Bacterial polysaccharide synthesis and gene nomenclature

Peter R. Reeves, Matthew Hobbs, Miguel A. Valvano, Mikael Skurnik, Chris Whitfield, David Coplin, Nobuo Kido, John Klena, Duncan Maskell, Christian R.H. Raetz and Paul D. Rick

Most bacteria produce surface and/or secreted polysaccharides that can act as prominent antigens. Many of these polysaccharides are also extremely variable, as shown for Salmonella and Escherichia coli. For a review of E. coli antigenic diversity see Ref. 1, and for reviews of the biochemical and genetic basis of the variation see Refs 2–6.

Diversity within Salmonella is equivalent to that within a species such as E. coli, and many researchers place all Salmonella within S. enterica. However, the nomenclature in common use still reflects the historical allocation of species names to each serovar, for example S. typhimurium. In the context of this review, we will use the genus name Salmonella where possible. On the basis of DNA relatedness, E. coli and the four species of Shigella are a single species, and therefore we will treat strains of Shigella as serovars within the species E. coli, for example E. coli dysenteriae.

Bacterial polysaccharides include lipopolysaccharide (LPS), lipooligosaccharide (LOS) and extracellular polysaccharide (EPS) (Fig. 1). LPS is present in most Gram-negative bacteria and, characteristically, comprises three components: lipid A, core oligosaccharide and O antigen. The lipid A component is composed of sugars and fatty acids, which anchor the LPS in the outer leaflet of the outer membrane and where, in some species at least, it replaces phospholipid. The core is made of sugars and sugar derivatives, such as 3-deoxy-d-manno-octulosonic acid (Kdo). The O antigen is a polysaccharide that extends from the cell surface and consists of repeating oligosaccharide units generally composed of 3–6 sugars (O units; often repeated 10–30 fold). LOS lacks a polymeric O antigen.

LPS is highly variable, as shown for S. enterica. However, the number of structures and genes is very large, as shown for E. coli (30-fold). LOS lacks a polymeric O antigen. The lipid A component is composed of 3–6 sugars (O units; often repeated 10–30 fold). LOS lacks a polymeric O antigen.

Gene nomenclature for bacterial surface polysaccharides is complicated by the large number of structures and genes. We propose a scheme applicable to all species that distinguishes different classes of genes, provides a single name for all genes of a given function and greatly facilitates comparative studies.

E. coli and S. typhimurium have been studied, the O antigen has been found to be highly polymorphic and in general even related species have few or no O-antigen types in common. For example, the approximately 50 and 173 O antigens found in S. typhimurium and E. coli, respectively, only three are common to both species. Similarly, there are few O antigens common to E. coli and Klebsiella. The total number of types present is still likely to be many thousands. Likewise, there are 30 capsule types in E. coli. Many other species have a large repertoire of O antigens and capsules. Capsules also occur in many Gram-positive species, and EPS are also widespread.

Kenne and Lindberg’s list 103 sugars found in bacterial polysaccharides and 233 structures. Many other polysaccharides have been identified antigenically but not characterized chemically. The variation is caused not only by the diversity of monosaccharide components, but also by the diversity of linkages between sugars. Further levels of variation are imparted by the addition of non-sugar moieties (such as O-acetyl residues or amino acids) and variation in the modal length of the polysaccharide chain. A comprehensive list would be extremely long; readers may wish to refer to the complex carbohydrate structural database (CCSD)9.

EPS may be present in both Gram-negative and Gram-positive bacteria and, like O antigen, is made of repeating units. EPS can be released from the cell into the environment as a slime or remain attached to the cell to form a capsule. The distinction between LPS and different forms of EPS is not always clear and the allocation of names is sometimes close to arbitrary.

Variation within and between species

In the majority of the species studied, the O antigen has been found to be highly polymorphic and in general even related species have few or no O-antigen types in common. For example, the approximately 50 and 173 O antigens found in S. typhimurium and E. coli, respectively, only three are common to both species. Similarly, there are few O antigens common to E. coli and Klebsiella. The total number of types present is still likely to be many thousands. Likewise, there are 30 capsule types in E. coli. Many other species have a large repertoire of O antigens and capsules. Capsules also occur in many Gram-positive species, and EPS are also widespread.

Kenne and Lindberg’s list 103 sugars found in bacterial polysaccharides and 233 structures. Many other polysaccharides have been identified antigenically but not characterized chemically. The variation is caused not only by the diversity of monosaccharide components, but also by the diversity of linkages between sugars. Further levels of variation are imparted by the addition of non-sugar moieties (such as O-acetyl residues or amino acids) and variation in the modal length of the polysaccharide chain. A comprehensive list would be extremely long; readers may wish to refer to the complex carbohydrate structural database (CCSD)9.
**REVIEWS**

The Demerec system of genetic nomenclature for bac-

Current nomenclature for polysaccharide genes

The Demerec system of genetic nomenclature for bac-

A proposal for renaming bacterial surface polysaccharide genes

In the case of polymorphic loci, such as rfb and eps, the Demerec nomenclature allows only 26 symbols to de-

Fig. 1. Bacterial polysaccharides. Lipopolysaccharide (LPS) is present in most Gram-negative bacteria and comprises lipid A (black), core oligosaccharide (red) and O antigen (OAg; light green). Extracellular polysaccharides (EPS) can be released from the cell into the environment as a slime (dark green) or remain attached to the cell to form a capsule (blue). The bacterial cell wall is depicted in orange. Abbreviations: OM, outer membrane; PP, periplasm; PG, peptidoglycan; CM, cytoplasmic membrane.

**TRENDS IN MICROBIOLOGY**

Vol. 4 No. 12 December 1996

496

4-dehydrorhamnose reductase, presumably because of the confusing nomenclature. Assignment was based on a misunderstanding caused by the devolution of a similar function in the original rfbD gene of E. coli K-12. However, the rfbD gene of K. pneumoniae is not related to the rfbD gene of E. coli K-12, and the assignment was based on a misunderstanding caused by the confusing nomenclature.

Clearly, we have run out of letters for the fourth letter of the rfb gene symbol, and many of the rfb sym-

It is desirable for a given gene symbol to have the same meaning in all strains within a species, and there seems considerable merit in making it possible to use the same set of symbols in any of the large number of bacterial species. We put forward a proposal that would make this possible.

A proposal for renaming bacterial surface polysaccharide genes

In the case of polymorphic loci, such as rfb and eps, the Demerec nomenclature allows only 26 symbols to de-

...
recommend that the use of rfb, cps, exo, eps, cap, amR, kps and similar gene names be abandoned, and we propose a new scheme for a bacterial polysaccharide gene nomenclature (BPGN) as follows:

1. Most genes should be given names in the form w***, which provides 17,576 gene names. All genes of any cluster should, in general, have the same first three letters, except where the function is the same as that of a gene in another cluster. In this case, the same name is used for both genes even if it means that genes with different three-letter names co-exist in one cluster. The w*** set was chosen, rather than any other set, because there are very few genes already using this group of symbols. It is proposed that all genes of any block defined by the first two letters are of the same general type. For example, all wbf** genes (wbfA, to wbfZ) will be O-antigen genes. Likewise, wbc**, wc** and wcl** can be used for capsular, exopolysaccharide and LOS genes, respectively. By choosing these names we have aimed to maintain a link between current and proposed usage. For example, wbfA replaces mtaA (Fig. 2) and wbfP replaces rfbP (Fig. 3). Many other examples are given in Figs 2 and 3. A full list of the gene names assigned at the time of submission is shown in Table 1.

2. Genes for certain families of homologous proteins involved in saccharide processing, which are common to many gene clusters, are given names of the form uc**.

3. Genes involved in the synthesis of saccharide precursors, mostly nucleotidyl sugars, have names related to the pathway, which are applied to homologous genes in all gene clusters in all species (see Fig. 4).

Ideas concerning the implementation of this scheme can be found in Box 1.

Pathway genes

Some of the sugars found in bacterial polysaccharides are present elsewhere in the cell and their biosynthetic pathways are part of general metabolism, but for the remaining sugars, the genes for biosynthesis will generally be in the gene cluster for the polysaccharide. Most of the sugars occur in more than one structure and many of the pathways have steps in common. As there are relatively few genes in each pathway and each pathway has its own three-letter symbol, the total number of genes involved should be low. The use of names related to the pathway should make it easier to associate genes and functions and to identify homologous genes. For example, genes for the dTDP-L-rhamnose pathway will be named rmlA-D. This approach has already been used for pathway genes located within region B of the kps cluster, as well as for individual genes such as galE (involved in the synthesis of UDP-galactose), which has been found within polysaccharide gene clusters.

Transf erase genes

In contrast to the number of pathway genes anticipated, the number of possible linkages, and hence of specific transferases, is very large indeed, and many gene symbols will be required if each unique function is to have a unique name. It is largely because of the number of specific transferases that we chose to use the w*** set of gene names for bacterial polysaccharide gene names. Each cluster will, in general, have one such gene for each linkage in the structure, although bifunctional transferases are known.

A given gene symbol is used for genes carrying out the same function and/or displaying obvious homology. This principle can be illustrated by some of the glycosyl transferases of Salmonella (see Fig. 3). All wbaP (rfbP) genes encode an enzyme that transfers Gal-1-P to undecaprenol-P in the initiating step of some O units, and all have sequences that are readily aligned. The gene wbaV (rfbV) from group D and group B strains encodes a tyvelose transferase and an abequose transferase, respectively, but each can carry out both functions in the presence of appropriate precursors. As they display evident homology they are both named wbaV; future studies may show that they have a considerable degree of specificity, but that they were
simply not distinguished on their ability to complement in vivo in antigen synthesis. Thus, they are given the same name for now. Finally, the 4-mannosyl transferases, which have been shown to form different linkages \([\text{Man-}(\beta 1-4)-\text{Rha}\] for \(\text{wbaO}(\text{rfbO})\), \([\text{Man-}(\alpha 1-4)-\text{Rha}\] for \(\text{wbaU}(\text{rfbU})\); \([\text{Man-}(\alpha 1-2)-\text{Man}\] for \(\text{wbaW}(\text{rfbW})\) and \([\text{Man-}(\alpha 1-3)-\text{Gal}\] for \(\text{wbaZ}(\text{rfbZ})\) and have barely detectable sequence similarity for the four genes, are each given their own names.

**Saccharide-processing genes**

The genes for saccharide processing (including export, polymerization and assembly of complex polysaccharides such as LPS) commonly occur in families of homologous genes that perform the same general function. The proposal is that all genes of a family are given the same name and that it is in the \(\text{wz}^*\) subset of \(\text{w***}\) names. This will enable ready identification of such genes. Examples are given in Figs 2 and 3.

Two genes with the same name may have specificity for different oligosaccharides and need to be distinguished by including species and/or other relevant information as a subscript (see below).

**Genome clusters containing \(\text{wzy}\) (repeat unit polymerase), \(\text{wzx}\) and \(\text{wzz}\) genes**

Most O-antigen gene clusters have genes currently known as \(\text{rfc}\), \(\text{rfbX}\) and \(\text{cld}\) or \(\text{rol}\). Each can be recognized by topological features of the encoded protein, although there is usually little sequence similarity for the proteins in the first two families. \(\text{wzy}\) genes from different sources have different specificity and the same may apply to \(\text{wzx}\) genes. The pre-existing names of each gene (\(\text{rfc}\), \(\text{rfbX}\) and \(\text{cld}\) or \(\text{rol}\)) are anomalous in some way and we suggest that they be renamed \(\text{wzy}\), \(\text{wzx}\) and \(\text{wzz}\), respectively, to be consistent with the nomenclature for other genes.

The \(\text{wzy}\) genes are generally found within the O-antigen gene cluster but were not given \(\text{rfb}\) names for historical reasons, as in the classical Salmonella strain LT2 the gene maps separately. As we propose to abandon the use of \(\text{rfb}\), \(\text{rfbX}\) now seems anomalous, and in the case of \(\text{rol}\) and \(\text{cld}\) we have two names for the same gene. A further reason for adopting new terminology is that as a \(\text{wzx}\) gene has been reported for a capsule gene cluster and, because \(\text{wcy}\) may also be found in capsule gene clusters, \(\text{rfbX}\) and \(\text{rfc}\) are inappropriate names.
Gene clusters containing ABC-type transporter genes, wzm and wzt genes

Some polysaccharide gene clusters have a pair of genes that encode proteins involved in polysaccharide export belonging to the ABC-2 subfamily of ABC-type transporters\(^*\). Included in this group are genes in clusters for the production of wzy-independent O antigens, such as orf261 and orf431 in E. coli O9 (Ref. 15), rfbH in Vibrio cholerae O1 (Ref. 16), rfbAB in K. pneumoniae O1 and O8 and Serratia marcescens O16 (Refs 17,18), and rfbDE in Yersinia enterocolitica O3 (Ref. 19). Equivalent gene pairs have also been described in several clusters responsible for capsule production and for which a common molecular origin has been

<table>
<thead>
<tr>
<th>3-letter gene names</th>
<th>4th letters of gene names</th>
<th>General description of genes(^*)**</th>
</tr>
</thead>
<tbody>
<tr>
<td>waa</td>
<td>A–C,E–I–N,P,Q,S,T,Z</td>
<td>Escherichia coli K-12 and Salmonella LT2 lipid A core (waaA was kdtA, waaM and waaN were htrB and msbB, respectively)</td>
</tr>
<tr>
<td>wbc</td>
<td>A–D</td>
<td>Salmonella groups A,B,C,D,E,F,G,S,E,Z</td>
</tr>
<tr>
<td>wbo</td>
<td>A–G</td>
<td>Yersinia pseudotuberculosis O1 antigens (incomplete information for several serovars)</td>
</tr>
<tr>
<td>wcc</td>
<td>A–M</td>
<td>E. coli K-12 colanic acid</td>
</tr>
<tr>
<td>wcd</td>
<td>B,F,G,J–N</td>
<td>E. coli and Salmonella ECA (rff and rfe genes)</td>
</tr>
<tr>
<td>wce</td>
<td>A–D</td>
<td>Salmonella group O54 serovar Borreze</td>
</tr>
<tr>
<td>wcf</td>
<td>A–D</td>
<td>Salmonella group O54 serovar Borreze</td>
</tr>
<tr>
<td>wcr</td>
<td>A–D</td>
<td>E. coli O21 (K-12) O antigen</td>
</tr>
<tr>
<td>wcs</td>
<td>A–D</td>
<td>Klebsiella pneumoniae O1 and O8 and Serratia marcescens O16 O antigens</td>
</tr>
<tr>
<td>wct</td>
<td>A–D</td>
<td>E. coli O16 (K-12) O antigen</td>
</tr>
<tr>
<td>wcv</td>
<td>A–D</td>
<td>Y. enterocolitica O8 O antigen</td>
</tr>
<tr>
<td>wcw</td>
<td>A–D</td>
<td>Y. enterocolitica O3 outer core</td>
</tr>
<tr>
<td>wcx</td>
<td>A–D</td>
<td>Y. enterocolitica O9 O antigen</td>
</tr>
<tr>
<td>wcy</td>
<td>A–D</td>
<td>Y. enterocolitica O11 O antigen</td>
</tr>
<tr>
<td>wcz</td>
<td>A–E</td>
<td>Yersinia pseudotuberculosis O antigens</td>
</tr>
</tbody>
</table>

\(^*\)Some pathway names have not yet been used (e.g. par, col and fuc) but are reserved for the uses shown in Fig. 4. References can be found in the bacterial polysaccharide gene nomenclature (BPGD; see Box 2).

\(^*\)Description of w** genes, species, serovar and saccharide type for other w** genes and pathway end products for pathway genes. Refer to box 2.

\(^*\)Abbreviations: Kdo, 3-deoxy-D-manno-octulosonic acid; ECA, enterobacterial common antigen; OM, outer membrane.

---

Table 1. Gene name allocations\(^*\)

<table>
<thead>
<tr>
<th>3-letter gene names</th>
<th>4th letters of gene names</th>
<th>General description of genes(^*)**</th>
</tr>
</thead>
<tbody>
<tr>
<td>waa</td>
<td>A–C,E–I–N,P,Q,S,T,Z</td>
<td>Escherichia coli K-12 and Salmonella LT2 lipids A core (waaA was kdtA, waaM and waaN were htrB and msbB, respectively)</td>
</tr>
<tr>
<td>wbc</td>
<td>A–D</td>
<td>Salmonella groups A,B,C,D,E,F,G,S,E,Z</td>
</tr>
<tr>
<td>wbo</td>
<td>A–G</td>
<td>Yersinia pseudotuberculosis O1 antigens (incomplete information for several serovars)</td>
</tr>
<tr>
<td>wcc</td>
<td>A–M</td>
<td>E. coli K-12 colanic acid</td>
</tr>
<tr>
<td>wcd</td>
<td>B,F,G,J–N</td>
<td>E. coli and Salmonella ECA (rff and rfe genes)</td>
</tr>
<tr>
<td>wce</td>
<td>A–D</td>
<td>Salmonella group O54 serovar Borreze</td>
</tr>
<tr>
<td>wcf</td>
<td>A–D</td>
<td>Salmonella group O54 serovar Borreze</td>
</tr>
<tr>
<td>wcr</td>
<td>A–D</td>
<td>E. coli O21 (K-12) O antigen</td>
</tr>
<tr>
<td>wcs</td>
<td>A–D</td>
<td>Klebsiella pneumoniae O1 and O8 and Serratia marcescens O16 O antigens</td>
</tr>
<tr>
<td>wct</td>
<td>A–D</td>
<td>E. coli O16 (K-12) O antigen</td>
</tr>
<tr>
<td>wcv</td>
<td>A–D</td>
<td>Y. enterocolitica O8 O antigen</td>
</tr>
<tr>
<td>wcw</td>
<td>A–D</td>
<td>Y. enterocolitica O3 outer core</td>
</tr>
<tr>
<td>wcx</td>
<td>A–D</td>
<td>Y. enterocolitica O9 O antigen</td>
</tr>
<tr>
<td>wcy</td>
<td>A–D</td>
<td>Y. enterocolitica O11 O antigen</td>
</tr>
<tr>
<td>wcz</td>
<td>A–E</td>
<td>Yersinia pseudotuberculosis O antigens</td>
</tr>
</tbody>
</table>

---

Gene clusters containing ABC-type transporter genes, wzm and wzt genes

Some polysaccharide gene clusters have a pair of genes that encode proteins involved in polysaccharide export belonging to the ABC-2 subfamily of ABC-type transporters\(^*\). Included in this group are genes in clusters for the production of wzy-independent O antigens, such as orf261 and orf431 in E. coli O9 (Ref. 15) (see Fig. 2), rfbH in Vibrio cholerae O1 (Ref. 16), rfbAB in K. pneumoniae O1 and O8 and Serratia marcescens O16 (Refs 17,18), and rfbDE in Yersinia enterocolitica O3 (Ref. 19). Equivalent gene pairs have also been described in several clusters responsible for capsule production and for which a common molecular origin has been
The subscripts need only be used

\( \text{GDP-L-Fuc} \)

REVIEWS

Rha, rhamnose; Tyv, tyvelose; ViNAc, N-acetyl viosamine (4-acetamido-4,6-dideoxy-\( \text{D}-\)glucose).

\( \text{thimidine-5'-diphosphate} \)

\( \text{Glc-1-P} \)

Abbreviations: Abe, abequose; Alt, \( \text{L}-\)altrose; Asc, ascarylose; Col, colitose; d, deoxy; dTDP, deoxy-

\( \text{N-acetyl-D-glucosamine} \)

Hep, L-glycero-\( \text{D}-\)manno-heptose; Man, mannose; ManNAc, N-acetyl-\( \text{D}-\)mannose; ManNAcA, N-acetyl-\( \text{D}-\)mannuronic acid; P, phosphate; Par, paratose; Per, perosamine; Glc, \( \text{D}-\)glucose; GlcNAc, N-acetyl-\( \text{D}-\)glucosamine; Hep, L-glycero-\( \text{D}-\)manno-heptose; Man, mannose; ManNAc, N-acetyl-\( \text{D}-\)mannose; ManNAcA, N-acetyl-\( \text{D}-\)mannuronic acid; P, phosphate; Par, paratose; Per, perosamine.

\( \text{UDP-Galp} \)

\( \text{UDP-GlcNAc} \)

\( \text{UDP-ManNAcA} \)

\( \text{UDP-ManNAc} \)

\( \text{UDP-Glf} \)

\( \text{UDP-Galp} \)

\( \text{UDP-Gly} \)

\( \text{UDP-Asc} \)

\( \text{UDP-Mann} \)

\( \text{UDP-Asc} \)

\( \text{UDP-Mann} \)

\( \text{UDP-Asc} \)

\( \text{UDP-Mann} \)

\( \text{UDP-Asc} \)

\( \text{UDP-Mann} \)

\( \text{UDP-Asc} \)

\( \text{UDP-Mann} \)

\( \text{UDP-Asc} \)

\( \text{UDP-Mann} \)

\( \text{UDP-Asc} \)

\( \text{UDP-Mann} \)

\( \text{UDP-Asc} \)

\( \text{UDP-Mann} \)

\( \text{UDP-Asc} \)

\( \text{UDP-Mann} \)

\( \text{UDP-Asc} \)

\( \text{UDP-Mann} \)

\( \text{UDP-Asc} \)

\( \text{UDP-Mann} \)

\( \text{UDP-Asc} \)

\( \text{UDP-Mann} \)

\( \text{UDP-Asc} \)

\( \text{UDP-Mann} \)

\( \text{UDP-Asc} \)

\( \text{UDP-Mann} \)

\( \text{UDP-Asc} \)

\( \text{UDP-Mann} \)

\( \text{UDP-Asc} \)

\( \text{UDP-Mann} \)

\( \text{UDP-Asc} \)

\( \text{UDP-Mann} \)

\( \text{UDP-Asc} \)

\( \text{UDP-Mann} \)

\( \text{UDP-Asc} \)

\( \text{UDP-Mann} \)

\( \text{UDP-Asc} \)

\( \text{UDP-Mann} \)

\( \text{UDP-Asc} \)

\( \text{UDP-Mann} \)

\( \text{UDP-Asc} \)

\( \text{UDP-Mann} \)

\( \text{UDP-Asc} \)

\( \text{UDP-Mann} \)

\( \text{UDP-Asc} \)

\( \text{UDP-Mann} \)

\( \text{UDP-Asc} \)

\( \text{UDP-Mann} \)

\( \text{UDP-Asc} \)

\( \text{UDP-Mann} \)

\( \text{UDP-Asc} \)

\( \text{UDP-Mann} \)

\( \text{UDP-Asc} \)

\( \text{UDP-Mann} \)

\( \text{UDP-Asc} \)

\( \text{UDP-Mann} \)

\( \text{UDP-Asc} \)

\( \text{UDP-Mann} \)

\( \text{UDP-Asc} \)

\( \text{UDP-Mann} \)

\( \text{UDP-Asc} \)

\( \text{UDP-Mann} \)

\( \text{UDP-Asc} \)

\( \text{UDP-Mann} \)

\( \text{UDP-Asc} \)

\( \text{UDP-Mann} \)

\( \text{UDP-Asc} \)

\( \text{UDP-Mann} \)

\( \text{UDP-Asc} \)

\( \text{UDP-Mann} \)

\( \text{UDP-Asc} \)

\( \text{UDP-Mann} \)

\( \text{UDP-Asc} \)

\( \text{UDP-Mann} \)

\( \text{UDP-Asc} \)

\( \text{UDP-Mann} \)

\( \text{UDP-Asc} \)

\( \text{UDP-Mann} \)

\( \text{UDP-Asc} \)

\( \text{UDP-Mann} \)

\( \text{UDP-Asc} \)

\( \text{UDP-Mann} \)

\( \text{UDP-Asc} \)

\( \text{UDP-Mann} \)

\( \text{UDP-Asc} \)

\( \text{UDP-Mann} \)

\( \text{UDP-Asc} \)

\( \text{UDP-Mann} \)

\( \text{UDP-Asc} \)

\( \text{UDP-Mann} \)

\( \text{UDP-Asc} \)

\( \text{UDP-Mann} \)

\( \text{UDP-Asc} \)

\( \text{UDP-Mann} \)

\( \text{UDP-Asc} \)

\( \text{UDP-Mann} \)

\( \text{UDP-Asc} \)

\( \text{UDP-Mann} \)

\( \text{UDP-Asc} \)

\( \text{UDP-Mann} \)

\( \text{UDP-Asc} \)

\( \text{UDP-Mann} \)

\( \text{UDP-Asc} \)

\( \text{UDP-Mann} \)

\( \text{UDP-Asc} \)

\( \text{UDP-Mann} \)

\( \text{UDP-Asc} \)

\( \text{UDP-Mann} \)

\( \text{UDP-Asc} \)

\( \text{UDP-Mann} \)

\( \text{UDP-Asc} \)

\( \text{UDP-Mann} \)

\( \text{UDP-Asc} \)

\( \text{UDP-Mann} \)
dual copies of a gene within a genome is not uncommon; several instances have already been observed. For example, in Salmonella LT2, two genes for GDP-mannose synthase are found in both the O-antigen and the colanic acid clusters. Until now they have been known as rfbB and manC. Other recent statistical considerations to avoid giving two genes the same name within a strain but these have their own drawbacks. Where necessary the genes can be distinguished by subscripts. When it is necessary to use a form that is easily handled by computers, for example in database entries, the information can be put in parentheses instead of in subscripts. As the species and strain are defined elsewhere in such entries it is only necessary to include information on polysaccharide types instead of in subscripts. As the species and form that is easily handled by computers, for example distinguishing by subscripts. When it is necessary to use a form that is easily handled by computers, for example in database entries, the information can be put in parentheses instead of in subscripts. As the species and strain are defined elsewhere in such entries it is only necessary to include information on polysaccharide types instead of in subscripts.

Box 1. Implementation of the BPGN scheme

We suggest that gene names should be allocated after discussion with whoever is running the database, currently Peter Reeves in Sydney. This suggestion is made, first, to avoid the possibility that two research groups find the same gene at about the same time and give it two different names or that the same name is taken by two different groups for different genes and, second, to provide an opportunity to discuss any problems in name allocation.

The scheme has been applied to O-antigen genes of Escherichia coli, Salmonella, Klebsiella pneumoniae, Yersinia enterocolitica, Yersinia pseudotuberculosis, LBS of Bordetella pertussis, lipid A and LPS core of E. coli and Salmonella. The O-antigen of E. coli and Salmonella and capsule gene clusters of Elwina stewardii and E. coli K12. Details, and the principles underlying name allocation, are given in a web paper released in conjunction with this paper.

The scheme could be applied to other gene clusters, such as those for the type 2 capsules of E. coli and the related clusters for capsules of Neisseria meningitidis and Haemophilus influenzae, but as yet no agreement has been reached on this. In Table 1, we have included only genes in clusters in which we are, or have been, actively involved. The BPGN proposal allows for all appropriate genes to be named, but some of those researchers approached did not wish to join in the proposal and will, for the time being, continue with the old nomenclature. Nonetheless, the 29 gene clusters listed in Table 1 include 22 for which all genes are known from sequence, and we have great confidence that the scheme will be widely adopted. Now that the BPGN is launched we hope that many others use it in naming and renaming genes. We did not find time to contact all people working in the area and concentrated on the names of genes for which the whole cluster is known. We now welcome proposals to add new gene names to the scheme, regardless of how many genes of the cluster are known, and hope that new genes will be named according to the BPGN scheme.

Box 2. Database and more information on the web

We have set up a Bacterial Polysaccharide Gene Database (BPGD) to include the new name, old names where relevant, and function where known of all genes entered into the database, together with diagrams of the gene clusters and the oligosaccharides. The database currently has all genes referred to in this paper and we aim to add all relevant genes for which there is sufficient information. The database, written in FoxPro, is available on the world wide web (http://www.angis.su.oz.au/BacPolGenes/BPGD.html). We hope the database will facilitate the distribution of information about these genes and also ease the introduction of the new scheme. Information on the submission of suggestions for new names is available at the same site. We have also put an expanded version of this paper on the web because, owing to space limitations, we have only been able to discuss some aspects of the proposal here.

**Box 2. Database and more information on the web**

We have set up a Bacterial Polysaccharide Gene Database (BPGD) to include the new name, old names where relevant, and function where known of all genes entered into the database, together with diagrams of the gene clusters and the oligosaccharides. The database currently has all genes referred to in this paper and we aim to add all relevant genes for which there is sufficient information. The database, written in FoxPro, is available on the world wide web (http://www.angis.su.oz.au/BacPolGenes/BPGD.html). We hope the database will facilitate the distribution of information about these genes and also ease the introduction of the new scheme. Information on the submission of suggestions for new names is available at the same site. We have also put an expanded version of this paper on the web because, owing to space limitations, we have only been able to discuss some aspects of the proposal here.
REVIEWS

• How far should we go in other areas to standardize gene names,
• How should duplicate genes within a genome be named?
• How do we know when two genes with a high degree of sequence
  similarity are the same gene?
• What information should a gene name convey?

and more genes are being described in large-scale sequencing?
both within and between bacterial species, especially now that more
similarity are the same gene?

here is generally adopted, it will facilitate analysis and
to nomenclature are applied. If the proposal presented
particular classes of genes and in particular groups of
different species. Increasingly, there will be a desire for,
distinctive names to genes for families of saccharide-
surface polysaccharide gene, (2) allows genes with the
have a name that is unique but identifies it as a bacterial

Benefits and drawbacks of the BPGN scheme
The BPGN scheme: (1) allows each distinctive gene
to have a name that is unique but identifies it as a bacterial
surface polysaccharide gene, (2) allows genes with the
same function to have the same name, (3) gives pathway
genomes that relate to the pathway and (4) gives
distinctive names to genes for families of saccharide-
processing proteins.
The BPGN will remove the confusion caused by the
use of different names for genes of one function and the
same name for genes of different functions. The use of
specific names for pathway and saccharide-processing
genes will make it easier to compare genes from dif-
ferent species. Increasingly, there will be a desire for,
and the possibility of, computer analysis of patterns in
particular classes of genes and in particular groups of
organisms. This will be difficult if varying approaches to
nomenclature are applied. If the proposal presented
here is generally adopted, it will facilitate analysis and
help those outside this exciting area to access the data
without a long apprenticeship in nomenclature.

There are inevitably some drawbacks. By giving the
same name to genes with the same function, all genes
in a gene cluster will not have the same three-letter
symbol. What is clear is that change is
needed because the existing terminology cannot cope.
We hope that the BPGN will be widely adopted and
that as it becomes familiar we will all reap the benefits
of unambiguous names and ease of cross-cluster and
cross-species comparisons.

Conclusions
A draft of this manuscript was discussed at the American
Society for Microbiology (ASM) meeting in Las Vegas,
NV, USA in April 1994. It has been through much dis-
cussion and many drafts since then. Everything included
in the proposal, and many alternatives not adopted,
was discussed at length. Not all decisions were easy
and there are still some reservations, in particular about
replacing rfa with acy. What is clear is that change is
needed because the existing terminology cannot cope.
We hope that the BPGN will be widely adopted and
as it becomes familiar we will all reap the benefits
of unambiguous names and ease of cross-cluster and
cross-species comparisons.

Acknowledgements
The work in each lab. was supported by grants, which are acknowledged
in our research publications. We thank the many people who com-
dmented on drafts and in particular the contributions of Mary Berlyn,
Ken Sanderson and Kevin Rudd.

References
1. Lin, H. (1994) in Escherichia coli in Domestic Animals and
Humans (Gyles, C.L., ed.), pp. 31–72, CAB International
Pathogenesis: Cellular and Molecular Biology (2nd edn)
4. Reeves, P.R. (1994) in Bacterial Cell Wall (Neuberger, A. and
van Dijken, L.M., eds), pp. 283–314, Elsevier Science
615–482
Escherichia coli in Domestic Animals and Humans (Gyles, C.L.,
ed.), pp. 437–494, CAB International
699–704

UDPG-Gal, UDP-Glc and UDP-GlcNAc are the precursors,
but as all of these are involved in other pathways
in these species, the genes for their synthesis are present
elsewhere on the chromosome. The transferase genes for
the Sal monella and E. coli dysenteriae O antigens
have been identified and named wabaN, O,W,V,R,W,Z
and wbbFQ,W,R, respectively. Only one transferase gene
(wbbH) has been identified for E. coli K-12 and the
others will be among the genes named wbbH-L. Each
transferase is unique and the details can only be found
from the literature or, more easily, from the BPGD
(see Box 2).

These O antigens are polymerized and exported by
different pathways from those in the first group of O anti-
gen discussed. The gene clusters contain acz genes,
which encode a membrane protein thought to be the
O-unit flippase, and most contain acy, the O-antigen
polymerase gene. The gene acz, which controls the
chain length of the O antigen, is present in an adjacent
location in each case. It is noteworthy that the acy
gene of Sal monella groups B and D1 is anomalous be-
cause it occurs elsewhere on the chromosome.
The examples illustrated in Figs 2 and 3 include
cases of pathway genes common to different gene clus-
ters and of homologous genes of some of the saccharide-
processing protein families. It is now easy to see these
patterns, which before were obscured by the variety
of nomenclatures used. In addition, there are several
examples of rfa names being used for genes of differ-
ent function, which will be avoided under the proposed
scheme.

The examples illustrated in Figs 2 and 3 include
cases of pathway genes common to different gene clus-
ters and of homologous genes of some of the saccharide-
processing protein families. It is now easy to see these
patterns, which before were obscured by the variety
of nomenclatures used. In addition, there are several
examples of rfa names being used for genes of differ-
ent function, which will be avoided under the proposed
scheme.

The examples illustrated in Figs 2 and 3 include
cases of pathway genes common to different gene clus-
ters and of homologous genes of some of the saccharide-
processing protein families. It is now easy to see these
patterns, which before were obscured by the variety
of nomenclatures used. In addition, there are several
examples of rfa names being used for genes of differ-
ent function, which will be avoided under the proposed
scheme.

The examples illustrated in Figs 2 and 3 include
cases of pathway genes common to different gene clus-
ters and of homologous genes of some of the saccharide-
processing protein families. It is now easy to see these
patterns, which before were obscured by the variety
of nomenclatures used. In addition, there are several
examples of rfa names being used for genes of differ-
ent function, which will be avoided under the proposed
scheme.
11 Fraser, C.M. et al. (1995) Science 270, 397–403
20 Frosch, M. et al. (1993) Mol. Microbiol. 5, 1221–1263