Universidade do Minho
Escola de Engenharia

Tiago Filipe Pereira Isabelinho

Comparative analysis and characterization of industrial *Streptococcus thermophilus* genomes

April 2019

Universidade do Minho
Escola de Engenharia

Tiago Filipe Pereira Isabelinho

Comparative analysis and characterization of industrial *Streptococcus thermophilus* genomes

Dissertação de Mestrado
Mestrado em Bioinformática

Trabalho efetuado sob a orientação de

Oscar Dias, Universidade do Minho

Hüseyin Demirci, Universidade do Minho

Martin Holm Rau, Chr. Hansen A/S

Ahmad Zeidan, Chr. Hansen A/S

April 2019

Acknowledgements / Agradecimentos

Este trabalho não seria possível sem um grande grupo de pessoas que me apoiou durante esta etapa. Agradeço o apoio a todas estas pessoas que sem elas não teria conseguido superar este objetivo.

Em primeiro lugar, gostaria de agradecer aos meus orientadores, professor Oscar Dias e doutor Huseyin Demirci por todos os conselhos e disponibilidade demonstrada desde o inicio do trabalho. Obrigado por todo o suporte no desenrolar do trabalho e tempo despendido para me ajudar e rever o meu trabalho.

Um especial obrigado ao Ahmad Zeidan e Martin Rau por esta oportunidade na Dinamarca, por todas as horas dispensadas em orientar-me, ensinar e apoiar quando situações complicadas surgiam. Obrigado por me ajudarem a crescer enquanto profissional, a poder trabalhar autonomamente sempre com os vossos conselhos.

A todos os meus colegas na Dinamarca, pelo companheirismo e amizade que vou guardar mesmo depois do final deste mestrado. Obrigado por me darem a conhecer a Dinamarca, a sua cultura e o seu estilo de vida. Foram um suporte fora do meu país. Obrigado Lisandra, Iuliana e Seawen.

Obrigado muito especial a todos os meus amigos por todos os momentos fantásticas que me proporcionaram. Vocês estiveram sempre lá para me animar e descontrair ao longo desta jornada. Ajudaram-me a superar obstáculos e sempre tiveram palavras de ânimo para não desistir dos meus objetivos.

Especial agradecimento para a minha família, principalmente aos meus pais e à minha irmã, por todo o carinho, força e conselhos que me deram ao longo da vida. Por todas as vezes que lidaram comigo nos momentos menos bons e pela paciência e dedicação.

A todas estas pessoas digo um enorme obrigado, sem vocês não teria conquistado o que consegui até hoje. Obrigado!

Resumo

A evolução de técnicas de sequenciação de nova geração originou um crescimento exponencial do número de genomas sequenciados. O pan-genoma permite uma visão global de múltiplos genomas. Este é definido como o completo conjunto de genes presente nos genomas em estudo, sobre os quais estabelecem-se relações de ortologia. Neste trabalho, uma visão geral e avaliação da performance de quatro diferentes ferramentas (FindMyFriends, OrthoMCL, Proteinortho e Roary) de deteção de ortólogos foi efectuada com uma similaridade de 0.95 e a não separação de parálogous em 21 genomas públicos. As ferramentas apresentaram uma performance global positiva com diferentes aplicabilidades dependendo do objectivo do trabalho.

Streptococcus thermophilus é uma bactéria ácido-láctica amplamente usada na produção de iogurte e queijo, associada com o seu sabor e textura. O estudo do seu pan-genoma envolvendo 333 genomes foi estabelecido pelo FindMyFriends com o objectivo da caracterização genómica para perceber a diversidade da espécie e características para o seu desenvolvimento. Correspondência entre o pan-genome e a base de dados KEGG foi efectuada através do K number, de modo a, identificar e/ou reconstruir o metabolismo dos carbohidratos, vias de biossíntese dos aminoácidos, transportadores ABC de aminoácidos e enzimas proteolíticas. A disrupção do gene glucokinase nas estirpes S e GA faz delas bons candidatos a aumentar a doçura do iogurte. Cerca de 21% das estirpes são auxotróficas para a histidina. Situações pontuais de auxotrofia foram encontradas em relação a outros aminoácidos. A protease extracelular PrtS não foi identificada em 153 estirpes, devendo estas ser co-cultivadas ou depender de outras proteases extracelular para atingir uma taxa de crescimento ótimo e acidificação. Em relação às peptidases intracelulares, quatro estirpes podem ter uma menor taxa de crescimento e menor composição de aminoácidos, percussores de compostos aromáticos. Uma maior capacidade de transporte de polipeptídeos poderá ser notória em 10 estirpes pela identificação de uma terceira cópia do gene *oppA*.

Existe sequências de DNA que influenciam e ajudam na iniciação do processo de transcrição e tradução de genes. Identificação dos promotores, caixa de pribnow e sequência -35, e da sequencia de Shine-dalgarno (SDS), dos genes pertencentes ao pan-genoma mostrou que a localização dos promotores é variável ao invés da SDS que se localiza nos 20bp a montante do sítio de iniciação da tradução. Este estudo do pan-genoma facilitará estudos fenotípicos para a fermentação do leite, melhorando os produtos lácteos.

Palavras chave: *Streptococcus thermophilus*; Pan-genome; Melhoramento; Síntese de Aminoácidos; Protéases; Transportadores; Identificação; Motifs

Abstract

The evolution of next-generation sequencing techniques gave rise to an exponential growth of the number of sequenced genomes. Pan-genome allows a global vision of multiple genomes. It is defined as the entire set of genes present in a group of genomes under study, on which orthology relationships are established. In this work, an overview and performance assessment of four different orthologous detection tools (FindMyFriends, OrthoMCL, Proteinortho e Roary) was done with a 0.95 similarity parameter and not splitting paralogous, in 21 public genomes. Tools show a good overall performance with different applicabilities depending on the work' goals.

Streptococcus thermophilus is a lactic acid bacterium widely used as a dairy starter to yogurt and cheese production, associated with their flavour and texture. Pan-genome study involving 333 genomes was stablished by FindMyFriends with the goal of genomic characterization to understand strain diversity and characteristics for their development. Matching between pan-genome and KEGG database through K numbers was done in order to identified and/or reconstructed carbohydrate metabolism, amino acid biosynthesis pathways, amino acid ABC transporters and proteolytic enzymes. Glucokinase gene disruption on strain S and GA make them good candidates to increase naturally yogurt sweetness. About 21% strains are histidine auxotrophic. Other residual auxotrophic situations were found regarding other amino acids. Extracellular protease PrtS was not identified in 153 strains, which must be co-cultivated or rely in other extracellular proteases to reach an optimal growth rate and acidification. Regarding intracellular peptidases, four strains could have a lower growth rate and a low amino acid composition, precursors of aromatic compounds. A greater transport capacity may be notorious in 10 strains by the identification of a thirth copy gene *oppA*.

There are DNA sequences that affect and help in the start of the gene transcription and translation process. Identification of the promoters, pribnow box and sequence -35, and Shine-Dalgarno sequence (SDS) from genes belonging to pan-genome show that promoters' location is variable instead of SDS that is in the 20bp upstream of the translation initiation site. This pan-genome study will facilitate phenotypic studies for milk fermentation, improving dairy products.

Keywords: *Streptococcus thermophilus*; Pan-genome, Improvement; Amino Acids Biosynthesis; Protéases; Transporters; Identification; Motifs

Contents

List of Figures

List of Tables

List of Abbreviations and Acronyms

BBHs	Bidirectional Best Hits
BR	KEGG Brite
BLAST	Basic Local Alignment Search Tool
BSR	BLAST Score Ratio
COG	Clusters of Orthologous Groups
CDS	Coding Sequence
CGN	Conserved Gene Neighborhood
GO	Gene Ontology
HGT	Horizontal Gene Transference
HMM	Hidden Markov Model
IGRs	Intergenic Regions
KEGG	Kyoto Encyclopedia of Genes and Genomes
KO	KEGG orthology
LAB	Lactic Acid Bacteria
MCL	Markov Clustering
MD	Module
NGS	Next-generation Sequencing
NCBI	National Center of Biotechnology Information
ORFs	Open Reading Frames
PATH	KEEG pathway maps
RBS	Ribosomal Binding Site
rRNA	Ribosomal Ribonucleic Acid
SDS	Shine-Dalgarno Sequence
SNPs	Single Nucleotide Polymorphism
TIS	Translation Initiation Site
TSS	Transcriptional Start Site
UniProt	Universal Protein Resource

1. Introduction

1.1. Context and Motivation

Milk and dairy products are consumed by billions of people around the world, playing a key role in human nutrition as a source of macro- and micronutrients. However, the usage of these products in human nutrition has been increasingly debated in recent years. Consumption of dairy products are determined by economic factors, such as income levels and relative prices, demographic factors, such as, urbanization, and social and cultural factors. As salaries increase, it is expected that dairy products sales will increase more than any other food product. In the recent decades, this fact was observed in developing countries where levels per capita consumption has increased rapidly, driven by economic growth. Technological development has resulted in major increases in productivity [1]. So, trading of dairy products has a great economic impact with an annual market of around 40 billion United States Dollars (USD)[2].

Streptococcus thermophilus is an industrially significant specie widely used as a major dairy starter for products, such as, cheese and yogurt. *Streptococcus* genus comprises commensal and opportunistic pathogen species, with *Streptococcus thermophilus* being the only food species, and also, the only one considered as 'Generally Regarded As Safe' specie [3]. Normally, it is combined with *Lactobacillus bulgaricus* for yogurt production but it also could be used solely or in combination with genus *lactobacilli* for cheese production [2]. A larger number of different strains are used commercially in order to ensure robust production and provide product diversity [4]. These strains are associated with food biopreservation, quality of fermented food products, for example, with their flavour and texture and has also been linked to probiotic effects like alleviation of lactose intolerance and modulation of intestinal microbiota. It can be linked directly or indirectly to the gene content of the strains used in fermentation [5].

In an evolutionary context, it has to cope with changing environments, with biotic and abiotic constraints in milk, yogurt and in digestive tract following ingestion, all of which being involved in the shaping of its genome [3]. It follows an evolutionary path divergent of the pathogenic species through loss-of function events and lateral gene transfer with other dairy species that contribute to their adaptation to the milk

environment [2]. *S. thermophilus* is still undergoing a process of reductive evolution towards an adapted bacterium to a relatively new environment, the milk niche [5]. Numerous genomic islands have unique gene features encoding several important industrial phenotypic traits. In effect, considerable strain variability exists and that has yet to reach an equilibrium [3]. Bacterial strains isolated from natural habitats can show phenotypic differences or interstrain phenotypic variability [6].

Next-generation sequencing (NGS) technologies led to the sequencing of thousands of bacterial genomes of commercially significant organisms, which are available in public domains [7]. With the increase genomic information, comparative genomic analyses provides powerful tools to analyze multiple genomes, which allows to extract information about strains' variability and the relations between them [8].

Chr. Hansen is a global bioscience company that develops and produces natural solutions, such as, cultures, enzymes, probiotics and natural colors for a rich variety of foods, confectionery, beverages, dietary supplements and even animal feed and plant protection. The company is the global market leader in dairy ingredients where their natural ingredients are consumed every day more than 1 billion people around the world. It is the owner of one the world's largest commercial collections of bacteria where bacteria are screened, selected and improved to meet specific requirements in their application. *Streptococcus thermophilus* strains from dairy environments are isolated being part of their bacteria collection. The company has been sequencing these strains that might be analyzed using several bioinformatics tools to perform comparative genomics and genomic motif prediction, together with the interpretation of physiological data and literature analysis, to perform a genomic characterization of industrially relevant *S. thermophilus* strains from the Chr. Hansen Culture Collection along with publicly available *S. thermophilus* genomes. This will yield a better understanding on strain variability and aid in the rational improvement and selection of commercial dairy strains used for making yogurt.

1.2. Objectives

The main goal of this work is the characterization of the sequenced *S. thermophilus* strain genomes belonging to the Chr. Hansen Culture Collection to understand strain diversity and characteristics for their development. Together with these genome strains, the work also includes publicly available genome strains.

In detail, this work aims at performing the following approaches:

- Comparative genomics analysis of *S. thermophilus* from the Chr. Hansen strain collection and publicly available strains, such as pan, core, accessory genome and unique genes;
- Comparing the performance of selected comparative genomics methods in the context of *S. thermophilus*;
- Improved annotation of accessory genes for predicting strain-specific functional characteristics;
- Gene structural elements analysis, i.e. identifying promoter motifs and subsequent genome-wide prediction of promoters;
- Genome-wide prediction of other regulatory motifs (e.g. transcription factor binding sites);
- Identifying strain variations in gene structural elements (promoter, operator and Shine Dalgarno sequences);
- In depth description of inter-strain variation in genomic regions encoding industrially relevant traits and investigation of correlations between genomic signatures and known phenotypes.

1.3. Structure of the Document

Chapter 2

State of the Art

Introduction of comparative genomics and pan-genome concepts. Overview about the general approach to create a pan-genome and applications. Brief presentation of available pan-genome creation tools, bioinformatic databases and platforms.

Chapter 3

Streptococcus thermophilus

Description about the lactic acid bacteria *Streptococcus thermophilus*, their food importance and its relevance to Chr. Hansen company.

Chapter 4

Material and Methods

Brief tutorial of four selected orthologous detection tools. Parameters settings, inputs and outputs.

Description of the data, pre-processing, *Streptococcus thermophilus* pan-genome creation, annotation improvement, and their match to desired genes as well as identification of promoters and shine-dalgarno sites.

Chapter 5

Results and Discussion

Assessment of orthologous detection tools.

Presentation of the main results generated in the thesis and discussion of how these finding could affect *Streptococcus thermophilus* metabolism.

Chapter 6

Conclusion and further work

Main conclusion of the thesis followed by a description of the possible future work.

2. State of the Art

Before the genomic era, genetic engineering enhanced production yields of naturally isolated organisms by random mutagenesis followed by screening and selection of desired phenotype. To further explore strain improvement, this approach does not answer what specific genetic perturbation led to this [9]. Listing all genes and proteins is not enough to understand the complex systems behind living organisms. It is necessary to know how all parts are assembled within each component, by gene regulatory networks and their biochemical interactions, and how each component dynamically interact with each other. System biology field studies an organism as an unique entity, allowing us to understand their behavior to the level of the system in response to stimuli [10].

The advent of NGS technologies brings a substantial reduction in costs and an increase in throughput and accuracy allowing that many *de novo* sequencing projects of new species started. It had resulted an exponential increase in the number of complete genomes deposited in biological databases [11, 12]. This ignited the development of several computational tools that can cover all steps of data analysis in a genome sequencing pipeline. A typical pipeline for this purpose includes processing reads (error correction and quality filters), sequence assembly and gap filling. Accuracy of these tools are very important to achieve more complete genomes, which will influence the following analysis [13]. The output of the high-throughput sequencing technologies is a set of short sequence reads that need to be assembled to produce a most probable reconstruction of genome through a *de novo* genome assembly process. Sequencing errors complicate assembly, so quality of reads might be improved avoiding incorrect assembly probability. Later, reads are grouped into contigs and contigs into scaffolds by overlapping regions in a graph data structure. The length of contigs are limited by repetitive sequences, polymorphisms and missing data [11, 14].

These technologies bring novel scientific branches, such as, exomics, metagenomics, epigenomics, and transcriptomics. Other, like comparative genomics had a fast progress with an increase of the availability of genomic sequences [12] because it enables to compare multiple genomes instead of focusing in just one genome [15].

Comparative genomics is a field of biological research in which the genome sequences of same or different species are compared [16]. On the one hand, it studies the process of organism evolution, like tracing their origin, understanding how evolutionary forces govern their changes on genome features, and what makes the related species

unique. On the other hand, it determines gene function based in the principle that DNA sequences are conserved among species [17, 18].

Inspite of the availability of the genomes, the following further analysis and biological explanation slowdown in understanding genomes because a single DNA sequence provides very few information. Genes do not function as isolated units, they need interaction with noncoding DNA and neighboring genes. To understand the gene function and the wider processes of genetics and inheritance, we must look to similar genes in other organisms to determine how function and position has changed over the course of evolution [12, 19].

2.1. Pan-genome Analysis

A single strain genome does not reveal much by itself, it needs to be studied in comparison with other species in the phylogenetic context of the evolutionary process [19]. With database expansions and improvements of high-throughput genomic sequencing, scientific community require new advanced software to post-sequencing functional analysis. Whole-genome comparisons are the next step to determining coding sequences (CDS), discovering regulatory signals, deducing mechanisms and genome evolution [20]. Comparative genomics among multiple genomes had revealed that there is an extensible genomic intra-species diversity [21], identifying all sequence differences among them and which are responsible for phenotypic shift [22]. To answer the question how many genomes are necessary to fully describe the entire gene repertoire of a bacterial specie, it was noted that for each new strain sequenced, new genes will be found in each genomic sequence. Tettelin *et al* coined the concept of the pan-genome to describe a specie [23].

2.1.1. Concepts and Definitions

Pan-genome is the entire gene content present in a group of genomes from the same genus and/or species. The intersections between genomes represent pan-genome partitions (Figure 1). Commonly, a pan-genome has three partitions, core, accessory and unique genome. Core genome comprises the set of genes shared by all genomes in study. Dispensable genome is divided in accessory genome, representing the set of genes shared by only some genomes and the unique genome, composed only by strain-specific genes [24]. Then, genes that define a specie are present in core genome which include genes

responsible for all basic housekeeping functions related with survival. While dispensable genome characterizes a strain diversity like strain-specific phenotype. This diversity is generated by horizontal gene transference (HGT) assigning selective advantages, such as, niche adaptation, antibiotic resistance, and the ability to colonize new hosts [23-25]. Some authors still differentiate accessory genome even more in smaller partitions. Soft-core, comprise of genes shared among 95% of the genomes, the cloud, bundling genes present only in a few genome and the shell, including the remaining genes, present in various genomes [26].

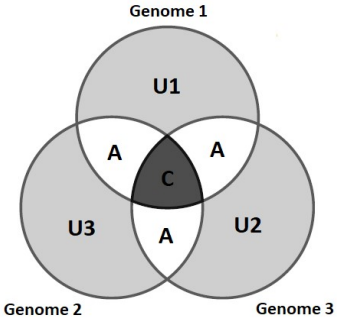


Figure 1: Pan-genome representation of genomes sequences from three organisms. It is represented core (C), accessory (A) and unique (U1, U2, U3) partitions. Partition contains orthology groups that cover a desired number of genomes according to the partition that represents

Therefore, a pan-genome can be 'open' where the size is infinite and grow with the number of new strains genomes added by the contribution of distinct genes. Dispensable genome is larger than core genome with the increase of the number of genomes sequenced. On the other hand, a 'close' pan-genome has their boundaries well defined being a fixed number of genomes enough to fully characterize the pan-genome of the specie. This is noted when the number of distinct genes added converge to zero after the addition of some genomes [24, 26]. Species with an 'open' pan-genome are capable to colonize several niches in different environments because they have the capacity of a quick and economical response to fluctuating environments. This is guaranteed due to a constant opportunity to exchange material genetic by multiple ways with related strains while maintaining stability. Whereas closed pan-genome species have a lifestyle more limited having a low capacity of acquire foreign genes together with fact of living in isolated niches result in a more restrict access to the microbial gene pool [25, 27]. *Buchnera aphidicola* is an extreme example of genome stability living 50 million years without any no chromosome rearrangements, duplications or horizontal gene transfer [29].

Pan-genome study has several applications, such as, study of pathogenicity [30], vaccine development [31], total of all mobile elements [32], all the antibiotic resistance genes [33], taxonomy reclassification [34], among others.

2.1.2. Online Databases

Comparative genomics through pan-genome analysis involve multiple genomes of the desired strain. After this understanding and interpreting genome variation is necessary, retrieving more information from protein sequences on genomes, such as, function, reactions and pathway, through of the access of several databases descripted in Table 1.

Table 1: List of databases with relevant. information for the pan-genome analysis.

Name	Contents	URL
NCBI	Genes, protein functions Taxonomic Data Genome sequences	www.ncbi.nlm.nih.gov/
Uniprot	Proteins Functions	www.uniprot.org
KEGG	Genes, protein function Metabolic data	www.genome.jp/kegg/

The National Center of Biotechnology Information (NCBI) is a repository of biological databases, providing tools for the analysis and retrieval of this information. It englobes GenBank nucleic acid sequence database which provides publicity nucleotide sequences. It receives data from scientific community, DNA Databank of Japan (DDBJ) and European Molecular Biology Laboratory (EMBL). Data exchange ensures worldwide coverage [35]. The Universal Protein Resource (UniProt) is the largest repository of protein sequences and their annotations. This database is the combination of three databases (UniProt Knowledgebase, UniProtReference Clusters, and UniProt Archive). The section that contains manually curated and reviewed entries is UniProt-SwissProt. All others unreviewed and automatically annotation sequences are in UniProtKB/TrEMBL [36]. The Kyoto Encyclopedia of Genes and Genomes (KEGG) is a knowledge-base that have information of individual metabolites or genes and how they are networked in terms of reactions on pathways [37].

2.1.3. Gene Orthology Identification

Comparative genomic studies are based on orthologous identification. They can give us information about evolutionary history, variability of sequences, how and when they gain or loss genes. It is also important to functional annotation in system biology [38].

Table 2 summarizes the definition of the most common types of homology relationships between genes. Homology refers to a relationship between genes sharing common origin being these entities homologous. Orthologous and paralogous are two major types of homology relationship.

Table 2: Definition of types of homologous relationships between genes.

Term	Definition
Homologous	Genes sharing a common origin.
Orthologous	Genes arising by speciation from a single and last common ancestral gene.
Pseudoorthologous	Paralogous genes that appear to be orthologous due to differential, lineage-specific gene loss.
Paralogous	Genes arising by duplication from a single and last common ancestral gene.
Pseudoparalogous	Orthologous genes that appear to be paralogous due to vertical inheritance and HGT.
In-/Out-paralogous	Paralogous genes resulting from a lineage-specific duplication(s) after/before a given speciation event.
Xenologous	Genes arising by HGT from another organism.
Co-orthologous	In-paralogous genes that are collectively, but not individually, orthologous to genes in other lineages (due to their common origin by speciation).
Orthologous group	Group of all descendants of an ancestral gene that diverged from a given speciation event.

When homology is the result of gene duplication, both copies have descended side by side in an organism, they are paralogous. While orthologous are homologous resulting of a speciation event so that the history of gene is the same than the evolution of species [39]. Paralogous usually display less functional similarity than orthologous, because its redundancy enables changes in the sequence causing modifications in the structure. In contrast, an ortholog has greater structural similarity and the same function to the reference [40].

Although speciation and duplication events are considered major factors to gene evolution, there are other events that contribute to that, such as, gene loss, HGT, fusion, fission and other rearrangements of genes. Therefore, orthologous and paralogous are

inside of a complex system of relationships [41]. There are several relationships what make it hard to differentiate orthologous from other types of relationships. Then, this task is always an inference because speciation event that raised them cannot be observed directly. Hence inference should be performed cautiously because in close organisms can occur two inverse situations, of the orthologous genes have different function or genes with the same function not being orthologous. In distantly related organisms this analysis is more tough and susceptible to errors because genomes are in a continuous evolution through several events that complicate knowledge of evolutionary scenarios [38].

A great heterogeneity of so many orthologous prediction tools and databases available without standard formats represent an obstacle to compare or integrate different datasets in order to identify the best set of orthologous of a sequence [42]. OrthoXML format was created to standardise input and output data formats [43].

2.2. Methods for Identification of Orthologous Groups

There is a variety of pan-genome tools established for the analysis and storage of pan-genomes, maintain large-scale datasets and providing automatized pan-genome construction. Some of them were designed for clustering genes from distantly related eukaryotes, not closely related strains or species [44]. Tools available cover different spectra of functionalities. Some of them concern more in reducing memory usage and runtime to maintain large datasets, others offer functional analysis modules allowing to gain biological knowledge of the dataset.

2.2.1. General Strategy

How this methodology is based on the gene content, regions of genomic DNA that encode genes must be identified. Specially in newly sequenced strains or the ones lacking genome annotation, a preprocessing step must be performed to predict genes. It is also important that all genomic sequences used in pan-genome analysis had been annotated with the same annotation method not to influence downstream analysis with bias of individual annotation.

Overall concept of the pan-genome analysis tools is initially very centered on orthologous in which each protein sequence between organisms are compared to finding orthologous [15]. Basic Local Alignment Search Tool (BLAST) is the most used sequence similarity search algorithm. Similarity based algorithms are relatively slow,

having approximately a quadratic complexity on the number of genomes added. Some tools perform a pre-clustering step to cut down time consuming. Alignment relative score of two sequences needs to be filtered to exclude insignificant scores. Minimum score, maximum e-value, and/or minimum length are used as thresholds. Other tools follow an approach based on presence functional domains. Markov Clustering (MCL) algorithm is the most used for orthologous clustering across multiple genomes. Markov matrix is constructed representing transition probabilities generated from weighted sequence similarity scores between two proteins for which similarity was detected. At the beginning, it considers all relationships represent in the graph globally and simultaneously using a flow simulation on the edges by random walks through graph. High flow value indicates that many random walks pass through this area and iteratively algorithm promotes flow where it is strong and removes flow where it is weak, ending when the matrix remains unchanged. At the end, it returns clusters of orthologous groups (COG) containing sequences at least two species [45]. Pan-genome matrix structure is built representing the orthologous genes across genomes under study. From that, several downstream analyses can be done, and each partition can be analyzed and extracted. These analyses include multiple sequence alignment of the core genes, construction of phylogenetic trees, estimation of the pan-genome size, statistics and plots.

2.2.2. Gene-based Computational Tools

Pan-genome tools through research of orthologous genes in the predicted CDS between the set of genomes build the pan-genome. These tools enumerated on Table 3 use graph-based methods for orthology attributing while some tools introduce their own orthology criteria and others use established orthology prediction algorithms [44]. For the set of protein sequences is performed an all pairwise sequence similarities. BLAST is most often used to align the sequences [46]. Nodes are the protein sequences and weighted edges bound nodes representing values from all pairwise sequence similarities. The chosen sequence comparison method and the scoring scheme have influence on the sensitivity and specificity of the estimation. Orthology criteria such as similarity cut-offs or overlap criteria highly impact the orthology assignment and consequently pan-genome structure [44].

Table 3: Description of tools used to do a pan-genome analysis.

Tool	Description	Reference
OrthoMCL	All-against-all BLASTP comparison and Bidirectional best hits (BBHs) are performed to found orthologous and paralogous. COG is performed with MCL algorithm	[47]
Efficient Database framework for comparative Genome Analyses using BBH score Ratios (**EDGAR**)	All-against-all BLASTP to find BBHs. A sliding window is used to find a cutoff that discriminates different thresholds based on normalized BLAST Score Ratio (BSR).	[42, 43]
Prokaryotic-genome Analysis Tool (**PGAT**)	Identify gene families using BLASTP. Choose one gene reference for each family and they are aligned against all open reading frames (ORFs) in their six-frame translation for each genome sequence.	[50]
Pan-genome Analysis Pipeline (**PGAP**)	An all-against-all BLASTP comparison and group the orthologs into multi-strain clusters. Protein sequences of each strain are mixed with the genes of the given strain. BLASTALL is performed and then cluster with MCL algorithm.	[51]
Pan-genome Ortholog Clustering Tool (**PanOCT**)	Combine information from conserved gene neighborhood (CGN) together with BSR values. Compute CGN scores as a metric to retrieve orthologous clusters	[52]
GET_HOMOLOGUES	BBH using BLAST to cluster orthologous. Clustering stringency can be adjusted by scanning protein domains.	[26]
PAN-genome analysis based on FUNctional PROfiles (**PanFunPro**)	Protein families are defined based on functional domain content. It is combined with the Hidden Markov Model (HMM)-based families to generate a protein profile for each genome	[53]
Integrated Toolkit for Exploration of microbial Pan-genome (**ITEP**)	All-against-all comparison using BLASTP and BLASTN. The results are clustered using MCL algorithm	[54]
Pan-Genome Profile (**PanGP**)	Pan-genome profile analysis performed by mathematical models using sampling algorithms	[55]
Large Scale BSR (**LS-BSR**)	Each gene sequence is aligned against all input genomes and itself in order to retrieve the BSR value.	[56]
Roary	CD-HIT is used to pre-clustering similar sequences and remove core sequences from dataset. All-against-all comparison with BLASTP is performed and then clustered with MCL. These results are combined with the pre-computed CD-HIT clusters	[57]
micropan	Includes two approaches (1) all-against-all BLAST and then uses a hierarchical clustering approach. (2) searching for protein domains in the genes using HMMER3 against the Pfam database	[58]

Pan- and Core-Genomic profiling (**PanCoreGen**)	All-against-all BLASTN between reference genome and other genomes. Iteratively, reference genome is replaced by another genome.	[59]
FindMyFriends	CD-HIT grouping and then an own algorithm splits groups based on three factors	[60]
Bacterial Pan Genome Analysis (**BPGA**)	COG are performed with one chosen tool (OrthoMCL, USEARCH, CD-HIT)	[61]
Proteinortho	High-ranked bidirectional best hits (BBH) are combined using MCL and MultiParanoid to retrieve the orthologous groups	[62]

OrthoMCL [47] is a genome-scale algorithm for grouping orthologous protein sequences. Starting from desired genomes, it is done an all-against-all BLASTP, on WU-BLAST tool. By reciprocal best similarity pairs across any two genomes, potential orthologous pairs are founded and reciprocal best similarity pairs within each genome as potential in-paralog/recent paralog pairs. Either for putative orthologous or paralogous a parameter of p-value cut-off must be defined. Similarities between proteins are calculated as normalized BLAST p-values in order to correct for differences in evolutionary distance. Based on similarity matrix, a graph is built and then MCL algorithm [45] is applied to split ortholog clusters [47]. Although this method does not provide downstream analysis, it is a well-established method which improves the accuracy of the ortholog group assignments. Their integration in KBase [63] provides added value.

EDGAR [42,43] contains information about pan-genome analysis in a precomputed database containing 8079 genomes from 322 genera where the user can only choose available pre-processed strains for analysis. It uses a fast, straightforward and simple approach called BBHs [64] to infer probable orthologous based on sequence similarity. BBHs are more likely to be formed by orthologs obtained in a pairwise sequence similarity search through blastp that relies on the assumption that orthologous are more similar to each other than they are to any other sequence from the compared genome. For orthology estimation, it estimates thresholds based on BSR [65] of a gene *a* against a gene *b* is defined as BSR(a, b) = $S_{a,b}/S_{a,a}$, where $S_{a,b}$ refers to the BLAST alignment bit score, in which *a* is the query and *b* the reference/database sequence. All hits are normalized regarding to maximum score. To discriminate between BBH, a heuristic approach through a sliding window was used to find a cut-off value. EDGAR web interface has a limitation which provides only access to data stored in their databases. It is also possible to create private projects with own data, but a request is required.

However, it supports several downstream analysis options. Multiple sequence alignment and phylogenetics trees can be performed for the core genes. Venn diagrams of the gene content can be depicted for up to five genomes and the content of orthologous genes can be analyzed using comparative view or creating synteny plots.

PGAT [50] is a web-based database application for comparing sequences across multiple microbial genomes. Gene sequences from the input genomes are aligned using BLASTP to identify gene families. For each gene family, a gene is chosen as reference. All gene references are aligned against all ORFs in their six-frame translation of each genome sequence. Orthologous genes are taken from the sequence alignments if they have a sequence identity percentage greater than 91-92% and include more than 80% of the gene length. Novel genes predicted using Glimmer [66] are added to the set of reference genes. This method is only applied to highly similar genome sequences where the arbitrary choice of the reference gene has impact on the results. PGAT is entirely web-based, not having a stand-alone version, which restricts the analysis to only nine species in their database. Another disadvantage is the fact that it is not possible to upload genome data for analysis. Beyond building pan-genome, it has an option to identify genes present or absent in a chosen subset of genomes, being the list of core gene only available when selecting all genomes. It is also possible to identify single nucleotide polymorphisms (SNPs) for a desired gene by multiple sequence alignment of orthologs using Muscle [67]. The presence of genes in different metabolic pathways can be studied due to this tool is linked to KEGG [37]. A synteny map allows the visualization of islands organization comparing the order of genes among different strains within the genomic neighborhood of a selected gene.

PAGP [51] is a pan-genome-analysis pipeline composed by Perl scripts for Linux platform. It is possible to input data to identify the pan-genome. InParanoid [68] method finds orthologs between each pair of strains performing an all-against-all BLAST comparison among all protein sequences. It uses pairwise alignment score cut-off and overlap criteria to group orthologs. In the next step MultiParanoid [69] method, will group the orthologous into clusters covering multiple strains. The Gene Family method mixes protein sequences for each strain with the genes of the given strain. A BLASTALL search is performed among the previous group and the results are clustered with MCL algorithm [45]. With the pan-genome information is possible to plot the pan-genome size against the number of genomes, detect the genetic variations (indels, synonymous and non-synonymous mutations) in COG and compute the ratio of divergence at non-synonymous

and synonymous sites. Phylogenetic tree also could be constructed based on gene profiles for each strain and in the detected variations in the core gene clusters. This tool stands out by permitting to perform a functional enrichment analysis, annotating each gene cluster with the most frequent function description and COG classification of its members. The usage of this pipeline might be a little more difficult to users who have not command line knowledge[51].

PanOCT [52] is the first tool using CGN information directly in addition to homology to generate clusters containing single orthologous genes from each of multiple genomes. Requirement of an input file containing all-against-all BLASTP comparison of protein sequences to compute BSR is a disadvantage. After this step that computes BSR, potential frame-shift genes are detected. For each potential orthologous pair, CGN score is calculated, taking in account homology scores of *n* genes upstream and downstream of the target and query protein. Finally, complete linkage clustering is performed using CGN scores as a metric to retrieve orthologous clusters. This algorithm merges sequence similarity with CGN complicating the choices. PanOCT only covers the orthologous gene assignment which could not identify the pan-genome and perform downstream analysis. It is a single application written in Perl having command line interface which could be a little tricky for an unexperienced user.

GET_HOMOLOGOUES [26] is a command line tool available for Linux and Mac OS X systems. It is scalable being able to analyze hundreds of genomes without the computational time increases exponentially. GenBank or FASTA are the input files. This tool offers three sequence clustering algorithms based on the BBH using BLAST: OrthoMCL [47], COGtriangles [70] and another version of the BBH algorithm. Sequences are scanned against the Pfam database [71] with the goal of filtering out clusters containing sequences with different domain architectures. Together with this tool some scripts are provided to do downstream analysis. It is possible to compute the intersection between three cluster groups, an estimation of size of the core and compute pan-genome. Pan-genome matrices can be generated and used to create phylogenetic trees. It is also possible to identify sets of gene for pan-genome analysis. To avoid an underestimation of the core size in a dataset containing draft genome with some missing or fragmented genes, the authors defined the soft-core, composed of genes shared among 95% of the genomes. Although it has a command line interface, it provides an installation script to simplify the installation and a detailed manual with tutorial how to use the analysis scripts makes easier for users without strong bioinformatics skills.

PanFunPRO [53] is a pan-genome analysis tool available as a web server or can be used as a stand-alone tool. Each genome is represented by a FASTA file where only predicted protein sequences of the input genomes are used. If there is not annotations available, Prodigal [72] a tool to gene prediction is applied to the genome sequence to identify potential ORFs. In the next step, potential CDS are scanned against three protein databases, Pfam [71], THFAM [73] and Superfamily [74], in order to retrieve the functional domain content and defined protein families. All sequences having no match are collected and clustered using CD-HIT [75] with a minimal sequence identity of 60%. The resulting clusters are considered as protein families and are combined with the HMM-based families to generate a protein profile for each genome. Pan-genome is computed based on protein families being composed of all unique functional profiles observed. A pan-matrix is generated and visualized where each entry contains of the ratio of non-shared sequences. Gene Ontology (GO) information can also be extracted.

ITEP [54] is a collection of scripts that runs on Linux. It takes three inputs, GenBank files of genomes, a user-defined groupings of input organisms to identify protein families and clustering parameters. ITEP interfaces with a database that it is generated performing an all-against-all comparison between the sequences using BLASTP and BLASTN. The results are clustering using the MCL algorithm, for which maximal minbit or maxbit can be used as metrics. The alignment bit score is either normalized by the minimum or maximum of the self bit scores. A user can also import results from any other orthologous family prediction methods, allowing flexibility. ITEP provides interfaces to different analysis steps, such as, generation of multiple alignments, phylogenetic trees or appending gene neighborhood information. Other functions less common in these tools are provided such as, annotation curation, identification of gene presence or absence within a phylogeny or a generation of draft metabolic networks. This tool has a disadvantage that it requires the use of a different script for each analysis. There are tutorials available and each script has a manual, facilitating it is usage, however, basic programming skills are useful. It relies on multiple dependencies and workflows and it has not a good performance, scaling quadratic with additional genomes.

PanGP [55] is a tool for pan-genome profile analysis. It has dependency to other software to generate COG, binary matrix or pan-genome data that could be passed as the input data, being this the main disadvantage. It has integrated two sampling algorithms, totally random and distance guide, so that it could calculate the pan-genome profile of hundreds of strains at low time-cost. The two algorithms sample *s* combinations

consisting of *n* strains, in order to avoid computing all strain combinations. Genome diversity is characterized by three proposed mathematical models. In the model A genome diversity is characterized by the discrepancy of evolutionary distance on phylogenetic tree. Model B characterizes this diversity by the difference of gene number for each strain and model C calculates it by the discrepancy among gene clusters. It has available few analysis, such as, pan- and core-genome profile and new gene profile. PanGP provides a graphical user interface.

LS-BSR [56] method to Linux/ Mac OSX platforms allows the comparison of the gene content between a large set of bacterial genomes. This method can use a defined set of genes or predict CDS using Prodigal [72]. In order to retrieve the BSR value, each gene sequence is aligned against all input genome and itself with BLAT, BLASTN or TBLASTN. It is generated a matrix as output that contains for each gene the BSR value in each genome. Through a script, it is possible to identify gene sets filtering gene sequences based on a user-defined threshold. It can rapidly compare the gene content of a large number of bacterial genomes with a modest runtime when compared with GET_HOMOLOGUES [26], PGAP [51] and ITEP[54]. Downstream analysis steps are not possible. Matrix can be used as an input for other tools. There is an option to create a compatible matrix with PanGP to perform statistics relating to gene distribution and conservation at different genome depths.

Roary [57] is a tool to build a pan-genome quickly with thousands of prokaryote samples without compromising the quality of the results. The input is a GFF3 format with one annotated assembly per sample. Targeting inefficient runtime of all-against-all comparisons with BLAST, dataset is reduced by pre-clustering similar sequences with CD-HIT [75] and filtered to remove partial sequences. Core sequences are removed from the dataset. On the reduced dataset, all-against-all protein comparison with BLASTP is performed and they are clustered with MCL algorithm [45]. These results are combined with the pre-computed clusters. In a secondary step, using CGN information true orthologous are separated from paralogous which increases accuracy from using it. Output matrix can be used as input for other visualization tools. It provides options to identify the core and the pan-genome computing gene intersections and unions, relative order of the clusters by creating a graph or clustering the input genomes based on gene presence/absence. This method shows an increased accuracy and a high-speed scaling consistently when more samples are added which makes it possible to use with large databases on desktop computer. Although Roary is a UNIX based tool, it is incorporated

in the Galaxy [76] platform being more easier of usage [57]. There is not a parameter on Roary to define a certain minimum length of identity span. It is an issue when specie genomes analyzed have a lot of pseudogenes being themselves joined together in same ortholog group of the full-length genes.

Another R package available from Comprehensive R Archive Network (CRAN) [77] for microbial pan-genomes analyzes is **micropan** [58]. Protein sequence comparisons could be made by BLAST or HMMER3. The FASTA-file input for each genome containing all protein sequences is required. Download of genomes are also possible being necessary to call the gene finder Prodigal [72]. The first method uses BLAST to do all-against-all comparison between the set of protein sequences for each genome. Gene families are found by a hierarchical clustering where linkage functions and cutoff thresholds can be specified. The second method uses HMMER3 to search protein domain in the genes against Pfam database [71]. Sequences are then clustered by same protein domain content in the same non-overlapping order. This last method facilitates analysis for larger datasets in terms of runtime. A pan-genome matrix is constructed saving the number of copies of gene cluster found in each genome. This matrix could be used to plot the number of clusters found in each genome, pan-genome size estimation, principal component analysis or the construction of a phylogenetic tree. R requires only very modest programming skills and inside it there are so many packages available, statistical analysis and graphical displays of all kinds can be made.

PanCoreGen [59] is a standalone software with graphic user interface that generates a pan-genomic profile of genes. As input files, it requires GenBank file for annotated genome sequence, that can be input manually or downloaded from GenBank database, and FASTA file for draft genomes. It also allows the downloading of genomes directly from NCBI. This tool also implements a BLAST approach with the difference that each iteration has another reference genome. The annotated genes from the reference are compared with other genomes using BLASTN. In next iteration, another genome is considered as reference and only genes of this reference that are not identified as orthologous in previous iterations are used as queries. In each iteration, unique genes belonging to the current reference are identified, therefore, core and accessory sets are expanded by new hits relative to the new reference. If genes are duplicated in one genome the number of orthologous founded might be overestimated. Comparing annotated and unannotated genes enable the detection of new genes that remain not unannotated. As output, it provides the list of all gene groups and other with nucleotide sequences of all

genes constituting the pan-genome. For each gene, it also provides the sequences of all matching orthologous. PanCoreGen is only available for Windows.

FindMyFriends [60] is a tool available on R released being part of the Bioconductor [78]. It creates and analyzes pan-genome with thousands of genomes being fast and accurate. Their scope is broader than micropan [58]. The followed approach is not based on BLAST avoiding the computational time. At first step, protein sequences are grouped with CD-HIT [75] and then groups are splitted based on similarity of the neighboring genes, the similarity of the sequences and if genes are from the same genome. Each gene group is again splitted extracting clips from a graph representation. In the end, groups that failed to be part of the same pre-clustering group are merged. It allows to define a certain minimum length of identity span. This tool is integrated within Bioconductor that has a vast of genomic tools and many packages available, statistical analysis and graphical display in R.

BPGA [61] is a command line tool available on Windows and Linux for comprehensive pan-genome studies and downstream analysis. It accepts as input GenBank file, protein sequence file and binary matrix. To perform COG, user can chose between three clustering tools USEARCH [79] defined as default, CD-HIT [75] and OrthoMCL[47]. Once orthologous are detected, BPGA offers several analysis modules. Pan-genome curve, gene family distribution plot and statistics and new gene distribution plot can be made. The second module identifies core, accessory and unique protein families and extracts one representative sequence for each ortholog family. It can also identify orthologous families that contain genes from all genomes expect one specific genome or only containing unique genes. It has a unique feature for atypical GC content analysis, extracting gene sequences which show a deviating GC content when compared with average genome GC content. Pan-genome functional analysis compare the representative protein sequences against COG [80] and KEGG [37] database using BLASTP. Species phylogenetic tree is constructed based on concatenated core gene alignments and binary matrix or user-selected housekeeping genes. It has an option to select a subset of genomes enabling to identify genes exclusively present or absent.

Proteinortho [62] is a stand-alone tool conceived to large datasets reducing the required time and memory for ortholgoy analysis. Pairwise comparisons are done in all complete gene collection using blast. The result is ranked by similarity, such as, blast statistics, evolution distances or genome rearrangements. High ranking bidirectional best hits need to be combined in order to identify orthologous groups. MCL algorithm and

MultiParanoid are used to determine cluster of orthologous groups (COG). Bit score derived from the blast alignent serves as edge weight while e-value cut-off is used on the alignments not included on the graph. Certain minimum length of identity span is possible to define as a parameter. Althought it does not provide downstream analysis, it is straightforward and well-established method to work with large datasets.

2.3. Bioinformatic Tools

Workbench platforms can be an asset for users, integrating several tools in different areas enabling to do a complete workflow in the same platform avoiding difficulties that might come from standalone applications.

The Department of Energy Systems Biology Knowledgebase (KBase) [63] is a web browser platform for software and data designed to integrate them in a unified graphical interface. This ends the need for integration of heterogeneous, distributed, and error-prone primary and derived data into a variety of computational tools to analyze complex and heterogeneous data sets. Therefore, for the creation and running of a workflow, the user does not need to access and learn multiple tools and convert multiples file formats. KBase's reference database includes all public genome sequences from RefSeq [81], Phytozome [82] and selected plants genomes from Ensembl [83]. These genomes are maintained with their original gene IDs and annotation. Their annotation pipeline updated gene calls and annotations. It also contains biochemistry data, biochemical compounds, their function, reactions, growth media formulations, ontologies and metabolic pathways. All this information is always updated from source databases and beyond public data, it is possible upload private data. KBase provides data integration and search, along with easy access to shared user analyses of public microbial reference data from external resources like NCBI and the DOE Joint Genome Institute. Data objects supported is extensible including, reads, contigs, genomes, metabolic models, growth media, RNA-seq, expression, growth phenotype data, and flux balance analysis solutions. Currently, there is a diversity of released applications for genome assembly, genome annotation, sequence homology analysis, tree building, comparative genomics, metabolic modeling, community modeling, gap-filing, RNA-seq processing, and expression analysis. Any user could develop applications to add in the KBase. As the platform grows and adopted widely it is expectable that the data, analysis tools, and computational

experiments contributed will increase, bring wider biological applications with richer and more sophisticated support for functional prediction and comparison.

With a unified platform, it is easy create, execute, collaborate on, and share sophisticated reproducible analysis of their own biological data, in the form of narratives, to propagate new results and compare similar approaches for quality control. Data and tools are provided in a Narrative, a simple graphical user interface of KBase, built on Jupyter Notebook [84]. It enables scientists easily work together within the same platform that it is user-friendly, dynamic and an interactive document that includes all the data, analysis steps, parameters, visualizations, scripts, commentaries, results and conclusions of an experiment. Tutorials are available in the form of public narratives. With this implementation, users can create and run scripts within a narrative using a 'code cell'. All narratives are maintained private by default unless the users choose to share them publicly or selected collaborators.

The integration of OrthoMCL to KBase enables to do several analyses which was not possible as a stand-alone tool. All proteins present in the pan-genome object could be compared, grouping proteins into protein families putatively sharing function and sequence-similar protein families. In many cases, it supports the identification and correction of annotation errors, new and interesting biology, assessing the extent of conservation among genomes and how genomes have evolved and adapted to their distinctive environments. Pan-genome circle plot to view the overlapping membership of genes against a reference genome and phylogenetic pan-genome accumulation could be done.

Galaxy platform [76] is a web-based genomic workbench for scientific workflows, data integration, data analysis and publishing platform that aims to make computational biology accessible to research scientists that do not have programming skills. It was created to finish the problems of accessibility of computational tools, reproducibility and combination of multiple tools together in an analysis workflow.

Reproducibility is essential for the understanding and extending results toward new discoveries. NGS technologies made it difficult with a lack of standards, large dataset sizes and complex computational tools. Although Bioconductor [78] and Bioperl [85] improve the accessibility of computation, scientists without programming skills cannot include a computational tool in an analysis workflow. Galaxy pretends transparency, accessibility and reproducibility of scientific research showing the precise computational

details of the analysis, context and narrative. It provides a wide availability for analysis tools, genomic data, tutorials, workspaces and publication services.

Regarding accessibility, Galaxy has a unified and simple web-based interface for obtaining genomic data and for creating complex workflows. User could import data from established database or upload his own data. It does not require any programming skills or tools implementation knowledge. Any developer may add a new tool in this platform. It is essential to track provenance to ensure reproducibility when user performs any analysis. This is automatically done by capturing and storing descriptive information about datasets and tools in each step. In a history, all datasets are viewable, and the user can rerun and copy any analysis steps. User descriptions or notes explaining the intent of the analysis and why a step is needed or important are also saved in history. All galaxy items have a tag to help users find them easily. Moreover, in order to facilitate transparency, galaxy have a sharing model for Galaxy items (datasets, histories and workflows) and public repositories of published items, a web-based framework for reporting shared or published Galaxy items and Pages, a custom web-based documents to communicate their experiment at every level of detail where the community can reproduce and extend their workflows.

Several applications of NGS technologies are available such as, quality check (QC) and manipulation, mapping reads, genome assembly, RNA analysis, peak calling, variant analysis, annotation, restriction-site associated DNA sequencing (RAD-Seq), variant calling, de novo reconstruction of transcriptomes from RNA-seq data, analysis of high-throughput sequencing data (ChIP-seq, RNA-seq or MNase-seq), correct handle paired-end data, methylation analysis, metagenomic analysis, proteomics, phylogenetic and sequence analysis.

Roarys' integration on Galaxy enables its use in an easier way for people without programming skills. Firstly, it is necessary to upload a set of bacterial genomes to perform pan-genome analysis. After that summary statistics, core gene alignment and gene presence/absence are shown. It is also possible to infer phylogeny using core genes. Although there is a visualization framework on Galaxy, it does not serve the purpose to generate plots using the output data. Therefore, user must download the created files for posterior visualization in another tool.

3. *Streptococcus thermophilus*

3.1. Characterization

S. thermophilus strains are chemoorganotrophic, nonsporulating, catalase negative, devoid of cytochromes, facultative aerobic and acid tolerant. They perform lactic fermentation being the milk its natural environment. Cells are spherical to ovoid with a diameter in 0.7-0.9 μm and grow in pairs to long chains. These strains are moderately thermophilic and require a nutrient-rich environment to grow [86].

Streptococcus belong to the kingdom *Bacteria*, subgroup *Firmicutes*, class *Bacilli*, order *Lactobacillales*, family *Streptococcaceae* and genus *Streptococcus* [81]. The genus comprises a wide variety of pathogenic species and commensal gram-positive bacteria. They have great capacity to adapt getting to inhabit a wide range of hosts (e.g. humans, horses, pigs and cows) and for each host to colonize diverse habitats. The common diseases caused by *Streptococcus* species are pneumonia (*Streptococcus pneumoniae*), pharyngitis and acute rheumatic fever (*Streptococcus pyogenes*) meningitis and septicemia in neonates (*Streptococcus agalactiae*) and tooth decay (*Streptococcus mutans*) [88]. Other species like *Streptococcus oralis*, are opportunistic pathogens living normally on human organism without cause any disease in healthy individual while other like *Streptococcus thermophilus* are harmless saprophytes with major economic importance [89].

Strains of *S. thermophilus* are one of the most economically important of the lactic acid bacteria (LAB). Inside *Streptococcus* genus, it is the unique specie considered as safe to be used in food industry. It is a starter culture widely used in the manufacture of dairy products [5]. Normally, it is conjugated with *Lactobacillus* species, especially *Lactobacillus delbrueckii* subsp. *bulgaricus* (*Lb. bulgaricus*) or *Lb. helveticus* for manufacturing products like yogurt or cheese, but it may also be used alone for yogurt production [2]. It has a great impact on the qualities of the fermented food products, e.g., flavor and texture, that can be linked directly or indirectly to the gene content of the strains used in fermentation [5]. It also has a probiotic effect in the alleviation symptoms of lactose intolerance and other gastrointestinal disorders, maintaining remission of ulcerative colitis, prevention of the postoperative recurrence of Crohn's disease. L-lactic acid is the main product resulting from the fermentation of subtracts such as fructose,

glucose, lactose and sucrose. Vitamin folate is also produced in high levels with an importance for iron metabolism and maintaining cardiovascular function and fetal development [86].

In an evolutionary context, it has to cope with changing environments, with biotic and abiotic constraints in milk, yogurt and in digestive tract following ingestion, all of which involved in shaping its genome. *S. thermophilus* still undergoes a process of reductive evolution towards an adapted bacterium to a relatively new environment, the milk niche. Among other things, this has occurred through acquisition of foreign genetic material by horizontal gene transfer events from other dairy species, so that, it has unique gene features that contributes to its adaptation. In effect, considerable strain variability exists and that has yet to reach an equilibrium [2, 5].

3.2. Genome Features

There are currently 22 fully sequence strains of Streptococcus thermophilus deposited in NCBI database. Makarova *et all* [90] were the first authors to describe *Streptococcus thermophilus* LMD-9 considered as a strain reference of its species. The major genomic features, updated on NCBI, are summarized on Table 4.

Table 4: Genomic features of *Streptococcus thermophilus* LDM-9 strain;

Feature	Value
Genome size	1,86 Mb
G+C content	39.1%
Plasmids (number of genes)	pSTER1 (2) pSTER2 (5)
Genes	1.993
Pseudogenes	230
rRNA	18
tRNAs	67
Proteins	1,674

The circular genome of *S. thermophilus* is 1,8 Mb, being their size around 60 kb larger than the previous genome sequenced of plasmid-free strains [5]. The molar G + C ratio is 39,1 % containing 1993 genes. There are 1 674 proteins encoded on genome. It harbors 230 pseudogenes which indicates reductive evolution in which occurs active genome degeneration with loss of metabolic pathways and transport systems that are non-essential in dairy niches rich in nutrients [91].It has 18 rRNA and 68 tRNA. This strain

harbors two small plasmids, pSTER1 and pSTER2. Plasmid pSTER1 is larger than plasmid pSTER2 having a length of 4449 bp and 3361 bp encoding two and five genes, respectively [90]. Plasmid pSTER2 encodes a small heat shock protein (sHsp) that is linked with the increase of the survival in LMD-9 under several stress conditions [5].

3.3. Importance for Chr. Hansen Company

Chr. Hansen is a global bioscience technologic company that develops natural solutions in Food Cultures & Enzymes, Health & Nutrition and Natural Colors areas. The Food Cultures & Enzymes division develops and produces meat and wine cultures, dairy enzymes and dairy cultures including probiotics for the food industry in general and the dairy industry where this area represents 58% of the company work. Natural ingredients determine taste, appearance, nutritional value and health benefits.

The company is the global market leader where one per two cheeses in the world contains its products. Chr. Hansen Culture Collection is one of the world's largest commercial collections of bacteria, numbering almost 30 000 strains. Each strain from this collection are screened, selected and improved to meet specific requirements in food, dietary supplement or feed products.

Genomic characterization of industrially S. thermophilus strains from the Chr. Hansen Culture Collection is really essential for the company to yield a better understanding of strain variability to do rational improvement and selection of commercial dairy strains used for making yogurt [92].

4. Material and Methods

4.1. Assessment of Orthologous Detection Tools

4.1.1. Genomes and Annotation

Streptococcus thermophilus genomes were retrieved from NCBI assembly database [93] on February 2018. Data is composed by a set with 21 genomes with a "Complete Genome" status regarding the assembly level and a refseq status (Table 5). Based on the previous conditions, the files were downloaded through 'Download Assemblies' option on Genomic GenBank format type from a RefSeq source database. The resulting folder have the compressed files into a single tar format file and must be decompressed to access their information.

Refseq files are annotated by the Prokaryotic Genome Annotation Pipeline (PGAP) [94], an exclusive method developed by NCBI to annotate bacterial and archaeal genomes (chromosomes and plasmids). It is not publicly available and it can be used only inside of the NCBI environment in genomes submitted to GenBank. It was the single annotation method used. If you use multiple gene prediction softwares, you are wasting your time looking at noise and errors since they rarely completely agree.

Table 5: *Streptococcus thermophilus* complete genomes in RefSeq format available on NCBI. Each genome has a mask code associated with itself (See on Section 4.2.1).

Strain	Mask Code	RefSeq assembly accession
JIM 8232	AT	GCF_000253395.1
LMG 18311	IU	GCF_000011825.1
CNRZ1066	FO	GCF_000011845.1
LMD-9	HU	GCF_000014485.1
ND03	DH	GCF_000182875.1
MN-ZLW-002	BB	GCF_000262675.1
ASCC 1275	BR	GCF_000698885.1
SMQ-301	FJ	GCF_000971665.1
MN-BM-A02	DG	GCF_001008015.1
MN-BM-A01	CE	GCF_001280285.1
S9	CQ	GCF_001514435.1
KLDS SM	CG	GCF_001663795.1
CS8	KZ	GCF_001685375.1
KLDS 3.1003	EO	GCF_001705585.1
ND07	EV	GCF_001855705.1
APC151	GJ	GCF_002012365.1
ST3	EG	GCF_002286255.1
B59671	HG	GCF_002443035.1
GABA	LH	GCF_002846075.1
EPS	KI	GCF_002846155.1
ACA-DC2	GG	GCF_900094135.1

4.1.2. Pan-genome Creation Tools

Detection of the orthologous genes across the different genomes was done with four open access tools, FindMyFriends (v. 1.3.3) [95], OrthoMCL (v.0.0.7 in KBase)[47], Roary (v. 3.12.0) [96] and Proteinortho (v. 5.16) [62]. These tools were selected because the last three are among most commonly used and FindMyFriends represents a novel non-blast based approach.

FindMyFriends is an R package available on Bioconductor. Its installation can be done with simple commands to install Bioconductor and FindMyFriends from there (see documentation on Bioconductor). The easier way to use OrthoMCL and Roary are in KBase and Galaxy platforms, respectively, since they do not require installation. However, they are available as stand-alone versions in UNIX-like systems (see respective documentation to install). Roary was used as a stand-alone version. Proteinortho installation is quite simple and can easily be ready for use just uncompressing the tar.gz file downloaded from the website.

4.1.3. Inputs

Input file format differs across different tools and it can contain sequences either represented as nucleotide or as amino acid sequences. Fasta and GFF3 file formats comprising amino acid sequences were obtained from GenBank file format to ensure that there are no information differences across distinct formats used by tools and to do not include identified pseudogenes. Genome input files are taken by tools in different formats (Table 6). OrthoMCL accepts GenBank format, Roary takes the GFF3 format and both FindMyFriends and Roary use fasta format.

Although it is not mandatory, FindMyfriends needs chromosomal position of the gene, such as, beginning and end positions and strand. Be aware, that the lack of this information can compromise results'quality and it can be passed in different ways, like as, dataframe or a function. Fasta file format and gene location file were created from GenBank files through GetFAAandGC.py script (Supplementary Material https://1drv.ms/f/s!ApASY1Jo99MLb0ba4XsI0OMCyBY).

GFF3 file format from NCBI can not be used by Roary since they only contain annotation sequence. Thus, a perl script bp_genbank2gff3.pl [97] were used to embed the nucleotide sequence and CDS amino acid sequences.

Table 6: Summary of input format files across orthologous detection tools used (FindMyFriends, OrthoMCL, Roary and Proteinortho).

Tool	FindMyFriends	OrthoMCL	Roary	Proteinortho
Input	Fasta Gene Coordinates	GenBank	GFF3	Fasta

4.1.4. Parameters and Execution

OrthoMCL only distinguishes "recent" paralogs from "ancient" paralogs. Thus, homology groups contains "recent" paralogs together with orthologs. For this reason, tools were set up to do not to split paralogous.

OrthoMCL has two adjustable parameters with a default setting of $1e^{-5}$ e-value and 1.5 Markov clustering inflation index. Subsequent analysis were done using this default values which show a high sensitivity and specificity in orthology detection performance [98].

Roary was set up with 95 default percentage identity (-i option), in order not to split paralogous (-s option). Output directory was defined (-f option) as well as all input files with GFF file extension on working directory were passed.

```
roary -f ./roary_out –i 95-s *.gff
```

FindMyFriends and Proteinortho are algorithms that utilize a cut-off parameter on the sequence length difference to avoid grouping full-length genes with gene fragments. This parameter is represented as maxLengthDif on FindMyFriends and coverage on Proteinortho. Regarding FindMyFriends, at first a genomes vector was produced with the names of fasta files in the directory where they are. Pan-genome object was created passing genomes vector, a translated logical argument indicating that the file contains amino acid sequences and a genLocation argument with the dataframe containing gene coordinates. Initial grouping was performed with a start similarity threshold of 0.95, a maximum deviation in sequence length of 0.2 (i.e. 0.8 coverage threshold) and a 0.95 start similarity threshold. Lastly, groups were refined splitting the members into new groups by neighborhood synteny considering paralogoues similarity (argument forceParalogues)

```
genomes <- list.files(pattern = '*.faa')
Strepangem <- pangenome(genomes, translated = TRUE, geneLocation = gene_loc)
Strepangem <- cdhitGrouping(Strepangem, maxLengthDif = 0.2, from = 0.95)
```

Strepangem <- neighborhoodSplit(Strepangem, forceParalogues=FALSE)

Proteinortho program was set up with 0.95 amino acid and similarity (must be in percentage). Best blast alignments must have 80 minimum coverage percentage. Genes without orthologoues were also reported (-singles option). Last argument passes all fasta files found in the directory.

./proteinortho5.pl –cov=80 -sim=0.95 -identity=95 -singles genomesFAA/*.faa

4.1.5. Outputs

Pan-genome matrix is a common output across different tools. This matrix represents homology groups on rows containing orthologous identified across different genomes on columns.

Since FindMyFriends creates an R object containing entire information, the package has a set of tools to handle it and access the desired information. The most important functions will be highlighted (to see more about it see Bioconductor documentation). GroupInf function allows to get and set information about each homology group, such as, their name as well as genomes and genes present. Statistics can be calculated to each gene group or organism. OrgStat function calculates organism statistics such as, minimum length of gene, maximum length of gene standard deviation of gene lengths, residue frequency, number of gene groups and number of paralogues. GroupStat returns statistics and positional information about each gene group, such as, maximum number of genes from the same organism (paralogues), shortest and longest sequence length, standard deviation of sequence lengths, index of genes in group, downstream and upstream gene groups. Number of gene groups (nGeneGroups), genes (nGenes) and organisms (nOrganisms) represented in a pan-genome can be queried. Accession to gene location (geneLocation function), gene names (geneNames function) and sequence length (geneWidth function) to each gene is also possible as well as extracting gene sequences (gene function). Their pan-genome matrix has a different formatting: each line has the homology group, one orthologue and their genome. Thus, matrix has to be saved from pan-genome R object (“write.csv(as(Strepangem, 'matrix'), file = "matrixFindMyFriends.csv")”) and converted to a standard matrix through a Get_FMF_Gene_Matrix python script (Supplementary Material https://1drv.ms/f/s!ApASY1Jo99MLb0ba4XsI0OMCyBY).

OrthoMCL output downloaded from KBase contains two matrices. One of them contains the number of distinct protein families identified to each pair of genomes while another matrix is a standard matrix with annotation and identifiers to each homology group.

Proteinortho creates 3 output files. Standard pan-genome matrix *myproject.proteinortho* contains additional information, such as, number of genomes covered by homology group, number of genes present in each homology group and a parameter that refers how well the genes are connected inside graph. It also generates graph information files, myproject.blast-graph and myproject.proteinortho-graph, containing all pairwise orthology relationships including similarity scores for both genome directions.

Roary creates several output files. A summary statistics file has an overview of how genes are distributed across different pan-genome partitions and how frequently they occur in the organism. Gene_presence_absence csv file contains standard matrix with annotation, number of organisms, sequences in the homology group, average number of sequences per organism, information about gene linkage and indication of their order within graph to core and accessory genes, comments on the quality control, minimum, maximum and average length in nucleotides of the homology group. The previous file is also presented as an R tab format type containing only the binary matrix with the presence and absence of each gene in each organism. A pan_genome_reference fasta file contains a representative nucleotide sequence from each homology group in the pan-genome. Graphs in DOT format type of how genes are linked together at contig level in the pan-genome (core_accessory_graph.dot) and accessory genomes (accessory_graph.dot) are also created. A multi-fasta alignment of all core genes can be created with –e parameter. Pan-genome matrices of the different tools are presented on pan-genome_matrix_tools file (Supplementary Material https://1drv.ms/f/s!ApASY1Jo99MLb0ba4XsI0OMCyBY).

4.2. *Streptococcus thermophilus* Pan-genome

This section describes methodology used to do a pan-genome analysis. The workflow from pan-genome creation to a genomic analysis is depicted in Figure 2. In this first part, genome set was annotated with Prokka and filtered to remove genome repetitions. Pan-genome was constructed and annotated against KEGG database. At the end, K number matching was done to group the homology groups.

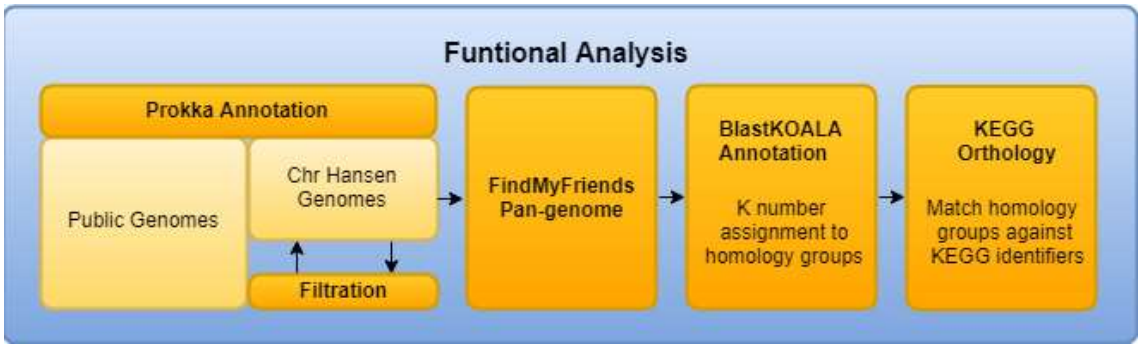


Figure 2: Overview from pan-genome to functional analysis. Public and Chr. Hansen genomes were annotated with the Prokka tool. To remove redundancy, Chr. Hansen genomes were filtered. A pan-genome was created with the FindMyFriends pipeline, where represents all homology groups across the genomes. The annotation of each homology group was improved assigning to each one a k number with the BlastKOALA tool. Finally, the homology groups were matched against the KEGG identifiers and the genes which would be studied were retrieved.

4.2.1. Data

This study has been executed with 541 *Streptococcus thermophilus* genomes. This number consists of 21 public genomes and 520 genomes coming from Chr. Hansen's bacterial private collection. Genomes are alphabetically labelled from A to TU. Between these private genomes, there is a redundancy arising from multi-sequencing of the same strain resulting in more than one file for the same strain with different assembly levels. There are 9 complete genomes named as AB, TQ, TS, GE, BJ, CR, HC, IM and LR. The other sequences are incomplete because of different assembly levels consisting of contigs and scaffolds.

4.2.2. Genome Annotation and Filtration

Genomes were annotated automatically using Prokka software [99], version 1.12, with defaults parameters. Comparisons between all 520 Chr. Hansen genomes were made with the FindMyFriends pan-genome pipeline [100] using fasta files produced by Prokka.

The aim of doing genome filtration is to remove redundancy arising from multi-sequencing which creates genome repetitions. FindMyFriends was carried out using amino acid sequences (translated argument set to true) and the gene coordinates. The parameters settings were 0.95 start similarity, a 0.2 maximum deviation allowed in sequence length and forceParalogues argument set to true in order not to split paralogues considering their similarity zero- A file is created for each gene composed of the gene identifier, location and strand were created from the genbank files using Get_Gene_Coordinates.py script (Supplementary Data https://1drv.ms/f/s!ApASY1Jo99MLb0ba4XsI0OMCyBY). The pan-genome results,

containing all homology groups across the genomes, were saved from FindMyFriends R object and converted through the Get_FMF_Gene_MATRIX.py script (Supplementary Data https://1drv.ms/f/s!ApASY1Jo99MLb0ba4XsI0OMCyBY) to a matrix containing the gene identifiers, where each row represents a homology group and each column the genome identifier. A binary matrix was obtained from the previous one whose cells were filled with 1 if there is a gene in the genome belonging to a homology group and 0 otherwise. Pearson correlation coefficient was calculated for all the genome pairs present on the binary matrix. The correlation matrix was reordered according to the correlation coefficient using hierarchical clustering order with the default complete linkage agglomeration method and the correlation plot was depicted. Clusters containing genomes with a correlation coefficient more than 0.95 were identified. To remove the redundancy between the genomes, the genomes with the highest assembly levels were kept and the others were discarded. After that, low-level sequencing genomes still remain on the resulting set. They were identified by the small core genes content and excluded from it.

4.2.3. Pan-genome Analysis

A *Streptococcus thermophilus* pan-genome was constructed using 21 public and 315 Chr. Hansen's filtered genomes. The methodology was quite similar with the previous one. The input was the protein fasta files of the translated CDS sequences created by Prokka and two files containing the gene coordinates to the two genomes sets. The software has been run with the parameters translated sequences set to true, 0.95 start similarity threshold, 0.2 maximum deviation and forceParalogues set to false. The last parameter to split the paralogous from the real genes is the main difference from the previous created pan-genome. Script Add_Columns.py (Supplementary Data https://1drv.ms/f/s!ApASY1Jo99MLb0ba4XsI0OMCyBY) was used to add two columns with the consensus annotation and amino acid average length to each homology group in the standard matrix obtained from FindMyFriends R object, previously described.

4.2.3.1. 16S rRNA sequence analyses

Genomes that deviated more than 15% from the average gene number, 1800 genes, for its specie were selected. The ribosomal RNA (rRNA) genes were predicted by Barrnap, version 0.7 [101], from the genome assemblies creating a GFF file containing the rRNA locations. The 16S rRNA sequences were extracted from the genome

assemblies using the getfasta command from the BEDTools tool, version 2.18 [102]. These sequences were blasted [46] against the SILVA [103] and NCBI 16S ribosomal RNA sequence (Bacteria and Archaea) databases.

4.2.3.2. Functional Annotation (Annotation Improvement)

In view of work on a curated annotation basis, not relying on prokka annotation labels, pan-genome was annotated against KEGG database. A representative sequence was considered to be the first gene in every homology sequence. Its amino acid sequence was retrieved from the inferred pan-genome to create a fasta file. These protein sequences were assessed against KEGG´s annotation tool, BlastKOALA tool [104], an automatic annotation server to classify genes or proteins sequences. The 1308 taxonomy group, corresponding to *Streptococcus thermophilus*, was set to search against to KEGG GENES database at prokaryote species level. KEGG orthology (KO) assigns KO entry identifiers named K numbers to identify the KEGG orthologs in order to characterize individual gene functions and reconstruct of KEGG pathways (or modules).

4.2.3.3. Identification and Selection of Genes

The aim of this section is to retrieve the homology groups matching each gene in order to assemble them to each metabolic pathway or functional module (in the case of transporters and proteinases). From KEEG BRITE functional hierarchies (BR) or KEEG pathway maps (PATH) identifiers, the K numbers were extracted. Modules (MD) are functional units, such as, structural complexes often forming molecular machineries, transporters, or other types of essential sets on pathways. What underlies this step is K number matching between the KEGG database and homology groups inferred on the pan-genome. Thus we are going to have all the homology groups joined that contain one specific gene. In light of this, we can have assembled pathways and functional modules.

4.2.3.4. BLAST

Blast was done to know why some genes are not present in some genomes and subsequently not represented on the homology group. Representative sequence, of the respective homology group, was blasted as a query against the genome where the gene was not detected. In order to simplify the blast process, a database was created to each query only with the genomes where were not detected a gene.

4.2.4. Motif Characterization

This section describes methodology used to do motif characterization. The process of filtering the pan-genome matrix and creating motifs to scan for new sites is presented in Figure 3.

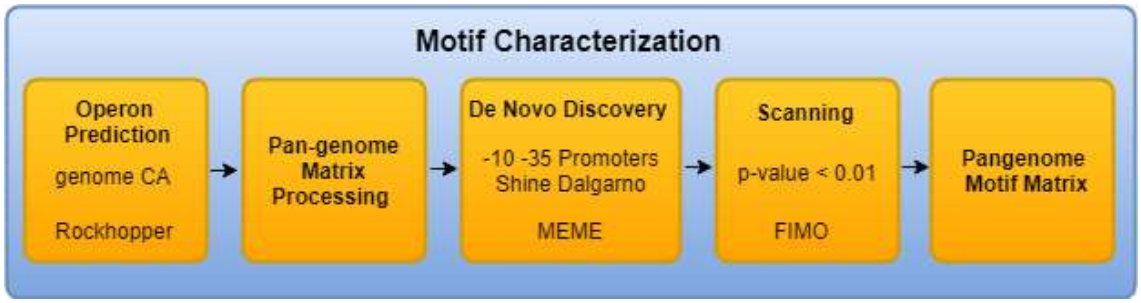


Figure 3: Overview of the process of motif characterization from a 333 genomes pan-genome. The operon structures were predicted with expression data from the genome CA. The pan-genome matrix was processed by filtering out the homology groups that contain the genes present inside operons from CA genome. Promoters, -35 and -10, and SDS motifs were built with MEME tool and then scanned for all occurrences with a p-value < 0.01 in the 200 nucleotides upstream on the genes for the remaining genomes using the FIMO tool.

4.2.4.1. Operon Prediction

Operons are a structure of a group of genes controlled by the same promoters. The putative operon prediction was made with the goal of excluding such genes inside of an operon that do not have immediate upstream promoters. Rockhopper (version 2.0.3) that led to do various stages of bacterial RNA-seq data analysis was used to determine the operon structures [105]. The input was the genome file and Chr. Hansen generated RNA sequencing reads in a FASTQ format from the complete genome CA. These step was performed by supervisor Martin Holm Rau. From the output file *_operons.txt containing information about each multi-gene operon predicted, the first gene was kept and the remaining genes inside the operon structure were identified and filtered out of the 333 genome pan-genome matrix, removing each homology group that comprise the gene.

4.2.4.2. De novo Motif Discovery

A fasta file were created containing the 200 nucleotides upstream for each gene in every genome. In cases in which the gene does not have an enough upstream length were excluded.

Identification of hexanucleotide sequence elements such as the promoters, -35 and -10, and Shine-Dalgarno sequences (SDS) were proceeded using MEME Suite [106]. Shine-Dalgarno (SD) motif of the ribosomal binding site (RBS) were identified creating

a background model with 20 nucleotides upstream of translation initiation sites (TISs) on genome AB. SD motif was discovered by searching only in the given strain one motif with a 6 nucleotides width. A promoter prediction tool called PePPER [107] was used on the public genome *Streptococcus thermophilus LMG 18311* fasta file, labelled as IU. Through the transcriptional start site (TSS) determination, promoter sequences were predicted and extracted with an 28bp length. Separated motifs were built for promoters, which includes a -35 region and -10 pribnow box (analogous to the TATA box), using the previous promoter sequences as the background nucleotide distribution. The two motifs were searched only in the given strand with the 6 nucleotide width option. MEME minimal motif text format containing the background frequencies and motif letter probability matrix were saved to SDS and promoters to use in the following step. This step was manually performed on the web server.

4.2.4.3. Motif Scanning

Once sequence motifs have been built, they might be scan on new genomes to find new sequence sites. MEME minimal motif text format for motifs built and upstream sequence was used to scan new SDS and promoters sites using FIMO tool. SDS and promoters MEME minimal motif text format were input separately. Fasta file containing the collection of upstream regions for each genome was used as an input as well. Scanning options were set to scan only in the given strand (--norc parameter) and output only the matched sites found that have p-values less than 0.01 (--thresh parameter). The --verbosity parameter regulates the level of output information. The process was automated by a script that ran individually the command line, fimo -oc <directory output fodler> --verbosity 1 --thresh 0.01 --norc < directory motif file> < directory sequence file>, to each motif in every 333 genomes. The output file reports information about motifs found, like the sequence, strand, start, end positions and p-value. Afterwards, a post-processing step is crucial to filter out the new sequence sites found for both SDS and promoters and choose the best sequences matches found in each genome. Two conditions were set: (i) right order: promoter -35, -10 and SDS sorted according to distance for the of translation initiation sites (TISs); (ii) spacer length: promoter -35 and -10 are separated by 15-19 nucleotides spacer [108].

5. Results and Discussion

5.1. Assessment of Orthologous Detection Tools

Tools were set up to "do not split" paralogous since orthoMCL does this option by default. Orthology groups defined cover different number of genomes being part of different pan-genome partitions (Table 7). Orthologous grouping is done on different extent by different tools used.

As OrthoMCL and Roary do not have a cut-off length parameter, it is expected them to create fewer orthology groups because gene fragments should be kept together with full-length genes. Considering total number of orthology groups, Roary has the smallest number of total groups with 2582 homology groups. Other tools created a similar number of homology groups: OrthoMCL has 2731, FindMyFriends has 2706 and Proteinortho has 2783. Naturally, Roary has less orthology groups across other different partitions. Roary is the second tool with more orthology groups on core, 1097 orthology groups, and the tool with least number of unique groups, with 403 orthology groups.

Table 7: Overview of orthology groups created by each orthologous detection tool (OrthoMCL, Roary, FindMyFriends and Proteinortho) across different pan-genome partitions covering different genome numbers.

Partition / **Tool**	Core	Accessory	Soft-Core	Shell	Cloud	Unique	**Total**
OrthoMCL	1104	*931*	203	566	162	696	2731
Roary	1097	1082	202	662	218	403	2582
FindMyFriends	1070	1186	209	700	277	450	2706
Proteinortho	1071	1197	192	740	265	515	2783
Genomes	100%	10-95%	90-95%	16-90%	6-15%	5%	
	21	2-20	19-20	4-18	2-3	1	

OrthoMCL was the second tool that creates more number of total groups. It is reflected on the 1104 core groups, the highest core group across tools. On the other hand, it has smallest accessory partition with 931 groups and the highest unique partition with 696. Soft-core is very similar across different tools with about 200 homology groups. As it creates more unique groups, their cloud on accessory genome is the smallest one. It has a soft behavior since it was expected to create fewer orthology groups due to not having coverage parameter. Thus, it is clear that orthoMCL is able to define core groups but that

it splits orthology groups towards unique groups. For this reason, it is evident that OrthoMCL have an intermediary behavior on splitting paralogous. Naturally, it keeps the most recent paralogous with orthologous and splits the ancient paralogous [47]. It makes it difficult to get closer to other tools since it has no intuitive parameters to do not split paralogous at all. On the other hand, Roary is more conservative. It does not split paralogous from orthologous groups and also does not have coverage parameter which leads to variation in relation to OrthoMCL. It could justify their lowest orthology group number. No splitting behavior has as consequence a greater accessory number and fewer unique group number.

FindMyFriends and Proteinortho has a similar performance on the orthology groups creation across pan-genome partitions. As these two tools use a cut-off parameter, it is expected to create a higher number of homology groups mainly on accessory and singleton groups possibly at the expense of core groups. Indeed, these two tools created the higher number of accessory groups and the smaller number of core groups. For this reason, these two tools were the only ones where the number of accessory groups was higher than core groups. FindMyFriends created 2706 homology groups and Proteinortho created more 77 homology groups with a total value of 2783 homology groups. Since they identified about the same number of core groups (1070), this difference is shown mainly on singleton groups created by Proteinortho. Proteinortho created more 11 accessory groups and 65 singleton groups than FindMyFriends. FindMyFriends seems slightly stricter than proteinortho since it creates less orthology groups.

The progress of the different partitions during the addition of genomes to the pan-genome was plot to each tool (Figure 4). Since, this type of plot is very biased towards the order of genomes, a bootstrap genome order was created. The more genomes are added to the pan-genome is natural being created more orthology groups and therefore total number of homology groups (blue line) increases. Core partition (dark red line) decreases because when more genomes are added there are genes no longer shared across all genomes being now part of the accessory (dark green) or even to unique partition (yellow). Tools identify orthology groups and, naturally, split or not group them on different levels. FindMyFriends and Proteinortho has a similar performance on the creation of orthology groups number. As seen on Figure 4 A) and C), both tools have a greater accessory partition than core partition which could show that they split more the orthology groups. However, FindMyFriends is slightly stricter on to split option compared with proteinortho since it creates less number of orthology groups reflecting its

effect on the creation of less singleton groups. As seen on Figure 4 D), Roary is strict on grouping orthologous. It creates less orthology groups than the other tools, and even on the orthology groups created, they cover more genomes, as it is possible to be observed on the core (dark red line) and accessory (dark green) partitions close to each other and the most singleton partition (yellow line). Possibly, it may be explained by the no coverage parameter that sieve pseudogenes from full-length genes. OrthoMCL is the tool that splits more the orthology groups creating more unique groups. As seen on Figure 4 B), accessory partition (dark green) is smaller than core partition (dark red) resulting in increased singleton partition (yellow). Actually, this partition line has a high gradient compared with the other tools that even exceed shell partition as well as soft-core (light green) exceeding to cloud (light red). Previous cases were not observed on other tools. Another fact around the increasing of singleton groups is the exceeding soft-core (light green) to cloud (light red) meaning that the tool progresses from the splitting of the high covering homology groups towards singleton groups.

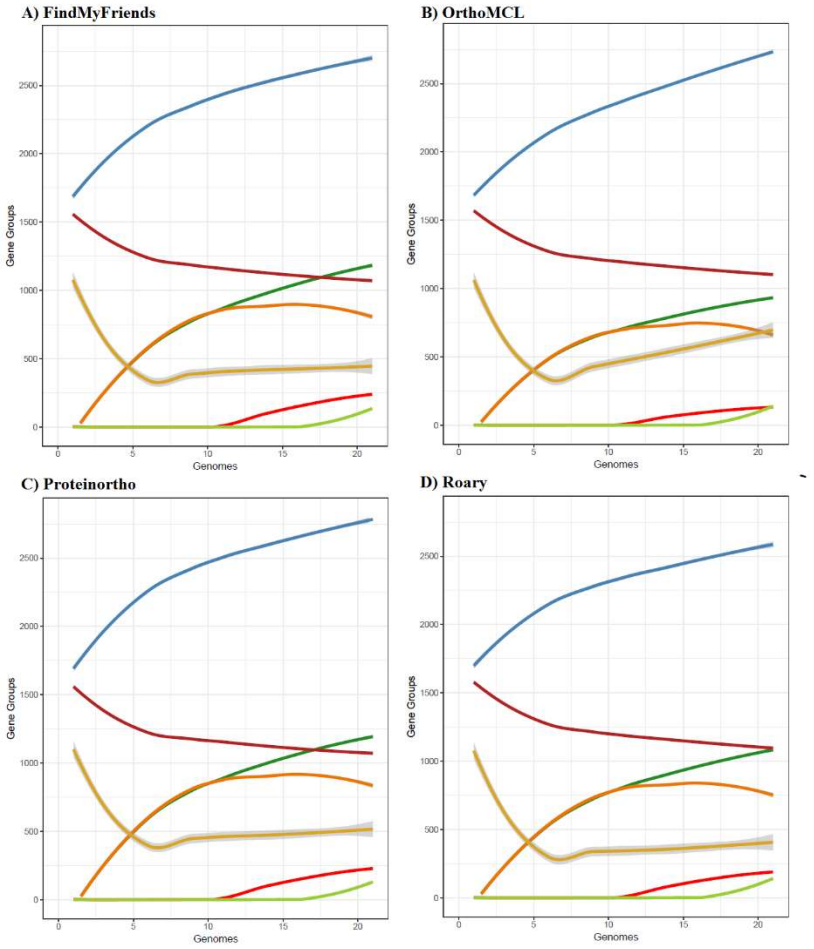


Figure 4: Evolution of pan-genome partitions evolution as more genomes are added. Total gene groups (blue), core (dark red), accessory (dark green), soft- core (light green), shell (orange), cloud (light red) and singleton (yellow).

Functional dendrograms were created for each tool based on their different homology groups assignment (Figure 5). There is an agreement across different tools on a large strain clustering. Strain JIM is connected but not clustered together with the large cluster. Large cluster is split in two clusters: strains MN BM A02, KLDS SM, ASCC 1275, ND07, SMQ 301 and LMD9 are together while the strains ND03, APC151, MN ZLW 002 and MN BM A01 are in another cluster. Short cluster containing strains ST3 and GABA is connected with this large cluster. FindMyFriends and OrthoMCL also had strain KLDS 3 1003 connected with all previous ones but more distanced (Figure 5 A and B). Strains CSS and CNRZ1066 were clustered together being the following strains not clustered together and more far away from them: strains S9, EFS, LMG 18311 and ACA DC2. FindMyFriends and Proteinortho also has strain B59671(Figure 5 A and C). These strains are clustered at same level with the large cluster on FindMyFriends, Proteinortho and Roary. OrthoMCl has these strains clustered together at same level with strain KLDS 3 1003 and large cluster. B59671 is the outest strain on OrthoMCL and Roary (Figure 5 B) and D)). Strain KLDS 3 1003 is the outest clustered strain on Proteinortho (Figure 5 C). Strain KLDS 3 1003 and B59671 have been clustered differently across the tools. It could suggest that these genomes had a variation on orthologous detection. Overall, genome clustering across different tool is very similar.

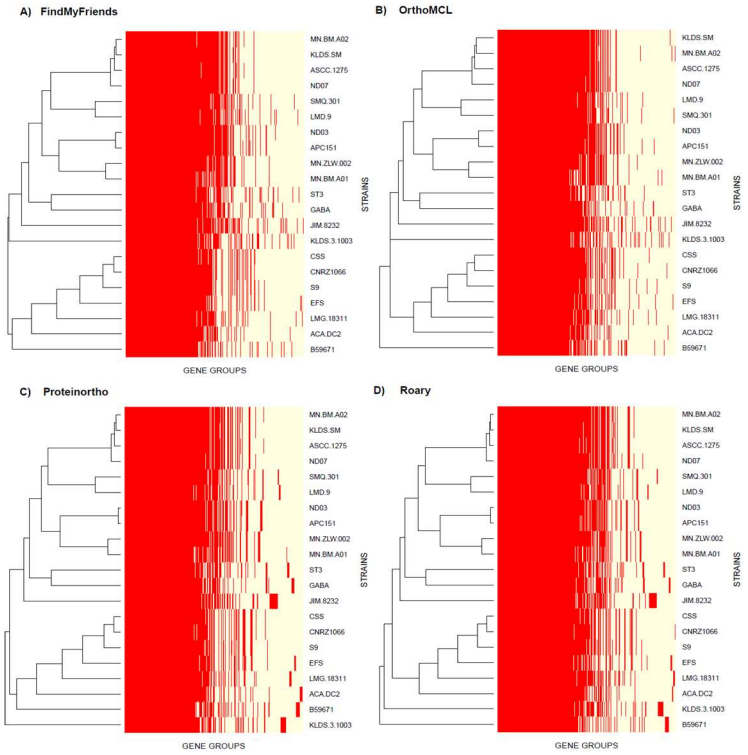


Figure 5: Heatmap pan-genome of 21 *Streptococcus thermophilus* strains. Each coloured block represents homology (identical gene) in the pan-genomes created by FindMyFriends (A), OrthoMCL (B), Proteinortho (C) and Roary (D).

Genome LMD9 was taken as a reference on the pan-genome from different tools in order to study how differently genes were grouped. For each tool, it was taken pan-genome matrix containing only the strain LMD9. Thus, it contains LMD9 gene identifiers, orthology groups identifiers and the number of genomes included on the respective orthology group, named isolates or coverage (join_result_tools_LMD9_genome file on Supplementary Material https://1drv.ms/f/s!ApASY1Jo99MLb0ba4XsI0OMCyBY). Through a gene-by-gene comparison was possible to compare how the different tools differ on the coverage, number of genomes included on the respective orthology group, where gene is assigned. This genome has 1,911 total CDS of which 230 are pseudogenes and 1681 coding CDS. Pseudogenes were not considered by tools. Out of 1681 coding CDS there are 254 CDS features annotated as hypothetical protein.

Taking into account the LMD9 gene identifier and respective number of genomes coverage by each orthology group across tools was measured the standard deviation It shows how widely coverage orthology group value are dispersed across tools. Tools are concordant in 1398 genes about their assignment on homology groups across the genomes. The remaining 283 genes, which is about 17% of all coding CDS, were assigned differently in at least one tool. Correlation coefficient between tools was calculated based on their coverage orthology group assignment to determine the relationship across tools (Table 8).

Table 8: Correlation coefficients based on the number of genomes covered by each orthology group across the different tools.

TOOLS	Roary	Proteinortho	OrthoMCL
FindMyFriends	0,895	0,938	0,817
Roary		0,934	0,803
Proteinortho			0,796

All tools have a strong relationship which might mean that the difference about their orthology group coverage assignment are not much dispersed. OrthoMCL had the lowest correlation coeficients. It is the furthest away tool having a more disparate coverage assignment. On the other hand, Proteinortho had the strongest relationship with FindMyFriends and Roary with about 0.93. The closest relationship was confirmed between FindMtFriends and Proteinortho. In spite of this FindMyFriends does not have a close relationship as Proteinortho has to with Roary.

Venn diagram summarizes the results obtained from comparing the four tools in relation to homology group coverage assignment (Figure 6). Count has done for each gene where tools either agree or disagree: intersections and disjunctions were calculated based on homology group coverage.

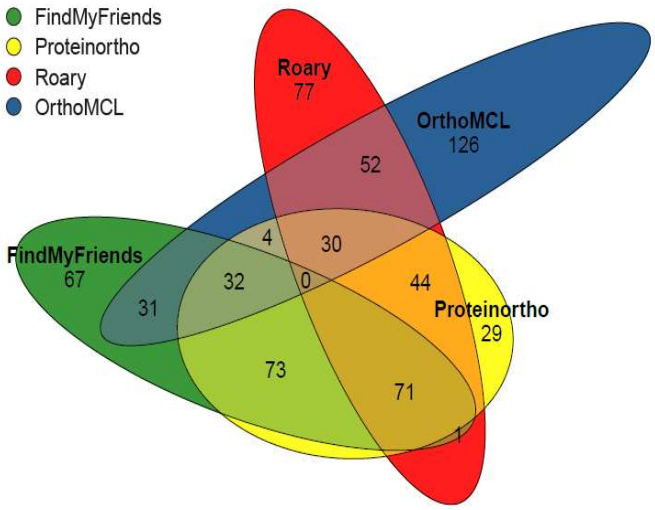


Figure 6: Homology group coverage overview. Venn diagram shows the common and unique homology group coverage of the different tools. Tools are represented by different colors: FindMyFriends (green), Proteinortho (yellow), Roary (red), OrthoMCL (blue). Since venn diagram is proportional, 1398 homology group coverage concordant with all tools was not represented. FindMyFriends, Roary and OrthoMCL intersection with a 8 value was not plotted [109].

OrthoMCL was the tool with a greatest difference at level of the orthology group coverage, i.e, number of genomes contained within it. It assigned to 126 orthology groups a different coverage than other tools. It could mean that it is careful when creating and grouping orthology groups or just to split paralogous from orthology groups established. Regarding similarity, Roary has more similarity with OrthoMCL than other tool. They agree in 90 orthology groups which is in agreement with the reason of these two tools do not have coverage parameter. Despite this, there is a slight difference of the correlation coefficient between OrthoMCL to FindMyFriends (0,817) and Roary (0.803). Although Roary agrees more with OrthoMCL, the difference between the number of genomes included on orthology groups is slighter higher than FindMyFriends. OrthoMCL is not so well related as the other tools are, having a correlation coefficient around 0.8, because creates more orthology groups than other tools. Roary has 77 orthology groups where the number of genomes included is unique. On the other hand, 145 orthology groups are in agreement with other tools. These results are in accordance with its correlation coefficients having a better relationship with Proteinortho than FindMyFriends.

Proteinortho has a broad overlaid over the other tools, meaning that have a better agreement over the other tools. It occurs in 254 homology groups with at least one tool. On the other hand, it disagrees 29 times with the other tools. Among them, FindMyFriends shows more agreement, 176 times, than all the other. Even that correlation coefficients of Proteinortho to the Roary and FindMyFriends are roughly the same value of 0.93, the different between the number of genomes cover by the orthology groups is greater on Roary than FindMyFriends. This can be explained by the fact that Roary has no coverage parameter. FindMyFriends and Roary have a good correlation coefficient with Proteinortho being these tools closest to Proteinortho. However, between each other there is a higher disagreement on the number of genomes included in each orthology group, with a correlation coefficient value of 0,895.

Computational time comparison is not fair because tools were run in different platforms. Clustering algorithms is the main important factor that causes time consuming. Even though genome sample can influence, it has not more weight than strategy implemented since pan-genome analyses is usually made up with hundreds of different genomes. Obviously, there must be an equilibrium point between the quality of results and the time consumed to choose the best tool according to goals. True computational cost of the different algorithms relies on their complexity. BLAST has an execution complexity of O(nˆ2) being relatively slow algorithm taking most of the computational time on tools. OrthoMCL took 1 hour and 30 minutes to run on the KBase platform. Proteinortho took 4 minutes and 20 seconds using 32 available CPU threads and Roary took 1 minute and 57 seconds. FindMyFriends took 2 minutes and 10 seconds on a personal computer. Clustering step of OrthoMCL takes a lot of memory and CPU consumption which makes it further in terms of time consumption. Proteinortho and Roary took lower computational time than OrthoMCL because they have run with a better processing capacity computer. If Proteinortho and OrthoMCL would be run with the same processing capacity, Proteinortho would answer with a lower computational time since it scales better [62].

Roary and FindMyFriends were the faster tools approaching a linear complexity. Once FindMyFriends does not have a blast approach it will answer with a better performance in terms of computational time. Despite of that, blast-approach Roary had a good computational time. The reason that it has a pre-clustering gene step to cut down time, or otherwise it would take as much time as OrthoMCL or Proteinortho. However, with the increasing heterogeneity, Roary reaches a non-linearity complexity which it

takes more time [60]. To briefly summarise, FindMyFriends is the fastest tool, Roary is the second and Proteinortho and OrthoMCL are slowest one slightly tied at the time.

5.2. *Streptococcus thermophilus* Pan-genome

This chapter seeks to explain the differences found on amino acid biosynthesis pathways, transporters and proteolytic system and figure out how these genomic variations could influence *Streptococcus thermophilus* phenotype. Furthermore, shine-dalgarno sequence, sequence -35 and pribnow sites were identified.

This work is based on pan-genome analysis, enabling to study several genomes at once. FindMyFriends built pan-genome and KEGG database annotation were used to annotate the pan-genome. This means that gene inside of an orthology group can be easier identified through K number matching. Repetitive, low-sequencing level and contaminant genomes were filtered out. Blast were done to know why a gene is absence on orthology group. Lastly, a bibliographic review was done to know how a gene disruption has influence on *Streptococcus thermophilus* metabolism. Then, prediction of new sites of shine-dalgarno sequence, sequence -35 and pribnow site were done in the upstream sequence and their distance was characterized.

5.2.1. Genome Filtration

Multi-sequencing strains arises redundancy that need to be filtered out. FindMyFriends established a pan-genome comprising 520 Chr. Hansen using 0.95 start similarity, 0.2 maximum deviation in sequence length and paralogous were not split. Correlation coefficients between genomes were calculated based on present or absence of the genes previously established within homology groups. Upper triangular of the correlation matrix displays in each line the correlations coefficient from one genome to another (Figure 7)**:** Excerpt from correlation matrix of the 520 Chr. Hansen genomes. Circles represents the correlation values between genomes. Red circles represent genomes correlated more than 95% while blue circles represent the other cases. Red triangles represent a group of genomes correlated more than 95%. ø represents excluded genomes. Complete figure on Supplementary Data 520CHR_corrplot_upper https://1drv.ms/f/s!ApASY1Jo99MLb0ba4XsI0OMCyBY.. A strain cluster, highlighted as a triangle, have a correlation coefficient great or equal than 0.95. There are other situations where these genome groups are not so well defined. For each line on correlation

matrix, it was counted the consecutive values greater than or equal to 0.95, being this value the size and the content of the cluster. The process continued either after cluster is found or in the next line. From the 520 Chr. Hansen genomes represented on the matrix, 49 clusters were identified which to each case it was select to keep one genome with the highest assembly level. For the biggest clusters, it was selected more than one genome (Supplementary Data Genome_Filtration https://1drv.ms/f/s!ApASY1Jo99MLb0ba4XsI0OMCyBY). Hence, 205 genomes were excluded, identified from A to D and from MC to TU, remaining 315 genomes.

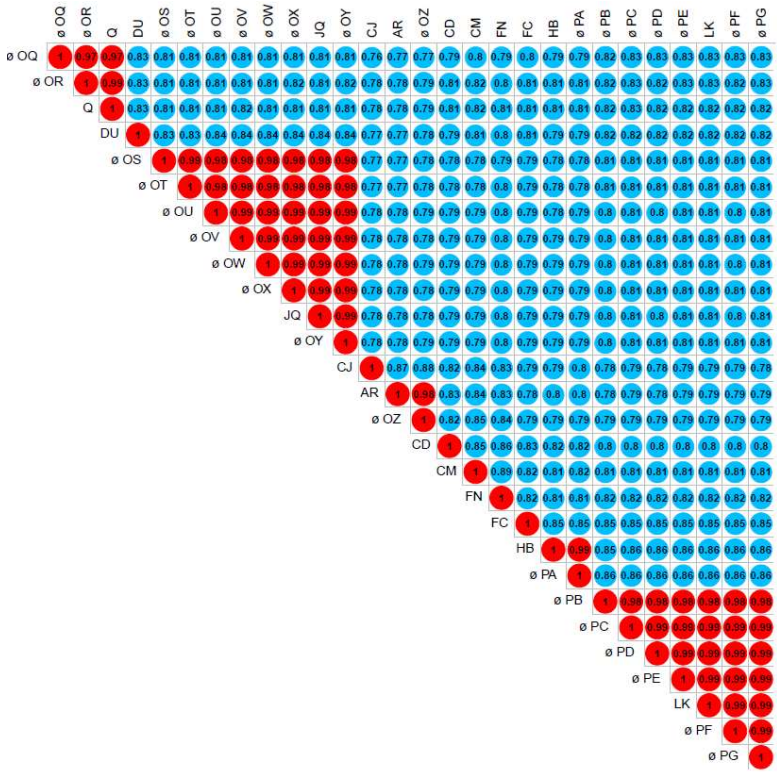


Figure 7: Excerpt from correlation matrix of the 520 Chr. Hansen genomes. Circles represents the correlation values between genomes. Red circles represent genomes correlated more than 95% while blue circles represent the other cases. Red triangles represent a group of genomes correlated more than 95%. ø represents excluded genomes. Complete figure on Supplementary Data 520CHR_corrplot_upper https://1drv.ms/f/s!ApASY1Jo99MLb0ba4XsI0OMCyBY.

After genome filtration, looking to the pan-genome matrix it is possible to see genomes with a small gene content because of their low sequencing level (Supplementary Data FMF_520CHR_unfiltered https://1drv.ms/f/s!ApASY1Jo99MLb0ba4XsI0OMCyBY). These genomes TQ, TR, TS, TT and TU were excluded.

5.2.2. Genome Diversity

The relatedness of the 336 strains were visualized based on their gene contents through a heatmap generated from pan-genome binary matrix (Figure 8). There is a great cluster diversity with several clusters containing few genomes and some genomes were not grouped with any genome. From the left to right it is possible note different degrees of orthology group presence across genomes. It is possible to observe a red large bar corresponding to core-genes, covering all the genomes. Afterwards, this bar begins to vanish appearing represented in some genomes, representing the accessory group. Finally, isolates single lines appears to represent the unique genes. Some genomes are grouped far away point to a dissimilarity to others. Genomes JC, HY and KQ fall in these characteristics. It is possible to see several gene groups exclusive for these genomes, raising questions about their specie. For this reason, a 16S ribosomal analyzes was done.

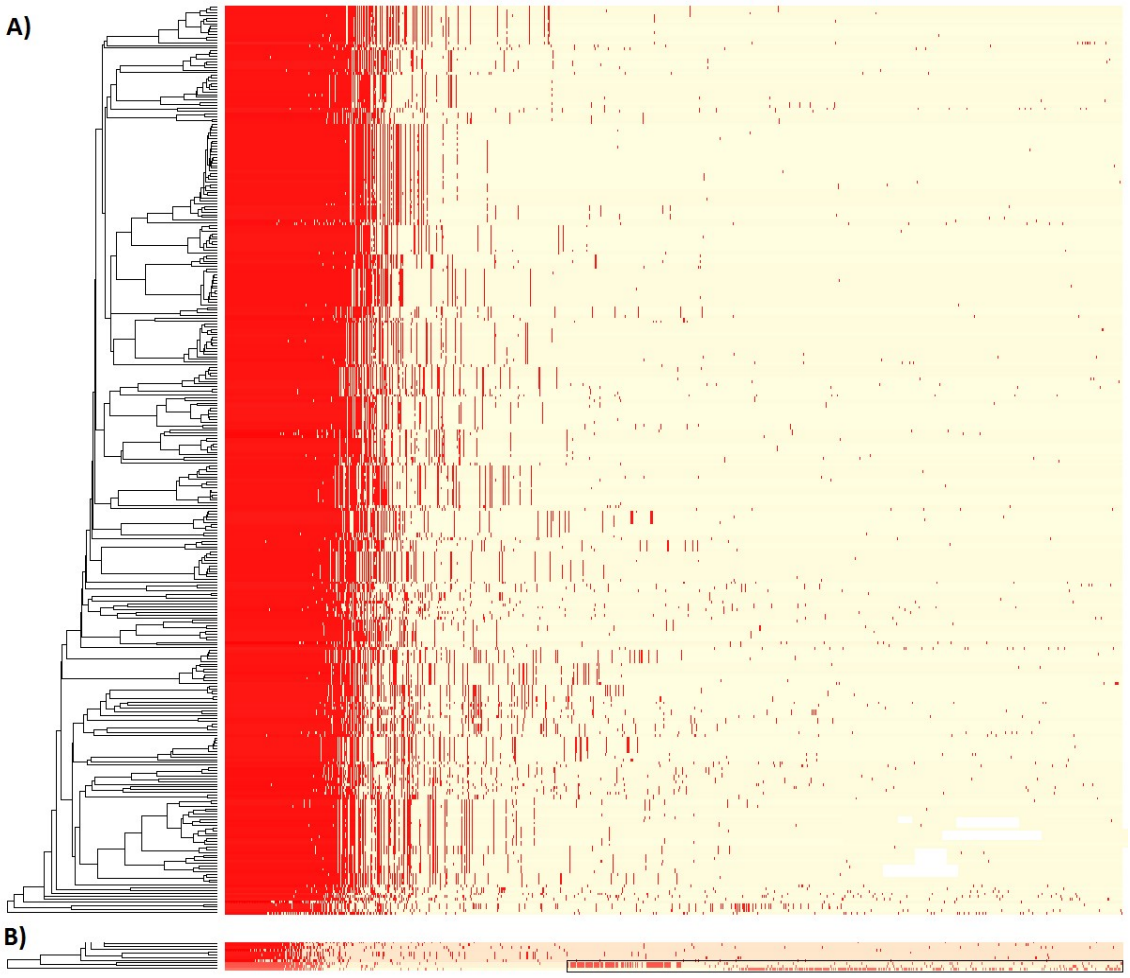


Figure 8: Gene content heatmaps across 333 strains (A). From the 336 strains heatmap (B), it was possible to identify 3 strains with a relative high number of genes (black box), which were excluded from the following analysis. Strains were ordered by hierarchical clustering. Columns correspond to each gene group across strains while each row to one genome. Red and beige means presence and absent relative to gene content, respectively. Complete pictures are represented on supplementary data and Heatmap_333 (A) and Heatmap_336 (B) https://1drv.ms/f/s!ApASY1Jo99MLb0ba4XsI0OMCyBY.

5.2.3. 16S ribosomal RNA Analysis

Suspicions have been raised for the genomes HY, JC e KQ due to an unsual gene number and their exclusivity for the genomes mentioned. They have 4041, 3801 and 5784 genes, respectively. Barrnap has predicted the location of two 16S ribosomal RNA genes to genomes HY and HQ and one to genome JC. Regarding taxonomy classification, rRNA alignment against SILVA database reveals that one 16S rRNA derived from genome KQ belongs to Bacillales order, Bacillaceae family and Bacillus genus, being the other four 16S rRNA genes from Lactobacillales order and Streptococcaceae family. Among these, one HY 16S rRNA sequence belongs to *Lactococcus* genus and the other three resulting from each genome belong to *Streptococcus* genus. Blast results against 16S ribosomal RNA sequences (Bacteria and Archaea) NCBI database (Table 9) have check this taxonomy classification results, where the previous *Streptococcus* genus sequences match the partial sequences of *Streptococcus thermophilus* strain ATCC 1925 with 100% identity. The *Bacillus* genus sequence from KQ match with the complete sequence in Bacillus subtilis strain IAM 12118 and the *Lactococcus* genus sequence from HY match with the partial sequence on *Lactococcus lactis* subsp. *tructae* strain L105. These genomes were excluded from the following analysis. The reasons behind it could be an inappropriate strain isolation for sequencing or contamination with other species.

Table 9: Blast hits from predicted 16s rRNA against 16S ribosomal RNA sequences (Bacteria and Archaea) NCBI database.

Genome	Gene Length (bp)	Sequence	Identity (%)	Accession
HY	1545	*Streptococcus thermophilus* strain ATCC 19258, partial sequence	100	NR_042778.1
HY	1545	*Lactococcus lactis* subsp. tructae strain L105, partial sequence	99	NR_116443.1
JC	1545	*Streptococcus thermophilus* strain ATCC 19258, partial sequence	100	NR_042778.1
KQ	1545	*Streptococcus thermophilus* strain ATCC 19258, partial sequence	100	NR_042778.1
KQ	1548	*Bacillus subtilis* strain IAM 12118, complete sequence	99	NR_112116.2

5.2.4. Pan-genome, Core-genome, and Evolution of Genome Composition

FindMyFriend pan-genome were plotted taking into account their gene content either on group evolution (Figure 9) or on final state (Table 10).

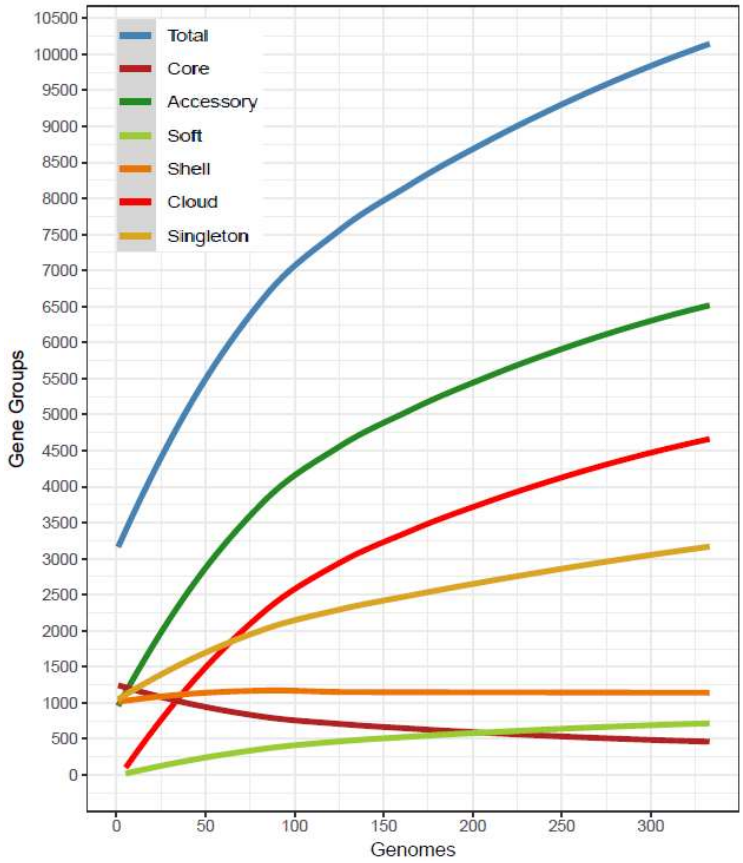


Figure 9: Pan-genome evolution in gene groups. It shows how the number of the different pan-genome partitions evolved as the size of the pan-genome increases. Total number of genes (blue), core genes (dark red), accessory genes (dark green), singleton genes (yellow). The accessory genes englobe the soft genes (light green), shell genes (orange) and cloud genes (red). A clear visualization (Zoom-in) on the starting pan-genome partitions on the first 21 genomes (Supplementary Data PlotEvolution_21out333 https://1drv.ms/f/s!ApASY1Jo99MLb0ba4XsI0OMCyBY).

Streptococcus thermophilus pan-genome appears not to have been reached the total number of possible genes, as depict in their trend to keep increase. Indeed, incomplete genomes could contribute to it. An estimation of the Streptococcus pan-genome surpasses 6000 genes [110]. On the other hand, there are 458 genes shared between all strains, core-genome, that appears to reach a plateau around this number. Accessory group englobes the soft, shell and cloud groups. Soft and shell groups represent 718 and 1144 genes, respectively. Both groups also seem to stabilize their increasing in gene numbers as many genomes are added, around 750 and 1250 genes, respectively. Cloud group are the most common group, representing around 46 % of the genes, equivalent to 4666 genes. After the cloud group, the singleton group is the second most common representing around 31 % of the genes, equivalent to 3175 genes. In the pan-

genome of this species, here are represented lineage specific genes. It also split paralogous, pseudogenes and incomplete sequences.

How many more genomes the pan-genome partitions represent, more genes are represented there, suggesting that they form a coherent group. The pattern of the total gene numbers relies on accessory group, specially cloud group, and unique group containing strain specific genes. It might be still influenced by the inclusion of new genome sequences. They might be biased since pan-genome were built with incomplete genomes. These analyses suggest that it is possible to determine a core genome at species level. Pathogenic streptococcal genus possesses an open pan-genome for which they need to extent its genomic content, through multiple ways of exchanging genetic material, in order to adapt and survive in multiple environments [89]. These results demonstrate an open pan-genome for *Streptococcus thermophilus*. Nevertheless, incomplete genomes create uncertainty being not possible to conclude about their open/closed pan-genome. On the comparison of tools to cluster orthologous groups chapter, pan-genome size has been increasing what makes it an open pan-genome. However, it has done with a small sample of 21 genomes and then more genomes need to be added to confirm it. An assumption about *Streptococcus thermophilus* pan-genome is that it might be or progress to a close pan-genome. This could be mainly due to milk environment adaptation for which *Streptococcus thermophilus* has a high level of gene decay related mainly with virulence potential but also with carbon source utilization [111].

Table 10: *Streptococcus thermophilus* pan-genome gene content. Their partitions, with the number genomes covered by them as well the number of genes group defined are represented.

PARTITION	GENOMES	GENE GROUPS
TOTAL	1-333	10161
CORE	333	458
SOFT	317-332	718
SHELL	50-316	1144
CLOUD	2-49	4666
UNIQUE	1	3175

From *S. thermophilus* pan-genome (Supplementary Data FMF_333ST https://1drv.ms/f/s!ApASY1Jo99MLb0ba4XsI0OMCyBY), the first nucleotide sequence from each homology group was picked up as a representative. After that, BlastKOALA was used to annotate the representative sequence against KEGG, assigned them a K number (Supplementary Data BlastKOALA_annotation

https://1drv.ms/f/s!ApASY1Jo99MLb0ba4XsI0OMCyBY). Orthology groups annotated with BlastKOALA were matched against desired pathways (Supplementary Data Pathways https://1drv.ms/f/s!ApASY1Jo99MLb0ba4XsI0OMCyBY). In most cases, representative sequence was used as a query to do a blast against genomes where it was not identified within orthology groups (Supplementary Data Blast_queries https://1drv.ms/f/s!ApASY1Jo99MLb0ba4XsI0OMCyBY) being the blast results saved (Supplementary Data Blast_results https://1drv.ms/f/s!ApASY1Jo99MLb0ba4XsI0OMCyBY). A gene not sequenced, it will be always absente either on the pan-genome or on the blast results. On the other hand, once a match is found, it will enable to know why the gene is missing from the orthology group. The main reason for a gene be absent is naturally fragmentation occurred in the edges of the contigs. This information will not be reported forward from here, but it can be consulted in the blast results.

After this point, the main features on the analyzed metabolisms in *Streptococcus thermophilus* will be highlighted and subsequently discussed.

5.2.5. Metabolism in Study

5.2.5.1. Central metabolism and pyruvate dissipating pathways

Homolactic fermentation is the only way to *S. thermophilus* regenerate NAD^+ for glycolysis reaching an equilibrated redox balance from glucose metabolism in anaerobic conditions [112]. One lactate dehydrogene has 328 amino acids length (OG203). Another paralogous, probably not functional, has 316 amino acids. The number of pyruvate dissipation enzymes is limited [112]. Besides lactate dehydrogenase, it was found the genes required to produce formate, acetate and acetoin. Regarding acetate production, genome EE has a shorter 258 amino acids acetyltransferase (OG11441) instead of a full-gene found in the remaining genomes coding 327 amino acids protein (OG606). This shorter protein resulted of one gap opening and mismatch that caused gene disruption. Genomes CQ (public strain S9) and KV also have an acetate kinase disrupted due to nucleotide mismatches, not having a 397 amino acids protein (OG789). In relation to acetoin prodution, acetolactate decarboxylase is not sequenced on genome KW while on genome JS a 60 bp nucleotide deletion occurred. This deletion is located within the nucleotide sequence starting from the 80bp to 140bp on the 720bp full-gene (OG804). A

second paralogous identified as acetolactate synthase large subunit (OG1036) is located is located next to this gene.

S. thermophilus are not able to produce acetaldehyde, from acetyl-coA, and ethanol and butanediol, from acetaldehyde. Fragmented alcohol/acetaldehyde dehydrogenase (*adhE* (K04072); *adhA* (K13953)) and butanediol dehydrogenason (*butA* (K03366)) are in agreement that these genes are pseudogenes [111]. A multicomponent acetoin dehydrogenase complex (AcoABCL) was reported in *B.subtilis* to be involved in acetoin utilization, as a carbon source, when the sugar source is drained [113]. High similarity is shared between this complex and pyruvate dehydrogenase complex [114]. It is present in every genome (although fragmented on contig ends in some genome), howerver, further studies need to be done to clarify their activity. This complex produce acetaldehyde, an important compounds related with aromatic qualities of yogurts [114].These changes are highlighted on Figure 10.

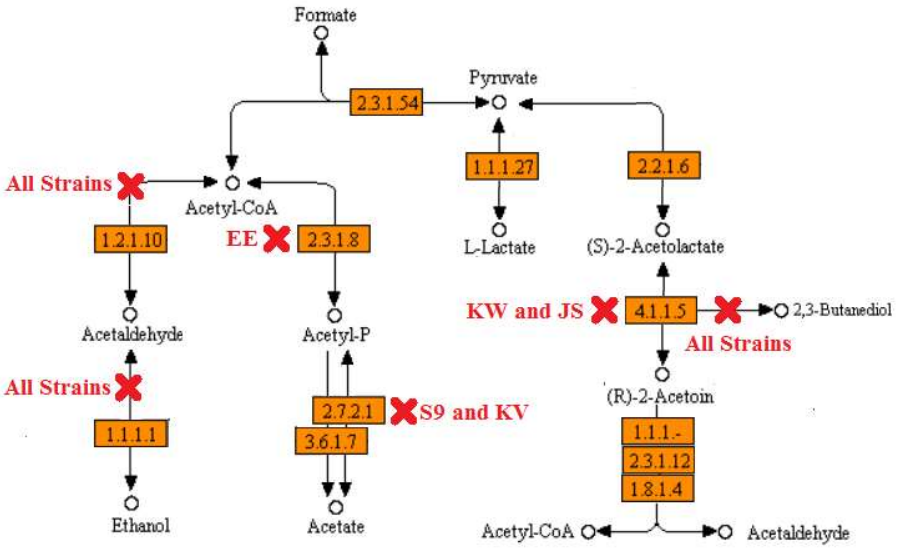


Figure 10: Defective strains found on pyruvate fermentation to different products and acetoin degradation metabolism. Ethanol and butanediol is not possible in *Streptococcus thermophilus*. Lactic fermentation is the main process to consume lactose present on milk. Although that, strains could not be able to produce acetate or acetoin as a minor end-product Red cross means that was found an evidence that the enzyme is not functional because somehow the respective gene is disrupted or absent. It is followed by the strains where it is not present.

5.2.5.2. Sugar Metabolism and Biosynthesis of Nucleotide Sugars

On the glucose catabolism branch, full-length glucokinase (*GlcK)* is disrupted in genomes S and GA, although a shorter protein was identified in these genomes (Figure 11). A single nucleotide deletion occurs in the genome S and an 9bp deletion in the genome GA. On the other galactose branch, it is either normally secreted or catabolized

by the Leloir pathway by certain strains, through operon *GalKTEM*. The inability of many strains metabolize galactose would be due to the low activity of *GalK* and *GalM* [114]. Either obtained from glucose or galactose, glucose-6-phosphate can be converted into pyruvate, through glycolysis.

Nucleotide sugars (UDP-galactose, UDP-glucose and TDP-rhamnose) are synthesized from the precursor glucose-1-phosphate and incorporated in polysaccharides synthesis, such as, extracellular polysaccharide (EPS). Three copies of gene *GalE* coding UDP-glucose 4-epimerases responsible by interconvertion between UDP-glucose and UDP-galactose were identified. Two copies encode two functional enzymes: one OG1157 is part of the galactose operon coding a 333 amino acids protein while another OG1124 code a 337 amino acids protein. Thirth copy OG1602 is a pseudogene. UDP-glucose can also be interconverted to UDP-glucuronate by UDP-glucose 6-dehydrogenase (EC 1.1.1.22), encoding ugd. It was identified a CDS in 69 genomes, including 7 public genomes. It was identified 8 ortology groups having a great variability in amino acid length. Orthology group OG2318 is the largest covering 50 genomes (including public DH EG, FJ, GJ and HU), containing a 387 amino acids protein, which it could suggest that is complete. Orthology group OG4337 could also contain a functional protein with 411 amino acids length in 7 genomes. The fact of the orthology groups OG4065 (include public strains BB and CE). and OG4336 contain CDS located side by side suggest that they are pseudogenes, in accordance with not be identified in most genomes and its length variability.

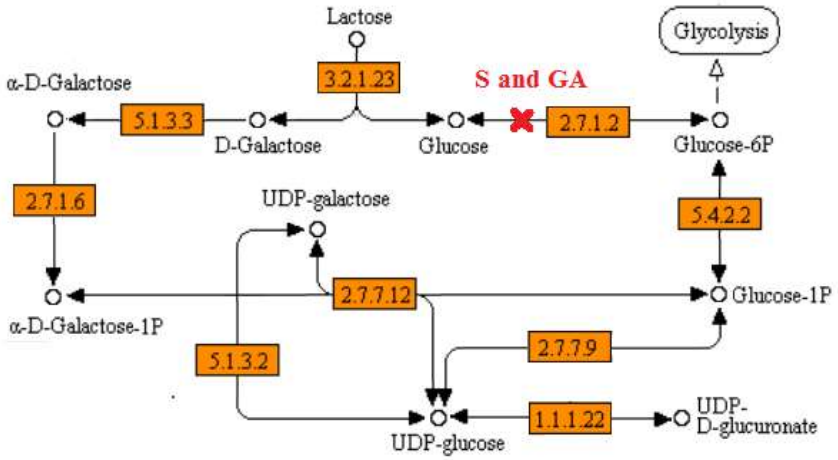


Figure 11: Defective strains found on sugar metabolism, Leloi pathway and biosynthesis of bucleotide sugars. Strains S and GS have a disrupted glucokinase gene. They are not able to use glucose moiety. Red cross means that an evidence was found that the enzyme is not functional because somehow the respective gene is disrupted or absent. It is followed by the strains where it is not present.

5.2.5.3. Amino Acids Biosynthesis

Alanine, Aspartate and Glutamate Metabolism

Alanine, aspartate and asparagine amino acids can be synthetized by all strains. Gene *alaA* (OG188) codes a 404 amino acids alanine-synthesizing transaminase (EC 2.6.1.2) being able to produce alanine from pyruvate. Gene *ald* (OG1312), which codes alanine dehydrogenase (EC 1.4.1.1) catalyzing the same reaction, is truncated in all genomes either in 112 amino acids or 95 amino acids fragments. Alanine can also be produced also from aspartate in 24 genomes. It is possible due to the presence of the gene *asdA* (OG2883) codes an aspartate 4-decarboxylase (EC 4.1.1.12) which have a 533 amino acids length. In relation to aspartate, it is produced from oxaloacetate through 393 amino acids aspartate aminotransferase (EC 2.6.1.1). Asparagine is produced from aspartate through a 330 amino acids aspartate-ammonia ligase (EC 6.3.1.1).

Glutamate dehydrogenase (EC 1.4.1.4) has 450 amino acids and produce glutamate from 2-oxoglutarare. From this amino acid glutamine is produced by a 447 amino acids glutamine synthetase (EC 6.3.1.2). All strains have genes to produce glutamate and glutamine (Figure 12).

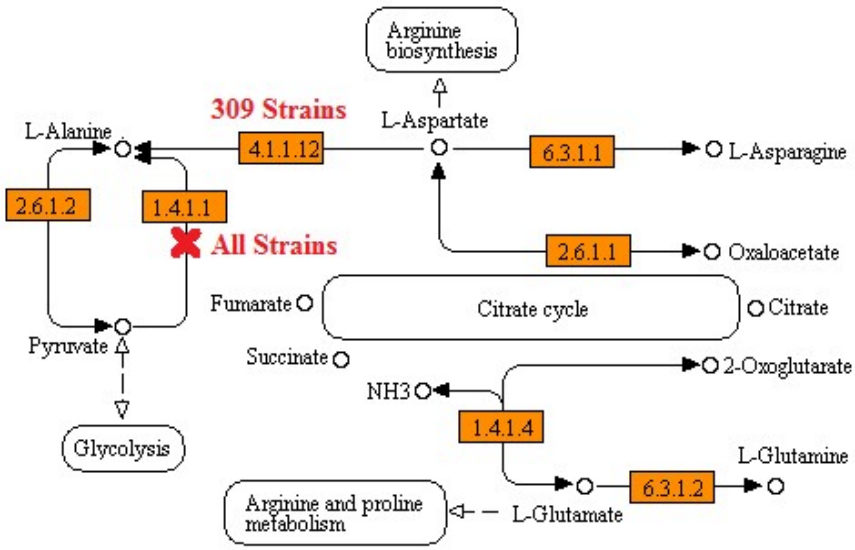


Figure 12: Defective strains on alanine, aspartate and glutamate metabolism. Alanine can be produced from aspartate in 24 strains. Red cross means that an evidence was found that the enzyme is not functional because somehow the respective gene is disrupted or absent. It is followed by the strains where it is not present.

Arginine and Proline Metabolism

Arginine is produced either from glutamate or aspartate, inside the urea cycle. This pathway is well conserved across genomes. Genome KW has a disrupted gene *argD*. Two protein fragments with 249 and 99 amino acids length were identified. A complete gene (OG864) match was found with 3 mismatches and 1 gap opening. These nucleotide changes are causing the disruption of the gene, possibly resulting in the substitution of one amino acid codon for a stop codon. The last reaction inside the urea cycle (EC 3.5.3.1) that converts arginine to ornithine were not found (Figure 13).

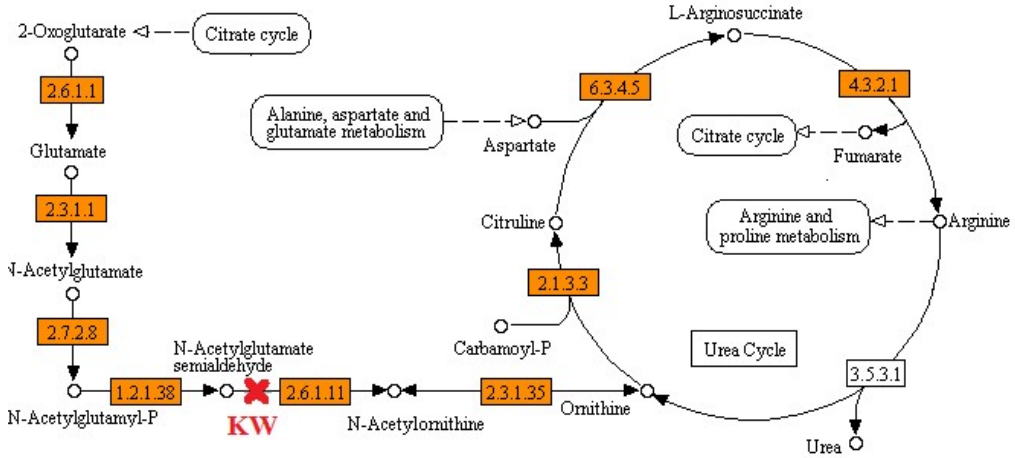


Figure 13: Defective strains on arginine metabolism. Strain KW could not produce arginine from glutamate because enzyme 2.6.1.11 could not be functional. Red cross means that an evidence was found that the enzyme is not functional because somehow the respective gene is disrupted or absent. It is followed by the strains where it is not present.

Glutamate is a precursor to produce proline. It requires an operon *proBA* (coding the enzymes EC 2.7.2.1 l and EC 1 .2,1.41, respectively) and another gene *proC* (coding the enzyme EC 1.5.1 .2). All genomes have these genes to produce proline. *Streptococcus thermophilus*, ornithine cannot be directly converted to proline either through ornithine cyclase (EC 4.3.1.12, K01750), as reported in *Clostridium sporogene*, or converted to intermediate pyrroline-5-carboxylate to produce proline (EC 2.6.1.13, K00819), as reported in *Bacillus subtilis* [115]. These genes were not identified in *S. thermophilus*.

Biogenic amines are toxic compounds produced at the end of metabolic activities. Amino acids decarboxylation removing the alpha-carboxyl group produce the respective biogenic amine. This process under stress conditions increase survival, restoring external pH and obtaining energy, however, it is not a desirable property on lactic acid bacteria [116]. Agmatine and putrescine are biogenic amines produced from arginine and

ornithine, respectively. *Streptococcus thermophilus* strains are not able to produce these two biogenic amines due to the absence of enzymes to produce ornithine (EC 3.5.3.1), precursor of putrescine (EC 4.1.1.17), and agmatine (EC 4.1.1.19). Despite that, 310 strains have genes to convert agmatine to putrescine (EC 3.5.3.11).

Histidine Metabolism

The total of the genes required to synthesis histidine from phosphoribosyl diphosphate (PRPP) where found in 265 genomes. On the genome GQ, pontual genes are present. Gene *hisG* codes a 216 amino acids ATP phosphoribosyltransferase (EC 2.4.2.17) that requires a regulatory subunit, encoding gene *histZ*. Subunit is disrupted in genome KW being found 11 mismatches and 1 gap opening. Gene hisE codes a 104 amino acids phosphoribosyl-ATP pyrophosphohydrolase (EC 3.6.1.31), which was also identified in genome GQ (GQ.fa_01139). Gene *hisI* (EC 3.5.4.19) and *hisA* (EC 5.3.1.169) were found in all genomes. Following gene *hisF* (EC 4.3.2.10), it was not identified on genome DM and apart from this genome, a larger 345 amino acids protein (GQ.fa_01137) was found in genome GQ, in comparison with 253 amino acids full length protein. Genes *hisH* (EC 4.3.2.10), *hisB* (EC 4.2.1.19) were found in these 265 genomes. The next gene *hisC* (EC 2.6.1.9) was also identified in genome GQ and it is disrupted in genome LY in two fragments (LY.fa_01021; LY.fa_01020). Indeed, complete nucleotide sequence was found through blast on contig 24 with 8 mismatches that caused disruption. Histidinol-phosphatase (PHP family) (EC 3.1.3.15) was found in 328 genomes. Histidinol dehydrogenase (EC:1.1.1.23), encoded *hisD*, was found in all genomes of the set. No strains have the capacity to produce glutamate from histidine, reaction EC 3.5.2.7 is absent (Figure 14).

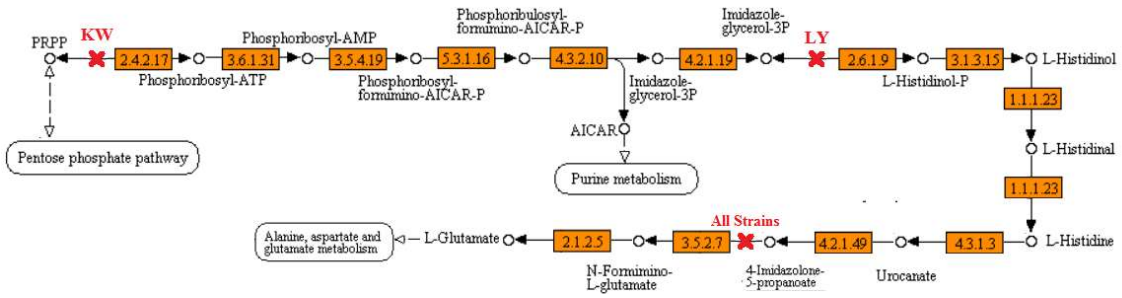


Figure 14: Defective strains on histidine metabolism. Apart from the 68 strains which histidine biosynthese genes were not identified, strains KW and LY could have a non-functional enzyme EC 2.4.2.17 and EC 2.6.1.9, respectively. Red cross means that an evidence was found that the enzyme is not functional because somehow the respective gene is disrupted or absent. It is followed by the strains where it is not present.

Phenylalanine, Tyrosine and Tryptophan Biosynthesis

Branched-chain amino acids enzymes were identified through the strains (Figure 15). Two genes, *aroF* (OG786) and *aroH* (OG1123) are located side by side, encode a 343 amino acids two 3-deoxy-7-phosphoheptulonate synthase (EC 2.5.1.54) isoenzymes. Genomes have at least one out of the two isoenzymes. Gene *aroH* was not identified in genomes EG and GS while the other gene *aroF* was not identified in genome CC. Next gene *aroB* codes a 355 amino acids 3-dehydroquinate synthase (EC 4.2.3.4), which is disrupted in genome CH due to a one nucleotide deletion, giving rise two protein fragments identified (CH.fa_00116; CH.fa_00115). Gene *aroD* coding a 225 amino acids 3-dehydroquinate dehydratase I (EC 4.2.1.10) is disrupted on genome BG. Complete length gene match was found with 4 nucleotide substituions that lead to a shorter 194 amino acids protein identified.

Regarding the tryptophan branch, anthranilate synthase component I (EC 4.1.3.27), encoded *trpE*, has 451 amino acids on 330 genomes. On genomes AB, BT and KI, it is disrupted with 21 mismatches and 1 gap opening being identified two proteins of 233 and 228 amino acids. The last enzyme on this branch is tryptophan synthase (EC 4.2.1.20). Alpha chain has 260 amino acids. On genomes FS, gene is disrupted with about 10 mismatches. Beta chain has 402 amino acids, and it is disrupted on genomes AB, BT and KI (public strain EPS) with 42bp mismatch. On tyrosine branch, prephenate dehydrogenase (EC 1.3.1.12) encode gene *tyrA2* is disrupted on genomes DS and HG (public strain B59671) with about 14 mismatches nucleotides. On phenylalanine branch, the last reaction is common on the two branch is catalyzed by a 350 amino acids histidinol-phosphate aminotransferase (EC:2.6.1.9), encoded *hisC*, which was found in 265 genomes, not being found on the remaining 67 genomes. Genome LY, 8 nucleotides mismatches disrupted gene, being identified two proteins with 176 and 163 amino acids.

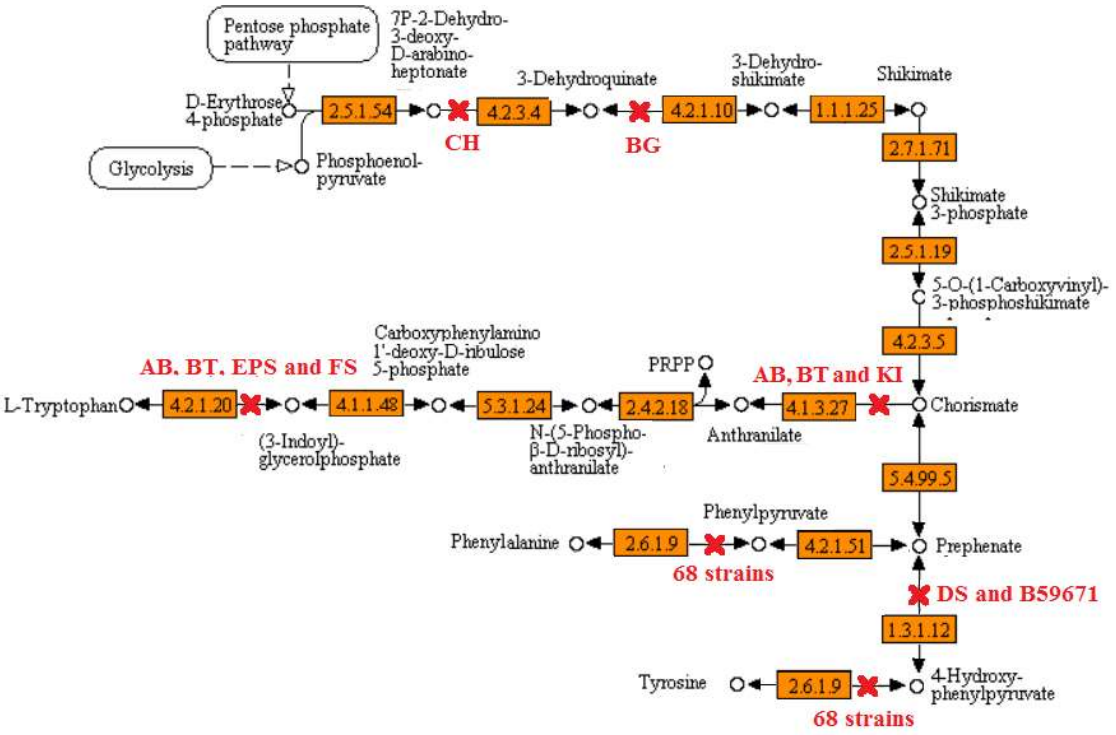


Figure 15: Defective strains on phenylalanine, tyrosine and tryptophan biosynthesis. Aromatic amino acids could not be produced in 77 strains. Strain CH and BG could not produce BCAA. Phenylalanine and tyrosine are the amino acids that could not be produce in 68 strains because of the gene was not identified, coding the enzyme EC 2.6.1.9, the same than histidine metabolism. Five strains could not produce tryphtophan and more 2 strains could not produce tyrosine. Red cross means that an evidence was found that the enzyme is not functional because somehow the respective gene is disrupted or absent. It is followed by the strains where it is not present.

Glycine, Serine and Threonine Metabolism

Serine is produced from 3P-glycerate, derived from glycolysis by gene *serA* (codes a EC 1.1.1.95), gene *serC* (codes a EC 2.6.1.52) and gene *serB* (codes a EC 3.1.3.3). Gene *serC* was not identified in genomes DM and IM. Genome HG (Public strain B59671) has the gene *serB* disrupted with 6 mismatches and 1 gap opening, being identified only a 173 amino acids shorter protein fragment. Conversion of serine to glycine is possible in all genomes by the presence of gene *glyA*, coding a hydroxymethyltransferase (EC:2.1.2.1). On the other hand, production of threonine from glycine is not possible in any of the genomes, not being found the respective reaction (EC 4.1.2.48, 4.1.2.5).

Instead of that, threonine is produced from aspartate through five enzymes. Genomes AF and GS could not produce threonine. Homoserine kinase (EC 2.7.1.39) was found disrupted on genome AF with 7 mismatches and 1 gap opening that lead to gene disruption being identified a shorter 111 amino acids protein. On genome GS, the end

381 amino acids of threonine synthase (EC:4.2.3.1) are encoded on contig end 62 being the initial part not sequenced (Figure 16).

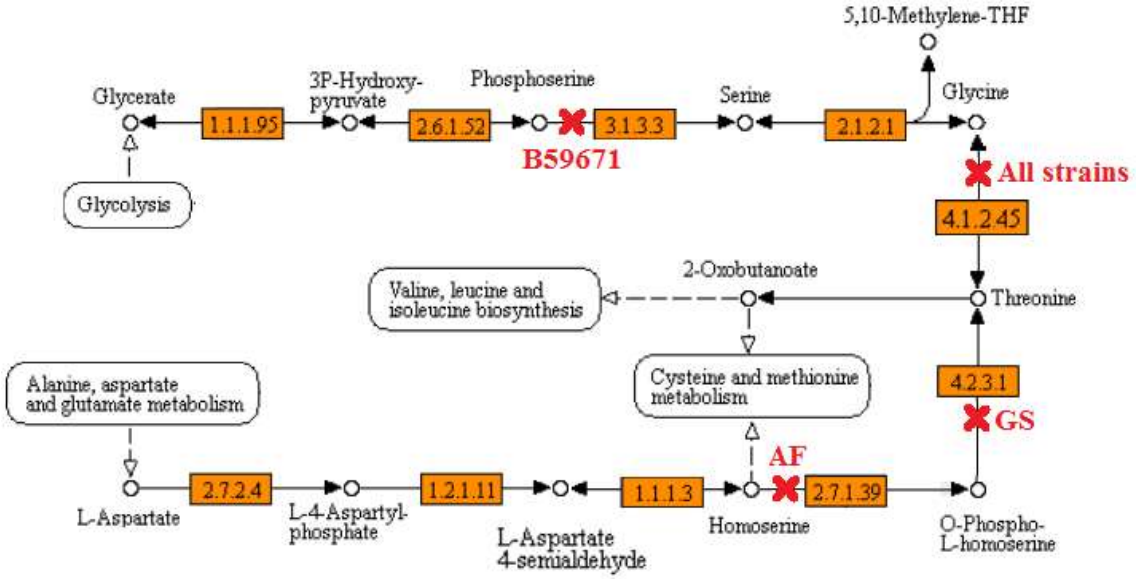


Figure 16: Defective strains on glycine, serine and threonine metabolism. Threonine production is not possible on the strains AF and GS due to the full-length gene coding the enzymes EC 2.7.1.39 and EC 4.2.3.1, respectively. Public strain B59671 has the gene coding for enzyme EC 3.1.3.3 disrupting being not able to produce serine. Red cross means that an evidence was found that the enzyme is not functional because somehow the respective gene is disrupted or absent. It is followed by the strains where it is not present.

Valine, Leucine and Isoleucine Biosynthesis

Threonine is converted to 2-oxobutanoate by threonine dehydratase (EC:4.3.1.19). The next four enzymes are common to both isoleucine and valine production. Two large subunits and one small subunit were identified to acetolactate synthase (EC 2.2.1.6). Homology group OG838 contains 566 amino acids large subunit that it is located near of the other ilv genes and side by side with small subunit. Another homology group OG1036 contains a paralogous protein. Small subunit has 158 amino acids. Gene *ilvC* codes a 340 amino acids ketol-acid reductoisomerase (EC 1.1.1.86). Genome AT (public strain JIM 8232), was two 214 and 98 amino acids fragments due to being disrupted with one mismatch and one gap opening. The last three genes are organized together on operon *ilvBNC*.

It was identified two copies of gene *ilvD* coding a dihydroxy-acid dehydratase (EC 4.2.1.9). Homology group OG481 contains an intact 567 amino acids protein while another homology group OG1518 contains a paralogous protein. It is located between ilvBNC operon and gene from homology group OG840. This gene looks not to be intact, however, their functionality needs to be established [114]. This second paralogous protein

was also found in different genomes: 574 amino acids in 42 genomes (OG2409) and 573 amino acids in 9 genomes (OG3962).

Valine-isoleucine branch splits from the penultimate reaction for a four sequential reaction to produce leucine. Enzyme 3-isopropylmalate/(R)-2-methylmalate dehydratase (EC 4.2.1.33) enzyme has a small and a large subunit. Identified with 460 and 196 amino acids, respectively (Figure 17).

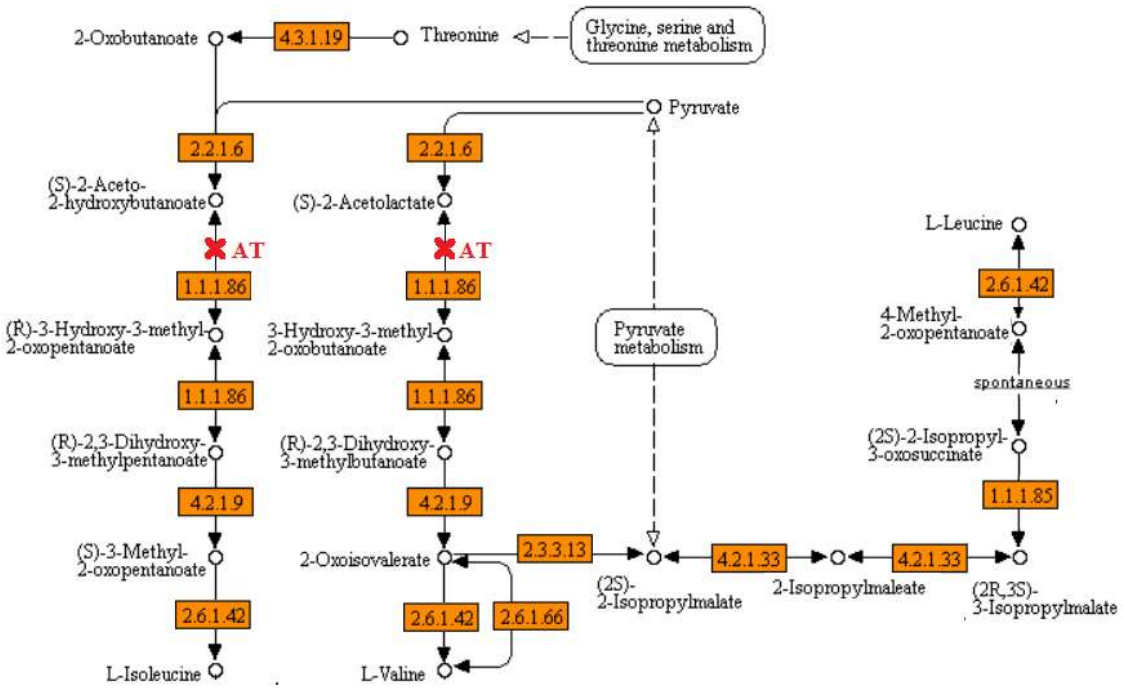


Figure 17: Defective strains on branched-chain amino acids metabolism. Strain AT has a disrupted gene coding enzyme EC 1.1.1.86 being unable of produce none of the three branched-chain amino aicds. Red cross means that an evidence was found that the enzyme is not functional because somehow the respective gene is disrupted or absent. It is followed by the strains where it is not present.

Cysteine and Methionine Metabolism

Cysteine is produced from serine. Serine O-acetyltransferase (EC 2.3.1.30), encoded *cysE*, has 206 amino acids. Shorter 174 amino acids protein (OG8558) was found on genome CJ being disrupted with 2 mismatches and 1 gap opening. A 193 amino acids paralogous protein was identified in 182 genomes (OG1621).

Methionine is produced from aspartate, which is converted to homoserine. Gene *metB* (OG599) codes a cystathionine gamma-synthase (EC 2.5.1.48) with 364 amino acids length that produce cysthationine either from succinul-homoserine or acetyl-homoserine. On genome EG (public strain ST3), one mismatch causes gene disruption being identified two shorter proteins of 236 and 107 amino acids (EG.fa_00394; EG.fa_00393). Two genes, *metC* and *patB*, codes isoenzyme cysteine beta-lyase (EC

4.4.1.13). Gene *patB* (OG969) is located side by side with gene *metB*. This gene is disrupted on genomes AB, BT, KI (public strain EPS) and FS with about 4 mismatches. Another gene *metC* (OG1265) is a paralogous gene, located near the paralogous gene cysK (OG1260), that was not identified in 33 genomes. Methionine is then produced on the SAM cycle.

Regarding SAM cycle reactions, both homocysteine S-methyltransferase (EC 2.1.1.10) and homocysteine methyltransferase (EC 2.1.1.14) catalyze the reaction that produce methionine from homocysteine are present. The enzyme DNA (cytosine-5)-methyltransferase 1 (EC 2.1.1.37) has a great variability across the genomes being identified across 131 genomes several sequences with different length. This enzyme could be fragmented on contig end or even not sequenced on the other genomes because the other SAM cycle enzymes are present. It could also be catalyzed by other enzymes not identified. Inside the cycle, homocysteine is produced back. Adenosylhomocysteinase (EC 3.3.1.1) was only identified in 258 genomes with about 548 amino acids. On genome CE (public strain MN-BM-A01), gene is disrupted with one gap opening. On the other hand, homocysteine is produced back through adenosylhomocysteine nucleosidase (EC 3.2.2.9) and S-ribosylhomocysteine lyase (EC 4.4.1.21) (Figure 18).

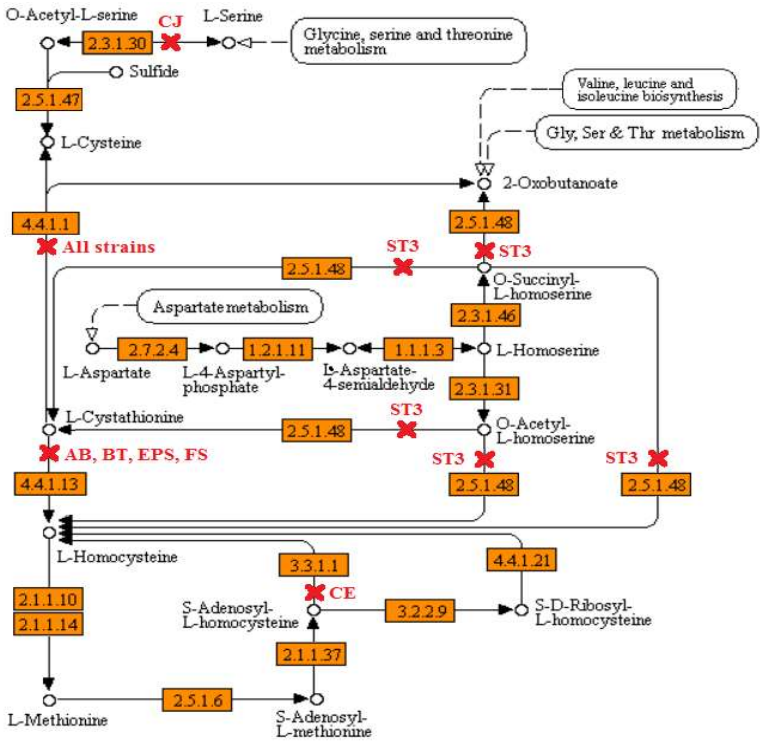


Figure 18: Defective strains on cysteine and methionine metabolism. Only one strain CJ is not able to produce cysteine. having the gene coding Serine O-acetyltransferase (EC 2.3.1.30) disrupted. Strains CJ and, AB, BT, EPS and FS are not able to produce methionine since they have disrupted genes coding

cystathionine gamma-synthase (EC 2.5.1.48) and cysteine beta-lyase (EC 4.4.1.13). Red cross means that an evidence was found that the enzyme is not functional because somehow the respective gene is disrupted or absent. It is followed by the strains where it is not present.

Recapitulation of the amino acids metabolisms is done, highlighting the strains that for some reason do not have a functional gene to produce respective amino acid (Figure 19). A larger number of strains do not have a full-length gene unabling them to produce histidine, aromatic amino acids and lysine. Special case, regarding lysine synthesizes genes, which are not yet fully understood in *Streptococcus thermophilus*. These genes not present could be due to a poor sequencing genomes or for some a gene is missing gene.

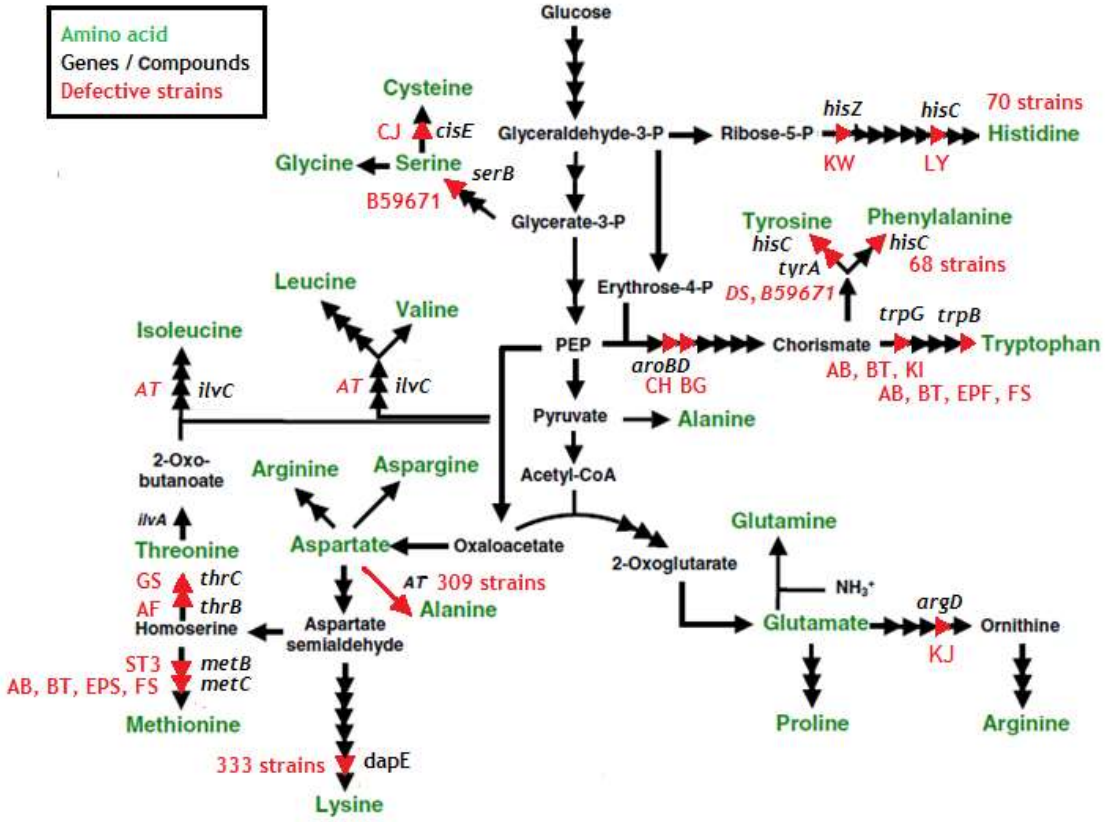


Figure 19: Overview of the potential auxotrophic strains on the amino acids biosynthetic pathways. Followed by the name (or number) of the strains that are not able to produce an amino acid from a percursor (red), it is also represented the respective disrupted gene (black). black arrows correspond to hipotetical active enzymes while red arrows correspond to potential inactive enzymes due to the absence of the corresponding gene. amino acids (green), genes and compounds (black) and defective strains (red) are represented. Adapted from hols *et all*, 2005 [114]. argD, acetylornithine aminotransferase; aroB, 3-dehydroquinate synthase; aroD, 3-dehydroquinate dehydratase; AT, undefined aminotranferase; cysE, serine acetyltransferase; dapE, succinyl-diaminopimelic descuccinlyasadipeptidase; hisC, atp phosphoribosyltransferase regulatory subunit; hisZ, atp phosphoribosyltransferase; ilvC, ketol-acidreductoisomerase; metb, cystathionine c-lyase; metc, cystathionine b-lyase; serb, phosphoserine phosphatase; thrB, homoserine kinase; thrc, threonine synthase; trpB, tryptophan synthase, b subunit; trpG, anthranilate synthase component ii; tyrA, prephenate dehydrogenase;

5.2.5.4. Amino acids ABC-Transporters

Lysine transporter (M00589) belongs to the polar amino acid uptake transporter. Both 516 amino acids permease (gene *lisC*) and 246 amino acids ATP-binding protein (gene *lysY*) were identified in all genomes. Branched-chain amino acid transporter (M00237) was identified in all genomes. It is composed of 391 amino acids substrate binding protein, two 294 and 316 amino acids permeases and two 254 and 236 amino acids ATP-binding proteins.

The sulfur atom of methionine is provided by cysteine or homocysteine, while the methyl group derive from a serine molecule. Thus, methionine biosynthesis involves three metabolic pathways: sulfur utilization, carbon backbone formation, and methylation [117]. Cystine uptake involves three different systems on *Bacillus subtilis*: two ABC-binding cassette (ABC) transporters, *tcyABC* and *tcyJKLMN*, and a symporte *tcyP* [118]. Regarding L-cystine transport *tcyABC* (M00234), it was identified the substrate-binding protein and permease. Through BlastKoala approach, not ATP-binding protein was identified but it is possible that another transport-system ATP could take over his function. Longest 265 amino acids substrate binding protein (OG2041) was identified in 84 genomes while the longest 190 amino acids permease protein (OG2011) on 88 genomes. These two genes are located side by side. There are great diversity of homology groups containing these non full-length gene, being identified a CDS feature coding them in other genomes. Further studies need to be done to clarify its activity, however, these preliminary data point to be probably pseudogenes. L-cystine transporter (M00585) encoded by operon *tcyJKLMN* were identified in 50 genomes. Genes coding their subunits are located together. One of the two substract binding protein, coding *tcyJ*, was not identified, however, it could not be essential if they do not work in synergy. Further studies should be done to understand that it is presence and if gene *tcyJ* is strictly necessary or other protein is fulfilling their absence. Cystein symporte transporter TcyP (K06956) was not identified. Putative S-methylcysteine transport system (M00586), encoded genes *YxeMNO*, is involved in the uptake of another sulfur source S-methylcysteine in *B.subtilis* [118]. Complete system transport was identified in 84 genomes. In addition to those genomes, regarding permease genome EN has it fragmented in contig end while genome LD was it disrupted due to a gap opening. Only substrate-binding protein was identified in genome GS. This transporter is complete in these genomes, hypothetically also in genome EN, and probably it is functional.

In relation to putative polar amino acid transport systems (M00236) identified is not clear their amino acid specificity. Polar amino acids include arginine, asparagine, aspartate, glutamine, glutamate, histidine, lysine, serine, threonine and tyrosine. In an attempt to establish these transporters, they were matched against a previous work that establish them [114] (Table 11). Several homology groups were identified containing their components. Four complete transporters are identified, and others are incomplete, probably due to transposons which disrupt them. There are other homology groups containing incomplete or disrupted sequences from the previous genomes were these sequences were not identified. There is a variety of putative polar amino acid transport system and further studies need to be done to clarify the characteristics of these transporters.

Regarding methionine transporter (M00238), their genes coding a permease and ATP-binding are located side by side while gene coding for substrate-binding protein is separated of them by pseudogenes. Substracte-binding protein with 300 amino acids length (OG963) was not identified in 4 genomes (AB, BT, GM and KI). Another paralogous gene (OG1240) coding a shorter 20 amino acids protein was identified. Incomplete or uncoupled transporter identified matched with the methionine transporter (M00238). Substract-binding protein (OG963) of methionine transporter (M00238) is coded side by side polar amino acid substract-binding gene (OG1778). These two genes are not coupled with permease (OG681) and ATP-binding (OG306) genes belonging to methionine transporter (M00238), which are also located side by side being seperated by a pseudogene. For this reason, this transporter could be incomplete being this pseudogene interfering with this transporter transcription, not allowing their activitivy.

Glutamine can be obtained either from peptides and casein or as a free amino acid. Glutamine synthesis is essential for *S. thermophilus* utter growth since these sources are not enough to reach their requirements [119]. Then, glutamine transporter system plays its role on glutamine uptake or releasing triggered by their internal concentration control [120]. Operon *glnHPQ* encodes a high-affinity transport system of glutamine [121], putative glutamine transporter (M00228).Gene *glnH* coding a glutamine binding protein (OG935) is disrupted on genomes EY, IH and IN being identified two CDS features side by side. There are two copies of the gene *glnP* (OG435 and OG936) located side by side coding a permease with a 216 and 232 amino acids length, respectively. Genome AW and DY have one gene (OG936) disrupted.

Table 11: Grouping of the polar amino acid transporters. ABC transporters annotated by KEGG as part of the putative polar amino acid transporter (M00236) based on strain Streptococcus thermophilus LMG 18311 were grouped as reported on supplementary data in [114].

Orthology Group a)	Strain LMG 18311	Description
Complete		
OG69	stu0875	polar amino acid ABC uptake transporter membrane-spanning protein
OG527	stu0876	polar amino acid ABC uptake transporter ATP-binding protein
OG46	stu0877	polar amino acid ABC uptake transporter substrate binding protein
OG734	stu1579	polar amino acid ABC uptake transporter substrate binding protein
OG135	stu1580	polar amino acid ABC uptake transporter ATP-binding protein
OG524	stu1581	polar amino acid ABC uptake transporter membrane-spanning protein
OG956	stu1582	polar amino acid ABC uptake transporter membrane-spanning protein
OG340	stu1652	polar amino acid ABC uptake transporter ATP-binding protein
OG418	stu1653	polar amino acid ABC uptake transporter membrane-spanning protein
OG678	stu1654	polar amino acid ABC uptake transporter substrate binding protein
OG1221	stu1501	glutamine ABC uptake transporter membrane-spanning protein
OG561	stu1502	glutamine ABC uptake transporter ATP-binding
Incomplete		
OG348	stu0605	polar amino acid ABC uptake transporter membrane-spanning protein
OG562	stu0606	polar amino acid ABC uptake transporter ATP-binding protein
OG902	stu1492	polar amino acid ABC uptake transporter substrate binding protein
OG901	stu1494	polar amino acid ABC uptake transporter substrate binding protein
OG134	stu1495	polar amino acid ABC uptake transporter substrate binding protein
Probably complete		
OG34	stu0288	cobalamin biosynthesis protein, putative transporter
OG512	stu0289	ABC transporter membrane-spanning protein
OG463	stu0290	ABC transporter ATP-binding protein
OG1349	stu0291	ABC transporter substrate binding protein
Incomplete or stu0296-297 uncouple with stu0301-302		
OG1778	stu0296	polar amino acid ABC uptake transporter substrate binding protein
OG963	stu0297	ABC transporter substrate binding protein
OG306	stu0301	polar amino acid ABC uptake transporter ATP binding protein
OG681	stu0302	ABC transporter membrane-spanning protein

[a)] Underlined Homology Groups: not identified as part of the putative polar amino acid transporter (M00236).

Quorum sensing is a process of control gene expression by signaling molecules on the environment. Oligopeptide transport system (M00439), named Opp, plays a role on signaling pathway by uptake these short signaling peptides. It also supplies auxotrophic bacteria with peptides that are digested by peptidases to free amino acid. It is composed of five subunits organized in an operon *oppABCDF*. Single genome can have multiple copies of the operon *opp* and/or genes [122]. The reason of this redundancy is not yet documented, however, it might be due to their location on a dynamic zone surrounding by mobile elements. This lead to a several truncated sequence. Oligopeptide transport gene were identified across genomes (Table 12). Given that it is essential, strain have at least one gene coding each subunit. Some full-length genes were not identified since they are either not sequenced or fragmented on contig ends. There are genomes that do not have a full-gene length on oligopeptide transport system sequence identified due to be fragment on contig end: 51 genomes in oppA, 3 genomes (HD, KW and LY) in oppB and oppC, 5 genomes (BT, DM, HD, KW and LY) in oppD and 28 genomes in oppF.

Dipeptide transport system (M00239) was identified in 23 genomes and it is responsible to uptake the dipeptide D-Ala-D-Ala into the cytoplasm [123]. It is composed of one 551 amino acids substrate-binding protein, two 317 and 352 amino acids permeases and one 565 amino acids ATP-binding protein. These genes are consecutively located. A second ATP-binding protein that constitute it was not identified being previous set of genes surrounded by transposable elements. Given its function on antibiotic resistance, lowering cell walls affinity [123], it makes sense to have missed this transporter in a non-pathogenic *S. thermophilus*.

Osmoprotectant transport system (M00209), encoded by *opuABDC* operon, was identified in 148 genomes. Full-gene length was identified in these genomes. Permease and ATP-binding protein are either not sequenced or fragmented in contig ends on genomes GS and EN, problably, it is not sequenced. Another osmoregulatory system, encoded *proXWV*, that transport glycine, betaine and proline was not found (M00208).

Electrochemical potential-driven serine/threonine transporter was found in all genomes. Other transporters were found without being linked to any functional hierarchy (KEGG BRITE). Amino acid transporter (AAT) family (K03293) is a subclass from amino acid/polyamine/organocation (APC) superfamily. There are 10 genomes in which it is disrupted. Proton-dependent oligopeptide transporter (POT family) (K03305) was

also found, being disrupted in 10 genomes. Branched Chain Amino Acid:Cation Symporter (LIVCS Family) (K03311) was found in all genomes

Table 12: Oligopeptide transport system (M00439) identification based on strain Streptococcus thermophilus LMG 18311 as reported on supplementary data in [114].

Orthology Group	Strain LMG 18311	Average Protein Size	Gene	Description
Complete, splitted 1				
OG1368	stu1438	309	*oppF1*	oligopeptide ABC uptake transporter ATP-binding protein
OG1002	stu1439	361	oppD	oligopeptide ABC uptake transporter ATP-binding protein
OG888	stu1440	308	oppC	oligopeptide ABC uptake transporter membrane-binding protein
OG881	stu1441	497	oppB	oligopeptide ABC uptake transporter membrane-binding protein
OG3938	stu1442	655	*oppA1*	oligopeptide ABC uptake transporter substrate-binding protein
Splitted2				
OG1515	stu1445	657	oppA3	oligopeptide ABC uptake transporter substrate-binding protein
Splitted3				
OG1426	stu0125	653	*oppA2*	oligopeptide ABC uptake transporter substrate-binding protein
Incomplete/others members truncated				
OG1653	stu0178	312	oppF	oligopeptide ABC uptake transporter membrane-spanning protein

Phosphate specific transport system Pst (M00222) is complete consisting of five subunits: a 291 amino acid substrate binding protein pstS, two permeases 304 amino acid pstC1 and 294 amino acids pstC2 and two ATP binding protein with 267 amino acids pstB1 and 252 amino acids pstB2. It is responsible to phosphate assimilation indispensible for energy supply and nucleic acid and phospholipid biosynthesis [124]. Public genome GJ (strain APC151) has this gene disrupted being found a complete match length sequence comprising two CDS (GJ.fa_00136; GJ.fa_00135).

The presence of the amino acid ABC transporters in strains are summarized in TABLE 13.

Table 13: Overview on ABC transporters analyzed. Transporters are identified with their module identifier as well as strain numbers where it is present.

Transporter	Identifier	Strains
Lysine	M00589	333
BCAA*	M00237	333
L-cystine	M00234	84
L-cystine	M00585	50
S-methylcysteine	M00586	84
Methionine	M00238	329

Transporter	Identifier	Strains
Polar amino acid **	M00236	333
Glutamine		309
Glutamine	M00228	330
Oligopeptide	M00439	333
Dipeptide	M00239	23
Osmoprotectant	M00209	148

* Branched chain amino acids (BCAA): Isoleucine, Valine and Leucine
** Four polar amino acid ABC uptake transporter (M00236), unknown amino acid specificity. One of them is glutamine transporter.

5.2.5.5. Peptidases

Since milk is poor in low molecular weight peptides and amino acids, *S. thermophilus* growth depends essential on extracellular serine proteinase (PrtS) to hydrolyze casein into oligopeptides. Even though many strains do not have it, they should rely in other serine extracellular peptidases to growth. These strains do not have a higher growth rate leading to a faster milk acidification. Oligopeptides are transported and degraded by a set of intercellular peptidases [125].

Lactopepsin (PrstS) is a serine cell envelope proteinase (CEP) that initiates casein cleavage. It was identified in 180 genomes with 1610 amino acids. Ten genomes have it fragmented on contig ends. Even so, strains containing the same PrtS allele have different acidification rates as a consequence of different regulation expression [126]. A limited set of intracellular peptidases were characterized (Table 14).

Two endopeptidases that break peptide bonds within the protein were identified. Oligoendopeptidase PepO (K07386) has 631 amino acids with endopeptidase activity belonging to metallo catalytic family. Their activity is over oligopeptides ranging from 5 to 30 amino acids length with a hydrophobic amino acid on the first position preferentially Phe or Leu. It is a monocistronic gene [127] not belonging to oligopeptide transport system (opp) operon as occurs in *L. Lactis* [128]. Strains have another oligoendopeptidase, named PepF or PepB (K08602). This one has a broad specificity cleaving substrates ranging from 5 to 23 amino acids length [129]. One copy of the gene *PepF1* was identified coding a 601 amino acids protein. The following peptidases are exopeptidases that catalyze the cleavage of the terminal peptide bond releasing single amino acid or dipeptide from the peptide. They are aminopeptidases that cleave the amino terminal and they are classified by their specificity.

Table 14: Proteolytic enzymes identified on *Streptococcous thermophilus* genomes. Genomes where it is present (total of 333 genomes), their amino acids length, catalytic class of peptidase and substract specificity are summarized.

Peptidase	Absence Strains	Amino acid length (aa)	Type [a]	Substract-Specificity [b]	Reference
Endopeptidase					
PepO	-	631	M	NH2–Xn↓Xn–COOH	[127]
PepF/PepB	-	301	M		[129]
Aminopeptidase					
PepA	-	355	M	NH_2–Glu/Asp↓X_n– COOH	[130]
PepC	-	445	C	NH_2–X↓X_n–COOH	[131]
PepN	CD	846	M	NH_2–X↓X_n–COOH	[132]
PepS	DQ,DS,ID	413	M	NH2–Arg/Ω↓Xn– COOH	[133]
Tripeptidase					
PepT	-	407	M	NH_2–X↓X–X–COOH	[134]
Dipeptidase					
PepV	-	468	M	NH_2–X↓Beta-Ala–COOH	[135]
Proline-specific (prolinase)					
PepP	-	353	M	NH_2–X↓Pro–X_n–COOH	[136]
PepQ	-	353	M	NH_2–X↓Pro–COOH	[137]
PepX/PepXP	-	755	S	NH_2–X–Pro↓X_n–COOH	[138]

[a]Catalytic class of peptidase according to sequence analysis or biochemical characterization
[b]The arrow indicates the cleavage site
M Metallopeptidase, C cysteine-peptidase, and S serine peptidase; Ω Aromatic Amino Acids

Glutamyl aminopeptidase, PepA (K01261), is a metallopeptidase with substrate specificity to glutamyl and aspartyl residues on the amino terminus. Serine can also be released with a less extent [130]. There are two general exopeptidases, aminopeptidase PepC and aminopeptidase PepN. Aminopeptidade PepC (K01372) belongs to cysteine family. It has an intracellular location with a broad aminopeptidase specificity releasing acid, neutral or basic amino acids. It is mandatory an unblocked N-terminal residue on the subtract [139]. Aminopeptidase PepN (K01256) is an intracellular metallopeptidase that hydrolyzes small peptides with a broad specificity to N-terminal amino acid, except for acidic amino acid, glycine and proline either last or penultimate position [132]. Genome CD has this gene disrupted being found a complete match length sequence with 6 mismatches. Aminopeptidase PepS (K19689) is intracellular metallopeptidase with an intermediate specificity hydrolyzing arginine or aromatic amino acids from peptides ranging in size from 2 to 10 residues. It also cleaves dipeptides with an uncharged amino acid (glycine, alanine) at first position [133].It was identified on 330 genomes. On the remaining genomes DS, DQ and ID, it is disrupted being identified two shorter proteins with 281 (OG5702) and 123 (OG5788) amino acids length coded side by side and matching complete gene length with 9bp mismatch length.

As the following peptidases only act on shorter peptides they act on the final steps of intracellular protein degradation. Tripeptidase PepT (K01258) is an intracellular metallopeptidase that hydrolyzes a variety of tripeptides cleaving their amino terminal, unless if they have proline and tyrosine on the last residues [140]. It is present in every genome with 407 amino acids length. Dipeptidase pepV is a metallopeptidase with a relative unspecific specificity, however, the majority of the dipeptides possesses a beta-alanine or D-alanine residue on the amino terminal. It was not associated with any KO number. It was identified on every genome with 468 amino acids length.

Proline specific peptidases have an important role since this amino acid is present in a high amount in casein derived peptides [141]. Because of its subtract specificity to proline, it may be provided by the combined action of several enzymes. Xaa-Pro Aminopeptidase PepP (K012629 specifically hydrolyzes N-terminal residue when proline residue is in the second position. It is also involved in the maturation of nascent but it does not replace PepM [136]. It is present in 332 genomes with 353 amino acids length. X-prolyl-dipeptidyl aminopeptidase PepX (K01281) has 755 amino acids plays a role on casein degradation releasing X-proline dipeptides from the N-terminal when proline is in the second position [138]. Xaa-Pro dipeptidase PepQ (K01271) has 361 amino acids and it has a strict specificity to dipeptidases with a proline on the last position [20]. Enzymes with carboxypeptidase activity was reported recently in LAB. Cell-associated extracellular, membrane and intracellular carboxypeptidases activities were reported not being so well characterized [142].

Pseudogenes proteolytic enzymes were also found on *S. thermophilus*. General dipeptidase, named pepDA or pepD (K08659) is a truncated cysteine peptidase. This dipeptidase first identified in *L. helveticus CNRZ32* has 474 amino acids length. A study indicated that pepDA is not well conserved on lactic acid bacteria and its deletion had not effect on growth rate in *L. helveticus CNRZ32* in milk [143]. Pseudogene dipeptidase PepE (K05995) also known as aspartyl dipeptidase hydrolyse dipeptides with an aspartate residue on the first position [144].

Other peptidases involved in quorum sensing, protein maturation, degradation of abnormal proteins and short-lived proteins, environmental stress resistance and response, antibiotic resistance and post-transcriptional modification of ribosomal RNA were found. It was also found pseudogenes related with pathogenesis and response to stress induced by antibiotics. These genes were undergoing to deletion on their genome lost in *S.*

thermphilus which could indicate that their function is not important and required in their milk environment [145].

5.2.6. Motif characterization

Motifs were discovered in upstream sequences and represented as position-dependent letter-probability matrices that describe the probability of each possible nucleotide at each position in the pattern. In order to discovery new sites of the motifs, these probability matrices were used to scan the sequences.

Pan-genome matrix was filtered out for the homology groups containing genes which were predicted being located inside of an operon. The promoter (in particular the −35 and the −10 regions) and the Shine-Dalgarno (SD) motifs were built (Figure 20). Motifs scanned on the upstream regions were post processed and saved (Supplementary Data https://1drv.ms/f/s!ApASY1Jo99MLb0ba4XsI0OMCyBY motifs_mat200bp_0.01). It still contains noisy sequences, such as, fragmented sequences, paralogous or pseudogenes that were not filtered out and therefore care should be taken when this information is taken on a gross way.

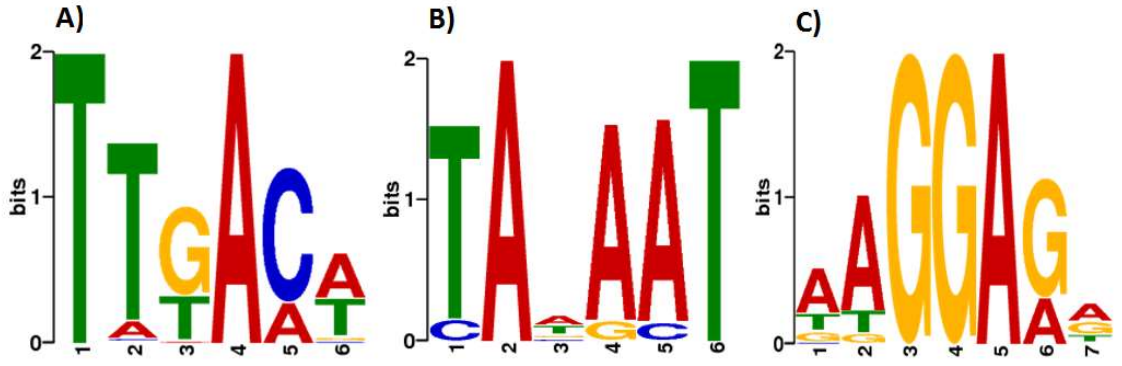


Figure 20: Sequence logo representing consensus sequence found for promoter -35 (**A**), promoter -10 (**B**) and shine-dalgarno sequence **(C)**. The relative sizes of the nucleotides (A, C, T or G) indicate their frequency in the sequences. The total height of the letters depicts the information content of the position, in bits. Meme output files from the discovered motifs are present on supplementary data (https://1drv.ms/f/s!ApASY1Jo99MLb0ba4XsI0OMCyBY Meme_ouput_promoters_motifs and Meme_ouput_SDSmotif).

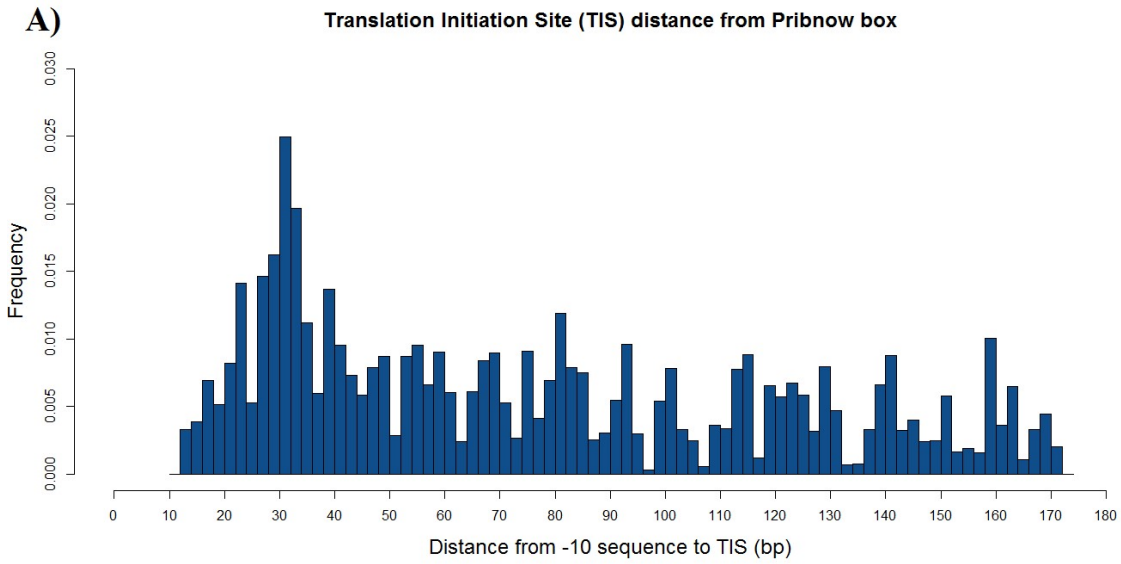
Transcription initiation is possible due to sigma factors association to RNA polymerase leading to promoter recognition. Thus gene expression modulation is possible due to the capacity of the sigma factors recognize specific DNA patterns [146]. Two DNA box promoters are located 10bp, named Pribnow box, and 35bp upstream of the transcription start site (TSS). They are roughly separated by 14 to 19 base pairs. Once DNA transcript to messenger RNA (mRNA), it could be translated to protein. Protein synthesis is facilitated by a nucleotide sequence, called shine-dalgarno sequence (SDS), located upstream from start codon on mRNA. During translation initiation, it is

complementary with 16S rRNA sequence helping to correctly position the ribosome [147].

Previous conserved motifs were discovered on genome AB (Figure 20) and scanned on the remaining genomes. All motifs built show a high-degree of conservation. Promoters motifs were built from 553 promoter predicted sequences. Promoter -35 motif TTGACW were found. It suggests every thymine at position 1 and adenine at position 4 in -35 sequence. Other positions are slightly different with thymine at position 2 (0.88), guanine (0.65) at position 3, cytosine (0.76) at position 5 and adenine (0.5) at position 6. Pribnow box hexamer motif, TAWAAT, is more conserved than the promoter -35. Positions 2 and 6 are completely probable to be an adenine and thymine, respectably. Nucleotide at position 3 were not found well defined on a motif. Usually, this position is a well conserved thymine. It can be a larger variation at this position in this specie, variating among the most frequent nucleotides, an adenine (0.39) and thymine (0.35).

Shine-dalgarno sequence was predicted from the 20 base pairs upstream on 1381 sequences. It is a conserved AG-rich hexamer motif AAGGAG. First two positions are more probable to contain an adenine either on the first (0.6) or the second position (0.78). Positions 3 to 5 are completely probable to be a conserved sequence GGA. Last position contains a purine base, although more probable to have a guanine (0. 74).

Spacer sequence length either between promoter -10 and shine-dalgarno or from this one to translation start site (TSS) were analyzed in order to if there is certain tendency (Figure 21). It was done broadly considering all sequences where the three motifs were identified and not to a subset of genes.



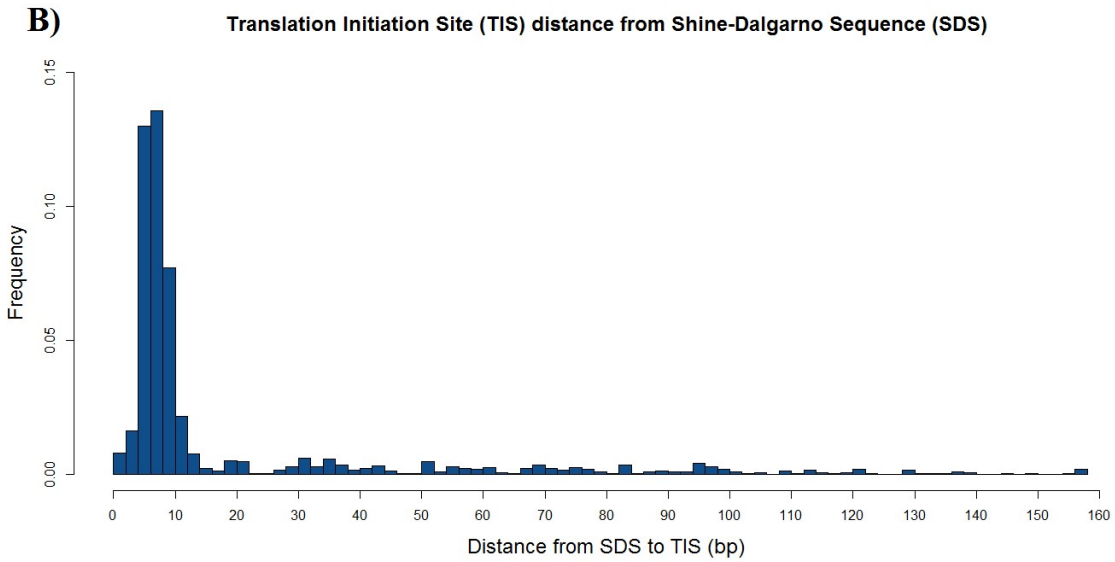


Figure 21: Histogram of distances of motifs to TSS in the promoter regions. **(A)** Distances for the Pribnow box (TAWAAT) motif, commonly called the -10 sequence, to TIS. **(B)** Distances for the SDS motif (AAGGAG) to TIS.

The 5′ untranslated region (5′UTR) is a region of the mRNA that begin at the transcription start site and ends right before of the initiation codon of the coding region. It was identified on a range of 11 to 173bp. At this range, this space length is much variable not being possible to identify a pattern where it is more centred. However, it is noticed that a space length between 20 and 40bp is found with a greater frequency. Shine-dalgarno sequence (SDS) correctly located lead to an efficient translation. Nucleotide composition and spacer length between SDS and translational initiator codon influences the SDS activity [148]. This spacer length is more stable than spacer that separates promoter -10 and the initiator codon. Overall, SDS is located within 3-12 nucleotides upstream from the start codon [147]. Although, SDS sequence was identified over the 158bp upstream sequence. Either on spacer length from promoter -10 or SDS to translation start codon, there are false positives which for the most part match fragmented sequences that were not filtered out. Meme suite will always find the motifs on the query sequences and thus there is a need to ensure the right balance between statistically and biological significance.

5.2.7. Discussion

Multi-sequencing strains lead us to have multiple strain sequenced genomes with different assembly levels. Filtration is needed to do not work with genome repetitions. This approach had a good performance in grouping several genomes of the same strain. However, in some cases was not perfect since it appears groups representing two strains together and there is a need to look them manually to pick up the high-assembly level genomes. Gene content heatmap in conjunction with 16s rDNA analyzes was a good strategy to caught genomes with an excessive gene content even considering that we are working with incomplete genomes. The reasons that lead to this was species contamination or defective strain isolation.

In dairy fermentation, it is desirable a high growth rate of *S. thermophilus* with a rapid lactose oxidation to lactate. Along with this, production of flavor and texture related compounds are also important. Although *S. thermophilus* is able to ferment several sugars, it is an obligate homolactic bacterium. Regarding fermentation end-products, lactacte is the main product (>95%) revealing a high adaption to grow on lactose. Other low level products were detected: acetaldehyde, acetate, acetoin, diacetyl and formate. Milk environment adaption lead to a regressive evolution of *S. thermophilus* revealing a low flexibility on sugar metabolism with several pseudogenes and non-functional genes related to uptake and sugar utilization [114]. Overall, pyruvate dissipating routes were found complete for end-products synthesis. Lactate and formate pathways were found in all strains. As in other, acetate and acetoin pathways, although with specific situations. Diacetyl is produced by a non-enzymatic reaction from acetolactate. It has been reported acetaldehyde production from threonine catabolism by enzyme *GlyA* (K00600) as well as from acetoin. As expected, ethanol is not produced. During growth in milk under anaerobic conditions, *pfl* protein, involved on formate production, is one of the most abundant proteins and also had a stimulatory on *Lb. bulgaricus* growth in a co-operation process [114]. Strains EE, CQ and KV could have some variations on acetate end-product production. Disruption of phosphate acetyltransferase (K00625) on strain EE could lead to inability to produce acetate. Strain CQ (public strain S9) and KV could have different kinetic parameters on acetate production due to use of acylphosphatase (K01512) in compensation for disrupted acetate kinase (K00925). Therefore, short 92 amino acids acylphosphatase functionality needs to be studied, however, their length is around the

same than other species, such as, *Bacillus subtilis* subsp. *subtilis 168*, *Streptococcus salivarius JIM8777*, *Escherichia coli K-12 MG1655*, etc. Strains JS and KW could not produce the end-product acetoin since acetolactate decarboxylase (K01575) is disrupted. Since small amounts of acetoin is produced from pyruvate, its function is less important in maintaining pH homeostasis and NAD^+ regeneration. Large amounts of diacetyl, a flavour compound, is produced either in the disruption of acetolactate decarboxylase or branched-chain amino acid deprived medium. Another function of acetolactate decarboxylase is the regulation of branched acid amino acids biosynthesis by the deviation of acetolactate to acetoin production. The absence of this enzyme could have effect on growth rate due to not occur growth stimulation [149]. These two strains could have a greater capacity to produce branched amino acids and diacetyl. In theory, the other strains containing a full-gene, once lactose, sugar source, is drained on medium, they are able to produce highest amounts of acetaldehyde, a flavor compound, and acetyl-CoA from the small amounts of acetoin end-product.

Commonly, lactose is broken in glucose, which is metabolized, and galactose, which is expelled. Shorter 278 amino acids glucokinase (K00845) seems to be not functional on strains S and GA. This mutation on glucokinase could be an advantage enabling them the hypothesis to metabolize galactose rather than glucose moieties. They are not able to grow on exogenous glucose source. Limited number of strains are able to ferment galactose sugar. They need to be good in other properties which will require further studies to identify mutations either on glucokinase and other elements, such as, galactose operon promoter and glucose phosphotransferase system. From the combination of *L. delbrueckii* subsp. *bulgaricus* with these preliminary *S. thermophilus*, a sweeter yogurt could be obtained with residuals lactose level and high levels of glucose, responsible for the sweetness enhancement [6, 53].

Lactic acid bacteria have different amino acid requirements for growth. *Streptococcus thermophilus* is less strict regarding exogenous amino acids source than other lactic acid bacteria. However, it also needs a exogenous source of amino acids in order to growth, being the essential amino acids strain-dependent [50, 51]. While some strains do not have any amino acid requrements, others are auxotrophic [152]. For a minimal growth, they need one sulfur amino acid and histidine, which is required only in some strains [112].

It has a wider amino acid metabolism than other lactic acid bacteria. Overall, strains have the genes required for amino acid biosynthesis pathway: sulphur containing

amino acids (cysteine and methionine), aromatic amino acids (phenylalanine, tyrosine, tryptophan), branched-chain amino acid (leucine, isoleucine, and valine), glutamate, glutamine, alanine, arginine, aspartate, asparagine, glycine, serine and threonine. As milk is a protein rich environment and *S. thermophilus* has been used for centuries on milk fermentation, it might be expected a loss of some amino acid pathways [112]. The importance of amino acid biosynthesis for growth in milk reflects the conservation of functional amino acid syntheses genes being found few pseudogenes associated with them.

Any strain is able to produce lysine despite its pseudogene dapE (K01439) as well as lacking gene *dapF* (K01778). Regarding alanine synthesis be guaranteed by alanine transaminase (K00814) from pyruvate, they also have a pseudogene alanine dehydrogenase *alaD*, catalyzing the same reaction. Other 24 strains could have an advantage to be able to produce alanine from aspartate due to the presence of gene *asdA*. Branched acid amino acid biosynthesis could not be possible on strain AT (public strain JIM 8232). Once confirmed gene *ilvC* interruption, this strain will not reach an optimal [153]. A second pseudogene *ilvD2* was found in all strains. Other 9 strains could be unable to produce leucine due to absence of gene *leuC*. Regarding aromatic amino acids, strains CH and BG have a disrupted gene *aroB* and *aroD*, respectively, on chorismate pathway which could mean their inability to produce none of them. Some strains have a disrupted gene being unable to produce one of the aromatic amino acids: 1) tryptophan: strains AB, BT and KI disrupted gene *trpE*; strains FS disrupted gene *trpA*; KI (public strain EPS), AB and BT disrupted gene *trpB*; 2) tyrosine: strains DS and HG (public strain B59671) disrupted gene *tyrA2*. 3) Both tyrosine and phenylalanine: 68 strains do not have gene *histC*, also involved on histidine biosynthesis. About 263 strains, corresponding to 79%, have genes that code for enzymes required to histidine biosynthesis. The remaining strains are in opposition with the affirmation that *Streptococcus thermophilus* are prototrophs to histidine [152]. Probably, from a common ancestral of the two strains sets were deleted the histidine cluster. Single histidine omission on medium lead to a no growth phenotype [112]. Once it is an essential amino acid in some, histidine prototrophic strains could have a lower histidine requirement on medium.

Regarding sulphur amino acids, there are strains cysteine or methionine auxotrophic. Strain CJ have a disruption on gene cysE leading to an incapacity to produce cysteine. If cysteine amounts, from degradation of peptides or import, were not enough to cover their demands, as they prefer cysteine over methionine, this strain will have a

lower growth rate [112]. Another sulfur amino acid, methionine, could not be produced due to gene disruptions either on gene metB on strain EG and gene patB (K14155) on strains AB, BT, KI (public strain EPS) and FS.

Strains have the genes to incorporate ammonia into oxoglutarate, citric acid cycle component, to produce glutamate and then glutamine. Growth experiments have shown that the addiction of glutamate improves growth rate on a minimal medium [112]. However, authors point absence of glutamate synthase gene as the reason for absence of growth in glutamate and glutamine omission experiments and glutamate auxotrophy of some strains [114]. After fermentation in a medium without glutamine, the extracellular glutamine concentration was approximately equal to the one found in the yoghurt whey. This situation reports a regulatory system restricting the inside glutamine levels and a glutamine transporter. For fermented products, it could have nutritional consequences [120].

Incomplete full-length gene *serB* involved in serine biosynthesis was not found on strain HG (Public strain B59671). *S. thermophilus* genotyping tool has been used based on the nucleotide variation within of this housekeeping gene [154]. Further studies on this gene are important to understand either their functionally or other present genes. Strain HG could be a serine auxotrophic. Strain AF and KW could be a threonine and arginine auxotrophic, respectively, since they have a disrupted gene *thrB1* and gene *argD*. Strains are proline prototroph, however their low-level constitutive expression does not provide sufficient proline. This explains why the addition of proline to the medium increase the growth rate. Casein proteolysis can supply proline, however, growth rate can still be enhanced with addition of proline to medium [115]. Strains are aspartate and asparagine prototrophic. Aspartate biosynthesis is essential for the growth since aspartate in milk is not enough for growth [155]. As they require high amount of asparagine, it could be supplemented with their synthesis [120].

Phosphate uptake transporter (M00222) is present in all genomes. Public genome GJ (strain APC151) acquire mutations that cause disruption of gene *PstB2* coding phosphate-transport-ATP-binding protein [156]. It could result in a less capacity of phosphate uptake. Complete lysine ABC transport system (M00589) and the common transport system for the uptake of branched-chain amino acids (M00237), leucine, valine and isoleucine, is present in every strain [157]. This branched-chain amino acid transporter has a complementary role with BCAA biosynthesis pathway since they are essential to optimal growth and scarce on milk [153].

Regarding uptake of a sulfur source, cysteine and methionine are two sulfur containing amino acids. Strains do not have some functional transport systems to uptake cystine, dimer form of the amino acid cysteine. Another sulfur source, S-methylcysteine, could be uptaken for 85 strains through a S-methylcysteine transport system (M00586). This could mean an advantage on taking other sulfur sources when available. Strains AB, BT, GM and KI could not be able to uptake methionine since they do not have subtract-binding protein, part of D-Methionine transporter (M00238). Further analysis must be done to see the functionality of paralogue gene coding a substrate-binding protein. Despite these transporters absences, every strain should be able to obtain sulfur containing amino acids either hydrolyze available sulfur-containing peptides or synthesize them.

Four complete polar amino acid transporter (M00236) found need to be studied in detail in order to know their specificities to amino acids/peptides. One of the transporters could not be functional on strains LY, KW, FN and CM. Glutamine is essential to growth on milk environment [119]. This transporter is used to glutamine exchange either in uptake when present as a free amino acid or in the liberation, mainly throughout log phase [120]. Two glutamine transporters were found. Glutamine-glutamate transport system (glnPQ) is absent in about 20 strains. Putative glutamine transporter (M00228) is absent in strains AW, DY, EY, IH and IN. Further studies should be done to validate their functionality and specificity to glutamine. These two glutamine transporters could fulfill the not functionality of each one. These absences on strains could mean a less capacity or even an incapacity to exchange free amino acids. However, strains can obtain glutamine/glutamate from peptides [158]. After growth, glutamine enrichment environment may influence nutritional potential of the milk fermented products [120].

Oligopeptide ABC transport system is essential for optimal growth in milk [159]. OppA (OG1426) and oppBCDF genes are located together and organized in an operon structure. Other copies of the gene oppA (OG1515 and OG3938) are dissociated and surrounded by traces of insertion sequence elements suggesting a mobilization from oppA. The copies also work with the components of the transport system. Organization could be different in some genomes because some of the opp genes are pseudogenes. Functional extra or absence of copies should have an impact on growth in milk [114]. Inactivation of entire set of genes oppA led to severe decrease on growth rate in milk or even the non-growth using as nitrogen source peptides [159]. Single mutation on oppA genes pointed different negative level effect being the homologous oppA3 (OG1515) the

worst. As oligopeptide-binding proteins are attached to membrane, their mobility is reduced being restricted to subtract in their surroundings. Then, strains with a larger copy number of functional oppA genes are more advantageous since oligopeptides transporter is facilitated through stoichiometry changes[159]. Oligopeptide binding proteins triple mutant had a very slow and limited growth rate indicating that there is no other oligopeptide binding protein [159].

Strains contain a di/tripeptide proton dependent transporter (K03305), homologous DtpT from *Lc. Lactis* [111]. It was reported a present of at least one dipeptide transporter since opp system mutant grew in a methionine-containing dipeptides medium [159]. Their co-existence with ATP-dependent di-tripeptide transport Dpp system (M00239) occurs in *Lc. Lactis* [160] but not happen in *Streptococcus thermophilus*. However, Dpp fragments were found in *Streptococcus thermophilus*. In some strains, such as *S. pyogenes*, these di- and tripeptide transport systems ensure enough di- and tripeptides and amino acids to normal growth in a manner that opp system is not essential for growth [159].

Osmoprotectant transport system (M00209) was found in 146 strains. These strains have the advantage to have another osmoprotectant, which protect them against osmotic pressure. However, their functionality needs to be tested. Even if transporter has not activity, it should not be an issue regarding survival, since there are other osmoprotectants. Then, strains without this osmoprotectant must rely on other ones. Other amino acid transport system-encoding genes are present and predicted to be complete and functional. Those include branched-chain amino acid ABC transporter (Liv; K03311), serine/threonine transporter (K07862), amino acid transporter (AAT) family (K03293) and di- or tripeptide:H+ symporter (K03305).

Transport system for amino acid and/or peptides were identified on genomes and some of the genes appear to be subject to duplication and breaking actions.

The majority of proteolytic enzymes were identified as expected since it mainly influences the absorption of the nitrogen source as well as the formation of flavor compounds on fermented dairy products [161]. Endopeptidases (PepO and PepF), tripeptidase PepT, dipeptidase PepV, prolinases (PepP, PepX, PepQ) and aminopeptidases (PepA and PepN) are present in all strains.

As milk is a protein rich environment in caseins, its hydrolyses is vital for bacterial growth providing free amino acids. Cell envelope-associated protease PrtS identified in 180 strains has the advantage of an optimal growth and acidification rates in milk when

grow alone in milk [125]. Despite the presence of the *PrtS* gene, these strain could show a *PrtS*-negative phenotype due to a change of their regulation [162]. *PrtS*-absent strains are viable with a slower growth due to the existence of cell-associated extracellular carboxypeptidase and peptidyl dipeptidase activities that have a weaker proteolytic activity than *PrtS* [142]. PrtS-absence strains to reach an optimal growth need to be co-cultivated with other PrtS-positive strains, such as *Lb. bulgaricus*, to use their released peptides through a cooperation process [114].

A study on *L. lactis* shows that no peptidase is essential but that the combination of multiple peptidase deletions had a detrimental effect on bacterial growth, not growing at all. Single peptidase deletion activity may be replaced by another peptidase with low level of activity against the same amino acids residues, what it is enough to sustain growth but not to reach optimal growth rate [163].

Strain CD does not have peptidase N, even though complete match sequence was found on contig end, any CDS was predicted due to maybe be disrupted with six nucleotide substitutions. Since peptidase N is a housekeeping gene [164], further analysis should be done do detect this gene functionality. This absence could lead to a lower acidification ratio due to a slower optimal growth. Since it supplies majority of free amino acids due to broad specificity. Latter on cheese ripening, it also contributes to lower bitterness contributing to flavor development through amino acids addition precursor of components related in flavor and taste, such as phenylalanine, leucine, tyrosine but particularly methionine [132]. Peptidase S was found disrupted with 9 nucleotide substitutions and 1 gap opening on strains DQ, DS and ID. These strains will not reach wild-type growth because this absence is not completely compensated by other peptidases. It could mean a pleiotropic role through its involvement in growth via nitrogen nutrition due to their highly specificity to a limited number of subtracts and other cellular function on peptidoglycan metabolism [165]. It is involved on flavor development on dairy products through liberating aromatic amino acids, precursors of aroma compound, from bitter peptides [133]. Once confirmed these functional genes, strains CD, DS, DQ and ID could not be so advantageous to milk fermentation due to their lower growth-rate. This causes a lower composition of free amino acids released which are important to production of flavor compounds.

The data arising from the motif characterization could be used together with transcriptional and proteomic data to study different classes of genes. A method on *L. lactis* shows that strength of promoters can be modulate either by their promoter sequence

or spacer length which separate the two promoters [166]. Sigma factors tolerate mismatches on -35 and -10 sequences promoters. However, promoter strength has been reported as correlated with the similarity to the matching consensus sequence [146]. Prediction of an efficient translation can provide useful information to optimize bacterial host for the production of compounds [167]. SDS effectiveness is determined by hybridization of mRNA-rRNA and space length from start codon. Correlation between SDS and predicted expression levels based on start codon biases, functional gene classes, distance between successive gene and type of start codon can be studied [168]. Rate synthesis are correlated with SDS location and mRNA-rRNA hybridization stability which could also be used to expose genome annotations errors [147].

6. Conclusion and Future Work

6.1. Assessment of Orthologous Detection Tools

This is an overview of different tools for clustering orthologous genes across different genomes. The choose of the best program is not straightforward since it relies on user goals. It is possible to set a distance point between different tools, however, a reference point is needed to achieve to best tool. In order to do that, gene specific analyses are required to understand if the genes are grouped correctly on orthologous groups. In the case where there are some differences, it will be important to describe in which sequences, for example, on pseudogenes or types of mobile genetic elements. An idea was obtained how many homology groups were assigned equally on tools, however, a more specific analyze might be done to see how different they are regarding coverage assignment and if it makes sense on the genes involved.

It is better to use a tool with a coverage parameter, such as, FindMyFriends or Proteinortho, which will split the pseudogenes of the full-length genes. FindMyFriends is a good option because it is easy-to-use and the time complexity increases linearly with the genome size without the requirement of an advanced hardware. Considering new sequenced genomes or pan-genome studies across different species, FindMyFriends is also a good package since it allows to save even more time. It has addGenomes and mergePangenomes functions that allow to join more genomes to a pan-genome already built or merge pan-genomes without the need of rerun again. Proteinortho can also be an option since it shows similar results but in spite of that it takes more time to run. Otherwise, if you know in advance that the considered genomes have a residual value of pseudogenes, a tool with no coverage parameter can be a good performance since they will not create too much confusion when grouped with the orthologous. Roary can be a good choice since it has a parameter to perfectly split or not the paralogous. OrthoMCL could be a little bit tricky to use because it was an intermediary behavior related with the paralogous. It does not have a split paralogous option and moreover it keeps together recent paralogous and orthologous, splitting only the ancient paralogous.

Parameter setting is an important factor in these tools comparison studies. Since tools have their own algorithms, it is tough to set equally the parameters across tools because in some cases they are based on different ones. Parameters on OrthoMCL, e-

value and Markov clustering inflation, are unique and different than other tools. More strict detection (lower e-values) and tightness clustering (larger inflation index) lead to lower false-positive rate [98]. Thus, it makes it more difficult to adjust other tools that rely on similarity and identity. Other tools even have cut-off length parameters which it was made an attempt to smooth it removing the pseudogenes identified.

6.2. *Streptococcus thermophilus* Pan-genome

Lactic acid bacteria play a key role in in the production of fermented foods and beverages. Several studies have been done to create added-value products in order to fulfill consumer demands regarding sweetness, appearance, texture, and flavor. *Streptococcus thermophilus* is a major starter lactic acid bacteria for yoghurt and cheese providing the enzymes in the formation of specific flavors. Amino acids are the major precursors for aroma compounds.

For this work, bioinformatics tools were used to search in genomes for essential components, such as proteinases, peptidases, aminotransferases, enzymes for biosynthesis of amino acids, and transport systems for peptides and amino acids. Full complement of proteins involved in flavour-forming reactions, and hence the potential for formation of specific flavour compounds could provide insight into it. It could provide useful information regarding different ability to produce fermentation end-products and also precursors of compounds that contribute to flavor and texture. KEEG-based strategy is subject to errors on KO assignment. Genes without an KO assignment are not retrieved, but even so, they are grouped together in orthology groups being more tough to retrieve them through prokka annotation.

Differences about amino acid dependency, transporters and proteolytic system were also described in specific strains that could influence their way to growth and demands. Genomic differences can be used to expand with new insights genome-scale metabolic models. Related to genome sequence, it might also be useful to increase genome assembly level and review some genomes in order to check if there are sequencing errors.

Data generated in this work will allow a deeper analysis relative to the hypotheses raised and open doors to a future work:

- Exopolysaccharide biosynthesis cluster characterization: once homology groups are defined they may be used to do characterize it;

- Gene functionality: transcriptional data could be coupled together with genomic content in order to check activity of some genes previously reported;

- In depth description of inter-strain variation in genomic regions encoding industrially relevant traits and investigation of correlations between genomic signatures and known phenotypes. Characterization of the intergenic spacer between 16S-23S: type of the intergenic spacer is correlated with phenotypic properties, such as, proteolytic and acidifying activity [162].

- Nucleotide variability characterization, i.e. identification of mutation (SNP, indels, etc.) and how they could affect gene expression. As an example, it has been reported a variation in a substrate binding segment on *PrtS* which could interfere with their subtract specificity [140, 190].

- Transcriptional and proteomic data may be used together with promoter and SDS motif identified. How transcriptional and translation effectiveness relies on several characteristics, such as, motif nucleotide variation, spacers length and start codon [53, 58, 74];

Bibliography

[1] K. Burgess, 'Milk and Dairy Products in Human Nutrition (2013), by E. Muehlhoff, A. Bennett and D. McMahon, Food and Agriculture Organisation of the United Nations (FAO), Rome. E-ISBN: 978-92-5-107864-8 (PDF). Available on web-site (publications-sales@fao.org).', *International Journal of Dairy Technology*, vol. 67, no. 2, pp. 303–304, 2014.

[2] P. Hols *et al.*, 'New insights in the molecular biology and physiology of Streptococcus thermophilus revealed by comparative genomics', *FEMS Microbiology Reviews*, vol. 29, no. 3 SPEC. ISS., pp. 435–463, 2005.

[3] C. Delorme, 'Safety assessment of dairy microorganisms: Streptococcus thermophilus', *International Journal of Food Microbiology*, vol. 126, no. 3, pp. 274–277, 2008.

[4] T. B. Rasmussen, M. Danielsen, O. Valina, C. Garrigues, E. Johansen, and M. B. Pedersen, 'Streptococcus thermophilus Core Genome: Comparative Genome Hybridization Study of 47 Strains', *Appl Environ Microbiol*, vol. 74, no. 15, pp. 4703–4710, Aug. 2008.

[5] Y. J. Goh, C. Goin, S. O. Flaherty, E. Altermann, and R. Hutkins, 'Specialized adaptation of a lactic acid bacterium to the milk environment : the comparative genomics of Streptococcus thermophilus LMD-9', *Microbial Cell Factories*, vol. 10, no. Suppl 1, p. S22, 2011.

[6] G. Giraffa, A. Paris, L. Valcavi, M. Gatti, and E. Neviani, 'Genotypic and phenotypic heterogeneity of Streptococcus thermophilus strains isolated from dairy products.', *J Appl Microbiol*, vol. 91, no. 5, pp. 937–943, Nov. 2001.

[7] D. J. Edwards and K. E. Holt, 'Beginner's guide to comparative bacterial genome analysis using next-generation sequence data', *Microbial Informatics and Experimentation*, vol. 3, no. 1, p. 2, Apr. 2013.

[8] A. Ali *et al.*, 'Bacteriology & Parasitology Microbial Comparative Genomics : An Overview of Tools and Insights Into The Genus Corynebacterium', vol. 4, 2013.

[9] J. M. Otero and J. Nielsen, 'Industrial systems biology', *Biotechnology and Bioengineering*, vol. 105, no. 3, pp. 439–460, 2010.

[10] H. Kitano, 'Systems biology: a brief overview.', *Science*, vol. 295, no. 5560, pp. 1662–1664, Mar. 2002.

[11] M. Wojcieszek, M. Pawełkowicz, R. Nowak, and Z. Przybecki, 'Genomes correction and assembling: present methods and tools', *Proc. SPIE*, vol. 9290, pp. 92901X-92901X–8, 2014.

[12] L. Liu *et al.*, 'Comparison of next-generation sequencing systems', *Journal of Biomedicine and Biotechnology*, vol. 2012, 2012.

[13] P. H. C. G. de Sá *et al.*, 'GapBlaster—A Graphical Gap Filler for Prokaryote Genomes', *PLOS ONE*, vol. 11, no. 5, p. e0155327, May 2016.

[14] M. Baker, 'De novo genome assembly: what every biologist should know', *Nature Methods*, vol. 9, pp. 333–337, 2012.

[15] H. Ellegren, 'Comparative genomics and the study of evolution by natural selection.', *Mol Ecol*, vol. 17, no. 21, pp. 4586–4596, Nov. 2008.

[16] J. Touchman, 'Comparative Genomics', *Nature Education Knowledge 3(10):13*, 2010.

[17] X. Xia, *Comparative Genomics*. Berlin Heidelberg: Springer-Verlag, 2013.

[18] T. Primrose, S.B R. M., *Principles of Genome Analysis and Genomics*. Wiley-Blackwell, 2003.

[19] M. S. Clark, ‘Comparative genomics: the key to understanding the Human Genome Project.’, *Bioessays*, vol. 21, no. 2, pp. 121–130, Feb. 1999.
[20] P. Chain, S. Kurtz, E. Ohlebusch, and T. Slezak, ‘An applications-focused review of comparative genomics tools: capabilities, limitations and future challenges.’, *Brief Bioinform*, vol. 4, no. 2, pp. 105–123, Jun. 2003.
[21] M. J. Pallen and B. W. Wren, ‘Bacterial pathogenomics.’, *Nature*, vol. 449, no. 7164, pp. 835–842, Oct. 2007.
[22] B. Hu, G. Xie, C.-C. Lo, S. R. Starkenburg, and P. S. G. Chain, ‘Pathogen comparative genomics in the next-generation sequencing era: genome alignments, pangenomics and metagenomics.’, *Brief Funct Genomics*, vol. 10, no. 6, pp. 322–333, Nov. 2011.
[23] H. Tettelin *et al.*, ‘Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial “pan-genome”.’, *Proc Natl Acad Sci U S A*, vol. 102, no. 39, pp. 13950–13955, Sep. 2005.
[24] H. Tettelin, D. Riley, C. Cattuto, and D. Medini, ‘Comparative genomics: the bacterial pan-genome’, *Current Opinion in Microbiology*, vol. 11, no. 5, pp. 472–477, 2008.
[25] C. Donati *et al.*, ‘Structure and dynamics of the pan-genome of Streptococcus pneumoniae and closely related species’, *Genome Biology*, vol. 11, no. 10, 2010.
[26] B. Contreras-Moreira and P. Vinuesa, ‘GET_HOMOLOGUES, a versatile software package for scalable and robust microbial pangenome analysis.’, *Appl Environ Microbiol*, vol. 79, no. 24, pp. 7696–7701, Dec. 2013.
[27] S. Bentley, ‘Sequencing the species pan-genome.’, *Nat Rev Microbiol*, vol. 7, no. 4, pp. 258–259, Apr. 2009.
[28] D. Medini, C. Donati, H. Tettelin, V. Masignani, and R. Rappuoli, ‘The microbial pan-genome.’, *Curr Opin Genet Dev*, vol. 15, no. 6, pp. 589–594, Dec. 2005.
[29] I. Tamas *et al.*, ‘50 million years of genomic stasis in endosymbiotic bacteria.’, *Science*, vol. 296, no. 5577, pp. 2376–2379, Jun. 2002.
[30] J. Chun *et al.*, ‘Comparative genomics reveals mechanism for short-term and long-term clonal transitions in pandemic Vibrio cholerae’, *Proc. Natl. Acad. Sci. U.S.A.*, vol. 106, no. 36, pp. 15442–15447, Sep. 2009.
[31] A. Muzzi, V. Masignani, and R. Rappuoli, ‘The pan-genome: towards a knowledge-based discovery of novel targets for vaccines and antibacterials.’, *Drug Discov Today*, vol. 12, no. 11–12, pp. 429–439, Jun. 2007.
[32] H. C. den Bakker *et al.*, ‘Comparative genomics of the bacterial genus Listeria: Genome evolution is characterized by limited gene acquisition and limited gene loss’, *BMC Genomics*, vol. 11, p. 688, Dec. 2010.
[33] J. Olivares, A. Bernardini, G. Garcia-Leon, F. Corona, M. B Sanchez, and J. L. Martinez, ‘The intrinsic resistome of bacterial pathogens’, *Front Microbiol*, vol. 4, p. 103, 2013.
[34] A. Caputo *et al.*, ‘Pan-genomic analysis to redefine species and subspecies based on quantum discontinuous variation: the Klebsiella paradigm’, *Biol Direct*, vol. 10, Sep. 2015.
[35] E. W. Sayers *et al.*, ‘Database resources of the National Center for Biotechnology Information.’, *Nucleic Acids Res*, vol. 37, no. Database issue, pp. D5-15, Jan. 2009.
[36] ‘UniProt: a hub for protein information.’, *Nucleic Acids Res*, vol. 43, no. Database issue, pp. D204-212, Jan. 2015.
[37] M. Kanehisa and S. Goto, ‘KEGG: kyoto encyclopedia of genes and genomes.’, *Nucleic Acids Res*, vol. 28, no. 1, pp. 27–30, Jan. 2000.

[38] D. M. Kristensen, Y. I. Wolf, A. R. Mushegian, and E. V. Koonin, 'Computational methods for Gene Orthology inference.', *Brief Bioinform*, vol. 12, no. 5, pp. 379–391, Sep. 2011.
[39] W. M. Fitch, 'Distinguishing homologous from analogous proteins.', *Syst Zool*, vol. 19, no. 2, pp. 99–113, Jun. 1970.
[40] M. E. Peterson, F. Chen, J. G. Saven, D. S. Roos, P. C. Babbitt, and A. Sali, 'Evolutionary constraints on structural similarity in orthologs and paralogs.', *Protein Sci*, vol. 18, no. 6, pp. 1306–1315, Jun. 2009.
[41] E. V. Koonin, 'Orthologs, paralogs, and evolutionary genomics.', *Annu Rev Genet*, vol. 39, pp. 309–338, 2005.
[42] T. Gabaldon, C. Dessimoz, J. Huxley-Jones, A. J. Vilella, E. L. Sonnhammer, and S. Lewis, 'Joining forces in the quest for orthologs.', *Genome Biol*, vol. 10, no. 9, p. 403, 2009.
[43] 'OrthoXML'. [Online]. Available: http://www.orthoxml.org.
[44] A. Kuzniar, R. C. H. J. van Ham, S. Pongor, and J. A. M. Leunissen, 'The quest for orthologs: finding the corresponding gene across genomes.', *Trends Genet*, vol. 24, no. 11, pp. 539–551, Nov. 2008.
[45] A. J. Enright, S. Van Dongen, and C. A. Ouzounis, 'An efficient algorithm for large-scale detection of protein families.', *Nucleic Acids Res*, vol. 30, no. 7, pp. 1575–1584, Apr. 2002.
[46] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, 'Basic local alignment search tool', *Journal of Molecular Biology*, vol. 215, no. 3, pp. 403–410, 1990.
[47] L. Li, C. J. Stoeckert, and D. S. Roos, 'OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes', *Genome Res*, vol. 13, no. 9, pp. 2178–2189, Sep. 2003.
[48] J. Blom *et al.*, 'EDGAR: a software framework for the comparative analysis of prokaryotic genomes.', *BMC Bioinformatics*, vol. 10, p. 154, May 2009.
[49] J. Blom *et al.*, 'EDGAR 2.0: an enhanced software platform for comparative gene content analyses.', *Nucleic Acids Res*, vol. 44, no. W1, pp. W22-28, Jul. 2016.
[50] M. J. Brittnacher, C. Fong, H. S. Hayden, M. A. Jacobs, M. Radey, and L. Rohmer, 'PGAT: a multistrain analysis resource for microbial genomes.', *Bioinformatics*, vol. 27, no. 17, pp. 2429–2430, Sep. 2011.
[51] Y. Zhao, J. Wu, J. Yang, S. Sun, J. Xiao, and J. Yu, 'PGAP: pan-genomes analysis pipeline.', *Bioinformatics*, vol. 28, no. 3, pp. 416–418, Feb. 2012.
[52] D. E. Fouts, L. Brinkac, E. Beck, J. Inman, and G. Sutton, 'PanOCT: automated clustering of orthologs using conserved gene neighborhood for pan-genomic analysis of bacterial strains and closely related species.', *Nucleic Acids Res*, vol. 40, no. 22, p. e172, Dec. 2012.
[53] O. Lukjancenko, M. C. Thomsen, M. Voldby Larsen, and D. W. Ussery, 'PanFunPro: PAN-genome analysis based on FUNctional PROfiles', *F1000Research*, vol. 2, p. 265, Dec. 2013.
[54] M. N. Benedict, J. R. Henriksen, W. W. Metcalf, R. J. Whitaker, and N. D. Price, 'ITEP: an integrated toolkit for exploration of microbial pan-genomes.', *BMC Genomics*, vol. 15, p. 8, Jan. 2014.
[55] Y. Zhao *et al.*, 'PanGP: a tool for quickly analyzing bacterial pan-genome profile.', *Bioinformatics*, vol. 30, no. 9, pp. 1297–1299, May 2014.
[56] J. W. Sahl, J. G. Caporaso, D. A. Rasko, and P. Keim, 'The large-scale blast score ratio (LS-BSR) pipeline: a method to rapidly compare genetic content between bacterial genomes.', *PeerJ*, vol. 2, p. e332, 2014.

[57] A. J. Page *et al.*, 'Roary: rapid large-scale prokaryote pan genome analysis', *Bioinformatics*, vol. 31, no. 22, pp. 3691–3693, Nov. 2015.
[58] L. Snipen and K. H. Liland, 'micropan: an R-package for microbial pan-genomics.', *BMC Bioinformatics*, vol. 16, p. 79, Mar. 2015.
[59] S. Paul, A. Bhardwaj, S. K. Bag, E. V. Sokurenko, and S. Chattopadhyay, 'PanCoreGen - Profiling, detecting, annotating protein-coding genes in microbial genomes.', *Genomics*, vol. 106, no. 6, pp. 367–372, Dec. 2015.
[60] T. Pedersen, 'FindMyFriends: A Framework for Fast and Accurate Pangenome Analysis of Thousands of Diverse Genomes', *Nature Methods*, 2016.
[61] N. M. Chaudhari, V. K. Gupta, and C. Dutta, 'BPGA- an ultra-fast pan-genome analysis pipeline', *Scientific Reports*, vol. 6, p. 24373, Apr. 2016.
[62] 'Proteinortho: Detection of (Co-)orthologs in large-scale analysis | BMC Bioinformatics | Full Text'. [Online]. Available: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-12-124. [Accessed: 21-Dec-2018].
[63] A. P. Arkin *et al.*, 'The DOE Systems Biology Knowledgebase (KBase)', *bioRxiv*, p. 096354, Jan. 2016.
[64] R. Overbeek, M. Fonstein, M. D'Souza, G. D. Pusch, and N. Maltsev, 'The use of gene clusters to infer functional coupling.', *Proc Natl Acad Sci U S A*, vol. 96, no. 6, pp. 2896–2901, Mar. 1999.
[65] R. D. Page and M. A. Charleston, 'From gene to organismal phylogeny: reconciled trees and the gene tree/species tree problem.', *Mol Phylogenet Evol*, vol. 7, no. 2, pp. 231–240, Apr. 1997.
[66] A. L. Delcher, D. Harmon, S. Kasif, O. White, and S. L. Salzberg, 'Improved microbial gene identification with GLIMMER.', *Nucleic Acids Res*, vol. 27, no. 23, pp. 4636–4641, Dec. 1999.
[67] R. C. Edgar, 'MUSCLE: multiple sequence alignment with high accuracy and high throughput.', *Nucleic Acids Res*, vol. 32, no. 5, pp. 1792–1797, 2004.
[68] E. L. L. Sonnhammer and G. Ostlund, 'InParanoid 8: orthology analysis between 273 proteomes, mostly eukaryotic.', *Nucleic Acids Res*, vol. 43, no. Database issue, pp. D234-239, Jan. 2015.
[69] A. Alexeyenko, I. Tamas, G. Liu, and E. L. L. Sonnhammer, 'Automatic clustering of orthologs and inparalogs shared by multiple proteomes.', *Bioinformatics*, vol. 22, no. 14, pp. e9-15, Jul. 2006.
[70] D. M. Kristensen *et al.*, 'A low-polynomial algorithm for assembling clusters of orthologous groups from intergenomic symmetric best matches.', *Bioinformatics*, vol. 26, no. 12, pp. 1481–1487, Jun. 2010.
[71] R. D. Finn *et al.*, 'The Pfam protein families database: towards a more sustainable future.', *Nucleic Acids Res*, vol. 44, no. D1, pp. D279-285, Jan. 2016.
[72] D. Hyatt, G.-L. Chen, P. F. Locascio, M. L. Land, F. W. Larimer, and L. J. Hauser, 'Prodigal: prokaryotic gene recognition and translation initiation site identification.', *BMC Bioinformatics*, vol. 11, p. 119, Mar. 2010.
[73] D. H. Haft, J. D. Selengut, and O. White, 'The TIGRFAMs database of protein families.', *Nucleic Acids Res*, vol. 31, no. 1, pp. 371–373, Jan. 2003.
[74] J. Gough, K. Karplus, R. Hughey, and C. Chothia, 'Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure.', *J Mol Biol*, vol. 313, no. 4, pp. 903–919, Nov. 2001.
[75] L. Fu, B. Niu, Z. Zhu, S. Wu, and W. Li, 'CD-HIT: accelerated for clustering the next-generation sequencing data.', *Bioinformatics*, vol. 28, no. 23, pp. 3150–3152, Dec. 2012.

[76] J. Goecks, A. Nekrutenko, and J. Taylor, ‘Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences.’, *Genome Biol*, vol. 11, no. 8, p. R86, 2010.
[77] K. Hornik, ‘The Comprehensive R Archive Network’, *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 4, no. 4, pp. 394–398, 2012.
[78] R. C. Gentleman *et al.*, ‘Bioconductor: open software development for computational biology and bioinformatics.’, *Genome Biol*, vol. 5, no. 10, p. R80, 2004.
[79] R. C. Edgar, ‘Search and clustering orders of magnitude faster than BLAST.’, *Bioinformatics*, vol. 26, no. 19, pp. 2460–2461, Oct. 2010.
[80] R. L. Tatusov, M. Y. Galperin, D. A. Natale, and E. V. Koonin, ‘The COG database: a tool for genome-scale analysis of protein functions and evolution.’, *Nucleic Acids Res*, vol. 28, no. 1, pp. 33–36, Jan. 2000.
[81] N. A. O’Leary *et al.*, ‘Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation.’, *Nucleic Acids Res*, vol. 44, no. D1, pp. D733-745, Jan. 2016.
[82] D. M. Goodstein *et al.*, ‘Phytozome: a comparative platform for green plant genomics.’, *Nucleic Acids Res*, vol. 40, no. Database issue, pp. D1178-1186, Jan. 2012.
[83] T. Hubbard *et al.*, ‘The Ensembl genome database project.’, *Nucleic Acids Res*, vol. 30, no. 1, pp. 38–41, Jan. 2002.
[84] F. Perez and B. E. Granger, ‘IPython: A System for Interactive Scientific Computing’, *Computing in Science & Engineering*, vol. 9, no. 3, pp. 21–29, Jun. 2007.
[85] J. E. Stajich *et al.*, ‘The Bioperl toolkit: Perl modules for the life sciences.’, *Genome Res*, vol. 12, no. 10, pp. 1611–1618, Oct. 2002.
[86] J. M. Sturino and T. R. Klaenhammer, ‘Bacteriophage Defense Systems and Strategies for Lactic Acid Bacteria’, in *Advances in Applied Microbiology*, vol. 56, Academic Press, 2004, pp. 331–378.
[87] ‘Streptococcus thermophilus (ID 420) - Genome - NCBI, https://www.ncbi.nlm.nih.gov/ genome/?term=Streptococcus thermophilus[Organism]&cmd=DetailsSearch.’
[88] G. B. Golding, P. R. Marri, and W. Hao, ‘Gene Gain and Gene Loss in Streptococcus: Is It Driven by Habitat?’, *Molecular Biology and Evolution*, vol. 23, no. 12, pp. 2379–2391, Sep. 2006.
[89] X.-Y. Gao, X.-Y. Zhi, H.-W. Li, H.-P. Klenk, and W.-J. Li, ‘Comparative Genomics of the Bacterial Genus Streptococcus Illuminates Evolutionary Implications of Species Groups’, *PLOS ONE*, vol. 9, no. 6, p. e101229, Jun. 2014.
[90] K. Makarova *et al.*, ‘Comparative genomics of the lactic acid bacteria.’, *Proceedings of the National Academy of Sciences of the United States of America*, vol. 103, no. 42, pp. 15611–6, 2006.
[91] H. M. Ja *et al.*, ‘Comparative Genomic and Functional Analysis of 100 Lactobacillus rhamnosus Strains and Their Comparison with Strain GG’, vol. 9, no. 8, 2013.
[92] ‘www.chr-hansen.com.’
[93] P. A. Kitts *et al.*, ‘Assembly: a resource for assembled genomes at NCBI’, *Nucleic Acids Res*, vol. 44, no. D1, pp. D73–D80, Jan. 2016.
[94] T. Tatusova, M. DiCuccio, A. Badretdin, V. Chetvernin, S. Ciufo, and W. Li, *Prokaryotic Genome Annotation Pipeline*. National Center for Biotechnology Information (US), 2013.

[95] ‘Pedersen TL (2018). FindMyFriends: Microbial Comparative Genomics in R. R package version 1.12.0, https://github.com/thomasp85/FindMyFriends.’ .
[96] ‘Roary: rapid large-scale prokaryote pan genome analysis | Bioinformatics | Oxford Academic’. [Online]. Available: https://academic.oup.com/bioinformatics/article/31/22/3691/240757. [Accessed: 21-Dec-2018].
[97] ‘bp_genbank2gff3.pl - Genbank-gtgbrowse-friendly GFF3 - metacpan.org’. [Online]. Available: https://metacpan.org/pod/distribution/BioPerl/scripts/Bio-DB-GFF/bp_genbank2gff3.pl. [Accessed: 21-Dec-2018].
[98] Z. Zhou, J. Gu, Y.-Q. Li, and Y. Wang, ‘Genome plasticity and systems evolution in Streptomyces’, *BMC Bioinformatics*, vol. 13, no. Suppl 10, p. S8, Jun. 2012.
[99] T. Seemann, ‘Prokka: rapid prokaryotic genome annotation’, *Bioinformatics*, vol. 30, no. 14, pp. 2068–2069, Jul. 2014.
[100] Pederson TL, ‘FindMyFriends: Microbial Comparative Genomics in R. R package version 1.12.0, https://github.com/thomasp85/FindMyFriends.’ 2018.
[101] T. Seemann, ‘barrnap 0.7: Rapid Ribosomal RNA Prediction, https://github.com/tseemann/barrnap’, 2013.
[102] A. R. Quinlan, ‘BEDTools: the Swiss-army tool for genome feature analysis’, *Curr Protoc Bioinformatics*, vol. 47, pp. 11.12.1-11.12.34, Sep. 2014.
[103] E. Pruesse *et al.*, ‘SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB.’, *Nucleic Acids Res*, vol. 35, no. 21, pp. 7188–7196, 2007.
[104] M. Kanehisa, Y. Sato, and K. Morishima, ‘BlastKOALA and GhostKOALA: KEGG Tools for Functional Characterization of Genome and Metagenome Sequences.’, *J Mol Biol*, vol. 428, no. 4, pp. 726–731, Feb. 2016.
[105] B. Tjaden, ‘De novo assembly of bacterial transcriptomes from RNA-seq data’, *Genome Biol*, vol. 16, no. 1, 2015.
[106] T. L. Bailey, J. Johnson, C. E. Grant, and W. S. Noble, ‘The MEME Suite’, *Nucleic acids research*, vol. 43, no. W1, pp. W39–W49, Jul. 2015.
[107] A. de Jong, H. Pietersma, M. Cordes, O. P. Kuipers, and J. Kok, ‘PePPER: a webserver for prediction of prokaryote promoter elements and regulons’, *BMC genomics*, vol. 13, pp. 299–299, Jul. 2012.
[108] J. Slager, R. Aprianto, and J.-W. Veening, ‘Deep genome annotation of the opportunistic human pathogen Streptococcus pneumoniae D39’, *Nucleic Acids Res*, vol. 46, no. 19, pp. 9971–9989, Nov. 2018.
[109] J. Larsson, A. J. R. Godfrey, P. Gustafsson, D. H. E. (geometric algorithms), and E. H. (root solver code), *eulerr: Area-Proportional Euler and Venn Diagrams with Ellipses*. 2018.
[110] T. Lefébure and M. J. Stanhope, ‘Evolution of the core and pan-genome of Streptococcus: positive selection, recombination, and genome composition’, *Genome Biol*, vol. 8, no. 5, p. R71, 2007.
[111] A. Bolotin *et al.*, ‘Complete sequence and comparative genome analysis of the dairy bacterium Streptococcus thermophilus.’, *Nat Biotechnol*, vol. 22, no. 12, pp. 1554–1558, Dec. 2004.
[112] M. I. Pastink, B. Teusink, P. Hols, S. Visser, W. M. de Vos, and J. Hugenholtz, ‘Genome-Scale Model of Streptococcus thermophilus LMG18311 for Metabolic Comparison of Lactic Acid Bacteria’, *Appl. Environ. Microbiol.*, vol. 75, no. 11, pp. 3627–3633, Jun. 2009.

[113] N. O. Ali, J. Bignon, G. Rapoport, and M. Debarbouille, 'Regulation of the acetoin catabolic pathway is controlled by sigma L in Bacillus subtilis', *J. Bacteriol.*, vol. 183, no. 8, pp. 2497–2504, Apr. 2001.
[114] P. Hols *et al.*, 'New insights in the molecular biology and physiology of Streptococcus thermophilus revealed by comparative genomics', *FEMS Microbiol Rev*, vol. 29, no. 3, pp. 435–463, Aug. 2005.
[115] D. Limauro, A. Falciatore, A. L. Basso, G. Forlani, and M. De Felice, 'Proline biosynthesis in Streptococcus thermophilus: characterization of the proBA operon and its products.', *Microbiology*, vol. 142 (Pt 11), pp. 3275–3282, Nov. 1996.
[116] Y. Gezginc, I. Akyol, E. Kuley, and F. Özogul, 'Biogenic amines formation in Streptococcus thermophilus isolated from home-made natural yogurt', *Food Chemistry*, vol. 138, no. 1, pp. 655–662, May 2013.
[117] G. Y. Kovaleva and M. S. Gelfand, 'Transcriptional regulation of the methionine and cysteine transport and metabolism in streptococci', *FEMS Microbiol Lett*, vol. 276, no. 2, pp. 207–215, Nov. 2007.
[118] P. Burguière, S. Auger, M.-F. Hullo, A. Danchin, and I. Martin-Verstraete, 'Three Different Systems Participate in l-Cystine Uptake in Bacillus subtilis', *Journal of Bacteriology*, vol. 186, no. 15, pp. 4875–4884, Aug. 2004.
[119] C. Monnet, D. Mora, and G. Corrieu, 'Glutamine Synthesis Is Essential for Growth of Streptococcus thermophilus in Milk and Is Linked to Urea Catabolism', *Applied and Environmental Microbiology*, vol. 71, no. 6, pp. 3376–3378, Jun. 2005.
[120] C. Guimont, 'Change of free amino acids in M17 medium after growth of Streptococcus thermophilus and identification of a glutamine transport ATP-binding protein', *International Dairy Journal*, vol. 12, no. 9, pp. 729–736, Jan. 2002.
[121] T. Nohno, T. Saito, and J. Hong, 'Cloning and complete nucleotide sequence of the Escherichia coli glutamine permease operon (glnHPQ)', *Molec. Gen. Genet.*, vol. 205, no. 2, pp. 260–269, Nov. 1986.
[122] R. Gardan, C. Besset, A. Guillot, C. Gitton, and V. Monnet, 'The Oligopeptide Transport System Is Essential for the Development of Natural Competence in Streptococcus thermophilus Strain LMD-9', *J. Bacteriol.*, vol. 191, no. 14, p. 4647, Jul. 2009.
[123] I. A. D. Lessard and C. T. Walsh, 'VanX, a bacterial d-alanyl-d-alanine dipeptidase: Resistance, immunity, or survival function?', *PNAS*, vol. 96, no. 20, pp. 11028–11032, Sep. 1999.
[124] H. W. van Veen, 'Phosphate transport in prokaryotes: molecules, mediators and mechanisms.', *Antonie Van Leeuwenhoek*, vol. 72, no. 4, pp. 299–315, Nov. 1997.
[125] M. D. Fernandez-Espla, P. Garault, V. Monnet, and F. Rul, 'Streptococcus thermophilus Cell Wall-Anchored Proteinase: Release, Purification, and Biochemical and Genetic Characterization', *Applied and Environmental Microbiology*, vol. 66, no. 11, pp. 4772–4778, Nov. 2000.
[126] W. Galia, N. Jameh, C. Perrin, M. Genay, and A. Dary-Mourot, 'Acquisition of PrtS in Streptococcus thermophilus is not enough in certain strains to achieve rapid milk acidification', *Dairy Science & Technology*, vol. 96, no. 5, pp. 623–636, Sep. 2016.
[127] F. Chavagnat, J. Meyer, and M. G. Casey, 'Purification, characterisation, cloning and sequencing of the gene encoding oligopeptidase PepO from Streptococcus thermophilus A', *FEMS Microbiol Lett*, vol. 191, no. 1, pp. 79–85, Oct. 2000.
[128] 'Transcriptional Pattern of Genes Coding for the Proteolytic System of Lactococcus lactis and Evidence for Coordinated Regulation of Key Enzymes by

Peptide Supply | Journal of Bacteriology'. [Online]. Available: https://jb.asm.org/content/183/12/3614.long. [Accessed: 28-Jan-2019].
[129] V. Monnet, M. Nardi, A. Chopin, M. C. Chopin, and J. C. Gripon, 'Biochemical and genetic characterization of PepF, an oligopeptidase from Lactococcus lactis', *J. Biol. Chem.*, vol. 269, no. 51, pp. 32070–32076, Dec. 1994.
[130] F. Rul, J.-C. Gripon, and W. Monnet, 'St-PepA, a Streptococcus therrnophilus aminopeptidase with high specificity for acidic residues', p. 7, 2019.
[131] M. P. Chapot-Chartier, F. Rul, M. Nardi, and J. C. Gripon, 'Gene cloning and characterization of PepC, a cysteine aminopeptidase from Streptococcus thermophilus, with sequence similarity to the eucaryotic bleomycin hydrolase', *Eur. J. Biochem.*, vol. 224, no. 2, pp. 497–506, Sep. 1994.
[132] F. Chavagnat, M. G. Casey, and J. Meyer, 'Purification, Characterization, Gene Cloning, Sequencing, and Overexpression of Aminopeptidase N from Streptococcus thermophilus A', *Appl Environ Microbiol*, vol. 65, no. 7, pp. 3001–3007, Jul. 1999.
[133] M. D. Fernandez-Espla and F. Rul, 'PepS from Streptococcus thermophilus. A new member of the aminopeptidase T family of thermophilic bacteria', *Eur. J. Biochem.*, vol. 263, no. 2, pp. 502–510, Jul. 1999.
[134] I. Mierau *et al.*, 'Tripeptidase gene (pepT) of Lactococcus lactis: molecular cloning and nucleotide sequencing of pepT and construction of a chromosomal deletion mutant.', *Journal of Bacteriology*, vol. 176, no. 10, pp. 2854–2861, May 1994.
[135] K. F. Vongerichten, J. R. Klein, H. Matern, and R. Plapp, 'Cloning and nucleotide sequence analysis of pepV, a carnosinase gene from Lactobacillus delbrueckii subsp. lactis DSM 7290, and partial characterization of the enzyme', *Microbiology*, vol. 140, no. 10, pp. 2591–2600, 1994.
[136] J. Matos, M. Nardi, H. Kumura, and V. Monnet, 'Genetic Characterization of pepP, Which Encodes an Aminopeptidase P Whose Deficiency Does Not Affect Lactococcus lactis Growth in Milk, Unlike Deficiency of the X-Prolyl Dipeptidyl Aminopeptidase', *Appl Environ Microbiol*, vol. 64, no. 11, pp. 4591–4595, Nov. 1998.
[137] F. Morel, J. Frot-Coutaz, D. Aubel, R. Portalier, and D. Atlan, 'Characterization of a prolidase from Lactobacillus delbrueckii subsp. bulgaricus CNRZ 397 with an unusual regulation of biosynthesis', *Microbiology (Reading, Engl.)*, vol. 145 (Pt 2), pp. 437–446, Feb. 1999.
[138] R. Anastasiou, M. Papadelli, M. D. Georgalaki, G. Kalantzopoulos, and E. Tsakalidou, 'Cloning and sequencing of the gene encoding X-prolyl-dipeptidyl aminopeptidase (PepX) from Streptococcus thermophilus strain ACA-DC 4', *J. Appl. Microbiol.*, vol. 93, no. 1, pp. 52–59, 2002.
[139] M.-P. Chapot-Chartier, F. Rul, M. Nardi, and J.-C. Gripon, 'Gene Cloning and Characterization of PepC, a Cysteine Aminopeptidase from Streptococcus thermophilus, with sequence Similarity to the Eucaryotic Bleomycin Hydrolase', *European Journal of Biochemistry*, vol. 224, no. 2, pp. 497–506, 1994.
[140] B. W. Bosman, P. S. Tan, and W. N. Konings, 'Purification and Characterization of a Tripeptidase from Lactococcus lactis subsp. cremoris Wg2', *Appl. Environ. Microbiol.*, vol. 56, no. 6, pp. 1839–1843, Jun. 1990.
[141] E. Tsakalidou, R. Anastasiou, K. Papadimitriou, E. Manolopoulou, and G. Kalantzopoulos, 'Purification and characterisation of an intracellular X-prolyl-dipeptidyl aminopeptidase from Streptococcus thermophilus ACA-DC 4', *Journal of Biotechnology*, vol. 59, no. 3, pp. 203–211, Jan. 1998.

[142] Z. Hafeez *et al.*, 'New Insights into the Proteolytic System of Streptococcus thermophilus: Use of Isracidin To Characterize Cell-Associated Extracellular Peptidase Activities.', *J Agric Food Chem*, vol. 63, no. 34, pp. 7522–7531, Sep. 2015.
[143] E. G. Dudley, A. C. Husgen, W. He, and J. L. Steele, 'Sequencing, distribution, and inactivation of the dipeptidase A gene (pepDA) from Lactobacillus helveticus CNRZ32.', *Journal of Bacteriology*, vol. 178, no. 3, pp. 701–704, Feb. 1996.
[144] R. A. Lassy and C. G. Miller, 'Peptidase E, a peptidase specific for N-terminal aspartic dipeptides, is a serine hydrolase.', *J Bacteriol*, vol. 182, no. 9, pp. 2536–2543, May 2000.
[145] M. Bek-Thomsen, K. Poulsen, and M. Kilian, 'Occurrence and Evolution of the Paralogous Zinc Metalloproteases IgA1 Protease, ZmpB, ZmpC, and ZmpD in Streptococcus pneumoniae and Related Commensal Species', *mBio*, vol. 3, no. 5, pp. e00303-12, Nov. 2012.
[146] 'Detection of prokaryotic promoters from the genomic distribution of hexanucleotide pairs | BMC Bioinformatics | Full Text'. [Online]. Available: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-7-423. [Accessed: 01-Mar-2019].
[147] 'Predicting Shine–Dalgarno Sequence Locations Exposes Genome Annotation Errors'. [Online]. Available: https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.0020057. [Accessed: 01-Mar-2019].
[148] A. Al-Qahtani and K. Mensa-Wilmot, 'A 5′ Untranslated Region Which Directs Accurate and Robust Translation By Prokaryotic and Mammalian Ribosomes', *Nucleic Acids Res*, vol. 24, no. 6, pp. 1173–1174, Mar. 1996.
[149] C. Monnet, M. Nardi, P. Hols, M. Gulea, G. Corrieu, and V. Monnet, 'Regulation of branched-chain amino acid biosynthesis by alpha-acetolactate decarboxylase in Streptococcus thermophilus.', *Lett Appl Microbiol*, vol. 36, no. 6, pp. 399–405, 2003.
[150] K. I. Sørensen, M. Curic-Bawden, M. P. Junge, T. Janzen, and E. Johansen, 'Enhancing the Sweetness of Yoghurt through Metabolic Remodeling of Carbohydrate Metabolism in Streptococcus thermophilus and Lactobacillus delbrueckii subsp. bulgaricus', *Appl. Environ. Microbiol.*, vol. 82, no. 12, pp. 3683–3692, Jun. 2016.
[151] E. Neviani, G. Giraffa, A. Brizzi, and D. Carminati, 'Amino acid requirements and peptidase activities of Streptococcus salivarius subsp. thermophilus', *J. Appl. Bacteriol.*, vol. 79, no. 3, pp. 302–307, Sep. 1995.
[152] C. Letort and V. Juillard, 'Development of a minimal chemically-defined medium for the exponential growth of Streptococcus thermophilus', *J. Appl. Microbiol.*, vol. 91, no. 6, pp. 1023–1029, Dec. 2001.
[153] P. Garault, C. Letort, V. Juillard, and V. Monnet, 'Branched-chain amino acid biosynthesis is essential for optimal growth of Streptococcus thermophilus in milk', *Appl. Environ. Microbiol.*, vol. 66, no. 12, pp. 5128–5133, Dec. 2000.
[154] 'Genotyping of Streptococcus thermophilus strains isolated from traditional Egyptian dairy products by sequence analysis of the phosphoserine phosphatase (serB) gene with phenotypic characterizations of the strains - El-Sharoud - 2012 - Journal of Applied Microbiology - Wiley Online Library'. [Online]. Available: https://onlinelibrary.wiley.com/doi/full/10.1111/j.1365-2672.2011.05212.x. [Accessed: 19-Feb-2019].
[155] S. Arioli *et al.*, 'Aspartate Biosynthesis Is Essential for the Growth of Streptococcus thermophilus in Milk, and Aspartate Availability Modulates the Level

of Urease Activity', *Appl. Environ. Microbiol.*, vol. 73, no. 18, pp. 5789–5796, Sep. 2007.
[156] D. M. Linares, S. Arboleya, R. P. Ross, and C. Stanton, 'Complete Genome Sequence of the Gamma-Aminobutyric Acid-Producing Strain Streptococcus thermophilus APC151', *Genome Announc.*, vol. 5, no. 17, pp. e00205-17, Apr. 2017.
[157] K. M. Akpemado and P. A. Bracquart, 'Uptake of Branched-Chain Amino Acids by Streptococcus thermophilus', *Appl Environ Microbiol*, vol. 45, no. 1, pp. 136–140, Jan. 1983.
[158] G. K. Schuurman-Wolters and B. Poolman, 'Substrate specificity and ionic regulation of GlnPQ from Lactococcus lactis. An ATP-binding cassette transporter with four extracytoplasmic substrate-binding domains', *J. Biol. Chem.*, vol. 280, no. 25, pp. 23785–23790, Jun. 2005.
[159] P. Garault, D. L. Bars, C. Besset, and V. Monnet, 'Three Oligopeptide-binding Proteins Are Involved in the Oligopeptide Transport of Streptococcus thermophilus', *J. Biol. Chem.*, vol. 277, no. 1, pp. 32–39, Apr. 2002.
[160] Y. Sanz, F. C. Lanfermeijer, P. Renault, A. Bolotin, W. N. Konings, and B. Poolman, 'Genetic and functional characterization of dpp genes encoding a dipeptide transport system in Lactococcus lactis', *Archives of Microbiology*, vol. 175, no. 5, pp. 334–343, 2001.
[161] Y. Cui, T. Xu, X. Qu, T. Hu, X. Jiang, and C. Zhao, 'New Insights into Various Production Characteristics of Streptococcus thermophilus Strains', *Int J Mol Sci*, vol. 17, no. 10, Oct. 2016.
[162] W. Galia, C. Perrin, M. Genay, and A. Dary, 'Variability and molecular typing of Streptococcus thermophilus strains displaying different proteolytic and acidifying properties', *International Dairy Journal*, vol. 19, no. 2, pp. 89–95, Feb. 2009.
[163] I. Mierau *et al.*, 'Multiple-peptidase mutants of Lactococcus lactis are severely impaired in their ability to grow in milk.', *Journal of Bacteriology*, vol. 178, no. 10, pp. 2794–2803, May 1996.
[164] J. Yu *et al.*, 'Multilocus sequence typing of Streptococcus thermophilus from naturally fermented dairy foods in China and Mongolia', *BMC Microbiol*, vol. 15, Oct. 2015.
[165] S. Thomas, C. Besset, P. Courtin, and F. Rul, 'The role of aminopeptidase PepS in the growth of Streptococcus thermophilus is not restricted to nitrogen nutrition: Role of the streptococcal aminopeptidase PepS', *Journal of Applied Microbiology*, vol. 108, no. 1, pp. 148–157, Jan. 2010.
[166] 'The sequence of spacers between the consensus sequences modulates the strength of prokaryotic promoters. - PubMed - NCBI'. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/9435063. [Accessed: 01-Mar-2019].
[167] S. W. Seo *et al.*, 'Predictive design of mRNA translation initiation region to control prokaryotic translation efficiency', *Metabolic Engineering*, vol. 15, pp. 67–74, Jan. 2013.
[168] J. Ma, A. Campbell, and S. Karlin, 'Correlations between Shine-Dalgarno Sequences and Gene Features Such as Predicted Expression Levels and Operon Structures', *Journal of Bacteriology*, vol. 184, no. 20, pp. 5733–5745, Oct. 2002.
[169] A. Abolbaghaei, 'Shine-Dalgarno Anti-Shine-Dalgarno Sequence Interactions and Their Functional Role in Translational Efficiency of Bacteria and Archaea', Thesis, Université d'Ottawa / University of Ottawa, 2016.