Title
Systems Biology and Pangenome of Salmonella O-Antigens.

Permalink
https://escholarship.org/uc/item/1mz9x6gp

Journal
mBio, 10(4)

ISSN
2150-7511

Authors
Seif, Yara
Monk, Jonathan M
Machado, Henrique
et al.

Publication Date
2019-08-27

DOI
10.1128/mbio.01247-19

Peer reviewed
O-antigens are made up of glycan repeat units which contain a combination of two to eight sugar residues. Variations in the final structure occur as a result of the order in which the glycans are linked, the type of linkage between residues, and the type of linkage between O units. These glycans are frequently O-antigen specific, and their biosynthesis is well studied in Salmonella, Escherichia, and Shigella (1, 2). There are 47 known O-antigen structures and 21 different O glycans in Salmonella (2). The Kauffmann-White-Le Minor serotyping scheme is based on the combination of O and H antigens expressed at the surface (3). Serogroups are defined based on the characterized O-antigen structure alone and are subdivided into serovars based on the H-antigen structure. A differentiating feature across Salmonella serogroups is the first sugar found in the O-antigen unit, which can either be an N-acetylgalcosamine/N-
acetylgalactosamine or a galactose residue. Salmonella strains falling under the latter category are the most frequently isolated and typed (2).

Most O-antigen biosynthetic genes are colocated in a genomic island known as the rfb cluster, where sequence diversity faithfully reflects the diversity of O-antigen structures. The island has dissimilar G+C content with respect to the rest of the genome, reflecting the fact that it is a hot spot for homologous recombination and horizontal gene transfer (4, 5). Currently, the sequences for 47 rfb clusters corresponding to each Salmonella serogroup have been identified, a subset of which are also found in Escherichia coli strains. The rfb cluster is a genomic locus subject to high selection pressure as is evidenced by the extensive variability of lipopolysaccharide (LPS) structures across Salmonella strains. Although genetic variability within serogroups has been reported in a few cases, no study has systematically analyzed this diversity across all serogroups. In addition, due to their genetic diversity, the O-antigen and LPS metabolic subsystems have long presented a challenge when propagating metabolic networks across strains using comparative genomics (6–8). Understanding the consequences of O-antigen sequence diversity constitutes a step forward to understanding the diversity in host-pathogen and host-phage interactions.

Here, we both assemble an O-antigen gene island pangenome and reconstruct the metabolic pathways for O-antigen biosynthesis incorporating them with a pangenome reconstruction of Salmonella metabolism (8). Genome-scale models of metabolism (GEMs) offer a powerful way to quantitatively analyze the energy and carbon demands of metabolite synthesis. We add here a much-needed functionality to the Salmonella GEM, which can also be extended to other organisms, i.e., the prediction of a strain’s serogroup and its ability to synthesize different LPS and O-antigen structures. Taking advantage of the structural similarity of O-antigens across different organisms, we show that these pathways can be exported to other Gram-negative strains.

RESULTS

O-antigen biosynthetic pathways reveal unexpected nonglycolytic extracellular nutrient sources. To date, only a subset of O-antigen precursor biosynthetic pathways have been formally described through computational models (and are available in BiGG and other metabolic databases [9], with LPSO:4 (group O:4) being the only lipopolysaccharide structure for which a full metabolic pathway description exists in Salmonella. However, there are 46 known serogroups in Salmonella, each characterized by a unique O-antigen structure, LPS structure, and rfb cluster (with the exception of serogroup O:28, which is subdivided into subserogroups O:28ab and O:28ac). To complete the O-antigen metabolic subsystem, we curated and reconstructed the biosynthetic pathways for 46 out of 47 O-antigen structures (Fig. 1A; see also Table S1 in the supplemental material). For this purpose, we conducted an extensive literature review and followed an established reconstruction protocol (10). This endeavor uncovered gaps of knowledge in O-antigen metabolism, specifically that (i) the substrate specificity and pathway order of glycosyltransferases and acetyltransferases and that (ii) the O-antigen precursor biosynthetic pathways for group O:61 have not been fully characterized. The resulting O-antigen metabolic model (Oag.v1) yielded 348 genes, 695 reactions, and 590 metabolites.

We subsequently updated the pangenomic reconstruction of Salmonella pan-STM.v1.1 (which contains all known and modeled Salmonella metabolic pathways) (8) with the content from Oag.v1. As a result, the content of pan-STM.v1.2 was 23% larger and encompassed 1,718 genes and 3,372 reactions. We observed that glucose-1-phosphate, ribulose-5-phosphate, and fructose-6-phosphate serve as common precursors to the biosynthesis of different subsets of O-antigen glycans, with CDP-ribitol being the only glycan capable of being synthesized from ribulose-5-phosphate. In addition, various extracellular nutrients which bypass the reliance on glycolytic flux can serve as precursors to O-antigen glycan biosynthesis. These include D-mannose, D-mannose-6-phosphate, D-gluconate, D-glucosamine, and D-glucosamine-6-phosphate, as well as products obtained from anhydro muropeptide recycling (which serve to
To assess the relative carbon yield of each LPS structure, we fixed the growth rate at 80% of wild-type growth on glucose plus M9 minimal medium (see Materials and Methods and Fig. 1B). The predicted yield varied as much as 56.5% across structures, with the LPS structure for serogroup O:28ab having the lowest yield and serogroup O:1,3,19 having the highest. These results reflect the stoichiometric differences between metabolic pathways. We proceeded to computationally simulate gene essentiality for the production of each LPS structure and found that, on average, 39/110002.57 genes become essential when LPS is included in the biomass objective function (11) (see Table S1 in the supplemental material).

Pan-STM.v1.2 accurately predicts the serogroup of Salmonella strains. To validate our reconstruction, we used a data set of 663 genomic sequences covering a range of 28 serogroups (Table S2). Of these, 410 are complete (gapless) and have existing metabolic reconstructions (12), and 253 are incomplete (more than one contig) but are experimentally serotyped and sequenced by the same group (13). We reconstructed or
updated strain-specific networks for all 665 strains using the approach highlighted by Seif et al. (8). We proceeded to predict each strain’s capability to synthesize any of the 46 O-antigen structures using flux balance analysis (see Materials and Methods).

As a result, serogroup predictions for 647 of 659 were accurate, amounting to an overall agreement of 98% (4 strains were not serotyped). Because of the similarity in biosynthetic pathways, strains of serogroups O:9 and O:3,10 were predicted to be able to synthesize two O-antigen structures each (structures O:9 and O:2 and structures O:1,3,19 and O:3,10, respectively). Interestingly, strain NCTC 12419 lacked wzzST, which is involved in regulating the O-antigen polymer length and could not synthesize long O-antigen polymers (14). There were four instances of misclassified strains that were annotated as serogroup O:54. rfbO54 is plasmid encoded and has been previously reported for serovar Borreze, which also harbors an inactivated chromosomal O-antigen locus (15). All four strains carried a genomic rfb cluster and the strain-specific models predicted that three strains (annotated as serovar Montevideo) could synthesize O:7, whereas the latter strain could synthesize O:4 (annotated as serovar Borreze). However, none of the plasmid sequences were available for these strains, and we could not verify O:54 biosynthesis capability. Strain GTN-01 (serogroup O:9,46) was predicted to be incapable of synthesizing any O-antigen structure. Upon further inspection, we noticed that the rfb cluster was missing eight genes essential for O-antigen metabolism.

For the remaining eight inconsistencies, we could not find any of the biosynthetic genes corresponding to the annotated serogroup in the O-antigen gene island. Instead, we found the intact rfb cluster matching that of the predicted serogroup. For example, two strains, FORC_015 and FORC_020 (annotated as serogroup O:4), were predicted to synthesize LPSO8 and LPSO6,7 and carried complete rfbO8 and rfbO6,7 clusters, respectively. However, they both lacked all of the rfbO4-specific genes. It is possible that rfbO4 was plasmid encoded in these strains; however, we did not have access to the corresponding plasmid sequences. Similar cases were observed for strains CF-SAN000624, ATCC 35640, RSK2980, 81-2806, SA20092095, and 1422-74 (Fig. 1C). We serologically tested strain CFSAN000624 experimentally, and the results agreed with the model-driven group O:4 prediction (Table 1).

A total of four strains were not annotated for a serogroup in our data set. Our strain-specific model approach predicted (i) S. bongori NCTC 12419 as serogroup O:66, (ii) S. enterica strain C629 as serogroup O:4, and (iii) S. enterica strain FORC_019 as serogroup O:9. We could not assign a serogroup for the fourth strain, however, because of a single gene deletion of nnaD (an O-acetyltransferase) from the O:48 gene island. We hypothesize that the strain (Salmonella bongori strain N268-08) synthesizes a modified O:48 antigen structure with a nonacetylated L-fucosamine residue. In general, these results show that our metabolic-model-based approach achieves high predictive accuracy while enabling a mechanistic interpretation of model predictions.

### TABLE 1 Experimental serotyping assays outcome

<table>
<thead>
<tr>
<th>Strain</th>
<th>Test type</th>
<th>Polyvalent 42-67 (81061), Salmonella</th>
<th>Monovalent O:4,5 (23839), Salmonella</th>
<th>Monovalent O:15 (48243), E. coli</th>
<th>Polyvalent A:S+Vi (73519), Salmonella</th>
</tr>
</thead>
<tbody>
<tr>
<td>E. coli K-12 MG1655</td>
<td>NC</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>S. enterica strains</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>subsp. diarizonae CFSAN044910</td>
<td>Test</td>
<td>P</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>Serovar Stanleyville CFSAN000624</td>
<td>Test</td>
<td>N</td>
<td>P</td>
<td>N</td>
<td>P</td>
</tr>
<tr>
<td>Serovar Enteritidis EC20100101</td>
<td>PC</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>P</td>
</tr>
<tr>
<td>Serovar Enteritidis SA20094177</td>
<td>PC</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>P</td>
</tr>
<tr>
<td>Serovar Bovismorbificans 3114</td>
<td>PC</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>P</td>
</tr>
</tbody>
</table>

aPC, positive control; NC, negative control.
bP, presence of seroagglutination; N, absence of seroagglutination.

O-antigen glycan biosynthesis is variably shared across several Gram-negative species. O-antigen biosynthesis is shared across Gram-negative bacteria and can be acquired through cross-species horizontal gene transfer (5, 16). We thus extended our
workflow to other Gram-negative genera for which highly curated GEMs exist, including *Escherichia* (*n* = 408), *Klebsiella* (*n* = 264), *Yersinia* (*n* = 91), and *Shigella* (*n* = 36) spp. and *Pseudomonas putida* (*n* = 24). We considered strains for which complete genomic sequences are available on PATRIC (12) and built strain-specific draft models of metabolism. Upon simulating O-antigen and O-antigen precursor biosynthesis, we found that only two glycolipids are commonly synthesized across all strains: UDP-*N*-acetyl-*d*-glucosamine and UDP-glucose (see Table S3 in the supplemental material). The former also serves as a precursor to peptidoglycan biosynthesis, while the latter is a precursor to colanic acid and capsular polysaccharide biosynthesis. Similarly, the biosynthesis of dTDP-*d*-glucose, dTDP-4-dehydro-6-deoxy-*d*-glucose, UDP-*N*-acetyl-*d*-mannosamine, and UDP-galactose was conserved in more than 97% of the strains. About 21% of *E. coli* strains synthesized UndPP-GalNAc, whereas only 8.3% of *Shigella* and 1.5% of *Salmonella* strains could. Eleven *E. coli* strains were predicted to completely synthesize and assemble *Salmonella* O-antigens, including serotypes O:35 (strains 95JB1, 95NR1, 11128, and 268-78-1), O:58 (strain DSM 103246), and O:6,14 (strains 042, C4, CF5AN061770, and UMN026). Most glycolipids (e.g., GDP-colitose, dTDP-3-acetamido-*d*-fucose, and dTDP-3,6-dideoxy-3-formamido-*d*-galactose) could only be synthesized by a reduced number of strains probably due to their specificity to the O-antigen metabolic subsystem and to a subset of O-antigen structures. Two biosynthetic genetic pathways—(i) *rmlABCD* and (ii) *manAB, gmd*, and *fcl*—were highly shared across genera (Fig. 2A).

The most discriminating capability was an ability to synthesize UndPP-galactose, ADP-ribose, dTDP-4-dehydro-6-deoxy-*l*-mannose, GDP-*d*-mannose, and dTDP-*d*-rhamnose (Fig. 2B). ADP-ribose biosynthesis was highly conserved across *Klebsiella* strains (261 of 264), and GDP-*d*-mannose biosynthesis was shared across 19 of 24 *P. putida* strains. Only *Salmonella* strains could produce UndPP-galactose (which is the first sugar in the O-antigen unit of multiple clinically relevant *Salmonella* serovars), indicating that this capability may have developed across *Salmonella* strains after species diversification. In addition, *Shigella* and *Escherichia* strains were not distinguishable, suggesting common ancestry in their O-antigen biosynthetic module. Finally, seven glycolipids were *Salmonella* specific: CDP-abequose, CDP-paratose, CDP-tyvelose, UDP-*N*-acetyl-*d*-galactosamine, UDP-*N*-acetyl-*d*-mannosamine, UDP-3-acetamido-*d*-fucose, and dTDP-3,6-dideoxy-3-formamido-*d*-galactose (Fig. 2C).

**Extending the knowledge base to over 11,000 *Salmonella* strains.** Our systems biology approach is a bottom-up approach that is highly accurate, reliable, and encompasses many attractive features, such as the power to predict a strain’s O glycan and the identification of the active metabolic pathways feeding into O-antigen biosynthesis in any user-defined nutritional environment and genetic background. However, this approach relies on high-quality genomic sequences to reach its full predictive power at the strain level. We designed an alternative approach that bypasses this requirement to assess the full genetic diversity across *Salmonella rfb* clusters. Making use of our curated and nonredundant reconstructed metabolic network, we selected one representative *rfb* cluster for each serogroup and assembled them into the “O-antigen reference database” or “rOag” (Fig. 3; see also Table S4 in the supplemental material). We linked our curated annotations to all genes across rOag, resulting in a total of 372 annotated genes.

Next, we selected a subset of signature genes per serogroup which fully distinguish the 46 serogroups (including O-antigen polymerases, *wzy* and *flippases*, and *wzx* [see Materials and Methods]). We proceeded to retrieve and quality check O-antigen gene islands for a total of 11,117 *Salmonella* genomes from PATRIC and curated the metadata available to obtain the strains’ serogroup. We found the necessary information to associate a serogroup to 5,184 strains (46.2%). We constructed the O-antigen pangenome (PanOag) by pooling together all 104,628 O-antigen genes and clustered them into 498 orthologous groups (see the supplemental material). Using the selected signature genes, we predicted the serogroups for all strains and obtained 98.5% agreement. To benchmark our approach, we ran SeqSero on our data set (13), which
yelled a 98.1% agreement (Table S5). There were 34 strains for which prediction discrepancies between SeqSero and PanOag occurred. PanOag and SeqSero made no predictions for 15 strains each. The failure cases for PanOag are discussed below and may constitute a pitfall of this approach. However, 14 of the 15 cases in which SeqSero made no predictions were predicted correctly by PanOag.

**Gene duplication, decay, and deletion contribute to intraserogroup genetic diversity and lead to loss of function.** Serogroup predictions and annotations matched well under the signature genes approach. However, when we attempted to predict a strain’s serogroup by matching the full gene content between O-antigen islands, our accuracy dropped to 94.2%. This drop was due to variations in the gene content between O-antigen islands of the same serogroup. While gene order was highly conserved, 3.2% of *Salmonella* strains were affected by either gene decay and/or gene deletion. Tandem gene duplication occurred in 706 *rfb* clusters, with serogroups O:4 and O:2 exhibiting the highest frequency of duplications. The most commonly duplicated genes in serogroup O:2 were UTP–glucose-1-phosphate uridylyltransferase...

---

**FIG 2** O-antigen glycan biosynthesis capability across Gram-negative genera. (A) Conservation of O-glycan biosynthesis genes across genera. (B) Predicted capability to synthesize O-antigen glycans across various genera: *Escherichia*, *Salmonella*, *Yersinia*, *Klebsiella*, *Pseudomonas*, and *Shigella*. The most conserved feature is the ability to synthesize UDP-N-acetylglucosamine and UDP-glucose. (C) *Salmonella*-specific O-glycan metabolic pathway. Abbreviations: Glc-1-P, glucose-1-phosphate; dTDP-D-Glc, dTDP-D-glucose; dTDP-L-Rha, dTDP-L-rhamnose; Fru-6-P, fructose-6-phosphate; d-Man-6-P, d-mannose-6-phosphate; d-Man-1-P, d-mannose-1-phosphate; GDP-D-Man, GDP-D-mannose; GDP-L-Fuc, GDP-L-fucose; Glc-1-P, glucose-1-phosphate; CDP-Abe, CDP-abequose; CDP-Par, CDP-paratose; CDP-Tyv, CDP-tyvelose; UdcDP-gal, undecaprenyl diphospho-galactose; dTDP-4d6dm, dTDP-4-dehydro-6-deoxy-L-mannose; ADP-rib, ADP-ribose; CDP-Glc, CDP-glucose.

---

July/August 2019 Volume 10 Issue 4 e01247-19 mbio.asm.org 6
and mannosyltransferase. Gene duplication events constitute one of the many mechanisms that drive the acquisition of new gene functions (17) and can lead to elevated intraspecific gene expression variation (18).

Another evolutionary route after gene duplication is the decay of one of the two copies. We found a total of 1,189 decaying genes across 612 strains (see Materials and Methods and Table S5 in the supplemental material), with the most highly represented orthologous groups being \textit{wbaU} (n/H11005 182), \textit{rmlB} (n/H11005 62), \textit{ddhD} (n/H11005 60), \textit{wzx} (n/H11005 41), \textit{manB} (n/H11005 39), and \textit{manC} (n/H11005 39) (Fig. 4A and B). The frequency of occurrence of at least one gene decay across serogroups ranged between 3.1% (group O:18) and 100% (group O:2). Of these, 458 genes (across 392 strains) both were duplicated and exhibited gene decay in at least one copy; of these, 65 exhibited gene decay in both copies. In the latter case, the gene is likely nonfunctional and may affect the strain’s ability to synthesize O-antigen macromolecules. Gene deletion affected 175 strains across 20 serogroups, in which case one of the O-antigen genes was completely missing from the O-antigen gene island (Fig. 4C). Notable, \textit{wzy} was the only gene found between \textit{galF} and \textit{gnd} in strain CFSAN000197 (serovar Bareilly). The most commonly missing functions were \textit{wbaL} (n = 74), \textit{wbaN} (n = 40), and \textit{wbaU} (n = 36). We hypothesize that
the corresponding strains exhibit the rough-type phenotype, in which the lipopolysaccharides produced and embedded in the outer membrane only consist of lipid A and core oligosaccharide but lack an O-antigen.

Of the orthologous groups in PanOag, 25.3% fall outside the orthologous groups in rOag. We found that 341 strains carried a gene family (orthologous group) in their \textit{rfb} cluster that was absent from the reference \textit{rfb} cluster for that serogroup. The majority of unexpected orthologous groups (77.8%) were not annotated with a catalytic activity. In total, 126 orthologous groups were not present in any of the reference clusters and were searched against the EggNOG database for a putative functional annotation. We assigned general metabolic categories to 74 orthologous groups; 22 of these were glycosyltransferases, 15 were transposases, and another 15 were annotated with a non-O-antigen-related metabolic function (Table S6). The rest of the orthologous groups \((n = 42)\) could not be assigned to any function or general metabolic category. We observed that insertion elements (including transposases and insertion sequences) appeared in the O-antigen loci of 184 strains spanning seven serogroups and often resulted in the disruption of genes involved in O-antigen biosynthesis such as acetyltransferases and glycosyltransferases (Fig. 4D). For example, all Minnesota strains \((n = 16, \text{ serogroup O:21})\) carried an insertion element disrupting \textit{wdaL}, an acetyltransferase. Transposons flanking genes that have yet to be annotated with a catalytic activity were distributed across other serogroups as well and may be the genetic basis for unknown/novel serovar variants.

PanOag uncovers the group O:4 B2 variant in novel serotypes, as well as an additional genotype variant for group O:4. Of the 341 strains, 184 carried at least one insertion element, with 150 of these predicted to be serogroup O:4. Transposon insertions in group O:4 occurred predominantly in the \textit{wbaV}-\textit{wbaU} region. All Schwarzengrund strains \((n = 103, \text{ serogroup O:4})\) carried three transposases, a 391-amino-acid oligosaccharide repeat unit polymerase (\textit{wzy}), and a putative protein that contained only 46 amino acids. O-antigen polymerases catalyze the stacking of O-antigen units.
(made up of three to eight glycans) onto a lipid carrier in the periplasm, generating specific linkages between the O units. Polymerases vary highly in sequence across serogroups because of their specificity to the O unit and linkage type. The polymerase sequence found across Schwarzengrund strains was homologous to \( wzy_{O9,46,27} \) (that catalyzes an \( \alpha 1-6 \) linkage type instead of the \( \alpha 1-2 \) linkage type). Strains of Bredeney (all 23), Altendorf (1), Cerro (1 of 15), Eko (all 3), and Heidelberg (103 of 104) also carried \( wzy_{O9,46,27} \) flanked by two transposases, which were absent across all other 3,865 strains of serogroup O4. The group O4 O-antigen polymerase is located outside the \( rfb \) cluster and shares no similarity with the polymerase identified here. We searched all of the strains carrying \( wzy_{O9,46,27} \) for \( wzy_{O4} \) and found it to be present in all cases. Reeves et al. made a similar observation across strains of serovars Schwarzengrund, Schleissheim, Sloterdijk, and Vellore, all of which were classified as group B2 variants and have the O27 epitope (20). Epitope O27 has been identified across strains of Bredeney and Altendorf (3), and the presence of \( wzy_{O9,46,27} \) is likely the corresponding genetic basis. However, the O27 epitope has yet to be reported across strains of Cerro, Eko, and Heidelberg, despite \( wzy_{O9,46,27} \) being carried in the majority of Heidelberg strains (except one). In addition to \( wzy_{O9,46,27} \), all three Eko strains carried the O3,10 acetyltransferase (\( wbaK \)) downstream of \( wbaP \) (21), suggesting that their O-antigen structure may be acetylated. This constitutes yet another group O4 genetic variant, here named B3.

**PanOag distinguishes non-Salmonella strains and uncovers two new Salmonella gene islands.** We could not assign a serogroup annotation to 24 \( rfb \) clusters using the signature gene-based approach. Of these, 15 clusters were missing \( wzx \) and/or \( wzy \), and many exhibited multigene deletion events, with the most extreme cases being the \( rfb \) clusters for (i) strain CFSAN000197 (serovar Bareilly), which contained only one gene (\( wzx \)) in its O-antigen island instead of the expected 7 genes, and (ii) strain 818 (serovar Cerro), with three genes deleted (\( wzy \), \( wbaB \), and \( wbaD \)) in addition to an acetyltransferase (\( wfbX \)) insertion. Closer examination of the remaining eight \( rfb \) clusters revealed that they contained multiple orthologous groups that fell outside \( rOag \). To confirm that the genomic sequences really corresponded to *Salmonella* strains, we calculated their tetranucleotide signature correlation against the *JSpecies* internal reference database (22) and found that only two of the eight strains were predicted to belong to the *Salmonella* genus while the other six were identified as either *Citrobacter* species (\( n = 5 \)) or *Raoultella* species (\( n = 1 \)) (Table S7). Disagreement between species-level classification by routine laboratory methods and classification by genomic similarity has been previously observed to occur at a frequency of 12% (23). In light of these results, we discarded the six sequences from further analysis.

The three remaining strains were NCTC10436 (subsp. *salamae*, serogroup O:44), strain CFSAN044910 and strain SA20051472 (subsp. *diarizonae*, serogroup O:59) (Table S8). Strain NCTC10436 contained 10 genes in its cluster, 5 of which did not fall under any of the annotated orthologous groups, including novel variants of \( wzx \) and \( wzy \), and three novel glycosyltransferases. A BLAST search against the nonredundant NCBI database identified close orthologs carried by *E. coli* strain ESNIH1. However, ESNIH1 did not have an annotated serogroup. We thus screened all 10 genes against the *E. coli* O-antigen database assembled by DebRoy et al. using pBLAST and found matches for all 10 genes in *E. coli* strain G1630 (group O:18ab, Fig. 5A) (24). Interestingly, the percentage of sequence homology was highest at the edges and lowest toward the center.

The \( rfb \) cluster for strains CFSAN044910 and SA20051472 contained eight genes, with only three falling outside the O-antigen knowledge base, including new variants of \( wzx \), \( wzy \), and \( wbuS \). We found orthologs for all genes in the *E. coli* strain P12b O-antigen gene island, annotated as serovar O15:H17 (Fig. 5B). However, the sequence of the genes was not conserved. Instead, the order of two syntenic blocks (\( wzx-wbuS-wzy \) and \( fnlA-fnlB-fnlC-wbuB-wbuC \)) was flipped, with a higher percentage sequence identity shared with \( fnlA-fnlB-fnlC-wbuB-wbuC \) than with \( wzx-wbuS-wzy \), suggesting that these two blocks were acquired at different times. We named these two genotypes O:68 and O:69, respectively. We acquired and tested strain CFSAN044910 for seroagglutination with *E. coli* O:15 antisera and *Salmonella* polyvalent O-antisera OMF...
Table 1. Seroagglutination occurred upon O-antiserum OMF exposure but not upon O:15 antiserum exposure. We searched the entire genomic sequences of CFSAN044910 and SA20051472 but found no matches for \textit{wdcF}, \textit{wdcG}, \textit{wdaZ}, \textit{wzxO59}, or \textit{wzyO59}. We conclude that O:69 was not expressed and that a second O-antigen island could be present in an unsequenced area of the genome or plasmid.

**DISCUSSION**

This study represents the first comprehensive attempt to elucidate the systems biology of O-antigen biosynthesis in Gram-negative bacteria, including the construc-
tion of a knowledge base and associated computational models. We use the knowledge base to predict the strain-specific capability to produce various O-antigen structures and O-antigen precursors in *Salmonella* and more than 1,400 Gram-negative strains, and as a reference database to exhaustively analyze the genetic diversity in O-antigen gene islands across *Salmonella* strains.

Genome-scale metabolic model-driven serogroup predictions covering 665 *Salmonella* strains yields 98% agreement. Failure cases in which strains carry both a genomic and a plasmid-borne *rfb* cluster highlight one of the pitfalls of *in silico* approaches. We successfully validate experimentally one of the eight remaining inconsistencies, with the model-driven prediction proving to be correct. The addition of the O-antigen metabolic subsystem yields a 23% increase in metabolic gene content in pan-STM.v1.1 (*Salmonella* pan reactome), nearly doubling the size of the accessory reactome of *Salmonella* (in = 382 reactions that are not shared across all *Salmonella* strains) (8). This result indicates that the *Salmonella* accessory reactome currently underrepresents the full metabolic diversity of the species and may still increase further in size when additional network reconstruction efforts are undertaken. Furthermore, this effort highlights the fact that metabolic reconstructions should no longer be restricted to a single strain of a species. Instead, to assess the full metabolic capabilities of a species, reconstruction endeavors should proceed at the species level, one metabolic subsystem at a time. In addition, despite being a well-studied species, our reconstruction efforts reveal knowledge gaps in *Salmonella* O-antigen biosynthesis. Finally, merging O-antigen biosynthesis to central metabolism will allow for the application of strain design approaches and serve as a mechanistic tool for designing glycoconjugate-based *Salmonella* vaccines (25).

Because the O-antigen metabolic network is converted into a mathematical format, it is easily transferable to available genome-scale reconstructions (regardless of the species). Notably, not only is O-antigen biosynthesis shared across Gram-negative bacteria, but the biosynthetic gene clusters are also often horizontally transferred across species. We constructed strain-specific metabolic models enriched with the *Salmonella* O-antigen biosynthetic pathways across five other species and simulated both O-antigen and O-antigen precursor biosynthesis capability. Unsurprisingly, a subset of *E. coli* strains share the ability to synthesize multiple *Salmonella*-specific O-antigens. The results demonstrate that O-antigen precursor biosynthesis capability can differentiate some Gram-negative species and shed some light onto the evolutionary trajectory of O-antigen biosynthesis, indicating that some O-antigen pathways existed before species diversification, while others evolved or were acquired later. Given the ubiquity of O-antigen pathways, we argue that O-antigen metabolism should be treated as a panspecies metabolic module.

To conduct a comprehensive study of the genetic diversity of *Salmonella* O-antigen gene islands we used a comparative genomics approach. We assembled a database of O-antigen gene islands, identified serogroup-specific orthologous groups, and predicted the serogroup for more than 11,000 strains of *Salmonella* (achieving an agreement with experimental metadata of 98.2%). The *rfb* cluster database reveals extensive intraserogroup genetic modifications, with evidence of both added and/or subtracted O-antigen metabolic functions with respect to known metabolic pathways. We observed novel functions (such as acetyltransferases) in strains across multiple serogroups (including serogroup B2) and loss of function due to pseudogene formation, gene deletion, or gene disruption by insertion elements. Diversification of O-antigen structures can arise from phage lysogenization (evidenced by insertion and transposon elements disrupting open reading frames in the O-antigen gene island or by the carriage of glycosyltransferase operons outside the *rfb* cluster causing phase variation) (3, 26). This phenomenon results in the induction of bacteriophage resistance in *E. coli* and *Salmonella* strains (27, 28). Alternatively, loss of function can affect the strains’ ability to synthesize an O-antigen. Although O-antigen expression is not always necessary for a strain’s survival, it does affect the strain’s ability to survive in some niches (29, 30). For example, rough *Salmonella* mutants lose the ability to penetrate the lymph
nodes, spleen, and liver and instead colonize the intestinal tract, as opposed to their isogenic counterpart (31). In total, tandem gene duplication, pseudogene formation, deletion, and transposon/prophage-mediated insertion affect the genetic composition of 994 \textit{rfb} cluster spanning 27 serogroups.

We identify two novel \textit{Salmonella} O-antigen genotypes in \textit{S. enterica} subsp. \textit{salamae} strain NCTC10436 and \textit{S. enterica} subsp. \textit{diarizonae} strain CFSAN044910. Strains of subspecies \textit{diarizonae} and \textit{salamae} are less often isolated and sequenced than strains of subspecies \textit{enterica}, suggesting that there may be other O-antigen genotypes in these and other understudied \textit{Salmonella} subspecies that have yet to be identified. Both genotypes are found across \textit{E. coli} strains, suggesting a common ancestry for the O-antigen gene islands between the two species. However, follow-up experiments reveal that one of the two genotypes may not be expressed, suggesting that the gene island is inactivated, as in serogroup O:54 (32).

Altogether, our results constitute the most comprehensive pangenome analysis of \textit{Salmonella} O-antigen gene islands, which comes as a result of combining the bottom-up metabolic reconstruction approach with the top-down pangenome approach. The resulting reconstructions and databases can be used to further study different aspects of O-antigen metabolism evolution and genotype-phenotype associations and can be combined with additional factors contributing to structural variations of O-antigens, such as the insertion of glycosyltransferase operons.

\textbf{MATERIALS AND METHODS}

\textbf{Construction of a \textit{Salmonella} O-antigen biosynthesis knowledge base.} Reference \textit{rfb} clusters were obtained from various literature sources and online databases, with the priority given to those strains which were serotyped experimentally, curated with gene annotations and sequenced by the same authors. The selected reference \textit{rfb} clusters were further curated against the reported \textit{rfb} clusters by Liu et al. (2). We henceforth refer to this collection of curated gene islands as the reference O-antigen gene cluster (Oag). We obtained sequences for 46 of 47 complete \textit{rfb} clusters but only found the sequences for \textit{wbaV} and \textit{wzy} for serogroup O:9,46,27 (33). For each serogroup, we reconstructed a full O-antigen biosynthetic pathway complete with LPS assembly and transport. While most O-antigen-specific biosynthetic genes are located within these clusters, some are present elsewhere in the genome. We used literature review to locate and identify these genes and complete the corresponding metabolic pathways. When the glycosyltransferase substrates were not known, the O-unit assembly was lumped into one reaction. The biosynthetic pathway for one of the glycans for group O:61 is not yet fully characterized and was therefore left out. Two different pathways were built for O:28ab and O:28ac (which are categorized as the same serogroup) due to significant variation in O-antigen structure and corresponding genetic island. As part of the reconstruction process, we corrected the annotation for 182 genes (\textit{tyv}), which are actually pseudogenes. The repeat unit for O:2 and O:9 are similar save for the presence of tyvelose in O:9, which is replaced by paratose in O:2 due to a frameshift mutation in \textit{tyv} (CDP-paratose 2-epimerase).

The degree of O-antigen polymerization is regulated by a combination of genes (\textit{wzzB, wzzST}, and \textit{wzzVL}) (14, 34). We therefore reconstructed three pathways for each O-antigen structure to accommodate an average of three observed polymer length distributions: short (15 O units), long (25 O units), and very long (100 O units). The chemical formula for each O polymer and downstream metabolite was calculated using mass balance analysis. We included all biosynthetic pathways in one metabolic model separately (Oag.v1), and updated the pangenome metabolic model of \textit{Salmonella} (pan-STM.v1.1) with the O-antigen biosynthesis module (pan-STM_v1.2). We adjusted the chemical formulae for 21 metabolites to correct for the 21 reactions which were not mass balanced in the benchmark model. Metabolite charges were computed using ChemAxon (35), setting the modeled pH at 7.1. We used the COBRApy package to complete all of the steps described above (36). To ensure that all pathways were complete, we simulated the capability of pan-STM_v1.2 to produce each LPS structure iteratively by setting the lower bound to the biomass function at 80% of wild-type growth on M9 plus glucose minimal medium and the optimization objective LPS export to 1. For the synthesis of each LPS structure, we ran flux-balance analysis using the model.optimize() functionality of the COBRApy package to compute the relative yield of each LPS structure (Fig. 4B), as well as the cobra.flux_analysis.find_essential_genes function to simulate gene essentiality.

\textbf{Predicting the LPS and O-glycan biosynthesis capability across \textit{Salmonella} strains.} We used three datasets to validate and test the O-antigen biosynthesis model: (i) a data set of incomplete genomic sequences for which laboratory serovar annotation exists, (ii) a previously published data set (8) consisting of 410 \textit{Salmonella} strain-specific models built from an analysis of complete genomic sequences of \textit{Salmonella}, and (iii) 1,285 complete genomic sequences for Gram-negative strains retrieved from PATRIC and quality checked. We retrieved genome-scale models of metabolism (GEMs) for \textit{E. coli} (iML1515) (37), \textit{Salmonella} (pan-STM_v1.2) (8), \textit{Yersinia} (iPC815) (38), \textit{Pseudomonas putida} (JUN1411) (39), and \textit{Klebsiella} (iYL1228) (40) strains and updated them with pathways from STM_Oag. We subsequently built strain-specific models according to a similar workflow, as previously reported (8). Briefly, homolo-
gous genes were identified via bidirectional best BLAST hit between each strain and modeled genes in the corresponding reference GEM. When a gene was not found but was essential for a reaction (e.g., did not have a modeled homolog), the reaction was removed. Gene presence or absence was thus used to tailor the reconstructed networks to each strain. The genomic sequences for the first data set was quality checked for the completeness of the rbf cluster, while those of the third data set were quality checked by excluding strains with a high ratio of unidentified nucleotide bases (>0.5%). In this analysis we did not exclude plasmids from homology searches. In fact, we screened all of the Salmonella plasmid sequences (n = 376) available on PATRIC (12) for O-antigen biosynthesis genes, but we found no matches. The biomass function of each reference model was updated to exclude O-antigen-specific metabolites (such as glycolipids). After we adjusted the networks to each strain's genetic background, the models were gap filled such that growth on M9 plus glucose minimal medium could be achieved. The process of gap filling consists of adding the minimum number reactions necessary to achieve growth (41). LPS and O-antigen glycan biosynthesis capability was tested by iteratively adding a sink reaction for each of the 46 LPS structures and O-antigen glycans, setting the objective function to 1, and simulating for growth (by running flux balance analysis [42]). Yield values exceeding 0.001 mmol/g (dry weight)/h were considered to indicate that a strain is able to synthesize the corresponding glycolipid.

Experimental serotyping assays. Strains CF5AN049410 and CF5AN000624 were acquired from the Center for Food Safety and Applied Nutrition (43, 44). Serotyping was performed by sereagglutination using commercial antisera (SSI Diagnostica, Denmark) and following the supplier’s instructions. Briefly, strains were streaked onto tryptic soy agar medium from –80°C stocks and grown overnight at 37°C. A single colony was mixed with a drop (20 μl) of antiserum for each different serum. Agglutination observed within the first 10 s after mixing was considered a positive result. E. coli K-12 MG1655 was used as a negative control because it has the rough-type phenotype. S. enterica serovar Enteritidis strain EC20100101, S. enterica serovar Enteritidis strain SA20094177, and S. enterica serovar Bovismorbificans strain 3114 were used as positive controls. Table 1 for the serotyping assay results.

Construction of an O-antigen pangenome database and rbf sequence analysis. We parsed the PATRIC database (12) to extract the rbf clusters cross strains spanning the Salmonella genus. For this purpose, we searched for the two signature genes (UPT–glucose-1-phosphate uridylyltransferase and 6-phosphogluconate dehydrogenase) known to flank the O-antigen cluster by using a keyword search of open reading frame annotated functions. Genomic regions were selected when both keywords were located on the same contig, with fewer than 35 coding DNA sequences separating them. We searched a total of 15,080 genomic sequences, which yielded 11,117 rbf clusters. Next, we assembled all sequences into a pan-O-antigen gene island database (PanOag) and proceeded to identify the set of nonredundant orthologous genes by running CD-HIT (minimum sequence identity of 0.90) (45). Briefly, CD-HIT clusters gene sequences into orthologous groups when sequence homology meets a certain threshold. Once serogroup predictions were made, we proceeded to compare the gene family content of each rbf cluster with the expected gene content for that serogroup (obtained from the curated reference genomic islands) for deletion, duplication, and insertion events. To identify decaying genes, we started by assuming that all genes annotated in the Oag.vo and rOag are functional. Then, for each annotated gene, we searched for the same gene family across serogroups as a signature orthologous groups because their genetic sequence is highly variable across serogroups as a result of their specificity to the O-antigen structure (13). In fact, they generally formed distinct orthologous groups for each serogroup. This approach alone worked to distinguish most serogroups except for O:54, O:2, O:4, O:9, O:46,37, O:9,46, O:1,3,19, O:3,10, and O:67. To distinguish O:2 from O:9 strains, we identified the strains that harbor an L2A substitution in the CDP-paratose 2-epimerase (rpy) amino acid sequence which is specific to group O:2 (46). Similarly, wbaK served to distinguish serogroup O:3,10 from O:1,3,19. Most Salmonella O-antigens are synthesized via the Wzy-dependent pathway in which an O unit is assembled in the cytoplasm, flipped across the periplasmic membrane via Wzx, polymerized in the periplasm via Wzy, and finally transported to the outer membrane. In contrast, the O-antigen biosynthetic pathway for group O:67 is ATP-binding cassette transporter dependent and therefore does involve Wzy or Wzx (2, 47). In this pathway, polymerization occurs in the cytoplasm, and the heteropolymer is subsequently transported across the membrane. As such, we chose wzt as the signature gene family for group O:67. Similarly, the O:54 O polysaccharide is a homopolymer which Keenleyside and Whitfield reported as being likely transported while being polymerized (15). We chose wbeE as the signature gene family for group O:54. We proceeded to compare our predictions with experimental observations. The serogroup was extracted from each strain’s metadata either through the strain’s name, its serotype, its antigenic formula, or its annotated serogroup (when available). To map strain names and serotypes to a serogroup, we used the White-Kauffman-Le Minor scheme reported by Institut Pasteur (3). We found sufficient metadata to assign a serogroup for 5,170 strains. After comparing our predictions with the annotations, we noticed that a subset of strains of serogroup O:4 could initially not be distinguished from strains of serogroup O:9,46,27 because a subset of strains also carried wzyO114O4627. We subsequently added obe as an additional signature gene family for group O:4, as well as a B2 and putative B3, O:68 and O:69 representative gene island in rOag. We chose strain NCTC6026 (serovar Bredeney) as a B2 representative because its O-antigen island gene content was the most common across B2 strains and because its genomic sequence is complete.
SUPPLEMENTAL MATERIAL

Supplemental material for this article may be found at https://doi.org/10.1128/mBio.01247-19.

<table>
<thead>
<tr>
<th>TABLE</th>
<th>Type</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td>CSV file</td>
<td>0.4 MB</td>
</tr>
<tr>
<td>S2</td>
<td>CSV file</td>
<td>1.3 MB</td>
</tr>
<tr>
<td>S3</td>
<td>CSV file</td>
<td>0.9 MB</td>
</tr>
<tr>
<td>S4</td>
<td>CSV file</td>
<td>0.3 MB</td>
</tr>
<tr>
<td>S5</td>
<td>CSV file</td>
<td>2.3 MB</td>
</tr>
<tr>
<td>S6</td>
<td>CSV file</td>
<td>0.03 MB</td>
</tr>
<tr>
<td>S7</td>
<td>XLSX file</td>
<td>0.1 MB</td>
</tr>
<tr>
<td>S8</td>
<td>XLSX file</td>
<td>0.01 MB</td>
</tr>
</tbody>
</table>

ACKNOWLEDGMENTS

Y.S., H.M, J.M.M., and B.O.P. were supported by NIH NIAID grant 1-U01-AI124316-01. E.K. and B.O.P. were supported by the Novo Nordisk Foundation through the Center for Biosustainability at the Technical University of Denmark (NNF10CC1016517). The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

We thank Rebecca Bell and Roberson Dwayne at the Center for Food Safety and Applied Nutrition for generously providing bacterial strains for this study and Marc Abrams for editing this article.

All data generated in this study are included in this published article (and its supplementary information files). All genomic sequences analyzed in this study are publicly available from PATRIC (supplementary information files). All genomic sequences analyzed in this study are publicly available from PATRIC (supplementary information files). All genomic sequences analyzed in this study are publicly available from PATRIC (supplementary information files). All genomic sequences analyzed in this study are publicly available from PATRIC (supplementary information files). All genomic sequences analyzed in this study are publicly available from PATRIC (supplementary information files). All genomic sequences analyzed in this study are publicly available from PATRIC (supplementary information files).

REFERENCES


