The glycan alphabet is not universal: a hypothesis

Jaya Srivastava[1]*, P. Sunthar[2] and Petety V. Balaji[1]

[1]Department of Biosciences and Bioengineering, Indian Institute of Technology Bombay, Powai, Mumbai 400076, India
[2]Department of Chemical Engineering, Indian Institute of Technology Bombay, Powai, Mumbai 400076, India

*Corresponding author
Email: jaya_srivastava@iitb.ac.in

Abstract

Several monosaccharides constitute naturally occurring glycans but it is uncertain if they constitute a universal set like the alphabets of proteins and DNA. Based on the available experimental observations, it is hypothesized herein that the glycan alphabet is not universal. Data on the presence / absence of pathways for the biosynthesis of 55 monosaccharides in 12939 completely prokaryotic genomes are presented in support of this hypothesis. Pathways were identified by searching for homologs of biosynthesis pathway enzymes. Substantial variations are observed in the set of monosaccharides used by organisms belonging to the same phylum, genera and even species. Monosaccharides are grouped as Extensive, Intermediate and Rare based on their prevalence in prokaryotes. It is observed that fewer enzymes suffice to biosynthesize the Extensive group. It appears that the Extensive group originated before the formation of three domains of life. In contrast, the Rare group are confined to a few species in a few phyla, suggesting that they evolved much later. Fold conservation, as observed in aminotransferases and SDR superfamily members involved in monosaccharide biosynthesis, suggests neo- and sub-functionalization of genes leading to the formation of Rare group monosaccharides. Non-universality of the glycan alphabet begets questions about the role of different monosaccharides in determining an organism's fitness.

Introduction

Living organisms show enormous diversity in organization, size, morphology, habitat, etc., but are unified by the highly conserved processes of central dogma: replication, transcription and translation. The enormous diversity seen in life forms is encoded by DNA and decoded primarily by proteins. Both DNA and proteins use the same set of building blocks (nucleotide bases and amino acids, respectively) in all organisms; yet, they store the requisite information by merely varying the (i) set/subset of building blocks used, (ii) number of times each building block is used and (iii) sequence in which the building blocks are linked [collectively referred to as the 'sequence' (Table 1)]. The information required for several other biological processes are stored by glycans, the third group of biological macromolecules (Varki 1993). It has been found that glycans evolve rapidly in response to changing environmental conditions, especially in prokaryotes, and thus serve as diversity generating machines (Mostowy and Holt 2018) (Mostowy et al. 2017). The question is, do glycans use the same set of building blocks (viz., monosaccharides) in all organisms, the way proteins and nucleic acids do?

Monosaccharides show lot more structural variations than amino acids in terms of the enantiomeric forms (both D and L), size (5 to 9 carbon atoms), ring type (pyranose, furanose), and type and extent of modification (deoxy, amino, N-formyl, N-acetyl, etc.). Some pairs of monosaccharides differ from each other merely in the configuration of carbon atoms. The sequence [as defined above] of monosaccharides brings about diversity even in the primary structure of glycans. DNA and protein are linear polymers and the linkage type that connects monomers remains the same throughout. In contrast, glycans can be branched and have alternative isomeric linkages (e.g., $\alpha1\rightarrow3$, $\beta1\rightarrow4$, $\alpha2\rightarrow6$ and so on) (Gabius and Roth 2017), two features that enhance diversity in glycans. Repeat length heterogeneity (the number of occurrences of a sequence repeat) is observed in glycans (Bravo et al. 2008; Kalynych et al. 2014), as well as DNA and proteins, although there are no data on the frequency of occurrence of this feature in these three classes of biomolecules. An additional factor contributes to the diversity in the primary structure of glycans viz., microheterogeneity (Johannessen et al. 2012), a feature not seen in DNA or proteins (Table 1). These structural variations demand the use of multiple analytical

techniques for sequencing and hence there are not any automated methods for sequencing glycans. Biosynthesis of DNA and proteins is template-driven but not that of glycans. Consequently, there is no equivalent of polymerase chain reaction or recombinant protein expression to 'amplify' glycans to obtain samples in amounts required for structural / functional analysis. These constraints have largely limited data on glycan sequences.

Table 1 Sources of diversity in primary structures of DNA, proteins and glycans

Feature	DNA	Protein	Glycan
Structural diversity of building blocks	(i) Low (four nucleotides). (ii) Nucleotide modifications (known but rare): N7-methylation of Ade/Gua	(i) Higher, relative to DNA (20 amino acids). Has structurally similar pairs: Asp/Glu, Asn/Gln, Phe/Tyr, Leu/Ile/Val. (ii) Amino acid modifications (known but rare): hydroxylation of Pro and Lys, selenocysteine, pyrrolysine	(i) Highest. Several pentoses and hexoses, many of which are configurational isomers. (ii) Pyranose and furanose forms (e.g., Gal). (iii) Both enantiomeric forms (e.g., Gal). (iv) Modifications extremely common (deoxy, uronic acid, deoxyamino and its derivatives, acetylation, sulfation, …)
Linkage	3',5'-phosphodiester. 5',5'-phosphodiester occurs but very rare	Amide bond. γ-COOH of Glu and ε-NH2 group of Lys used but very rare	Alternative isomeric linkages are very common (α1→3, β1→3, α1→6, β1→4, α2→3, α2→6 and so on)
Sequence	(i) Set/subset of building blocks used (ii) Number of times each building block is used (iii) Sequence in which the building blocks are linked		
Branching	Absent	Absent	Quite common
Sequence repeat heterogeneity	Present	Present	Present
Microheterogeneity[1]	Absent	Absent	Present

[1]Microheterogeneity refers to the presence of multiple forms of glycans (minor but distinct variations) present in different molecules of a protein synthesized by a cell at the 'same' time. This feature is unique to glycans just as the presence of splice variants is unique to proteins.

Carbohydrates are viewed as the third alphabet of life (Kaltner et al. 2019). How large is this alphabet? The number of monosaccharides used collectively by living systems is at least 60. An analysis of the prokaryotic glycan structural data showed a distinct difference in the set of monosaccharides used by prokaryotes and mammals (Herget et al. 2008). Is this difference evidence of absence i.e., monosaccharides found in databases are true representations of monosaccharides used by these organisms, and those not found are not used by organisms? Or, is it just absence of evidence i.e., the glycan alphabet is indeed universal and the observed differences are merely due to inadequate sequencing? With the availability of the whole genome sequence of a large number of prokaryotes, it has now become possible to resolve this issue.

In this study, it is hypothesized that the glycan alphabet is NOT universal, i.e., different organisms use different sets of monosaccharides. This is in contrast to those of DNA, RNA and proteins. This hypothesis is put forward based on the observations that >60 monosaccharides are found in living systems, the database of glycan structures shows differential usage of monosaccharides and that several serotypes differ from each other in the monosaccharides they use. Results obtained by mining whole genome sequences of nearly 13000 prokaryotes are presented herein in support of this hypothesis.

Results

Glycan alphabet size is not the same in prokaryotes: The number of monosaccharides used by different species is significantly different (Figure 1) and is independent of proteome size (Figure S3). Data for the prevalence of monosaccharides in 12939 genomes is very similar to that in 3384 species (Figure S4) indicating that the outcome is not biased by the skew in the number of genomes (strains) sequenced for a given species (Figure S1). In fact, none of the organisms use all 55 monosaccharides: the highest number of monosaccharides used by an organism is 23 [*Escherichia coli* 14EC033]. Just 1 and 2 monosaccharides are used by 188 and 117 species, respectively. Glucose, galactose and mannose, and their 2-N-acetyl (Glc2NAc, Gal2NAc, Man2NAc) and uronic acid (GlcA, GalA, Glc2NAcA, Gal2NAcA) derivatives are the most prevalent besides L-rhamnose, as

the biosynthesis pathways for these monosaccharides are found in >50% of genomes (Figure 2). These monosaccharides are thus categorized as 'Extensive' group. However, none of them are used by all organisms (Supplementary_data.xlsx:Worksheet3).

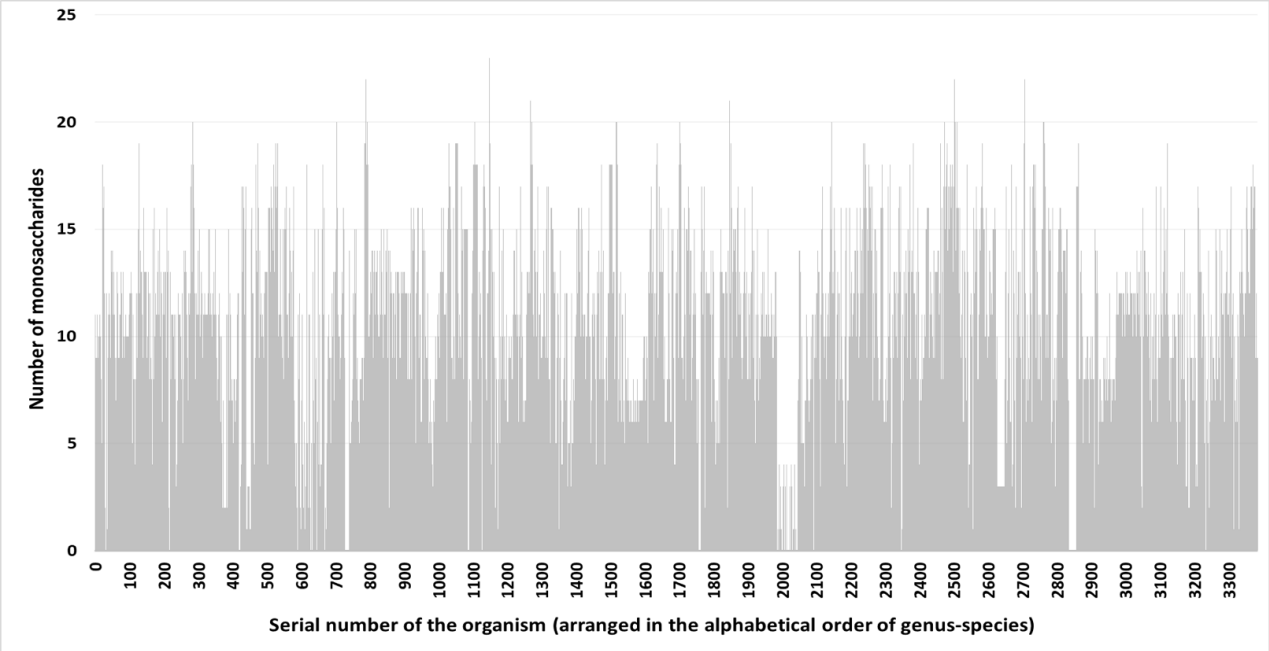


Figure 1 The number of monosaccharides for which biosynthesis pathways are found in a species. More than one strain is sequenced for several species (Figure S1). In such cases, data for the strain which has the highest number of monosaccharides is plotted. Total number of species = 3384

EXTENSIVE GROUP	INTERMEDIATE GROUP	RARE GROUP			
"Simple" sugars	**"Simple" sugars**	**"Simple" sugar**			
Glucose	Xylose	L-Galactose			
Galactose	L-Arabinose				
Mannose		**Amino / N-acetyl derivatives**			
	Furanose form	Qui2NAc	Fuc2NAc	Per	Qui4NFo
C2-N-Acetyl derivatives	Galactofuranose	Qui3NAc	Fuc3NAc	Per4Ac	L-Ara4NFo
Glc2NAc		Qui4NAc	Fuc4Nac	L-Rha2NAc	
Gal2NAc	**Deoxy derivative**	L-Qui2NAc	L-Fuc2NAc	Bac2Ac4Ac	
Man2NAc	L-Fucose				
		Uronic acid [+N-acetyl] derivatives			
Uronic acid derivatives	**C4-amino derivative**	ManA	Man2NAc3NAcA		
GlcA	L-Ara4N				
GalA		**Deoxy derivatives [+ furanose form]**			
Glc2NAcA		Cereose	Rhamnose	Fucofuranose	6-Deoxy-L-talose
Gal2NAcA		Cillose	Fucose	Yelosamine	6-Deoxytalose
Man2NAcA					6-Deoxygulose
		Dideoxy derivatives			
Deoxy derivative		Paratose	Abequose	Tyvelose	
L-Rhamnose		L-Colitose	L-Ascarylose		
		Heptoses			
		L-*glycero*-β-D-*manno* heptose		D-*glycero*-α-D-*manno* Heptose	
		6-Deoxy-α-D-*manno* heptose		6-Deoxy-α-D-*altro* heptose	
		9-Carbon sugars			
		Neu5Ac	L-Pse5Ac7Ac	Leg5Ac7Ac	

Figure 2 Classification of monosaccharides into three groups based on prevalence in prokaryotic genomes. These groups are Extensive (found in >=50% of genomes), Intermediate and Rare (found in <=10% of genomes). Abbreviated names are used for some of the monosaccharides. Full names of these are given in Supplementary_data.xlsx:Worksheet4.

Evolution and diversification of glycan alphabet: It is observed that only a limited set of enzymes suffice to biosynthesize the Extensive group monosaccharides e.g., nucleotidyltransferases (activation), amidotransferase and N-acetyltransferase (Hex2NAc from a hexose), C4-epimerase (Glc to Gal) and C6-dehydrogenase (uronic acid) belonging to the SDR superfamily, non-hydrolyzing C2-epimerase (Glc2NAc to Man2NAc), mutase (6-P to 1-P) and isomerase (pyranose to furanose) (Figure 3 and Supplementary_data.xlsx:Worksheet5). Using this limited set of monosaccharides, organisms seem to achieve structural diversity by mechanisms such as alternative isomeric linkages, branching and repeat length heterogeneity. Some organisms use an additional set of monosaccharides, viz., L-fucose, galactofuranose, xylose, L-Ara4N and L-arabinose. These monosaccharides are categorized as Intermediate group. Organisms using this group of monosaccharides have enhanced the glycan repertoire by acquiring C3/C5-epimerase, 4,6-

dehydratase, C4-reductase, C6-decarboxylase and C4-aminotransferase. Rest of the monosaccharides are used by very few organisms and thus constitute the 'Rare' group (Figure 2).

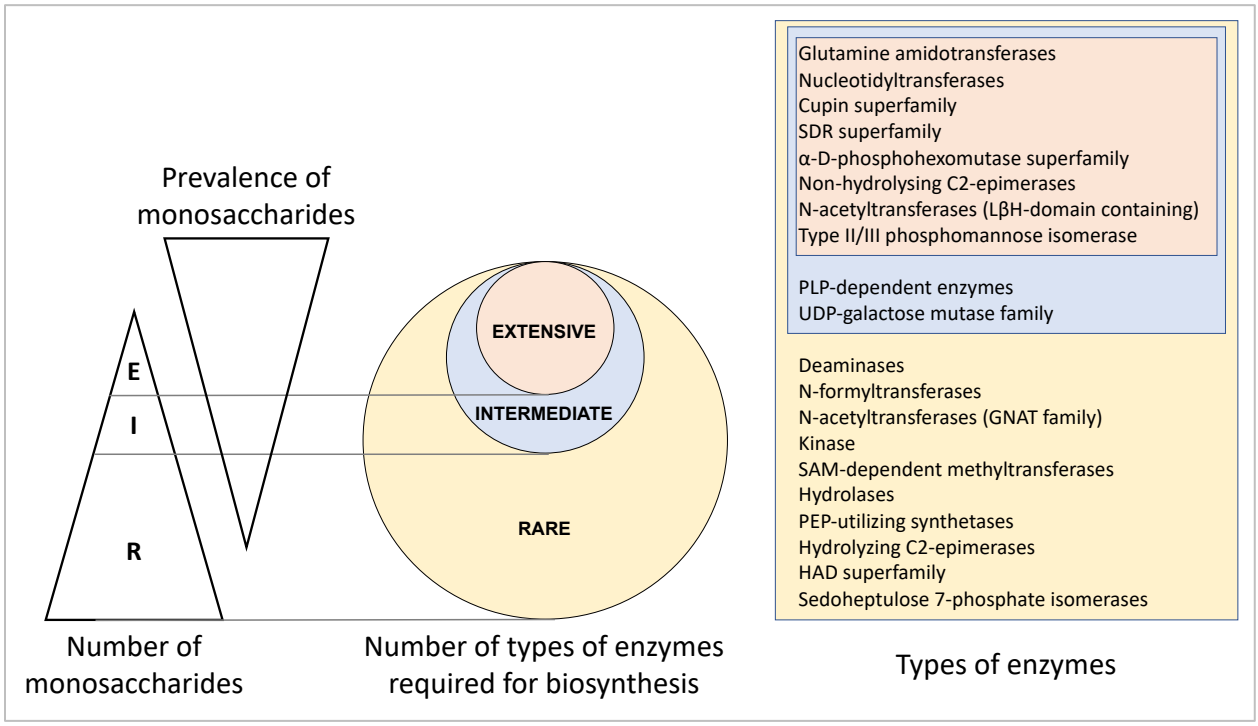
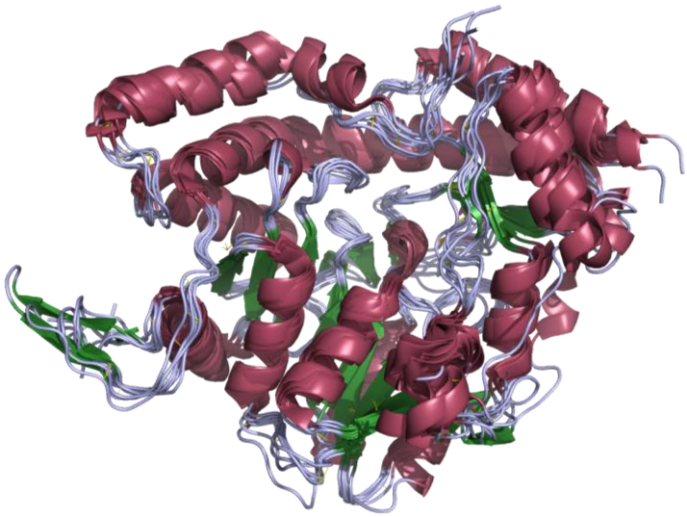


Figure 3 A qualitative comparison of the number of monosaccharides of the three groups viz., Extensive (E), Intermediate (I) and Rare (R) with their prevalence in prokaryotic genomes and the number of types of enzymes required for their biosynthesis. The size of a group is inversely related to the prevalence of the corresponding group of monosaccharides. Enzymes required for the biosynthesis of Extensive group monosaccharides are required for the biosynthesis of Intermediate and Rare groups also; similarity, those for the Intermediate group are required for the biosynthesis of Rare group also. Different enzymes belonging to each of the superfamily mentioned above are listed in the file Supplementary_data.xlsx:Worksheet5. Note that the group sizes are not to scale.
HAD, haloalkanoic acid dehalogenase
GNAT, GCN5-related N-acetyltransferases
LβH, left handed β helix
PEP, phosphoenolpyruvate
PLP, pyridoxal 5'-phosphate
SAM, S-adenosyl-L-methionine
SDR, short chain dehydrogenase reductase
UDP, uridine diphosphate

Occurrence of the Extensive group of monosaccharides in all three domains of life points to their presence early on during evolution. Neo- and sub-functionalization of horizontally acquired and duplicated genes during the course of evolution have been widely reported (e.g.,(Ohno 2013; Copley 2020)). It is envisaged that the enzymes required for the biosynthesis of Rare group monosaccharides have arisen by such neo- and sub-functionalization. Aminotransferase and short-chain dehydrogenase reductase (SDR) superfamily enzymes involved in the biosynthesis of monosaccharides lend support to this inference. Superposition of a few C3- and C4-aminotransferases show remarkable conservation of the 3D structures despite differences in the pyranose ring position at which the amino group is transferred as well as the nucleotide sugar substrate (Figure 4). 3D structures are conserved even among SDR superfamily enzymes despite catalysing different reactions viz., epimerization (at C2 or C4), removal of water (dehydratase at C4, C6) and reduction (at C4).

(A)



(B)

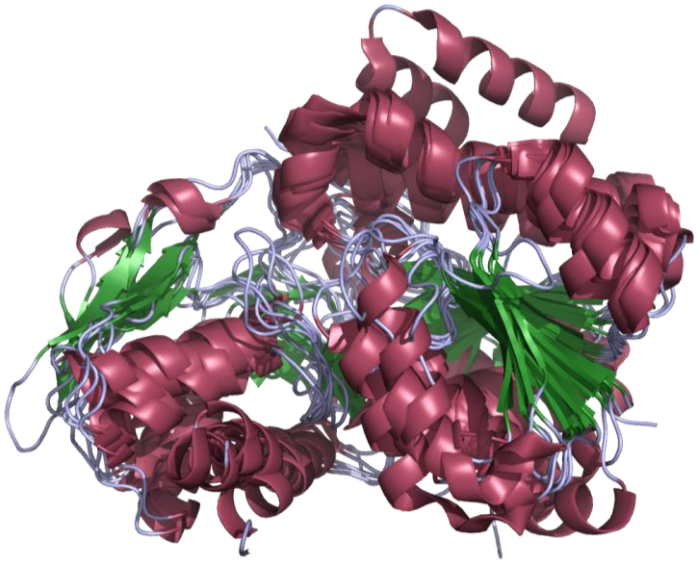


Figure 4 3D structural superimposition of enzymes belonging to aminotransferase (A) and SDR (B) superfamilies involved in the biosynthesis of monosaccharides. Color scheme: helices, raspberry red; sheet, forest green; loops, light blue. **Panel A**: Aminotransferase superfamily enzymes: 1MDO_A: ArnB from UDP-L-Ara4N biosynthesis; 2FNI_A: PseC from CMP-L-Pse45Ac7Ac biosynthesis; 2OGA_A: DesV from (d)TDP-desosamine biosynthesis; 3BN1_A: perA from GDP-per biosynthesis; 3NYU_A: WbpE from UDP-Man2NAc3NAcA biosynthesis; 4PIW_A: WecE from (d)TDP-Fuc4NAc biosynthesis; 4ZTC_A: PglE from CMP-Leg5Ac7Ac biosynthesis; 5U1Z_A: wlaRG from (d)TDP-Fuc3NAc/Qui3NAc biosynthesis. ArnB, PseC, perA, WecE and PglE are C4-aminotransferases whereas DesV, WbpE and wlaRG are C3-aminotransferases. **Panel B**: 1ORR_A: RfbE, C2-epimerase from CDP-tyvelose biosynthesis; 2PK3_A: Rmd, 4-reductase from GDP-rhamnose biosynthesis; 1KBZ_A: rmlD, C4-reductase from (d)TDP-L-rhamnose biosynthesis; 1T2A_A: gmd, C4,C6-dehydratase from GDP-L-fucose biosynthesis; 1SB8_A: WbpP, C4-epimerase from UDP-Gal2NAc biosynthesis; 5BJU_A: PglF, C4,C6-dehydratase from UDP-Bac2Ac4Ac biosynthesis.

Glycan alphabet varies even across strains: Remarkably, variations in the size of glycan alphabet are significant even at the strain level (Figure 5). Strain-specific differences are pronounced in species such as *E. coli*, *Pseudomonas aeruginosa* and *Campylobacter jejuni* (Figure 6) possibly reflecting the diverse environments that these organisms inhabit. Among organisms which inhabit the same environment, strain-specific differences show mixed pattern: among the 71 strains of *Streptococcus pneumoniae*, the maximum and minimum number of monosaccharides utilized by a strain are 4 and 12, respectively. Such a variation could have evolved as a mechanism to evade host immune response. In contrast, strains of *Streptococcus pyogenes* and strains of *Staphylococcus aureus* inhabit the same environment (respiratory tract and skin, respectively) and show very little variation in the monosaccharides they use. Both are capsule producing opportunistic pathogens suggesting that they might bring about antigenic variation by variations in linkage types, branching etc. (Keinhörster et al. 2019), even with the same set of monosaccharides. Strains of *Mycobacterium tuberculosis, Brucella melitensis, Brucella abortus* or *Neisseria gonorrhoeae*, all of which are human intracellular pathogens, also show insignificant variation. It is possible that different strains of a pathogen are a part of distinct microbiomes and microbial interactions within the biome/with the host determine the glycan alphabet of the organism. Availability of additional characteristics such as phenotypic data and temporal variations in glycan structures is critical for understanding the presence/absence of strain-specific variations.

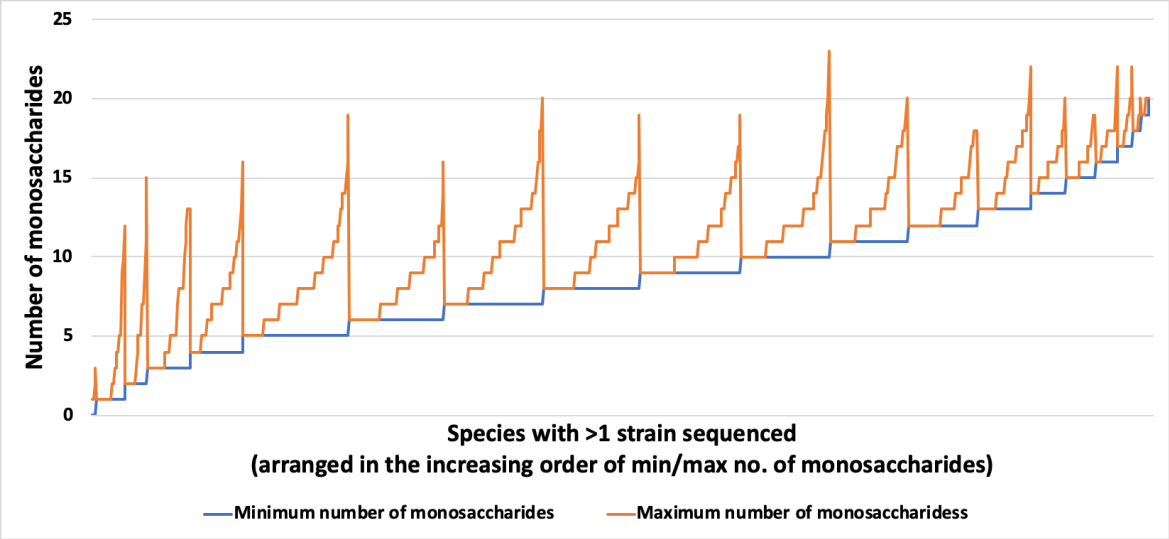


Figure 5 Variations in the number of monosaccharides used by different strains of a species. Species with more than one sequenced strain and at least one monosaccharide predicted in one of the strains are considered. Only the smallest and largest numbers are shown.

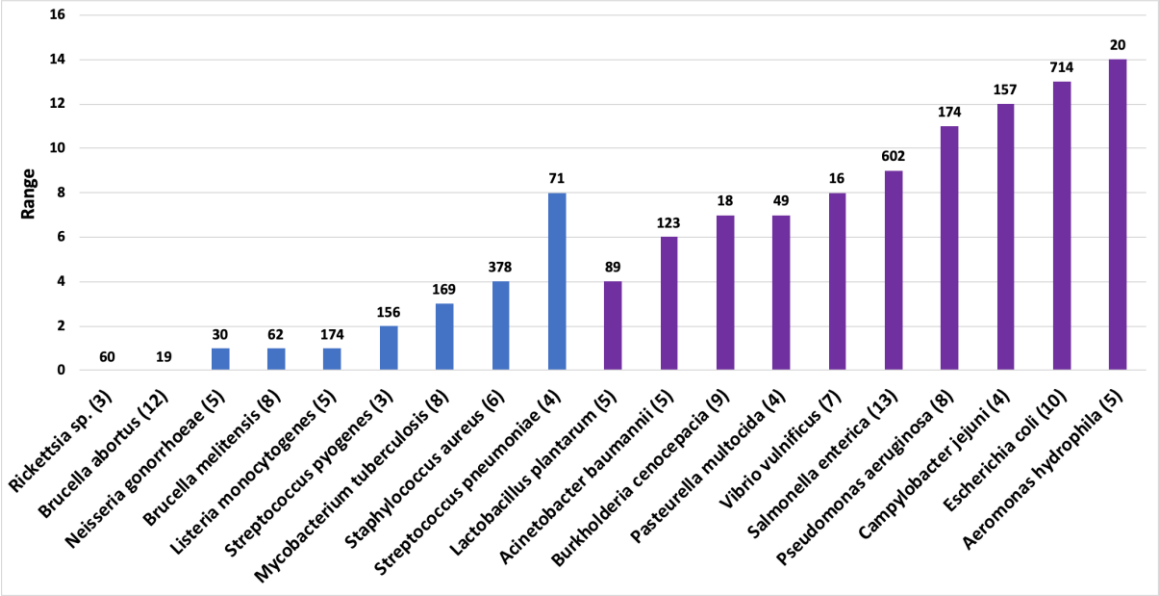


Figure 6 Different strains of some of the species do not use the same number of monosaccharides. The range of the number of monosaccharides used by various strains of some of the clinically important species are shown here. The number of sequenced strains for each organism is shown above the corresponding bar. Number in parenthesis after the name of each organism represents the minimum number of monosaccharides used by one of the strains of this organism. Note that the set of monosaccharides encoded by different strains utilizing the same number of monosaccharides may vary. Organisms associated with narrow habitat are shown in blue, while those with broad habitat are shown in purple.

Prevalence of monosaccharides across Phyla: Not all sugars of the Extensive group (Figure 1) are found across all phyla whereas Neu5Ac belonging to the Rare group is found across all phyla. GlcA and GalA (Extensive group) are absent in Thermotogae suggesting that pathways for their biosynthesis are lost in this phylum. A similar conclusion is drawn for the absence of L-fucose and L-colitose in TACK group phylum. Most of the Rare group sugars are limited to very few species in a few phyla (Figure 7). For instance, Fuc4NAc and L-*glycero*-β-D-*manno*-heptose (ADP-linked) are found only in Gamma-proteobacteria, a class that comprises of several pathogens. The other three heptoses, which are GDP-linked, are absent in Gamma-proteobacteria. Recently, it was found that *Helicobacter pylori*, belonging to the class Epsilon-proteobacteria, synthesizes ADP-*glycero*-β-D-*manno*-heptose for activating the NF-κβ pathway in human epithelial cells (Pfannkuch et al. 2019). This pathway has been experimentally characterized in very few organisms. Consequently, homologs for this pathway are found by BLASTp queries and not by HMM profiles. In the present study, this pathway turned out to be a false negative because of the high stringency set for BLASTp thresholds. In view of this, it is possible that such sugars which appear restricted to a few phyla are also found in others.

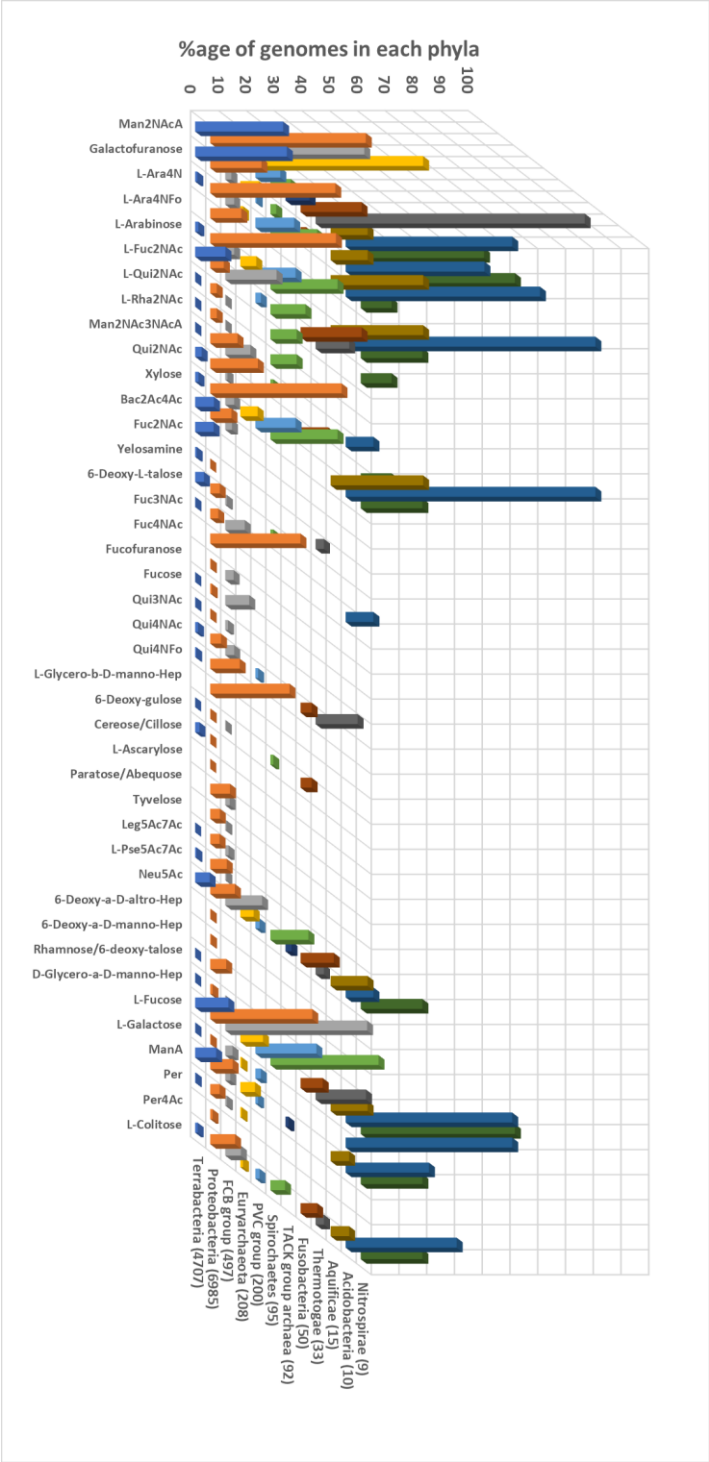


Figure 7 Prevalence of Intermediate and Rare group monosaccharides in different phyla. Data for phyla with less than five sequenced genomes are not shown to avoid visual clutter. Only names of monosaccharides are used for annotation even though all are biosynthesized as nucleotide sugars. Abbreviated names are used for some of the monosaccharides. Full names of these are given in Supplementary_data.xlsx:Worksheet4.

Why some eubacteria do not biosynthesize any monosaccharide? None of the monosaccharides are biosynthesized by some mollicutes (e.g., Mycoplasma) and endosymbionts (e.g., *Ehlrichia sp.* and *Orientia sp.*) because the biosynthesis pathways are completely absent. Mollicutes lack cell wall (Trachtenberg 1998) which could explain the absence of monosaccharides. Endosymbionts have reduced genomes which is seen as an adaptation to host dependence (Khachane et al. 2007) (Amano et al. 1987). Biosynthesis pathway enzymes are lost / are being lost as part of the phenomenon of genome reduction. This is illustrated by the endosymbiont *Buchnera aphidicola*: 13 of the 25 strains have the pathway for the biosynthesis of UDP-Glc2NAc, 7 have partial pathway and 5 do not encode any gene of this pathway. Pathway for none of the other monosaccharides are found in this organism. Pathways are incomplete i.e., enzymes catalysing one or more steps of the pathway are absent in some organisms. Some species of Mycoplasma, Ureaplasma and Spiroplasma lack mannose-1-phosphate guanylyltransferase because of which GDP-mannose is not biosynthesized. GlmU which converts Glc2N-1-phosphate to UDP-Glc2NAc is absent in *Chlamydia sp.* However, Glc2N is found in the LPS of *Chlamydia trachomatis* (Rund et al. 1999). Whether this is indicative of the presence of a transferase which uses Glc2N-1-phosphate instead of UDP-Glc2N needs to be explored.

Do *Rickettsia sp.* and *Chlamydia sp.* source monosaccharides from their host? *Rickettsia sp.* (60 strains), *Orientia tsutsugamushi* (7 strains), and *Chlamydia sp.* (143 strains) are obligate intracellular bacteria. *O. tsutsugamushi* does not contain pathways for the biosynthesis of any of the monosaccharides. This is in consonance with the finding that it does not contain extracellular polysaccharides (Amano et al. 1987). Rickettsia species have pathways for the biosynthesis of Man2NAc, L-Qui2NAc and L-Rha2NAc. L-Rha2NAc is the immediate precursor for L-Qui2NAc (Figure S2e). Rickettsia are known to use Man2NAc and L-Qui2NAc but not L-Rha2NAc (Peturova et al. 2015) implying that UDP-L-Rha2NAc is just an intermediate in these organisms. The pathway for the biosynthesis of UDP-Glc2NAc, precursor for these Hex2NAcs, is absent suggesting partial dependence on host (human). Notably, genes for the biosynthesis of Man2NAc and L-Qui2NAc have so far not been reported in humans, which explains why Rickettsia have retained these pathways (human genome was scanned and these pathways are not

found; Jaya Srivastava, Petety V. Balaji, unpublished data). Both Rickettsia and Orientia belong to the same order, Rickettsiales. Symptoms caused by these two are similar (Theunissen et al. 2017). In spite of similarities in host preference and pathogenicity, *Rickettsia sp.* continues to use certain monosaccharides while diverging from *O. tsutsugamushi* (Tamura et al. 1995) which uses none. Is this because Rickettsia use ticks as vectors whereas Orientia use mites (Fuxelius et al. 2007)? *Rickettsia akari*, the only Rickettsial species which uses mites as vectors and contains pathways for Man2NAc and L-Qui2NAc biosynthesis, has been proposed to be placed as a separate group because its genotypic and phenotypic characteristics are intermediate to those of Orientia and Rickettsia (Fuxelius et al. 2007).

Absence of Glc2NAc in organisms other than endosymbionts: UDP-Glc2NAc is the precursor for the biosynthesis of several monosaccharides (Figures S2e, S2f). However, pathways for its biosynthesis are absent in ~10% of the genomes excluding endosymbionts. None of the organisms in FCB group and Spirochaetes contain this monosaccharide. Further analysis revealed the loss of first (GlmS) or last (GlmU) enzyme of the pathway in several of their genomes. This pattern suggests that organisms of this phyla are in the process of losing UDP-Glc2NAc pathway. Incidentally, some of these genomes do contain its derivatives. They include host-associated organisms such as *Bacteriodes fragilis, Flavobacterium sp., Tannerella forsythia, Akkermansia muciniphila, Bifidobacterium bifidum, Leptospira interrogans*, etc., suggesting that they obtain Glc2NAc from their microenvironment. However, a few free-living organisms which contain derivatives of UDP-Glc2NAc but not UDP-Glc2NAc were also identified. For instance, GlmU is not present in *Arcticibacterium luteifluviistationis* (arctic surface seawater) and its C-terminus (acetyltransferase domain) is absent in *Chlorobaculum limnaeum* (freshwater). Nonetheless, both organisms contain the UDP-L-Qui2NAc pathway cluster.

Prevalence of enantiomeric pairs and isomers of N-acetyl derivatives: Both enantiomers of a few monosaccharides are reported in natural glycans. However, the present analysis shows that both enantiomers are found in only a small number of organisms, that too in specific genera, class or phyla (Table 2). Three isomeric N-acetyl derivatives of fucosamine (6-deoxygalactosamine) and of quinovosamine (6-deoxyglucosamine) are found in living systems. The N-acetyl group is present at C2, C3 or C4 position in these isomers. Only few organisms use more than one of these three isomers (Table 2). One such organism is *E. coli* NCTC11151 which contains both Fuc4NAc and Fuc3NAc. In contrast, *E. coli* O177:H21 uses L-Fuc2NAc along with Fuc3NAc. Genomic context analysis showed that Fuc4NAc biosynthesis genes are part of the O-antigen cluster in both these strains. On the other hand, genes for the biosynthesis of Fuc3NAc (in NCTC11151) and L-Fuc2NAc (in O177:H21) are present as part of the colanic acid cluster. Four genomes (strains) of *Pseudomonas orientalis* use Qui4NAc, Qui2NAc and L-Qui2NAc; genes required for the biosynthesis of these three monosaccharides are all in the same genomic neighbourhood.

Table 2 Presence of enantiomeric pairs and isomeric N-acetyl derivative pairs§

Monosaccharide	Where present	Number of organisms (genomes) in which these monosaccharide pairs are used
Both D- and L-enantiomers		
Galactose	Extremophiles	33
Fucose	Phylum FCB group Phylum Deferribacteres Class Gammaproteobacteria	7 1 10
Rhamnose	Genus Pseudomonas	338
6-Deoxytalose	Genus Pseudomonas	140
Qui2NAc†	Phylum Proteobacteria	64
Fuc2NAc	Genus Staphylococcus¶	300
Isomers of N-acetyl derivatives		
Fuc3NAc and Fuc4NAc‡	Family Enterobacteriaceae	33
Qui2NAc, Qui4NAc	Several phyla	193

§ The diastereomeric pair of Fuc4NAc and L-Fuc2NAc are found in some strains of *Escherichia coli*.

¶ Both Fuc2NAc and L-Fuc2NAc are components of capsular polysaccharides (Jones 2005)

† Abbreviated names are used for some of the monosaccharides. Full names of these are given in Supplementary_data.xlsx:Worksheet4.

‡ Glucose-1-phospate is the precursor for both Fuc3NAc and Fuc4NAc and UDP-Glc2NAc is the precursor for Fuc2NAc (the isomer that is absent in these organisms).

Why are some pathways not found in Archaea? Pathways for the biosynthesis of GDP-mannose, UDP-Glc2NAc and (d)TDP-L-rhamnose are partial in some archaeal species, particularly those belonging to the TACK group. For instance, (d)TDP-L-rhamnose biosynthesis pathway in *Saccharolobus sp.*, *Desulfurococcus sp.* and *Sulfolobus sp*. is incomplete due to the absence of (d)TDP-glucose 4,6-dehydratase (RmlB). Analysing the genomic context of other enzymes of these pathways revealed the presence of RmlB which is not identified by the HMM profile because of sequence divergence: RmlB sequences score (~300-350 bits) far below

the threshold (=400 bits) [profile GPE05430; Supplementary_data.xlsx:Worksheet1]. This is in contrast to other cases of absence wherein none of the proteins in the genome score even the default bit score of HMMER (i.e., 10 bits). A cursory glance at the source organisms of input sequences used for profile generation shows that only ~4% of the 789 sequences are from Archaea.

Use of more than one nucleotide derivative/alternative pathways: L-rhamnose and Qui4NAc are biosynthesized as both UDP- and (d)TDP-derivatives (Figures S2a, S2c). However, the (d)TDP-pathways are found in prokaryotes, but not the UDP-pathways. (d)TDP-6-deoxy-L-talose is biosynthesized via reduction of (d)TDP-4-keto-L-rhamnose or C4 epimerization of (d)TDP-L-rhamnose (Figure S2a). The former pathway occurs in 141 genomes belonging to multiple phyla and notably in *Pseudomonas sp.*, *Streptococcus sp.* and *Streptomyces sp*. The latter pathway is found in 255 genomes belonging to Proteobacteria and Terrabacteria, and notably in *Burkholderia sp.*, *Mycobacterium sp.* and *Xanthomonas oryzae*. N,N'-diacetyl legionaminic acid can be biosynthesized either from UDP-route or GDP-route (Figure S2f). The latter pathway is found in 93 of 96 genomes of *Campylobacter jejuni* whereas the former is found in 10 other genomes primarily belonging to Bacteriodetes/Chlorobi class.

Discussion

The importance of glycans, especially in prokaryotes, is well documented. Establishing the specific role of glycans and studying structure-function relationship is largely hindered by factors such as non-availability of high-throughput sequencing methods, inadequate information as to which genes are involved in non-template driven biosynthesis, phase variation (Lukácová et al. 2008) and microheterogeneity (Johannessen et al. 2012). In this study, completely sequenced prokaryotic genomes were searched for monosaccharide biosynthesis pathways using sequence homology-based approach. It is found that the usage of monosaccharides is not at all conserved across prokaryotes. This is in stark contrast to the alphabets of DNA and proteins which are universal. In addition, marked differences are observed even among different strains of a species. The range of monosaccharides used by an

organism seems to be influenced by environmental factors such as growth (nutrients, pH, temperature, …) and environmental (host, microbiome, …) conditions. For instance, high uronic acid content in exopolysaccharides of marine bacteria imparts anionic property which is implicated in uptake of Fe^{3+} thus promoting its bioavailability to marine phytoplankton for primary production (Hassler et al. 2011) and against degradation by microbes (Zhang et al. 2015). Thus, organisms, even at the level of strains, seem to evolve to modify their monosaccharide repertoire to increase fitness in the environmental niche that they inhabit.

Neu5Ac is found in 5% and 0.6% genomes of Alpha-proteobacteria and Actinobacteria, respectively; the bacterial carbohydrate structure database had no Neu5Ac-containing glycan from organisms belonging to this class/phylum (Herget et al. 2008). L-Rhamnose and L-fucose are found in 16% of Delta- and Epsilon-proteobacteria genomes and in 25% of actinobacteria genomes. However, very few L-rhamnose- and L-fucose-containing glycans from these classes/phyla are deposited in the database leading to the inference that these are rare sugars in this class/phylum. Thus, inferring monosaccharide usage based on an analysis of experimentally characterized glycans can at best give a partial picture.

Rare group monosaccharides are those which are found only in a few species, few genera and few phyla. Reasons for acquiring Rare group sugars can at best be speculative. For instance, Bac2Ac4Ac occurs at the reducing end of glycans N- and O-linked to proteins (Morrison and Imperiali 2014) but the presence of Bac2Ac4Ac is not mandatory for *C. jejuni* PglB, an oligosaccharyltransferase, since it can transfer glycans which have Glc2NAc, Gal2NAc or Fuc2NAc also at the reducing end (Wacker et al. 2006). Perhaps, Bac2Ac4Ac provides resistance to enzymes like PNGase F that cleave off N-glycans. L-rhamnose, Neu5Ac, L-Qui2NAc, Man2NAc and L-Ara4N are not used by *Leptospira biflexa* (a non-pathogen) but are used by *Leptospira interrogans* (a pathogen). It is tempting to infer that these monosaccharides impart virulence to the latter but analysis of monosaccharides used by *E. coli* strains belonging to multiple pathotypes (enterohemorragic, enteropathogenic, uropathogenic) did not reveal any relationship between monosaccharides and their phenotype! Tyvelose, paratose and abequose are 3,6-dideoxy sugars that belong to the Rare group. These are found primarily in

Salmonella enterica, *Yersinia pestis* and *Yersinia pseudotuberculosis*. These are present in the O-antigen of *Y. pseudotuberculosis* (Kenyon et al. 2017). *Y. pestis*, closely related to and derived from *Y. pseudotuberculosis*, lacks O-antigen (rough phenotype) due to the silencing of O-antigen cluster (Skurnik et al. 2000). *Y. enterocolitica*, also an enteric pathogen like *Y. pseudotuberculosis*, does not contain these monosaccharides. Hence the role of these 3,6-dideoxy sugars in the O-antigen of *Y. pseudotuberculosis* does not seem to be related to enteropathogenicity.

Besides answering the question of the universality of glycan alphabet, this study also has led to certain beneficial outcomes. The presence of L-rhamnose, mannose and L-Pse5ac7Ac in *B. cereus, B. mycoides* and *B. thuringeinsis* but not in *B. subtilis, B. amyloliquefaciens, B. licheniformis, B. velezensis and B. vallismortis* can be exploited towards taxonomic identification of metagenomic samples. Enzymes synthesizing monosaccharides that are exclusive to a pathogen vis-à-vis its host can be identified as potential drug targets. An illustrative example is of the non-hydrolyzing C2 epimerase: it mediates the synthesis of UDP-Man2NAc, UDP-L-Qui2NAc, UDP-L-Fuc2NAc and UDP-Man2NAc3NAc and is found in 60% of the prokaryotic genomes but not in humans (human genome was scanned for the presence of these pathways; unpublished results). It has already been reported that inhibitors of this enzyme are effective against methicillin-resistant *S. aureus* and a few other bacteria (Xu et al. 2013). Based on the prevalence of this enzyme in all other phyla, inhibitors against this enzyme would be promising broad spectrum antimicrobial therapies. As already noted (Mariño et al. 2010), knowledge of monosaccharide composition is also useful in ensuring consistency of recombinant glycoprotein therapeutics. Knowledge of biosynthesis pathways also allows cloning the entire cassette in a heterologous host for large scale production of monosaccharides for commercial and research applications.

Thus, glycans show least evolutionary conservation among these three macromolecules (Varki 2017). Owing to their virtue of endowing distinction, existence of a universal glycan alphabet is antithetical. It is perhaps due to the lack of a universal alphabet that glycan synthesis is template independent.

Materials and Methods

Methods are described in brief here; details are included as Supplementary_material. Completely sequenced genomes of 12939 prokaryotes were considered for analysis in this study; of these 303 are archaeal genomes (Figure S1). Pathways for the biosynthesis of 55 monosaccharides have been elucidated to date (Table 3, Figure S2). Homologs of enzymes that catalyse each step of these pathways were searched in completely sequenced genomes using HMM profiles or BLASTp. Presence of a homolog for each and every enzyme of the biosynthetic pathway of a monosaccharide is taken as evidence of the utilization of this monosaccharide by the organism. On the other hand, absence of a homolog for even one enzyme of the pathway is interpreted as the absence of the corresponding monosaccharide from the organism's glycan alphabet.

Table 3 Summary of the pathways for the biosynthesis of monosaccharides[a,b]

Details about the end product of biosynthesis pathways	Precursor[c]					
	Glc-1-P	Fru*f*-6-P	GDP-Man	UDP-Glc2NAc	Glc2NAc-1-P	Sed-7-P
Number of nucleotide sugars[d]	27	2	8	16	1	4
Number of monosaccharides[e]	25	2	8	16	1	4
Number of monosaccharides with different number of backbone carbon atoms						
Pentose	4	-	-	-	-	-
Hexose	21	2	8	13	-	-
Heptulose	-	-	-	-	-	4
Nonulose	-	-	-	3	1	-
Number of monosaccharides of the two enantiomeric forms[f]						
D	19	2	5	12	1	3
L	6	-	3	4	-	1
Number of monosaccharides of the two ring forms						
Pyranose	23	2	8	16	1	4
Furanose	2	-	-	-	-	-
Number of monosaccharides with different nucleotides						
ADP	-	-	-	-	-	1
CDP	7	-	-	-	-	-
CMP	-	-	-	3	1	-
GDP	-	1	8	-	-	3
(d)TDP[g]	9	-	-	-	-	-
UDP	11	1	-	13	-	-

[a] The monosaccharide L-Iduronic acid has not been considered in this study since there is no separate pathway for its biosynthesis. Dermatan sulfate epimerase-1 or -2 (DS-epi1 or DS-epi2) catalyses C5-epimerization of glucuronic acid to L-iduronic acid in chondroitin sulfate polymeric chains (Malmström et al. 2012).

[b] Enzymes catalysing one or more steps of the biosynthesis pathway are not characterized experimentally for some of the monosaccharides. Such monosaccharides were not considered in this study.
[c] Glc-6-P is the precursor for Glc-1-P (conversion catalysed by phosphoglucomutase), Fru*f*-6-P (catalysed by phosphoglucose isomerase) and Sed-7-P (formed in the non-oxidative phase of the pentose phosphate pathway). Fruf-6-P is the precursor of GDP-Man and UDP-Glc2NAc.
[d] There are two pathways for the biosynthesis of CMP-Leg5Ac7Ac, one starting from UDP-Glc2NAc and the other from Glc2NAc-1-P. Hence, the total number of nucleotide sugars will be 57 even though row sum is 58.
[e] L-Rhamnose and Qui4NAc are biosynthesized as both UDP- and (d)TDP-derivatives. Hence, the number of monosaccharides is less than the number of nucleotide sugars by 2.
[f] The prefix D is omitted for D enantiomers whereas the prefix L is explicitly mentioned for L enantiomers.
[g] No distinction is made between TDP and dTDP in this work since literature suggests that both ribo- and deoxyribo-substrates are used by enzymes, albeit with varying extents of specificity depending upon the source organism. In fact, dTDP and TDP have been used synonymously by some authors.

HMM profiles were generated using carefully curated datasets to ensure that all proteins considered for generating a profile do have the corresponding enzyme activity. More than one profile is used for some steps as for example when non-orthologous enzymes catalyse the same reaction. Each profile was given an annotation based on the enzyme activities of proteins that were used to generate the profile. Thresholds for most HMM profiles were set using ROC curves or bit-score plots. Protein sequences were used as BLASTp queries when the number of experimentally characterized sequences are inadequate to generate a profile.
The pathways for the biosynthesis of monosaccharides in various genomes as predicted in this study may be viewed at http://www.bio.iitb.ac.in/glycopathdb/.

Acknowledgments and funding information
We thank Nitesh Kumar, Ruchi Kumari, Tejas Shah and Tejas Vaidya for technical assistance on the development of GlycoPathDB. We thank Shradha Khater and Toshi Mishra for useful discussion. Jaya Srivastava is thankful to Council of Scientific and Industrial Research, Government of India for fellowship.

References

Amano K, Tamura A, Ohashi N, Urakami H, Kaya S, Fukushi K. 1987. Deficiency of peptidoglycan and lipopolysaccharide components in Rickettsia tsutsugamushi. Infect Immun. 55(9):2290–2292.

Bravo D, Silva C, Carter JA, Hoare A, Alvarez SA, Blondel CJ, Zaldívar M, Valvano MA, Contreras I. 2008. Growth-phase regulation of lipopolysaccharide O-antigen chain length influences serum resistance in serovars of Salmonella. J Med Microbiol. 57(Pt 8):938–946.

Copley SD. 2020. Evolution of new enzymes by gene duplication and divergence. FEBS J. 287(7):1262–1283.

Fuxelius H-H, Darby A, Min C-K, Cho N-H, Andersson SGE. 2007. The genomic and metabolic diversity of Rickettsia. Res Microbiol. 158(10):745–753.

Gabius H-J, Roth J. 2017. An introduction to the sugar code. Histochem Cell Biol. 147(2):111–117.

Hassler CS, Schoemann V, Nichols CM, Butler ECV, Boyd PW. 2011. Saccharides enhance iron bioavailability to Southern Ocean phytoplankton. Proc Natl Acad Sci USA. 108(3):1076–1081.

Herget S, Toukach PV, Ranzinger R, Hull WE, Knirel YA, von der Lieth C-W. 2008. Statistical analysis of the Bacterial Carbohydrate Structure Data Base (BCSDB): characteristics and diversity of bacterial carbohydrates in comparison with mammalian glycans. BMC Struct Biol. 8:35.

Johannessen C, Koomey M, Børud B. 2012. Hypomorphic glycosyltransferase alleles and recoding at contingency loci influence glycan microheterogeneity in the protein glycosylation system of Neisseria species. J Bacteriol. 194(18):5034–5043.

Jones C. 2005. Revised structures for the capsular polysaccharides from Staphylococcus aureus Types 5 and 8, components of novel glycoconjugate vaccines. Carbohydr Res. 340(6):1097–1106.

Kaltner H, Abad-Rodríguez J, Corfield AP, Kopitz J, Gabius H-J. 2019. The sugar code: letters and vocabulary, writers, editors and readers and biosignificance of functional glycan-lectin pairing. Biochem J. 476(18):2623–2655.

Kalynych S, Morona R, Cygler M. 2014. Progress in understanding the assembly process of bacterial O-antigen. FEMS Microbiol Rev. 38(5):1048–1065.

Keinhörster D, George SE, Weidenmaier C, Wolz C. 2019. Function and regulation of Staphylococcus aureus wall teichoic acids and capsular polysaccharides. Int J Med Microbiol. 309(6):151333.

Kenyon JJ, Cunneen MM, Reeves PR. 2017. Genetics and evolution of Yersinia pseudotuberculosis O-specific polysaccharides: a novel pattern of O-antigen diversity. FEMS Microbiol Rev. 41(2):200–217.

Khachane AN, Timmis KN, Martins dos Santos VAP. 2007. Dynamics of reductive genome evolution in mitochondria and obligate intracellular microbes. Mol Biol Evol. 24(2):449–456.

Lukácová M, Barák I, Kazár J. 2008. Role of structural variations of polysaccharide antigens in the pathogenicity of Gram-negative bacteria. Clin Microbiol Infect. 14(3):200–206.

Malmström A, Bartolini B, Thelin MA, Pacheco B, Maccarana M. 2012. Iduronic acid in chondroitin/dermatan sulfate: biosynthesis and biological function. J Histochem Cytochem. 60(12):916–925.

Mariño K, Bones J, Kattla JJ, Rudd PM. 2010. A systematic approach to protein glycosylation analysis: a path through the maze. Nat Chem Biol. 6(10):713–723.

Morrison MJ, Imperiali B. 2014. The renaissance of bacillosamine and its derivatives: pathway characterization and implications in pathogenicity. Biochemistry. 53(4):624–638.

Mostowy RJ, Croucher NJ, De Maio N, Chewapreecha C, Salter SJ, Turner P, Aanensen DM, Bentley SD, Didelot X, Fraser C. 2017. Pneumococcal Capsule Synthesis Locus cps as Evolutionary Hotspot with Potential to Generate Novel Serotypes by Recombination. Mol Biol Evol. 34(10):2537–2554.

Mostowy RJ, Holt KE. 2018. Diversity-Generating Machines: Genetics of Bacterial Sugar-Coating. Trends Microbiol. 26(12):1008–1021.

Ohno S. 2013. Evolution by Gene Duplication. Springer Science & Business Media.

Peturova M, Vitiazeva V, Toman R. 2015. Structural features of the O-antigen of Rickettsia typhi, the etiological agent of endemic typhus. Acta Virol. 59(3):228–233.

Pfannkuch L, Hurwitz R, Traulsen J, Sigulla J, Poeschke M, Matzner L, Kosma P, Schmid M, Meyer TF. 2019. ADP heptose, a novel pathogen-associated molecular pattern identified in Helicobacter pylori. FASEB J. 33(8):9087–9099.

Rund S, Lindner B, Brade H, Holst O. 1999. Structural analysis of the lipopolysaccharide from Chlamydia trachomatis serotype L2. J Biol Chem. 274(24):16819–16824.

Skurnik M, Peippo A, Ervelä E. 2000. Characterization of the O-antigen gene clusters of Yersinia pseudotuberculosis and the cryptic O-antigen gene cluster of Yersinia pestis shows that the plague bacillus is most closely related to and has evolved from Y. pseudotuberculosis serotype O:1b. Mol Microbiol. 37(2):316–330.

Tamura A, Ohashi N, Urakami H, Miyamura S. 1995. Classification of Rickettsia tsutsugamushi in a new genus, Orientia gen. nov., as Orientia tsutsugamushi comb. nov. Int J Syst Bacteriol. 45(3):589–591.

Theunissen C, Cnops L, Van Esbroeck M, Huits R, Bottieau E. 2017. Acute-phase diagnosis of murine and scrub typhus in Belgian travelers by polymerase chain reaction: a case report. BMC Infect Dis. 17(1):273.

Trachtenberg S. 1998. Mollicutes-wall-less bacteria with internal cytoskeletons. J Struct Biol. 124(2–3):244–256.

Varki A. 1993. Biological roles of oligosaccharides: all of the theories are correct. Glycobiology. 3(2):97–130.

Varki A. 2017. Biological roles of glycans. Glycobiology. 27(1):3–49.

Wacker M, Feldman MF, Callewaert N, Kowarik M, Clarke BR, Pohl NL, Hernandez M, Vines ED, Valvano MA, Whitfield C, et al. 2006. Substrate specificity of bacterial oligosaccharyltransferase suggests a common transfer mechanism for the bacterial and eukaryotic systems. Proc Natl Acad Sci USA. 103(18):7088–7093.

Xu Y, Brenning B, Clifford A, Vollmer D, Bearss J, Jones C, McCarthy V, Shi C, Wolfe B, Aavula B, et al. 2013. Discovery of Novel Putative Inhibitors of UDP-GlcNAc 2-Epimerase as Potent Antibacterial Agents. ACS Med Chem Lett. 4(12):1142–1147.

Zhang Z, Chen Y, Wang R, Cai R, Fu Y, Jiao N. 2015. The Fate of Marine Bacterial Exopolysaccharide in Natural Marine Microbial Communities. PLoS ONE. 10(11):e0142690.