approach to uncover the genomic history of major adaptations in metazoan lineages. Trends Genet 2007;23(11):533–9.
[39] Domazet-Lošo T, Carvunis AR, Mar Albà M, Sebastijan Šestak M, Bakarić R, Neme R, et al. No evidence for phylostratigraphic bias impacting inferences on patterns of gene emergence and evolution. Mol Biol Evol 2017;34(4):843–56.
[40] Domazet-Loso T, Tautz D. An ancient evolutionary origin of genes associated with human genetic diseases. Mol Biol Evol 2008;25(12):2699–707.
[41] Domazet-Lošo T, Tautz D. A phylogenetically based transcriptome age index mirrors ontogenetic divergence patterns. Nature 2010;468(7325):815–8.
[42] Domazet-Lošo T, Tautz D. Phylostratigraphic tracking of cancer genes suggests a link to the emergence of multicellularity in metazoa. BMC Biol 2010;8(1):66.
[43] Futo M, Opašić L, Koska S, Čorak N, Široki T, Ravikumar V, et al. Embryo-like features in developing *Bacillus subtilis* biofilms. Mol Biol Evol 2021;38(1):31–47.
[44] Šestak MS, Božičević V, Bakarić R, Dunjko V, Domazet-Lošo T. Phylostratigraphic profiles reveal a deep evolutionary history of the vertebrate head sensory systems. Front Zool 2013;10(1):18.
[45] Šestak MS, Domazet-Lošo T. Phylostratigraphic profiles in zebrafish uncover chordate origins of the vertebrate brain. Mol Biol Evol 2015;32(2):299–312.
[46] Shi L, Derouiche A, Pandit S, Rahimi S, Kalantari A, Futo M, et al. Evolutionary analysis of the *Bacillus subtilis* genome reveals new genes involved in sporulation. Mol Biol Evol 2020;37(6):1667–78.
[47] Tautz D, Domazet-Lošo T. The evolutionary origin of orphan genes. Nat Rev Genet 2011;12(10):692–702.
[48] Xia S, Chen J, Arsala D, Emerson JJ, Long M. Functional innovation through new genes as a general evolutionary process. Nat Genet 2025;57(2):295–309.
[49] Kanehisa M, Goto S. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res 2000;28(1):27–30.
[50] Boutet E, Lieberherr D, Tognolli M, Schneider M, Bansal P, Bridge AJ. UniProtKB/Swiss-Prot, the manually annotated section of the UniProt KnowledgeBase: how to use the entry view. In: Edwards D, editor. Plant bioinformatics. New York City: Springer New York; 2016. p. 23–54.
[51] Berbee ML, James TY, Strullu-Derrien C. Early diverging fungi: diversity and impact at the dawn of terrestrial life. Annu Rev Microbiol 2017;71(1):41–60.
[52] Hughes LC, Ortí G, Huang Y, Sun Y, Baldwin CC, Thompson AW, et al. Comprehensive phylogeny of ray-finned fishes (Actinopterygii) based on transcriptomic and genomic data. Proc Natl Acad Sci USA 2018;115(24):6249–54.
[53] Irisarri I, Baurain D, Brinkmann H, Delsuc F, Sire JY, Kupfer A, et al. Phylotranscriptomic consolidation of the jawed vertebrate timetree. Nat Ecol Evol 2017;1(9):1370–8.
[54] Misof B, Liu S, Meusemann K, Peters RS, Donath A, Mayer C, et al. Phylogenomics resolves the timing and pattern of insect evolution. Science 2014;346(6210):763–7.
[55] Morris JL, Puttick MN, Clark JW, Edwards D, Kenrick P, Pressel S, et al. The timescale of early land plant evolution. Proc Natl Acad Sci USA 2018;115(10):E2274–83.
[56] Regier JC, Shultz JW, Zwick A, Hussey A, Ball B, Wetzer R, et al. Arthropod relationships revealed by phylogenomic analysis of nuclear protein-coding sequences. Nature 2010;463(7284):1079–83.
[57] Shen XX, Zhou X, Kominek J, Kurtzman CP, Hittinger CT, Rokas A. Reconstructing the backbone of the Saccharomycotina yeast phylogeny using genome-scale data. G3 Genes|Genomes|Genetics 2016;6:3927–39.
[58] Wang H, Moore MJ, Soltis PS, Bell CD, Brockington SF, Alexandre R, et al. Rosid radiation and the rapid rise of angiosperm-dominated forests. Proc Natl Acad Sci USA 2009;106(10):3853–8.
[59] Richter DJ, Fozouni P, Eisen MB, King N. Gene family innovation, conservation and loss on the animal stem lineage. eLife 2018;7:e34226.
[60] Domazet-Loso T, Tautz D. An evolutionary analysis of orphan genes in *Drosophila*. Genome Res 2003;13(10):2213–9.
[61] Moyers BA, Zhang J. Toward reducing phylostratigraphic errors and biases. Genome Biol Evol 2018;10(8):2037–48.
[62] Vakirlis N, Carvunis AR, McLysaght A. Synteny-based analyses indicate that sequence divergence is not the main source of orphan genes. eLife 2020;9:e53500.
[63] Koska S, Leljak-Levanić D, Malenica N, Bigović Villi K, Futo M, Čorak N, et al. Developmental phylotranscriptomics in grapevine suggests an ancestral role of somatic embryogenesis. Commun Biol 2025;8(1):265.
[64] Kasalo N, Domazet-Lošo M, Domazet-Lošo T. Massive outsourcing of energetically costly amino acids at the origin of animals. 2024. bioRxiv 590100.
[65] Kasalo N, Domazet-Lošo M, Domazet-Lošo T. Convergence in amino acid outsourcing between animals and predatory bacteria. IJMS 2025;26(7):3024.
[66] R Core Team. R: a language and environment for statistical computing. R Core Team; 2021.
[67] Wickham H. ggplot2: elegant graphics for data analysis. 2nd ed. Cham: Springer International Publishing; 2016.
[68] Wilke CO. cowplot: streamlined plot theme and plot annotations for "ggplot2" 2015:1.1.3.
[69] Reily C, Stewart TJ, Renfrow MB, Novak J. Glycosylation in health and disease. Nat Rev Nephrol 2019;15(6):346–66.
[70] Lindahl U, Couchman J, Kimata K, Esko JD. Proteoglycans and sulfated glycosaminoglycans. In: Varki A, Cummings RD, Esko JD, Stanley P, Hart GW, Aebi M, editors. Essentials of glycobiology. Cold Spring Harbor (NY): Cold Spring Harbor Laboratory Press; 2017.
[71] Zeng L, Zhang A, Gu W, Zhou J, Zhang L, Du D, et al. Identification of haplotype tag single nucleotide polymorphisms within the receptor for advanced glycation end products gene and their clinical relevance in patients with major trauma. Crit Care 2012;16(4):R131.
[72] Taylor ME, Drickamer K, Schnaar RL, Etzler ME, Varki A. Discovery and classification of glycan-binding proteins. In: Varki A, Cummings RD, Esko JD, Stanley P, Hart GW, Aebi M, editors. Essentials of glycobiology. Cold Spring Harbor (NY): Cold Spring Harbor Laboratory Press; 2015.
[73] Beihammer G, Maresch D, Altmann F, Strasser R. Glycosylphosphatidylinositol-anchor synthesis in plants: a glycobiology perspective. Front Plant Sci 2020;11:611188.
[74] Paulick MG, Bertozzi CR. The glycosylphosphatidylinositol anchor: a complex membrane-anchoring structure for proteins. Biochemistry 2008;47(27):6991–7000.
[75] Vogt MS, Schmitz GF, Varón Silva D, Mösch HU, Essen LO. Structural base for the transfer of GPI-anchored glycoproteins into fungal cell walls. Proc Natl Acad Sci USA 2020;117(36):22061–7.

[76] Kinoshita T. Glycosylphosphatidylinositol (GPI) anchors: biochemistry and cell biology: introduction to a thematic review series. J Lipid Res 2016;57 (1):4–5.
[77] Nosjean O. No prokaryotic GPI anchoring. Nat Biotechnol 1998;16(9):799.
[78] Frappaolo A, Karimpour-Ghahnavieh A, Sechi S, Giansanti MG. The close relationship between the Golgi trafficking machinery and protein glycosylation. Cells 2020;9(12):2652.
[79] Hadley B, Litfin T, Day CJ, Haselhorst T, Zhou Y, Tiralongo J. Nucleotide sugar transporter SLC35 family structure and function. Comput Struct Biotechnol J 2019;17:1123–34.
[80] Mikkola S. Nucleotide sugars in chemistry and biology. Molecules 2020;25 (23):5755.
[81] Silberstein S, Gilmore R. Biochemistry, molecular biology, and genetics of the oligosaccharyltransferase. FASEB J 1996;10(8):849–58.
[82] Mohanty S, Chaudhary BP, Zoetewey D. Structural insight into the mechanism of N-linked glycosylation by oligosaccharyltransferase. Biomolecules 2020;10(4):624.
[83] Gill S, Catchpole R, Forterre P. Extracellular membrane vesicles in the three domains of life and beyond. FEMS Microbiol Rev 2019;43 (3):273–303.
[84] Toyofuku M, Nomura N, Eberl L. Types and origins of bacterial membrane vesicles. Nat Rev Microbiol 2019;17(1):13–24.
[85] Flood BE, Leprich DJ, Hunter RC, Delherbe N, MacGregor B, Nieuwenhze MV, et al. Outside-in: intracellular vesicles in giant sulfur bacteria contain peptidoglycan. 2022. bioRxiv 487978.
[86] Forterre P, Gribaldo S. Bacteria with a eukaryotic touch: a glimpse of ancient evolution? Proc Natl Acad Sci USA 2010;107(29):12739–40.
[87] Grant CR, Wan J, Komeili A. Organelle formation in bacteria and archaea. Annu Rev Cell Dev Biol 2018;34(1):217–38.
[88] Kapteijn R, Shitut S, Aschmann D, Zhang L, De Beer M, Daviran D, et al. Endocytosis-like DNA uptake by cell wall-deficient bacteria. Nat Commun 2022;13(1):5524.
[89] Lonhienne TGA, Sagulenko E, Webb RI, Lee KC, Franke J, Devos DP, et al. Endocytosis-like protein uptake in the bacterium Gemmata obscuriglobus. Proc Natl Acad Sci USA 2010;107(29):12883–8.
[90] Saier M. Membrane-bound compartments in bacteria: membrane-bound structures within Gram-negative photosynthetic and magnetotactic bacteria help overturn an old dogma about prokaryotes. Microbe Magazine 2014;9 (9):368–72.
[91] Errington J, Mickiewicz K, Kawai Y, Wu LJ. L-form bacteria, chronic diseases and the origins of life. Philos Trans R Soc Lond B Biol Sci 2016;371 (1707):20150494.
[92] Liu X, Meng Y, Fu B, Song H, Gu B, Zhang Y, et al. GlycoPro: a high-throughput sample-processing platform for multi-glycosylation-omics analysis. Engineering. In press.
[93] Yang K, Li X, Lai M, Zhao W, Song W, Chen S, et al. Ablation of ST6Gal-I downregulates BACE1 expression and suppresses production of Aβ42 plaques in Alzheimer's disease. Engineering. In press.
[94] Zuniga-Banuelos FJ, Hoffmann M, Reichl U, Rapp E. New avenues for human blood plasma biomarker discovery via improved in-depth analysis of the low-abundant N-glycoproteome. Engineering. In press.
[95] Chiang AWT, Baghdassarian HM, Kellman BP, Bao B, Sorrentino JT, Liang C, et al. Systems glycobiology for discovering drug targets, biomarkers, and rational designs for glyco-immunotherapy. J Biomed Sci 2021;28(1):50.
[96] RodrÍguez E, Schetters STT, van Kooyk Y. The tumour glyco-code as a novel immune checkpoint for immunotherapy. Nat Rev Immunol 2018;18 (3):204–11.