Main Manuscript for

Is there a universal glycan alphabet?

Jaya Srivastava[1]*, Papanasamoorthy 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
*Lab 402, Department of Biosciences and Bioengineering, Indian Institute of Technology Bombay, Ph: (+91)8291530926, ORCID: 0000-0002-1657-4004
Email: jaya_srivastava@iitb.ac.in

Author Contributions

PVB designed research. JS performed research. PS assisted in designing, structuring and functioning of GlycopathDB. PVB and JS analyzed data and wrote the paper.

The authors declare no conflict of interest

This PDF file includes:

Main Text
Figures 1 to 6
Table 1

Abstract

The fact that the sizes of DNA and protein alphabets are 4 and 20 became textbook material a few decades ago. In contrast, several monosaccharides are known to constitute naturally occurring glycans but it is not clear if they constitute a universal set. Pathways for the biosynthesis of 55 monosaccharides have been delineated from a variety of organisms. We exploited this set of experimentally characterized sequences to search for their homologs in nearly 13,000 prokaryotic genomes. Our results indicate that the usage of monosaccharides is not conserved across prokaryotes, unlike those of nucleic acids and proteins. Substantial variations exist in the set of monosaccharides within phyla and genera; in fact, marked variations are found even among different strains of a species. Based on prevalence, these monosaccharides are classified into Extensive, Intermediate and Rare group. A limited set of enzymes suffice to biosynthesize the Extensive group of monosaccharides and it appears that these monosaccharides originated before the formation of the three domains of life. In contrast, Rare group monosaccharides are confined to a few species in a few phyla, suggesting that they have evolved much later in the course of evolution. Endosymbionts and mollicutes have lost these pathways altogether, as a consequence of host dependence and lack of cell wall, respectively. These variations beget questions concerning the contribution of different sets of monosaccharides to an organism's fitness.

Keywords

Glycobiology, monosaccharides, glycoinformatics, glycan alphabet

Introduction

Living beings show enormous diversity in organization, size, morphology, habitat, etc., but are unified by the use of the same set of bases and amino acids for the three processes of central dogma: replication, transcription and translation. Structurally, DNA and proteins illustrate storing distinctly different types of information: variations in sequence and number of bases have proven adequate for DNA to encode the enormous diversity seen in life forms. Proteins additionally use co-factors, oligomerization and dynamics to carry information required to perform a whole gamut of functions. Glycans may also be viewed as carriers of certain types of information since they are involved in a variety of biological processes by acting as masks, depots, traitors, etc. (1). Diversity in the primary structure of glycans arises from the sequence and number of monosaccharides, alternative isomeric linkages, branching and repeat length heterogeneity (2). Prokaryotic cell surface glycans contribute to phase and antigenic variations. Thus, glycans show least evolutionary conservation among these three macromolecules (3). Additionally, chemical complexity has made glycan sequencing non-trivial. Because of these two reasons, knowledgebase of glycans, vis-à-vis nucleic acids and proteins, is limited (3).

Natural proteins are made of only L-enantiomers of amino acids that primarily differ from each other in their side groups. Glycans are made of both D and L enantiomers. Monosaccharides show lot more variations than amino acids in terms of size (5 to 9 carbon atoms), ring type (pyranose, furanose), and type and extent of modification (deoxy, amino, N-formyl, N-acetyl, etc.). With a few minor exceptions, sizes of DNA and protein alphabets remain 4 and 20, respectively, across the three domains. What about the alphabet size of glycans?

Pathways for the biosynthesis of 50+ monosaccharides have been elucidated to date (Figures 1, S1; Table 1). Are all these monosaccharides used by all organisms? An analysis of the composition of bacterial glycans revealed differences in the usage of enantiomers, anomers, ring types, etc. (4). Is this 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., organisms do use additional monosaccharides which are awaiting discovery? How universal is the glycan alphabet? Is the size of glycan alphabet same across all organisms? This study addresses these questions with respect to prokaryotes.

HMM profiles were generated using carefully curated sets of homologs for 57 families of enzymes that catalyse various steps of 55 monosaccharides (Dataset S1). Sequences were used directly as BLASTp queries when the number of enzymes characterized experimentally is not sufficient for a HMM profile (Dataset S2). 12939 whole genome sequences corresponding to 3384 species of prokaryotes were searched for homologs of enzymes involved in the biosynthesis of monosaccharide (Figure S2). It is found that only a few monosaccharides are used across all phyla whereas a large number of monosaccharides have highly restricted distribution (Figure 1). Possible reasons of such a large variation in the alphabet size of glycans and implications of these observations are presented herein.

Results

Glycan alphabet size is not the same in prokaryotes: The number of monosaccharides used by different species is significantly different (Figure 2) and independent of proteome size (Figure S3). In fact, none of the organisms use all 55 monosaccharides: highest number of monosaccharides used by an organism is 23 [*Escherichia coli* strain 14EC033]. Just 1 or 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 ('Extensive' group), found in > 50% of the genomes (Figure 3), but none of them are used by all organisms

(Dataset S3). A limited set of enzymes suffice to biosynthesize these monosaccharides viz., C4-epimerase (Glc to Gal), C6-dehydrogenase (uronic acid), amidotransferase and GlmU (C2-NAc), C2-epimerase (Glc2NAc to Man2NAc), mutase (6-P to 1-P), isomerase (interconversion of furanose and pyranose) and nucleotidyltransferase (activation). 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 ('Intermediate' group) to enhance structural diversity 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 hence form the 'Rare' group (figure 3).

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 most organisms contain only one enantiomer. Both enantiomers are found in only a small number of organisms, that too in specific genera, class or phyla: galactose in extremophiles, fucose in Gamma-proteobacteria, FCB group and Deferribacteres, rhamnose in mostly Pseudomonas, 6-deoxytalose in Pseudomonas, Qui2NAc in Proteobacteria and Fuc2NAc in mostly Staphylococci. Even the prevalence of isomers of N-acetyl derivatives is restricted: Fuc4NAc and Fuc3NAc, both derived from glucose-1-phosphate, are found mostly in Enterobacteriaceae, none of which contain Fuc2NAc (derived from UDP-Glc2NAc). Fuc4NAc and Fuc3NAc appear to be components of O-antigen and colanic acid, respectively, in *E. coli* NCTC11151. *E. coli* O177:H21 is one of the 70 genomes that contain L-Fuc2NAc (derived from UDP-Glc2NAc) and Fuc4NAc: Fuc4NAc biosynthesis genes are part of the O-antigen cluster whereas those of L-Fuc2NAc are part of the colanic acid cluster. Thus, while the strain NCTC11151 has Fuc3NAc in the colanic acid cluster, O177:H21 has L-Fuc2NAc. L-Fuc2NAc and D-Fuc2NAc, both of which are components of capsular polysaccharide (5), are encoded by several strains of *S. aureus*. Unlike fucosamine isomers, no genomes were found to encode both Qui3NAc and Qui4NAc. However, Qui4NAc and Qui2NAc are found in 193 genomes belonging to several phyla. Four genomes of *Pseudomonas orientalis* contain Qui4NAc, Qui2NAc and L-Qui2NAc, all in the same neighbourhood. The biological implications of variations in such isomers and enantiomers need further investigation.

Prevalence of monosaccharides across Phyla: Not all sugars of the Extensive group (Figure 1) are found across all phyla whereas Neu5NAc belonging to the Rare group is found across all phyla. Neu5NAc and those in the Extensive group (except Man2NAc and Man2NAcA) have been reported to be present in eukaryotes (6) implying that they originated before the formation of the three domains. 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 4). For instance, Fuc4NAc and L-*glycero*-beta-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.

Why do some eubacteria not biosynthesize any monosaccharide? The number of monosaccharides is zero in 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 (7) which could explain the absence of monosaccharides. Endosymbionts have reduced genomes which is seen as an adaptation to host dependence (8) (9). 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* (10). Whether this is indicative of the presence of a transferase which uses Glc2N-1-phosphate instead of UDP-Glc2N needs to be explored.

Absence of Glc2NAc in organisms other than endosymbionts: UDP-Glc2NAc is the precursor for the biosynthesis of several monosaccharides (Figure S1, panels E and F). 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.

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 (9). Rickettsia species have pathways for the biosynthesis of Man2NAc, L-Qui2NAc and L-Rha2NAc. L-Rha2NAc is the immediate precursor for L-Qui2NAc (Figure S1, panel E). Rickettsia are known to use Man2NAc and L-Qui2NAc but not L-Rha2NAc (11) 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; unpublished data). Both Rickettsia and Orientia belong to the same order, Rickettsiales. Symptoms caused by these two are similar (12). In spite of similarities in host preference and pathogenicity, *Rickettsia sp.* continues to use certain monosaccharides while diverging from *O. tsutsugamushi* (13) which uses none. Is this because Rickettsia use ticks as vectors whereas Orientia use mites (14)? *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 (14).

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; Dataset S1]. 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.

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 *Escherichia 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: the maximum and minimum number of monosaccharides utilized by 71 strains of *Streptococcus pneumoniae* 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 *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. (15), 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 contribute to glycan alphabet. Clearly, a complex interplay of multiple factors determines the glycan alphabet size of an 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.

Use of more than one nucleotide derivative/alternative pathways: L-rhamnose and Qui4NAc are biosynthesized as both UDP- and (d)TDP-derivatives (Figure S1, panels A and C). However, only the (d)TDP-pathways are found in prokaryotes, 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 S1, panel A). 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*. Leg5NAc7NAc can be biosynthesized either from UDP-Glc2NAc or Glc2NAc-1-phosphate (Figure S1, panel F). The latter pathway is found in 93 of 96 genomes of *C. jejuni*. 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 (16) and microheterogeneity (17). In this study, completely sequenced prokaryotic genomes were searched for monosaccharide biosynthesis enzymes using sequence homology-based approach. Usage of monosaccharides is not conserved across prokaryotes, unlike those of nucleic acids and proteins. 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 pH and temperature of growth, nutrient media, host interaction, interactions within the biome, etc. For instance, high uronic acid content in exopolysaccharides of marine bacteria imparts anionic property which is implicated in uptake of Fe(III+) thus promoting its bioavailability to marine phytoplankton for primary production (18) and against degradation by microbes (19).

Neu5NAc is found in 5% and 0.6% genomes of Alpha-proteobacteria and Actinobacteria, respectively; the bacterial carbohydrate structure database had no Neu5NAc-containing glycan from organisms belonging to this class/phylum (4). Similarly, L-rhamnose and L-fucose are found in 16% and 25% genomes of Delta/Epsilon-proteobacteria and Actinobacteria, respectively. However, very few L-rhamnose- and L-fucose-containing glycans from these classes/phyla were

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 found only in a few species, few genera and few phyla. Reasons for acquiring Rare group sugars can at best be speculative as of now. For instance, Bac2NAc4NAc occurs at the reducing end of glycans N- and O-linked to proteins (20) but this is not mandatory for *Campylobacter jejuni* PglB, an oligosaccharyltransferase, since it can transfer glycans which have Glc2NAc, Gal2NAc or Fuc2NAc also at the reducing end (21). Perhaps, Bac2NAc4NAc provides resistance to host PNGase F-like enzymes that cleave off N-glycans. L-rhamnose, Neu5NAc, 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* (22). *Y. pestis*, closely related to and derived from Y. pseudotuberculosis, lacks O-antigen (rough phenotype) due to the silencing of O-antigen cluster (23). *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-Pse5NAc7NAc 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 *Staphylococcus aureus* and a few other bacteria (24). 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 (25), 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.

Materials and Methods

A dataset of 493 experimentally characterized enzymes was used for sequence homology-based search for homologs in 12939 completely sequenced prokaryotic genomes. Source of sequences, software and databases used, procedure used for generating HMM profiles and setting bit score thresholds, BLASTp query sequences and the corresponding similarity and coverage thresholds are all described in detail in SI Appendix. The results of genome scan which include predictions of monosaccharides as well as the biosynthesis pathway enzymes predicted in 12939 genomes is available at http://www.bio.iitb.ac.in/glycopathdb/.

Acknowledgments

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

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

2. D. Bravo, *et al.*, Growth-phase regulation of lipopolysaccharide O-antigen chain length influences serum resistance in serovars of Salmonella. *J. Med. Microbiol.* **57**, 938–946 (2008).

3. A. Varki, Biological roles of glycans. *Glycobiology* **27**, 3–49 (2017).

4. S. Herget, *et al.*, 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 (2008).

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

6. A. Varki, S. Kornfeld, “Historical Background and Overview” in *Essentials of Glycobiology*, 3rd Ed., A. Varki, *et al.*, Eds. (Cold Spring Harbor Laboratory Press, 2015) (December 21, 2019).

7. S. Trachtenberg, Mollicutes-wall-less bacteria with internal cytoskeletons. *J. Struct. Biol.* **124**, 244–256 (1998).

8. A. N. Khachane, K. N. Timmis, V. A. P. Martins dos Santos, Dynamics of reductive genome evolution in mitochondria and obligate intracellular microbes. *Mol. Biol. Evol.* **24**, 449–456 (2007).

9. K. Amano, *et al.*, Deficiency of peptidoglycan and lipopolysaccharide components in Rickettsia tsutsugamushi. *Infect. Immun.* **55**, 2290–2292 (1987).

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

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

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

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

14. H.-H. Fuxelius, A. Darby, C.-K. Min, N.-H. Cho, S. G. E. Andersson, The genomic and metabolic diversity of Rickettsia. *Res. Microbiol.* **158**, 745–753 (2007).

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

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

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

18. C. S. Hassler, V. Schoemann, C. M. Nichols, E. C. V. Butler, P. W. Boyd, Saccharides enhance iron bioavailability to Southern Ocean phytoplankton. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 1076–1081 (2011).

19. Z. Zhang, *et al.*, The Fate of Marine Bacterial Exopolysaccharide in Natural Marine Microbial Communities. *PLoS ONE* **10**, e0142690 (2015).

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

21. M. Wacker, *et al.*, Substrate specificity of bacterial oligosaccharyltransferase suggests a common transfer mechanism for the bacterial and eukaryotic systems. *Proc. Natl. Acad. Sci. U.S.A.* **103**, 7088–7093 (2006).

22. J. J. Kenyon, M. M. Cunneen, P. R. Reeves, Genetics and evolution of Yersinia pseudotuberculosis O-specific polysaccharides: a novel pattern of O-antigen diversity. *FEMS Microbiol. Rev.* **41**, 200–217 (2017).

23. M. Skurnik, A. Peippo, E. Ervelä, 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**, 316–330 (2000).

24. Y. Xu, *et al.*, Discovery of Novel Putative Inhibitors of UDP-GlcNAc 2-Epimerase as Potent Antibacterial Agents. *ACS Med Chem Lett* **4**, 1142–1147 (2013).

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

Figures and Tables

EXTENSIVE	INTERMEDIATE	RARE
"Simple" sugars	**"Simple" sugar**	**"Simple" sugar**
Glucose (Glc)	Xylose	L-Galactose
Galactose (Gal)	L-Arabinose	**Amino / N-Acetyl derivatives**
Mannose (Man)	**Furanose form**	Qui2NAc, Fuc2NAc, Per4N, Qui4NFo
C2-N-Acetyl derivatives	Galactofuranose	Qui3NAc, Fuc3NAc, Per4NAc, L-Ara4NFo
Glc2NAc	**Deoxy derivatives**	Qui4NAc, Fuc4NAc, L-Rha2NAc
Gal2NAc	L-Fucose	L-Qui2NAc, L-Fuc2NAc, Bac2NAc4NAc
Man2NAc	**C4-amino derivative**	**Uronic acid derivatives**
Uronic acid derivatives	L-Ara4N	ManA, Man2NAc3NAcA
GlcA		**Deoxy derivatives**
GalA		6-Deoxy-L-talose, Fucofuranose, Cereose, Rhamnose
Glc2NAcA		6-Deoxytalose, Fucose, Cillose
Gal2NAcA		6-Deoxygulose, Yelosamine
Man2NAcA		**Dideoxy derivatives**
Deoxy derivative		Paratose, Abequose, Tyvelose, L-Colitose, L-Ascarylose
L-Rhamnose		**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
		Neu5NAc, L-Pse5NAc7NAc, Leg5NAc7NAc

Figure 1. Classification of monosaccharides into three groups based on prevalence in prokaryotic genomes

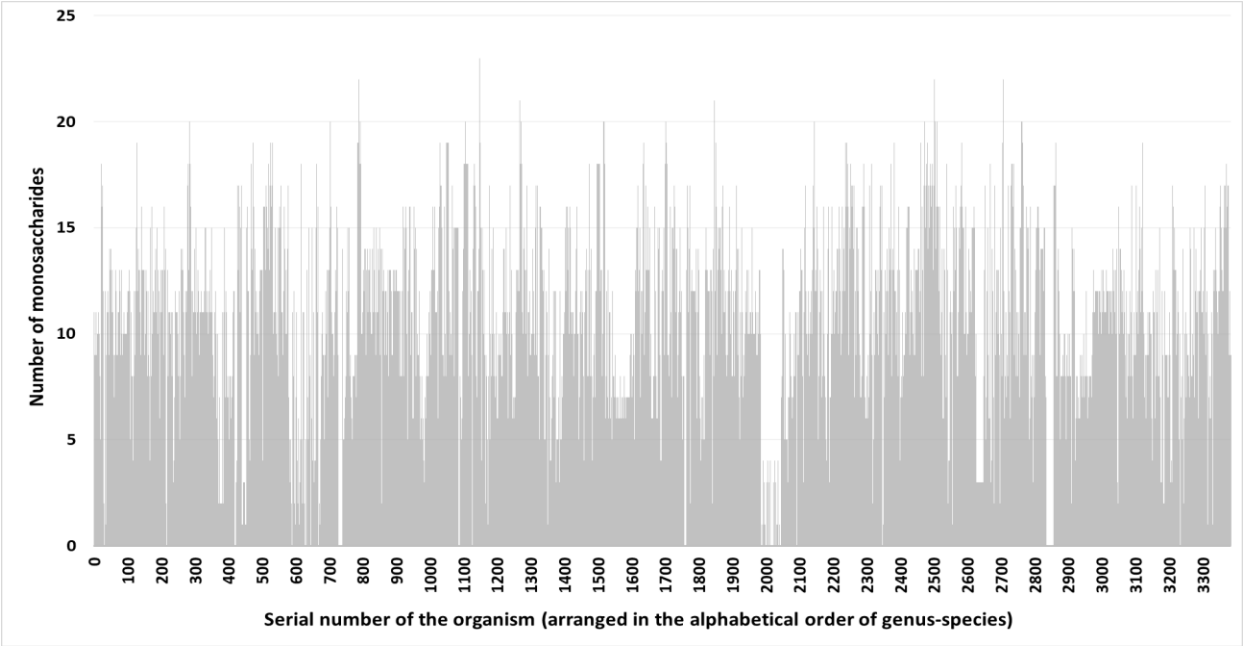

Figure 2. The number of monosaccharides for which biosynthesis pathway is found in a genome. More than one strain is sequenced for several species (figure S2). In such cases, data for the strain which has the highest number of monosaccharides is plotted. Total number of species = 3384.

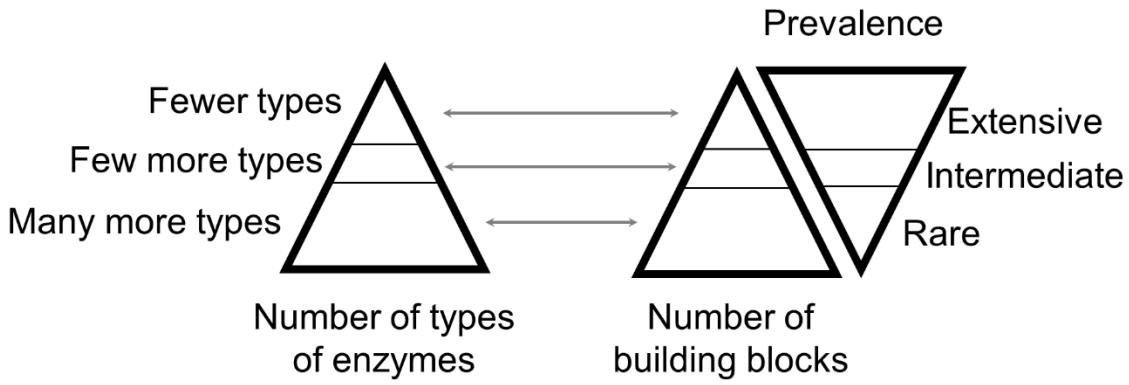

Figure 3: A schematic depiction of the prevalence of the three groups of monosaccharides viz., Extensive (found in >=50% of genomes), Intermediate and Rare (found in <=10% of genomes). Types of enzymes required for the biosynthesis corresponds to the size of the three groups.

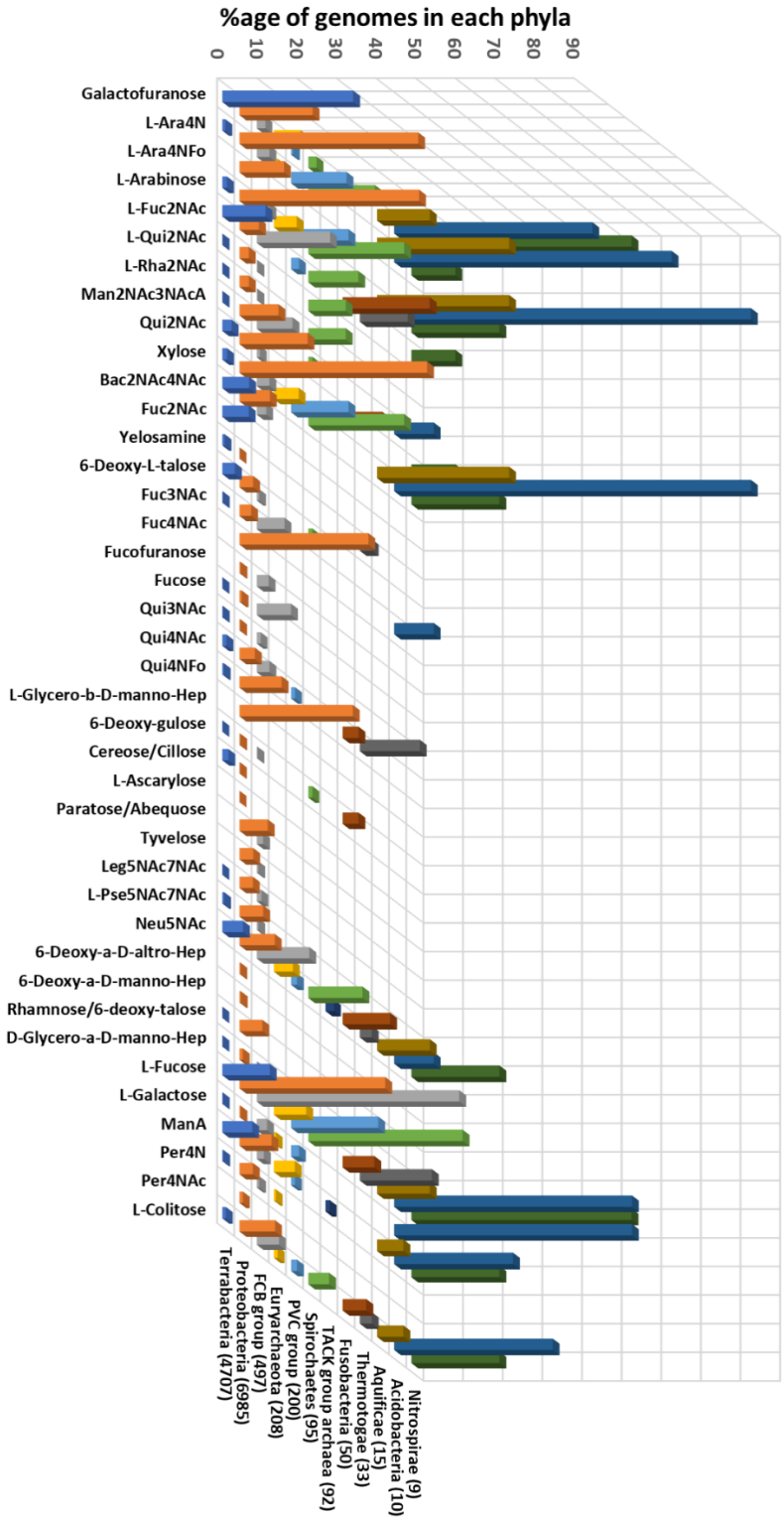

Figure 4: Distribution of Intermediate and Rare group monosaccharides in different phyla. Number of sequenced genomes in each phylum are mentioned in parenthesis. Data for phyla with less than five sequenced genomes are not shown.

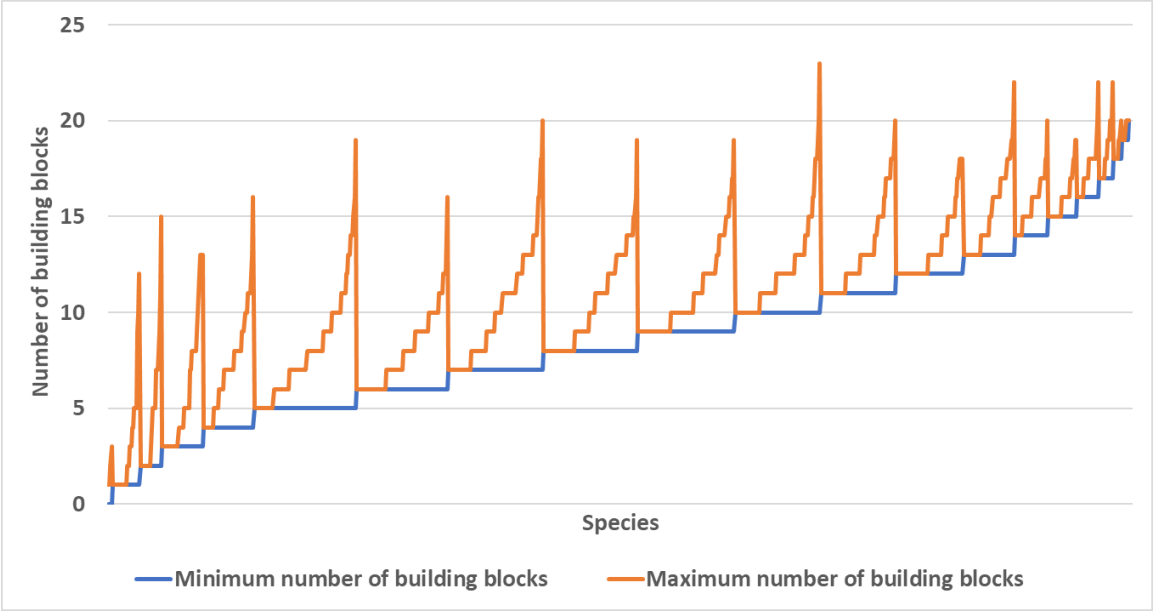

Figure 5: Variations in the number of monosaccharides among strains of a species. Species with more than one sequenced strain and at least one monosaccharide predicted in one of the strains are considered.

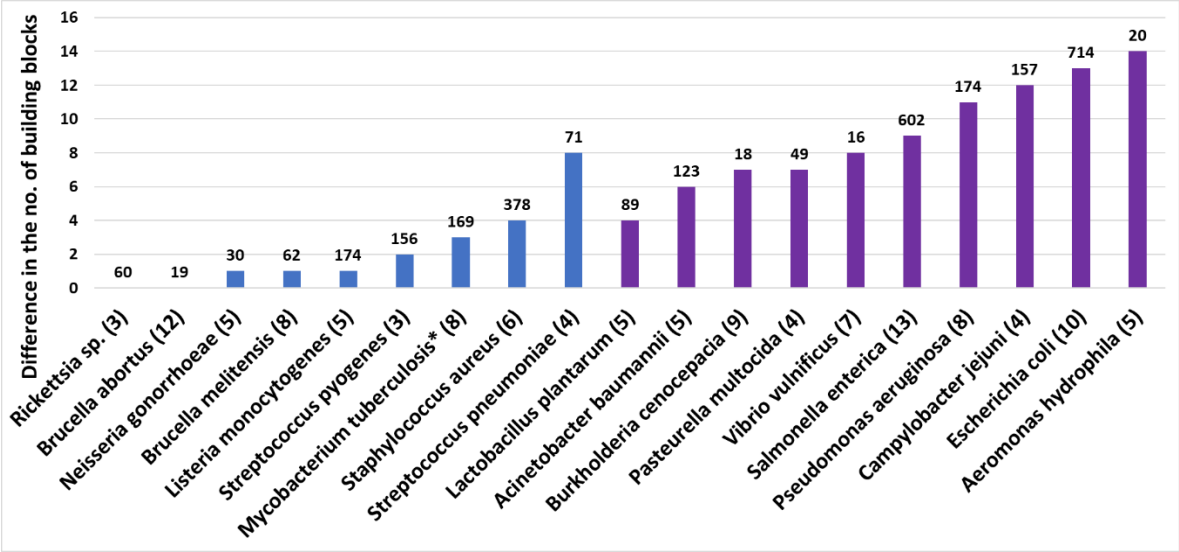

Figure 6: Difference between the maximum and minimum number of monosaccharides among strains of clinically important species. Number of sequenced strains for each organism is specified above the corresponding bar. Note that the set of monosaccharides encoded by different strains having the same number of monosaccharides may vary (minimum number encoded by an organism is given in parenthesis). Organisms associated with narrow habitat are shown in blue (including *Rickettsia sp.* and *Brucella abortus*), while those with broad habitat are shown in purple. Note: Out of 170 total strains of *M. tuberculosis,* one strain (Haarlem/NITR202) was predicted to have only two monosaccharides, while in the rest the number ranges from 8 to 11. Data for this strain was excluded in this plot as an outlier.

Table 1. Summary of pathways for the biosynthesis of monosaccharides

Details about the end product of biosynthesis pathways	Precursor[1]					
	Glc-1-P	Fru*f*-6-P	GDP-Man	UDP-Glc2NAc	Glc2NAc-1-P	Sed-7-P
Number of nucleotide sugars[2]	27	2	8	16	1	4
Number of monosaccharides[3]	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						
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[4]	9	-	-	-	-	-
UDP	11	1	-	13	-	-

[1] It is to be noted that glucose-6-phosphate is the precursor for glucose-1-phosphate (phosphoglucomutase), fructofuranose-6-phosphate (phosphoglucose isomerase) and sedoheptulose-7-phosphate (non-oxidative phase of the pentosephosphate pathway). Fructofuranose-6-phosphate is the precursor of GDP-mannose and UDP-Glc2NAc.
[2] There are two pathways for the biosynthesis of CMP-Leg5NAc7NAc, one starting from UDP-Glc2NAc and the other from Glc2NAc-1-phosphate. Hence, the total number of nucleotide sugars will be 57 even though row sum is 58.
[3] L-Rhamnose and Qui4NAc are biosynthesized as UDP- and (d)TDP-derivatives. Hence, the number of monosaccharides is less than the number of nucleotide sugars by 2.
[4] 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.