Development and implementation of multilocus sequence typing to study the diversity of the yeast *Kluyveromyces marxianus* in Italian cheeses

Fabrizia Tittarelli,1,2 Javier A. Varela,2 Loughlin Gethins,2 Catherine Stanton,3,4 R.P. Ross,6 Giovanna Suzzi,1 Luigi Grazia,5 Rosanna Tofalo1 and John P. Morrissey2,4,*

**Abstract**

The yeast *Kluyveromyces marxianus* possesses advantageous traits like rapid growth, GRAS (generally regarded as safe) status and thermostolerance that make it very suitable for diverse biotechnological applications. Although physiological studies demonstrate wide phenotypic variation within the species, there is only limited information available on the genetic diversity of *K. marxianus*. The aim of this work was to develop a multilocus sequence typing (MLST) method for *K. marxianus* to improve strain classification and selection. Analysis of housekeeping genes in a number of sequenced strains led to the selection of five genes, *IPP1*, *TFC1*, *GPH1*, *GSY2* and *SGA1*, with sufficient polymorphic sites to allow MLST analysis. These loci were sequenced in an additional 76 strains and used to develop the MLST. This revealed wide diversity in the species and separation of the culture collection and wild strains into multiple distinct clades. Two subsets of strains that shared sources of origin were subjected to MLST and split decomposition analysis. The latter revealed evidence of recombination, indicating that this yeast undergoes mating in the wild. A public access web-based portal was established to allow expansion of the database and application of MLST to additional *K. marxianus* strains. This will aid understanding of the genetic diversity of the yeast and facilitate biotechnological exploitation.

**DATA SUMMARY**

Accession numbers are provided in Table 1. The software used was as follows:


**INTRODUCTION**

*Kluyveromyces marxianus* is a homothallic yeast that, along with its sister species *Kluyveromyces lactis*, belongs to the subphylum Saccharomycotina. Some strains are able to utilize lactose as a carbon source but, unlike *K. lactis*, it can also secrete enzymes that degrade inulin, a fructan found in plants such as Jerusalem artichokes and agave [1, 2]. The capacity to assimilate lactose means that *K. marxianus* is frequently isolated from dairy products, such as fermented milk, yoghurt, kefir, koumis and cheese, although it can also be isolated from fermenting fruit and plant material [3–7]. The broad metabolic capacity combined with a rapid growth rate, thermostolerance (up to 52 °C) and a high secretory capacity (e.g. secretion of lytic enzymes) have led to multiple applications in biotechnology; for example, production of bioingredients and aroma compounds, production of ethanol by fermentation, removal of lactose from food, and bioremmediation [8, 9]. Because of its long and safe association with foods, *K. marxianus* is a GRAS (generally regarded as safe) organism in the USA and is on the European Food Safety Authority QPS (qualified presumption of safety) list.

There is not a unique *K. marxianus* strain that has been extensively characterized as a ‘model’ and those strains that...
have been studied display substantial physiological diversity [2, 10]. Although a challenge for detailed molecular and metabolic studies, this diversity is likely to be a very important enabler of a wide spectrum of biotechnological applications and, thus, can be considered positively. Nonetheless, full exploitation of this potential requires an improved understanding of the genetic diversity, and a means to identify and categorize strains. Sequence variation in the D1/D2 region of the 26S rRNA gene is universally accepted as the best way to distinguish yeast species from each other [11], but since this sequence is invariant within a species, this is not useful for intraspecific studies. Several researchers in recent years have applied molecular approaches to address this question. Differences at the intraspecific level in \textit{K. marxianus} have been studied by restriction analysis of the non-transcribed spacer (NTS) region of the rRNA gene cluster, RAPD (random amplified polymorphic DNA) PCR fingerprinting and variability of the insertion of the long terminal repeat (LTR) retrotransposon Tkm1 [12–16]. Karyotyping of \textit{K. marxianus} strains from different isolation sources and geographical origins suggest diverse chromosome numbers, although analysis of genomes of sequenced strains is consistent with a standard complement of eight chromosomes [17, 18]. This fits the emerging view that stable haploid and diploid \textit{K. marxianus} strains occur in natural environments [19]. It is notable, however, that a wide haplotype diversity was observed through RFLPs of mitochondrial DNA [15, 20]. A rapid, easy to use, scalable method in yeast and fungi is microsatellite analysis. Microsatellites are short tandem DNA repeats that can be found in all genomes and are used as molecular markers able to distinguish strains. They have been applied for typing a variety of yeasts, such as \textit{Saccharomyces uvarum} and \textit{Saccharomyces cerevisiae}. These studies showed that the patterns are stable and that the method is discriminatory; thus, it can be used for population structure studies [21, 22].

Comparison of the relationships among \textit{K. marxianus} strains is difficult and does not yield consistent patterns [23]. Furthermore, none of those methods is based on phylogeny, and so including new isolates and strains is problematic. These drawbacks can be overcome with the use of methods that are based on genomic sequence variation, making it relevant for exploitation. In addition, genome engineering tools are now available to allow construction of new strains. There is, however, little information on the genetic diversity of the species, or on the relationship between strains. This is resolved in this study through the development of an multilocus sequence typing system. To facilitate use by the research community, a web-based database and tool was implemented to allow any researchers to easily type and classify more \textit{K. marxianus} strains in the future.

**IMPACT STATEMENT**

\textit{Kluiveromyces marxianus} is emerging as one of the most important non-\textit{Saccharomyces} yeasts under development for biotechnological applications. It has traits such as rapid growth, high yield and thermostolerance that make it relevant for exploitation. In addition, genome engineering tools are now available to allow construction of new strains. There is, however, little information on the genetic diversity of the species, or on the relationship between strains. This is resolved in this study through the development of an multilocus sequence typing system. To facilitate use by the research community, a web-based database and tool was implemented to allow any researchers to easily type and classify more \textit{K. marxianus} strains in the future.

**Table 1.** \textit{K. marxianus} strains used in this study and their origin

<table>
<thead>
<tr>
<th>Strain</th>
<th>Isolation source</th>
<th>Culture collection/accession no.</th>
</tr>
</thead>
</table>
especially multilocus sequence typing (MLST), which is
defined by comparing polymorphic sites in a number of con-
served housekeeping genes within a species, permits easy
classification. As proof of principle, this method was applied
to study the diversity of a collection of *K. marxianus* strains
that had been isolated from two Italian cheese
fermentations.

**METHODS**

**Origin of yeast strains**

The 83 *K. marxianus* strains included in this study are listed
in Table 1, which also lists the strains for which genome
sequences were available for this study. Most of the isolates
are from dairy products from different geographical areas
in Italy, and were obtained from previous studies [13, 33].
In addition, 19 *K. marxianus* strains were from the Wester-
dijk Fungal Biodiversity Institute (Utrecht, The Nether-
lands) and from the National Collection of Yeast Cultures
(Norwich, UK), and 1 was from the University College Cork
collection. Yeasts were maintained as culture stocks in YPD
(1 %, w/v, yeast extract, 2 %, w/v, peptone, 2 %, w/v, dex-
trose) medium containing 20 % (v/v) glycerol at −80 °C. All
the strains were routinely grown in YPD medium overnight
at 28 °C in shake flasks.

**Choice of loci and design of primers**

Five housekeeping genes were retrieved from the *K. marxia-
nus* genomes of seven sequenced strains (CCT 7735, DMB1,
DMKU 3-1042, NBRC 1777, KCTC 17555, IPP1, GPH1,
GSY2 and GSA1), and a multiple sequence alignment was carried out
to identify single nucleotide polymorphisms (SNPs). These
genes, IPP1, TFC1, GPH1, GSY2 and GSA1, showed good
SNP distribution and were used in this study. Primers were
designed with SnapGene software to hybridize to non-variable
regions of each gene, thereby allowing amplification of the
polymorphic region in any *K. marxianus* strain. The
primer sequences and amplicon sizes are shown in Table 2.
A BLAST (basic alignment search tools) search was used to
check the specificity for *K. marxianus* of each primer set.
Furthermore, the properties of each primer were verified by
Oligo Analyzer 3.1. *K. lactis* CBS 683T was used as a nega-
tive control to evaluate primer specificity by PCR.

**DNA extraction, amplification and nucleotide
sequencing**

DNA was extracted with the QIAprep Spin miniprep kit,
following the manufacturer’s guidelines (Qiagen), modified
as follows: 5 ml overnight culture was resuspended in 250 µl
resuspension buffer (50 mM Tris-HCl (pH 8.0), 10 mM
EDTA and 10 µg RNase A ml−1) and vortexed for 5 min
with glass beads (0.15–0.6 mm in diameter; Sigma). DNA
quantification was carried out with a NanoDrop ND-1000
spectrophotometer. PCR analysis was performed with a
ProFlex PCR system thermal cycler (Applied Biosystems).
The reaction mixture (25 µl) contained 100 ng genomic
DNA, 13 µl GoTaq Green master mix (Promega) and 10 µM
each primer. The PCR conditions were: initial denaturation
at 95 °C for 5 min; followed by 30 cycles of denaturation at
95 °C for 30 s, annealing at 56 °C for 15 s and elongation at
72 °C for 1 min; followed by a final extension step at 72 °C
for 5 min. PCR products were separated in 1 % (w/v) agar-
ose gel in 1× TAE buffer (40 mM Tris (pH 7.6), 20 mM ace-
tetic acid, 1 mM EDTA (pH 8.0)) at 80 V for 1 h, stained with
SafeView (NBS Biologicals) and visualized under UV light.
DNA fragment sizes were determined using a 1 kb DNA
HyperLadder (BioLine). The amplified products were puri-
ified with a commercial PCR purification system (QIAquick
PCR purification kit; Qiagen), and amplicons were
sequenced in each direction by GATC (Constance,
Germany).

**Data analysis**

The sequence analysis for each gene fragment was per-
formed with Mega 7 [34] and Unipro UGENE programs [35].
For each fragment, the sequences obtained from the 83
strains were manually edited and compared with those from
the 7 *K. marxianus* reference sequences. Defined regions in
all sequences obtained from a single locus were aligned
using Muscle [36]. All the polymorphic sites were visually
inspected, and the polymorphic nucleotide sites within these
sequences were identified. In some cases, there were two
coincident, equivalently sized peaks in the forward and
reverse sequences, indicating heterozygosity at that position.
To avoid erroneous interpretation of double peaks, we fol-
lowed previous suggestions, and routinely took account of
noise in neighbouring positions and compared the chroma-
tograms for both sequenced strands [37]. The one-letter
code for nucleotides from the International Union of Pure
and Applied Chemistry (IUPAC) nomenclature was used to
designate homozygous and heterozygous polymorphic sites.
Genotype numbers were assigned to each unique sequence.
Each strain was defined by the combination of numbers cor-
responding to the genotype at the loci analysed, which is a
genotype profile or diploid sequence type (DST). Sequences
different even at a single nucleotide site were considered dis-
tinct genotypes.

Similarities among allelic profiles were displayed with a
dendrogram generated by using the unweighted pair-group
method with average (UPGMA) algorithm. This data analy-
sis was performed by using start software [38]. For the
computation of the phylogenetic tree, based on the
concatenated sequence of the five loci, a distance matrix was
created and a phylogenetic tree was reconstructed with the
program Unipro UGENE by using the maximum-likelihood
Table 2. Characteristics of the five housekeeping genes and primer sequences

<table>
<thead>
<tr>
<th>Gene</th>
<th>Molecular function</th>
<th>Gene length (bp)</th>
<th>Primer (5'→3')</th>
<th>Amplicon size (bp)</th>
<th>5' (3') position start</th>
<th>Size of sequenced fragment (bp)</th>
<th>No. of genotypes identified</th>
<th>No. of variable sites</th>
</tr>
</thead>
<tbody>
<tr>
<td>IPP1</td>
<td>Inorganic diphosphatase activity</td>
<td>864</td>
<td>Fwd: ATCGGTGCAAGAA-CACCTT Rev: TTGTCGATGGCTCG TCTGG</td>
<td>803</td>
<td>44 (807)</td>
<td>763</td>
<td>28</td>
<td>25</td>
</tr>
<tr>
<td>TFC1</td>
<td>RNA Pol III transcription factor activity</td>
<td>1857</td>
<td>Fwd: AAGGCCGATTTGGG TCAAAC Rev: TCTGCGGACTCAGAG TTATGC</td>
<td>832</td>
<td>992 (1783)</td>
<td>791</td>
<td>37</td>
<td>44</td>
</tr>
<tr>
<td>GPH1</td>
<td>Glycogen phosphorylase activity</td>
<td>2712</td>
<td>Fwd: TGGAACACTGTGAAG-CAGCA Rev: TTGTCAGCGTAC TCTGG</td>
<td>902</td>
<td>1797 (2659)</td>
<td>862</td>
<td>30</td>
<td>34</td>
</tr>
<tr>
<td>GSY2</td>
<td>Glycogen synthase activity</td>
<td>2106</td>
<td>Fwd: CACGCCATGAGA TCCCTCA Rev: CGTCTCTGTCGTCG CATGCC</td>
<td>891</td>
<td>1197 (2068)</td>
<td>871</td>
<td>34</td>
<td>49</td>
</tr>
<tr>
<td>SGAI</td>
<td>Glucan 1,4-α-glucosidase activity</td>
<td>1788</td>
<td>Fwd: CTCCGATGGTGCGG TCAAT Rev: TCATGCTCAAGGTAC TGGC</td>
<td>926</td>
<td>179 (1085)</td>
<td>906</td>
<td>26</td>
<td>44</td>
</tr>
</tbody>
</table>
Table 3. Genotypes of the 83 strains of *K. marxianus* analysed by MLST

<table>
<thead>
<tr>
<th>Strain name</th>
<th>Genotype</th>
<th>DST</th>
</tr>
</thead>
<tbody>
<tr>
<td>LM127*, LM30*, LM38*</td>
<td>IPP1 1 TFC1 1 GPH1 1 GSY2 1 SGA1 1</td>
<td>1</td>
</tr>
<tr>
<td>LM3*, LM50*, LM96*</td>
<td>IPP1 1 TFC1 1 GPH1 1 GSY2 14 SGA1 2</td>
<td>2</td>
</tr>
<tr>
<td>LM47*</td>
<td>IPP1 1 TFC1 1 GPH1 1 GSY2 18 SGA1 3</td>
<td>3</td>
</tr>
<tr>
<td>LM32*</td>
<td>IPP1 1 TFC1 1 GPH1 2 GSY2 1 SGA1 4</td>
<td>4</td>
</tr>
<tr>
<td>LM6*</td>
<td>IPP1 1 TFC1 1 GPH1 4 GSY2 2 SGA1 5</td>
<td>5</td>
</tr>
<tr>
<td>LM42*</td>
<td>IPP1 1 TFC1 1 GPH1 17 GSY2 16 SGA1 6</td>
<td>6</td>
</tr>
<tr>
<td>LM36*</td>
<td>IPP1 1 TFC1 1 GPH1 6 GSY2 1 SGA1 7</td>
<td>7</td>
</tr>
<tr>
<td>CBS 1555, CBS 712, CBS 7858, CBS 608</td>
<td>IPP1 1 TFC1 8 GPH1 1 GSY2 1 SGA1 8</td>
<td>8</td>
</tr>
<tr>
<td>6M2*</td>
<td>IPP1 1 TFC1 10 GPH1 3 GSY2 4 SGA1 9</td>
<td>9</td>
</tr>
<tr>
<td>CBS 834†</td>
<td>IPP1 1 TFC1 11 GPH1 1 GSY2 4 SGA1 10</td>
<td>10</td>
</tr>
<tr>
<td>VG6†</td>
<td>IPP1 1 TFC1 16 GPH1 17 GSY2 2 SGA1 11</td>
<td>11</td>
</tr>
<tr>
<td>K326*</td>
<td>IPP1 1 TFC1 26 GPH1 16 GSY2 4 SGA1 12</td>
<td>12</td>
</tr>
<tr>
<td>M169†, M38‡</td>
<td>IPP1 1 TFC1 26 GPH1 17 GSY2 4 SGA1 13</td>
<td>13</td>
</tr>
<tr>
<td>LM139*</td>
<td>IPP1 4 TFC1 1 GPH1 1 GSY2 1 SGA1 14</td>
<td>14</td>
</tr>
<tr>
<td>LM133*</td>
<td>IPP1 4 TFC1 1 GPH1 1 GSY2 15 SGA1 15</td>
<td>15</td>
</tr>
<tr>
<td>CBS 7894</td>
<td>IPP1 3 TFC1 9 GPH1 22 GSY2 1 SGA1 16</td>
<td>16</td>
</tr>
<tr>
<td>ISC4*</td>
<td>IPP1 6 TFC1 13 GPH1 12 GSY2 15 SGA1 17</td>
<td>17</td>
</tr>
<tr>
<td>LM116*</td>
<td>IPP1 7 TFC1 1 GPH1 17 GSY2 4 SGA1 18</td>
<td>18</td>
</tr>
<tr>
<td>LM72*</td>
<td>IPP1 12 TFC1 4 GPH1 1 GSY2 19 SGA1 19</td>
<td>19</td>
</tr>
<tr>
<td>LM28*</td>
<td>IPP1 19 TFC1 5 GPH1 28 GSY2 1 SGA1 20</td>
<td>20</td>
</tr>
<tr>
<td>LM169*</td>
<td>IPP1 20 TFC1 27 GPH1 13 GSY2 24 SGA1 21</td>
<td>21</td>
</tr>
<tr>
<td>LM142*</td>
<td>IPP1 21 TFC1 7 GPH1 1 GSY2 25 SGA1 22</td>
<td>22</td>
</tr>
<tr>
<td>CBS 2762</td>
<td>IPP1 25 TFC1 28 GPH1 8 GSY2 14 SGA1 23</td>
<td>23</td>
</tr>
<tr>
<td>CBS 1596</td>
<td>IPP1 32 TFC1 21 GPH1 1 GSY2 1 SGA1 24</td>
<td>24</td>
</tr>
<tr>
<td>LM141*</td>
<td>IPP1 33 TFC1 1 GPH1 1 GSY2 25 SGA1 25</td>
<td>25</td>
</tr>
<tr>
<td>LM92*</td>
<td>IPP1 33 TFC1 8 GPH1 20 GSY2 1 SGA1 26</td>
<td>26</td>
</tr>
<tr>
<td>LM174*</td>
<td>IPP1 34 TFC1 2 GPH1 19 GSY2 22 SGA1 27</td>
<td>27</td>
</tr>
<tr>
<td>LM9*</td>
<td>IPP1 35 TFC1 1 GPH1 1 GSY2 14 SGA1 28</td>
<td>28</td>
</tr>
<tr>
<td>LM14*</td>
<td>IPP1 2 TFC1 1 GPH1 1 GSY2 3 SGA1 29</td>
<td>29</td>
</tr>
<tr>
<td>LM48*</td>
<td>IPP1 3 TFC1 1 GPH1 5 GSY2 1 SGA1 30</td>
<td>30</td>
</tr>
<tr>
<td>LM136*</td>
<td>IPP1 4 TFC1 9 GPH1 6 GSY2 15 SGA1 31</td>
<td>31</td>
</tr>
<tr>
<td>LM17*</td>
<td>IPP1 5 TFC1 1 GPH1 10 GSY2 23 SGA1 32</td>
<td>32</td>
</tr>
<tr>
<td>LM141*</td>
<td>IPP1 6 TFC1 23 GPH1 18 GSY2 1 SGA1 33</td>
<td>33</td>
</tr>
<tr>
<td>LM153*</td>
<td>IPP1 7 TFC1 3 GPH1 11 GSY2 21 SGA1 34</td>
<td>34</td>
</tr>
<tr>
<td>VG1†, VG4‡</td>
<td>IPP1 8 TFC1 22 GPH1 22 GSY2 11 SGA1 35</td>
<td>35</td>
</tr>
<tr>
<td>M41†</td>
<td>IPP1 8 TFC1 22 GPH1 17 GSY2 4 SGA1 36</td>
<td>36</td>
</tr>
<tr>
<td>M83‡</td>
<td>IPP1 8 TFC1 22 GPH1 17 GSY2 4 SGA1 37</td>
<td>37</td>
</tr>
<tr>
<td>M12*, M123‡, M131†, M135†, M166†, M48‡, M68‡</td>
<td>IPP1 9 TFC1 26 GPH1 17 GSY2 4 SGA1 38</td>
<td>38</td>
</tr>
<tr>
<td>M14‡</td>
<td>IPP1 9 TFC1 26 GPH1 17 GSY2 5 SGA1 39</td>
<td>39</td>
</tr>
<tr>
<td>M81‡</td>
<td>IPP1 10 TFC1 22 GPH1 17 GSY2 4 SGA1 40</td>
<td>40</td>
</tr>
<tr>
<td>LM44*</td>
<td>IPP1 11 TFC1 1 GPH1 15 GSY2 1 SGA1 41</td>
<td>41</td>
</tr>
<tr>
<td>CBS 397</td>
<td>IPP1 12 TFC1 8 GPH1 1 GSY2 1 SGA1 42</td>
<td>42</td>
</tr>
<tr>
<td>LM161*</td>
<td>IPP1 13 TFC1 26 GPH1 14 GSY2 20 SGA1 43</td>
<td>43</td>
</tr>
<tr>
<td>LM15*</td>
<td>IPP1 13 TFC1 27 GPH1 12 GSY2 15 SGA1 44</td>
<td>44</td>
</tr>
<tr>
<td>LM154*</td>
<td>IPP1 13 TFC1 37 GPH1 16 GSY2 32 SGA1 45</td>
<td>45</td>
</tr>
<tr>
<td>FM96§</td>
<td>IPP1 14 TFC1 31 GPH1 29 GSY2 29 SGA1 46</td>
<td>46</td>
</tr>
<tr>
<td>NCYC 179</td>
<td>IPP1 15 TFC1 19 GPH1 1 GSY2 7 SGA1 47</td>
<td>47</td>
</tr>
<tr>
<td>CBS 6432</td>
<td>IPP1 15 TFC1 24 GPH1 1 GSY2 7 SGA1 48</td>
<td>48</td>
</tr>
<tr>
<td>UCKM1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CCT 7735</td>
<td>IPP1 15 TFC1 26 GPH1 30 GSY2 15 SGA1 50</td>
<td>50</td>
</tr>
<tr>
<td>NCYC 1424</td>
<td>IPP1 16 TFC1 14 GPH1 9 GSY2 14 SGA1 51</td>
<td>51</td>
</tr>
</tbody>
</table>
method. The bootstrap analysis was conducted on 1000 replicates.

The MLST data were used to predict genetic exchanges with the program SplitsTree [39]. The split decomposition method was used to assess the degree of tree-like structure present in the alleles found in all the five loci for the strains isolated from the Italian cheeses. Three types of statistical analysis were applied to the data: (1) the $p$-distance to check mean identity to estimate reliability of the alignment; (2) the standardized index of association ($I_{SA}$) to estimate multilocus linkage disequilibrium [40, 41]; (3) the phi-test for detecting the presence of recombination. The tests were carried out with Mega 7, START and SplitsTree programs, respectively.

**Website development**

A website was created to allow users to compare a strain of interest against the 83 *K. marxianus* strains shown in this study, through a maximum-likelihood phylogenetic tree. The website takes a multi FASTA file containing the IPP1, TFC1, GPH1, GSY2 and SGA1 sequences as input. Then, it runs a custom Biopython script that computes the phylogenetic tree. The program checks for sequence format, runs a multiple alignment against a prebuilt alignment profile using MUSCLE v 3.8 and reconstructs the phylogenetic tree using PhyML v 3.1. The website was constructed using the HTML v 5.0 mark-up language and CSS v 3. Complementary scripts were written in Biopython v 1.69 and implemented into the website through a CGI Python script.

### RESULTS AND DISCUSSION

The availability of an accurate and reproducible method to classify strains within a species is necessary to establish phylogenetic relationships and study biogeography, population structure and evolution. There are a wide variety of methods that have been developed and applied for this purpose but, depending on the method, there can be problems with technical difficulty, poor reproducibility between groups and limited scalability [42, 43]. The most powerful methods are based on DNA sequence variation; thus, comparison of whole genome sequences can be considered the gold standard. However, despite the reduction in sequence costs, this is not always feasible when large numbers of strains are involved and other methods that are rapid, easy to use and scalable, while still accurate, have been applied over the years to study diversity [16–18, 20].

### Choice of loci and sequence polymorphisms

The focus of this study was to develop an MLST analysis method for *K. marxianus* that could be used to distinguish strains. The utility of MLST relies on the selectively neutral variability of the housekeeping genes selected [24]; therefore, the choice of genes is critical.

To identify loci that would be suitable for MLST, five genes that had been previously described as housekeeping genes and used as controls in molecular studies in *S. cerevisiae* were considered [44]. The sequences of these genes were retrieved from the seven *K. marxianus* genome sequences available at that time. The dN/dS (non-synonymous to

### Table 3. cont.

<table>
<thead>
<tr>
<th>Strain name</th>
<th>Genotype</th>
<th>DST</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>IPP1</td>
<td>TFC1</td>
</tr>
<tr>
<td>CBS 5795</td>
<td>16</td>
<td>29</td>
</tr>
<tr>
<td>LM54*</td>
<td>17</td>
<td>12</td>
</tr>
<tr>
<td>CBS 5668</td>
<td>19</td>
<td>24</td>
</tr>
<tr>
<td>BO339</td>
<td>18</td>
<td>3</td>
</tr>
<tr>
<td>KCTC 17555, CBS 6556</td>
<td>20</td>
<td>5</td>
</tr>
<tr>
<td>NBRC 1777</td>
<td>21</td>
<td>1</td>
</tr>
<tr>
<td>CBS 4857</td>
<td>21</td>
<td>10</td>
</tr>
<tr>
<td>DMMKU 3-1042</td>
<td>22</td>
<td>11</td>
</tr>
<tr>
<td>LM148*</td>
<td>23</td>
<td>11</td>
</tr>
<tr>
<td>IPE453</td>
<td>22</td>
<td>11</td>
</tr>
<tr>
<td>CBS 745</td>
<td>24</td>
<td>30</td>
</tr>
<tr>
<td>DMB1</td>
<td>25</td>
<td>11</td>
</tr>
<tr>
<td>LM20*</td>
<td>26</td>
<td>18</td>
</tr>
<tr>
<td>LM167*</td>
<td>27</td>
<td>16</td>
</tr>
<tr>
<td>CBS 2233, CBS 4354</td>
<td>28</td>
<td>8</td>
</tr>
</tbody>
</table>

*From Parmigiano Reggiano, Italy.
†From cow’s milk whey, Italy.
‡From Pecorino di Farindola, Italy.
§From fermented milk, Italy.
||From unpasteurized milk, Ireland.

Tittarelli et al., *Microbial Genomics* 2018;4
(synonymous amino acid changes) in each of the five genes, IPP1 (0), TFC1 (0.03±0.013), GPH1 (0.006±0.003), GSY2 (0.04±0.017) and SGA1 (0.04±0.036) was less than 1, indicating that they were not under positive selection; therefore, they were valid choices for MLST. These genes were individually compared and it was confirmed that they satisfied the criterion of displaying sufficient polymorphisms within regions that could be comfortably amplified in a single PCR reaction. Seventy-nine SNPs were found within the selected sequences of the five genes and primers to conserved regions were then designed to PCR amplify these loci from 76 additional K. marxianus strains. The identical loci were successfully amplified from all strains and analysed for polymorphisms. In total, there were 196 polymorphic sites (4.6 %) where nucleotide substitutions were seen. No insertions or deletions were seen and 133/196 of the changes were synonymous, not changing the predicted encoded amino acid. The remaining 63 SNPs resulted in a changed predicted amino acid at that position, of which 39 were substantive, such as basic to acid side chains and aliphatic to aromatic side chains (Table S1, available in the online version of this article). The most common change was from a non-polar aliphatic side chain to a polar aliphatic side chain.

The number of polymorphic sites ranged from 25 in IPP1 to 49 in GSY2 (Table 2). Phylogenetically informative sites (at least two bases present in two or more of the alleles) ranged from 67 % in TFC1 and GPH1 to 41 % in GSY2, and are also shown in Table S2, which also details polymorphic nucleotide sites, the different genotypes identified at each of the five genes and their frequency of detection among the 83 K. marxianus strains.

All fragments of a given locus were aligned by multiple sequence alignment and a reference sequence was defined on the basis of the most common nucleotide at each polymorphic site (Table S2). Many of the strains were diploids and heterozygosity was quite common, in which case it was necessary to use IUPAC nomenclature to describe the sequence. It was not possible to determine which allele was associated with which nucleotide at heterozygous positions, a limitation common to all methods using nucleotide polymorphism analysis at several heterozygous loci in diploid organisms [30]. The reference sequence was given an arbitrary genotype number of 1, with variant sequences numbered successively (Table S2). The number of genotypes identified for each of the five loci investigated ranged from 26 for SGA1 to 37 for TFC1 (Table 2). Using the five

![Fig. 1. Phylogenetic tree reconstructed on the basis of the aligned concatenated sequence of the five housekeeping genes used in this work (IPP1, TFC1, GPH1, GSY2, SGA1) with the maximum-likelihood method. The position of the strains originating from the two Italian cheeses are indicated by circles. The bootstrap analysis was conducted on 1000 replicates.](image-url)
genotype numbers, it was possible to describe the DSTs as the unique combination of the genotypes of the five loci studied. From the 83 strains, 66 DSTs were defined, whereby 58 strains were the sole representative of their DST, while the other 8 DSTs contained 2–7 representatives. This high level of differentiation confirmed the discriminatory power of MLST as a strain-typing procedure (Table 3).

To investigate the relationship between the 83 strains, the 5 loci of each strain were concatenated and aligned using Muscle. Allelic profiles evaluated by the UPGMA algorithm are shown in Fig. S1. A phylogenetic tree was then reconstructed by the maximum-likelihood method using Unipro UGENE (Fig. 1). This illustrated the diverse nature of the *K. marxianus* species, and showed that the culture collections and sequenced strains are interspersed in different clades. Split decomposition analysis was applied to further analyse the population structure of the 52 *K. marxianus* strains that were isolated from two different Italian cheeses, Parmigiano Reggiano and Pecorino di Farindola. These strains were part of sets previously analysed by different typing methods that did not yield consistency in classification [13, 18, 33]. The concatenated IPP1, TFC1, GPH1, GSY2 and SGA1 gene sequence was analysed and graphically displayed with SplitsTree, which calculates a phylogenetic network that permits the visualization of genetic exchanges in addition to mutations (Fig. 2). The SplitsTree graphs for both the Parmigiano Reggiano strains (Fig. 2a) and the Pecorino di Farindola strains (Fig. 2b) presented network-like structures, separating the strains into groups.

The same groups are found in the maximum-likelihood tree shown in Fig. 1. Concerning the origin of the strains, all *K. marxianus* isolated from Pecorino di Farindola clustered together, although strains M41, M81 and M83 appeared in a separated branch in a similar way as in Figs 1 and 2(b). However, Parmigiano Reggiano *K. marxianus* strains were found to be more different, showing the presence of different groups in Fig. 1. In Fig. 2(a) the majority of the strains clustered together and just seven strains (LM154, LM169, LM161, LM15, LM20, LM167 and LM148) were differentiated. Two different analyses were employed to check for recombination in the two sets of strains. First, the phi-test gave $p$-values of $1.337^{-5}$ with $K=2$ (Parmigiano Reggiano strains) and $2.158^{-5}$ with $K=1$ (Pecorino di Farindola strains), which provides statistical support for recombination in both groups of strains. Second, the level of linkage disequilibrium between alleles was assessed using the $I_A$ value [45], which is calculated as the ratio of the observed variance with the expected variance [40]. To avoid dependence on the number of loci, the $I_A^S$, which is expected to be zero when alleles are in linkage equilibrium (free recombination), was used [41]. The $I_A^S$ value was 0.993 for the entire population, indicating the presence of groups of strains corresponding to distinct populations. The $I_A^S$ was then calculated for the two sub-populations separately (Parmigiano Reggiano and Pecorino di Farindola strains). The values were 0.028 and 0.069 for Parmigiano Reggiano and Pecorino di Farindola strains, respectively, indicating recombination within the subpopulations. Both the phi-test

![Fig. 2. Split decomposition analysis based on the allelic profiles of the 39 strains isolated from Parmigiano Reggiano (a), and the 13 strains isolated from Pecorino di Farindola (b).](image-url)
and the linkage disequilibrium measure $I^S_N$ are consistent with the idea that there is recombination within these populations of strains. This should not be unexpected since most strains were diploids and \textit{K. marxianus} possesses an intact mating system and can readily be crossed in laboratory studies \cite{19, 46}.

To allow the MLST developed here to be applied to more strains, a straightforward website was set up where custom Biopython scripts carry out the typing and generate a phylogenetic tree as well as a PhyML text output file in Newick format. That output file can be imported into other tree visualization software such as MEGA or TreeMe. This website is hosted on a University College Cork server at \texttt{(kmarxianusMLST.ucc.ie)}. To type any additional \textit{K. marxianus} strains, the \texttt{IPPI, TFC1, GPH1, GSY2 and SGAI} genes need to be sequenced using the primers specified in Table 2. The sequences are then uploaded in a multi \texttt{FASTA} file and the analysis is automatically performed. Illustrations of the outputs are available as supplementary data (Figs S2, S3 and S4).

In conclusion, in addition to shedding light on the population structure of the \textit{K. marxianus} species, this study provides a framework useful to understanding genetic divergence. Our results showed that MLST of \textit{K. marxianus} is highly reproducible and discriminatory, and superior to other methods used to date to classify strains. To enhance the applicability of the method, a website was constructed to allow typing and comparison of additional \textit{K. marxianus} strains in the future.

Funding information
L.G. was supported by Teagasc Walsh fellowship grant number 2013044. The research leading to these results has received funding from the People Programme (Marie Curie Actions) of the European Union’s Seventh Framework Programme FP7/2007-2013, under REA grant agreement no. 606795.

Acknowledgements
We thank Paddy Mulligan for assistance with installing and hosting the MLST server.

Conflicts of interest
The authors declare that there are no conflicts of interest.

Data bibliography
1. The accession numbers of the genome sequences are listed in Table 1.

References
23. Pérez-Brito D, Tapia-Tussell R, Quijano-Ramayo A, Larqué-Saavedra A, Lappe P. Molecular characterization of...


