2nd year, Semester 4
Internship Report

Study of genetic diversity of *Lactobacillus delbrueckii*: description of recombination.

Name: Blin
First name: Camille

Previous and current internships: (Indicate current internship in bold)

Master Biosciences	**Semester**	**Host team, laboratory**	**Internship supervisor**
1st year	S2	Muséum National d'Histoire Naturelle Department of Systematics and Evolution UMR-CNRS 7205	Thierry Wirth
2nd year	S3	Environmental Research Institut University College Cork, Irland	Mark Achtman
	S4	**GenEvol team,** **Mathematics, informatics and Genomics Laboratory,** **INRA Jouy en Josas**	**Philippe Bessière,** **Mahendra Mariadassou**

Number of characters (maximum 40 000) : 39 588

Study of genetic diversity of *Lactobacillus delbrueckii*: description of recombination

Abstract

Lactobacillus delbrueckii (*L. delbrueckii*) is a bacterial species extensively used as a starter in dairy industry. The strains of *L. delbrueckii* present highly variable immunomodulatory properties and an ongoing project tries to elucidate the molecular basis of those differences using, among others, comparative genomics approaches. The internship used the genomes produced within the project to study and assess the importance of homologous recombination in *L. delbrueckii.* This evolutionary mechanism consists in an exchange of homologous DNA material between closely related bacteria. It has an important impact on the genetic diversity because each recombination event introduces up to several substitutions and thus clouds the clonal relationship that exists between strains.

10 *L. delbrueckii* complete genomes, equally distributed among the two subspecies *L. delbrueckii lactis* and *bulgaricus* were available. First we reconstructed the phylogeny of the strains which turned out to be strongly supported by the genomic data. Then, we applied an algorithm based on Monte Carlo Markov Chains that is able to reconstruct the clonal frame of a strain set while accounting for recombination and to infer recombinant tracts. We have shown that recombination occurs in *L. delbrueckii* and is responsible for up to 40% of the observed diversity. The subspecies *lactis* seems to be more affected by recombination than *bulgaricus*. The analysis of the genes affected did not reveal a specific enrichment of one of the functional categories but showed an heterogeneity of the recombinant tract distribution among the genes and over the genome. The origin of the tracts could not be determined and remains unknown.

1. Introduction

1.1. Background

Lactobacillus delbrueckii is a bacterial strain extensively used in dairy industry and consumed by humans in high amounts. Its capacity to transform galactose into lactic acid makes it a good starter culture for the production of fermented milk. We can find large quantities of *Lactobacillus delbrueckii* in hard cooked cheese (such as Emmental, Comté or Beaufort) and in yogurts. Indeed, it is estimated that 10^9-10^{10} bacteria are consumed daily by every French people. There has been a recent growing interest in the bacteria resident of digestive tracts because it appears that they could be involved in maturation and modulation of the immune system and act as "probiotic". According to the Food and Agriculture Organization, the concept of "probiotic bacteria" refers to "live microorganisms which, when administered in adequate amount, confer a health benefit for the host" (FAO/WHO, 2001). There are

many studies of the immunomodulatory properties of probiotic bacteria, as those bacteria could be used to help treat chronic inflammatory disorders of the digestive tract. The idea of "probiotic" bacteria was first introduced at the beginning of the 20th century by Metchnikoff who hypothesized that yogurt bacteria consumption was health-beneficial. *L. delbrueckii* is not considered a probiotic but has been shown to exhibits some anti-inflammatory properties, with high variability between strains. In this context, a comprehensive scientific project called SURFING (Starter SURFace against INFlammation of the Gut) was initiated in 2011 to improve our knowledge of these daily consumed bacteria. Several strains presenting different patterns of immunomodulatory properties were selected and will be compared to find the genetic basis and the molecular mechanisms at the origin of their immunomodulatory properties.

Lactobacillus delbrueckii is a lactic acid bacteria species presenting two subspecies: *Lactobacillus delbrueckii bulgaricus* and *lactis* (*L. bulgaricus* and *L. lactis*). Despite its importance in the dairy industry, this species has been less intensively studied than others lactic acid bacteria because of its atypical GC content (about 50% compared to about 35% in the genus) (Nicolas et al., 2007) was inconvenient for experiments and because the transformation of these strains was difficult (Serror et al., 2002). The first complete genome of *L. bulgaricus* was sequenced in 2006 (Van de Guchte et al., 2006), initiating the sequencing of other strains (Makarova et al., 2006; Hao et al., 2011) and subsequent studies. Those studies have shown an ongoing evolution process leading to an extensive reduction of the genome, as illustrated by pseudogenization and incomplete metabolic pathways. This genome reduction is interpreted as an adaptation of the bacteria from its original plant-associated environment to its new lactose-rich environment, in protocooperation with *Streptococcus thermophilus.* (Van de Guchte et al., 2006) *L. lactis* has been even less studied: as of today, only two genomes are available. The SURFING project is based on the sequencing (30-40X) of additional *L. delbrueckii* strains presenting highly contrasted immunomodulatory properties for a total of 5 *L. bulgaricus* and 5 *L. lactis.* The project will integrate comparative genomics, transcriptomics and proteomics analyses to identify candidate genes for the interaction with the host immune system.

In this context, the main objective of my internship was to mine the genomic data produced within SURFING to study and characterize the extent of homologous recombination in *L. delbrueckii*. Even if bacteria multiply clonally, it is now known that genetic exchange between individuals exists and plays a major role in bacteria evolution (Thomas and Nielsen, 2005). Homologous recombination is one of the DNA exchange processes during which a bacterium replaces a fragment of its genome by a homologous sequence imported from another bacterium called donor. From a population point of view, this process has important effects on genetic diversity and could facilitate the response to natural selection (Vos, 2009). Almost all bacterial species are affected by this mechanism but the extent and the nature of recombination varies among species (Vos and Didelot, 2009). For example, *Helicobacter pylori* evolves mainly through recombination whereas recombination in species *Mycobacterium tuberculosis* (Namouchi et al., 2012) and other "genetically monomorphic" pathogens is less important: as their name states, they were at first thought to be monomorphic. Even though the recombination rate is usually lower than the mutation rate, a recombination event typically involves dozens or hundreds of base pairs (bp) and may introduce several substitutions, compared to a single bp and a single substitution for a mutation event, so that the overall effect of recombination in the observed genetic diversity could well overtake the effect of point mutations. In such a scenario, the recombination events could disrupt the clonal frame of the strains, weaken the phylogenetic signal present in complete genome alignments and eventually hinder the reconstruction of the strain genealogy.

The problems caused to phylogenetic reconstruction by the recombination led to the development of methods inferring the recombination history of a dataset (ClonalFrame (Didelot and Falush, 2007), RDP (Martin and Rybicki, 2000)). The decreasing cost of sequencing allowed us to apply recombination detection techniques at a genomic scale (Namouchi et al., 2012), instead of restricting ourselves to a few housekeeping genes as usually done in Multi Locus Sequence Typing. We applied an algorithm developed and implemented in ClonalFrame by X.Didelot (Didelot and Falush, 2007) to our dataset in order to assess the amount of recombination in our *L. delbrueckii* sample, to identify the recombinant parts of the genomes, and to verify whether the two subspecies behave in the same way regarding recombination. We attempted to retrace the possible origin of the imported tracts and cross-compared them to Gene Ontology (Tatusov et al., 2001) to check for significant enrichment in the number of imported tracts of any functional category. In the global context of SURFING, the knowledge of the recombination history is useful to reconstruct an effective genealogy for each gene. This genealogy is in turn useful to correct association tests of genes with properties of interest by the shared history of genes, and therefore for a rational selection of immunomodulatory candidate genes.

2. Materials and Methods

2.1. Dataset

The complete genomes of ten strains of *Lactobacillus delbrueckii* were used in this study. The genomes of three *L. bulgaricus* strains referred as "ATCC BAA-365", "ATCC 11842", and "LDB2038" were already publically available under the respective GeneBank Accession CP000412, CR954253, and CP000156. An additional *L. lactis* strain ("NDO2", gb CP002341) was also available but incorrectly referred as a *L. bulgaricus* strain. The 6 other genomes were sequenced during the project (Strains were selected for their immunomodulatory properties and provided by M.Van den Guchte). Among these ten strains, five belong to the subspecies lactis, and the others to the subspecies bulgaricus. The strains were grouped in 3 sets: the first contains all the 10 strains, the second regroups the 5 *L. bulgaricus,* and the third regroups the 5 *L. lactis*. The genomes were assembled and annotated before the onset of the internship.

For each of the 3 strain sets, two types of alignment were constructed, for a total of 6 datasets (Information is stored in the first part of table1):

- Coding DNA sequences (CDS) were predicted using the AGMIAL platform and clusters of orthologous genes were defined using orthoMCL before the onset of the internship. Only clusters that contain genes present in single copy in each strain were selected for analyses. The nucleotidic sequences were aligned using the software Tcoffee with no regards to the Open Reading Frame. The clusters alignments containing more than 10 polymorphic sites within a stretch of 15bp were checked visually (using Jalview version 2.7 (Waterhouse et al., 2009)) in order to detect and remove possible alignment problems. Hereafter, these alignments are referred to as coding genes.

- The complete genomes were aligned each of the strains sets. The core genomes of each subset were defined (using MOSAIC (Chiapello et al., 2008)) and used for analyses. Contigs shorter than 1kb were filtered. All the alignments were stored in xmfa format and imported in R for being analyzed.

2.2. Phylogenetic reconstruction

Phylogenetic trees were computed on the different datasets using two reconstruction methods. phyML version 3.0 (Guindon et al., 2010) was used to reconstruct maximum likelihood trees (substitution model: GTR +I +G selected using modelgenerator, 100 bootstrap replications were performed to infer the support values of the clades). UPGMA trees were reconstructed using the

package Phylip: distance matrixes were computed with DNAdist (Jukes-Cantor distance) and Neighbor was used to get the UPGMA tree. This was done on 100 bootstraps alignments. FigTree version 1.3.1 and R package ape were used to visualize phylogenies.

2.3. Clonal Frame analysis

We used ClonalFrame version 1.1 to infer the clonal genealogy and to detect tracts of homologous recombination between the strains or from an external donor organism.

For each data set, at least three independent runs were computed: two with flat priors (default values, see ClonalFrame user guide), and one with disadvantageous priors on R and ν values. The disadvantageous priors favor high values of R and ν far apart from expected values. The runs consist in 150,000 iterations, excepted for the 10strains – CDS dataset where only 100,000 iterations were performed. The first 50,000 iterations were discarded and the rest of the chain was sampled every 100 iterations. $\boldsymbol{\theta}$ was fixed to the estimate $\boldsymbol{\theta}$ of Watterson ($\boldsymbol{\theta}_{\text{Watt}}$) and the topology was fixed to the UPGMA.

2.4. Output analysis

All the ClonalFrame outputs analysis and interpretations were performed with scripts written in R (with the use of package Coda to monitor the Markov Chains Monte Carlo (MCMC) convergence and Ape to handle the phylogenies). The convergence of the runs was visually checked (figure 3) and statistically assessed using a Gelman-Rubin test. The output contains for each branch and each reference sites the probability of being recombined. The reference sites correspond to the first and last sites of each block (contigs for complete genomes alignments, genes for CDS alignments), polymorphic sites, and sites sampled every 50bp. The recombinant tracts were defined as segments containing only reference sites with a recombination probability over 0.5 and at least one site with a recombination probability over 0.95 (mean over the 3 runs). This extraction method differs from the one used in other studies (which consists in extracting the recombinant tracts as all the segments containing only referenced sites with a recombination probability over 0.95 (den Bakker et al., 2008)). This last method was used at first but, we observed close recombinant segments separated by stretches of sites with intermediate probability of recombination. These segments likely belong to the same tract, which was split thanks to the stringent 0.95 threshold used. To avoid this artifact we fused this kind of segments with the two thresholds extraction.

For the *L. bulgaricus* CDS datasets, COG classes were available (Tatusov et al., 2001). In a context of a homogenous distribution of recombination events along the genome, we expect the number of recombination events in occuring a gene (N_{gene}) to depend on its sequence length (L_{gene}). According to the model used in ClonalFrame, N_{gene} should follow a Poisson law of parameter $(\frac{R}{2} \times l_{tree} \times \frac{L_{gene}}{L_{total}})$ where l_{tree} represents the total tree length and L_{total} the alignment length. The total number of called events (N_{total}) is lower than the number $\frac{R}{2} \times l_{tree}$ expected under the Poisson model (see results). To filter out this skepticism when calling recombination and to assess only the homogeneity of the distribution of recombination events along the genome, we fit the distribution to the observed events and replace $\frac{R}{2} \times l_{tree}$ by N_{total}, the total number of observed recombination events. For each gene we tested whether it was likely or not to obtain the observed value of recombination events according to the Poisson law of parameter $(N_{total} \times \frac{L_{gene}}{L_{total}})$. We were then able to class the genes as over-recombined (p-value < 0.05 after FDR (false discovery rate) correction for multiple tests) or non-over recombined (pvalue ≥ 0.05).

2.5. Research of potential donor organisms

To infer the organisms recombining with *Lactobacillus delbrueckii*, we compared the recombinant sequences longer than 40bp against a large set of nucleotide sequences (the "nr" database of NCBI containing All GenBank + EMBL + DDBJ + part of PDB sequences) using Blast with default options. This was performed only for the recombinations occurring on the tips of the phylogeny because we did not have a direct access to the sequences at the internal nodes. First and unlike tip sequences, they need to be reconstructed, and second, in case of consecutive recombination events concerning the same tract, information about the imported sequence has been wiped out from observed sequences by subsequent recombinations.

2.6. SHOW

Show (Ibrahim et al., 2007) was used to infer recombination events based on the use of Hidden Markov Chain Model. The entire alignment was recoded as follow: 0 for monomorphic sites, 1 for polymorphic sites in agreement with the topology (obtained with phyML), 2 for homoplasic polymorphic sites, 3 for missing data. The algorithm was run to detect two hidden states that should correspond to the recombined or non-recombined sequences. The program show_emfit first determines the best fitting model which describes the composition in terms of sites of each hidden state. As a byproduct, the EM algorithm also gives the posterior probability that each site to belong to one or the other state and the transition using an EM algorithm. Then, the program show_viterbi generates an outfile containing the most likely hidden-state path.

3. Results

3.1. Phylogeny of Lactobacillus delbrueckii

The annotated complete genomes of ten strains were compared during this study (see material and methods). They equally sample the two subspecies *Lactobacillus delbrueckii bulgaricus* and *lactis*. We extracted two types of alignments: the core genome which consists of all the sequences shared by all the strains of the subset and the CDS alignments (all the sequences coding for orthologous genes present in single copy in all the strains of the subset). The first part of Table 1 summarizes the (2x3) different datasets obtained.

We first generated the phylogeny using the maximum likelihood (ML) algorithm implemented in the software phyML. The tree obtained is represented on figure 1. We obtained the same topology with the UPGMA algorithm. The two subspecies are well distinguished with the distance between the two clades at least 5 times greater than the depth of the clades. When adding an external strain of *Lactobacillus acidophilus*, the root is placed on the branch separating the two clades.

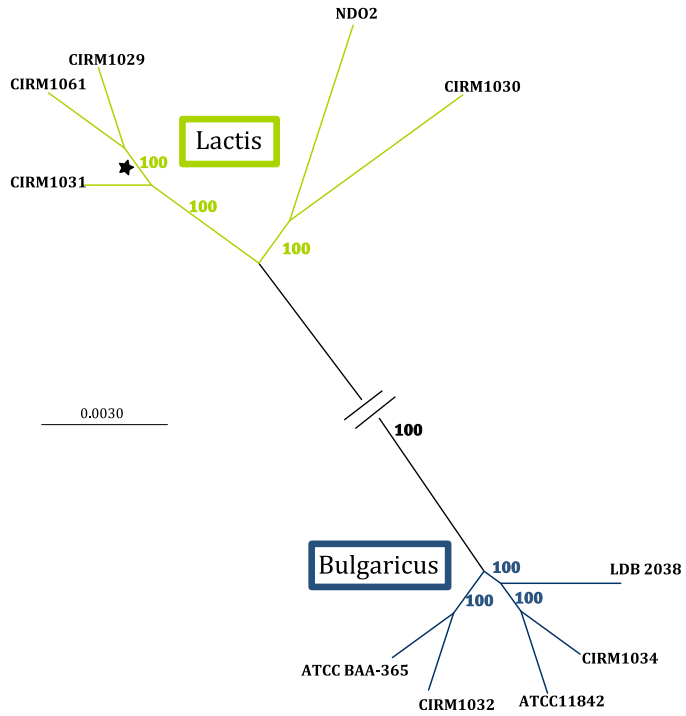

Figure 1: Maximum likelihood tree obtained with phyML on core genome sequence of the ten strains used in this study. The numbers on nodes indicate the bootstrap values. The scale is the number of mutations per site.

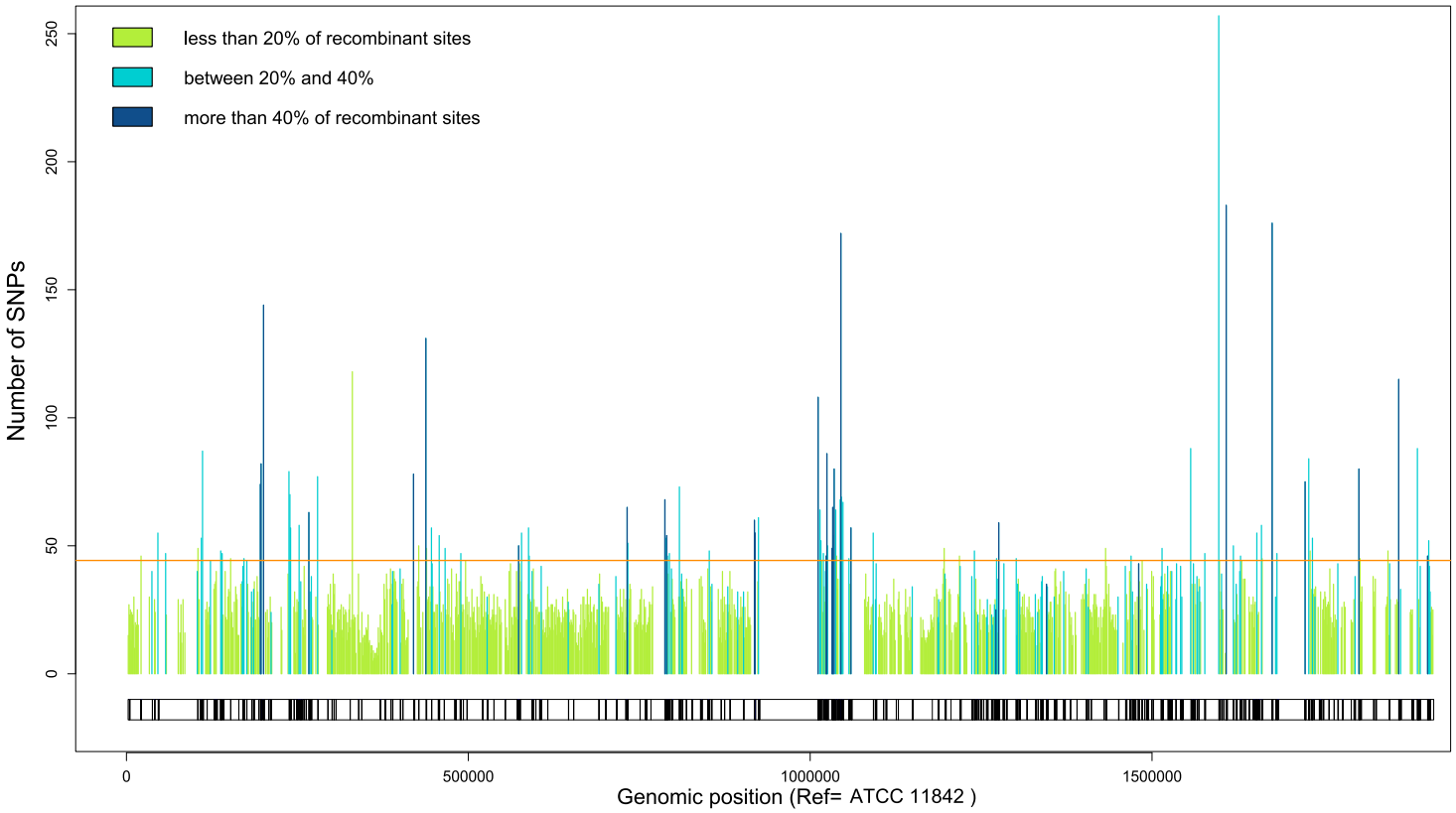

Figure 2: SNPs density along the core genome of 10 *LB delbrueckii* strains. Density of SNPs in non-overlapping 1-kb windows colored according to the amount of recombinant sites detected by ClonalFrame. The white zones correspond to gaps between contigs non-included in the core genome. The orange threshold line represents the amount of SNPs over which a region is significantly enriched in SNPs (binomial test, $p < 0.05$ after fdr correction). The black bars indicate the actual position of recombination events detected by ClonalFrame.

All the datasets confirm that the strain NDO2 is actually a *L. lactis* contrary to the information available from GenBank.

The topology obtained is robust, as confirmed by high bootstrap values, with the exception of the node indicated by a star which is resolved similarly for 3 datasets and a bit differently for the core genome of *L. lactis* strains only.

The different strains are strongly related with at least 97% of identity. The diversity appears to be non-homogeneously distributed over the core genome as presented on figure 2 (visualization of SNP density along the genome using a 1kb non-overlapping window). The orange threshold represents the limit over which a region is significantly enriched in SNP (Binomial Test, p-value<0.05 after FDR correction). 102 windows over 1016 exceed this limit. The content of these regions has not been studied yet.

3.2. Recombination history

Even if the clonal history obtained with ML is pretty well established, it is likely that recombination happened during the evolution of the clade as hinted by the non-homogeneous distribution of SNPs along the genome, and as suggested by previous studies (Nicolas et al., 2007). To test this hypothesis and detect putative recombinant tracts, we used the software ClonalFrame. This software is based on the assumption that recombination events introduce substitutions in a continuous region at a higher rate than expected under a pure mutation scheme. ClonalFrame works under a Bayesian framework and uses MCMC methods to infer parameters of interest and detect the recombinant stretches, whose

genealogy is probably in disagreement with the clonal relationship. The parameters used in the model are the following:

T: the genealogy, including topology and branch lengths measured in units of coalescent times;

θ: the mutation rate over the genealogy: the number of mutation events occurring on a given branch of length l follows a Poisson distribution with mean $\frac{\boldsymbol{\theta}}{2} \times l$;

R: the recombination rate over the genealogy: the number of recombination events occurring on a given branch of length l follows a Poisson distribution with mean $\frac{\mathrm{R}}{2} \times l$;

ν: the amount of nucleotide differences between the imported and the original sequences;

δ: the parameter of the exponential distribution of the tracts length (1/ δ gives represents the expected mean length of imported tracts);

All these parameters give access to the value "r/m" which is a value commonly used to measure the relative effect of recombination on the genetic diversity (Vos and Didelot, 2009; Namouchi et al., 2012). r/m=$\frac{R \times \boldsymbol{\nu}}{\boldsymbol{\theta} \times \boldsymbol{\delta}}$ is the ratio of the number of substitutions resulting from recombination and the number of substitutions resulting from mutation.

As ClonalFrame is based on MCMC, it is crucial to monitor the convergence of the chains by different methods, to assess the sensitivity of the results to the priors, and more generally to use several convergence diagnostic tools. To do so, we did three independent runs for each experimental condition with different priors values (see material and methods). The first runs of ClonalFrame were computed with flat priors for all the parameters (see ClonalFrame manual, defaults values). The chains did not converge because there was a strong correlation between R and θ. This is because a high amount of recombination combined with a high amount of punctual mutation and short branches can lead to the same pattern than lower values for both parameters combined with longer branch lengths. To fight against this internal mixing problem, we fixed the value of **θ** to the estimate of Watterson ($\boldsymbol{\theta}_{\mathrm{Watt}}$) which is an estimator of the population mutation rate based on the coalescent theory.

$\boldsymbol{\theta}_{Watt} = \frac{N_{seg}}{\sum_{k=1}^{n-1} \frac{1}{k}}$ where n is the number of strains and N_{seg} is the number of segregation sites.

We computed $\boldsymbol{\theta}_{Watt}$ on 20 bootstrap alignments to estimate the dispersion of theta. Values ranged in [7602-7825] quite close to the observed 7710 for the CDS alignment of the 10 strains. The algorithm was then able to converge but important computing efforts remained necessary, for example, 4 days of computing for the 150000 iterations on the CDS-5 *L. bulgaricus* dataset. To accelerate the convergence of the chains the topology was fixed to the UPGMA tree, which was the same as the ML tree for all the dataset (and the same as the topology obtained when running a short chain with no constraint on the topology). This concerned the topology only, branch lengths and nodes ages were still estimated. With the fixations the topology the runs were twice faster. We were then able to compute multiple long runs (150,000 iterations) for all the datasets. For each of the six datasets we computed 3 independent runs (see material and methods) to detect a possible influence of the priors on the posterior results. The convergence of the runs was satisfactory when tested with the Gelman-Rubin test (potential scale reduction factors < 1.12 for all parameters in all runs excepted for nu in the Lactis-CDS dataset which reaches 1.27), the mixing was also satisfactory as indicated by to good values of Effective Sample Size (ESS > 220) for each parameters. Figure 3A. illustrates the good congruency for the posterior density of r/m over the 3 runs. We also checked whether the different runs predicted the same recombinant tracts on the CDS datasets. We extracted the recombined sites for each of the runs independently and plotted the number of sites predicted as recombinant by one, two or three runs (Figure 3B). It appears that the prediction of tracts was really consistent over the runs: more than 88% of the sites predicted by at least one runs were also predicted by the two

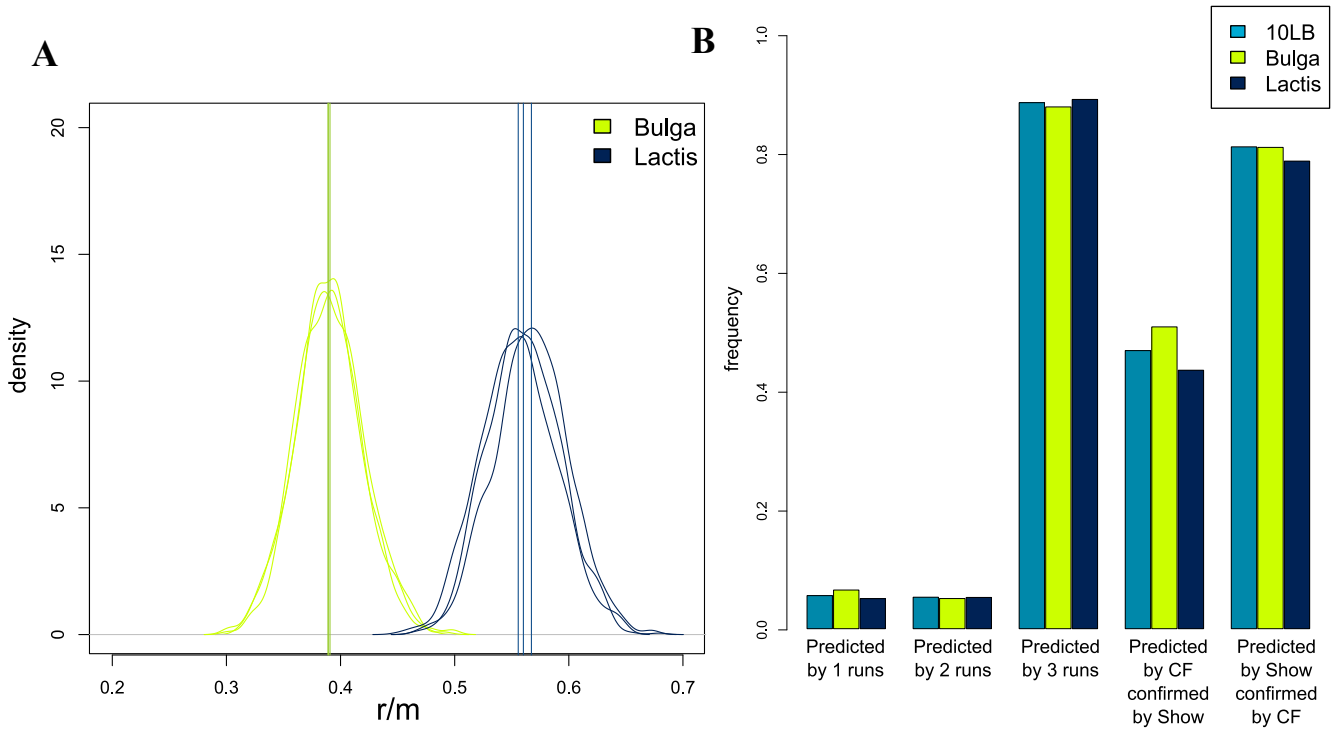

Figure 3: Convergence of ClonalFrame runs. A) Posterior density of the r/m parameter (relative importance of recombination and mutation) for the CDS alignments. Vertical lines indicate the median value for each run. B) Properties of the sites predicted as recombinant (see criteria in material and methods) by at least one run. (first three columns) Convergence of ClonalFrame and Show in terms of sites predictions.

others. We then extracted the tracts using the mean value of the 3 runs (see material and methods).

Table 1 summarizes the values obtained for each parameter and figure 4 shows the distribution of detected tracts over the genealogy. There is tremendous evidence for recombination, at least for recombination as it is modeled in ClonalFrame. This mechanism is around 15 times less frequent than point mutations (R compared to **θ**) but each single event introduces (1/δ)*ν = 6.5 substitutions in average so that overall, ClonalFrame evaluates at 28 to 40% the fraction of observed diversity that can be attributed to recombination (r/m). These value is close to the one obtained by Namouchi et al., 2012 with *Mycobacterium tuberculosis* but is lower than values obtained with highly recombinant species (e.g. r/m = 2.37-2.45 for *Bacillus cereus* (Didelot et al., 2010)). Furthermore, we can observe that *L. lactis* seems to be more affected than *L. bulgaricus*. We observe that the recombinant tracts detected for each datasets are quite homogeneous with comparable size (1/δ) and introduced diversity (ν). On figure 2, we observe that the recombinant tracts are not uniformly distributed along the genome. Some regions are almost unaffected by recombination (in green) when others present high amount of recombinant tracts (in blue). The regions enriched in SNPs correspond (with few exceptions) to these blue regions. This was also observed for the CDS datasets where the distribution of tracts allowed us to detect some "over-recombined" genes: genes with a significantly higher number of recombination than expected. The complete genomes datasets lead to a higher number of recombinations events, even when normalized by the alignment length; this can be explained by a more important conservative selective pressure on coding sequences. But as shown on figure 4, the proportions of recombinations on each branches of the tree are quite consistent. The detection is particularly efficient on the tips, but for the internal nodes there is a greater uncertainty about the sequences and the recombination events are more difficult to detect. This can explain why we obtain less recombinant tracts than expected under the Poisson law of parameter $\frac{R}{2} \times l_{tree}$. E.g: For the 5 *L. bulgaricus*-CDS dataset we expected around $\frac{R}{2} \times l_{tree} = 417$ events over the tree but we obtained only $N_{total} = 132$. This may also be partly due to our method of calling the events, which is highly conservative.

	Complete core genomes alignments			CDS alignments		
	10 Strains	**5 lactis**(1)	**5 bulga**(1)	**10 Strains**	**5 lactis**	**5 bulga**
Info datasets						
Total length	1,169,615	1,377,054	1,532,043	794,191	922,138	963,023
Number of contigs / CDS	322	308	210	769	920	946
Polymorphic sites	34,720 (3.00%)	22,703 (1.65%)	15,390 (1.00%)	21,813 (2.75%)	12,606 (1.37%)	7,962 (0.83%)
Θ_{Watt}/site (x10^{-3})	10.47	7.915	4.825	9.707	6.54	3.98
ClonalFrame Results						
R/site (x10^{-3})	0.928 [0.855-1.005]	0.681 [0.622-0.744]	0.454 [0.411-0.499]	0.632 [0.554-0.713]	0.54 [0.475-0.612]	0.241 [0.202-0.281]
ν (x10^{-2})	3.00 [2.87-3.14]	5.56 [5.31-5.84]	4.39 [4.15-4.64]	4.14 [3.84-4.45]	4.28 [3.94-4.63]	4.91 [4.49-5.36]
1/ δ	202 [188-218]	142 [130-155]	136 [125-150]	144 [130-160]	159 [141-180]	131 [115-151]
r/m (2)	0.537(35%)	0.679(40%)	0.563(36%)	0.388(28%)	0.55(35%)	0.389(28%)
Recombination events	559	419	350	301	229	132
Tracts Size (median-mean)	203-291	141-204	114-189	141-202	171-232	118-191
SHOW Results						
Recombination events	123	183	104	68	84	40
Tracts Size	142-240	116-209	126-254	249-420	185-298	124-274

Table1: Information about the different datasets used during the study and principal results obtained using ClonalFrame and SHOW programs.

[] 95% IC

(1) Bulga= *L. bulgaricus* Lactis= *L. lactis*

(2) the percentages indicate the fraction of observed diversity that can be attributed to recombination

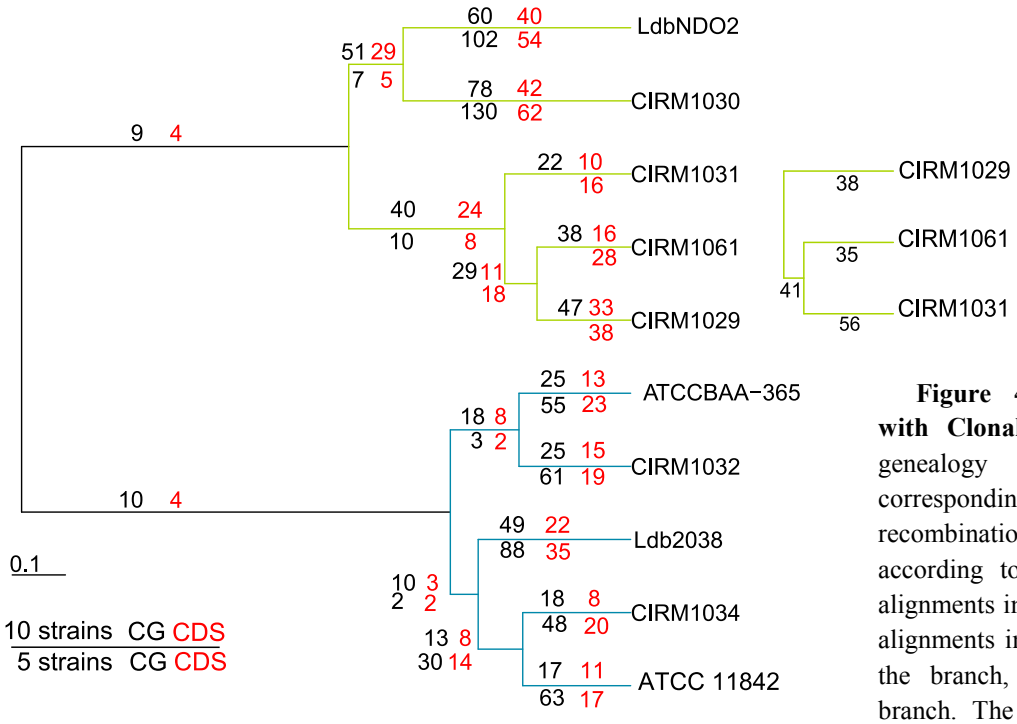

Figure 4: Genealogy reconstructed with ClonalFrame. Each branch of the genealogy is labeled with 4 values corresponding to the number of detected recombination events affecting the branch according to the different datasets: CDS alignments in red, Complete Genomes (CG) alignments in black, 10 strains dataset over the branch, 5 strains dataset under the branch. The alternative topology obtained with the 5 lactis-Complete genome dataset is represented on the right.

Beside ClonalFrame analysis, we tried to infer recombinant tracts using a simple method based on a recoding of the alignment. SHOW is a program using Hidden Markov Model (HMM). This Bayesian model assumes the existence of hidden states, that are not observable, and observed states (here the DNA sequence) whose distribution depends on the hidden states. Here, we considered a HMM with two hidden states, corresponding respectively to recombinant and non-recombinant regions, and aimed at determining the hidden state of each position. To do so, we coded the sequence as follows: 0 if the site is monomorphic, 1 if the site is polymorphic but in agreement with the phyML tree (i.e. not homoplasic), 2 if the site is homoplasic with respect to the tree, 3 if the site contains any gap. We expected the two hidden states. One state should correspond to non-recombinant segments where the diversity is low (only due to single substitutions) and should be in agreement with the clonal history. The other one, corresponding to recombinant tracts, should be characterized by a higher rate of substitutions and homoplasies in particular. SHOW can estimate the distribution of observed states (i.e. 0 to 3) conditional on the hidden states, the transition probability from one hidden state to the other, and the most likely hidden path given the observed states. We compared the results obtained to the ClonalFrame results (table 1 and figure 3B). Almost all the sites that turn out to be recombined according to SHOW are also detected by ClonalFrame. But this method infers less recombinant sites than ClonalFrame algorithm, only half of the recombinant sites detected by ClonalFrame are confirmed by SHOW analysis. There is a difficulty to clearly define the limits of the tracts in SHOW analysis which can explain this result, but it is also because the two algorithms are not able to detect the same kind of tracts as discussed below. Note that this analysis was only possible because there is a strong overall support in the alignment for a given topology.

3.3. Functional classes affected and origin of imports

For the CDS dataset of the 10 *L. delbrueckii* strains, we obtained a total of 301 recombination events over the 794,191bp alignment, affecting a total of 157 genes. As explained in "material and methods" section, we tested the homogeneity of the recombination event distribution over the genome. We determined that this distribution was not homogeneous. Actually, 41 clusters were more recombined than expected according to their size, and contain more than half the events. Furthermore, we tested whether certain functional categories had been enriched in recombination during the evolutionary history (COG ref). The figure 5 represents the composition of each functional category in terms of non-recombinant, non-over recombinant or over-recombinant genes. It appears that the class J ("Translation, ribosomal structure and biogenesis") is under-affected by recombination. On the contrary the N class ("Cell motility and secretion") presents a higher number of recombination events than expected under a homogeneous distribution. However, this is mainly due to the fact that this class is represented by only 6 genes: three of them contain imported tract and, more importantly, a gene coding for an N-acetyl muramidase presents 8 recombination events.

ClonalFrame is designed to detect segments with a mutation rate higher than expected and therefore likely to have undergone a recombination event. But there is no inference about the origin of the import. To test which organisms could have donated genetic material to *Lactobacillus delbrueckii* we compared the recombinant segments inferred by ClonalFrame with the genomes of other bacteria. We did not find any tract to match exactly to any bacterium outside of the *L. delbrueckii* species. This is maybe due to the lack of sequenced genomes of other bacteria closely related to, but different from *L. delbrueckii* in the nr database.

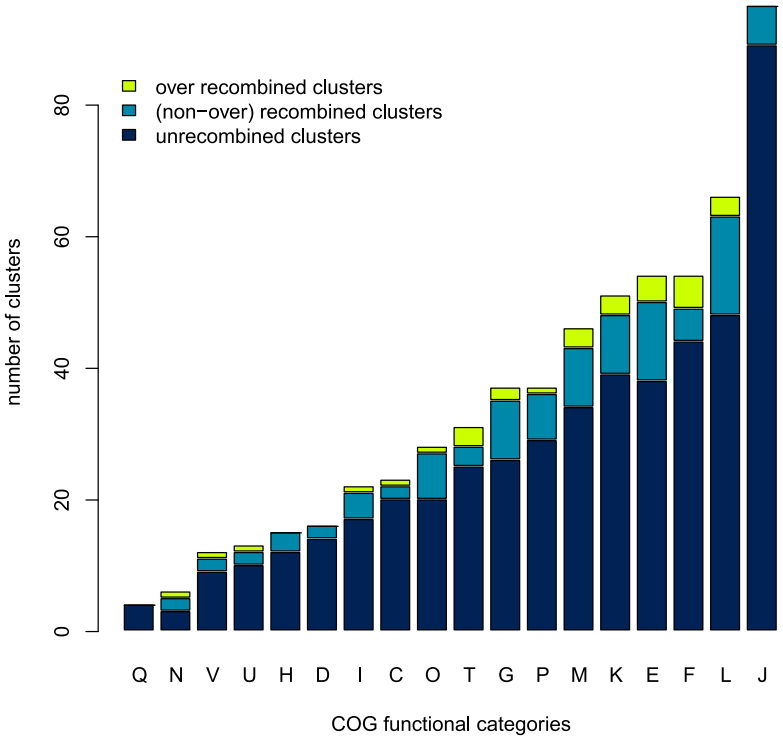

Figure 5: Distribution of recombination events over the functional categories. The barplot represents the composition of each COG functional category in terms of un-recombined, recombined, and over-recombined clusters. (N) Cell motility and secretion; (V) Defense mechanisms; (U) Intracellular trafficking, secretion, and vesicular transport; (H) Coenzyme transport and metabolism; (D) Cell division, cell cycle control; (I) Lipid transport and metabolism; (C) Energy production and conversion; (O) Post-translational modification, protein turnover, chaperones; (T) Signal transduction mechanisms; (G) Carbohydrate transport and metabolism; (P) Inorganic ion transport and metabolism; (M) Cell envelope biogenesis, outer membrane; (K) Transcription; (E) Amino acid transport and metabolism; (S) Function unknown; (L) DNA replication, recombination and repair; (R) General function prediction only; (J) Translation, ribosomal structure and biogenesis.

4. Discussion

Using ClonalFrame and SHOW algorithms, we were able to detect segments that are likely to have undergone recombination. Even though the clonal frame of the strains is quite well resolved and supported, the amount of recombination is significant. According to our results, up to 40% of the polymorphism could result from recombination events (r/m=0.68). We have observed a higher rate of recombination affecting *L. lactis* than *L. bulgaricus,* suggesting that recombination is a more powerful evolutionary mechanism in *L. bulgaricus.* This could be due to differences affecting their Restriction-Modification System which can detect and destroy foreign DNA. Indeed, Van de Guchte et al., 2006, have shown that this system is active in *L. bulgaricus.* It was one explanation of the difficulties encounter during transformation; it could also affect the recombination process. Assessing whether it is also active in *L. lactis* could therefore be interesting.

The program used to detect recombination failed to estimate all the parameters because of strong correlations between the parameters as explained in the results. We chose to estimate $\boldsymbol{\theta}$ with $\boldsymbol{\theta}_{Watt}$ which does not take into account the diversity generated by the recombination. The contribution of punctual mutations was then over estimated but it allowed us to be more stringent on the recombination detection and made the chains converge. The use of different datasets showed a relative good convergence so that it will not be necessary to perform this analysis on both CDS and complete genome in the future. Likewise, it seems that working on the entire set of strains is sufficient. This is because we disposed of a high quantity of genetic material for few strains contrary to MLST studies where the split into sub-lineage is preferable (den Bakker et al., 2008).

SHOW algorithm detects less recombinant tracts but almost all its predictions are confirmed by ClonalFrame predictions. This can be explained by the fact that there are only two hidden states, which can be interpreted as recombined or un-recombined. This involves homogeneity of recombinant tracts concerning their composition in homoplasic and polymorphic sites. But actually there are different possibilities for the origin of the tract. If the recombination occurs between strains of the data set, it will introduce homoplasies. On the contrary if the recombination happens with an external donor, it will introduce a higher rate of polymorphism but no homoplasies. With the code used, SHOW is likely to be efficient for detection of intra dataset recombinations but not for imports from an external strain. Furthermore, SHOW does not detect where, in the genealogy, the recombination happened. Finally, SHOW requires a good topology to be known in order to recode the genome. However, this algorithm is really faster than ClonalFrame (5 min in mean, 3 days for ClonalFrame) so it could be used as a preliminary analysis. It is still possible to test different encoding in order to obtain a better detection.

The tracts detected by ClonalFrame which are not confirmed by SHOW could be attributed to recombination events with organisms not included in the dataset. It has been proved that recombination occurs more frequently between closely related species (Didelot and Maiden, 2010). Actually, the need of a physical contact for the uptake of DNA to occur favors the exchange between bacteria of a same community, which are likely to be close because of geographical and ecological structuring. Moreover, the genetic proximity favors the acceptance of foreign DNA by the bacteria and recombination events affecting genes are more likely not to be purged by selection as their function should be adapted to the same conditions. Therefore, we could have expected clue of recombination between *Lactobacillus delbrueckii* and other bacteria of group Lactobacillus such as *Lactobacillus acidophilus.* Moreover, *Lactobacillus delbrueckii* is used as a starter in association with *Streptococcus thermophilus.* It would have been interesting to find evidence for recombination between these two organisms. However, we failed to infer the origin of

recombinant tracts by blasting them against the nr database. The lack of results is not surprising when working on such bacteria. Indeed, despite their importance in the dairy industry, *Lactobacillus* bacteria are not as extensively studied as pathogenic bacteria for example. Therefore, the nr database does not contain a lot of closely related bacteria. A study on a larger dataset containing more *Lactobacillus* and *Streptoccocus* genomes could reveal interesting results (Didelot et al., 2010).

Concerning the distribution of recombination, we have seen that COG Class J (Translation, ribosomal structure and biogenesis) is under affected: this is not surprising and could be due to a higher conservation on ribosomes sequences as observed for the mutations. But even if recombination occurs in same amount on these sequences and is not counter selected, the diversity introduced will be low because the sequences are more conserved among species, the algorithm will therefore be unable to detect it. There is no evidence for a significant enrichment of a functional category in recombination. But at the level of genes there is an important variability, with some genes being over affected. It would be interesting to study specifically these genes. The pattern of recombination events affecting each gene will be useful when trying to select candidate genes for immunomodulatory properties.

To conclude, we have shown the interest of using ClonalFrame to detect recombination in our *Lactobacillus delbrueckii* dataset. The next step of this study will be to continue the research of tracts origin by including other *Lactobacillus* and *Streptoccocus thermophilus* genomes as mentioned above. We will also study the dN/dS composition of the recombinant tracts in order to detect a possible effect of selection (Vos, 2009). Finally, the same kind of analysis will be performed on *Propionibacterium freundenreichii*, another dairy industry species studied in the SURFING project.

Acknowledgments

I would like to thank all the team who helped me to perform this internship and who welcomed me in such a friendly manner. I am especially grateful for the help and supervision of Mahendra Mariadassou and Philippe Bessières. I would also thank the members of the SURFING project: Julien Buratti, Hélène Chiapello, Valentin Loux and Pierre Nicolas for useful discussions and data providing.

Contributions: Mariadassou, M. and Bessières, P. designed and supervised the internship; Chiapello, H. made the genomics alignments; Loux, V. performed the CDS prediction; Buratti, J. constructed the orthologous clusters. Blin, C. performed the analysis and wrote the report with useful reading and comments from Mariadassou, M., Bessières, P. and Nicolas, P.

References

den Bakker, H.C., Didelot, X., Fortes, E.D., Nightingale, K., and Wiedmann, M. (2008). Lineage specific recombination rates and microevolution in Listeria monocytogenes. BMC Evolutionary Biology *8*, 277.

Chiapello, H., Gendrault, A., Caron, C., Blum, J., Petit, M.-A., and El Karoui, M. (2008). MOSAIC: an online database dedicated to the comparative genomics of bacterial strains at the intra-species level. BMC Bioinformatics *9*, 498.

Didelot, X., and Falush, D. (2007). Inference of bacterial microevolution using multilocus sequence data. Genetics *175*, 1251–1266.

Didelot, X., Lawson, D., Darling, A., and Falush, D. (2010). Inference of homologous recombination in bacteria using whole-genome sequences. Genetics *186*, 1435–1449.

Didelot, X., and Maiden, M.C.J. (2010). Impact of recombination on bacterial evolution. Trends in Microbiology *18*, 315–322.

FAO/WHO (2001). Report on Joint FAO/WHO Expert Consultation on Evaluation of Health and Nutritional Properties of Probiotics in Food Including Powder Milk with Live Lactic Acid Bacteria.

Van de Guchte, M., Penaud, S., Grimaldi, C., Barbe, V., Bryson, K., Nicolas, P., Robert, C., Oztas, S., Mangenot, S., Couloux, A., et al. (2006). The complete genome sequence of Lactobacillus bulgaricus reveals extensive and ongoing reductive evolution. Proc Natl Acad Sci U S A *103*, 9274–9279.

Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. (2010). New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. Syst Biol *59*, 307–321.

Hao, P., Zheng, H., Yu, Y., Ding, G., Gu, W., Chen, S., Yu, Z., Ren, S., Oda, M., Konno, T., et al. (2011). Complete Sequencing and Pan-Genomic Analysis of Lactobacillus delbrueckii subsp. bulgaricus Reveal Its Genetic Basis for Industrial Yogurt Production. PLoS One *6*,.

Ibrahim, M., Nicolas, P., Bessières, P., Bolotin, A., Monnet, V., and Gardan, R. (2007). A Genome-Wide Survey of Short Coding Sequences in Streptococci. Microbiology *153*, 3631–3644.

Makarova, K., Slesarev, A., Wolf, Y., Sorokin, A., Mirkin, B., Koonin, E., Pavlov, A., Pavlova, N., Karamychev, V., Polouchine, N., et al. (2006). Comparative genomics of the lactic acid bacteria. Proc. Natl. Acad. Sci. U.S.A. *103*, 15611–15616.

Martin, D., and Rybicki, E. (2000). RDP: detection of recombination amongst aligned sequences. Bioinformatics *16,* 562–563.

Namouchi, A., Didelot, X., Schöck, U., Gicquel, B., and Rocha, E.P.C. (2012). After the Bottleneck: Genome-Wide Diversification of the Mycobacterium Tuberculosis Complex by Mutation, Recombination, and Natural Selection. Genome Res. *22*, 721–734.

Nicolas, P., Bessières, P., Ehrlich, S.D., Maguin, E., and van de Guchte, M. (2007). Extensive horizontal transfer of core genome genes between two Lactobacillus species found in the gastrointestinal tract. BMC Evolutionary Biology *7*, 141.

Serror, P., Sasaki, T., Ehrlich, S.D., and Maguin, E. (2002). Electrotransformation of Lactobacillus Delbrueckii Subsp. Bulgaricus and L. Delbrueckii Subsp. Lactis with Various Plasmids. Appl. Environ. Microbiol. *68*, 46–52.

Tatusov, R.L., Natale, D.A., Garkavtsev, I.V., Tatusova, T.A., Shankavaram, U.T., Rao, B.S., Kiryutin, B., Galperin, M.Y., Fedorova, N.D., and Koonin, E.V. (2001). The COG database: new developments in phylogenetic classification of proteins from complete genomes. Nucleic Acids Res. *29*, 22–28.

Thomas, C.M., and Nielsen, K.M. (2005). Mechanisms of, and barriers to, horizontal gene transfer between bacteria. Nat. Rev. Microbiol. *3*, 711–721.

Vos, M. (2009). Why do bacteria engage in homologous recombination? Trends in Microbiology *17*, 226–232.

Vos, M., and Didelot, X. (2009). A comparison of homologous recombination rates in bacteria and archaea. Isme J *3*, 199–208.

Waterhouse, A.M., Procter, J.B., Martin, D.M.A., Clamp, M., and Barton, G.J. (2009). Jalview Version 2—a Multiple Sequence Alignment Editor and Analysis Workbench. Bioinformatics *25*, 1189–1191.