Genomics of *Weissella cibaria* with an examination of its metabolic traits

Kieran M. Lynch, Alan Lucid, Elke K. Arendt, Roy D. Sleator, Brigid Lucey and Aidan Coffey

1Department of Biological Sciences, Cork Institute of Technology, Rossa Avenue, Bishopstown, Cork, Ireland
2School of Food and Nutritional Sciences, University College Cork, Ireland

*Weissella* is a genus of lactic acid bacteria (LAB) consisting of species formerly included in the *Leuconostoc paramesenteroides* group. Similar to other LAB, they are commonly found in fermented foods but have also been isolated from environmental and human samples. Currently there are 20 recognized species. Herein, three *Weissella cibaria* genomes were sequenced using Illumia Mi-Seq and Roche 454 technologies. Annotation was performed using the Prokka and JGI IMG pipelines. A thorough analysis of the genomics of the *W. cibaria* strains was performed, in addition to brief comparative analyses of the genus *Weissella* as a whole. Genomic sequence data from the newly sequenced *W. cibaria* strains and data available in GenBank for other *Weissella* strains was used (*n*=10; four *Weissella cibaria*, one *Weissella ceti*, one *Weissella confusa*, one *Weissella halotolerans*, two *Weissella koreensis* and one *Weissella paramesenteroides*). The genomes had sizes varying from 1.3 to 2.4 Mb. DNA G+C contents ranged from 35 to 45 mol%. The core- and pan-proteome at genus and species levels were determined. The genus pan-proteome was found to comprise 4712 proteins. Analysis of the four *W. cibaria* genomes indicated that the core-proteome, consisting of 729 proteins, constitutes 69% of the species pan-proteome. This large core-set may explain the divergent niches in which this species has been found. In *W. cibaria*, in addition to a number of phosphotransferase systems conferring the ability to assimilate plant-associated polysaccharides, an extensive proteolytic system was identified.

INTRODUCTION

The genus *Weissella* was described by Collins *et al.* (1993) to include the *Leuconostoc paramesenteroides* group, including members of the atypical lactobacilli such as *Lactobacillus minor*, *Lactobacillus kandleri* and *Lactobacillus halotolerans*. The weissellas are Gram-positive, generally short rods with rounded or tapered ends, occurring in pairs or short chains. They are catalase-negative, facultatively anaerobic, heterofermentative bacteria, generally producing D- and L-lactate from glucose (Collins *et al.*, 1993). The sources of isolation of weissellas suggest an environmental origin. Strains have been isolated from a variety of niches including meat products (Diez *et al.*, 2009), dairy products (Van der Meulen *et al.*, 2007), fermented foods of vegetable origin (Björkroth *et al.*, 2002), sourdoughs (Galle *et al.*, 2010) and soil (Magnusson *et al.*, 2002). The species *Weissella confusa* and *Weissella cibaria* have also been found in clinical samples of human and animal origin (Björkroth *et al.*, 2002; Huys *et al.*, 2012). Production of the exopolysaccharide (EPS) dextran has been reported for several species including *W. cibaria*, *W. confusa*, *W. fabaria*, *W. ghanensis*, *W. kandleri* and *W. koreensis* (De Bruyne *et al.*, 2010; Huys *et al.*, 2012). Dextran production by these organisms has gained increasing interest in recent years, in particular for *W. confusa* and *W. cibaria*, due to the potential applications associated with this polysaccharide, for example as a replacement for commercial hydrocolloids in bakery products, coupled with its potential health benefit as prebiotic fibre (Arendt *et al.*, 2011; Ruas-Madiedo *et al.*, 2009; Zannini *et al.*, 2012).

To date, few genomic analyses have been performed on strains of this genus and, at the time of writing, sequencing data for only seven strains were available in public databases.
In addition, while there are currently 20 recognized species in the genus, the genomic data examined here represents only six species.

Our aim was to provide a genomic overview of this lesser-known genus of lactic acid bacteria LAB, with particular emphasis on *W. cibaria* due to high levels of dextran production by this species, which has been increasingly studied and applied in a number of food-related applications in recent years (Amari et al., 2013; Galle et al., 2010; Katina et al., 2009; Lynch et al., 2014; Zannini et al., 2013).

## METHODS

**Bacterial strains and DNA isolation.** Bacterial strains sequenced in this study were isolated during previous studies and stored in the Cork Institute of Technology LAB culture collection at −80 °C. For DNA isolation *W. cibaria* strains were grown in MRS medium overnight at 30 °C. Genomic DNA was isolating using an in-house phenol/chloroform extraction method.

**Sequencing strategy.** *W. cibaria* f53PR was sequenced using 454-GS-FLX+ titanium pyrosequencing technology by Roche 454 Life Sciences. Assembly of reads was performed using Newbler v2.5. Both *W. cibaria* strains MG1 and AB3b were sequenced using Illumina MiSeq technology by IMG Laboratories. Assembly of reads was performed using CLC Genomics Workbench 5.5.2. Table S1 (available in the online Supplementary Material) shows the number of reads and average genome coverage for each strain.

**Genomic data of other Weissella strains.** Genomic data of the additional *Weissella* strains examined in this study were obtained from the NCBI online database. GenBank accession numbers are shown in Table 1.

**Phylogenetic inference.** A phylogeny was created on the basis of multiple sequence alignment of a subset of available 16S rRNA data using the neighbour-joining method (Saitou & Nei, 1987) with MEGABLAST software (Tamura et al., 2013). Bootstrap analysis was carried out to test the statistical reliability of the topology of the neighbour-joining tree using 1000 bootstrap resamplings of the data. Phylogenetic analysis was also performed based on concatenation of the complete nucleotide sequences of seven housekeeping genes. These genes were *gyrB* (DNA gyrase, β subunit), *g6pd* (glucose 6-phosphate dehydrogenase), *pgm* (phosphoglucotase), *dihD* (d-Ala–d-Ala ligase), *dnaE* (DNA polymerase III, α subunit), *purK* (phosphoribosylamine-imidazole carboxylase) and *rpoB* (RNA polymerase, β subunit), and were selected according to a study of *Oenococcus oeni* (Bihlère et al., 2009).

**Genomic annotation.** Prediction of protein coding sequences and ORFs and subsequent gene annotation was done using Prokka (Seemann, 2014) and the JGI IMG (Joint Genome Initiative Integrated Microbial Genomes; Markowitz et al., 2012) automated annotation pipelines. This was done for all genomes, i.e. ORF re-prediction and re-annotation was also performed on the seven genomes obtained online. Specific genomic regions, genes or proteins of interest were manually checked and annotated using the NCBI BLAST program (Altschul et al., 1990). Accompanying supplementary data (see Table S4) give details of BLAST hits for all genes/proteins of interest in this study. Putatively identified transporters were also checked against the Transporter Classification Database (TCDB) and these results are shown in Table S5 (Saier et al., 2014).

Analysis of encoded metabolic pathways and capabilities (e.g. putative carbohydrate utilization) was performed using KAAS (KEGG Automatic Annotation Server; Moriya et al., 2007). Clusters of Orthologous Genes (COG) assignments were performed using the eggNOG v4.0 database (Powell et al., 2014). Localization of proteins (e.g. transmembrane, extracellular) was examined using Psort v3.0 (Subcellular Localization Prediction Tool; Yu et al., 2010) and InterProScan v5.0 (Jones et al., 2014). The presence of bacteriophage elements within the bacterial genomes was analysed using PHAST (Phage Search Tool; Zhou et al., 2011). CRISPRFinder (Grissa et al., 2007) was used to search the genomes for clustered regularly interspaced short palindromic repeat (CRISPR) sequences.

Visualization of genomic sequences was performed using SnapGene Viewer (GSL Biotech).

**Promoter prediction.** Prediction of promoter sequences for the dextranucrase genes was done using PPP (Groningen Biomolecular Sciences and Biotechnology Institute).

**Core and pan genomics.** The core- and pan-proteomes were predicted using PanOCT (Pan-genome Orthologue Clustering Tool; Fouts et al., 2012). Circular diagrams representing the pan-proteome (see Fig. 3c below and Fig. S1c) were constructing using CIRCOS circular visualization software (Krzywinski et al., 2009) and data from PanOCT using in-house scripts. Graphs of the predicted core- and pan-proteome size (Tettelin plots; Tettelin et al., 2005) were generated using the GET_HOMOLOGUES package (Contreras-Moreira & Vinuesa, 2013).

## RESULTS AND DISCUSSION

**General genomic features of the genus Weissella**

Online sequence data for seven *Weissella* strains (April 2014; one complete, six draft) and three *W. cibaria* strains sequenced in this study (all draft) were used to provide an overview of the genus. The ten genomes represent only six species, of which there are 20 recognized to date within the genus. This highlights the fact that little genomic sequence data are currently available for members of this genus. From the available data (summarized in Table 1) it can be seen that genomes vary in size from approximately 1.3 to 2.4 Mb. The number of predicted protein coding sequences (CDSs) in these genomes ranges from 1271 to 2349. Such variation could indicate significant gene loss or gain as a mechanism and as a consequence of niche adaptation. Similar observations have been made in the genus *Lactobacillus*, which has an even wider breadth of genomic variation (Kant et al., 2011; Makarova et al., 2006). The G+C content of the genomes varies from 35.5 to 45.1 mol%, which is within the range normally found for a defined bacterial genus. The phylogenetic relationship of members of the genus *Weissella* with respect to members of closely related LAB genera is shown in Fig. 1.

**Pan- and core-proteome**

The pan-proteome, defined as the full complement of proteins of the *Weissella* genomes, consists of 4712 *Weissella*-specific COGs (W-COGs) (Fig. 2). Although apparently low compared with the *Lactobacillus* pan-genome of 14 000 proteins (Kant et al., 2011), the difference may be,
Table 1. Overview of the origin and genome statistics of the *Weissella* genomes used in this study

Values shown are based on complete or draft genome data (strain-dependent) and Prokka (Seemann, 2014) and JGI IMG (Markowitz *et al.*, 2012) gene annotation pipelines. Psortb (Yu *et al.*, 2010) was used to compile the secretome data.

<table>
<thead>
<tr>
<th>Genome</th>
<th>Source</th>
<th>Length (bp)</th>
<th>Genome completeness (no. of scaffolds)</th>
<th>DNA G+C content (mol%)</th>
<th>Predicted ORFs</th>
<th>Coding DNA sequences (CDS)</th>
<th>Coding density (%)</th>
<th>*RNA genes</th>
<th>rRNA genes</th>
<th>tRNA genes</th>
<th>Genes assigned to COG (secretome (%; cell-wall associated + Extracellular))</th>
<th>GenBank accession number</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Weissella ceti</em> NC36</td>
<td>Rainbow trout spleen</td>
<td>1352640</td>
<td>7</td>
<td>40.76</td>
<td>1382</td>
<td>1271</td>
<td>88.71</td>
<td>111</td>
<td>16</td>
<td>68</td>
<td>1057</td>
<td>ANCA01000000</td>
</tr>
<tr>
<td><em>Weissella cibaria</em> AB3b</td>
<td>Buckwheat sourdough</td>
<td>2465158</td>
<td>88</td>
<td>44.56</td>
<td>2454</td>
<td>2349</td>
<td>88.09</td>
<td>105</td>
<td>7</td>
<td>57</td>
<td>1578</td>
<td>This study</td>
</tr>
<tr>
<td><em>Weissella cibaria</em> f3PR</td>
<td>Porcine faeces sourdough</td>
<td>2357128</td>
<td>60</td>
<td>44.86</td>
<td>2345</td>
<td>2228</td>
<td>88.24</td>
<td>117</td>
<td>7</td>
<td>70</td>
<td>1570</td>
<td>This study</td>
</tr>
<tr>
<td><em>Weissella cibaria</em> MG1</td>
<td>Sourdough</td>
<td>2436232</td>
<td>44</td>
<td>44.65</td>
<td>2378</td>
<td>2284</td>
<td>87.85</td>
<td>94</td>
<td>4</td>
<td>51</td>
<td>1607</td>
<td>This study</td>
</tr>
<tr>
<td><em>Weissella cibaria</em> KACC 11862</td>
<td>Kimchi</td>
<td>2317857</td>
<td>72</td>
<td>45.15</td>
<td>2225</td>
<td>2118</td>
<td>87.09</td>
<td>107</td>
<td>6</td>
<td>62</td>
<td>1560</td>
<td>AEKT01000000</td>
</tr>
<tr>
<td><em>Weissella cibaria</em> KCTC 3621</td>
<td>Wheat sourdough</td>
<td>2284024</td>
<td>71</td>
<td>44.79</td>
<td>2237</td>
<td>2141</td>
<td>85.62</td>
<td>96</td>
<td>11</td>
<td>46</td>
<td>1511</td>
<td>NZ_CAGH00000000.1</td>
</tr>
<tr>
<td><em>Weissella halotolerans</em> DSM 20190</td>
<td>Wheat sourdough</td>
<td>1358385</td>
<td>6</td>
<td>43.05</td>
<td>1425</td>
<td>1321</td>
<td>90.33</td>
<td>104</td>
<td>15</td>
<td>61</td>
<td>1157</td>
<td>NZ_ATUU00000000.1</td>
</tr>
<tr>
<td><em>Weissella koreensis</em> DSM 20190</td>
<td>Kimchi</td>
<td>1728940</td>
<td>17</td>
<td>35.50</td>
<td>1766</td>
<td>1661</td>
<td>86.39</td>
<td>105</td>
<td>17</td>
<td>60</td>
<td>1275</td>
<td>NZ_AKGG00000000.1</td>
</tr>
<tr>
<td><em>Weissella koreensis</em> KACC 15510</td>
<td>Kimchi</td>
<td>1441470</td>
<td>2</td>
<td>35.50</td>
<td>1439</td>
<td>1341</td>
<td>86.76</td>
<td>98</td>
<td>15</td>
<td>56</td>
<td>1091</td>
<td>NC_015759.1</td>
</tr>
<tr>
<td><em>Weissella paramesenteroides</em> ATCC 33313</td>
<td>Sausage</td>
<td>1962173</td>
<td>13</td>
<td>37.88</td>
<td>2398</td>
<td>2266</td>
<td>88.48</td>
<td>132</td>
<td>3</td>
<td>70</td>
<td>1477</td>
<td>NZ_ACKU00000000.1</td>
</tr>
</tbody>
</table>

*Includes tRNA, rRNA, tmRNA (transfer mRNA) and miscellaneous RNA genes.*
at least partly, due to the larger number of genomes used to estimate the Lactobacillus pan-genome. Furthermore, inherent variations in the computational programs used in the two studies, together with different sensitivity settings and cut-off values, must be considered when comparing outputs (Kant et al., 2011). Interestingly, the Tettelin mathematical model of the estimated Weissella proteome size (Fig. 3b) predicts that, for ten genomes, the pan-proteome consists of 3714 W-COGs. Again, different algorithms, cut-off values and applications used are probably responsible for the observed differences in pan-proteome size between the calculated and Tettelin-predicted value. The predicted value based on a theoretical input of 20 genomes increased the pan-proteome size by only 1210 COGs. This apparently small predicted pan-proteome compared with the pan-proteome based on 20 Lactobacillus strains suggests that there is lower diversity within the genus Weissella. However, note that in the current study the ten genomes analysed represented six species, whereas in the study by Kant et al. (2011), 20 genomes represented 13

**Fig. 1.** Neighbour-joining phylogenetic tree of species of the genera Weissella, Leuconostoc, Oenococcus and Fructobacillus, showing the position of the Weissella strains used in this study (*) based on 16S rRNA gene sequences. Bar, 2 nt substitutions per 100 nt. Bootstrap values (based on 1000 replications) are shown at nodes.
species and therefore a larger pan-proteome size would be expected.

The core-proteome at genus level, defined here as proteins present in all Weissella strains (i.e. those that are indispensable), was 729 W-COGs (Fig. 2), while the Tettelin-predicted value was 725 COGs (Fig. 3a). In addition, the core-proteome value is predicted to plateau after eight genomes, whereupon input of additional sequence data for other Weissella strains is predicted not to affect the core-proteome size. Even allowing for the fact that the dataset is representative of only six species, this may suggest stability in the Weissella genus.

To characterize the Weissella core-proteome pool, proteins were classified using the COG classification (Tatusov et al., 2003). This functional prediction of the Weissella core showed 16% of proteins belonging to ‘translation’, probably housekeeping proteins, 9% belonging to ‘replication and repair’, 29% to ‘unknown function or general function prediction only’, 5% to ‘transcription’ and 5% to ‘carbohydrate transport and metabolism’ (Fig. 4). All Weissella strains contained a mannose phosphotransferase system and a putative type III haemolysin. Apart from sharing of core genes required for central metabolism or ‘housekeeping’, the relative lack of shared traits between all species may reflect heterogeneity of the genus. The high percentage of proteins classified under ‘unknown function or general function prediction only’ also indicates the gap in genomic studies related to this genus compared with other LAB, and further illustrates the need for additional fundamental research on this group of organisms.

The pan- and core-proteome at species level was determined based on the four W. cibaria genomes in this study (Fig. 2). As expected, the calculated W. cibaria pan-proteome (2769 W. cibaria-specific COGs; Wc-COGs) was much smaller, and the core-proteome much larger (1912 Wc-COGs), than those values observed at the genus level. In total, 69% of the pan-proteome of the species constitutes the core-proteome. This proportionally large core-proteome may reflect the ability of this species to survive in the wide variety of niche habits in which strains has been found, including fermented meat and vegetable products (Ahmed et al., 2012; Kim et al., 2011; Kleppen et al., 2012; Park et al., 2007), dairy products (Björkroth et al., 2002; Van der Meulen et al., 2007), canine otitis (Björkroth et al., 2002), and human saliva, faeces and gall bladder (Björkroth et al., 2002; Kang et al., 2012). In addition, based on the core-proteome size of the genus (729 W-COGs) and species (1912 Wc-COGs), it is estimated that 1183 COGs in the W. cibaria core-proteome may be considered non-essential core proteins. However, it is important to realize that due to an over-representation of genomes from W. cibaria, the focus of our study, there may be a bias towards this species in the formation of COGs.

Tettelin plots of the estimated core- and pan-proteome size were also generated based on these species data (Fig. S1a, b); the predicted pan-proteome size was 2582 COGs and the core-proteome size was 1716 COGs. Again, variations in these values may be due to the factors outlined above.

The mathematical model generated by the Tettelin plot of the W. cibaria pan-proteome also predicts that with every new W. cibaria genome sequenced and/or added to the dataset (after the 10th genome), the number of new, unique, strain-specific COGs contributed by that genome will be ~18. This suggests that the W. cibaria pan-genome is open and that its size grows with the number of independent strains sequenced. Tettelin et al. (2005) observed a similar result with strains of Streptococcus agalactiae and Group A Streptococcus.

Figs 3(c) and S3(c) give a graphical representation of the pan-proteome at genus and species level, respectively. The relationship of one strain to another is depicted in terms of protein content.

**Secretome of Weissella**

Surface-associated and extracellular proteins play an important role in the interactions between LAB and their environment. These proteins form part of what is known as the bacterial secretome and have a role in processes such as recognition, binding and uptake of nutrients, signal transduction, communication with the environment and attachment of the bacterial cell to specific sites or surfaces (Desvaux et al., 2009; Zhou et al., 2010). The secretome of members of the genus Weissella, as indicated in Table 1, includes both cell-wall and extracellular (secreted) proteins. Certain cytoplasmic membrane proteins are often included as the secretome, but because the program used to analyse the data did not differentiate between proteins
intrinsic to the external or internal side of the plasma membrane, these proteins were excluded here. Table S2 shows a more comprehensive view of the Weissella secretome along with data from other LAB members for comparison. Additionally, InterProScan was used to analyse all proteins in W. cibaria MG1. Manual assessment of cellular protein localization (including cytoplasmic, cell-wall and extracellular proteins) indicated that 13 % were on the external side of the membrane and thus classified as part of the secretome; 46 % of these were hypothetical proteins.

As can be seen from Table 1, W. cibaria and closely related W. confusa have the largest secretome of all the weissellas. Many proteins of the MG1 secretome were large, multi-domain proteins of greater than 1000 aa. This has also been observed to be the case with proteins in the secretome of Lactobacillus species (Boekhorst et al., 2004). Many of the non-hypothetical proteins were identified as putative hydrolases, while four may be involved in peptide assimilation. The role of secretome proteins has gained particular interest in relation to probiotic micro-organisms, and it is interesting to observe that the species W. cibaria and W. confusa have a similar number of secretome proteins to well-known probiotic strains such as Lactobacillus rhamnosus GG (Table S2). Secreted proteins are postulated to play an important role in modulation and signalling to the host by these bacteria (Call & Klaenhammer, 2013; Zhou et al., 2010). Further studies and analyses are required to increase our understanding of the role of the secretome in food-associated LAB. Thus, we conclude that a large secretome does not imply probiotic potential upon a strain and it is the type and function of the secretome proteins that is more important in this respect.

Genomics of W. cibaria MG1

Carbohydrate metabolism. Consistent with it being classified as a heterofermentative LAB, W. cibaria has all the genes required for the phosphoketolase pathway. In silico analysis predicted that W. cibaria MG1 has the ability to metabolize galactose (via the Leloir pathway), maltose, fructose, ribose, xylose, sucrose and gluconate, for use as carbon sources. A β-galactosidase to enable lactose breakdown to glucose and galactose was putatively identified, indicating that lactose may be metabolized. These carbohydrates, however, must first be transported into the cell. Putative phosphotransferase systems (PTSs) were identified for N-acetyl-D-glucosamine, cellobiose, mannose, β-glucosides and fructose. One protein (MG1_0689) gave BLAST hits that suggest a role as a lactose and galactose permease, having homology to LacS (i.e. 68 % identity, over 99 % coverage to a lactose and galactose permease in Lactobacillus sakei). Depending on the organism, LacS can mediate lactose transport coupled to proton symport or by antiport with galactose (Aleksandrzak-Piekarczyk, 2013). The ability of W. cibaria MG1 to import and metabolize galactose and lactose is particularly interesting in light of a previous study by this group demonstrating the growth of this strain in 10 % reconstituted skimmed milk and survival in Cheddar cheese over a 3 month ripening period (Lynch et al., 2014). However, a study which examined the fermentation profiles of a number of W. cibaria strains found varying abilities to ferment galactose and lactose. In fact, only one of the six tested strains was able to ferment lactose (albeit weakly) (Bounaix et al., 2010). Thus, while the presence of a lactose permease–β-galactosidase system in strain MG1 is indicated, its functionality needs to be investigated further.

Two putative PTSs for the specific uptake of the β-glucoside, cellobiose, were annotated in MG1, reinforcing the plant-associated nature of this strain. Interestingly, it has been found that components of the cellobiose PTS in Lactococcus lactis IL1403 are also able to transport lactose, because the cellobiose-specific permease has an affinity for this disaccharide (Aleksandrzak-Piekarczyk et al., 2011). Despite the requirement for cellobiose to be present to allow lactose uptake, given the broad substrate specificity of PTS transporters of the Lac family, it is interesting to postulate that a similar system may function in MG1.

Another β-glucoside PTS of unknown specificity was also found in strain MG1 and may function in the uptake of other β-glucosides such as amygdalin, arbutin, aesculin, gentiobiose and salicin, carbohydrates widely distributed in plants. In addition, numerous phospho-β-glucosidases were found in the genome, highlighting its plant-associated nature and ability to utilize plant carbohydrates. Bounaix et al. (2010) found that all W. cibaria strains could ferment amygdalin, cellobiose, gentiobiose and salicin and could hydrolyse aesculin.

Interestingly, while strain MG1 was shown to possess a glycogen phosphorylase, responsible for glycogen and starch breakdown, putative enzymes conferring the ability to synthesize glycogen were also identified. Glycogen-synthesizing genes have recently been mentioned with respect to a Lactobacillus rhamnosus strain, but the presence of this trait in LAB has not been widely reported (Bove et al., 2012). A putative oligo-1,6-glucosidase (involved in the hydrolysis of α-1,6-D-glucosidic linkages) was annotated but contained an internal stop codon that probably abrogates activity. The presence of a non-functional oligo-1, 6-glucosidase pseudogene is interesting when considering that dextran (α-1,6-D-glucosidic linkages) EPS production is a key phenotype of this and other Weissella species. None of the W. cibaria strains examined by Bounaix et al. (2010) was able to ferment glycogen, which suggests an inability of this species to assimilate this polysaccharide.

Pyruvate metabolism. Pyruvate is the primary electron acceptor in LAB with the resultant formation of lactic acid. W. cibaria MG1 contains more than one copy of both D- and L-lactate dehydrogenases, re-enforcing their key role in NAD regeneration. Besides the reactions of the phosphoketolase (heterolactic acid fermentation) pathway, alternative pathways and fates of pyruvate are present in
coregenes (g) = 725 + 3304 \exp(-g / 0.87)

\text{pangenes (g)} = 1788 + 121.0(g - 1) + 2130 \exp(-2 - 1.29(g - 1)) / (1 - \exp(-1.29))
*W. cibaria* MG1, these mainly being dependent on the absence or presence of oxygen. This strain does not contain the enzymes of the pyruvate dehydrogenase pathway for the direct conversion of pyruvate to acetyl-CoA, nor does it contain pyruvate formate lyase, the enzyme responsible for the conversion of pyruvate to formic acid and acetyl-CoA in the so-called 'mixed acid fermentation' pathway (Huys *et al.*, 2012). Two putative pyruvate oxidase enzymes (and a third truncated enzyme) are, however, present in MG1. Pyruvate oxidase catalyses the conversion of pyruvate to acetyl-phosphate and ultimately acetate in the pyruvate oxidase pathway. This pathway requires the presence of oxygen, however, and will not occur under anaerobic conditions. Under anaerobic conditions an alternative route for pyruvate is the conversion to \( \alpha \)-acetolactate (diacetyl/acetoin pathway) catalysed by acetolactate synthase. \( \alpha \)-Acetolactate is subsequently converted to acetoin and 2,3-butanediol or non-enzymically and spontaneously transformed into diacetyl. Therefore, due to the requirement for oxygen in the pyruvate oxidase pathway, under anaerobic conditions the products of pyruvate will be lactic acid or those of the diacetyl/acetoin pathway. Under aerobic conditions,
however, in addition to the products of the diacetyl/acetoin pathway, due to the activation of the pyruvate oxidase pathway, additional acetate can be produced.

**Amino acid metabolism and proteolytic system.** *In silico* analysis suggests that *W. cibaria* MG1 can *de novo* synthesise alanine from pyruvate and aspartate from oxaloacetate. In addition cysteine can be synthesized from conversion pathways of aspartate, via homoserine. This suggests that MG1 is auxotrophic for the remaining 17 aa. Interestingly, only one of nine enzymes required to synthesise lysine from aspartate is not encoded in the genome of this strain, but could potentially be a missing gene due to gaps in sequencing.

*W. cibaria* appears to have an extensive proteolytic system, as also apparent in other LAB, although the exact enzymes that comprise this proteolytic system vary greatly, even at the strain level (see Table 2). *W. cibaria* MG1 possesses all three known LAB peptide transport systems, i.e. the di/tripeptide Dpp and DtpT systems and the oligopeptide Opp system. This includes two putative copies of the Opp system, arranged as two operons (MG1_0682..MG1_0686 and MG1_1992..MG1_1996). In addition to multiple, unidentified amino acid transporters, putative ABC transporters for glutamine, cysteine and methionine were identified. Specific transporters for lysine and tyrosine, as well as a serine/threonine transporter, an arginine/ornithine antiporter, an alanine/glycine permease, a choline/betaine transporter, and a γ-aminobutyrate permease were also found. Fifteen putative peptidases were identified, including genes with homology to aminopeptidases (pepC, pepN, PepM, PepA), endopeptidases (pepO, pepP), a dipeptidase (PepV), a tripeptidase (PepT) and proline preptidases (PepX, Pep/I, PepP, PepQ).

This large complement of peptide transporters and aminopeptidases (Table 2) may reflect the auxotrophy of this strain and its niche adaptation to the peptide-rich sourdough environment. A cell envelope proteinase was not present. This complement of amino acid transporters and aminopeptidases is also significant given that this strain has previously been shown to survive in a model cheese system over a 90 day ripening period (Lynch *et al.*, 2014). In addition, it is noteworthy that *W. cibaria* strains have been isolated from the dairy environment previously.

**Arginine deaminase pathway.** Our analysis shows that *W. cibaria* MG1 encodes the enzymes of the arginine deaminase pathway, the putative genes *arcABCD* being clustered together in an operon in a similar arrangement as is found in other LAB that can degrade arginine. The significance of this ability to hydrolyse arginine is not fully understood, but may aid survival in stressful environments by, for example, providing additional ATP when carbohydrates are scarce in the environment, or protecting from acid stress through ammonia production. A particular technological benefit of arginine deamination in the sourdough environment is the production of ornithine, which is an important precursor of crust aroma compounds (De Angelis *et al.*, 2002).

**EPS production in *W. cibaria*.** *W. cibaria* and *W. confusa* are known for their slime-producing properties, i.e. production of the homoeopolsaccharide (HoPS), dextran, in significant quantities (Ahmed *et al.*, 2012; Bouaïna *et al.*, 2010; Galle *et al.*, 2010; Huys *et al.*, 2012). HoPS production is also reported for some members of the genera *Leuconostoc, Lactobacillus* and *Streptococcus* (Korakli & Vogel, 2006; Monsan *et al.*, 2001; Tieking *et al.*, 2005). When grown on media containing sucrose, high-molecular-mass dextran is formed due to the catalytic action of an enzyme produced and secreted by a HoPS-producing LAB strain. This enzyme is a glucosyltransferase (GH70 family protein), known as a dextranase, that catalyses the hydrolysis of sucrose, coupled with the transfer of the glucose moiety to the growing dextran polysaccharide chain. In addition, if maltose is present in the medium it can act as an acceptor sugar and oligosaccharides of lower molecular mass will be formed (Ganzle & Schwab, 2009; Monsan *et al.*, 2001).

*W. cibaria* MG1 is of particular interest in our group due to its ability to produce a dextran HoPS in large quantities (~36 g l⁻¹). This strain and its associated EPSs have been studied in connection with potential bakery, brewing and dairy applications (Galle *et al.*, 2010; Lynch *et al.*, 2014; Wolter *et al.*, 2014; Zannini *et al.*, 2013). Genomic analysis confirmed the EPS-producing phenotype with the identification of a 1447 aa protein (MG1_1773) with 75% identity to a dextranase in *W. confusa*. In addition, the dextranase of MG1 shared 51% identity with the dextranase of *Leuconostoc mesenteroides* subsp. *mesenteroides* NRRL B-512(F). This *Leuconostoc mesenteroides* strain and its associated dextran have been used since the early 1950s as the reference strain for dextran research, and the industrial production of dextran, for application as a blood volume expander and use in other pharmaceutical products (Jeanes *et al.*, 1978). Fig. S2 shows an alignment of the dextranase amino acid sequences from both MG1 and B-512(F). Similarly conserved motifs, catalytic core residues and substrate binding residues are shown.

Analysis of the dextranase gene from *W. cibaria* MG1 has not revealed the reason for the hyper-production of EPS by this strain. However, the genomes of two additional *W. cibaria* strains (AB3b and f3PR) in the LAB collection at Cork Institute of Technology were sequenced in parallel with that of MG1, and this enabled additional analysis and comparison with the dextranase genes in these strains. Strain f3PR produced similar levels of EPS to MG1, in contrast to AB3b, which produced much lower levels under the same growth conditions. Fig. 5 shows an alignment of the amino acid sequences of the dextranase proteins from strains MG1 and AB3b. The catalytic site is highlighted as well as the proposed catalytic residues and those that interact with the sucrose/maltose substrate.

Due to incomplete sequencing data for strain f3PR, only a partial dextranase sequence was available for this strain.
However, over the total length of this partial sequence, the amino acid sequence of the ff3PR dextransucrase matches that of strain MG1 (100% identity over residues 796–1447). As mentioned, strain ff3PR produces similar levels of EPS to MG1 when grown under the same conditions on media containing 10% sucrose. In contrast, as can be seen in Fig. 5, there are a number of amino acid differences in the dextransucrase protein from strain AB3b in comparison to the protein produced by strain MG1. There are 37 aa substitutions in total, 19 within the proposed catalytic core; however, fundamentally, none of these substitutions occurs within the conserved dextransucrase motifs (I–IV) containing the catalytic and substrate-binding residues. It is indeed possible that one or more of these amino acid substitutions may be responsible for the lower levels of EPS produced by *W. cibaria* AB3b. Sequence differences in the promoter regions of the dextransucrase genes in both strains were also considered, but the proposed promoter regions are identical in the two strains (Fig. S3). It is also plausible that another element of regulation may be in action, or that the differences in the levels of EPS produced could be linked to, for example, the process of enzyme secretion from the cell, or indeed to differing specific growth rates of the strains.

**Table 2.** The proteolytic system of *W. cibaria* MG1, plus other LAB for comparison

<table>
<thead>
<tr>
<th>Peptidase</th>
<th>Substrate</th>
<th>MG1</th>
<th>ATCC 8293</th>
<th>IL1403</th>
<th>SK11</th>
<th>GG</th>
<th>DSM 20016</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Proteinase</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cell-wall-bound proteinase</td>
<td>PrtP</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>PrtM</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td><strong>Peptide transporters</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Oligopeptides ABC</td>
<td>OppA</td>
<td>4</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>transport system</td>
<td>OppB</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>OppC</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>OppD</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>OppF</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Di/tripeptides ABC</td>
<td>DppA/P</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>transport system</td>
<td>DppB</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>DppC</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>DppD</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>DppF</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Di/tripeptides ion-linked</td>
<td>DtpT</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>transporter</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Peptidases</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aminopeptidase</td>
<td>PepC</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>PepN</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>(Unique aminopeptidases)</td>
<td>PepM</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>PepA</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Pcp</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Endopeptidase</td>
<td>PepE/Pep G</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>pepO</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>pepF</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>Dipeptidase</td>
<td>PepD</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>PepV</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Tripeptidase</td>
<td>PepT</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Proline peptidase</td>
<td>PepX</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Pepl</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>PepR</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>PepL</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>PepP</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>PepQ</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

*MG1, Weissella cibaria MG1; ATCC 8293, Leuconostoc mesenteroides subsp. mesenteroides ATCC 8293; IL1403, Lactococcus lactis subsp. lactis IL1403; SK11, Lactococcus lactis subsp. cremoris SK11; GG, Lactobacillus rhamnosus GG; DSM 20016, Lactobacillus reuteri DSM 20016. Colour shading shows absence of a gene (white), a single gene (yellow) or multiple genes (green).*
Lactobacillus prophage attachment sites,
W. cibaria strains contain a gene that encodes a Streptococcus MG1. In addition, this strain also (2009) examined the haem- and consistently Lactococcus MG1. In both strains of Streptococcus agalactiae, Lactobacillus reuteri and Lactococcus lactis. However, while 95% of 88 Lactococcus lactis strains had the respiration trait upon testing, there were some strains that did not respond to haem (Brooijmans et al., 2009).

The presence of all necessary genes for haem-induced aerobic respiration in W. cibaria MG1, coupled with the fact that this trait has only been experimentally examined in a single isolate of this species, suggests that this should be further investigated, especially given the potential to increase strain robustness.

Bacteriophages and phage defence. PHAST (Zhou et al., 2011) was used to search for putative prophage elements in the genomes of W. cibaria. No phage-related proteins were identified in the genome of W. cibaria MG1. In both strains f3PR and KACC 11862, 14 putative phage-related ORFs were identified. These ORFs, however, did not represent complete prophages, and may be genetic vestiges of defective prophage. In contrast, PHAST identified a larger number (36 ORFs) of prophage-related elements in the genome of strain AB3b, in two genomic regions which were annotated as a putative temperate bacteriophage.

These prophages, occurring in two regions of the genome (1, 42 kb; 2, 21.3 kb) consisted of 31 and 21 ORFs, respectively (including hypothetical proteins). Manual annotation identified putative capsid, peptidase, lysis, holin, tail and terminase proteins in both prophage regions. The absence of an identifiable integrase protein in the smaller prophage suggests that this phage may be incapable of lysogeny. An integrase is present in the larger prophage, in addition to a second putative lysis/lysozyme and a methyltransferase [M subunit of a restriction modification (R/M) system]. This prophage is also flanked by regions with homology to attL and arrR prophage attachment sites, reinforcing that this region may represent a temperate phage. No phage repressor proteins were identified in either prophage region, but it is possible that one of a number of hypothetical proteins within each region could represent a novel repressor that has not yet been characterized. The most closely related phage to the putative temperate phage in region 1 was lactococcal phage 4268, a 36.5 kb Siphoviridae phage (Trotter et al., 2006). In addition, it is noteworthy that the ORF immediately preceding the prophage of region 2 was a protein annotated as an abortive infection phage resistance protein.

CRISPR-Finder (Grissa et al., 2007) was used to look for the presence of CRISPRs within the genomes of W. cibaria. No CRISPR sequences or CRISPR-associated proteins were despite a previous study demonstrating haem-induced respiration in this species (Sijpestein, 1970). The authors noted that the haem-induced respiration trait was highly species-specific and correlated with the presence of cydABCD genes; for example, all strains of Lactobacillus delbrueckii and Streptococcus thermophilus consistently lacked these genes, while they were present in all strains of Streptococcus agalactiae, Lactobacillus reuteri and Lactococcus lactis. However, while 95% of 88 Lactococcus lactis strains had the respiration trait upon testing, there were some strains that did not respond to haem (Brooijmans et al., 2009).

The presence of all necessary genes for haem-induced aerobic respiration in W. cibaria MG1, coupled with the fact that this trait has only been experimentally examined in a single isolate of this species, suggests that this should be further investigated, especially given the potential to increase strain robustness.

Bacteriophages and phage defence. PHAST (Zhou et al., 2011) was used to search for putative prophage elements in the genomes of W. cibaria. No phage-related proteins were identified in the genome of W. cibaria MG1. In both strains f3PR and KACC 11862, 14 putative phage-related ORFs were identified. These ORFs, however, did not represent complete prophages, and may be genetic vestiges of defective prophage. In contrast, PHAST identified a larger number (36 ORFs) of prophage-related elements in the genome of strain AB3b, in two genomic regions which were annotated as a putative temperate bacteriophage.

These prophages, occurring in two regions of the genome (1, 42 kb; 2, 21.3 kb) consisted of 31 and 21 ORFs, respectively (including hypothetical proteins). Manual annotation identified putative capsid, peptidase, lysis, holin, tail and terminase proteins in both prophage regions. The absence of an identifiable integrase protein in the smaller prophage suggests that this phage may be incapable of lysogeny. An integrase is present in the larger prophage, in addition to a second putative lysis/lysozyme and a methyltransferase [M subunit of a restriction modification (R/M) system]. This prophage is also flanked by regions with homology to attL and arrR prophage attachment sites, reinforcing that this region may represent a temperate phage. No phage repressor proteins were identified in either prophage region, but it is possible that one of a number of hypothetical proteins within each region could represent a novel repressor that has not yet been characterized. The most closely related phage to the putative temperate phage in region 1 was lactococcal phage 4268, a 36.5 kb Siphoviridae phage (Trotter et al., 2006). In addition, it is noteworthy that the ORF immediately preceding the prophage of region 2 was a protein annotated as an abortive infection phage resistance protein.

CRISPR-Finder (Grissa et al., 2007) was used to look for the presence of CRISPRs within the genomes of W. cibaria. No CRISPR sequences or CRISPR-associated proteins were
Fig. 5. Multiple sequence alignment of the amino acid sequences of dextranucrase enzymes from both W. cibaria strains, MG1 and AB3b. An asterisk indicates positions which have a single, fully conserved residue. A colon indicates conservation between groups of strongly similar properties. A period indicates conservation between groups of weakly similar properties. The catalytic site (cyan) is indicated. Conserved sequence motifs (Leemhuis et al., 2013) similarly present in glucansucrases from other bacteria are highlighted in yellow (I–IV). The three proposed catalytic residues, the nucleophilic aspartate (D556), the acid/base glutamate (E594) and the transition state stabilizing aspartate (D667) are shown in red. Residues in blue designate those that are proposed to interact with the substrate sucrose, while those in green (including the acid/base glutamine and the transition state aspartate) are proposed to interact with maltose. Residues in boxes are those that do not share similar properties. Alignments were created with CLUSTAL Ω (Sievers et al., 2011).
identified within the genomes of strains MG1, ff3PR and KACC 11862. CRISPR-Finder did identify questionable CRISPR sequences within the genome of strain AB3b, but no CRISPR-associated proteins were found, indicating that this is unlikely to be a functioning CRISPR system.

Two putative R/M systems were present in W. cibaria MG1. The BLAST results indicated that one of these systems is a type II R/M system containing a dam (DNA adenine) methyltransferase, and a restriction enzyme with putative homology to restriction endonuclease AlwI [recognition sequence: GGATC (4/5)]. The second R/M system appears to also be a type II system but its specificity is unknown. Interestingly, the putative AlwI R/M system is encoded by two of 66 ORFs that are uniquely present in the genome of W. cibaria MG1, and not in any other Weissella genome analysed in this study. In addition, two other ORFs in MG1 were annotated as putative DNA methyltransferases but no endonucleases were found near these ORFs.

W. cibaria AB3b contains one putative type II R/M system, of unknown specificity; unusually, one ORF in the putative temperate bacteriophage in region 1 was homologous to a DNA methyltransferase subunit of a type II R/M system. Similar to strain MG1, an endonuclease component was not identified in this region. Trotter et al. (2006) identified a methyltransferase gene in lactococcal phage 4268 and determined that it was likely to have originated from a host strain. It is plausible that the methyltransferase in the prophage of AB3b may similarly have originated from a host. Similar to strain MG1, strain ff3PR contains two putative type II R/M systems, of undetermined specificity.

Given that both strains MG1 and AB3b are sourdough W. cibaria isolates, it is interesting that MG1 contains no putative prophage elements within its genome, in contrast to AB3b which hosts two putative temperate phages. Whether these two strains were isolated from the same dough or doughs from the same back-slopping sourdough culture is not known. It is likely that strain AB3b evolved from an environment where it was likely to have been under repeated phage challenge. Few studies of Weissella bacteriophage have been performed to date, and these have been solely on phages of W. cibaria (Kleppen et al., 2012; Kot et al., 2014; Pringsulaka et al., 2011, 2012). The phages were of the families Podoviridae, Myoviridae and Siphoviridae. It was noted in these studies that the Myoviridae and Siphoviridae phages exhibited broad specificity, attacking other Weissella species and even crossing the genus barrier, lysing Lactobacillus brevis and Lactobacillus plantarum strains (Kot et al., 2014; Lu et al., 2012). None of these W. cibaria bacteriophages was isolated from sourdough, and very few phages have been isolated from this niche. It is thought that phage infections and spread in sourdough may be hampered by the dough’s solid nature and texture (Vogel et al., 2011).

Plasmids and encoded traits. Five putative plasmids were identified in W. cibaria MG1. Identification of contigs as plasmids was based on sequencing coverage, DNA G+C content and presence of a replication related protein (Table 3). The size of plasmids in MG1 ranged between 1.5 and 30.4 kb, and the DNA G+C content was between 36 and 41 mol%, lower than the value of 45 mol% of the chromosome. The lower G+C content suggests that these plasmids may have been transferred from related bacteria (Kim et al., 2013). There has been little research on plasmids of Weissella to date. Park et al. (2007) isolated three plasmids from a W. cibaria strain, KLC140 (pKLCA, 1.5 kb; pKLCB, 3.3 kb; pKLCC, 10 kb), reportedly the first isolation and characterization of plasmids from a member of the genus Weissella. Aside from a study which used another plasmid from this same strain (plasmid pKW2124, 2.1 kb, not reported in the 2007 study) to construct a surface display vector (Kim et al., 2013), no other research on plasmids of Weissella has been reported.

The replication protein (RepB) of the largest plasmid pKML1 in MG1, which encodes 29 ORFs, shows 64% similarity to a RepB protein from plasmid pSK11L of Lactococcus lactis subsp. cremoris SK11, a plasmid which replicates via the theta replication mechanism. Similarly, pKML2 and pKML3 show 71 and 73% similarity, respectively, to the theta-replicating plasmid, pCI305, of Lactococcus lactis subsp. lactis UC317. The single ORF on the smallest plasmid (pKML5), encoding the replication protein, showed 28% similarity to a replication protein on a plasmid (pFR18) from Leuconostoc mesenteroides that replicates via the rolling-circle mechanism (Shareck et al., 2004). In addition, this small plasmid shares 94% sequence identity (98% coverage) with a 1490 bp plasmid, pKLCA, in another W. cibaria strain (Park et al., 2007).

Annotation of plasmid-encoded ORFs in MG1 suggest that little genetic information is plasmid-encoded; the majority of genes on plasmids pKML2, pKML3 and pKML4 serve plasmid replication, but pKML5 is the only truly cryptic plasmid [i.e. extra-chromosomal DNA elements that encode no recognizable phenotype besides their replication functions (Shareck et al., 2004)]. In addition to replicative proteins, other ORFs are encoded on plasmids pKML2, pKML3 and pKML4, including a putative oxidoreductase, a xylose/proton symporter and an HNH endonuclease.

Among the encoded functions on the 30.4 kb plasmid, pKML1, including a number of hypothetical proteins, are proteins that appear to have diverse and unrelated functions. This includes proteins annotated as a putative glycosyltransferase, a copper chaperone, a metal transporter, a general stress response protein and a major facilitator protein.

Based on the criteria used to classify a contig as being a potential plasmid, it is questionable whether contig 9 (21.1 kb) represents a plasmid given its high sequence coverage and low G+C content. Despite no putative proteins involved in replication being annotated, a plasmid partition protein and a conjugative transfer relaxase
(similar function to plasmid mobilization) were identified. The highest BLAST hits for contig 9 were plasmids from Leuconostoc and Lactobacillus, but over a very short 108 bp region only. The apparent absence of encoded integrases and transposases and presence of only a single putative conjugal protein argue against this contig being a conjugal transposable element. It is possible, due to the draft nature of the genome, that other small contigs with high sequence coverage could be contiguous with this region. Two contigs of less than 1.5 kb in length, also having high sequence coverage, and containing ORFs coding solely for transposases, are represented in the draft genome of MG1.

Of the 21 ORFs present on contig 9, 11 encode glycosyltransferases and capsular polysaccharide biosynthesis proteins. Capsular polysaccharide biosynthesis has recently been observed in some strains of W. confusa but was not associated with W. cibaria strains (Malang et al., 2015). A single peptide transporter and seven hypothetical proteins were also present on this contig. The previously characterized plasmid, pKLCB, from W. cibaria KLC140 was also reported to lack a replication protein; the authors suggested that the GenBank database was too limited to allow a putative functional assignment to the ORFs (Kim et al., 2013).

Based on the criteria used to classify a contig as a putative plasmid, each of the other W. cibaria strains in this study contained five plasmids, with the exception of strain KACC 11862 which had a single 4.8 kb plasmid. The comparable lack of plasmids in this strain may reflect its divergent niche and geographical location with respect to the other W. cibaria strains (isolated from Kimchi in South Korea; Kim et al., 2011).

Interestingly, W. cibaria ff3PR, a porcine gut isolate, contained a near-identical plasmid (contig_36, 8116 bp, 36.40 mol% G+C, eight ORFs) to pKML3 in W. cibaria MG1. Of the eight ORFs on contig_36, seven have 100% identity to the ORFs on pKML3. The additional ORF on contig_36 (a mobilization protein, MobC) is also encoded on pKML3 but is truncated due to a frameshift mutation. Only three ORFs have a non-replicative function and were annotated as a putative xylose/proton symporter, a short-chain dehydrogenase and a hypothetical protein. Given the differing isolation sources of these strains, it is interesting that they share this plasmid. However, it is noteworthy that phylogenetic analysis of seven conserved housekeeping genes indicated that strains MG1 and ff3PR were most closely related within the W. cibaria clade (Fig. S4).

Mobile elements. Given the incomplete and draft nature of the W. cibaria genomes in this study, each containing a large number of gaps, it is likely that they contain a high number of repetitive sequences as these are the main reason for gaps in sequencing (Wassenaar & Lukjancenko, 2014). The presence of only seven transposases within strain MG1 would suggest a low level of genome plasticity; however, the incomplete nature of the genome and the presence of sequence gaps argue for a higher level of plasticity and it is possible that genes relating to transposable elements may be missing from the sequence. A similarly low number of transposasens was found in the other W. cibaria genomes. Conjugal transposons are one of the main drivers of gene gain and loss, and therefore if the lack of transposasens in strain MG1 truly reflects a low level of genome plasticity, this could indicate the stable adaptation of this strain to its sourdough niche (Douillard & de Vos, 2014; Hols et al., 2005). However, the low number of transposasens in these strains of W. cibaria is in stark contrast to the 111 found in another sourdough LAB, Lactobacillus sanfranciscensis TMW 1.1304, in which it was suggested that their high number facilitated niche adaptation of this strain (Vogel et al., 2011).

Species-specific protein set. In total, 231 species-specific COGs were identified within the four W. cibaria genomes; 115 of these were hypothetical proteins, while COG classifications were only assigned to 62 (27%). On the latter, 30.7% were assigned to ‘general functional prediction only or function unknown’, 16.1% to ‘carbohydrate metabolism and transport’, 9.7% to ‘cell-wall/membrane/envelope biogenesis’ and 6.5% to ‘amino acid metabolism and transport’. Some of the unique proteins in this species included PTSs for cellobiose and β-glucoside uptake, an acetoin diacetyl reductase, a putative glutamine ABC transporter and two proteins with homology to LysR-family transcriptional regulators.

An analysis of the niche-specific protein set within this species was performed. While the majority of niche-specific

<table>
<thead>
<tr>
<th>Contig</th>
<th>Sequencing coverage</th>
<th>Length (bp)</th>
<th>DNA G+C content (mol%)</th>
<th>Replication (initiation) protein</th>
<th>CDS</th>
<th>Designation</th>
</tr>
</thead>
<tbody>
<tr>
<td>16</td>
<td>&gt;1000×</td>
<td>30486</td>
<td>36.35</td>
<td>RepB</td>
<td>29</td>
<td>pKML1</td>
</tr>
<tr>
<td>9</td>
<td>&gt;1000×</td>
<td>21115</td>
<td>37.00</td>
<td>No</td>
<td>20</td>
<td>N/A</td>
</tr>
<tr>
<td>31</td>
<td>&gt;1000×</td>
<td>8457</td>
<td>36.54</td>
<td>RepB</td>
<td>8</td>
<td>pKML2</td>
</tr>
<tr>
<td>20</td>
<td>&gt;800×</td>
<td>8116</td>
<td>36.42</td>
<td>RepB</td>
<td>7</td>
<td>pKML3</td>
</tr>
<tr>
<td>19</td>
<td>&gt;5000×</td>
<td>4623</td>
<td>39.02</td>
<td>Replication initiation protein 4</td>
<td>4</td>
<td>pKML4</td>
</tr>
<tr>
<td>29</td>
<td>&gt;9000×</td>
<td>1550</td>
<td>41.16</td>
<td>Replication protein</td>
<td>1</td>
<td>pKML5</td>
</tr>
</tbody>
</table>
proteins were hypothetical, strains MG1 and AB3b, sourdough isolates, contained 38 potential sourdough-specific proteins. In total, 26 of these occur together within a similarly structured or conserved genomic region that contains a transposase, putative conjugal transfer proteins and a putative AAA ATPase. Interestingly, despite its isolation from a very different source (i.e. porcine faeces), strain ff3PR shares a larger number of unique proteins (114) with strain MG1 than does strain AB3b. This is also noteworthy, as phylogenetic analysis indicated that strains MG1 and ff3PR are the most closely related W. cibaria strains (Fig. S4). One of the shared regions between these two strains includes genes putatively involved in teichoic acid biosynthesis. Of all the W. cibaria isolates, both MG1 and ff3PR also contain a greater number of proteins that encode MucBP mucus-binding domains (Kleerebezem et al., 2010). In addition to their shared plasmid, as previously discussed, these findings raise interesting questions as to the evolution of both strains.

CONCLUDING REMARKS

To our knowledge, this study is the first to examine the genomics of the genus Weissella, with particular emphasis on W. cibaria. The core-proteome of the genus was determined to consist of approximately 729 W-COGs, and functional assignment of this core was similar to other LAB genera. A high proportion of these proteins, however, were classified as ‘unknown function or general function prediction only’, highlighting the current paucity of knowledge relating to this genus. Analysis of W. cibaria indicated that the core proteins constitute a large fraction of the pan-proteome of this species, which may be related to the diverse niche habitats in which this species has been found. In addition to a number of PTSs conferring the ability to assimilate plant-associated polysaccharides, an extensive proteolytic system was identified.

Nucleotide sequence accession numbers. The whole-genome shotgun project for W. cibaria MG1 has been deposited at DDBJ/EMBL/GenBank under accession number JWHU00000000. The version described in this paper is JWHU01000000.

The whole-genome shotgun project for W. cibaria AB3b has been deposited at DDBJ/EMBL/GenBank under accession number JWHT00000000. The version described in this paper is JWHT01000000.

The whole-genome shotgun project for W. cibaria ff3PR has been deposited at DDBJ/EMBL/GenBank under accession number JWHV00000000. The version described in this paper is JWHV01000000.

ACKNOWLEDGEMENTS

We thank Lynda Gunn for her help with the phylogenetic trees and also Caitriona Guinane for her assistance. K. M. L. and A. L. are supported by the Irish Research Council (Project refs. RS/2011/13 and RS/2012/219, respectively). This work was also supported by the Department of Agriculture, Food and the Marine, Dublin, Ireland (Project Ref. 08RDCIT600).

REFERENCES


Edited by: J. Kok