Università degli Studi di Padova

Sede Amministrativa: Università degli Studi di Padova

Dipartimento di Agronomia Animali Alimenti Risorse Naturali e Ambiente

SCUOLA DI DOTTORATO DI RICERCA IN : SCIENZE ANIMALI E AGROALIMENTARI
INDIRIZZO: PRODUZIONI AGROALIMENTARI
CICLO: XXVIII

WHEY VALORISATION BY MICROBIAL FERMENTATION

Genome analysis of eight *Streptococcus thermophilus* strains and study on their possible applications.

Direttore della Scuola : Ch.ma Prof.ssa Viviana Corich
Coordinatore d'indirizzo: Ch.ma Prof.ssa Viviana Corich
Supervisore :Ch.mo Prof. Alessio Giacomini

Dottorando: Dott.ssa Veronica Vendramin

There is grandeur in this view of life, with its several powers, having been originally breathed into a few forms or into one; and that, whilst this planet has gone cycling on according to the fixed law of gravity, from so simple a beginning endless forms most beautiful and most wonderful have been, and are being, evolved.

Charles Darwin

Somehow I can't believe that there are any heights that can't be scaled by a man who knows the secrets of making dreams come true. This special secret, it seems to me, can be summarized in four Cs. They are curiosity, confidence, courage, and constancy, and the greatest of all is confidence. When you believe in a thing, believe in it all the way, implicitly and unquestionable.

Walt Disney

INDEX

Abstract

Streptococcus thermophilus is a thermophilic lactic acid bacterium (LAB) of major importance in the dairy industry. This species is widely used as starter culture to produce fermented dairy products. It has been awarded the status of GRAS (Generally Recognized as Safe) in the USA and a Qualified Presumption of Safety (QPS) status in the European Union, due to its long history of safe use in food production. Increasing the number of starter available to producers by discovering new strains with desirable characters is important not only for identifying new properties that may better suited the needs of the industrial raising demand but also to preserve natural biodiversity, which is diminishing with the spread and overuse of commercial starters.

The progresses in high-throughput ‘omics’ technologies (‘Foodomics’) allows the development of more rational approaches aimed to improve fermentation processes both for the traditional foods productions and for new functional food products . Nevertheless, to date only few steps were made toward the in-depht analysis of the pan-genome and transcriptional regulation in species of food interest.

In this study the whole genome sequencing of eight *S. thermophilus* strains isolated from typical cheese-making processes in four Italian regions was performed using the Illumina platform. Genomic data were compared with the already available information in order to study the level of genetic biodiversity present within the species. In addition, some technological properties were analysed both genetically and phenotypically to integrate the knowledges at these two levels.

The applicative part of the study regarded the study of the strains during growth on milk whey, both from physiological and genetic (gene expression) standpoints. Particular effort was dedicated to production of vitamin, in particular folates. The obtained results, reported in this thesis, are interesting both from a scientific and applicative point of view.

Riassunto

Streptococcus thermophilus appartiene ai batteri lattici (LAB) termofili ed è un microrganismo di primaria importanza nel settore caseario. Questa specie è largamente usata come starter nella produzione di prodotti caseari fermentati. Grazie alla sua lunga storia nella produzione di alimenti, gli è stato conferito lo stato di GRAS (*Generally Recognized as Safe*) negli Stati Uniti d'America e di QPS (*Qualified Presumption of Safety*) nell'Unione Europea. Aumentare il numero di ceppi starter disponibili per i produttori caseari, scoprendo ceppi autoctoni che posseggano caratteri tecnologicamente rilevanti, è importante non solo per identificare nuove proprietà che possano rispondere maggiormente alla crescente domanda, ma anche per preservare la naturale biodiversità che sta diminuendo con il diffondere e il sempre maggiore degli starter commerciali.

I progressi attuali nelle tecnologie *high throughput* ('*Foodomics*') permettono lo sviluppo di approcci razionali per l'ottimizzazione del processo fermentativo sia nella tradizionale funzionalità alimentare sia nella nuova potenzialità dei prodotti nutraceutici. Tuttavia, non molti passi sono stati fatti verso l'analisi dettagliata della pan-genomica e della regolazione trascrizionale nelle specie di interesse alimentare.

In questo studio, è stato portato a termine il sequenziamento completo del genoma di otto ceppi di *S. themophilus* isolati da processi di caseificazione tradizionali in varie località italiane. I dati genomici sono stati comparati con l'informazione disponibile nei database pubblici nel tentativo di studiare il livello di biodiversità genetica presente all'interno della specie. Inoltre, alcune proprietà tecnologicamente rilevanti sono state analizzate sia geneticamente sia fenotipicamente in modo da integrare le conoscenze a questi due livelli

La parte applicativa dello studio ha riguardato lo studio dei ceppi durante la crescita in siero di latte, sia dal punto di vista fenotipico che dell'espressione genica, con particolare attenzione alla produzione di vitamine e specificamente di folati. I risultati ottenuti hanno prodotto informazioni interessanti sia dal punto di vista scientifico che, in prospettiva, da quello applicativo.

Introduction

1.1 Whey

Whey is one of the cheese production by-products, it is of principal importance in the dairy industry due to the large volumes produced and its nutritional composition. Worldwide whey production is estimated at around 180 to 190×10^6 ton/year and about 40×10^6 tons/year of whey is produced in the European Union alone (1) of this amount only 50% is then further processed (2). Approximately half of worldwide cheese-whey (CW) produced is treated and transformed into various food and feed products (3). About half of the recovered whey is used directly, 30% as powdered CW, 15% as lactose and lactose-derived products while the rest is transformed into cheese whey- protein concentrates (WPC, (4). Also whey permeate, obtained from whey ultrafiltration, is an important product because in the past it has been used as fermentation medium.

Whey re-use, for isolating of its component or directly as it is recovered during cheese making, is universally recognized to be advantageous both for the environment and sustainable economy (3). However, it is still often treated as dairy wastewater. The whey disposal represents a serious problem for industries due to its high organic compound load, which can achieve, for example, a chemical oxygen demand (COD) of 100,000 $mgO_2 \, l^{-1}$(2).

Whey derives from the manufacture of cheese. This processing is based on casein coagulating by rennet, an industrial casein-clotting preparation containing chymosin or other coagulating enzymes. Rennet-induced coagulation of casein occurs at approximately pH 6.5. Whey originated from this process is referred as sweet whey. The second type of whey results from either the usage of fermentation processes or the addition of organic (or mineral) acids for casein coagulation, typical in fresh cheese production (3). This kind of whey is called acid whey. The main differences between the two types are in the acidity, mineral content and composition of the whey protein fraction: (i) the acid coagulation approach results in substantially increased acidity (final pH approximately 4.5), required for casein precipitation.(ii) Rennet clotting produces fragment α-casein molecules, namely glycomacropeptides (GMPs), which end up in whey. GMPs constitute more less 20% of sweet whey protein fraction and are absent in acid whey.

Besides technological processes, the source of milk coupled with the physiological state of the animal (which undergoes seasonal variations) determinate whey composition. Generally, dry basis bovine whey contains 70–80% of lactose, 9% of proteins, 8–20% of minerals and other minor components, for instance some hydrolyzed peptides of k-casein, lipids and bacteria (5). Comparing the most diffused milk composition, it is clear that ovine and caprine whey are rather different from bovine, mainly in their lactose, protein, and fat concentrations (6). Other technological steps involved in the milk pre-treatment may influence the whey composition.

Whey retains more or less 55% of total milk nutrients (7) and therefore it may be considered a valuable by-product with several applications in the food and pharmaceutical industries (8).

To valorise this by-product, two different methods are being considered: the first is based on the recovery of its valuable compounds such as proteins and lactose by industrial methods. The second is to apply fermentation processes to obtain added value end products, mainly: organic acids (e.g. lactic, acetic and propionic ones, (9), single cell proteins and oils, biopolymers (polyhydroxyalkanoates,7) and bacteriocins (11). Nevertheless some pilot attempts aimed to re-entering whey into the food market have been successful. A particularly promising sector is the soft-drink products.

1.1.1 Whey beverages

Processing of whey to beverages began in the 1970ies, and one of the oldest soft drinks produced is Rivella from Switzerland. Until today a large scale of different whey beverages have been developed, which are derived from sweet or acid whey by processing, like deproteinized whey, diluted or powdered whey and fermented whey.

There are also some alcoholic beverages, like whey beer or wine and other drinks with low alcohol content (less than 1.5%).

Several difficulties occur during this kind of processes. First of all, the high water content makes fresh whey very susceptible to microbial spoilage and therefore heat treatments are recommended, although whey proteins are thermo-sensitive and denature when temperature overtakes 60 °C (12). Thus a certain amount of whey proteins precipitate after the usual thermal treatment of whey (72 °C for 15-20 sec).

Relatively high content of minerals in the dry matter represent a problem for whey beverage production because these minerals are responsible for an undesirable salty-sour flavour. This problem is especially found with acid whey due to its high amount of lactic acid and mineral content (13)

Therefore, fresh whey processing has proved to be the most interesting solution. Many efforts have been made in the development of beverages, mainly with addition of fruit concentrates in order to yield acceptable the taste of these drinks (14).

In recent two decades numerous patents for whey beverages production with variable fruit dry matter amounts (5-20%) addition of concentrates have been registered. Several works have speculated about the combination of different fruits and percentage ratio for whey based beverages. It was recorded that citrus-flavoured and tropical fruit aromas consumer acceptance (15).The addition of berries have been supposed to have useful properties because of their enrichment in antioxidants which protect whey protein from oxidation (16). More recently, it has been suggested the addition of CO_2 to overcome the undesirable flavour and odour of fermented whey (17). One of the better options to improve beverages acceptance is the manufacturing of fermented whey drinks. For whey fermentations, mainly starter and probiotic cultures of lactic acid bacteria are used, while in case of alcoholic fermentations mostly Kluyveromyces yeast are preferred. There are some indications that fermentation of whey using milk culture produces a yogurt flavour similar to the one obtained when milk is fermented (18)

Due to the low total mass content (6-7%) of liquid whey, the mouth feel of fermented whey beverages is watery in comparison with fermented milk. This characteristic may promote consumer acceptance of dietetic beverages based on whey. Whey seems a very good raw material for simple production of dietetic beverages because lactose hydrolyzation ends up in glucose and galactose production, two monosaccharides with higher sweetness, better solubility and better absorption than their precursor. In this way, other sweeteners can be excluded (i.e. Hedelmatarha produced in Finland). Products like this lessen the energy value (104-113 kJ/100 ml) enlarging their consumer acceptance.

Another kind of whey beverage is milk-like beverages, which included mixing liquid or powder whey with skim or whole milk, buttermilk, some vegetable oils, hydrocolloids and emulsifiers. Thereby the milk part is added to improve drink stability and density.

Alcoholic whey products are divided into beverages with low alcohol content (≤ 1.5%), whey beer and whey wine (13). The former production includes deproteinizing whey, whey concentration, lactose fermentation usually by yeast strains *Kluyveromyces fragilis* and *Saccharomyces lactis* with addition of sucrose until reaching the desired alcohol content (0.5 - 1%), flavouring, if needed sweetening and then bottling. Thus, a certain part of lactose is transformed into lactic acid which gives a refreshing taste to the product, while the sugar rest ferments to alcohol. Famous drinks belonging to this category are 'Milone' obtained by fermentation with kefir culture, and sparkling 'whey champagne' (Serwovit) produced in Poland (19). Beers can be produced with or without malt. They can be fortified with minerals, vitamins or can contain hydrolysed starch. The presence of milk fats can cause loss of beer foam while undesirable odour and taste depend from low solubility of whey proteins or inability of yeasts to consume lactose. Whey wine contains a relatively low alcohol amount (10-11%) and is commonly flavoured with fruit aromas. Production of whey wine includes clearing, deproteinazation, lactose hydrolysis by ß-galactosidase, decanting and cooling, addition of yeasts and fermentation, decanting, aging, filtering and bottling (13).

Today, consumers pay a lot of attention to the relation between food and health. As a consequence, the market of health-promoting foods, called functional foods, has shown a notable improvement in the last few years. It is universally recognized that whey can occupy a new and important role in this scenario.

1.1.2 Whey components and their nutraceutical application

Historically, whey has not just been considered as a poor substrate with a little value used only for animal feed. Instead, historically its health effects was well known and during the Middle Ages, whey was esteemed not only as medicine and a skin balm: it was a common unguent component to cure various illnesses (7). Recent researches have witnessed an increased interest in whey protein products, to their nutritional role and to their active role on human health. Several works covering this topic have been published(20)

Milk is constituted by two major families of protein system: caseins and whey proteins. Caseins account for 80% (w/w) of the whole protein amount, and can easily be recovered from skim milk via precipitation or coagulation. Whey proteins are mostly globular molecules with a substantial content of a-helix motifs, in which the acidic and basic and hydrophobic and hydrophilic amino acids are distributed in a balanced way along the polypeptide chains. From the functional point of view, important differences were detected between these two components. Indeed, whey proteins have been found to be more effective on satiety than other proteins such as caseins, and they have been associated with more rapid gastric emptying along with the resulting increase in serum amino acids, which can stimulate the hunger control system. Also other constituents were recognized as important actors of satiety, namely lactose and calcium (20) Whilst in the last years researches focused mainly on beneficial effects of whey proteins and hydrolysed protein, bovine whey also contains an interesting amount of non-proteic bioactive components. Several works on whey benefits have led to the discovery of nonessential trophic factors can promote health or prevent disease, or both (21).

In general today, whey should be considered a dietary protein supplement which provides important compounds that show antimicrobial activity, immune modulation, and act in cardiovascular disease and osteoporosis prevention. In addition, whey has display antioxidant, antihypertensive effects, antitumoral, hypolipidemic, antiviral, antibacterial effect, and it was recorded as chelating agent (4). To date, many studies have well described the functions belonging to different fraction of whey.

1.1.3 Healthy value of whey fractions

Whey proteins have a high nutritional value, due to the high content of essential amino acids, especially the sulphur containing ones. (4). Those proteins include β-lactoglobulin (β-LG), α-lactalbumin (α-LA), immunoglobulins (IG), bovine serum albumin (BSA), bovine lactoferrin (BLF) and lactoperoxidase (LP), together with other minor components.

The actual concentrations of whey proteins depend on different effectors: the type of whey (acid or sweet), the source of milk (bovine, caprine or ovine), time of the year, type of feed, stage of lactation and the quality of processing. A detailed classification of nutraceutical properties assigned to each whey protein is available (21).

The overall biological features of whey proteins have been described. They are known to play an important role in particular by antimicrobial function and antiviral action before, during and after some virus infection, i.e. human herpes simplex virus type 1. Then, whey proteins act at a different level of immune response system, suppressing in vitro lymphocyte mitogenesis and alloantigen-induced proliferation, increasing production of Gluthatione (GSH) which is important in immune regulation and cancer prevention and reduce oxidant-induced cell death. Whey proteins are particularly important to overcome GSH-deficiency in seropositive and Alzheimer's disease patients. In addition, they improve immune and liver functions.

In detail, β-lactoglobulin (β-LGthose is the main component of bovine milk whey (about 58%), is known to be a source of amino acids essential during childhood due to its role in muscle growth and in cysteine storage. Cysteine is important to prevent body oxidant stress. β-LG participates in the milk lipid digestion the neonate, activating pregastric lipases (22) but it still to be the major allergen of cow's milk.

α-lactalbumin (α-LA) is one of the most studied proteins: it is an of most important component of milk which is retained in whey at the end point process of cheese-making, and it contributes significantly to its physical, biological and nutritional characteristics. In human whey, α-LA is a major protein (1.7 mg/ml) and it is interesting due to its high mineral content and balanced amino acid composition. It is particularly enriched in essential amino acids, in fact it has a high content of lysine and cysteine and interesting high content of tryptophan (5.9% of the total amino acid content). α-LA and its derivate can be used as food supplements of essential amino acids to improve and maintain the immune system, to reduce stress, to enhance opioid activity and antihypertensive action, to regulate cell growth and immunomodulation. This protein may possess also bactericidal or antitumor activity. The high content in tryptophan makes α-LA a nutraceutical itself; in particular it helps improving mood, sleep, and cognitive performance. As a general consideration, thanks to the high content in essential amino acids, α-LA is an invaluable supplement for infant formulas (23).

Immunoglobulins (IGs) constitute a complex group of elements which concentration in whey is around 0.7 g/l. Generally, they are agents of passive immunity inherited from newborns, but in cross-species acquisition, they are potentially involved in removing toxic or undesirable dietary factors. As an example, naturally occurring antibody in milk can bind cholesterol in the human digestive tract and prevents its absorption into the bloodstream. Indeed, immune milk was suggested to lower blood pressure (24). Definitely, they play a role in antimicrobial and antiviral properties: it is known that concentration of colostrum whey antibodies against a particular pathogen can be raised by immunizing cows with the pathogen or its antigens.

Bovine serum albumin (BSA) has the impressive property of reversibly binding various ligands. It is the principal carrier of free fatty acids and other lipids, such as flavour compounds. BSA has the biological function of inhibiting tumor growth acting on the modulation of autocrine growth regulatory factor activities. It can bind fatty acids free in the human body as well, and it shown antioxidant activities.

Lactoferrin (BLF) displays a wide range of biological functions, many of which are connected with its iron binding ability. It plays a quite important role in iron metabolism (25), it seems to affect intestinal iron absorption in infants, it enhances the local iron accumulation at inflammation sites and it can have a bacteriostatic effect thanks to its ability to bind free iron, essential for the growth of bacteria. Then, LF has a bactericidal effect against Gram-positive and Gram-negative bacteria, which is iron-independent, and it acts also as growth factor activator. It is probably the most valuable biomedical protein present in whey due to the various therapeutic properties it exhibits.

Lactoperoxidase (LP) is characterized by antimicrobial activity: it catalyses the thiocyanate oxydation and generates intermediate products with a broad spectrum of antimicrobial effects against bacteria, fungi and viruses(26). Hence, it has been used in foods, cosmetics and in clinical applications because of completely safety. It must be kept in mind that LP inhibits Gram-negative, catalase positive organisms, such as pseudomonas, coliforms, salmonellae and shigella. Gram-positive, catalase negative bacteria, such as streptococci and lactobacilli are generally inhibited but not killed by this protein.

Aside of the proteic component, other milk component are studied for their healthy properties. Lipids, such as sphingolipids and fatty acids (FA), contain several bioactive factors exhibiting antimicrobial activity against bacteria, viruses, and fungi and regulate diverse biological functions, even at low concentrations. For example, sphingolipids and triglycerides enriched in capric and lauric acids exhibit bactericidal effects and thus may protect against food-borne gastroenteritis. Free FA-enriched fractions of whey inhibit the germination of *C. albicans* up to 80% (27)

Free oligosaccharides are key components of human milk and play multiple roles in the health of the infants, by stimulating growth of selected beneficial bacteria in the gut, participating in the development of the brain, and exerting antipathogenic activity. Oligosaccharide concentration is lower in mature bovine milk, normally used for infant formula, compared with human milk (28). It was revealed that milks coming from different cow races and different stages of lactation show interesting statistical differences in oligosaccharide composition for both quality and quantity, and last but not the least for the presence of sialic acid which is essential for brain development and cognitive function (29) and which can be industrially recovered from whey (30).

Functionalities of other interesting molecules are still unclear. For example, increasing attention might be focused on gangliosides. These sialic-acid-containing glycosphingolipids, found ubiquitously in cell membranes of higher animals but absent in lower animal and plants, are known to exert prebiotic functions by enhancing bifidobacterial growth, contributing intestinal immune response, interfering with the adhesion of several pathogenic bacteria, and are fundamental for the correct neuronal development (31). These molecules were found in cow cheese whey, even if in lower concentration than in human milk, and their potentiality is undefined to date.

1.1.3.1 Bioactive properties of fermented whey

Seeking different ways to improve whey value and its suitability in the global market, fermentation represents probably the best chance. During the last decades, a series of whey beverages were developed as described in 1.1.1.

Fermentation allows the usage of microbial metabolism to break down whey components into smaller polymers, which can thus exhibit reactive residuals masked before. For example, it leads protein fragmentation into bioactive peptides that further take part in important body regulation functions, mainly antihypertensive and antithrombotic activities, opioid and ileum contracting activities, antimicrobial and immunomodulatory functions or which act in the nutrition system, for example regulating the digestive process (32).

On the contrary, microbial metabolism can also combine and rearrange compounds present in whey obtaining higher value new compounds. One of the most explored fields is the production of beneficial polysaccharides from lactic acid bacteria (LAB) fermentation (33). Briczinski and colleagues have verified that whey can support microbial metabolism during the construction of constitutive membrane polysaccharides, about which there have recently been wide studies to determine their physico-chemical and bioactive properties (34). A new branch of studies have started to speculate on the production of other interesting lactose-derived nutraceuticals (35) For example, it is known that some yeast or engineered *Escherichia coli* for gene encoding a thermostabl β-galactosidase (36) are able to produce lactulose, a oligosaccharide recognized as prebiotic. It is used especially in commercial infant formulas because it promotes the intestinal Bifidobacterium proliferation. Lactulose is mainly produced by chemical synthesis from lactose and fructose. In the future, lactulose probably will be produced from microbial fermentation by means of whey combined with some fructose enriched agricultural waste, probably vegetables

Several studies have described the antioxidant activity of whey proteins. Nonetheless, some peptides derived from food hydrolysis have been shown to have worthy antioxidative activities against the peroxidation of lipids or fatty acids (37).

Furthermore, fermented products can intertwine with human body systems in ways which are unpredictable from *in vitro* experiments. Fermented whey showed a high anti-inflammatory effect on mice affected by atopic dermatitis (38). Metabolites with bioactive functions were also detected in fermented whey (39). *Lactobacillus gasseri* and *Propionibacterium freudenreichii* whey fermented compounds stimulate the function of the innate immune system in vivo in a murine model.

Finally, whey was evaluated as suitable media for probiotic bacteria growth and survival. Several works have confirmed these beverages may be attractive for the growing market of probiotics, suiting to requirements of consumer acceptability and food safety (40). To confer their beneficial effect, probiotics need to be in high number in food and to survive gastric and intestinal environments. Several studies have focused on the protective effect of whey on Lactobacillus and Bifidobacterium strains, which are the main genera ascribing potential probiotics, in both increasing cell growth (41) and enhancing survival of gastric and duodenal digestion (42).

1.2 Lactic acid bacteria (LAB)

The production of fermented foods is based on the use of starter cultures, essentially lactic acid bacteria (LAB) that initiate rapid acidification of the raw material. This group has a long history of application in the production of fermented foods and beverages. They cause rapid acidification toward organic acids production, mainly lactic acid. Also, they produce a low amount of other interesting products, namely ethanol, aroma compounds, bacteriocins, exopolysaccharides and several enzymes important for end product of fermentation process. The min properties of those bacteria is their promoting the shelf life and microbial safety of the final product also improving texture and contribute to the pleasant sensory outcome (43). The earliest production of fermented foods was based on spontaneous fermentation due to the development of the microorganism naturally on foods. Then, spontaneous fermentation was optimized by back-slopping, namely inoculation of fresh raw material with a small amount of a previously fermented product. This practice allows, alongside the shortening of the fermentation, the reduction of fermentation failure. Today the strictly controlled large-scale production of fermented foods has become of first relevance in the food industry, hence autochthonous strain have lost their role. The main advantage of the direct addition of selected starter cultures is definitely to promise a high control over the fermentation process, and thus standardization of the final product. Disadvantages are the loss of the original uniqueness and the limited possibilities to identify new characteristics in the final product itself. Examples are the cases of wild strain abilities in antimicrobials production, naturally developed in reply to natural competition pressure (44), or intense flavour developed from non-starter lactic acid bacteria (NSLAB), which belong to a secondary flora arising during maturation of all the cheeses and more important in the traditional products ripening (45). Such findings underlined the importance of the Designation of Protected Origin (DPO) products, which are crucial from economical aspect since they contribute to the survival of small-scale fermentation plants. A recent trend is the isolation of wild-type strains from traditional products to be used as starter cultures in food fermentation (46).

LAB were recently explored also as functional starters. Functional starter cultures are starters that possess functional properties, meaning they contribute to food safety and offer nutritional or health advantages. Examples are LAB able to produce sweeteners or pleasant aroma which reduced synthetic compounds addition, useful enzymes, nutraceuticals, or LAB with health-promoting properties, called probiotic strains (43). LAB showing functional properties can be distinguished in four categories: properties assuring food preservation and safety, characters enhancing product appeal, qualities allowing `technological advantage', and characteristics leading beneficial effects on health.

Concerning food preservation, some LAB display antimicrobial activity by production of organic molecules. A general bacteriostatic activity is assigned to organic acid production (lactic acid, acetic acid, formic acid, phenyllactic acid, caproic acid, etc.). Also, specific antimicrobial activities were developed by in situ bacteriocin production, which generally have an activity spectrum restricted to related Gram-positive bacteria, even if to date many bacteriocins are known to work against a wide range of undesirable microorganisms (47), including fungi (48). Undesirable microorganisms control occurred also by acidification of the fermented product. Nevetheless, in some cases, a negative effect follows the principal acidification, called the post-acidification effect. Generally, in yogurt production, lactose is converted into lactic acid until a final pH of 4.2– 4.5 is achieved. During the storage, pH can decrease below 4.0. This undesirable post-acidification effect, ascribed principally to *Lb. delbrueckii* subsp. *bulgaricus*, leads to an acid and bitter taste which must be cover by addition of aromas. Lactose-negative mutants of *Lb. delbrueckii* subsp. *bulgaricus* enable the production of mild yogurts, since they can give their proto-cooperation only until they are growing in couple with actively lactose fermenting *S. thermophilus* cells (43). LAB strains could enhance texture pleasurable increasing the mouthfeel by adding polysaccharides to the final product, which improves viscosity and firmness of yogurt, or synthetizing thermostable amylasis, which have potential application in cereal fermentations (49). Also, they can modify the aroma of final product, for example acidifying the food resulting in lactic acid taste, or exerting proteolytic and lipolytic activities, or producing aromatic compounds from, for instance, amino acids after further bioconversion. Homofermentative LAB convert the available energy source almost completely into lactic acid via pyruvate. Pyruvate can lead to the generation of many other metabolites such as acetate, ethanol, diacetyl, and acetaldehyde. In this way, LAB produce volatile substances that contribute to the typical, pleasant flavour of certain fermented products (50). This effect helps to design new 'low-calorie products' avoiding the addition of synthetic aromas, along to the production of sugar alcohols which are used to replace traditional sugars (51).

Several bacterial enzymes were recognised playing a role in the human nutrient absorption(51). Besides rational selection of the LAB starter and co-cultures, another key mechanism which can act enhancing these enzymatic reactions is inducing autolysis of cells toward release of intracellular enzymes. To find the process conditions for optimized endogenous enzyme activity, the addition of exogenous enzymes (enzyme-modified cheese) and the increased bacterial autolysis actually represent possible solutions (52).

Several nutraceuticals of bacterial origin have been added to food. As an example, fermented milks can be produced with LAB starter strains that produce high amounts of low-calorie alcohols in place of sugars to reduce their content (53)

Less known is the fact that LAB are recognized to be able to produce vitamins, in particular ones ascribed to the B-group (54), which actually increase the value of fermented food, as well as enzymes which exert potential synergistic effects on digestion and alleviate symptoms of intestinal malabsorption (51). As described in 1.1.2.2., specific LAB stains lead toward the removal of toxic or antinutritive factors, such as lactose and galactose from fermented milks to prevent lactose intolerance and accumulation of galactose or some proteic compound, i.e. β-lactoglobulin.

1.2.1 Streptococcus thermophilus

Streptococcus thermophilus is a thermophilic lactic acid bacterium of major importance in dairy industry. It is phylogenetically closer to streptococcal species of the viridans group, which is divided into five subgroups (i) the mutans group, (ii) the anginosus group, (iii) the sanguinus group, (iv) the mitis group and (v) the salivarius group. This includes *S. salivarius*, *S. vestibularis* and *S. thermophilus.* The taxonomic status of *S. thermophilus* has been controversial: for some years it was classified as a *S. salivarius* subspecies (*Streptococcus salivarius* ssp. *thermophilus*). Only in 1991, it was conferred full species status on *S. thermophilus* (55). This species is widely used as a starter to produce fermented milk products (56) mainly because of it is fermenting lactose, behaviour that contributes to milk acidification. *S. thermophilus* has the status of GRAS

(Generally Recognized as Safe) and of QPS (Qualified Presumption of Safety) because of its long tradition of safety use for food processing Today *S. thermophilus* is considered the second most important species of industrial LAB after *Lactococcus lactis*, with a market value of around 40 billion US$; over 10^{21} live cells are ingested annually by humans (57). The augment of available starter biodiversity in food development by mean of autochthonous strains usage is universally recognized as important not only to identify novel and desirable characteristics, which are of increasing interest in reply to modern industrial demands, but also to preserve natural diversity which diminishes with the overuse of limited industrial starters (58). *S. thermophilus* widely occurs as commercial starter cultures as well as in natural milk or whey cultures traditionally used in the manufacture of several protected designation of origin (PDO) and artisanal cheeses. Indeed, it is commonly used as natural leaven for manufactory of Italian cheeses such as Fontina, Grana Padano, Mozzarella, Pecorino Toscano, and other cheeses (59). It is clear that Italian microbial population occupy a major position in the biodiversity preservation when is kept in account the amount of media or small cheese factories involved in the total annual Italian agricultural production (60). Nonetheless, only partially artisanal Italian biodiversity were explored from the genetic point of view (60, 62).

One of the main roles of *S. thermophilus* in milk fermentation is to provide rapid acidification. Marino et collegues (62) found that *S. thermophilus* is the predominant species in milk fermentation among LAB, concluding that acidification rate depends from several factors, such as proteolytic system, ureolytic activity and sugar metabolism.

The proteolytic system of *S. thermophilus* involves more than 20 proteolytic enzymes. It is composed from an extracellular cell- anchored protease (CEPs) which is responsible for casein hydrolysis, several transports for amino acids and peptides necessary for amino acids import, and a group of intracellular peptidases whose are of main importance for various essential metabolisms (63). *S. thermophilus* is the only LAB displaying a significant urease activity. It is known that different concentrations of urea lead to unpredictable rates of acidification during the fermentation processes. The buffering effect of ammonia, indeed, reduces the rate of pH decrease, extending the fermentation time. This phenomenon may affect the final texture and moisture of the fermented products. Moreover, a delay in acidification may increase the costs of fermentation process. In general, bacterial capacity to metabolise nitrogen is of the highest importance for the efficacy of the acidification process in milk. It was demonstrated that strain engineered for proteolysis have significant enhanced their acidification properties (64). The extracellular protease PrtS of *S. thermophilus* is a sortase cell-wall-anchored serine. PrtS is present in only a few strains of *S. thermophilus*. Proteinases is essential for the optimal growth of *S. thermophilus* when it is alone in milk while when co-cultivated, as example with *Lb. bulgaricus* which is normally proteinases positive, *S. thermophilus* is capable to grow optimally using peptides released by the other species (63). Generally, LAB are nutritionally exigent, meaning that they need an exogenous supply of amino acids for the growth. In *S. thermophilus* amino acid requirements may be satisfied by its biosynthetic capacities and, principally, by cooperation with other bacteria. However, genome analysis of *S. thermophilus* has revealed a high conservation of functional amino acid biosynthetic genes, which was supposed to be reflection of their synthesis importance for the growth in its natural environment. Nevertheless, phenotypic tests demonstrated that amino acids auxotrophies are strains-specific. In fact some strains are auxotrophic for different amino acids, such as cysteine, glutamine, histidine, methionine, isoleucine, leucine, tryptophan and valine, but other LAB are known to be more exigent (65). The amino acid catabolism leads to the production of the main characteristic flavour component of yogurt, acetaldehyde (66). In *S. thermophilus*, threonine can be directly converted into acetaldehyde and glycine by the threonine aldolase activity (57). In addition, sugars play a role in aroma formation. Acetaldehyde can be produced either from lactose or other sugars. An example of well-studied flavours derived from those compounds are diacetyl and propionic acid, responsible for the characteristic flavour of butter and the aroma of Maasdam and Swiss type cheeses (66). Within sugar metabolism, the main pathway is involved in the rapid conversion of lactose into lactate, but attention might be paid also on the production of other compounds that contribute to the final taste. Five different sugars are fermented by these bacteria: lactose, sucrose, glucose, galactose, and fructose (63) *S. thermophilus* is deeply adapted to grow using lactose as carbon source, while the last two sugars are fermented by a limited number of strains. Thus, fermentation end-products diversity is generally limited. Besides L-lactate, the main fermentation product, low amount of formate, acetoin, diacetyl, acetaldehyde, and acetate have been recognised [78–80].

Sugars are essential to obtain energy for metabolic functions and to recover building blocks further used in cell structures. Galactose belongs to the first case (67). Generally, LAB are able to metabolise only the glucose moiety of lactose, whilst a small percentage of strains can use also the other moiety. Al least four different profiles of galactose consumption were recorded in *S. thermophilus* (67). Even though two pathways were described in LAB (namely Tagatose-6P pathway and Leloir pathway) only the second one seems to be ubiquitous in *S. thermophilus* (68). Cell structure building involves sugars which are, as examples, extracellular polysaccharides. They can be present as capsular polysaccharides bound to the cell surface (CPS and LPS), or release into the growth medium (EPS) In *S.thermophilus* they consist in heterosaccharide polymers (63). Many strains of *S. thermophilus* synthesise free EPS, but some others are encapsulated.(63) Production of EPS was demonstrated conferring any obvious advantage to the growth or the bacterial survival in milk. Several studies on the *eps* genes suggested a very complex evolution of the system, which probably involved a chimeric structure originated both from the acquisition of *L. lactis* sequences by horizontal transfer and from exchanges within the *S. thermophilus* species (69).

1.2.2 Nutraceutical application of S. thermophilus

Besides being a good starter, *S. thermophilus* recently was speculated as potential probiotic species. Several probiotic characteristics (deconjugation of bile salts, hydrophobicity and β-galactosidase activity) and the resistance to biological barriers (gastric juice and bile salts) have been recently reported in some strains (57). Although *S. thermophilus* is known to be sensitive to gastric acidic conditions, it has shown to survive Gastro Intestinal (GI) transit adhering to intestinal epithelial cells. The ability of this bacterium to survive passaging through the upper GI tract was investigated in animal models. In that work, living cells were detected at a magnitude of 10^6–10^7 per gram of intestinal contents. Similarly, Brigidi and colleagues (70) identified and estimated the number of *S. thermophilus* residual cells in faecal samples of subjects fed with a pharmaceutical preparation by a culture-independent polymerase chain reaction (PCR) assay. Successive investigations on bacteria viability in human faeces were performed and several studies revealed that a great number of yogurt bacteria can survive human GI transit (57). Positive effects on human health attributed to *S. thermophilus* ingestion were mainly reduction of diarrhoea in young children, enterocolitis in premature neonates and inflammatory gut disease. Alongside the improvement of lactose digestion in lactose intolerant individuals, it has displayed to produce antioxidants, stimulate the gut immune system (55) and reduce the risk of certain cancers and ulcers. It was recognised acting against intestinal and vaginal infections. Further, other beneficial effects have been linked to either non-viable cells or to its cell components and enzymes(57).

Across the multiple possibilities for bacterial application in the development of nutraceuticals, food enrichment in healthy compounds is considered one of the most promising sectors.

Besides the previously described nutraceutical molecules which can raise the biological value of fermented products, see paragraph 1.2.1, *S. thermophilus* was well studied for its production of one particular component, the γ-aminobutyric acid (GABA, 71).

GABA is a non-proteinogenic amino acid possessing well-known physiological functions such as neurotransmission and hypotension induction by its diuretic and tranquilizer effects. It is commonly used in the treatment for sleeplessness, depression and autonomic disorders alongside the chronic alcohol-related symptoms treatment. It stimulates immune cells and, recently, GABA has been hypothesized as substitute of insulin, that attributes to its a putative diabetic prevent function (72).

Within healthy compounds, the vitamins are probably the main interesting. They are essential micronutrients working as precursors of various enzymes required for the vital biochemical reactions in all the living cells. Humans are unable to synthetize vitamins and they must be obtained from an exogenous source. The use of vitamin-producing microorganisms represents a more natural and consumer-friendly alternative for food fortification in comparison with the chemically synthesized ones. Natural fortification would allow the production of foods gathering at the same time high vitamin concentrations and less probabilities to cause undesirable side-effects. The B-complex vitamins include thiamine (B1), riboflavin (B2), niacin (B3), pyridoxine (B6), pantothenic acid (B5), biotin (B7 or H), folate (B9–B11 or M) and cobalamin (B12). B-group vitamins act in synergy to maintain the body's homeostasis playing major roles in metabolic processes, such as energy production (54). These vitamins, normally are found in many foods and are easily removed or destroyed during cooking and food processing, therefore the insufficient intake are common in many societies (73).

The overproduction of vitamins by LAB provides a very attractive approach to improve the nutritional composition of fermented foods (74). Folates (folic acid and its related compounds) are essential for the growth and the reproduction in all vertebrates. Folates have a preventative function against several disorders, mainly the development of neural tube defects during the fetal stage, coronary heart diseases, some types of cancer and neuropsychiatric disorders (75). Indeed, they are involved in various essential metabolic functions such as DNA replication, repairing and methylation, nucleotide synthesis, other vitamins and some amino acids synthesis and thus, their derived compounds, such as neurotransmitters (54). Similarly, folic acid is an essential cofactor in bacterial metabolism and hence many bacteria used in food fermentations possess the biosynthetic capability to produce it. *S. thermophilus* was found being responsible for about a six-fold increase of folate content in fermented milk. However, great differences have been observed in the vitamin production by different strains (54).

Bacterial participation in the improvement of vitamin B2 (76) and both in folate and riboflavin enhancing (77) was witnessed recently. As for folate, two vitamin sources are available for humans: dietary source and microbiota local production in the intestine. From the metabolic point of view, riboflavin is the precursor of flavin mononucleotide (FMN) and flavin adenine dinucleotide (FAD), both are coenzymes for a total of hundreds of enzymes, called flavoproteins. Interesting updates related to riboflavin enhancing role of *S. thermophilus* during food processing were reported in 2015 (78). The possibility of produce other vitamins by *S. thermophilus* fermentation was tested. A slight (but not statistically significant) increase was recorded in the thiamine and pyridoxine concentration occurred as result of soy fermentation with *S.thermophilus* strain ST5 and *Lactobacillus helveticus* R0052 or *Bifidobacterium longum* R0175. The authors concluded that their strains not enough efficient in this task and that the chosen medium could determine the outcomes (54).

The logical selected of new strains as starter cultures in the fermentation processes may help to achieve greatest results in the expression of these properties, keeping a natural and healthy product. Seeking new starter strains required a lot of time and resources if made by traditional methods but, today, the genomics and metabolomics applications permit the overcome of conventional labour limitations.

1.3 Species exploration by genetic approach

1.3.1 Early genetic analysis on S. thermophilus

Recently, population structures and genetic diversity within the streptococci salivarius group have been studied using Multilocus Sequence Typing (MLST), a typing method involving identification of nucleotide variations in housekeeping genes (Enright and Spratt, 1999). Eight housekeeping genes were amplified and sequenced in 63 strains of salivarius group. Analysis of the allelic profiles and the phylogenetic clustering of each locus confirmed *S. thermophilus'* status as a distinct species. Amplification of the 16S–23S spacer region, random amplified polymorphic DNA (RAPD)-PCR and sequencing of the 16S rRNA gene allow reliable molecular typing for species identification (Moschetti et al., 1998; Flint et al., 1999; Langa et al., 2003). To investigate new *S. thermophilus* strains, a molecular approach is necessary to distinguish strains and assess their genetic diversity. Strains can be rapidly identified using species-specific PCR based on the amplification of an intragenic fragment of the sodA gene (Poyart et al., 1998). Pulsed-field gel electrophoresis (PFGE) and RAPD-PCR typing methods have shown a high degree of variability within the *S. thermophilus* species (Colmin et al., 1991; Giraffa et al., 2001; Mora et al., 2002; Moschetti et al., 1998). In contrast, investigation of genetic diversity by comparing the lacSZ operon sequences of 29 S. thermophilus strains showed only slight variability (Ercolini et al., 2005). Genetic diversity within the species has also been studied using multilocus typing approaches (MLST) based on the nucleotide sequence variations of eight housekeeping genes, namely ilvC, pepO, pyrE, glcK, ddlA, thrS, dnaE and tkt, in 27 strains isolated from different dairy products. The sequence divergence within the *S. thermophilus* MLST loci proved to be very low, with an average of 0.19%, close to the 0.15% polymorphism observed when comparing the whole genomes of two *S. thermophilus* strains (Hols et al., 2005; Bolotin et al., 2004). Both these two techniques have confirmed that the degree of polymorphism in this population is low. A microarray assay was set up and tested on 2250 genes (79) to satisfy two purposes, the functional analysis and a general description of species biodiversity.

The analyses of 47 permitted identification of 1271 genes belonging to the core genome, 302 noncore genes considered ‘conserved genes’, between 27 and 58 noncore genes detected in one to five of explored genomes (for a total of 183 genes) which were considered ‘recently acquired genes’ and other 431 genes, called ‘variable genes’, carry to 6 to 44 considered genomes. Phylogenetic analysis on the same dataset has allowed identification of some different groups, in particular a significant diversification of groups containing protease-negative strains, and a total genetic diversity involving between 35 to 270 genes. Analysis of the combination of alleles at each locus revealed no significant cluster that would allow a correlation to be made between the allelic profile and the geographical origin or type of product from which the strains were isolated (55).

A combination of phenotypic traits and genotypic information has been used to investigate microbial diversity within *S. thermophilus* species. Although data collected on genetic diversity between strains were useful principally from the applicative point of view, to outline all the genome information might permit not only the discovery of more about the evolution of the species but also unexpected unique characters. Since 2004 (80), genomic analyses are moved to a more complete and global approach, which was made possible by the developing of new sequencing technologies.

1.3.2 Next-generation sequencing revolution

Since the early 1990’s, the sequencing approach has become every day more common throughout the life sciences field. Early, the sequencing technology was based on the Sanger biochemistry. With this approach, the method contemplate that randomly fragmented DNA is cloned into a high-copy-number plasmid, which is then used to transform *Escherichia coli,* in the shotgun de novo sequencing, or PCR amplification is carried out with primers that flank the target, for targeted re-sequencing. Either way, the output of both is an amplified template which is further processed by cycles of template denaturation, primer annealing and primer extension is performed (called ‘cycle sequencing’ reaction). The primers are oligonucleotides complementary to known sequence flanking the target region. Each cycle of primer extension is terminated by the incorporation of fluorescently labels, dideoxynucleotides (ddNTPs). Then, sequence is recorded by high-resolution electrophoretic separation of the single-stranded extension products in a capillary polymer gel. Laser excitation of fluorescent labels provides the Sanger sequencing ‘trace’. Software decodes these traces in DNA sequence, computing also an error probability. Tis technique has the principal limit of low level of parallelization, represented from the simultaneous electrophoresis in only 96 or 384 independent capillaries.

Over the past years, the incentive for developing entirely new strategies for DNA sequencing has emerged on at four levels. First, optimization through a significant reduction in the DNA sequencing cost. Second, a new way for the data processing emerged. The potential utility of short-read sequencing has been strengthened by the availability of whole genome assemblies for all major model organisms, because these effectively provide a reference against which short reads can be mapped. Third, a growing variety of molecular methods have been developed, whereby a broad range of biological phenomena can be assessed by high-throughput DNA sequencing (e.g., genetic variation, RNA expression, protein-DNA interactions and chromosome conformation). And fourth, the progresses in technology, including microscopy, surface chemistry, nucleotide biochemistry, polymerase engineering, computation and others, have made possible different strategies for DNA sequencing (81). To date, Next Generation Sequencing (NGS) technologies have a great impact both at industrial and at research level, allowing an increase of data production alongside the cost reduction. These new kinds of techniques allow the sequencing of thousands of genomes and they open entirely new areas of biological inquiry, including the characterization of ecological diversity and the discover of new species (82)

The second-generation sequencing had made various implementations of conventional sequencing, mainly basing their core process on cycle-arrays. This has permitted the multiplication of the sequencing strategies. The most important are 454 sequencing (Roche Applied Science; Basel, Switzerland), Ion Torrent PGM System (Thermo Fisher Scientific, Waltham, MA, USA), SMRT sequencing (Pacific Bioscience of California, Menlo Park, CA, USA), Solexa technology (Illumina, San Diego, CA, USA) and the SOLiD platform (Applied Biosystems; Foster City, CA, USA). Although these platforms are quite diverse in sequencing biochemistry as well as in how the array is generated, their work flows are similar. Library preparation is accomplished by random fragmentation of DNA, followed by in vitro ligation of common adaptor sequences. Alternative protocols can be used to generate jumping libraries of mate-paired tags with control-lable distance distributions. The generation of clonally clustered amplicons to serve as sequencing features can be achieved by several approaches, but in all these strategies PCR amplicons derived from given single library molecule are spatially clustered, either to a single location on a planar substrate. The sequencing process consists of alternating cycles of enzyme-driven biochemistry and imaging data acquisition. These platforms are developed on 'sequencing by synthesis' method, which involves serial primed template extensions, the enzyme driving the synthesis can be either a polymerase or a ligase. Data are acquired by recording the full array image at each cycle (e.g., of fluorescently label incorporated by the polymerase) and elaborated by software. Global advantages of second-generation strategies include (i) in vitro construction of a sequencing library, followed by in vitro clonal amplification to generate sequencing, (ii) array-based sequencing, which enables a much higher degree of parallelism than previous capillary-based methods, (iii) dramatically lower cost, because array features are immobilized to a micro-planar surface, which drops the effective reagent volume per feature (81). A recent work has compared the three most recent sequencing platform (Illumina, Ion Torrent and PacBio,) in order to supply a guide to underline the more suitable pipeline for choosing one method over others (83). The PacBio platform shows a mean mapped read length of 1336 bases, longer than what obtained with both Ion torrent and Illumina technologies. Illumina carries to the lowest error rates (0.4% against 1.78% of Ion Torrent and 13% of PacBio) and the lower false single nucleotide polymorphism (SNPs) calls, while Ion Torrent leads to detected the highest number of single locus variation (82% against 68-76% recorded by Illumina platforms, for PacBio the recognition was unclear). Analysis of a complex genome has revealed that only 65% of the genome was covered with a high quality value when Ion Torrent technology was applied while the other method achieved between 98-99% of the total genome length. Summarizing, it could be concluded that definitively PacBio technology is not suitable for the purpose of whole genome sequencing of small genomes, Ion Torrent may be considered a good tool for this goal even if it could sometimes overestimate strain differences while, on the contrary, Illumina technology probably underestimate strain variation but introduces the smallest number of error.

PROJECT OUTLINE

The aim of this work was to improve *Streptococcus thermophilus* biodiversity knowledges. This species represents one of the most important starter bacteria in Italian dairy production, and in this work is suggests a new perspective for its utilization toward whey valorisation in the field of nutraceutical applications.

Eight strains originated from different environments, industrial processing and geographical regions, were chosen to perform genomic and phenotypic characterisations. Strain genomes were sequenced by Solexa NGS platform choosing the pair-end approach. The comparison of whole genomes permitted to identify the major differences in genome size and functional categories. The phylo-geographical analyses allowed inferring phylogenetic relationship between new sequenced strains and the previously sequenced ones.

Phenotypical description of the main important characters for the species was performed on all the sequenced strains, coupled with bioinformatics analysis, in order to define genetic mechanisms regulating the phenotype expression and to increase the available genetic knowledges.

Strains properties suggested their utilization in the development of vitamin enriched foods. A comparison between their behaviour in synthetic media, commercial sweet whey and modified whey were carried on to explore their potentiality for this purpose.

Finally, to connect genomic information to the displayed behaviour in vitamin production, six fermentation conditions were compared by RNAseq analysis. This analysis allowed the identification of metabolic changes describing the phenotypic diversity occurred between different grow conditions and between the different strains.

Genomic analyses of *S.thermophilus*

1.4 *S. thermophilus* available genomes

The first whole genome sequencing of *S. thermophilus* was performed in 2004 (80). In that work, two strains (CNRZ1066 and LMG18311) isolated from yogurt in France and United Kingdom respectively, were processed by the random shotgun strategy and reassembled by multiplex PCR. Gene identification was carried discovering the coding DNA sequences (CDS) by means of Glimmer software and function attribution by BLAST best hits analysis. Sequencing revealed an average genome size of around 1.8 Mbp. Comparison allowed the identification of about 1900 CDS in each genome, almost 80% presumptively orthologous genes with other streptococci and about 90% of the coding sequences shared between the analysed strains. Differences between strains involved mainly extracellular and capsular polysaccharide biosynthesis (eps and cps) and the bacteriocin synthesis and immunity system, called ‘clustered regularly interspaced short palindromic repeats’ (CRISPR) and their associated proteins (Cas). This system guarantees the prokaryotic defence against bacteriophage infections and represents the bacterial adaptive immunity. In 2005, a new comparison among three strains (*S. themophilus* LMD9, isolated in the USA, was added to the previous ones) permitted to better describe the principal metabolisms for carbohydrates, proteins and stress management (63). Authors sustained that lateral genes transfer (LGT) events, involving species sharing the ecological niches with *S. thermophilus* (manily *L. delbrueckii* subsp. *bulgaricus*), shaped species genome more than the single polymorphism (SNP) and the natural selection. The comparison of the heterogeneous groups of lactic acid bacteria (84) at genome level evidenced that during its evolution probably occurred a loss of genetic information. Evolution reconstruction suggests that the common ancestor of Lactobacillales had at least 2100–2200 genes, highlighting a loss of 600–1200 genes (almost 25–30%) inherited from the bacilli ancestor and an acquisition of about 100 new genes. Many of the mapped changes seem being related to the transition in a nutritionally rich medium. Indeed, a high number of genes for the cofactors’ biosynthesis were lost while several peptidases were acquired, apparently via LGT. In addition, at least 25% of the LaCOGs (Lactobacillales-specific clusters of orthologous genes) probably derived from LGT and show local acceleration in evolutionary rate.

After the studies on species evolution, genomic analyses were mainly carried on industrial strains selected for their applicative properties. *S. thermophilus* JIM 8232 sequence was determined by using Sanger and SOLiD sequencing technologies. It carries 9 unique regions: three of them correspond to hypervariable regions, such as the *eps* operon and CRISPR sequences; three regions contain genes potentially involved in metabolism, such as oxidative stress, and three regions contain integrases. Perhaps the two major islands has been acquired by LTG, one of them contains several proteins potentially involve in the yellow pigment synthesis which is not frequently detected in this genus, with the exception of *Streptococcus agalactiae* (85). S*. thermophilus* ND03 was isolated from naturally fermented yak milk in China and sequenced using combined methods of 454 sequencing and Solexa pair-ends. This strain was selected because of its excellent processing properties, such as flavour formation, acidification rate and viscosity and water retaining properties. This strain showed 73 unique genes, some of them are components of six large insertion islands, and encode for transposase, glutamate decarboxylase, acetyltransferase, glycosyltransferase, polysaccharide biosynthesis protein, and the exopolysaccharide (EPS) biosynthesis cluster. This cluster was detected particular enriched and supposed being the main responsible for the strain technological properties (86). With the same technology and for the same purpose, that was to identify a consensus gene cluster for a the production of EPS in the fermented media, another Chinese strain was sequenced in 2012 (87), namely strain MN-ZLW-002 isolated from a traditional fermented dairy food. This strain possesses a complex system, involving 24 open reading frames (ORFs) in the sense and one ORF in the anti-sense orientation direction. Upstream of the *eps* cluster, two ORFs oriented in the antisense direction encode for transposases, indicating that the system has been acquired by LTG.

In 2014, *S. thermophilus* ASCC 1275, a common dairy starter isolated in Australia, was sequenced for its impressive ability in exopolysaccharides production, both capsular and ropy EPS. Analyses revealed that this strain contains several CRISPR/Cas loci, more than the other strains, having 4 CRISPR loci and 24 CRISPR-associated protein (cas) genes (88).

The principal strain studied on phage infection response was *S. thermophilus* SMQ-301 and it was sequenced by a paired strategy of Illumina and PacBio sequencing in 2015

(89). The comparison of two DT1 phage resistant strains, LMD 9 and SMQ-301, allowed recognising a conserved pathway in CRISP/cas organization, confirming that phage infection is the main mark for acquired resistance in these bacteria. The last sequenced strain was MN-BM-A02, which was chosen for its high acid-producing rate and low post-acidification ability. The second ability extends the shelf life of fermented products avoiding sensory attributes' changes during transport and storage, since its acidification capacity depends from the proteolytic activity of *L. delbreuckii* subsp. *bulgaricus*. Sequencing was performed by 454 sequencing and Solexa pair-end strategies(90).

Strain ID	Country	Isolation source	Online available	Genome size (Mbp)	No. putative protein	No. Chr and plasmid	Assembly no.
ASCC 1275	Australia	Dairy starter	2014	1.85	1700	1	ASM69888v1
CNRZ1066	France	Yogurt	2004	1.80	1915	1	ASM1184v1
JIM 8232	France	Milk	2011	1.93	2145	1	ASM25339v1
LMD-9	USA	Industrial starter	2006	1.86	1716	3	ASM1448v1
LMG 18311	UK	Yogurt	2004	1.80	1889	1	ASM1182v1
MN-BM-A02	China	Traditional dairy products	2015	1.85	1895	1	ASM100801v1
MN-ZLW-002	China	Yogurt Block	2012	1.85	1910	1	ASM26267v1
ND03	China	Traditional dairy products	2010	1.83	1919	1	ASM18287v1
SMQ-301	Canada	Mozzarella whey	2015	1.86	2037	1	ASM97166v1

Table 2. 1 Summary of available genomes online (May 2015). A brief description of assemblies statistics are presented coupled with the GenBank assembly number.

Whilst to date principally good starter strains were selected for genome studies, natural diversity preservation might be imperative to counteract its dramatic drop due to overuse and spread of industrial starters.Autochthonous strain analysis may lead to identify novel and desirable characteristics, which can better answer to the industrial production demands. In this chapter, an overview on whole genome sequenced *S. thermophilus* strains is proposed.

1.5 Strain selection

For this project, were selected eight strains of *S. thermophilus* obtained from two important national strain collections. Two strains (M17PTZA496 and MTH17CL396) were supplied from the 'Institut Agricole Regional' of Aosta culture collection while the others (1F8CT, TH982, TH985, TH1435 and TH1477) were provided from 'Veneto Agricoltura, Istituto per la Qualità e le Tecnologie Agroalimentari' culture collection. Strains were selected in order to magnify the diversity, therefore they come from four different Italian region (fig 2.1). That was combined with the effort to choose strains coming from different processing product (in particular milk, curd, whey and cheese) and from the origin of the milk (cow, goat and buffalo).

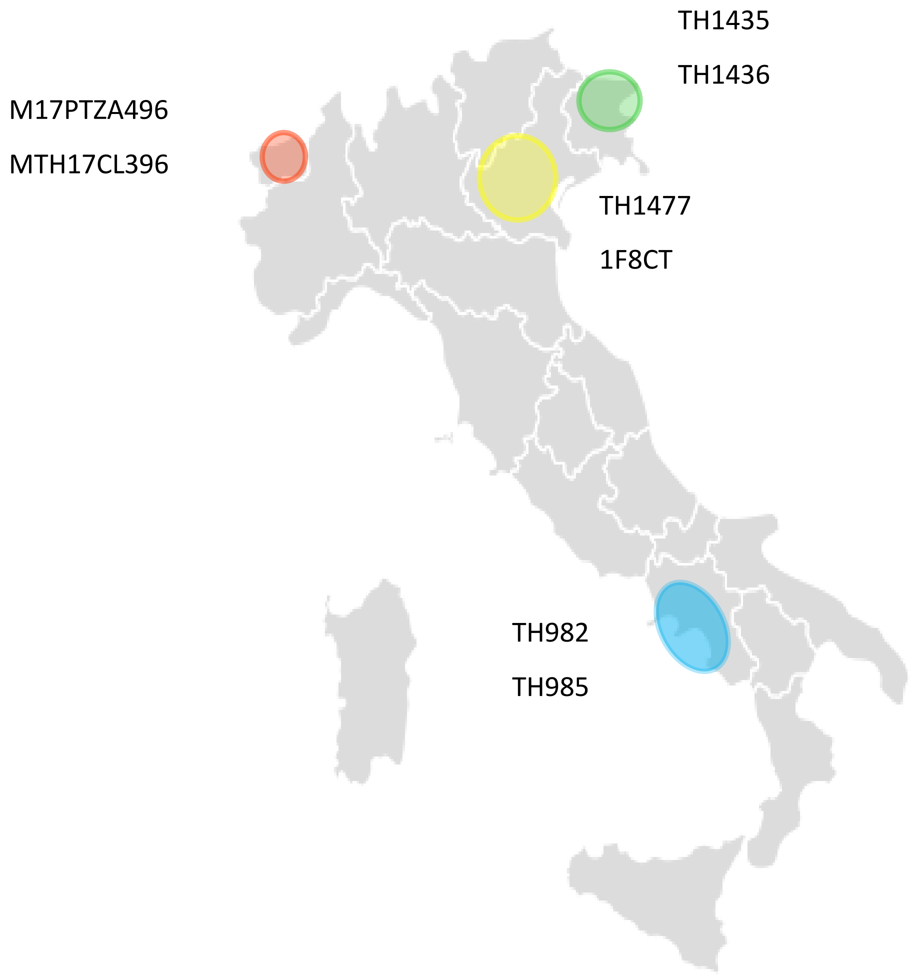


Figure 2. 1 Area of strains' isolation. Colours identify source and production process stage of collection: in red Fontina PDO cheese, in yellow raw cow milk, in green raw goat milk and in light blue buffalo mozzarella curd and whey.

1.6 Material and methods

1.6.1 Cells growth and DNA extraction

For the recovery of strains chromosomal DNA, an overnight culture was grown in 100 ml of M17L broth (Oxoid, Rodano, IT) at 37 °C. Cells were harvested by centrifugation, and the pellet was resuspended in 10 ml of TE buffer (10mMTris-hydrochloride, 1mMEDTA, pH7.5) containing lysozyme (0.5µg/ml). The suspension was incubated at 37 °C for 30 min. After that, the cells were collected and resuspended in 4ml of TE (50 mM Tris-hydrochloride, 20 mM EDTA, pH 8) containing 1% SDS. The suspension was gently mixed, and to ensure complete lysis, it was kept at 65 °C for 20 min; afterward, 2 ml of KAc 5M were added to the lysate, and the solution was maintained at 0 °C for 30 min. After centrifuging at 12,000g for 30min at 4°C, the supernatant was recovered, and the DNA was precipitated using two-volume of 96% (v/v) ethanol. The pellet was air-dried and resuspended in 1 ml of TE (50 mM Tris- hydrochloride, 20 mM EDTA, pH 8). RNA and proteins were removed, incubating the sample with RNAse A (0.2 µg/ml) for 30 min at 37 °C and with proteinase K (0.3 µg/ml) for 30 min at 56°C. DNA was precipitated with 1ml of absolute isopropanol and pelleted by centrifugation. Pellet was air-dried and resuspended in nuclease-freewater. Then, the DNA was purified using and of phenol-chloroform extraction. An iso-volume of phenol/chloroform/iso-amyl alcohol (in the ratio of 24:24:1) was added to each sample then the supernatant was recovered and washed with an equal volume of chloroform. Finally DNA was precipitated adding 130 µl of NaAc 5M (Sigma-Aldrich, Milano, Italy) and 800 µl of cold absolute ethanol and keeping samples at -20°C overnight. DNA was recovered by centrifuging for 30 minutes at 4°C. The supernatant was discharged and the pellet washed three times with 500µl of 75% (v/v) ethanol. Samples were air-dried and suspended in DNAse free water before lyophilized.

1.6.2 Extraction quality control

Genomic DNA quantification and purity were determined using NanoDrop (ThermoFisher Scientific, Waltham, MA, USA) and Qubit fluorometer (Life Technologies, San Diego, CA, USA). The former permits to evaluate the degree of contamination estimating by measuring the A260/A280 and A260/A230 absorbance ratios. For the latter, samples were prepared for dsDNA broad range assay, a fluorimetric assay which bind specifically double stand DNA, following the manufacturer instruction.

The absence of DNA degradation and its quantity were also visually estimated after agarose gel electrophoresis under UV illumination using Eurosafe (Euroclone, Milano, IT) as fluorescent dye. The signal for the DNA was compared to the intensity of a marker DNA with a known DNA concentration of the λ marker. Electrophoretic gel was prepared at a concentration of 0.8% agarose in TAE (Tris-Acetate-EDTA buffer) and DNA was let run for 30 min at 80 mV.

1.6.3 Sequencing and data quality control

Genomic DNA was sequenced at the Ramaciotti Centre for Gene Function Analysis (University of New South Wales, Sydney, NSW, AU) using the MiSeq Benchtop Sequencer (Illumina, San Diego, CA, USA). The paired-end reads of '250+ 250' bases strategy was chose. Libraries were produced using the 'Nextera XT' kit (Illumina, San Diego, CA, USA), and the DNA insert size was between ~350 bp and 1.5 kb. Sequence quality check, filtering, and conversion to FASTQ format were performed using the FASTX-Toolkit 0.0.13.

1.6.4 Assemblies and correction of genomes

High quality reads were used as input for the assembly. Abyss assembly software v.1.3.5 (91), 454 Newbler assembler (454 Life Sciences, Branford, CT, USA) and Velvet software v.1.2.10 (92) were compared for the *de novo* DNA sequence assembly. The first step of the *de novo* assembly process is a complete all-against-all reads comparison to identify the possible overlaps between fragments, while the second step is a contig optimization process that generates larger contigs (see Fig 2.2).

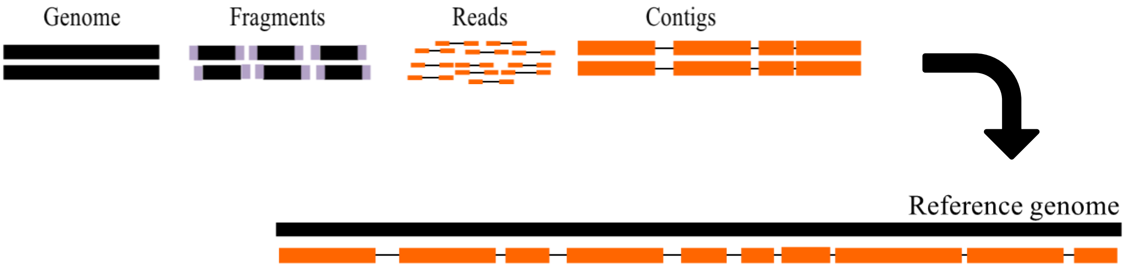


Figure 2. 2 Scheme summarising the pipeline used for whole DNA reconstruction by pair-ends sequencing strategy and contig reordering against reference.

A consensus sequence of the whole DNA was reconstructed assembling reads into contigs. Resulted contigs were aligned against the reference genome *S.thermophilus* CNRZ1066, assembly no. ASM1184v1, using MAUVE software v.2.3.1 and were finally reorganized (93). Mauve has been developed on the concept of LCBs (Locally Collinear Blocks) which represent homologous regions without rearrangements. LCBs allow the identification of conserved regions among the analysed genomes and highlight large scale rearrangements such as gain or loss, duplication and inversion of large segments. Small indels and SNPs do not interrupt the extension of the LCBs. The comparison between the resulting assemblies of each strain by visualization on Mauve allowed the manual finishing, creating a consensus genome in which long stretch of repeated sequenced, not identified bases and over represented sequences (mainly belonging to rRNA), which are difficult to place correctly, were considered contig breaks.

1.6.5 Gene annotation

Gene annotation was performed by RAST, a free automated annotation platform (94). This service produces high-quality assessments of gene functions and an initial metabolic reconstruction. The system is day-by-day improved from the increasing genetic information deposited in the online database. This platform refers to the SEED (95) as database and for the data organization. The SEED organizes the biological functions into a modular set of subsystems.

The protein function is based on attribution in families (*FIGfams*), which are further assigned into subsystems. Each subsystem is composed from a set of proteins sharing a globally similarity and, presumably, homology. All the members have a common function. The procedure takes as input protein sequences and returns a decision about whether or not these proteins might be assigned to a family, namely whether or not the proteins are globally similar to the members and are sharing their function. Two proteins are placed in the same family only if: (i) both have the same function and the similarity region shared covers over 70% of both the sequences. (ii) if they come from closely related genomes (i.e. genomes from two strains of the same species), the similarity is high (usually higher than 90% identity), and the context on the chromosome (i.e. the flanking genes) can easily be recognised as correspondents (fig. 2.3). The latter case allows placing the sequences in families even if their function is not determined. The dual directional approach (fig 2.3) reduces annotation errors propagation which frequently occurs. Annotation is the first step toward the genetic information organization into the SEED categorical structure.

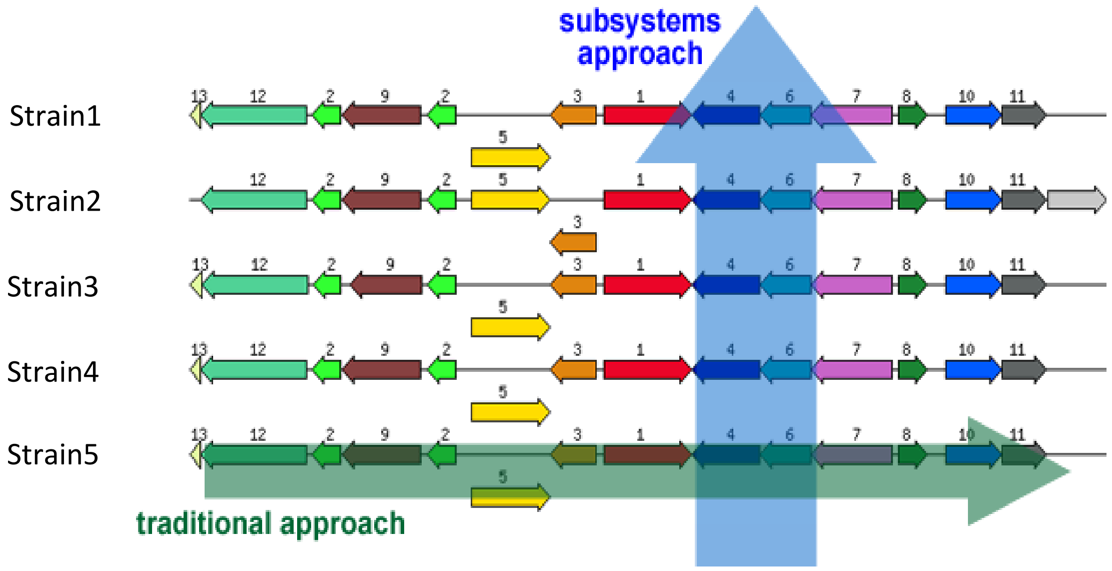


Figure 2. 3 The SEED functionality attribution compared with traditional automatic annotation service. Modified from http://www.theseed.org

1.6.6 Genome visualization

Artemis is a visualization tool that allows examination of the results placing the features in the sequence context (96). This tool serves several purposes and it represents the main tool for genome visualization coupled with Mauve. It allows the analysis of some interesting parameters on the raw sequences, i.e. GC content changes throughout genomes, as well as the specific features research, i.e. CDS and their relative positions on the genome. This software was used to simplify the analysis of LTG events due to its plasticity on different files type management and the possibility to identify CDS function using the BLAST best hits methods.

1.6.7 Phylogenetic analyses

Phylogenetic relationship between new sequenced strains and nine online available strains (tab. 2.1) has been computed integrating two approaches. The first method used was based on single-nucleotide variation profile. It was performed by the neighbour – joining method for draw phylogenetic tree of PHYLIP Package (97)

The program Neighbour takes as input a matrix of values representing SNPs distances between strains calculated considering all the possible couples of them. This program implements the Neighbour-Joining method and the UPGMA method of clustering and computes unrooted trees by successive clustering of lineages. Distances calculated by this program are then given as input to Drawtree that draw an unrooted tree diagram. The input matrix was computed by Mauve, this method exclude the un-paired regions from the calculation. This approach highlights genetic distances between strains because takes in account only the common portion of the genome. It is less suitable when different species are compared. To avoid under-estimations of phylogenetic distance, in the view of preliminary results obtained from the genome alignments, phylogenetic reconstruction was integrated with a second analysis based on a different approach. PhyloPhlAn is an automated method generating high-resolution microbial phylogenies by automatically detecting and combining of ubiquitously-conserved bacterial proteins. The phylogenetic tree reconstruction is based on the alignment of 400 conserved proteins (98). In the second analysis three outgroups were included, namely *S. macedonicus* 33MO (GeneBank assembly no. ASM71066v1), *S. pneumoniae* NT_110_58 (GeneBank assembly no. ASM81700v1) and *S. salivarus* JIM8777 (GeneBank assembly no. ASM25331v1).

1.6.8 Lateral gene transfer events

Acquisition of exogenous genetic material is frequent in prokaryotics. Traditionally compute system aim to identification of acquired genes are based the principle that the genomic islands (GIs) reflect the sequence composition of the donor genome, therefore the software were structured in order to record deviation, at various levels, from the host genome composition. It should be notice that this kind of prediction performs badly if the composition of the donor and the recipient genome are similar. Also, the composition of GIs could be similar to the host one if the LGT event has occurred in a relatively distant past, making LGT prediction more difficult. A different approach was studied for the prediction of putative LGT-derived region, based on variable compositional distributions. This approach does not require pre-existing annotation hence it can be considered not affected by annotation bias and errors. Alien Hunter v.1.7 (99) compute those compositional indices. The output could be uploading in Artemis to visualize the insertion regions into its original genomic context.

1.6.9 Identification of duplicated genes

Paralogous genes are generated via duplication events after the speciation. For practical purpose, paralogs are often defined as protein-coding sequences that have at least 30% sequence identity over more than 60% of their lengths (100). However, in several works identity threshold was changed toward a more conservative parameters those may allow the identification of strains specific events (101). In this study, analysis was performed clustering all the proteins encoded by the sequenced strains using CD-HIT platform (102). This software is based on a full parallelization and its core process can be simplified into two key steps: the checking procedure and the clustering procedure. It is pre-computed a 'word table' describing all the small sequences ('word', i.e. di- tripeptide and so on up to the complete sequence) paired with a score used to calculate the matching value between two sequences. Sequences are analysed in order to attribute the smaller number of word sufficient to describe their identity and then a ' word table' is computed. Given a word table, the former procedure checks each of the sequences against the table and determines if it is a redundant sequence. The clustering procedure makes a final determination of the status of a sequence, and which one is the representative sequence serving to update the original word table whether need. This approach was previous used for the same goal (103). Two different analyses were performed clustering proteins at 90 and 99% of sequence identity. Clustering at 90% of identity was retained the most suitable. The number of cluster with more than one

M17PTZA496 sequences were recorded and evaluated by the comparison with the other strains, in order to individuate uniqueness of M17PTZA496 genome respect its species background.

1.6.10 Gene content analysis

Analysis on gene content allows identifying strain peculiarities expressed as gene abundance variation in specific functional class. As described above, in this bacterium some gene classes seem to be more subjected to variation. The RAST annotation was used to analyse the gene content of each strain. Also, nine complete genomes available in GenBank (tab. 2.1) were used. Their raw sequences (fasta format) were downloaded from database and annottated by RAST. The ’Subsystem feature count’ lists were downloaded from the online service and gene functions organised by functional class. The resulting lists were analysed by MeV: MultiExperiment Viewer v.4.9.0 to elaborate the Hierarchical Clustering (HCL) of genome data (104). Aim of the hierarchical clustering is to build a dendrogram that enclose all the elements into a single tree. For any set of genes, a similarity matrix is computed, which contains similarity scores for all the pair genes. Resulting matrix is examined further to identify the highest value, representing the most similar set of genes. A node is created joining these two genomes, and a new profile is computed for the node by averaging observation for the joined elements. The similarity matrix is updated with the newly-formed node replacing the previous elements, and the process is repeated until only a single element remains. An important step is the selection of the method for measuring the distance between two nodes, which determines how the similarities are calculated. The software allows calculating the distance with several approaches, it has been chosen the Euclidean distance. It must be also set a parameter called ‘Linkage Method’ that defining the process used for determining cluster-to-cluster distances during the tree construction. It has been utilized the ‘average linkage’ method. Results of analyses were visualized as heat-map which represents the abundance of sequences (ordinated in rows) in each genome (described by the columns).

1.6.11 Variable genes attribution to categories

Variable genes identification among several strains belonging to the same species can facilitate the understanding of species ecological diversification and evolution. Annotation lists were used as dataset to manage the analysis of common and non-common species features by R costumer scripts (105). First, a non-redundant character lists were prepared enclosing the common function detected in all the analysed genomes. Then, each strain annotation list was compared with the consensus genome, excluding repeated functions. Variable genes were categorized into the SEED system and the results graphed as histogram. Both online available (tab 2.1) and new studied genome were analysed.

1.7 Results and Discussion

1.7.1 Extraction quantification and quality

The DNA extraction quality was evaluated by detection of contaminants and integrity of extracted genomic DNA. Absorbance ratio results displayed absence of contaminants, namely EDTA, carbohydrates and phenolic compounds (expressed as absorbance at the specific wavelength of 230 nm) and proteins (expressed as absorbance value at wavelength of 280 nm). All the samples indeed achieved OD 260/230 ratio >1.8 and an OD 260/280 ratio >2.0 as required for the library preparation protocol. Integrity of extracted DNA was also evaluated by visualization on agarose gel (fig 2.4)

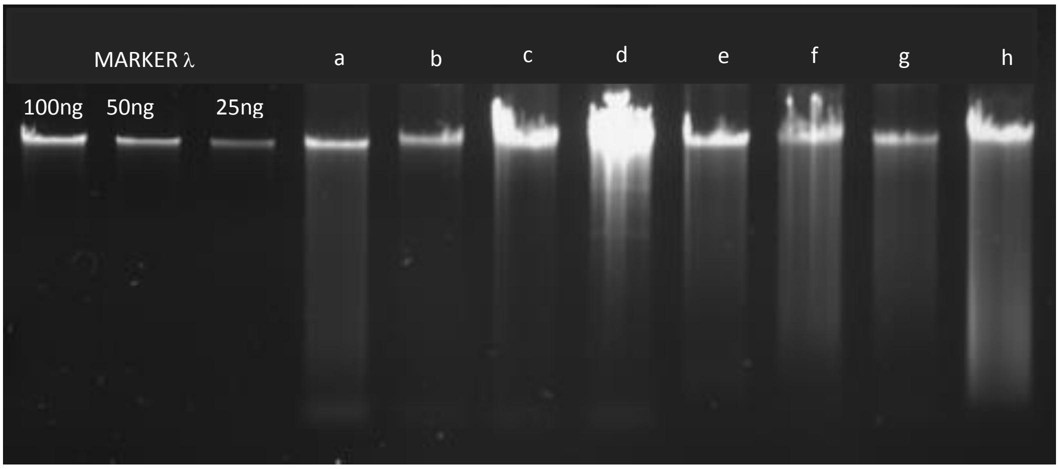


Figure 2. 4 Visualization on agarose gel of samples for sequencing. On the left, three markers λ at different DNA concentration. **a: TH1436, b:MTH17CL396, c:TH1435, d:M17PTZA986, e:TH1477, f: 1F8CT, g:TH985, h:TH982**

1.7.2 Sequencing and assemblies results

A common phenomenon is the reduction of quality in the terminal part of reads. The paired-end strategy limits the effect of this event because a second, generally more than one, read covers the same genomic region. The abundance of sequences covering the same genome fragment is an important parameter for evaluating the quality of the process. A good results is to achieve a number of reads sufficient to cover all the genome a high number of times, thus providing high attendance of the assembly results. The performed sequencing, with a total of about 2.0X10^7 paired reads each samples, and a yield of 96.15 % of paired reads after the first quality filtering, permitted to guarantee an average coverage of 338 folds.

Strain ID	Genome size (Mbp)	Final coverage (in fold)	No. final contig	No. large contig[¥]	Large contig (%)	Used reads (%)
1F8CT	1.75	261	60	46	0.77	0.80
M17PTZA496	2.07	107	72	71	0.99	0.80
MTH17CL396	1.82	407	49	33	0.67	0.84
TH982	1.73	218	52	30	0.58	0.73
TH985	1.84	183	84	80	0.95	0.79
TH1435	1.75	134	36	30	0.83	0.86
TH1436	1.78	159	27	25	0.93	0.85
TH1477	1.88	142	56	52	0.93	0.78

Table 2. 2 Summary of sequencing and assembly statistics. ¥Contigs>1000bp were referred as 'large'.

Assembly results revealed that almost all the genome sizes are comparable to what already reported in literature (Tab. 2.1). The only notable exception is represented by the strain M17PTZA496 which carries almost 15% more genetic information. In all the samples coverage reached at least 150 fold values and it indicate the good suitability of strategy chose. The final number of contigs is reduced, and almost all the genome is generally assembler into a few large contigs. The unique exception is TH982 which show a major rate of fragmentation probably due to many stretches of repeated sequence. Also in this case more than 50% of genome is built on large contigs.

Bacterial genomes exhibit a wide range of compositional diversity, most represented by variation in genome GC content (106). Base composition of genomic sequences varies widely, both across species and along chromosomes. For instance, the genomic GC content of cellular organisms ranges from 13% to about 75%, with vast intra-genomic heterogeneity (107). The nature of the biological processes underlying these differences has been long debated and two polarizing interpretations have been advanced, one proposing that GC content is driven by genome-specific mutational biases (the mutational hypothesis), and one that it reflects different selective processes in different organisms (the selectionist hypothesis (106). Nevertheless, similar GC content is observed in closed related species, supporting the idea that it should be dependent principally from environmental condition. All the new sequenced strains shown a GC content ranging between 38.8% and 39.1% (namely 39.1 in 1F8CT, 38.8 in M17PTZA496 and 38.9 in all the others) according with what previously found (80,85, 86,89,87,90)

1.7.3 Alignment visualization

Strain genomes were aligned using the program Mauve and the alignment was analysed using the viewer tool. From manual inspection of the alignments we identified acquired or lost sequences typical of a specific strain and inspected the degree of conservation throughout the genomes. From fig 2.5, it is evident that not only M17PTZA496 is characterised by unique regions, which are represented as white areas, but also MTH17CL396 and TH1477 carry portions with low level of identity. Mauve allows the identification of large rearrangements but it is still be complicated to recognise small modifications which probably could explain the presence of low conserved regions. It is notable, even though, that contigs (bound from red lines) were easily ordinated into eleven local colinear blocks (LCB, distinguished by different colours). When it wasn't possible, the rest of contigs were hold at the end of assembled genome.

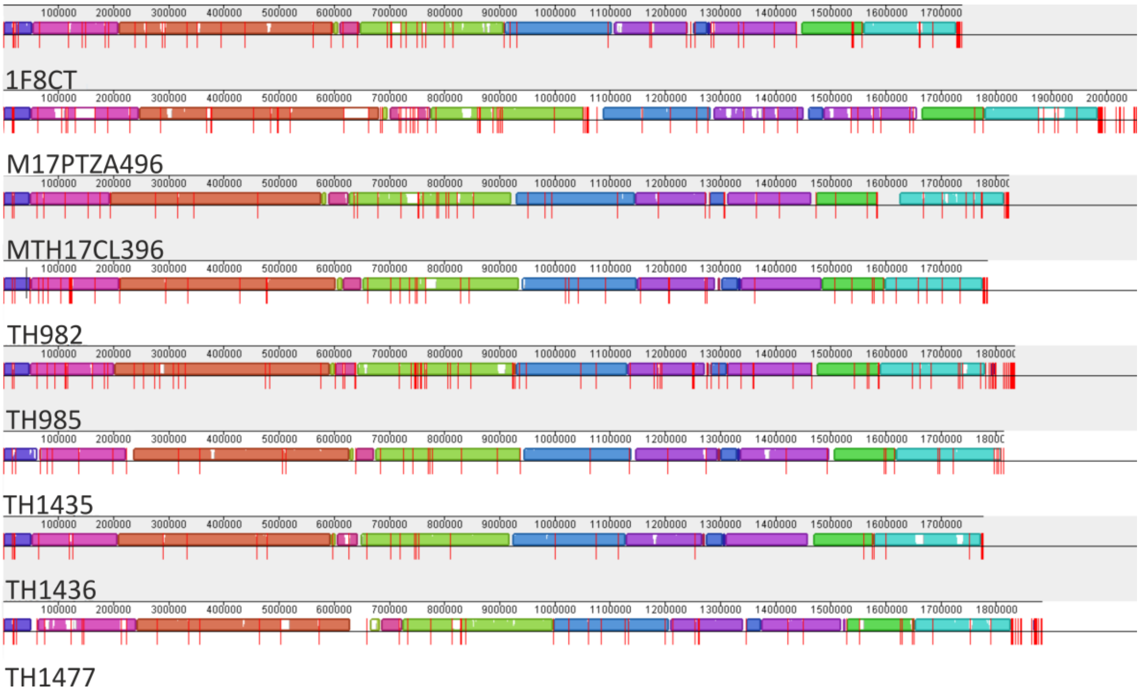


Figure 2. 5 Visualization with the viewer tool of Mauve of the eight genomes alignments. Red vertical lines separate contig. Local colinear blocks identified by the software are represented in different colours. The white spots within LCB individuate low conserved region while gaps between LCBs reflect strains unique regions.

1.7.4 Annotation statistics

Protein-coding open reading frames (ORFs) were identified and annotated using the RAST annotation server. Annotation consists in the identification of coding sequences over the genome and their matching against known sequences to find the higher similarity. The reference sequences support the functional recognition of the new discovered sequences. In case of strain 1F8CT, a total of 1.864 ORFs were speculated, and 51 RNA genes (involving mainly rRNA and tRNA) were identified. Strain M17PTZA496 genome was predicted to contain 2.221 ORFs and an unexpected higher number of RNAs, namely 89, while MTH17CL396 genome was predicted to contain 1.935 ORFs and 56 RNAs genes. Strains TH982, TH985, TH1435, TH1436 and TH1477 were detected to carry 1.924, 1.952, 1.925, 1.899 and 1.986 ORFS and 47, 69, 47, 48 and 55, RNA genes, respectively.

A focused analysis of CRISPRs/cas system permitted to detect a lower content in genes assigned to M17PTZA496 (see tab. 2.3). This system is particularly interesting because it is coding for the acquired phage resistance in bacteria and it was demonstrated that CRISPR–Cas system of archaea and bacteria mechanism showed high similarity to the RNA interference (RNAi) mechanisms of eukaryotes.

Strain ID	Cas$^{¥}$ family	Cas6	Cas3p	Cse2	Csn* family	Csm^{+} family	RAMP$^{\#}$ proteins
SMQ-301	7	1	1	-	3	3	2
1F8CT	7	1	-	-	3	4	2
M17PTZA496	3	-	-	-	1	-	-
MTH17CL396	7	1	1	-	3	3	2
TH982	7	1	-	-	3	3	2
TH985	8	1	-	1	3	3	2
TH1435	7	1	-	-	3	3	2
TH1436	7	1	1	-	3	3	2
TH1477	5	1	-	-	1	3	2

Table 2. 3 CRISPR/cas subsystem. Protein classification follows literature, sequenced strain systems are compared against *S. thermophilus* SMQ-301, strain which was long studies for phage resistance. *Cas*$^{¥}$: genes were assigned to *Cas1, Cas2* and *Cas7* family, *Csn**: genes belonged to *Csn1* and *Csn2* families, *Csm*$^{+}$: sequences were attributed to Csm1, Csm2 and Csm5 families, RAMP$^{\#}$: sequences were part of *Csm3* and *Csm4* families.

CRISPR/cas system classification was recently reviewed (108). Even if assignation of phage resistance proteins into few categories seems to be not easy because of their complex interaction within systems and their high rate of evolution, it was argued that Cas1 and Cas2 be present in all CRISPR–Cas systems that are predicted to be active, and are thought to be the information-processing subsystem that is involved in spacer integration during the adaptation stage. *Cse* sequences were detected firstly in *E. coli* and were considered unique of the species for long time but to date are considered orthologous to some Cas proteins. In particular Cse2 is a small α-helical protein involved in the Type I CRISPR–Cas systems. Csm proteins are specific of type II and type III systems representing the signature genes of the system. In this studies only Csm belonging to subtype III-A were found. Moreover, in the defence cascade, RAMP proteins with RNA endonuclease activity have been identified as the main enzyme catalysing the long spacer–repeat-containing transcript processing into mature crRNAs. In the studies strain, two RAMPs protein were detected in almost all the strains (as well as in the chosen reference) coupled with Cas6 which acts as CRISPR repeat RNA endoribonuclease. Interestingly, *Cas3p*, gene encoding for CRISPR-associated helicase, is present only in two of the sequenced strains (MTH17CL396 and TH1436) and in the genome reference.

1.7.5 Phylogenetic reconstruction

To clarify how far geographical origin can explain species diversity and in which amount it can contribute to genomic diversification, phylogenetic reconstruction were performed involving all the public *S. thermophilus* genomes known to date (tab 2.1).

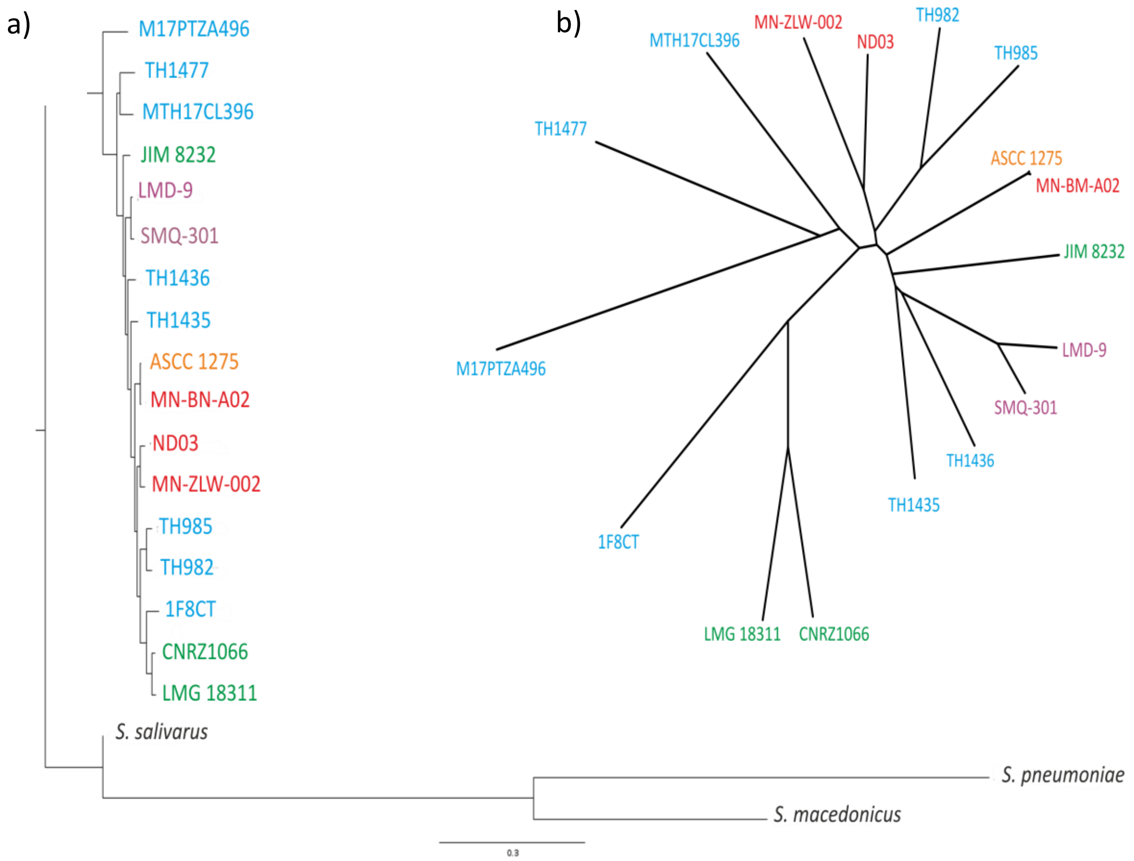


Figure 2. 6 Phylogenetic trees of *S.thermophilus*. a) Analysis performed considering 400 conserved genes. Strain IDs are colored according to their geographic origin: Italian strains, blue; European not-Italian strains, green; Asiatic strains, red, Australian strain, orange, American strains, purple. Outgroups are in black. b) Analysis performed using SNPs determined by whole genome alignment (outgroups are not included).

It should be noted that the two analyses are concordant even if the two approaches show small diversities in particular in the distance calculation. Strain M17PTZA496, not only differs in terms of genome size, but it also shows a higher difference compared to the others from evolutionary point of view both considering SNPs and conserved genes. As clearly shown in figure 2.6, the geographical origin of the strains is not linked to phylogeny since Italian strains are not clustered together while a shorter distance is evident for both 'Americans' and two out of three 'Asiatic' strains and the third remain extremely closer to Australian one. In some specific cases (i.e. TH982 and TH985; TH1435 and TH1436) the strains isolated from the same matrix are neighbouring (buffalo and goat milk respectively), although we cannot exclude also here a partial regional influence.

1.7.6 Lateral genes transfer analysis

The Lactobacillales have relatively small genomes, with the number of genes in different species ranging from 1,600 to 3,000 kb. This wide variation suggests that the evolution of LAB entails active processes of gene loss, duplication, and acquisition (109). Microevolution leads single strains to diverge form the average genomes of species. Mutations accumulation at a normal rate cannot explain big changes in the genome, and although in this study it was clear that SNPs variation and gene decay can be considered strongly enough to move this strain toward genetic shift and thus that it is probably undergoing to speciation, they still do not explain the widely increase of genetic material. Instead, several other mechanisms could be involved: acquisition from the environment, lateral gene transfer or phage-mediated incorporation, duplication of genes. To clarify this phenomenon, lateral gene transfer prediction was performed by means of the specific software Alien Hunter.

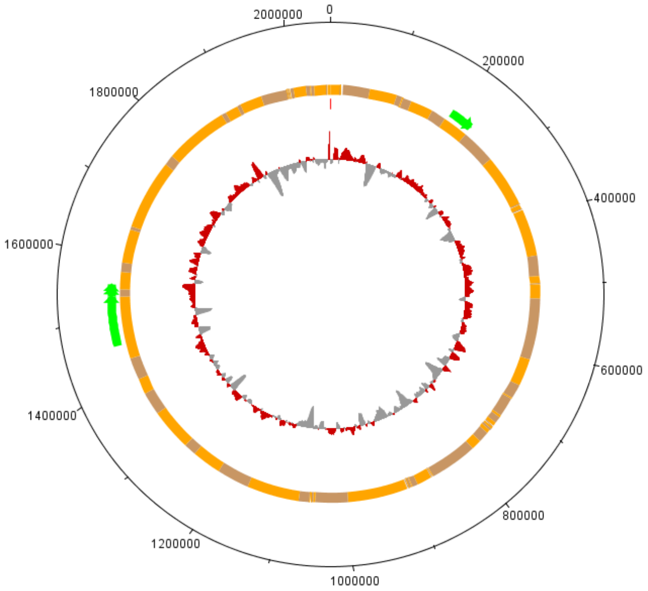


Figure 2. 7 Results of LTG areas prediction visualized by Artemis. Orange and dark pink tracts represent contigs of analysed genome. Green arrows identify LTG areas while the spike profile in the centre represents variation in GC content with respect to the mean GC content

The analysis revealed two big island of insertion, which were further explored by blastp analyses to discover their putative function. From these analyses it resulted that genes present in the two regions do not constitute a functional cluster but belong to different categories, i.e. transporter and stress related proteins (table S1). The larger island carries also some transposases. While the smaller region (Island1) encoded almost all features clearly recognised as part of the normal genetic pool of streptococci, the larger (Island2) region encodes genes attributed to different species. Similarity scores show a slight lowering in goodness of the features localized into the second island, indicating a higher decay which may be attributed to accumulation of mutation across generations. It is possible to speculate that probably this region was included in the bacterial genome before the other one.

1.7.7 Parologous genes identification

Alongside lateral genetic material acquisition, orthologues identification is a prerequisite for the evolutionary analysis of different bacterial groups while duplication events can actually better explain strain (relatively recent) diversification. Cluster analysis revealed that sixty clusters were detected to carry multiple copies of M17PTZA496 genes with 90% of identity and only six with 99% of identity, mainly assigned to mobile elements and related proteins- (table S2). Paralogs were assigned to arbitrary classes created on the number of copies of CDS carried from the other strains and placed in the cluster This classification has revealed that 30% of paralogs were shared with at least three other strains ('All' and 'Several' strains categories, 12% and 18% respectively) and that a small amount (5%) was shared only from the two strains isolated from the same environment, namely M17PTZA496 and MTH17CL396. More than half (57%) of the duplicates genes were found to be paralogs only in M17PTZA496. This group was detected to be particularly rich in RNA related genes, as previously detected by general overview of annotation (see paragraph 2.4.3) Nonetheless some peculiarly features which can make an advantage to strains proliferation were identified. Within these were identified anUDP-glucose 4-epimerase, involved in the galactose metabolism and a the transcriptional regulator, which belongs to a regulator family detecting to act on (i) quorum sensing-regulated protein, (ii) repressor of a multidrug efflux transporters, (iii) regulator of thermal resistance and (iv) the phenolic acid stress response (110). Also a gene coding for fructose-bisphosphate aldolase, enzyme involved in the glycolysis and in the microbial metabolism in of antibiotics biosynthesis, a choline binding protein, which acts in the osmo-protection system and an acyl carrier protein, employed in the lipopolysaccharide biosynthesis, were recorded. The rest (8%) was attributed to exclusive clusters, constituted only from genes belonging to this strain .Functions of these genes are related to mobile elements.

Eighteen out of sixty features are encoded from almost flanking sequences (placed in contig69, from peg.2001 to peg.2023) hence all this portion of genome was duplicated at the same moment. The second copy of this region is allocated in contig71, from peg.2064 to peg.2086. Four genes are included in the duplicated region but were not detected as paralog. Probably these genes are undergoing a more rapid decay (fig 2.7)

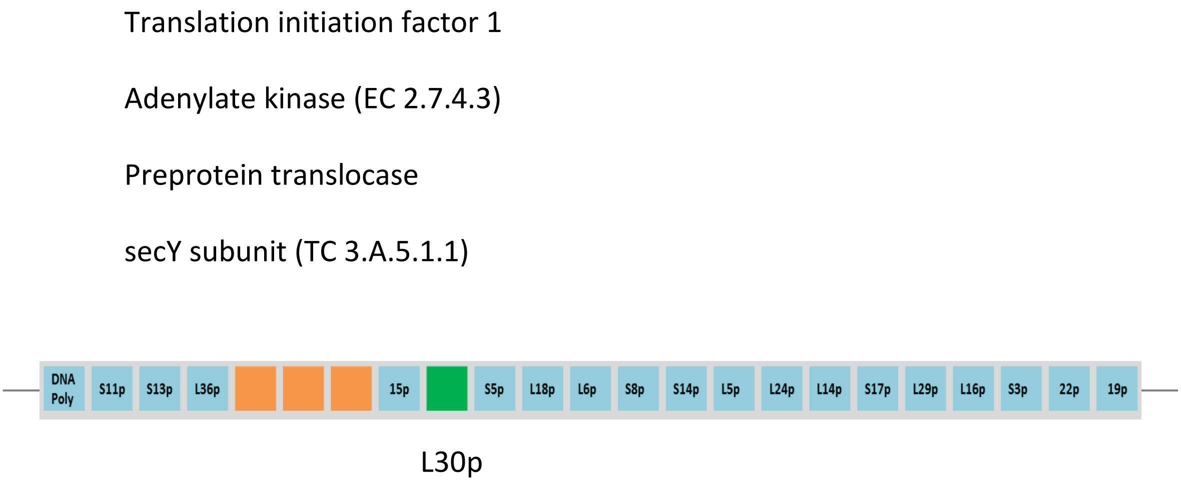


Figure 2. 8 Representation of duplicated region discovered in M17PTZA496. Above and below, genes identify as putative decaying. Within the figure genes are abbreviated. DNA poly: DNA-directed RNA polymerase alpha subunit (EC 2.7.7.6) while tags refer to ribosomal protein ID.

1.7.8 Gene content

Differences among strains can be reflected in the variable number of genes among functional categories. To explore this heterogeneity, the number of genes assigned to RAST subsystem was calculated and further visualized using MeV software. Hierarchic clustering was performed in order to evaluate similarities among strains and to identify possible correlations with the strain isolation source (environmental conditions) by testing Pearson correlation. It should be noted that functional clustering made in closer correlation strain distant from phylogenetic point of view: it is the case of MN-BM-A02 and ASCC 1275. Also Italian strains were placed in neighbouring each other, and M17PTZA496 was moved to a central position. Instead, strains isolated from the same environment (e.g. TH982 and TH985 or TH1435 and TH1436) were kept close toghether, sharing more similitudes than with the others.

Functional distribution revealed that, besides protein and DNA synthesis metabolism and regulation, the most variable gene classes can be ascribed to five main super classes which are related to: (i) resistance to biotic and abiotic stressors, (ii) transport of specific compounds, (iii) synthesis of secondary metabolites, (iv) synthesis of extracellular structures and (v) nitrogen related systems. The variability in the first superclass is determined mainly from the different management of oxygen, which can be speculated from diversity in 'Tetrapyrroles', compounds which are involved in the heme synthesis and which are widely conserved in all the species (111), and in 'Respiration' categories. Also 'Osmotic stress' play an important role, and it is poor in strains M17PTZA496, TH1436, 1F8CT and TH1477 for ABC-transporters of choline and betaine uptake, this compound indeed is accumulated inside the cell and acts as osmo-protective compound (112). Temperature modification response system is also involved. In fact, 1F8CT and TH1477 lack cold shock proteins CspA and CspG. These genes were studied in *S.thermophilus* main cooperative species, *L. delbrueckii* subsp. *bulgaricus,* which also shares the same environment. This two cold shock proteins transcription increased after a temperature downshift from 42 to 25°C and act as RNA chaperones to prevent secondary structure formation and to facilitate translation initiation or transcription antitermination (113). Genes assigned to the 'Resistance to antibiotics and toxic compounds' class are also highly variable, since strain 1F8CT lacks a cobalt-zinc-cadmium resistance gene, while on the contrary strains MTH17CL396 and TH1435 are enriched in genes involved in cadmium resistance. Within biotic stressor, the ‘Virulence, Disease and Defense’ category shows an expected high variation as already described (2.4.3.)

Within the second super-categories, some strains seem to miss all the related genes, probably because during the annotation process they were assigned to other category. Anyway, M17PTZA496 has more genes related to ‘Membrane transport’, particularly in the ‘Iron acquisition metabolism’. This enrichment is due to the presence in this strain of genes related to hemin transport. The variability among strains is extended also to the ABC-transporter genes class and to cation transporters.

Variability in the ‘Secondary metabolism’ is mainly related to uneven distribution of the Lanthionine biosynthesis gene cluster between strains, belonging to ‘Bacteriocins, ribosomally synthesized antibacterial peptides’. Finally, genes belonging to ‘Capsular and extrapolysaccharides’ class are more abundant in TH982 among new sequenced strains, cluster which is known to be involved in biofilm formation in specific environmental conditions, and slightly to ‘Adhesion’ class. It should be remembered that in dairy context EPS are known to be important for rheological properties (114). Also nitrogen source seems to be managed in different ways by each strain, as it can be deducted particularly from the variation in gene abundance of ‘Nitrogen Metabolism’ and ‘Histidine Metabolism‘ classes (fig 2.9)

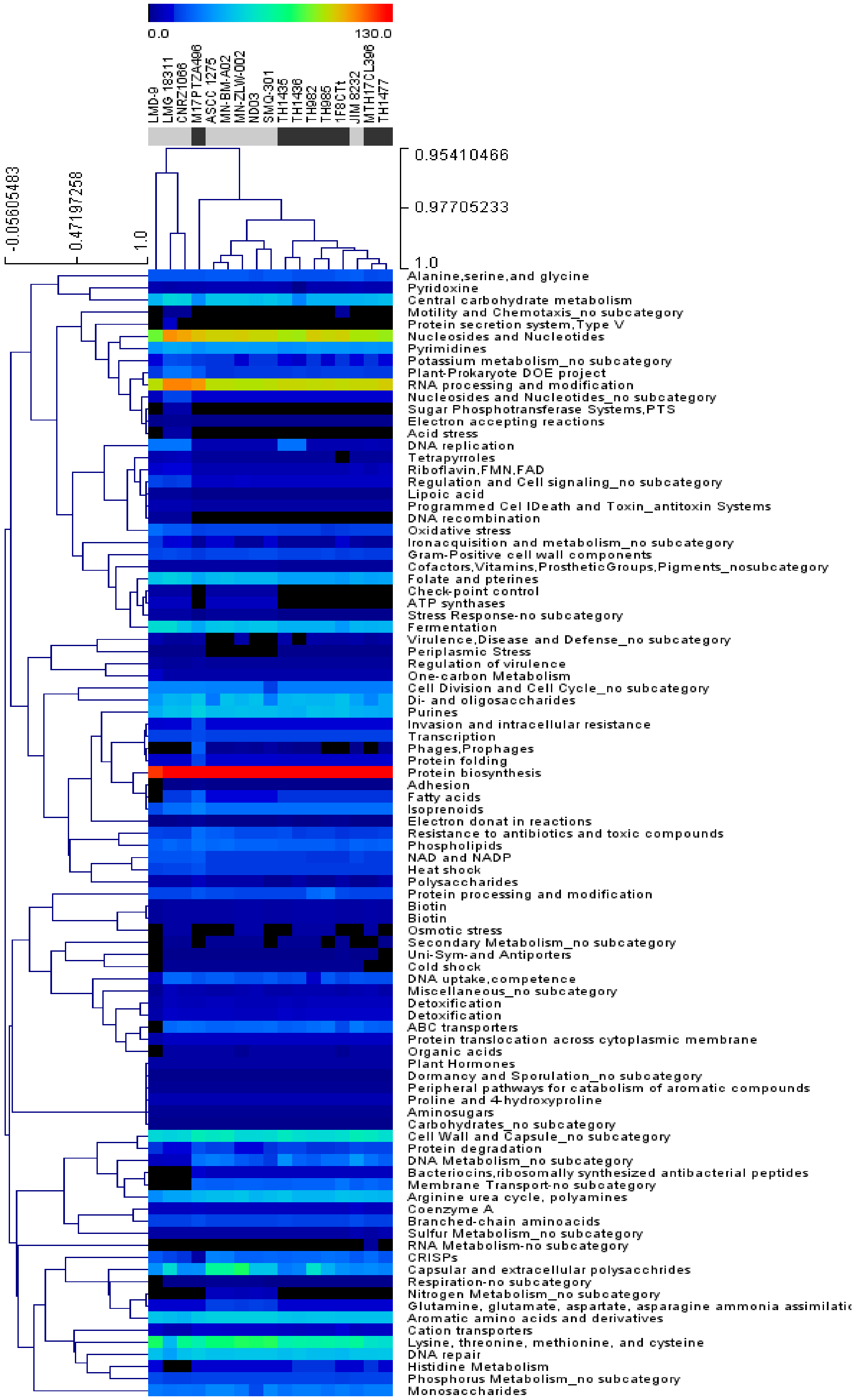


Figure 2. 9 Summary of functional categories individuated by annotation and highlight of differences between strains. On the left and on the top class and strain clustering results are presented. Dark grey: strains sequenced during this project, light grey: others

1.7.9 Specific features

It is known that each strain harbours a particular set of genes which characterizes the strain itself and that are needed to face particular environmental conditions changes. The analysis has the aim to detect whether geographical origin induced big variations on strain specific functionalities, purified from the genome size effect. Therefore, the categories distribution was analysed choosing non-redundant functions assigned from annotation.

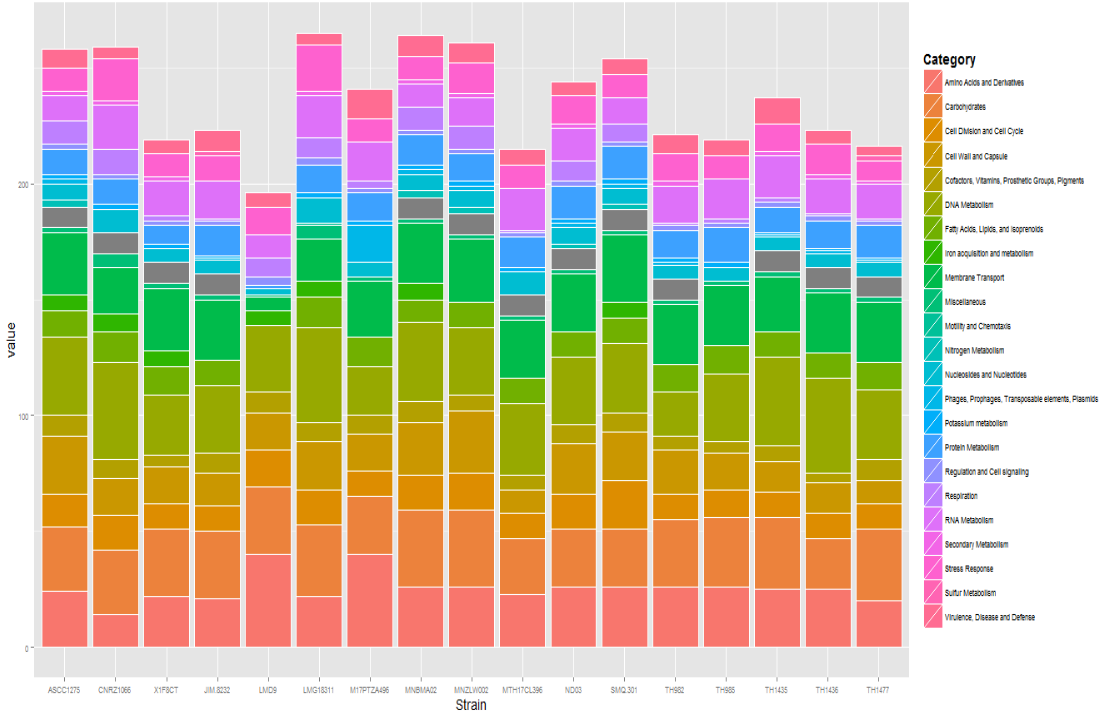


Figure 2. 10. Summary of the strain specific genes in the analyzed genomes. For each genome reported in the x axes, genes were assigned to 24 functional categories using the SEED. The number of genes for each functional category is proportional to each parcel height.

Specific genes varied between and 196 and 265 depending on the strain and representing on average 10% of the encoding regions. Among functional categories, four of the SEED categories account for most of the strain diversity, namely ‘Amino Acids and Derivatives’, ‘Carbohydrates’, ‘DNA Metabolism’ and ‘Membrane Transport’ which together describe almost 50% of the specific genes(11, 12, 13 and 14% respectively). The contribution of these categories to strain variability is known, since in comparative genome hybridization experiment (79) genes encoding for efflux/uptake pumps, EPS biosynthesis, peptide metabolism and phage related genes were classified as non-core. Interestingly, specific genes involved in ‘Iron acquisition and metabolism’ were identified only in seven strains while those involved in ‘Respiration’ vary from 1 to 11.

Technological properties insight

S. thermophilus is of major importance for the food industry since it is extensively used for the manufacturing of several dairy products. One major role of *S. thermophilus* in milk fermentation is to provide a rapid acidification. This is important because of, from one side it assures a good outcome of the dairy process and, from the other, the hydrolytic capacity of these bacteria can reduce the actual amount of lactose in the final product, as occurs in yogurt.

These organisms can also be used to increase the overall hydrolytic capacity in the small intestine alleviating lactose intolerance by their β-galactosidase enzymes (115). Similarly, galactose metabolism may enhance the utilization of such carbon sources, improving the yield of the fermentation process and reducing the amount of the free monosaccharides left in the medium. This character could be the core of a product line for galactosemic patients(68) since this character is uncommon among dairy bacteria.

The role of *S. thermophilus* in the fermentation of milk is not related only to sugar consumption but it has also several other important technological properties, such as long chain polysaccharides production, proteolytic activity and antibacterial compounds synthesis. One of the most important forms of saccharides originating from *S. thermophilus* fermentation is exopolysaccharides (EPS). EPS consist of heterosaccharide polymers, in this species principally constituted of galactose, glucose and rhamnose monomers, even if also N-polymers containing acetyl-galactosamine, fucose, and acetylated galactose moieties have been reported (57)

Generally, EPS gene clusters are considered strongly diverse although the modular gene organisation is conserved and the biosynthesis of EPS occurs via a common molecular mechanism. At least 28 distinct EPS clusters have been identified in this species (116,(88). The ability of *S. thermophilus* to produce extracellular polysaccharides (EPS) is important for the dairy industry, because it enhances the texture of fermented products: in situ EPS production typically imparts a desirable 'ropy' or viscous texture to fermented end products. Moreover, these compounds are useful as commercial stabilizers in yogurt manufacturing. *S.* Because hyaluronic acid was recently demonstrate being one of their secondary components, *S.thermophilus* EPS have been recently speculated in in the formulation of pharmaceuticals (117).

Moreover, also S. thermophilus proteolytic system has a role in the technological properties of the species, since acidification activity depends also from the specific protein breakdown capacity (64). Today the genomic sequence and evolution history of *S. thermophilus* cell-enveloped peptidase (CEP) are well documented. This CEP, namely PrtS, was first detected in about 15% of the INRA historical collection. That had indicated that this characteristic was not widespread in this species, but a more recent study on whole-genome hybridization discovered that 35 out of 47 industrial strains (close to 75%) contain *prtS* gene. In the last years, *S. thermophilus* proteinase has increase its popularity to hydrolyse the principal whey proteins, enhancing their digestibility and thus making whey-derived products more suitable for the development of healthy products (118).

Besides pH lowering, another important mechanism for the spoilage bacterial control in food is biocides compounds. Bacteriocins are compounds produced by bacteria that inhibit or kill closely related species. It is known that *S. thermophilus* dominate the fermentation processes, improving the end products safety, also when a natural leaven are used. Some previous studies have examined the usage of *S. thermohilus* bacteriocinogenic strains or their purified bacteriocins (119) in dairy production. From the molecular viewpoint, bacteriocins do not display often amino acid sequence similarity that could help recognising their function. This in part could be useful contributing to the wide bacteriocin diversity and probably derives from their high rate of evolution in the counter-act to bacteria resistance, while others reveal a conserved sequence similarity to other biocins. Bacteriocins of gram-positive bacteria can be quite different in their sizes, their modes of killing, their range of effect and their modes of release and transport into the cell. Frequently these molecules lack a specific receptor for adsorption and can be of relatively low molecular weight (120). In cases like this, genome sequence information gives an unprecedented view on the biodiversity of microbe properties and the research on the physiology of *S. thermophilus* has revealed important information on the genetic basis for these characters.

1.8 Material and methods

1.8.1 Growth curves

The description of strain specific metabolic properties should begin from definition of its growth rate. Several works were aimed to define statistic models to help microbiologists in this process. Population dynamics is described by bythe logarithm of the relative population size plotted against time and three parameters are universally used for it description (fig. 3.1): (i) lag phase (λ),the x-axis intercept of the tangent,(ii) the maximum specific growth rate, μmax, the tangent in the inflection point and (iii) the maximal value reached from population (N), namely its horizontal asymptote (121). Within the possible mathematic models developed over years, three were chosen for the analysis, namely Gompertz, which is the first and widest used model in ecology, Bayani, which is the first developed specifically for bacterial growth in food (122) (122) and Huang, which was postulated for both liquid and solid food matrix (123). *S. thermophilus* strains were cultivated in M17L broth (Oxoid, Rodano, IT) at 37°C. Growth was measured by plate counts (M17L with 20 g/l of agar (Oxoid, Rodano, IT). Particular attention was paid on inoculation: about 10^5 frozen cells adhered to CryoBeads (Pro-Lab Diagnostic, Neston, UK) were inoculated in 10 ml of pre-warmed medium. Intervals between data collection were defined depending on strain. Four indices were selected to identify the most suitable growth model (124): two describe the model performance, namely 'bias factor' (BF), an index of average deviation between predicted and observed, and the average accuracy of the estimates, the 'accuracy factor' (AF). The others describe the statistical significance of the difference between models in terms of the goodness-of-fit: the root mean square errors (RMSE) and Akaike's Information Criterion (AIC). A fitting could be considered good when the first two indices achieved a value near to 1 and when the other two parameters reached the lowest value.

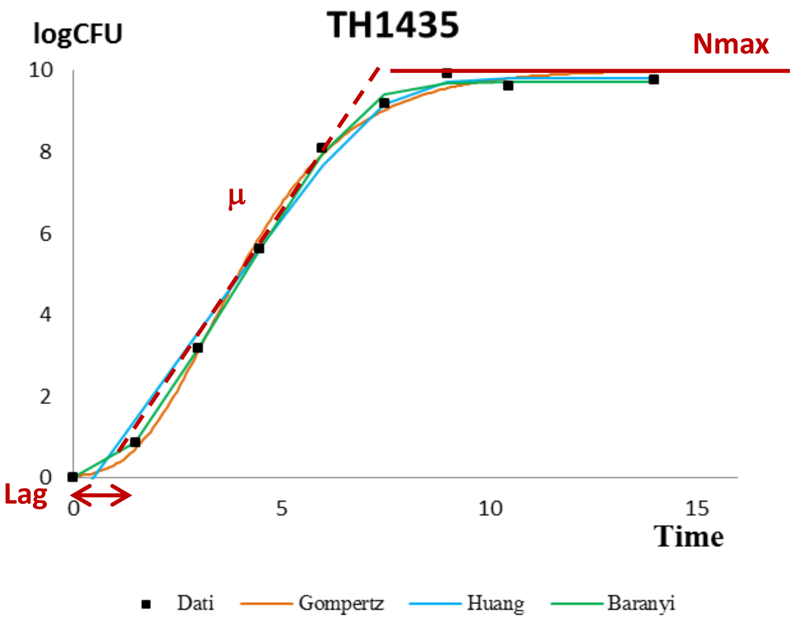


Figure 3. 1 Example of grow curve. The descriptive parameters of the curve are represented in dark red. In orange, blue and green are graphed hypothetical curves obtained from Gomperz, Huang e Bayani model prediction. Black dots are the experimental data

1.8.2 Fermentation in skim milk and commercial milk

Strain ability in milk coagulation was used as pre-test for determining whether strains may be considered suitable for industrial process. Strains were growth overnight at 37°C in M17L (Oxoid, Rodano, IT), and used to inoculate 10 ml of 10% (w/v) skim milk (Oxoid Rodano, IT), sterilised by autoclaving 10 min at 110°C. Samples were kept in water bath at 37°C as long was needed to detect milk coagulation. The time demanded was recorded and only the strains able to achieve the goal within 8 hours were further tested.

Fermentation parameters were evaluated by calculating the maximum acidification rate and time required to reach pH 5.2 and pH 4.6 (fig 3.2). Final media were inoculated with 2% (v/v) of a strain culture grown overnight at 37°C in M17L broth. Flasks were prepared with 250ml of fermentation media, either 10% (w/v) sterilized skim milk (10 min at 110°C) or fresh commercial pasteurised milk (Latterie Vicentine, Italy). The cultures were then incubated in a water bath at 37°C and the pH (pH electrode Mettler 405 DPAS SC, Toledo, Spain) was monitored during 24 hours as previously described (64). The pH was measured every second and values obtained during 3 minutes were averaged. The maximum acidification rate (V_m), defined as the maximum slope of the pH curve (dpH/dt), was expressed as pH units/minute. Experiments were repeated three times and data analysed by R.

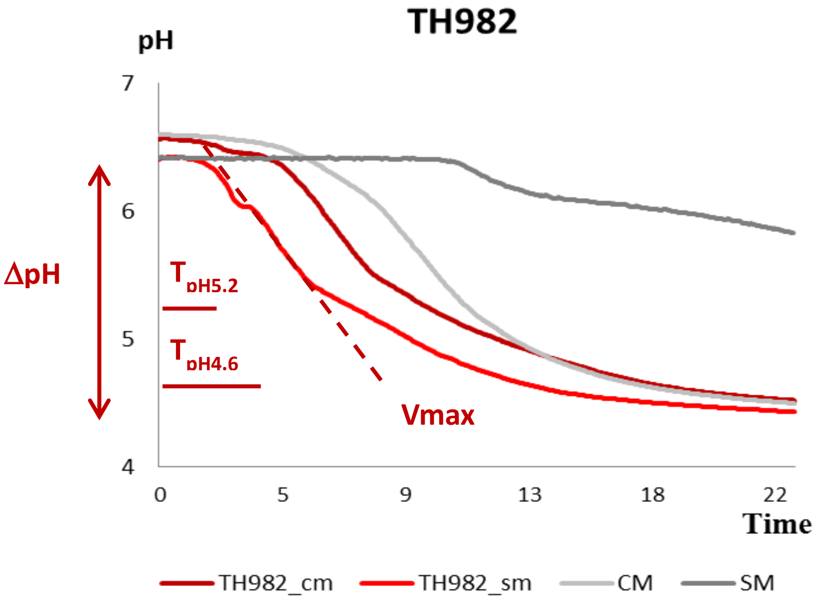


Figure 3. 2 Example of acidification curve. Parameters used for the comparison are graphed in dark red. Grey curves represented controls while red curves the experimental data collected. CM: Commercial milk, SM: Skim milk

1.8.3 gal-lac operon genes and proteinases sequence search

Since acidification rate remains the main technological property of the species, genetic analyses on two specific systems generally consider the main descriptor of fermentation metabolism were analysed. In this species, genes involved in galactose and lactose metabolism are localized on a single locus on the chromosome following the organisation *galKTEMlacSZ.* Strain LMG 18311 was chosen as reference for the comparison against new sequenced operons, according to van den Bogaard (125). Reference sequences were downloaded from the NCBI database while strains sequences were identified by matching using the blastp tool of RAST service. Each gene was processed separately, alignments of all strain sequences was performed by Muscle algorithms in MEGA (Molecular Evolutionary Genetics Analysis) v.6 (126).

Muscle algorithm is made of three steps: (i) the draft progressive, whose goal is to produce a multiple alignment, promoting speed over accuracy. It is done by calculating a distance tree base on UPGMA. (ii) The second one, the improved progressive, corrects errors made in the first step and re-estimates the tree using the Kimura distance, which is more accurate but requires an alignment. (iii) The refinement, during this stage the the new tree is divided into two subtrees of sequences. New alignments are computed both within and between the two subtrees until the scores converge. Moreover, nucleotide sequences of intergenic region of the two main regulation systems (*galR-galK* and *galM-lacS* intergenic regions) were compared in order to identify whether any variation occurred in the promoters. Alignments, obtained from Muscles, was analysed by manual inspection, seeking for the -35 and -10 sequences. After that, analyses of retained genes related to fermentation subsystems was performed as described below for all the strain, in order to evaluate the contribution of other fermentation features in the pH lowering. The search of proteinases sequence was made by searching sequences with high identity of the two component of the system, namely the strain specific cell-envelope proteinase, PrtS, and the membrane anchoring protein of proteinases, Sortase A, by blastp tool on RAST. The query sequences were recovered from the Pfam database and belonging to strain *S. thermophilus* MN-ZLW-002 (Ptrs reference no. YP_006340201.1, Sortase A reference number YP_006340309.1).

1.8.4 Protease activity

Protease activity was measured by a plate assay in order to record whether recognised sequences were expressed, according with Morris et al. 2011 (127). Lactose-free semi-skimmed milk powder was obtained from Valio Oy (Helsinki, FIN). A 10% solution (w/v) was made in deionised water and autoclaved at 110 °C for 10 min. Then, 1.5% (w/v) purified agar (Thermo Fisher Scientific, Rodano, IT) was prepared in water and autoclaved at 121 °C for 15 min. After sterilization, suspensions were cooled down to 55 °C. Semi-skimmed milk solution was added to the agar to a final concentration of 1% (w/v) and plated. Strains were grown overnight in M17L (Oxoid, Rodano, IT). For each culture a drop of 5 μl were placed on the surface of semi-skimmed plates keeping drops apart for at least 1 cm each other and from the border. Plates were incubated at 37°C and results recorded after 24 and 48 hours. Experiment was repeated three times. Results were recorded as halo sizes.

1.8.5 Exopolysaccharides subsystem analysis

Differences in *eps* gene cluster resulting from the gene content overview were further explored. Genes assigned by RAST to this subcategory were identified over the genomes, and then the number of copies and their amino acidic sequences were recorded. Sequences were clustered using CD-HIT (see 2.3.9.) at 50%, 80% or 90% of identity in order to record the degree of similarity and whether they can be considered orthologues.

1.8.6 Biofilm formation

The ability of bacterial strains to form a biofilm was analysed in two steps. First, it was evaluated in 96-well microtiter plates using the crystal violet assay as described by Maragkoudakis at al. (128). Strains were let to grow for 24 hours in 200 μl of M17L (Oxoid, Rodano, IT) at 37°C, monitoring the increase of absorbance at the wavelength of 600nm. When cultures reached their stationary phase, wells were emptied and washed three times with PBS buffer to remove floating cells. 100 μl of crystal violet 0.1% was dropped into each well and hold for 15 min at room temperature to guarantee complete material stain. The colorant was gently removed and plates were washed three times with PBS buffer. The stained biofilm was solubilised by using 95% ethanol. Finally, biofilm formation was quantified by measuring absorbance (OD_{590}) values with a spectrophotometric plate reader (TECAN, Männedorf, CH). Assays were repeated at least seven times for each strain and data analysed by R.

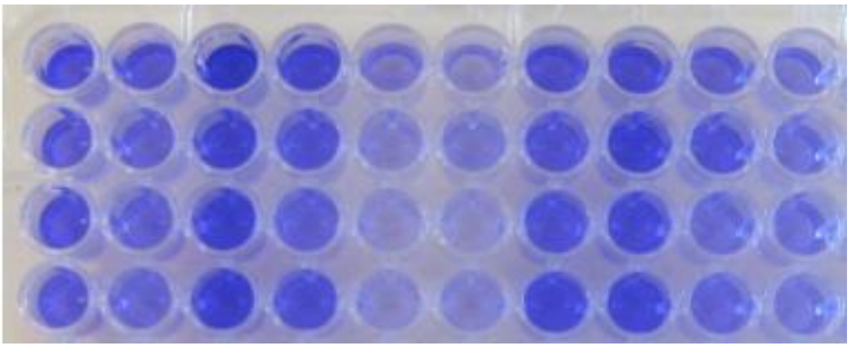


Figure 3. 3 Example of results obtained from crystal violet assay. In the plate is shown ethanol resuspension of the stain, different strains (organized one for each column) with replication (rows of the column) are presented.

After results evaluation, two representative strains, namely MTH17CL396 and TH985, were chosen for image analyses. For optical microscopy analysis and scanning electron microscopy (SEM), cells were grown 24 hours in M17L at 37 °C. ten-time diluted cultures were transferred in sterile Petri dishes and maintained at 37 °C for 24 hours on a glass coverslip which provided a mobile adhesion surface. After removal of the medium, half of the coverslips were gently washed using PBS buffer. Differential interference microscopy (DIC) was performed directly on the coverslips using a CTR 5000 microscope (Leica Microsystems, Wetzlar, DE). For SEM analysis, cells were fixed in a solution of 1% glutaraldehyde and 1% paraformaldehyde (in 0.1 M sodium cacodylate buffer, pH 7.2) for 12 hours at 25 °C. After this period, samples were washed three times with the same buffer, post-fixed in a 1% osmium tetroxide solution for further 12 hours at room temperature and washed again three times. Samples were then dehydrated in a series of solutions with increasing alcohol concentrations (from 70 to 100%) and dried in a CPD7501 critical point dryer (Polaron, Watford, UK). The samples were assembled on aluminium stubs with carbon tape and covered with gold using S 150B sputter coater (Edwards, Crawley, UK). The images were acquired using a Quanta 200 SEM (FEI, Hillsboro, OR).

1.8.7 Bacteriocin sequence detection

Identifying genes encoding bacteriocins and ribosomally synthesized and post-translationally modified peptides (RiPPs) can be difficult, especially for those peptides that do not shase sequence similarity to already identified peptides and heavily modified peptides like lanthipeptides, category recognised as interesting during strain comparison (paragraph 2.4.8). Therefore a specific software, BAGEL3, it was chosen for the mining task (129). Its identification approach combines direct and indirect mining, via context genes. The genetic context, in fact, harbours worth information that could be used for mining purposes. The main complexity in these kinds of task is the small size of the gene encoding for the target peptides. In fact, small ORFs are often omitted during automated annotation proceduers especially when their product do not show a strong similarity with known peptides. Another major advantage of BAGEL3 is its use of DNA sequence as input instead of annotated genomes, making it less dependent on ORF predictions. DNA sequences are analysed in parallel using two different approaches. (i)

The direct approach begins with a Glimmer ORF call. Then, ORFs are blasted against the databases and the context annotated using the Pfam database. (ii) The indirect approach starts performing a simple ORF search on the DNA. The products of these ORFs are subsequently screened for the presence of targets, called areas of interest (AOI). Then, an additional specialized ORF call is performed for finding small ORFs that encode for the targets individuated from the previous step in protein domains. Rules based on the discovered domains are further used to decide which part of the nucleotide sequence should be analysed in the AOI search. Eventually, the context is annotated using the Pfam database. The last step of the procedure is identifying the RiPP genes that are present in the AOI, using the results of a blast search against the BAGEL3 databases and a screening for known motifs. If any hit is obtained, BAGEL3 predicts a precursor peptide sequence based on sequence properties and genomic organization. Software output is actually the putative bacteriocin sequence and its context (overall of 20Kb).

1.8.8 Bacteriocin activities

Because all the strains were recognised as potentially bacteriocin producers, all were tested for the expression of antibacterial activity.

Antimicrobial activity was determined by spotting 5μl of tested strain grown cultures upon a M17L 20 g/l (w/v) agar as bottom-layer, let them grow at 37°C for 24 hours before gently pour above an upper layer with each indicator strain . Tested strain were grown overnight in M17L at 37°C. Instead, obtained dual-layer plates were incubated at the growth temperature of the indicator strain and results read after 24 and 48 hours. Two assays were set up, the first one aimed to verify inhibitory activity against mutualistic bacteria. For testing *S. thermophilus* -to- *S. thermophilus* compatibility, it was prepared by adding an overnight grown liquid culture of *S. thermophilus* into fresh prepared M17L agar to reach 5 g/l (w/v) of agar concentration (soft agar) and a the final concentration of 10^7 cell/ml or using MRS soft agar (following the same protocol) to examine *Lb. delbrueckii* subsp*. bulgaricus* and *L. delbrueckii* subsp*. lactis* interaction with Italian strains. The second had the goal to identify production of compounds able to inhibit undesirable bacteria. In this case, the upper layer were prepared with Baird-Parker broth for *Staphylococcus xylosus*, Brain heart Infusion broth for *Listeria innocua* and *E. coli*, Nutrient broth for *Bacillus amyloliquefaciens, Bacillus subtilis* and

Psudomonas fluorescens and MRS for *Enterococcus faecium.* All media were provided by Oxoid S.p.A. (Rodano, IT). Each test was repeated three times. Antimicrobial activity was considered in relation to inhibitory halo diameter.

1.8.9 Amino acids biosynthesis system analysis

The comparative analysis of genomic information related to amino acid biosynthetic gene clusters was performed by the extrapolation of a matrix containing genes belonging to those RAST subsystems assigned for each genome. Results were visualized by using MeV 4.9.0 (paragraph 2.3.10)

1.8.10 Amino acidic requirements

Strains were grown overnight at 37 °C in 10 ml of chemically-defined media (CDM) modified from Letort & Juillard (130), details in 'Media and solution'. For the test, 21 different broths were prepared omitting one amino acid in each one. Complete CDM was used as positive control for bacteria growth, while CDM without all the amino acids and the CDM lacking from amino acids and urea were used as negative controls. All the media were sterilized by filtering 0.22 µm. Pre-inocula were prepared to growing strains in complete CDM overnight at 37°C. Pre-inocula were washed twice with PBS buffer and inoculated 1% (v/v) in all the final media. Inocula viability was verified by cell count on M17L agar, in all the experiments inoculation achieved of 10^5 cell/ml. Growths were performed in 384-wells plates, covered by plastic lid and incubated at 37°C. Growth curves were monitored by a microplate reader (TECAN, Männedorf, CH), recording OD_{590} values every 10 min during 72h. Non inoculated media were used to control absorbance variation due to colour or volume modification over time. Each test was repeated at least 4 times.

1.9 Results and discussion

1.9.1 Grow parameters

It was recently reported how small variation in genomic information affect the metabolism expression and how bacterial growth represents one of the most susceptible index of this phenomenon (131). Because large differences were identified from the genome comparison, it was decided to test the overall metabolic functionality by analysing growth parameters in synthetic medium. Data were processed using three different statistic models selected from several used in literature. The goodness of fitting was tested using both the mean squared error and the AIC index, as suggested (132). When the two parameters were not in agreement only the latter index was considered (table S3). In almost all the cases Gompertz model was preferred for parameter estimation but for MTH17CL396 the Huang model was considered since it appeared to be more accurate (table S3).

Strain ID	1F8CT	M17PTZ A496	MTH17C L396	TH982	TH985	TH1435	TH1436	TH1477
log Nmax (cell/ml)	8.5	10.3	9.1	9.5	9.9	9.6	9.0	9.2
μ (h^{-1})	0.53	0.42	0.32	0.41	0.35	0.35	0.45	0.26
Lag (h)	2.93	1.12	1.16	2.05	1.46	1.46	1.53	2.21

Table 3. 1 Growth parameters calculated for the strains. Only results obtained by Gompertz model are presented because indices revealed that it represent the best fitting model. MTH17CL396 parameters were calculated by using the Huang model.

Strains show slight overall differences in growth rates. It should be noted that particularly interesting is the behaviour of strain 1F8CT which reaches a lower number of cell at the stationary phase, it has the lowest growth rate and the longest lag phase. It could be explain by taking in account the probable adaptation to its own isolation environment, which is characterised by relative high temperature of processing. Strain M17PTZA496 is characterised by the higher cell population number and show a reduced lag phase, as well as MTH17CL396. It is interesting to note that both these strains were isolated from the same matrix, the Fontina cheese PDO. The higher growth rate was expressed by TH1477, whilst it seems to require more time than the other before restore the replication activity after the metabolic stop.

1.9.2 Fermentation results

To evaluate strain acidification ability, pH variation during fermentation was monitored in two media, skim milk and fresh commercial milk. Before starting the measurement, a pre-test was performed by recording the time required to achieve the coagulation of skim milk and only the strain reaching the goal within 8 hours were selected for further analyses. Seven out of eight strains coagulated skim milk in few hours while 1F8CT revealed inability to perform this task at all: it was left in the water bath up to 24 hours but no changes in milk texture were recorded.

Strain ID	skim milk				commercial milk			
	ΔpH	V_{max}	$T_{pH\ 5.2}$	$T_{pH\ 4.6}$	ΔpH	V_{max}	$T_{pH\ 5.2}$	$T_{pH\ 4.6}$
M17PTZA496	0.99 (0.08)	29 (9)	10 (0.58)	-	1.38 (0.04)	34 (2)	10 (0.04)	-
MTH17CL396	0.98 (0.19)	61 (10)	16 (1.11)	-	1.25 (0.11)	25 (20)	13 (1.13)	-
TH982	1.74 (0.15)	33 (29)	7 (1.79)	13 (1.69)	1.67 (0.20)	159 (25)	10 (0.19)	17 (0.34)
TH985	1.26 (0.19)	24 (34)	12 (0.44)	23 (5.29)	1.53 (0.29)	92 (20)	9 (1.27)	19 (6.13)
TH1435	1.50 (0.18)	57 (10)	4 (0.59)	6 (0.87)	1.73 (0.12)	41 (2)	4 (0.30)	6 (0.58)
TH1436	1.62 (0.10)	52 (34)	4 (0.52)	7 (0.42)	1.74 (0.08)	106 (7)	4 (0.27)	6 (0.52)
TH1477	1.41 (0.29)	43 (8)	6 (1.27)	14 (6.13)	1.64 (0.05)	62 (13)	9 (0.67)	20 (1.20)

Table 3. 2 Acidification parameters describing strain activity during fermentation. Total variation in pH, Vmax acidification rate express ΔpH unit per 10^{-4} per minute and time required to achieve two pH point (in hours) particular interesting in dairy production are reported for both the media

Not all the strains reached pH 4.6, the important point for yogurt manufactoring, while all reached the first pH point. This was expected since it was the parameter for strain selection although it is clear that the experimental set up changes affected time demanded to lower the pH. Two strains, namely TH1435 and TH146, displayed outstanding performances lowering the pH below 4.6 within seven hours in both media. It can be speculated that this ability may be related to lactose consumption efficiency. Only M17PTZA496 displayed a significant difference in the final pH by comparing the M17PTZA496 it was detected a significant difference in the final pH comparing the results of two media fermentation (two tail t test, $t_{(2)}$= 4.30, p-value =0.02). Analysis on maximum acidification rate has shown that both ‘strain’ and ‘medium’ effects play a role on the determination of acidification profile (two-way ANOVA, p-value= 0.01 and p-value =0.001 respectively). T-test comparisons among media have shown that only two strains rate differences are statistically significant (p-value= 0.02 and p-value= 0.03 for TH982 and TH1477 respectively). While time points cannot be compared with what previously described from Dandoy at. (64), maximum acidification rate findings permit to recognise that tested strains have a good capacity in fermentation performance. In fact, TH982, TH985 and TH1436 reached a ΔpH closed to 0.1 point per minute. It should be noted that this results were achieved, in previous work, only from natural and artificial strains carrying the protease active form. Another similarity is interesting, namely that, similarly to TH982, a better performance was recorded in the low-heat treated media also for one of the strains reported in that study. It indicates that, for some strains, the thermal treatment of the media strongly influences the outcomes.

1.9.3 Gal-lac operon comparison

Despite that lactose operon was one of the first described and widely explored bacterial genetic system (125), it harbours a high genetic heterogeneity to date not exhaustively explained. For *S. thermophilus* it means that while several studies had the goal of defining which pattern of single locus variations result in galactose consumption (133), character which is going to vanish in this species, few information are available concerning the effect of those variation on the fermentation rate. The *gal-lac* operon of *S. thermophilus* is constituted of two parts working together. The first one is involved in galactose metabolism, it is composed of five genes: (i) *galR*, coding for a transcriptional regulator and which is transcribed in the opposite direction respect to the others, (ii) *galK* the galactokinase, (iii) *galT*, galactose-1-P uridyltransferase, (iv) *galE*, namely the UDP-glucose 4-epimerase, and (v) *galM*, the galactose mutarotase. The second part is actually the *lac* operon, it consist in two genes *lacS,* the lactose transporter, and *lacZ,* namely β-galactosidase. These genes, which play a fundamental role in the fermentation process, are well recognised for their wide sequence diversity among strains, so wide that a strain identification assay was based on these genes (134). In this work, a multi sequence comparison was performed on sequences found in the sequenced genomes. In particular, the attention was focused on the comparison of the two very good pH lowers, TH1435 and TH1436 against the others. Raw amino acidic sequence analyses has revealed only one variation shared between these strains and not with the other, an isoleucine to valine substitution at position 174 of *galM*. Interesting finding was the recording in *lacZ* of a wide deletion (60/1026 amino acids) in all the strains that seems not to alter gene functionality. Moreover, sequence comparison permitted to underline that three out of eight strains (M17PTZA496, MTH17CL396 and TH1477) are IIA^{LacS} group I while the other belong to the group II: for years it has been speculated that this genetic distinction was resulting in differences of the lactose transporter switch rate, but today its implication is less clear (135). Because of the number of similar point variations which seem not to affect the overall cluster functionality, analysis were further conducted on the promoter regions. Two intergenic region were compared: *galR-galK*, which was well studied and is considered the main regulator for all the *gal-lac* operon because of its participation in the transcription of *galR* (136), and *galM-lacS,* site of *lacS* promoter which slightly affects *lac* operon transcription rate (125) and from the literature considered acting on operon expression (137). Results highlight a point mutation in the -10 site of *galK* transcription factor of M17PTZA496, which is connected with the galactose consumption (125). In the *galM-lacS* intergenic regions several variations are reported for M17PTZA494, also one in the -35 sequence, which in any case cannot be clearly correlated with the expressed phenotype (table S5). Also changing in content of fermentation gene subsystems were evaluated, including ‘Fermentation: Lactate’ and ‘Fermentation: Mixed acid’ clusters in the analysis. It was performed in order to understand if the very good acidification abilities of the two strains isolated from goat milk could depend on the production of different organic acids but no significant difference was recoded.

1.9.4 Proteinases detection

The proteinase system was detected both at genetic and phenotypic level. From the genetic point of view, while all strains showed sequence matching with the Sortase A (in average: identity =97%, discrepancies depend mainly from SNPs, E-value = e-176), only in TH1435 the *prts* complete sequence was detected, with some slight differences against the reference sequence identified in point variations randomly distributed over the sequence, a DIP and a INDEL of three bases (identity = 97%, gaps = 5/4582, E-value =

0.0). Moreover, in this strain and in all the others, smaller fragments of sequences were detected but the sequence coverage (less than 20%) was evaluated insufficient to justify further investigation. The phenotypic assay, otherwise, has highlighted the absence of proteinase activity in all the strains.

1.9.5 Exopolysaccharides gene cluster

In a previous work (138) it was demonstrated that eps genes are involved in enzymatic reactions that are independent from the sugar since the EPS of *S. thermophilus* and capsular polysaccharides of closely related species conserve a high similarity level even if from a structural viewpoint are quite different. The eps gene structure could be summarized in a structure organized as follows: (i) regulation genes (*epsA, epsB*), (ii) genes regulating the chain length determination and the export (*epsC, epsD*), (iii) genes related to the biosynthesis of the repeating units for synthesizing exopolysaccharide (*epsE, epsF, epsG, epsH*, and *epsI*), and (iv) genes active in the polymerization and export (*epsK, epsL, epsM*) (44) It is known that *eps* show homology with another cluster of genes, conventionally called *cps*, which encodes for capsular polysaccharides (80). Therefore in the analyses both the systems were included.

Strain ID	*Glt1*	*Glt2*	*EpsA*	*EpsB*	*EpsC*	*EpsD*	*EpsE*	*CpsF*	*CpsG*	*CpsM*
1F8CT		1	1	1	1	1	1	1	1	1
M17PTZA496		1	1	1	1	1	1, 1	1	1	
MTH17CL396	1		1	1	1	1	1			
TH982	1	1	1	1,1	1,1, 1	1, 1, 1	1			1
TH985			1	1, 1	1, 1	1, 1	1			
TH1435			1	1	1	1	1			
TH1436			1	1	1	1	1			
TH1477	1		1	1	1	1	1			

Table 3. 3 Summary of genes assigned to the polysaccharides biosynthesis systems. Colours distinguish the aminoacid sequence in identity class: 100% identity, blue; 90%, red; 80%, green; 50%, yellow; black and grey distinguish genes clustered separately.

Strains show variable numbers of gene belonging to these subsystems, from 4 (in TH1435 and TH1436) up to 13 (TH982). As expected, all strains carry part of the two subsystems, which are partially coincident for the biological function. Five genes play an essential role in cell wall construction and adhesion between cells, which are ‘Exopolysaccharide biosynthesis transcriptional activator’ (*EpsA*), ‘Manganese-dependent protein-tyrosine phosphatase’ (EC 3.1.3.48, *EpsB*),’ Tyrosine-protein kinase transmembrane modulator '(EpsC), 'Tyrosine-protein kinase' (EC 2.7.10.2, *EpsD*), 'Undecaprenyl-phosphate galactosephosphotransferase' (EC 2.7.8.6, *EpsE*). *EpsA, EpsB* and *EpsC* are part of the subsystem normally carried both by ropy and non-ropy strains. An interesting difference has been recorded in the number of copies and in the degree of homology of these sequences. It seems to support the idea that this system is undergoing genomic decay, and may be deletion of an ancestral acquired gene cluster (138). Another element may contribute to modulate the expression of EPS is the presence of additional glycosyl-, galactosyl- or rhamnosyl transferase genes, generally called *glt*. From the analyses, six out of eight studied strains possess sequences encoding for this kind of enzyme, subdivided in glycosyltransferase, family group 1 or 2 (tab. 3.3). Even through genes of biosynthesis of the repeating units and genes of polymerization cannot be detected, homologous genes were found in 1F8CT, M17PTZA496 and TH982.

1.9.6 *Biofilm production*

In the effort to explain whether genetic differences reported in 3.2.5 actually enhance, in any case, the production of EPS, two phenotypic tests were performed. First, a colorimetric assay based on crystal violet was done.

This test permits to quantity the organic material anchored to the wall of the microtiter plate wells. Only when the bacterial growth in microtiter plate were considered reliable the test was conducted

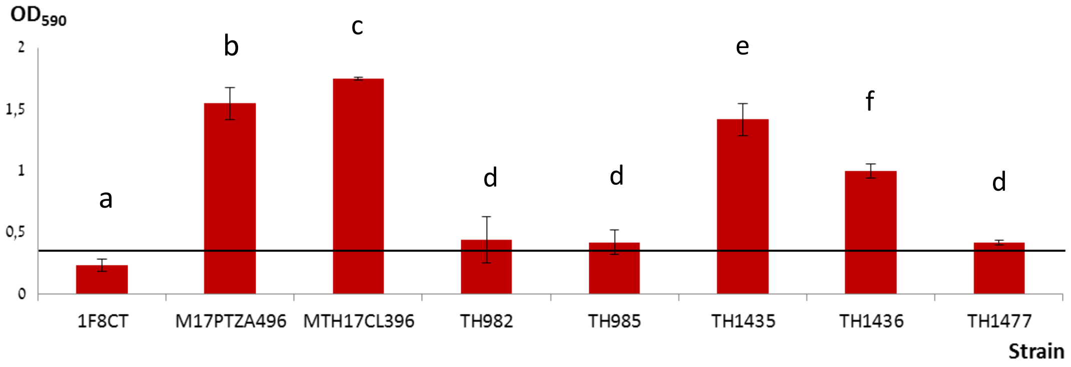


Figure 3. 4 Summary of results obtained from the colorimetric assay. Black line indicates the arbitrary threshold individuated to evaluate strains behaviour, namely the OD value recorded for empty wells used as control plus three standard deviations. Letters above the bars individuate the statistical groups.

Statistical analyses highlight that there were significant differences among strains (p-value <0.001). Even if replicates influence the results (p-value<0.01), when correct as factor into level, the statistical differences among strains keep significant (p-value <0.001). Analysis of contrast revealed that strains can be dived into six groups based on the amount of colour retained during crystal violet staining. Only 1F8CT was detected as not putative biofilm forming at all, while three strains (TH982, TH985 and TH1477) gave uncertain results, showing a slight ability only in some tests. Genetic analysis of the cluster revealed that 1F8CT was carrying also the accessories genes, both *glt* and polymerization genes, (paragraph 3.2.3) but it seems not correlated with their expression. On the contrary, MTH17CL396 which was detected to possess the simplest *eps* gene cluster, in the phenotypic test displayed the best performance.

Also significant is the fact that only one strain shows a strong consistency in replications, namely MTH17CL396 and this was supposed to be connected with an early production of EPS which, during the growth, had enough time to build multilayer structures, more robust than the others. To verify this hypothesis, a further analysis was set up. It was chosen to examine two strains, MTH17CL396 and TH985, in image processing. A series of differential interference microscopy (DIC) investigations were performed directly on coverslips to find the best moment in which to capture the forming structure by SEM. It was chosen to let the culture deposit over the coverslips for 24 hours before recovering the glasses and washing gently half of the samples. Both washed and unwashed glasses were processed and visualized at scanning microscopes but no significant differences were detected.

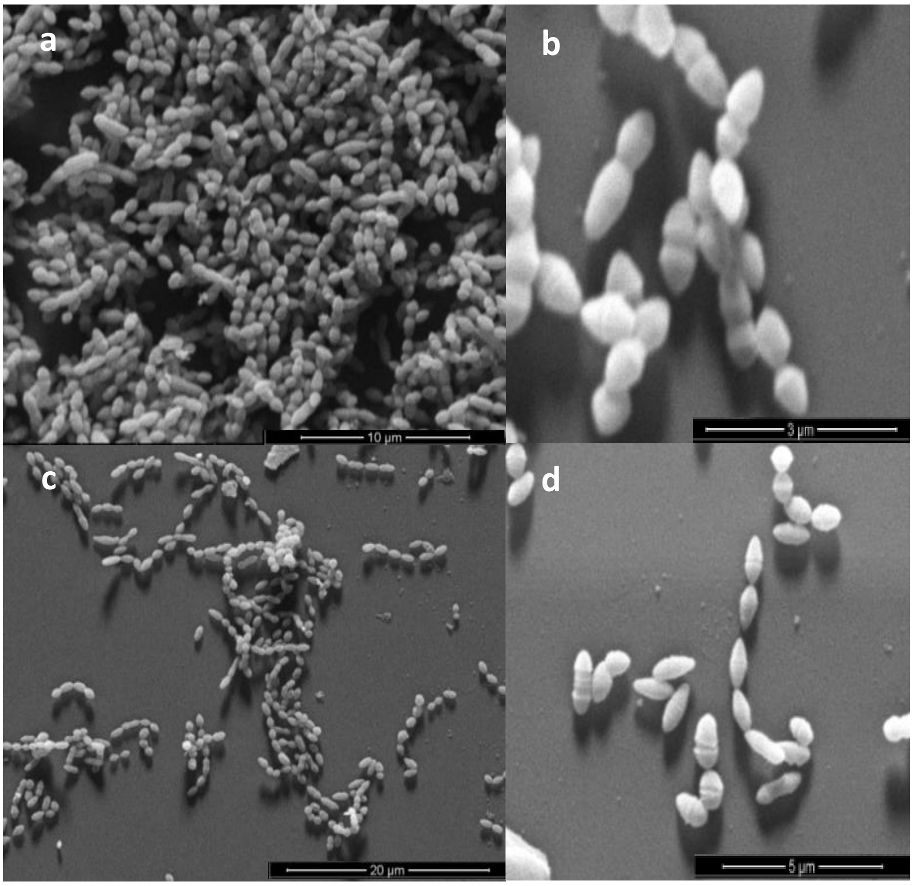


Figure 3. 5 SEM images of strains. a) MTH17CL396, 3000x b) MTH17CL396, 20000x c) TH985, 5000x d) TH985, 10000x magnification

Images reveal that no strains actually formed exopolysaccharid fibres and their differences were attributed to the degree of complexity in chains organization. The results obtained from the colorimetric test reflected, probably, the straightness of cell-to-cell anchoring system of cellular membrane. It is concordant with the genetic findings.

1.9.7 Bacteriocin genes mining

Because of the interesting results from whole genome functional comparison, which underlined differences in gene abundance among strains, in particular those belonging to bacteriocins and biosynthesis of lanthionine, presence of genes coding for these molecules was further investigated by a specific search. Lanthionine is a nonproteic amino acid involved in bacteriocin post-translational modifications, assigned to bacteriocins Class I, called lantibiotics. Bacteriocins are active against Gram-positive pathogens such as *L. monocytogenes* and *S. aureus*, and may be effective also against

Gram-negatives if the outer membrane is destabilized; some lantibiotics were recognised to act against a broad range of bacteria (139). Lantibiotics are small peptides, in average from 19 to 38 amino acids, which undergo extensive post-translational modifications. Until 2007 no fewer than 15 different post-translational modifications were documented in lantibiotics (140). The search for this type of molecule is difficult, therefore it was used a specific tool recently developed.

Strain ID	Lantibiotic related genes	ABC transporter	Lactococcin LcnD-like	Pore-forming peptide	Bacteriocin self-immunity protein
1F8CT	8	3	1		
M17PTZA496	1	2	1		1
MTH17CL396		5	1	1	
TH982	4	2			
TH985	2	1	1	2	
TH1435	2	4	1	2	1
TH1436	3	2	1	1	1
TH1477	1	2	1	1	1

Table 3. 4 Sequences recognised encoding for putative bacteriocins.

BAGEL3 provided an output in which the motif assigned belonging to the target molecules is highlighted in the putative bacteriocin sequence. Almost all the sequences were ascribed to the lantibiotic compounds, which are widely present in LAB and of major interest for dairy productions. It was demonstrated that in some cases, biocides are probably produced from the cell but the absence of specific transporter or maturating system actually prevent the activity (141). This function was recognised to be essential in Lactococcin D and for the correct expression of Lactococcin A. Even if in any system it wasn't found evidence for lactoccin production, the presence of LcnD-like sequence cannot be exclude it due to its possible involving in post-translational modification of other compounds, in particular if it is taking into account that it was identified in seven out of eight analysed strains. MTH17CL396 seems the unique strain not carrying sequences belonging to lantibiotics. It is known that some of the most studied lantibiotics (e.g. nisin) work as pore-forming protein. Therefore we cannot exclude that the pore-forming peptides recognised in this strains could belong to the same class. Antibiotic synthesis requires that producer organisms have a mechanism conferring resistance to their product. In the case of lantibiotics, immunity can be provided by a specific immunity protein, for example by masking the bacteriocin target molecule. Therefore the finding of this kind of molecule in four strains, namely M17TZA496, TH1435, TH1436 and TH1477, can be considered an indication of a probable antimicrobial activity.

1.9.8 Bacteriocin activity

The inhibition spectra of the eight *S. thermophilus* strains tested on an overall of seventeen bacterial strains of species commonly found in dairy products have revealed that none of the strains expressed antimicrobial activity. Antimicrobial compounds show, generally, species-specific range of action but, in the case of lantibiotic, which represent our main targets (see paragraph 3.2.5), are known to exhibit an antibiotic activity against a broad range of bacteria (120). Therefore, other analyses were excluded.

1.9.9 Amino acids biosynthesis

An overall of 168 genes distributed in fourteen pathways were studied to identify genetic differences carried from the sequenced strains on amino acids biosynthesis metabolism. Genes belong to the SEED subsystems and recognised in none of the strains were excluded (fig. 3.6). It should be noted that, as reported before (63), the amino acid biosynthetic pathways show high conserved at species level. Indeed, strains are sharing the same panel of genes whilst two genes, namely Cystathionine gamma-lyase (EC 4.4.1.1) and Cystathionine beta-synthase (EC 4.2.1.22), are absents in M17PTZA496, MTH17CL396 and TH1477. These genes, in two consecutive reactions, transform L-serine into L-cysteine. The first enzyme plays a role in the aroma formation (142) and therefore was well studied in relationship of the cheese-making process. There are two main pathways expressing the major variations, namely ‘Cysteine Biosynthesis’ and ‘Branched-Chain Amino Acid Biosynthesis’. The former system leads to the formation of cysteine, an important precursor of antioxidant compounds, which are key components of the regulation of cell metabolism. The latter was demonstrated to be essential for the correct growth of *S. thermophilus* in milk (143). In addition, both the systems participate to flavour compounds formation.

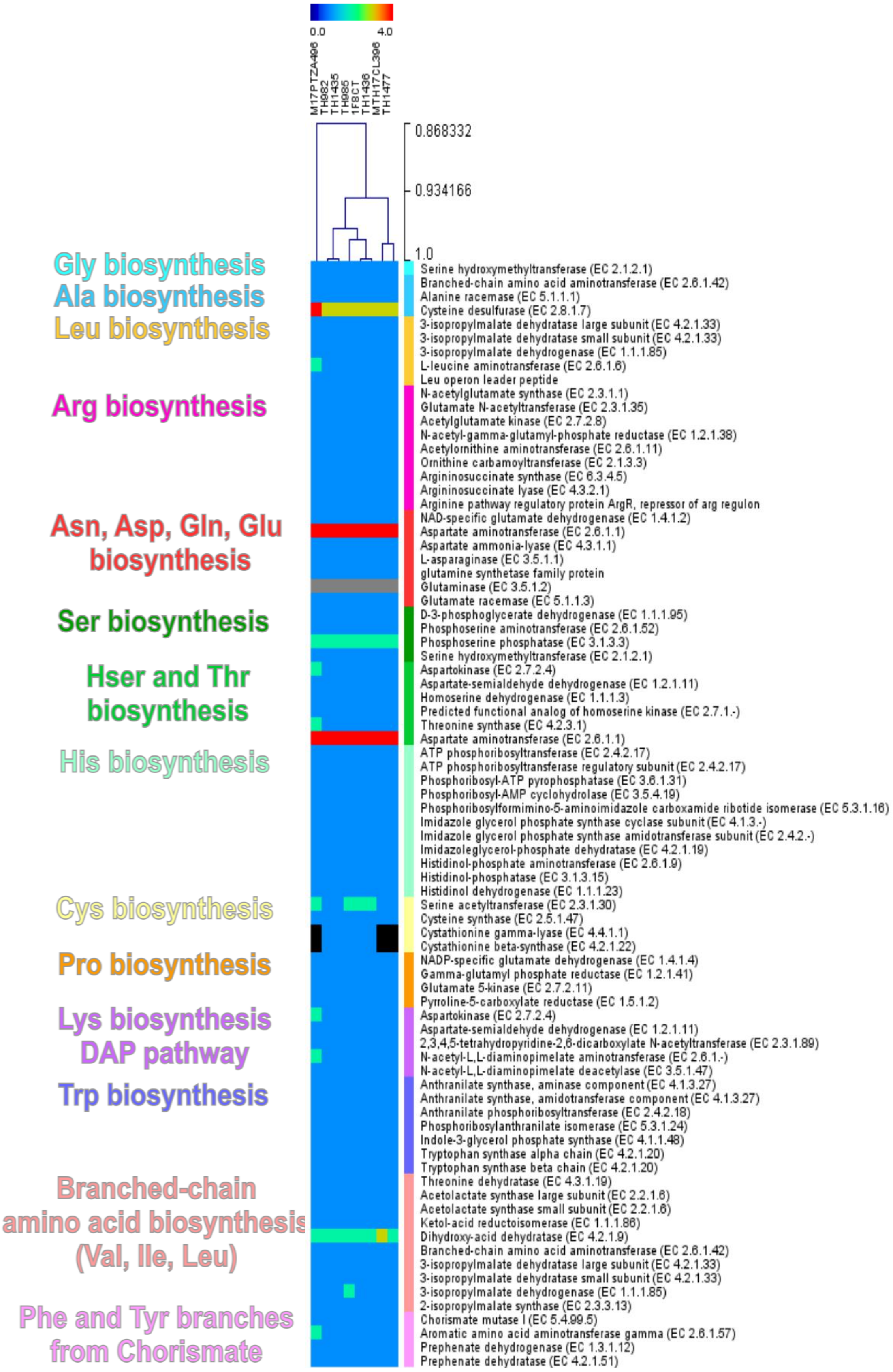
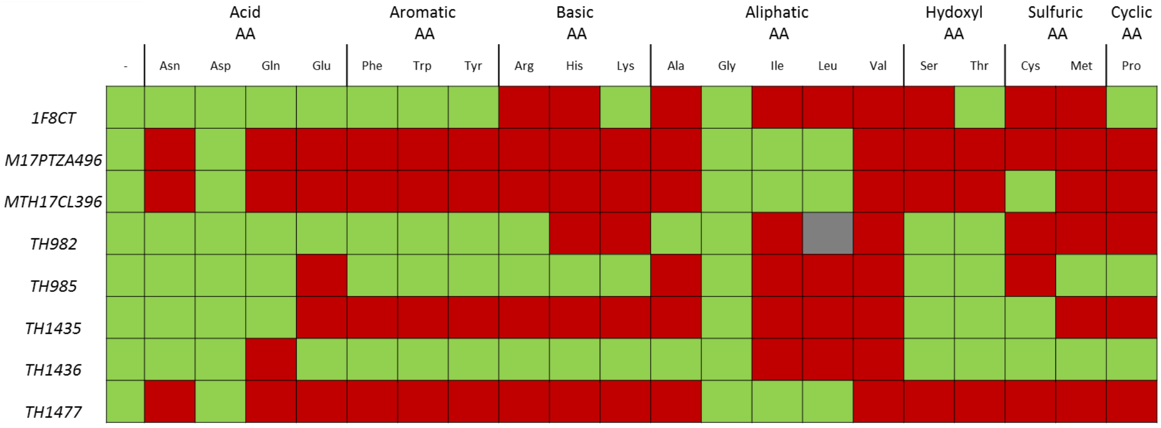


Figure 3. 6 Gene found to belong at *S.thermophilus* amino acid biosystems are present dived in the subsystems. Genes lacking are highlighted in black.

1.9.10 Amino acids biosystem expression

Since 1993 the amino acidic metabolism was well genetically described in LAB (144) and the species amino acidic requirements explored (65). In addition, interesting links between amino acids and technological properties were reported in starter cultures (145). Nevertheless, the phenotypic heterogeneity is not clearly explained from the genetic data and the mechanism regulating the biosynthesis of amino acids has been a poorly explored in this species.

Strains were grown in CDM (130) lacking one amino acid. In this way, it was determined whether, in case of amino acid absence, the biosynthetic pathway and exchange systems sustained growth requirements.



		Acid AA				Aromatic AA			Basic AA			Aliphatic AA					Hydoxyl AA		Sulfuric AA		Cyclic AA
	-	Asn	Asp	Gln	Glu	Phe	Trp	Tyr	Arg	His	Lys	Ala	Gly	Ile	Leu	Val	Ser	Thr	Cys	Met	Pro
1F8CT	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟥	🟥	🟩	🟥	🟩	🟥	🟥	🟥	🟥	🟩	🟥	🟥	🟩
M17PTZA496	🟩	🟥	🟩	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟩	🟩	🟩	🟥	🟥	🟥	🟥	🟥	🟥
MTH17CL396	🟩	🟥	🟩	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟩	🟩	🟩	🟥	🟥	🟥	🟩	🟥	🟥
TH982	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟥	🟥	🟩	🟩	🟥	▨	🟥	🟩	🟩	🟥	🟥	🟥
TH985	🟩	🟩	🟩	🟩	🟥	🟩	🟩	🟩	🟩	🟩	🟩	🟥	🟩	🟥	🟥	🟥	🟩	🟩	🟥	🟩	🟩
TH1435	🟩	🟩	🟩	🟩	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟩	🟥	🟥	🟥	🟩	🟩	🟩	🟥	🟥
TH1436	🟩	🟩	🟩	🟥	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟥	🟥	🟥	🟩	🟩	🟩	🟩	🟩
TH1477	🟩	🟥	🟩	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟩	🟩	🟩	🟥	🟥	🟥	🟥	🟥	🟥

Figure 3. 7 Summary of grow results. On top are expressed the omitted amino acid. Colours results were attributed based on the number of replication which showed a positive grow result. Green cells: at least 75% of curves were positive, yellow cell: 50% of grows were positive, red cells: at least 75% of curves were negative.

Phenotypic data show important differences among strains. Only whether TH982 could survive in absence of leucine was not determined. When strains were evaluated by the percentage of amino acid auxotrophy, it was noted that two strains M17PTZA496 and TH1477 are more exigent than the others, because of their inability to grow in 75% of the cases. The average of strains could grow without 25 to 75% of amino acids, while only one strain, TH1436, proliferated in more than 75% of one-amino acid omitted tests (actually 80%). Looking at the genetic information (see paragraph 3.2.9), it should be noted that the deep dependence of M17PTZA496 from amino acids was unexpected, taking into account its higher number of genes belonging to biosynthesis metabolism. The comparison between TH1436 and TH1477 gene content, which from the clustering resulted similar, has revealed a major change involving the twin isoleucine and valine pathways. TH1477 is probably able to derive these molecules from pyruvate precursor while TH1436 is not. It can be supposed that a fatal change occurred in acetolactate synthase (EC 2.2.1.6) because this is the only enzyme common of all the three biosynthetic pathways (table S4). Moreover, in TH1477 two enzymes could be responsible for the impossibility to convert glutamate into glutamine, namely NAD-specific glutamate dehydrogenase (EC 1.4.1.2) and glutamate racemase (EC 5.1.1.3).

All the positive tests are almost equally distributed among amino acid groups (strain positive response varies from 33% to 63%, for basic amino acids and acid ones respectively). Strains show high consistency in the response to some amino acid omission. Test results show that aspartic acid and glycine are not required from these strains while valine is essential.

In previous findings, different strains required aspartic acid, cysteine, histidine, isoleucine, methionine, leucine, tryptophan and valine (65) in single-amino-acid omission tests. Letort et colleagues (130) observed that two branched amino acids, leucine and valine, are essential, as previously discovered in *S. thermophilus* italian strains (65). These findings were partially confirmed from this new data, indeed valine is essential also in the new analysed strains. Variable requirements were recorded for the other amino acids.

Whey valorisation by vitamin production

Micronutrient malnutrition (MNM) is widespread in the industrialized nations, and even more in the developing regions of the world. It can affect all age groups, but young children, women of reproductive age and elderlies are the bracket most at risk for developing micronutrient deficiencies. The scale and impact of these deficiencies is difficult to quantify, even if an attempt is made from the governments promoting regular reports which quantify the cost of malnutrition diseases in terms of loss of life quality and medical costs (146). In the developed countries, greater access to a wider variety of micronutrient-rich or fortified foods and better health services are factors that contribute to lower risks and prevalence of MNM (73). Several study has provides evidences that vitamin fortification strategy could improve the public health also in Europe. Mainly, food fortification is based on two types of products, namely cereals and the milk derived. Traditionally the supplementation of vitamin or minerals is occurred adding synthetic compound to the final product, obtained from industrial processes. However natural enriched products have gained more attention in the past years. Several studies have demonstrated that natural vitamin could supply more suitable characteristics from the health point of view. It was well established, for example, in the case of folate. It is known, in fact, that the chemical forms of this molecule interact differently with the gastrointestinal tract (GI). It was observed a difference in plasma absorption kinetics: food folate (mainly constitute from 5-methyltetrahydrofolate and 5-formyltetrahydrofolate) is absorbed in the gut and pass through the epithetic cells to the bloodstream, while the synthetic form, folic acid, is metabolised mainly in the liver, causing an precocious saturation of in loco system causing the passage of great amount of vitamin to the blood. This origins two negative effects: (i) the conventional method for the identification of folate deficiency could overestimate the quantities of metabolized vitamin and (ii) other pathogenic states could be masked (75).

The bioavailability of food folate is estimated being 50% of folic acid one, but there are controversies. In a recent work, indeed, it was established that the bioavailability of food folate vary between 30% and 98% depending on the physiological state of the subject and on the way of providing. It was recorded that it adsorption rate putatively can achieve 86% when supplies as it milk natural form, compared with 76% of availability for the folic acid in fortified bread and 44% that was registered for spinach (147). This finding has increased the attention given to the milk-derived products. Milk cannot be considered a rich source of folate if compared with some vegetables. However, it was demonstrated that fermentation could enhance significantly the vitamin amount. It was described that certain yogurts contain more than five-folds the folate concentration compared to milk (148). Within LAB, *S. thermophilus* is widely recognised as one of the best folate producers. It should be notice that some auxotrophic bacteria, including many LAB, consume vitamins available in the medium (148). *S. thermophilus* displays, in average, much attitude to produce than to consume folate although this metabolism depends strongly on strain, fermentation time point and cultivation conditions. By application of different strains in yogurt making, the folate content varies from 20 to 160 ng/g(w/w) (149). The potentiality of converging costumer habits and advanced product formulation toward a healthy diet was displayed recently. Folate-rich fermented milks have shown to significantly increase the hemoglobin level in human blood. Hence, in the same work, high-folate–producing S*. thermophilus* strains demonstrated how they could play an important role in the novel approach aims to fortify naturally the food products (150). Several works has attempted to identify great vitamin producer in its natural environment, milk, in order to obtain an end product which can easily enter the market. A strong strain selection (151) and combination of producer strains (152) strategies were purposed. However, one of the first work has individuate in the mutualistic relationship between *S. thermophilus* and *L. delbrueckii* subsp*. bulgaricus* a negative factor for the final vitamin amount in the fermented milk due to the *L. delbrueckii* subsp*. bulgaricus* vitamin requirements during the growth (148). 0ther dairy products were hypothesized as carbon source for the bacterial metabolism. In fact, the recovery of whey permeates as suitable fermentation matrix for folate production was supposed in a tiny effort, adding this component to an already complete medium. A broad screening over different bacterial species were performed based on synthetic media enriched with permeate (153) and cereal flours, namely corn, wheat and barley (78).

Fermentation set-up strongly influences bacteria growth rate and, thus, folate production. It should be remember that folate is produced by bacteria because of its essential role in the nucleotides biosynthesis. Therefore it could be easily recognised that exceeding vitamin in the medium is determined from both the grow capability of the strains and its metabolic overproduction. Controlled culture condition are defined both chemical and physics parameters. A details study has demonstrated how an increase in folate production occurred in pH-controlled batch fermentations with excess of glucose or assuring the optimal growth condition in terms of oxygen control (149).

Chemicals compounds available in the medium are definitely influencing the fermentation outcomes. It was recently demonstrated that the optimized synthetic medium for *S. thermophilus*, i.e. modified M17 broth, is constituted from 3 g/l lactose, 20 g/l yeast extract (154). Either way, food biotechnology has represented a step forward in the product development. The biosynthetic pathway of this vitamin was deeply studied in plants (155) and it was recognised as conserved during the evolutionary process by the comparison with a different organisms (156). The study of folate metabolism in bacteria has permit to identify a precursor, *para*- aminobenzoic acid (PABA), which has shown a great capacity to improve folate production (149). This precursor is synthesized via glycolysis in the pentose phosphate pathway and shikimate pathway (fig 4.1). The shikimic acid pathway proceeds in 7 catalytic steps and combines carbohydrate metabolism with synthesis of aromatic amino acids. Its first step is a condensation of phosphoenol pyruvate (PEP) with erythrose-4-phospahte (E4P), which results in the formation of 3-deoxy-D-arabino-heptulosonate-7-phosphate. PEP is a metabolite originating from glycolysis, whereas E4P is an intermediate compound of the pentose phosphate pathway. This pathway ends with the synthesis of chorismate, which is further used in the synthesis of L-tryptophan, L-tyrosine, L-phenylalanine, and p-aminobenzoic acid. In *L. lactis*, folate production has shown to be influenced by the concentration of PABA in the medium, indeed its addition to minimal medium lacking aromatic amino acids, purines, and folate resulted in a two-fold increase of folate production (Fig. 3). Concentrations of PABA above 100 μM did not result in a further folate increase.

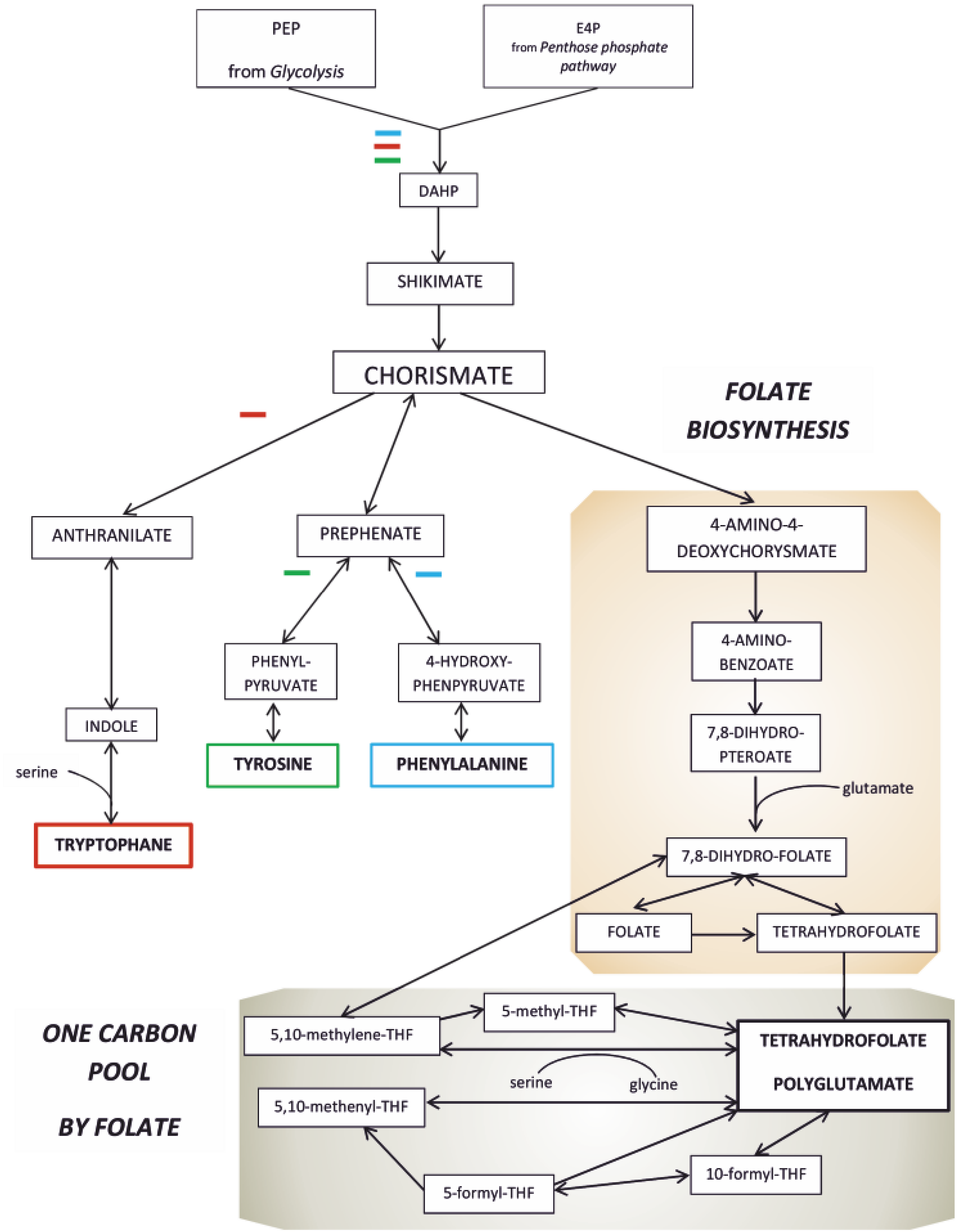


Figure 4. 1 Schematic representation of the folate biosynthetic pathway and one-carbon pool metabolism with their upstream pathways. Three amino acids are highlighted by colour and their sites of negative regulation are marked. PEP: phosphoenolpyruvate, E4P: erythrose-4-phosphate, DAHP 3-deoxy-D-arabinoheptulosonate-7-phosphate

It is known that, beside the direct effect of precursor on its productivity, other mechanisms are influenced its state, mainly related to the regulation of its upstream pathway. Several amino acids take part of the folate metabolism acting at different level of the interconversion reaction. In particular, three of them, namely the aromatic amino acids, have a role in the negative feedback regulation of upstream pathway because also their biosynthetic pathways derive from that metabolism.

While the regulation mechanism was well reported in plants, a few is known about intertwine between aromatic amino acids and this vitamin synthesis in bacteria. It was registered that folate production in *L. lactis* decreased two-fold by the addition of tyrosine (1.2 mM) to the synthetic medium in (149). Because of the differences in the genetic systems encoding for this metabolism between *L. lactis* and *S. thermophilus*, a complete study on aromatic amino acids and folate bonds in the latter species is presented in this chapter.

Biotechnologies allowed the investigation of other potential involvement of bacteria in vitamin food fortification scenarios. Concerning riboflavin, it was deeply explored the genetic mechanism lead to the selection, generation by generation, of riboflavin producing bacteria (157). The genetic discovery relative to the encoding system for this property was used mainly to engineer strains toward its overproduction. Even if some studies were carried out on food bacteria species (158), the general approach follow for its fortification is its industrial production for food and feed additive (159). Selection of strains by isolation of roseoflavin-resistant bacteria was applied only for a little number of application actually, mainly in the cereal-derived products, namely for improve this vitamin in bread (160), or in soymilk (161).

1.10 Whey characterisation

The composition of whey can widely vary depending on the processing techniques used. Because of this diversity in components and attempting to converge toward a potential application the final fermented product proposed, it was chosen to use commercial edible sweet whey (Lactalis, Laval, France) to perform the experiments. For the goals of the study, it was required additional information on its amino acid composition, and these parameters were analysed by HPLC Thermoscientific 3000 ultimate by an in house tuning protocol.

Amino acid	Quantity (% of dry mass)	Amino acid	Quantity (% of dry mass)
Ala	0.27 (0.003)	Lys	0.28 (0.001
Arg	0.28 (0.016)	Met	0.07 (0.014)
Asp	0.58 (0.038)	Phe	0.11 (0.001)
Cys	0.44 (0.192)	Pro	0.39 (0.036)
Glu	0.73 (0.021)	Ser	0.24 (0.015)
Gly	0.14 (0.004)	Thr	0.20 (0.018)
His	0.07 (0.006)	Tyr	0.11 (0.003)
Ile	0.12 (0.001)	Val	0.13 (0.012)
Leu	0.36 (0.001)		

Table 4. 1 Results of amino acid content of commercial whey. Quantities are expressed as percentage of amino acid for unit of dried mass. Means and SD (in brackets) are reported.

1.11 Material and methods

1.11.1 Bioinformatics analysis

To have a clear overview on the entire pathways which lead to vitamin production, the completeness of metabolic clusters were visual analysed by 'Compare Metabolic Reconstruction' tool of RAST, which permit to compute the comparison between the newly annotated genomes. Results were compared within new sequenced strains and against two reference strains, CNRZ1066 and JIM 8232. Further analyses were carried out on folate biosynthesis metabolism. In particular, map reconstruction of close related metabolisms were performed. When was analysed the chorismate pathway, a key enzyme was detected as lacking, i.e. chorismate mutase EC 5.4.99.5. Its presence in all the genomes was confirmed seeking the amino acidic sequence identity with the reference CNRZ1066 sequence (UniProt reference n. Q5LZH3) by mean of blastp. Genes involved in the *fol* operon were compared by alignment in MEGA, following the method previously described (paragraph 3.2.3).

1.11.2 Riboflavin screening test

Similar to that was previously reported in literature (151), this test was conducted starting from a an overnight growth in 10 ml of M17L at 37°C. Cultures were centrifuged and the supernatant liquid discharged. After that, samples were washed three times with sterile sodium chloride 0.85% (w/v), they were suspended in the same buffer and inoculate 10% (v/v) in the Riboflavin Assay Medium. Riboflavin Assay Medium (Difco, Leeuwarden, The Netherlands) is a culture medium free from riboflavin which contains all the other essential nutrients and vitamins (see 'Media and solution'). The qualitative analysis was based on the determination of raising turbidity against the negative control. This assay is particularly indicated for low amount of vitamin, ranged between 0.025 and 0.15 μg of riboflavin. Each test was repeated three times.

1.11.3 Fermentation in synthetic media

As already described, the cultural condition could strongly affect the overall metabolic expression and thus the amount of vitamin produced. To estimate the ability of analysed strains in the folate production, as first it was chosen to checked their capability in the standard condition of growth (149). Overnight grown culture were washed twice with PBS buffer (pH= 7.5), the optical density read at 600 nm in 200 μl of microplate wells (153) and adjusted to 0.3 before the inocula were transferred (1% v/v) in 10 ml of fresh medium. Also, inocula viability was checked by plate count method on M17L agar. Fermentation was carried out growing bacteria at 37°C in static condition. Several works have reported that the higher quantity of vitamin could be detected at different time points over bacterial fermentations, therefore vitamin production was monitored during a wide temporal interval. Samples were collected after 0, 6, 18 and 24 hours of incubation and then analysed for folate content. Optical density (OD) and pH data were collected for each time points. OD values were measure in 200 μl and recorded by multi-plate reader (TECAN, Männedorf, CH). Fermentation was repeated three times for each strain and data analysed by R.

1.11.4 Folates quantification in synthetic media

Two methods are the most frequently preferred. Most current studies determine food folate concentrations in response to growth of *L. rhamnosus* using a high throughput systems based on 96-well microtiter plates. The microbiological assay has been considered one of the best and most versatile methods for determining food folates. *Lactobacillus rhamnosus* ATCC 7469 (formerly known as *L. casei*) is the most commonly used and most accepted indicator strain for folate analysis of natural products. It responds to natural folate forms, avoiding the detection of its common degradation products*. L. rhamnosus* ATCC 7469 has greater capacity to respond to the glutamyl folate polymers compared to the other indicator organisms; however, its response is limited to short tailed folates, those are characterized from up to three glutamates in their tail, and a sensible much lower response to long forms. Hence, this assay required a treatment with pteroyl-y-glutamyl carboxypeptidase (folate conjugase, EC 3.4.19.9) in order to hydrolyze folate polyglutamates to folates with shorter glutamyl residues. The second method is by HPLC, which today have been refined and can successfully quantify naturally occurring folates. The major advantage of liquid chromatography analysis is its ability to quantify the different folate forms and the main limitation of methods is the need to identify all the known forms of folates in order to be able to quantify the real total folates amount (162). Folates were quantified by using a *L. rhamnosus* microbiological assay (163). The indicator microorganism was stored at -80°C in MRS medium supplemented with 15% glycerol. For use in the assay, *L. rhamnosus* was pre-grown in filter-sterilized Folic Acid Casei Medium (FACM, Difco, Leeuwarden, The Netherlands) supplemented with folate 0.3 μg/l; the culture was grown for 18 hours at 37°C. After that, cultures were cooled down on ice, and 40% cold, sterile glycerol was added. Aliquots were stored at –80 °C until use for folate determination. Samples were collected differently as whole broth or supernatant part after centrifuged 12000 g for 2 minutes and diluted 1:1 with 0.1 M sodium acetate buffer (pH 4.8)–1% ascorbic acid and boiled at 100°C for 5 min (149). Total folate concentration, including polyglutamyl folate, was determined after enzymatic deconjugation, which was performed adding deconjugation solution to samples and keeping them in dark condition for 4 hours at 37°C. The deconjugation solution was prepared adding 1 g of human plasma (Sigma-Aldrich Chemie, Zwijndrecht, The Netherlands) at 5 ml of 0.1 M 2-mercaptoethanol-0.1M sodium acetate buffer-1% ascorbic acid, the solution was clarified by centrifugation (10000 g, 2 min). This solution 2.5% (v/v) concentrated was added to the samples and incubated. The fraction of folates with short polyglutamyl tail was analysed following the same protocol but omitting the deconjugase step. To verify the good practice of assay, both folic acid standard solutions and M17L broth were analysed together with samples as positive control, the former for ensure the great indicator strain growth and the latter for check the human plasma enzymes efficiency. After incubation, samples were boiled at 95 °C for 5 min in order to inactivate the human plasma enzymes. Samples were transferred in new plates, where was added 4 iso-volumes of working buffer containing 0.1 M potassium phosphate buffer with sodium acetate buffer (pH=4.8)- 1% v/v of ascorbic acid. A series of 1:2 diluted plates were prepared from the first one. Of them, two plates were filled with sterile FACM as control and to provide OD references, while the others were treated as sample plates, hence an iso-volume of FACM 1% (v/v) with indicator strain was added. The growth of indicator strain in the working plates was determined by measuring the absorbance at 620 nm using the microplate reader (Universal Microplate Spectrophotometer, MQX200R PowerWave XS; Witec AG, Littau, Switzerland) after dark incubation at 37°C for 18h.

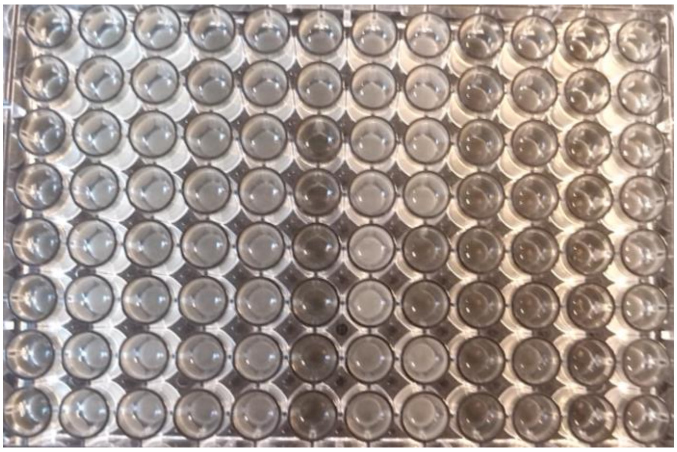


Figure 4. 2 Example of plate obtained after complete growth of *L. rhamnosus* ACTT 7469. Different samples (with replicates) are distributed on the plate

1.11.5 Whey fermentation

Whey is well known to be a rich source of all the biological components required from bacteria growth. Indeed, it was demonstrate that it could provide energy and building box for the cell construction so well to make *S. thermophilus* achieve up to 8·5 × 10^8 cfu/ml (41). Therefore it was selected as dairy substrate for testing strain folate production. Fermentation system was miniaturized and experiments were carried in 2 ml 96-wells microplates respecting the headspace proportion, and thus oxygen exposure, experimented in the synthetic medium fermentations. Commercial whey powder (Lactalis, France) was suspended in distilled water to obtain a final concentration of 10% w/v and sterilized autoclaving 10 min at 110°C. Overnight grown cultures were washed twice with PBS buffer (pH= 7.5), their optical density read at the wavelength of 600 nm and adjusted to 0.3 before the 1% v/v inoculation of the wells fulfilled with 1.8 ml of whey. Inocula viability was verified by spread plate method on M17L agar plates. Fermentations were conducted growing bacteria at 37°C and 45°C in static condition. After 0, 6, 18 and 24 hours of incubation samples were collected and analysed for folate content. Because of the natural turbidity of whey, only pH data were collected for each time points.

Not inoculated whey was used as negative control to check if pH or natural folate amount in whey varied over the time. Fermentations were repeated three times and data analysed by R.

1.11.6 Folates determination in whey fermentation

Because of the particular condition of fermentation, the folate assay was adjusted. For whey fermentation, 100 μl of samples were collected and transferred into a new 96-well microplate in which 0.1 M sodium acetate buffer–1% ascorbic acid were previously allotted 100 μl per well. After that, microplate was boiled 5 minutes at 95°C and centrifuged for 10 min at 5000 g. The supernatant was transferred into new working plates. For this analysis only the total folate quantification was performed, using the enzymatic deconjugation protocol as described above. After that, working microplate was boiled at 95 °C for 5 min to inactivate human plasma enzymes and centrifuged. Samples were moved into a new plate with 4 iso-volumes of working buffer containing potassium phosphate- sodium acetate buffer (pH=4.8) with 1% v/v of ascorbic acid. A series of dilution were prepared from the first one. Two plates were filled with sterile FACM as control and the others were treated filling them with an iso-volume of FACM 1% (v/v) and indicator strain.

1.11.7 Fermentation in amino acids enriched whey and folates detection

To verify the relationship between aromatic amino acid and folate synthesis, fermentations in whey were performed also changing the amount of amino acids. Three aromatic amino acids and glycine, used as control, were added in a large excess to whey. Amino acids stock solutions 100mM were prepared a part dissolving glycine, phenylalanine, tyrosine and tryptophan (Sigma-Aldrich Chemie, Zwijndrecht, The Netherlands) in distilled water and sterilized by filtration. These solutions were added to distinct whey in order to achieve final concentration of 10mM, 20mM and 50 mM, the concentrations were chosen to overcome the basic amino acid content. Whey was prepared as described above and fermentation carried on using the 2ml microplate assay. Total folate was determined using the modified assay tuning during this work.

1.12 Results and discussion

1.12.1 Bioinformatic analyses

Systematic analyses of metabolism in microorganisms are extremely useful to understand their potential applications. Several tools were developed recently with the aim to allow the exploration of the biochemical reaction networks that underlies cellular processes by reconstruction at genomic scale. The reconstruction process is organism-specific and is based on annotated genome sequences. RAST supplies an automatic metabolic model which ascribes each annotated function to its own pathway. This identifies when a panel of genes cold be considered sufficient to support an active variant of the subsystem. This method was chosen in order to estimate the degree of completeness of the vitamin biosynthetic pathways. None of the pathways has shown significant variations among the strains and in comparison to the reference strains CNRZ1066 and JIM 8232. Two vitamins were considered particularly interesting, namely riboflavin and folate. Both vitamin genetic systems were previously detected in other LAB species, and both well described in *L.* lactis. Nevertheless, for riboflavin the main reference organism is *Bacillus subtilis*, which is used for the industrial-scale production of this vitamin. In this species, riboflavin biosynthetic genes are organized in a single operon (159). This kind of structure was identified also in the folate biosynthetic pathway of *L. lactis,* while in *S. thermophilus* those genes are spread over the genome (164). Further analyses were carried out on folate biosynthesis metabolism. In particular, metabolism reconstruction was used for pathway analyses of closely related metabolisms. The chorismate pathway revealed to lack a key enzyme, chorismate mutase. Its presence was verified by seeking a sequence with high identity to the reference sequence on the genomes. In all the strains, a copy of chorismate mutase was found to have a perfect match against the reference. After that, organization and identity of all the sequences coding for genes involved in the folate biosynthetic pathway were inspected. Gene alignments revealed a high conservation of the sequences, both at nucleotide and amino acid level, all have an identity score higher to 97% at aminoacid level)

1.12.2 *Riboflavin production*

Even if genetic analyses didn't reveal the presence of riboflavin biosynthetic genes, phenotypical test were carried on because of the production of a yellow pigment by one of the studied strains, TH1477. It was possible that specific and highly different genes could encode for this vitamin synthesis, therefore its expression was checked by a screening test in a specific medium. After 24 hours, a slight turbidity was visible in all the samples, therefore they were incubated again at 37°C for 24 hours. The day after, samples were centrifuged to evaluate the amount and colour of pellets. No difference with respect to the control was detected, according to what reported from the manufacturer. Therefore it was concluded that strains are unable to synthetize this vitamin.

1.12.3 *Fermentation in synthetic media*

Microbial growth depends on the available compounds provided in the growth medium. Because of the strong relationship between folate synthesis and the physiological state of the cell, in particular related to the phase of cell division which is high demanding in terms of folate, is extremely important to correlate folate quantity with the state of microbial population growth.

		1F8CT	M17PTZA496	MTH17CL396	TH982	TH985	TH1435	TH1436	TH1477	p-value
t0	OD value	0.09 (0.000)	0.10 (0.000)	0.09 (0.000)	0.09 (0.000)	0.10 (0.000)	0.10 (0.000)	0.10 (0.000)	0.1 (0.001)	n.s
	pH	7.11 (0.006)	7.12 (0.010)	7.10 (0.006)	7.13 (0.001)	7.12 (0.015)	7.12 (0.006)	7.10 (0.006)	7.11 (0.010)	n.s
t6	OD value	0.60 (0.156)	0.58 (0.018)	0.26 (0.004)	0.80 (0.023)	0.69 (0.007)	0.49 (0.030)	0.63 (0.024)	0.52 (0.008)	<0.01*
	pH	6.41 (0.304)	5.88 (0.029)	5.97 (0.010)	6.05 (0.017)	6.21 (0.155)	5.82 (0.048)	6.19 (0.036)	6.01 (0.020)	<0.01$^{¥}$
t18	OD value	0.58 (0.006)	0.81 (0.008)	0.79 (0.009)	0.89 (0.018)	0.86 (0.005)	0.55 (0.009)	0.72 (0.023)	0.78 (0.002)	<0.01
	pH	6.02 (0.023)	5.15 (0.032)	5.84 (0.049)	5.82 (0.011)	5.91 (0.007)	5.19 (0.035)	5.40 (0.003)	5.27 (0.030)	<0.01
t24	OD value	0.57 (0.032)	0.87 (0.022)	0.33 (0.036)	0.92 (0.024)	0.89 (0.008)	0.48 (0.002)	0.70 (0.046)	0.84 (0.009)	<0.01
	pH	6.01 (0.053)	5.31 (0.022)	5.93 (0.010)	5.73 (0.016)	5.85 (0.002)	5.18 (0.024)	5.33 (0.022)	5.27 (0.007)	<0.01$^{¥}$

Table 4. 2 Results of fermentations carried out in M17L. OD$_{600}$ measures and pH are reported. Standard deviation of three replicates is reported in brackets. *: Welch one-way ANOVA, $^{¥}$ Kruskal-Wallis rank sum test, any marks: one-way ANOVA.

It should be noted that not all the strains achieved the same final OD values and as expected, each growth was characterized from a different profile. To compare the datasets, three statistical methods were used depending on the data distribution, namely one-way ANOVA, Welch one-way ANOVA and Kruskal-Wallis rank sum test. Statistical analyses revealed that there are significant differences in all the time point for both pH and absorbance parameters. Nevertheless, in all the cases the curve plateau is achieved after 18 hours of fermentation.

1.12.4 Folate production in synthetic media

Folate production was monitored by four measurements during culture growth in order to describe the trend of its expression in the studied strains. Folates were evaluated by comparing results of short- tailed folates with the total amount of folates, sum of long glutamyl -tail forms and ones. Also, the tendency to export folates outside the cell was evaluated comparing free folates in the supernatant against the whole broth amount after cell break.

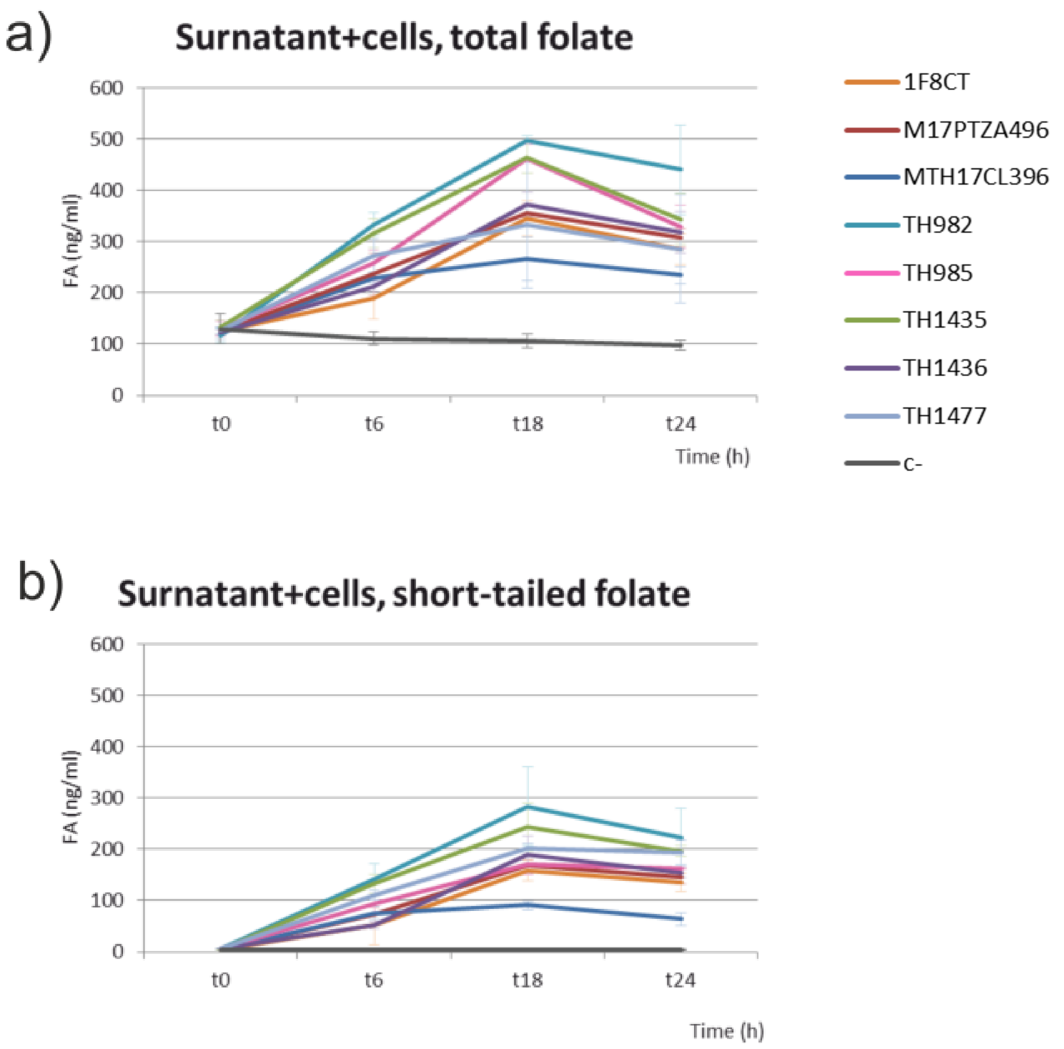


Table 4. 3 Profiles of folate (FA) production detected in the whole cultural media. a) Measures of total amount of folates, b) quantities of short polyglutamyl tail folates. Means value and SD are graphed, strains ID are on the right.

The whole broth inspection revealed that during fermention all the strains improved the amount of folates (fig 4.3). Total folate production ranged from 265 to 498 ng/ml while short-tailed folate between 90 to 296 ng/ml. Statistial analyses show that 'strain' play a significant effect in all the time point when the whole broth is considerd, excluding t0 point (one-way ANOVA, $p<0.01$, Welch one-way AANOVA $p<0.01$ and one-way ANOVA $p= 0.04$ respectively for t6, t18 and t24). Same findings were detected considering folates with a short polyglutamyl tail (one-way ANOVA $p=0.03$, one-way ANOVA $p<0.01$, one-way ANOVA $p<0.01$ for t6, t18 and t24). It should be noted that the higher value of folate was achieved after 18 hours of fermentation, which represents the late exponential phase. After that point, the detected folate decrease probably because consumed form the organism itself. In strain MTH17CL396, folate slightly increased. Four out of eight strains display a total folate production similar to what reported in literature, reaching about 300 ng/ml of folates (149) while three out of eight, namely TH982, TH985 and TH1435 show a higher biosynthetic capability.Concerning the short-tailed folates, strains showed trends similar to that recorded for the total folate over the time, almost all the strain achieved the concentration identified as mean of the species, i.e. 150 ng/ml. Unexpectedly two strains, TH982 and TH1435, exceeded 200 ng/ml.

Analyses of folates released in the medium shown also a significant effect of the ‘strain’ within the time point, t0 was also here the unique exception. Total folate detected ranged between 121 to 334 ng/ml while the short-glutaryl-chain folate ranged between 13 to 135 ng/ml. The comparison of total folate results in the supernatant was performed by one-way ANOVA $p<0.01$, ANOVA $p<0.01$ and one-way ANOVA $p<0.01$ for t6, t18 and t24, while for the comparison of short-tailed folate, ‘strain effect’ was recognised by Kruskal-Wallis $p=0.01$, Kruskal-Wallis = 21.30, $p=0.01$ and Kruskal-Wallis = 21.31, $p<0.01$ for t6, t18 and t24 resepectively. Profile displayed from M17PTZA496 is unusual, after 18 hours it seems to endure a different phenomenon which led to the release of almost all its folate content in the medium, probably due to cell lysis. Similarly, TH982 show an extensive vitamin exporting which, otherwise, is confined in the half amount of the whole folate pruduction (fig 4.3). It is interesting that almost all the folate released outside the cell is conserved during the last part of growth.

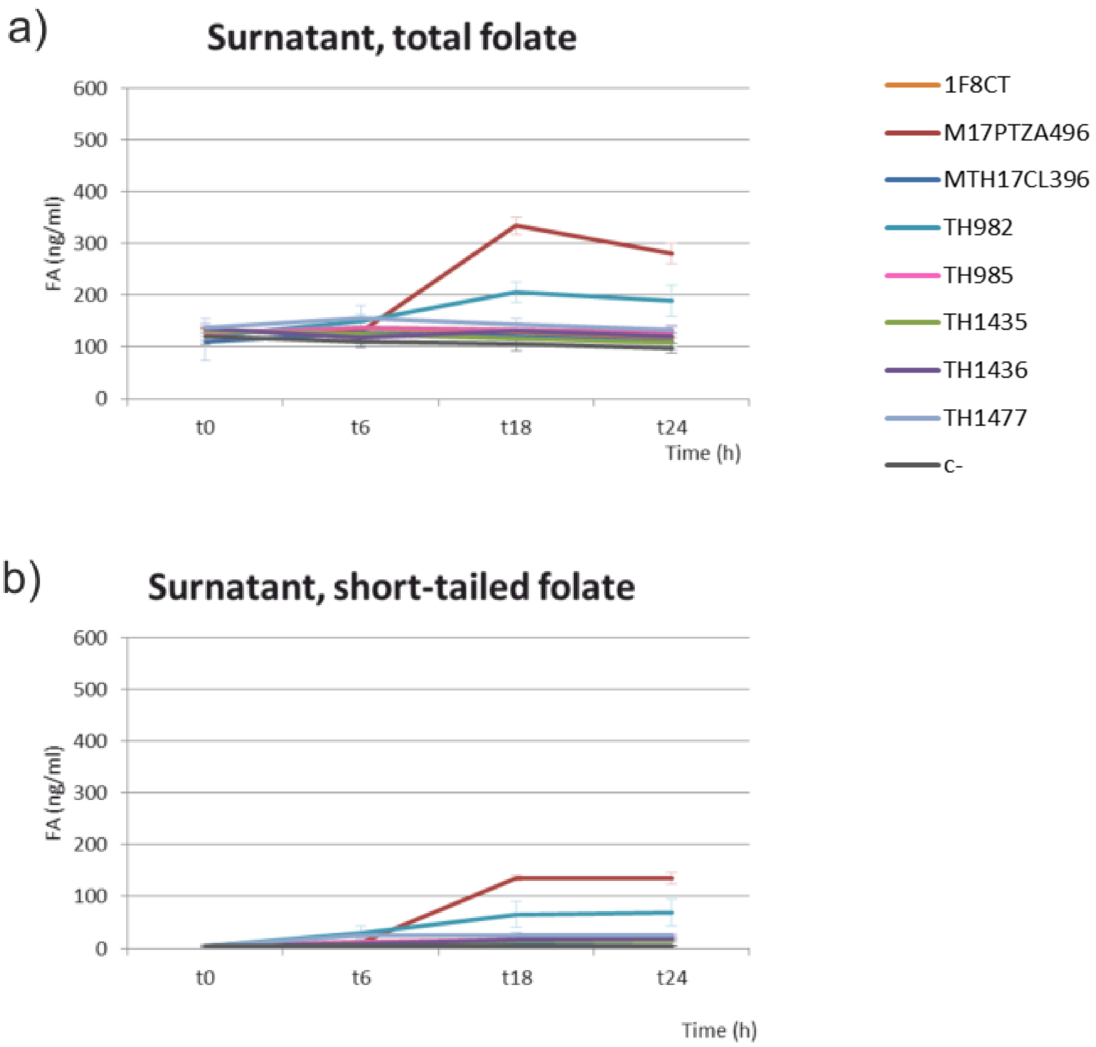


Table 4. 4 Folates (FA) released in the medium during fermentation**.** A) Total folates b) only the part of folate carrying a short glutamyl tail. Means and SD are reported in graphs.

1.12.5 Whey fermentation results

It was previously noticed that strains tested for the folate production in different medium have changed their phenotype significantly (151). It is known that whey represents a good source of nutrient for microbial metabolism, in particular for S. *thermophilus* (41), and therefore it was chosen as substrate to monitor strain fermentation abilities. This experiment was set up taking into account that, for some strains, in literature an improvement in vitamin synthesis was registered when bacteria were grown at higher temperature (152). Therefore, fermentations in whey were carried out at two different temperatures, suitable for the species.

.

		1F8CT	M17PTZA496	MTH17CL396	TH982	TH985	TH1435	TH1436	TH1477	p-value
37°C	t0	6.01 (0.099)	6.01 (0.020)	5.98 (0.020)	5.99 (0.015)	6.04 (0.046)	6.01 (0.025)	5.99 (0.017)	5.98 (0.015)	n.s
	t6	5.93 (0.018)	5.76 (0.358)	5.76 (0.413)	5.89 (0.511)	5.87 (0.496)	5.58 (0.627)	5.82 (0.55)	5.43 (0.435)	n.s.
	t18	5.56 (0.044)	4.73 (0.001)	4.70 (0.05)	4.85 (0.359)	4.80 (0.418)	4.43 (0.055)	4.52 (0.096)	4.41 (0.051)	0.01¥
	t24	5.33 (0.042)	4.64 (0.012)	4.57 (0.046)	4.49 (0.100)	4.94 (0.457)	4.35 (0.065)	4.37 (0.097)	4.30 (0.046)	<0.01¥
42°C	t0	5.98 (0.014)	6.04 (0.085)	6.02 (0.007)	5.99 (0.028)	6.045 (0.007)	5.98 (0.007)	6.02 (0.007)	5.98 (0.007)	n.s.
	t6	5.70 (0.014)	5.35 (0.014)	5.44 (0.043)	6.03 (0.064)	5.99 (0.007)	5.04 (0.077)	5.13 (0.049)	4.92 (0.014)	<0.01
	t18	5.20 (0.001)	4.69 (0.007)	4.64 (0.014)	5.04 (0.056)	5.50 (0.014)	4.40 (0.007)	4.35 (0.007)	4.35 (0.007)	0.04¥
	t24	5.18 (0.007)	4.60 (0.007)	4.45 (0.001)	4.90 (0.028)	5.31 (0.014)	4.27 (0.007)	4.23 (0.001)	4.26 (0.001)	0.04¥

Table 4. 5 Results of whey fermentation. Bacterial activity was evaluated as changing in the pH value over the time. Mean and SD (in parenthesis) are reported. ¥ Kruskal-Wallis rank sum test., any marks: one way ANOVA

Data analyses (tab. 4.5) show that a significant effect due to the strain was recorded with the exception of t0 and t6 points in the fermentations at the lower temperature. Instead, comparing fermentations performed at two different temperatures it is clear that higher temperatures enhance the metabolism rate leading to reduce the time required to lower the pH. From the results it could be seen how, also in these cases, fermentations achieved their plateau after about 18 hours.

1.12.6 Folate detection in fermented whey

Folate amounts were measured during the fermentation on whey, at four time points. Results were compared among fermentations carried out at two different temperatures (tab 4.6).

		1F8CT	M17PTZA496	MTH17CL396	TH982	TH985	TH1435	TH1436	TH1477	p-value
37°C	t0	13 (4)	12 (4)	13 (3)	13 (4)	14 (4)	14 (5)	14 (5)	14 (4)	n.s.
	t6	19 (6)	12 (3)	19 (11)	17 (12)	18 (8)	16 (2)	16 (3)	26 (1)	n.s.
	t18	17 (4)	11 (3)	16 (3)	12 (4)	14 (2)	11 (1)	15 (1)	20 (2)	n.s¥.
	t24	13 (2)	10 (2)	13 (3)	11 (3)	19 (7)	13 (2)	20 (2)	30 (3)	0.03¥
42°C	t0	11 (1)	11 (0)	11 (0)	10 (1)	12 (0)	11 (1)	12 (1)	12 (0)	n.s.
	t6	15 (0)	11 (2)	13 (0)	10 (1)	13 (3)	16 (2)	18 (1)	25 (0)	>0.01
	t18	15 (1)	9 (2)	14 (1)	9 (1)	14 (2)	10 (1)	15 (1)	21 (3)	>0.01
	t24	15 (2)	9 (1)	14 (1)	9 (2)	14 (2)	10 (1)	15 (2)	21 (1)	>0.01

Table 4. 6 Summary of results on folate production obtained in whey. Means and SD (in brackets) are expressed in ng/ml. ¥ Kruskal-Wallis rank sum test., any marks: one way ANOVA

Total folate production in whey at 37°C ranged between 12 and 30 ng/ml and in whey at 42°C between 10 and 25 ng/ml. Also in these cases, the tested strains didn't consume the vitamin supplied in the medium but not all of them enhance the final amount of vitamin, as it was registered for examples for strain M17PTA496. According with previous findings, the spike of folate production was recorded early in this medium, and as expected it did not achieve the same absolute values recorded for the synthetic medium.

Comparison of vitamin production among strains for each time point revealed that at 37°C only after 24 hours strains behaviour differ. Unexpectedly (152), raising the temperature didn't increase this phenomenon, which in average achieved the best results after 6 hours at 37°C. On the contrary, the major increase was obtained after 24 hours using TH1477, which permitted to increase folate amount by two fold with respect to the starting point at 37°C. This strain differs from the other because it didn't decrease vitamin concentration after 6 hours. The overall comparison of temperatures revealed that temperature affects dynamics in folate production but not its final yield. Results derived from whey fermentation showed how the whey can well support bacterial growth with a wide increase in the final concentration of folate, similarly to what previously reported for milk (165)

1.12.7 Folate production in amino acid enriched media

Several factors were supposed to affect folate synthesis rate in *S. thermophilus,* precursors of the biosynthetic pathway (154) and carbohydrates with prebiotic effects (166). In *L. lactis*, a putative effect was attributed to different amino acids, showing an interesting opposite response for tyrosine (149) and glutamine addition (165). One of the main changes between synthetic enriched media and natural milk is the availability of amino acids. The studied strains revealed to miss proteinase activity and, in absence of a co-operator species, they can use only the amino acids and peptides free in the medium. In addition, some amino acids share their biosynthetic pathway with folate (the shikimate pathway, see fig 4.1) and are responsible for negative regulation of key steps of this pathway. The Studied strains displayed interesting differences in response to amino acid lacking media (paragraph 3.2.10). From these evidences and from the genetic information available from the most studied species, *L. lactis*, and *S. thermophilus* show a substantial difference in the folate biosynthetic cluster organisation. The effects of aromatic amino acids addition in whey were studied in *S. thermophius,* and to evaluate if these amino acids could enhance folate production significantly, glycine was used as reference. Three different amino acid concentrations were tested for the four amino acids.

Firstly the, effects of amino acids on fermentation were evaluated considering the pH value achieved after 24 hours of growth in different media. A significant effect wasn't found for all the strains, indeed 1F8CT and TH985 did not show susceptibility to amino acid supplementation (S5). In the others, where a significant effect were recorded (M17PTZA496: Kruskal-Wallis $p<0.01$; MTH17CL396: $p<0.01$; TH982: $p<0.01$; TH1435: $p<0.01$; TH1436: $p<0.01$; TH1477: Kruskal-Wallis $p=0.01$), LSD Fisher post-hoc analyses registered in all the strains a significant difference of 50mM tryptophan fermentations compared to the others, which recorded highest pH values. In some cases also the addition of 50mM of glycine, 20mM of phenylalanine and 20mM of tryptophan decreased the acidification.

To simplify the analyses, data expresses the amount of folate produced during fermentation.

Gly (mM)	Time point	1F8CT	M17PTZA496	MTH17CL396	TH982	TH985	TH1435	TH1436	TH1477	p-value
10	t6	-3 (5)	-3 (5)	0 (5)	-3 (8)	1 (1)	0 (2)	1 (5)	3 (5)	n.s.
	t18	-1 (2)	-4 (5)	6 (7)	2 (9)	10 (5)	5 (4)	16 (6)	18 (10)	0.03
	t24	1 (6)	-4 (5)	1 (5)	-3 (5)	7 (7)	4 (6)	10 (9)	28 (15)	0.01
20	t6	-1 (1)	-2 (5)	0 (4)	-1 (4)	-1 (3)	0 (2)	-3 (6)	4 (6)	n.s.
	t18	-1 (8)	-3 (7)	5 (9)	1 (7)	2 (2)	0 (6)	11 (3)	14 (7)	n.s.
	t24	0 (7)	-4 (5)	5 (9)	-1 (7)	3 (2)	-2 (9)	5 (3)	21 (19)	n.s.
50	t6	2 (1)	-1 (5)	3 (4)	-2 (4)	-2 (3)	-2 (2)	-2 (6)	9 (6)	n.s.
	t18	4 (8)	-1 (7)	6 (9)	11 (7)	-1 (4)	-4 (6)	8 (3)	33 (7)	0.02
	t24	1 (7)	-4 (5)	-1 (8)	-2 (7)	-2 (2)	-4 (10)	4 (10)	36 (19)	0.01

Table 4. 7 Folate gained during fermentation in whey enriched with different amount of glycine. Mean and SD, in brackets, are reported in ng/ml. No marks: one-way ANOVA

Comparative strain analyses within condition and time point revealed interesting dissimilar profiles in response to amino acids concentrations. The majority of the analysed strains were not conditioned by the supplementation of glycine (tab. 4.7), for example M17PTZA496 displayed always the lowest value. TH1477 seems to be more influenced by the amino acid, its level of folate production increased when the highest concentration of glycine was added to the medium. Strain TH1436 seems to be conditioned more from low quantities of glycine than from the highest one. This phenomenon was mainly recorded in the late exponential phase of growth.

Phe (mM)	Time point	1F8CT	M17PTZA496	MTH17CL396	TH982	TH985	TH1435	TH1436	TH1477	p-value
10	t6	-1 (4)	-2 (4)	-3 (7)	1 (3)	-3 (5)	-4 (9)	-2 (6)	0 (8)	n.s.
	t18	2 (10)	-2 (5)	0 (4)	4 (6)	-1 (13)	-2 (10)	1 (12)	5 (18)	n.s.
	t24	8 (8)	-1 (5)	0 (4)	4 (4)	5 (6)	-1 (6)	3 (12)	15 (16)	n.s.
20	t6	5 (2)	7 (9)	3 (7)	10 (8)	2 (1)	1 (3)	1 (5)	6 (1)	n.s.
	t18	11 (5)	10 (6)	10 (8)	35 (24)	10 (5)	1 (5)	9 (2)	27 (7)	0.02
	t24	6 (11)	3 (6)	2 (8)	11 (11)	10 (4)	1 (4)	4 (6)	28 (5)	<0.01
50	t6	4 (3)	2 (1)	4 (3)	14 (12)	0 (3)	-1 (4)	-2 (4)	7 (3)	0.03
	t18	10 (9)	6 (5)	10 (11)	32 (23)	9 (6)	-4 (9)	-2 (7)	26 (4)	0.01
	t24	5 (12)	1 (7)	3 (10)	12 (15)	4 (6)	-4 (7)	-2 (8)	21 (12)	n.s

Table 4. 8 Folate gains during fermentation in different phenylalanine concentration media. Mean and SD, in brackets, are reported in ng/ml. No marks: one-way ANOVA

Analyses on the strains' response to large excesses of phenylalanine showed a general enhancement of folate particularly after 18 hours of growth with 20mM of this amino acid. The exception of TH1435 suggests that in this strain the regulation has modified its susceptibility to this compound. On the contrary, TH982, which previously displayed marked properties in this task fermenting the synthetic medium, revealed to be the most responsive strain to phenylalanine concentration. In its case, amino acids seems to improve the folate biosynthesis during the reproductive phase while, in the latest one, the vitamin drastically reduced, probably due to bacterial consumption. Also in this case TH1477 displayed a good performance and being conditioned from the amino acid concentrations, achieving its higher production in 20mM phenylalanine supplemented whey.

Tyr (mM)	Time point	1F8CT	M17PTZA496	MTH17CL396	TH982	TH985	TH1435	TH1436	TH1477	p-value
10	t6	1 (4)	1 (1)	3 (1)	0 (1)	3 (2)	11 (4)	1 (1)	21 (6)	<0.01
	t18	4 (3)	3 (2)	8 (3)	7 (2)	13 (1)	6 (3)	13 (4)	26 (12)	<0.01
	t24	5 (4)	2 (5)	9 (3)	4 (1)	12 (2)	12 (1)	9 (1)	29 (12)	<0.01$^{¥}$
20	t6	1 (2)	2 (2)	6 (3)	1 (1)	3 (3)	7 (1)	3 (1)	20 (4)	<0.01
	t18	4 (5)	6 (2)	7 (3)	8 (3)	7 (2)	8 (2)	12 (2)	20 (10)	0.01
	t24	7 (5)	4 (2)	9 (2)	10 (4)	10 (4)	13 (3)	11 (1)	27 (2)	<0.01
50	t6	3 (2)	2 (2)	3 (3)	1 (1)	4 (3)	11 (1)	3 (1)	20 (4)	<0.01
	t18	3 (4)	5 (2)	8 (3)	10 (2)	13 (3)	10 (2)	15 (2)	24 (10)	<0.01
	t24	4 (5)	5 (2)	5 (2)	10 (4)	9 (4)	14 (3)	12 (1)	23 (2)	0.01

Table 4. 9 Folate yield during fermentation with different quantities of tyrosine. Mean and SD, in brackets, are reported in ng/ml. $^{¥}$ Kruskal-Wallis rank sum test, no marks: one-way ANOVA

Although the biosynthetic gene cluster and growths in aromatic amino acid omitted CDMs (see paragraph 3.2.10) have witnessed symmetries in the mechanisms responsible for the phenylalanine and tyrosine management, this test demonstrated how the regulation is differently influenced from these amino acids. In fact, while TH1435 was registered as indifferent to phenylalanine presence, when the same strain is grown in condition of tyrosine excess, it shows a greater improvement in vitamin production. Nonetheless other strains react differently in this new condition, also TH982 which dramatically decreased its vitamin synthesis, moving the highest value of gained folate from 35 to 10 ng/ml. TH1477 recorded early vitamin overproductions in all the amino acid concentrations tested. Differently from what found for *L. lactis,* tyrosine did not show a marked inhibitory effect.

Trp (mM)	Time point	1F8CT	M17PTZA496	MTH17CL396	TH982	TH985	TH1435	TH1436	TH1477	p-value
10	t6	2 (0)	1 (5)	2 (5)	1 (2)	-1 (4)	-2 (4)	-2 (6)	1 (8)	n.s.
	t18	6 (4)	3 (7)	10 (8)	12 (7)	3 (10)	0 (10)	0 (10)	11 (20)	n.s.
	t24	8 (1)	3 (3)	9 (4)	9 (5)	15 (11)	5 (10)	5 (10)	16 (20)	n.s.
20	t6	-4 (4)	-3 (5)	-1 (14)	8 (12)	1 (4)	1 (0)	3 (2)	8 (4)	n.s*
	t18	0 (4)	-2 (4)	6 (5)	23 (11)	16 (1)	11 (4)	14 (3)	35 (10)	<0.01
	t24	0 (7)	0 (4)	8 (9)	14 (2)	14 (5)	7 (5)	13 (6)	37 (2)	<0.01
50	t6	-1 (4)	-1 (5)	0 (3)	6 (7)	-3 (5)	0 (2)	1 (3)	9 (3)	0.04
	t18	1 (4)	1 (4)	6 (6)	31 (25)	8 (5)	4 (5)	13 (5)	45 (6)	0.04¥
	t24	5 (9)	4 (9)	17 (2)	25 (23)	1 (6)	3 (8)	8 (5)	43 (12)	<0.01

Table 4. 10 Yield in folate production (in ng/ml) during fermentation with different concentration of tryptophan. Mean and SD, in brackets, are reported. *Welch one-way ANOVA ¥ Kruskal-Wallis rank sum test, no marks: one-way ANOVA

The yield in folate achieved with supplementations of tryptophan led to reach the highest value recorded in the experiments, namely when fermentation was conducted for 18 hours in whey enriched with 50mM of the amino acid and using TH1477. Other strains seem to not be influenced by amino acid variation, with the exception of TH982, that seemed to be conditioned from its presence when tryptophan is supplemented in the two highest concentrations. It should be noted that in this experiment, as for the phenylalanine one, while TH982 showed a significant increase in the registered folate during the late exponential phase and a successive decrease, in TH1477 the higher folate synthesis reached its maximum in the same growth phase but the vitamin content remained almost unaltered after.

Transcriptomic analysis of the folate synthesis

The first LAB genome was sequenced in the early 2000s. Ever since, the number of sequenced LAB genomes has grown exponentially and currently genomic data from over 100 LAB species and strains are available in public databases. These offer a wide amount of information, to further understand LAB in respect to their gene content and properties in human health as well as in food fermentations (167). Genetic analyses play an increasing in importance role in the assessment of desired or avoiding not desired effects of food microorganisms. This includes the functional prediction, the creation of genome-scale metabolic models and the pinpoint of complex food properties (168). These approaches are built on a strong knowledge-guided metabolic design which should take in account gene content as well as the expression of genes, which is essential to realize good predictions on new valued phenotypes. Hence, nowadays the tendency is to contextualise the desirable properties in a global metabolic network toward the understanding of the complex interconnection subtending characters expression.

Visualised the influence of the environmental on genetic expression requires the accurate quantification of all the expressed mRNAs. The transcriptome is the complete panel of transcripts in a cell and their quantity, which is determined for a specific developmental stage or physiological condition. Understanding the transcriptome is essential for interpreting the functional elements of the genome and revealing the molecular constituents of cells. The microarray technique provides an effective method for the in parallel analysis of thousands of transcripts, allowing the profiling of a genome transcriptome in a single experiment. Therefore it was so widely used in the past years. However, some technical problems, as examples the background required for setting up the experiment and the low reproducibility of results between laboratories, have limited their usage for transcriptome interpretation. The RNA sequencing (RNA- Seq) is a quite recent developed approach of transcriptome defining. It takes advantage of deep-sequencing technologies. RNA-Seq provides a far more precise measurement of levels of transcripts and allows the recording of new isoforms. Several comparative studies revealed a good correlation between the transcripts measured by microarrays and RNA sequencing. Moreover, these studies favoured RNA sequencing due of its higher reproducibility and higher accuracy of fold change detection in the expression levels (169), supporting the idea that RNA-seq has a much greater dynamic range for measuring variability in expression and, as consequence, that this technique can be more discriminatory at high levels and more sensitive at very low levels of gene expression. High-resolution RNA sequence allows the quantification of variation in expression levels for each transcript, during cultures development or under different growth conditions, and the identification of all the transcriptional features, promoters, terminators and operons among others, on any bacterial transcriptome.

To date, a few RNA-seq projects were occurred on LAB expression. Indeed, a large amount of main scientific questions still be analysed by means of microarrays although this is limiting for the discovery of unpredictable phenotypes. The NGS method was preferred recently for the deep analyse of the regulatory mechanism on growth and the global regulation system, and to clarify the differential process leading to the instauration of the stable microflora, in confront to the transit microbial population, during pathogenic events (167). Only recently the first work of RNA-seq were performed in the effort to explain the mutualist relationship connecting five LAB species in yogurt, one of them was a potential probiotic bacteria (170). For this work, the NGS approach was chosen in order to clarify the hidden mechanism determining the different responses between strains in the folate production tasks. The main object of study was two pathways: shikimate pathway and folate de novo synthesis.

As described in the previous chapter, chorismate is the key element due of its usage for the synthesis of both aromatic amino acids and p-aminobenzoic acid. More in details, the shikimic acid pathway begins with the condensation of PEP with E4P, which results in the formation of 3-deoxy-D-arabino-heptulosonate-7-phosphate. The enzyme, that catalysis this first step, is 3-deoxy-arabino-heptulo- sonate-7-phospahte synthase (DAHP synthase, EC 4.1.2.15). The new formed DAHP loses its phosphoryl group and undergoes to cyclization to 3-dehydroquinone (DHQ) by means of 3-dehydroquinone synthase (DHQ synthase, EC 4.6.1.3). The third enzyme, 3-dehydroquinone dehydratase (EC 4.2.1.10), introduces a double bond to the aromatic ring and, thus, activates the formation of an intermediate, 3-dehydroshikimate (DHS). Through reduction with NADP, the DHS is transformed to the shikimic acid (SA), this reaction is catalysed by shikimate dehydrogenase (EC 1.1.1.25). The next stage of the shikimic acid pathway is phosphorylation of shikimate to 3-phosphoshikimate, catalysed by shikimate kinase (EC 2.7.1.71). Afterwards, 3-phosphoshikimate is subject to condensation with a second molecule of PEP, producing 5-enolpyruvylshikimate-3-phosphate (EPSP) by means of EPSP synthase (EC 2.5.1.19). The produced intermediate compound loses its phosphoryl group and then is reduced to chorismate. In fact, the final step of the shikimate pathway is the synthesis of chorismate by chorismate synthase (EC 4.2.3.5), further used in the branch of aromatic amino acids and p-aminobenzoic acid synthesis.

The de novo folate synthesis consists of two branches, a pterin and a pABA one. The first enzyme of the pterin branch is GTP cyclohydrolase I (GCHY-I, EC 3.5.4.16), which catalyses a complex reaction in which the imidazole ring of GTP is opened, C8 is expelled as formate, and a six-membered dihydropyrazine ring is formed using C1 and C2 of the ribose moiety of GTP. The resulting 7,8-dihydroneopterin triphosphate is then converted to the corresponding monophosphate by a specific pyrophosphatase. O the contrary, removal of the last phosphate is mediated by a non-specific phosphatase. Then, dihydroneopterin aldolase (DHNA, EC 4.1.2.25) releases glycolaldehyde to produce 6-hydroxymethyl-7,8-dihydropterin. It is pyrophosphorylated by hydroxymethyldihydropterin pyrophosphokinase (HPPK, EC 2.7.6.3). Also, the DHNA interconverts the 7,8-dihydroneopterin and of 7,8- dihydromonapterin, and breaks the latter to 6- hydroxymethyl-7,8-dihydropterin. FolX converts the triphosphates of 7,8-dihydroneopterin and of 7,8-dihydromonapterin. In the pABA branch of the pathway, chorismate is aminated to aminodeoxychorismate (ADC) by ADC synthase (EC 6.3.5.8) using the amide group of glutamine as donor. ADC is then converted to pABA by ADC lyase (EC 4.1.3.38).

The outcomes of these processes, namely 6-hydroxymethyl-7,8-dihydropterin pyrophosphate and pABA moieties, are condensed by dihydropteroate synthase (DHPS, EC 2.5.1.15). The resulting dihydropteroate is glutamylated by dihydrofolate synthase (DHFS, EC 6.3.2.12) giving dihydrofolate (DHF), which is reduced by dihydrofolate reductase (DHFR, EC 1.5.1.3) to tetrahydrofolate (THF). Folylpolyglutamate synthase (FPGS, EC 6.3.2.17) then adds a γ-glutamyl tail. These metabolisms were deeply studied by the new sequencing technology for the comparison of two strains, TH1436 which represent the reference for species behaviour, and TH1477, which has showed high attitudes to produce folate in fermented whey.

1.13 Material and Methods

1.13.1 Fermentation and RNA extraction

Gene expression of two selected strains, TH1436 and TH1477, were compared in three fermentations set up, namely fermentation of M17L broth (Oxoid, Rodano, IT), whey (Lactalis, Laval, France) and whey with 50mM of tryptophan (Oxoid, Rodano, IT) in three replicates. Fermentations were conducted for 18 hours at 37°C in 100ml flasks respecting the headspace proportion used previously in the micro-fermentations (paragraph 4.2.5) before collected the samples. Cells were harvested by gently centrifugation, than the supernatant discharged and pellets were immediately frozen in liquid nitrogen. Samples were stored at -80°C until lysis and RNA extraction.

Cell lysis was obtained by successively steps. First, 100 µl of lysozyme solution (10 mM Tris-HCl, 0.1 mM EDTA, 10 mg/ml lysozyme, pH 8.0) were added to the samples, those were then mixed to assure the complete resuspension of the pellet. Then 0.5 µl of 10% (w/v) SDS were added. Furthermore, 350 µl of fresh prepared 1%(v/v) 2-mercaptoethanol lysis buffer, 5 ml of TRIzol (Invitrogen, Rodano, IT) and chloroform were added together to the suspension and mixed well with cold beads for an overall of 10 min. Samples were centrifuge at 12,000g for 20 minutes at 4°C. The supernatant was then collected and lysis mixture residuals were eliminated by a supplementary washing step with an iso-volume of chloroform. Afterward, RNA extractions were performed following the manufacturer's protocol using the Purelink RNA minikit (Invitrogen, Rodano, IT). Total RNA was finally treated with DNaseI (Qiagen, Milano, IT) and the suspended in RNase-free water. Samples were kept at -80°C until the successive processing.

1.13.2 mRNA enrichment

Since the prokaryote RNA pool contains a large amount of rRNA and tRNA, which may constitute more than 95% of the total RNA and can dramatically reduce the signal of mRNA in the sequencing process, they should be subtracted in a specific sample preparation step which involved the hybridization captures of 16S and 23S rRNAs. It was performed using the MICROBExpress kit (Ambion, Rodano, IT) following the supplier's instruction.

1.13.3 Extraction quality control

Before the RNA enrichment, RNA extraction were verified by means of the NanoDrop (ThermoFisher Scientific, Waltham, MA, USA) for the purity control and quantification, and by visualization on denaturing gel for assure the RNA integrity. Samples were denatured adding 1.8 ml of 37% of formaldehyde. The denaturing gel was prepared adding 37% formaldehyde to 0.2M MOPS, 50mM sodium acetate- 10mM EDTA solution (pH 7). This denaturing solution was added 10% (v/v) to 1% (w/v) agarose gel just before its cooling. Eurosafe (Euroclone, Milano, IT) was used as fluorescent dye.

After rRNA depletion, samples were analyzed by Agilent Bioanalyzer 2100. The bioanalyzer chip allows the evaluation of both RNA quantity and integrity, individuating also approximatively the RNAs sizes. It was performed at BMR genomics services.

1.13.4 Sequencing and data quality filtering

Extracted samples were sequenced at the Ramaciotti Centre for Gene Function Analysis (University of New South Wales, Sydney, NSW, AU) using the HiSeq Illumina 2000 (Illumina, San Diego, CA, USA) technology. It was chosen the paired-end reads '75+ 75' bases strategy. Libraries were produced using the 'Truseq' kit (Illumina, San Diego, CA, USA), and the RNA insert size was between ~200 bp and 1.5 Kb. Sequence quality check and filtering were performed by CLC Main Workbench 7.6.4 (CLC bio, Waltham, MA, USA), setting as parameters the quality score greater than 0.05 and reads lengths higher than 73 bp.

1.13.5 Read mapping

Read mapping is one of the essential tasks in this analyses and consists in the alignment of reads against a reference genome. Read alignment is a classic problem in bioinformatics, however, in this case it pose particular challenges because reads were short, for this analyses read length ranging between 36 and 125 bases, short reads respect to what is used for the genome sequencing, and error rates in this kind of approach is considerable (171). Due to the fact that reads cannot overlap the entire transcripts, the one from which they were derived cannot be always uniquely determined. In addition, paralogous genes and high similar sequences are the primary barriers contributing to the mapping uncertainty, together with polymorphisms and sequencing errors. Due to these factors, a significant number of reads is multireads, namely is constituted from reads that have high-scoring alignments to multiple positions in a reference genome. Two strategies are commonly used for resolve gene multireads. The first easily discards them, keeping only the uniquely mapping reads for expression estimation. The second strategy permits the recovery of multireads, allocating them in proportion to the calculated coverage of uniquely mapping reads (172). To be more conservative in the analyses results, it was chosen to map only reads with only one best-hit place.

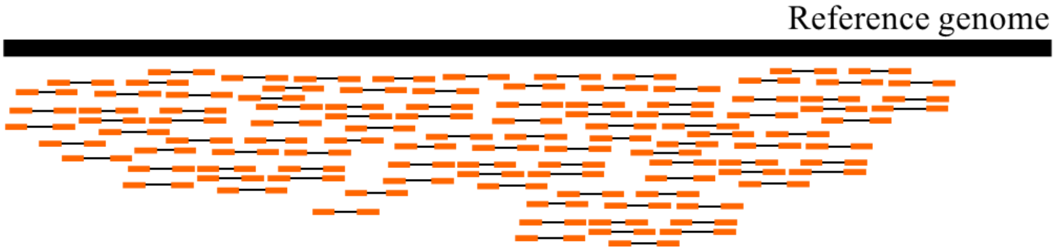


Figure 5. 1 Reads mapping performed using CLC Main Workbench. Raw pair-end reads were placed on the genome reference.

1.13.6 Expression analyses

The total mapped reads per CDS were calculated. Basically, it consists in a count of the times each transcript was sequenced, converting mapped reads to a base count data. Normalization of this base count data is a critical point in the data processing. Normalization leads to work with a relative dataset that permits to compare expression levels within a sample or between different samples. Basically, this transformation allows computing the variation in metabolic expression with more likelihood. It is need because more reads are required to cover the length of a longer gene and if their number is not corrected for the gene length, this bias could generate an expression overestimation. The normalized count data can be quantified by averaging the base count across a selected region of the genome. Since the average of the counts is used, the relative expression of any given transcription feature, independently of its length, can be expressed and compare with all the other (173). Different approaches to apply this correction were supposed, but today the RPKM calculation is widely considered the more suitable: this expression value is calculate as total exon read/ mapped reads for exon length (174).

Genes were annotated using SEED subsystems database, while homologies between the two strains CDS pools were identified by means of RAST genome comparison tool. Comparison of condition within strain and comparison between strains in the same medium were performed calculating fold change and significance, applying the tagwise dispersion using CLC Main Workbench 4.7.6 (CLC bio, Waltham, MA, USA). Deeply analyses data were carried on by costumer prepared R script, matching the expression data lists. Filtering parameters to exclude genes not significantly variant were the p-value >0.05 and an absolute Log_2 fold change value higher than 1.

1.14 Results and Discussion

1.14.1 Extraction protocol efficiently

Eighteen samples of total RNA were extracted independently from S.thermophilus cultures. Approximately 95% of total cellular RNA is constituted by large rRNA molecules that represent a interference during the sequencing. Therefore, this portion of RNA was removed before sequencing. After rRNA subtraction, the quality and quantity of the material were measured and RNA profiles produced by the usage of Bioanalyzer 2100, which returns the output presented in fig. 5.2.

Sequenced RNA molecules had length varying from 50 to 4000 nucleotides. Length distribution showed that the RNA was integer because more than half of the molecules were longer than 400 nucleotides even if a high number of molecules about 150 nucleotides long were detected. Peaks corresponding to 16S and 23S rRNAs are indicated in the graph. The profile shows that after subtraction rRNA contamination was significantly reduced.

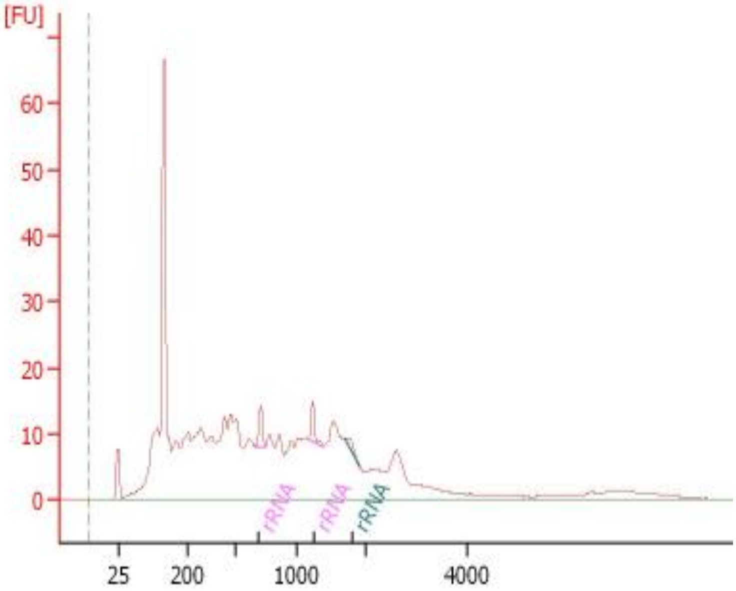


Figure 5. Example of RNA profiles obtained by Bioanalyzer. Length distribution shows that RNA was integer and that contaminating rRNAs were highly reduced (highlighted in pink and dark green).

1.14.2 Sequencing statistics

RNA-seq was performed using the Illumina technology. Sequencing produced an average of 2.0X10^7 paired reads for each sample with a yield of 99.00 % after the first quality filtering. Before assigning the reads to the corresponding genome, CLC Workbench (CLC bio, Waltham, MA, USA) was used for a further filtering to keep only high quality reads. Among total putative aligned reads, it was possible to discharge those that were not uniquely assigned on the genome. Reads uniquely aligned were chosen for computing the expression profiles of samples. Clearly, a dramatic reduction occurred in the number of useful reads during the mapping process, as reported in table 5.1, but visual inspection evidenced that also in the most critical case, i.e. TH1436 in M17L medium, less than 1% of genes were identified as not transcribed at all. Therefore, sequencing results were considered reliable.

Strain ID	Genome size (Mbp)	Fermented medium	No. raw reads	No. aligned reads	% aligned reads	No. filtered reads	% filtered reads
TH1436	1.78	M17L	1.93E+07	7.18E+05	4	1.27E+04	2
		Whey	1.94E+07	2.86E+06	15	4.09E+04	1
		Whey+50mM trp	2.02E+07	3.85E+06	19	6.34E+04	2
TH1477	1.88	M17L	1.59E+07	1.67E+06	12	5.46E+04	3
		Whey	2.02E+07	3.98E+06	20	8.61E+04	2
		Whey+50mM trp	2.12E+07	5.30E+06	25	1.03E+05	1

Table 5. 1 Summary of sequencing statistics and the alignment of the obtained reads to the corresponding genome performed by CLC. Results express the average of three replicated samples.

1.14.3 Gene expression in synthetic growth medium

The two strains analysed produced almost the same results in the phenotypical test of folate production during their growth in M17L. Transcriptomic analysis in this condition was performed in order to identify how far genetic differences between strains were responsible for basal variation in gene expression. Sequence similarity was verified by aligning the two genomes and setting 98% of identity. About 96% of the expressed genes were detected being orthologous, with the exception of TH1477 in M17L which showed a higher expression of unique genes, achieving 12% of unique gene expression. The expression comparison was performed by choosing as discriminant parameters a significant p-value (0.05), a fold-change greater than 1 and dropping sequences with non-identified function.

geneID	$\log_2$FC	p-value	Strain	Function
gene_0310	2.73	2.65E-02	TH1477	tRNA-dependent lipid II-AlaAla--L-alanine ligase
gene_1264	2.05	3.79E-02	TH1477	DNA-binding response regulator
gene_0331	1.74	2.20E-02	TH1436	Acetoin utilization acuB protein
gene_1676	2.08	1.74E-03	TH1436	Fumarate reductase, flavoprotein subunit precursor (EC 1.3.99.1)
gene_0394	2.11	4.02E-02	TH1436	PTS system, fructose-specific IIA component (EC 2.7.1.69)
gene_0332	2.12	1.16E-02	TH1436	Acetoin utilization protein AcuB
gene_0286	2.29	3.79E-02	TH1436	2',3'-cyclic-nucleotide 2';-phosphodiesterase (EC 3.1.4.16)
gene_1024	2.29	4.20E-02	TH1436	Lipoate-protein ligase A
gene_1306	2.30	3.19E-02	TH1436	UDP-glucose 4-epimerase (EC 5.1.3.2)
gene_1765	2.37	8.24E-03	TH1436	Acetate kinase (EC 2.7.2.1)
gene_1505	2.38	3.06E-03	TH1436	Sucrose operon repressor ScrR, LacI family
gene_1308	2.66	3.40E-02	TH1436	Galactokinase (EC 2.7.1.6)
gene_1307	2.69	8.24E-03	TH1436	Galactose-1-phosphate uridylyltransferase (EC 2.7.7.10)
gene_1732	2.70	7.70E-03	TH1436	Phage infection protein
gene_1775	2.88	2.48E-03	TH1436	DNA binding protein, FIG046916
gene_1503	3.10	5.31E-04	TH1436	PTS system, sucrose-specific IIB component (EC 2.7.1.69)
gene_0630	3.11	5.76E-06	TH1436	Phage shock protein C, putative; stress-responsive transcriptional regulator

Table 5. 2 Orthologous genes differentially expressed in M17L fermentation**.** Strain showing the overexpression is indicated in the fourth column. $\log_2$FC=logarithm of the fold-change

As expected, the comparison of strain behaviour in synthetic media reported low number of variations in gene expression. Only 17 genes satisfied the criteria and only two were more expressed in TH1477. The majority of them was related to sugar transport and utilisation systems.

1.14.4 Expression in whey

Strain-dependent modifications in whey fermentation were analysed by comparing gene expression of the two strains exposed to the same controlled environmental alteration, namely the addition of tryptophan. Comparing genes expression of the two strains in the same medium demonstrated that the genetic difference was described by 480 genes when bacteria were grown in whey, and 680 genes when they were cultured in the presence of the amino acid. Of them, only a set of 365 genes was recognised to be changing in both the system. This subset represents the set of function which are differently activated by the whey environment in the two strains. Among them, only a limited group of 104 genes responded to the analysis requirements, showing a fold-change variation greater than 1, highlighting clear variation between strains behaviour and encoding for a clearly identified function. Their detailed description is presented in table S7. It is evident that their functions are principally related to amino acid management in term of intake, transport and exchange process, beside to the cell division and energy related metabolism. Several genes belonging to the folate biosynthetic route were present, e.g. 5-formyltetrahydrofolate cyclo-ligase, all showing change in expression level in strain TH1477.

1.14.5 Expression in whey: TH1436

The comparison of differently expressed genes between strains grown in the two whey environments evidenced that expression of a set of genes is specifically changing in whey (115 genes) and in whey with tryptophan(315 genes). These behaviours depend entirely on the presence of the amino acids. In the first case presumably metabolism not repressed from the amino acids are more active or are aimed to supply functions which, when the amino acid is present, are no longer necessary. In the second case, tryptophan was provides in large excess and it forced the bacterial metabolism to move from its homeostatic equilibrium towards a new state, characterised by several mechanisms aiming to lower its concentration.

whey				wheyTRP			
geneID	log$_2$FC	p-value	Function	geneID	log$_2$FC	p-value	Function
gene_1625	2.86	3.52E-09	KH domain RNA binding protein YlqC	gene_0356	2.31	4.35E-17	Ammonium transporter family
gene_1626	2.69	6.81E-09	SSU ribosomal protein S16p	gene_0699	2.13	2.63E-06	Mobile element protein
gene_0133	2.13	7.81E-06	Protein YidD	gene_0822	1.97	7.48E-07	Type I restriction-modification system, specificity subunit S (EC 3.1.21.3)
gene_1392	2.04	2.84E-04	GTP-sensing transcriptional pleiotropic repressor codY	gene_0358	1.95	1.17E-07	Nitrogen regulatory protein P-II
gene_1155	1.99	1.79E-06	Phosphoenolpyruvate-protein phosphotransferase of PTS system (EC 2.7.3.9)	gene_0792	1.91	1.26E-09	ABC transporter membrane-spanning permease - glutamine transport
gene_1370	1.93	2.48E-06	putative ATP-dependent Clp proteinase (ATP-binding subunit)	gene_0574	1.91	1.25E-24	Topoisomerase IV subunit A (EC 5.99.1.-)
gene_1135	1.93	6.84E-07	membrane protein, putative	gene_0793	1.87	4.65E-17	Amino acid ABC transporter, ATP-binding protein
gene_0390	1.92	3.66E-05	Cysteinyl-tRNA synthetase related protein	gene_1680	1.82	4.40E-06	formate/nitrite transporter family protein, truncated
gene_1624	1.88	6.63E-07	**Dihydrofolate synthase (EC 6.3.2.12) / Folylpolyglutamate synthase (EC 6.3.2.**17)	gene_0890	1.79	3.62E-21	Fibronectin/fibrinogen-binding protein
gene_1775	1.82	3.84E-04	DNA binding protein, FIG046916	gene_0714	1.75	1.27E-08	Pneumococcal vaccine antigen A homolog
gene_0128	1.77	1.76E-06	FIG042801: CBS domain containing protein	gene_0573	1.73	1.06E-16	Topoisomerase IV subunit B (EC 5.99.1.-)
gene_0056	1.73	2.81E-04	Translation elongation factor Ts	gene_0716	1.73	2.30E-11	ABC transporter, ATP-binding/permease protein
gene_0511	1.70	2.78E-06	**5-Enolpyruvylshikimate-3-phosphate synthase (EC 2.5.1.19)**	gene_0715	1.68	6.83E-09	ABC transporter, ATP-binding/permease protein
gene_0334	1.68	4.78E-06	FIG000605: protein co-occurring with transport systems (COG1739)	gene_0617	1.64	5.69E-18	General stress protein
gene_0619	1.57	2.74E-03	Catabolite control protein A	gene_1763	1.62	6.31E-06	Transcriptional regulator, Cro/CI family

gene_0206	1.55	1.34E-05	Undecaprenyl-diphosphatase (EC 3.6.1.27)	gene_1218	1.59	5.87E-08	Two-component response regulator
				gene_0823	1.58	6.94E-18	tRNA:m(5)U-54 MTase gid
				gene_0706	1.56	8.21E-11	Adenosine deaminase (EC 3.5.4.4)

Table 5. 3 Selection of genes higher expressed by TH1436 during the two whey fermentation. Both whey (whey) and tryptophan enriched whey (wheyTRP) outcomes are reported. log$_2$FC= logarithm of fold change. In bold, genes encoding for enzymes contributing in the folate pathway

In should be noted that both the conditions affected more deeply TH1436 than TH1477. In whey, 52 genes were detected more expressed, and in table 5.3 are reported only those having a fold-change value higher than 1.5. Within those, particularly interesting are CodY pleiotropic repressor, which is a DNA-binding protein repressing the expression of many genes that are induced when cell transits from rapid exponential growth to stationary phase, and other genes involved in cell proliferation and homeostasis maintenance, i.e. CBS protein. In this subset a couple of genes were identified partecipating to the folate metabolism pathway. Genes related to the enriched whey fermentations assigned exclusively to TH1436 were 60 and similarly to what detected for the other strain, they belong mainly to the cell division system and membrane transport. Differently, no folate biosynthesis closely related gene was recorded.

1.14.6 Expression in whey: TH1477

The same analyses were performed on genes selectively over-expressed in the other analysed strains, TH1477, which in phenotypical tests displayed an elevated potential in folate production both in pure whey and even more in tryptophan enriched whey.

whey				wheyTRP			
geneID	log$_2$FC	p-value	Function	geneID	log$_2$FC	p-value	Function
gene_0745	1.71	3.27E-03	5-methyltetrahydropteroyltriglutamate--homocysteine methyltransferase (EC 2.1.1.14)	gene_0310	2.17	3.60E-21	tRNA-dependent lipid II-AlaAla--L-alanine ligase
				gene_1106	2.11	1.91E-04	Mobile element protein
				gene_1650	2.10	1.76E-12	Succinyl-CoA synthetase, alpha subunit-related enzymes
				gene_1158	2.03	1.17E-17	NADH peroxidase (EC 1.11.1.1)
				gene_1208	1.87	4.24E-18	ABC transporter, ATP-binding protein
				gene_1207	1.86	6.10E-19	ABC-type multidrug transport system, ATPase component
				gene_0141	1.82	2.55E-08	ABC transporter
				gene_1287	1.74	1.47E-07	putative coenzyme PQQ synthesis protein
				gene_1341	1.59	9.44E-09	**Tryptophan synthase alpha chain (EC 4.2.1.20)**
				gene_1206	1.56	5.15E-14	Transcriptional regulator, GntR family
				gene_1656	1.53	1.10E-12	SSU ribosomal protein S12p (S23e)
				gene_0350	1.50	1.94E-12	Triosephosphate isomerase (EC 5.3.1.1)
				gene_1348	1.50	9.65E-07	**Isochorismate pyruvate-lyase (EC 4.-.-.-)**

Table 5. 4 Selection of genes over-expressed by TH1477 during the two whey fermentation condition tested. Both whey (whey) and tryptophan enriched whey (wheyTRP) results are reported. log$_2$FC= logarithm of fold change. In bold, genes encoding for enzymes contributing in the folate pathway.

In table 5.4 are reported details about a set of genes discovered to be differently expressed between the two strains. Overall, only five genes were detected to be over-expressed in TH1477 during whey fermentation and of them only one was expressed more than 1.5 fold, namely 5-methyltetrahydropteroyltriglutamate-homocysteine methyltransferase. This gene is involved in the methyl group transfer in the methionine-cysteine exchange system. Addition of amino acids increased the number of genes specifically highly expressed in this strain up to 58 fold. A large amount of the discovered genes principally regarded the ABC transporters category, but also a peroxidase and a transcriptional factor, which putatively controls various biological processes, including antibiotic production, sensing of nutritional status, growth, proliferation and diverse metabolic processes, i.e. amino acids metabolism, were identified. Two key enzymes, marked in bold, of the folate pathway are highly expressed in this particular case.

1.14.7 Folate biosynthesis mechanism insight

Several mechanisms may affect phenotypes expression of bacteria. The defining of key steps occurred in the complex system of gene interaction may represent a challenge if some essential rules are not respected. In order to inspect expression variation in different media and to compare responses of the two strains, firstly homology analysis of the gene sequences was performed and non-homologous sequences were discharged. Expression changes were computed by analysing variation within strains and evaluating the number of fold-change respect to a zero condition, which was arbitrarily established to be the whey fermentation. This allowed identified both: (i) the genes following the same trend in both the strains moving from whey to M17L. This allowed localising the effectors of the deep reduction in folate synthesis in whey. (ii) The differences between strains which strongly conditioned the vitamin production passing from whey to tryptophan enriched whey. Data were elaborated by computing the pairwise comparison of gene expression, obtaining fold change values and significance of the data. Genes belonging to three main subsystems, namely the common pathway for the folate and aromatic amino acids synthesis, the folate biosynthesis cluster and the main route for its consumption, the one-carbon metabolism were considered. Redundant functions were dropped.

.

Subsystem	Function	TH1436		TH1477	
		Fold-change whey/wheyTRP	Fold-change whey/M17L	Fold-change whey/wheyTRP	Fold-change whey/M17L
Chorismate Synthesis	Chorismate mutase I (EC 5.4.99.5)	▬ 0.16	▬ 0.07	▬ 0.08	▬ 0.70
	Prephenate dehydratase (EC 4.2.1.51)	▬ -0.81	▼ -1.47	▬ -0.28	▬ -0.18
	Prephenate dehydrogenase (EC 1.3.1.12)	▼ -1.19	▬ -0.72	▬ -0.82	▬ -0.74
Chorismate: Intermediate for synthesis of Tryptophan, PAPA antibiotics, PABA, 3-hydroxyanthranilate and more.	Aminodeoxychorismate lyase (EC 4.1.3.38)	▬ 0.30	▼ -1.00	▬ -0.15	▬ -0.48
	Anthranilate phosphoribosyltransferase (EC 2.4.2.18)	▼ -1.13	▬ 0.33	▬ -0.99	▬ -0.25
	Anthranilate synthase, amidotransferase component (EC 4.1.3.27)	▼ -1.37	▬ -0.21	▬ -0.77	▼ -1.13
	Anthranilate synthase, aminase component (EC 4.1.3.27)	▼ -1.15	▬ -0.33	▬ -0.76	▬ -0.80
	Indole-3-glycerol phosphate synthase (EC 4.1.1.48)	▼ -1.12	▬ 0.02	▬ -0.68	▬ -0.67
	Isochorismate pyruvate-lyase (EC 4.-.-.-)	▬ -0.95	▬ -0.07	▬ -0.63	▼ -1.18
	Phosphoribosylanthranilate isomerase (EC 5.3.1.24)	▼ -1.50	▬ -0.12	▬ -0.81	▬ -0.66
	Phosphoribosylformimino-5-aminoimidazole carboxamide ribotide isomerase (EC 5.3.1.16)	▬ -0.74	▲ 1.61	▼ -1.83	▲ 1.24
	Tryptophan synthase alpha chain (EC 4.2.1.20)	▼ -1.14	▬ -0.17	▬ -0.43	▬ -0.20
	Tryptophan synthase beta chain (EC 4.2.1.20)	▼ -1.20	▬ 0.18	▬ -0.79	▬ -0.13
Common Pathway For Synthesis of Aromatic Compounds (DAHP synthase to chorismate)	2-keto-3-deoxy-D-arabino-heptulosonate-7-phosphate synthase I alpha (EC 2.5.1.54)	▬ 0.04	▲ 1.20	▬ -0.35	▬ 0.55
	2-keto-3-deoxy-D-arabino-heptulosonate-7-phosphate synthase I alpha (EC 2.5.1.54)	▬ 0.08	▬ 0.30	▬ 0.39	▬ 0.58
	3-dehydroquinate dehydratase I (EC 4.2.1.10)	▼ -1.25	▼ -1.00	▬ -0.71	▼ -1.43
	3-dehydroquinate synthase (EC 4.2.3.4)	▼ -1.33	▬ -0.72	▬ -0.85	▬ -0.72
	5-Enolpyruvylshikimate-3-phosphate synthase (EC 2.5.1.19)	▼ -1.29	▼ -1.13	▬ -0.44	▬ -0.01
	Chorismate synthase (EC 4.2.3.5)	▼ -1.08	▬ -0.83	▬ -0.68	▬ -0.91
	Shikimate kinase I (EC 2.7.1.71)	▼ -1.17	▼ -1.49	▬ -0.35	▬ -0.01
	Shikimate/quinate 5-dehydrogenase I beta (EC 1.1.1.282)	▼ -1.26	▬ -0.87	▬ -0.72	▼ -1.32
Folate Biosynthesis	5-formyltetrahydrofolate cyclo-ligase (EC 6.3.3.2)	▲ 1.20	▬ 0.52	▬ 0.71	▬ -0.83
	Dihydrofolate reductase (EC 1.5.1.3)	▬ -0.06	▼ -1.43	▬ 0.06	▬ -0.58
	Dihydrofolate synthase (EC 6.3.2.12)	▼ -1.88	▼ -1.84	▬ 0.15	▼ -1.16
	Folylpolyglutamate synthase (EC 6.3.2.17)	▬ 0.49	▬ -0.83	▬ -0.43	▬ -0.07
	Thymidylate synthase (EC 2.1.1.45)	▬ -0.31	▬ -0.50	▬ -0.31	▬ -0.03
Folate biosynthesis cluster	2-amino-4-hydroxy-6-hydroxymethyldihydropteridine pyrophosphokinase (EC 2.7.6.3)	▬ 0.37	▬ 0.34	▬ 0.54	▬ 0.38
	Cell division protein FtsH (EC 3.4.24.-)	▲ 1.20	▬ 0.52	▬ -0.11	▬ -0.46
	Dihydroneopterin aldolase (EC 4.1.2.25)	▬ 0.46	▬ 0.00	▬ 0.07	▬ 0.27
	Dihydropteroate synthase (EC 2.5.1.15)	▬ 0.51	▬ -0.05	▬ 0.13	▬ -0.57
	GTP cyclohydrolase I (EC 3.5.4.16) type 1	▬ 0.33	▬ -0.03	▬ 0.04	▬ -0.86
	Hypoxanthine-guanine phosphoribosyltransferase (EC 2.4.2.8)	▬ 0.38	▬ -0.08	▬ 0.03	▬ -0.82
One-carbon metabolism by tetrahydropterines	5,10-methylenetetrahydrofolate reductase (EC 1.5.1.20)	▲ 2.58	▬ -0.27	▬ 0.96	▼ -1.51
	Formate--tetrahydrofolate ligase (EC 6.3.4.3)	▬ -0.45	▼ -1.48	▬ 0.14	▬ -0.46
	Methylenetetrahydrofolate dehydrogenase (NADP+) (EC 1.5.1.5)	▬ -0.86	▲ 1.23	▼ -1.12	▲ 1.51
Phenylalanine and Tyrosine Branches from Chorismate	Aromatic amino acid aminotransferase gamma (EC 2.6.1.57)	▬ 0.68	▲ 1.12	▬ 0.33	▬ 0.70

Figure 5. 2 Fold changes (Log$_2$) of gene sequences belonging to folate biosynthesis, its upstream and downstream pathways. In grey are marked values with not significant difference (p-value cut off 0.05). Coloured marks indicate whether gene is more (green) or lower expressed (red) passing from whey to the second medium take in account in the comparison.

Data overview provided the idea that TH1477 has a reduced sensibility to tryptophan, expressed in two out of 36 genes compared to the 18 influenced genes in TH1436. As expected, strain TH1436 showed the lowering of transcription in a series of genes which are involved in the shikimate pathway, i.e. 3-dehydroquinate dehydratase I, 3-dehydroquinate synthase, 5-Enolpyruvylshikimate-3-phosphate synthase, chorismate synthase, shikimate kinase I, shikimate/quinate 5-dehydrogenase I beta passing from whey to the tryptophan enriched whey. This confirms that tryptophan played a role in the negative regulation of shikimic acid pathway, similarly to what is recognised for other aromatic amino acids (175). It was also registered a slight reduction in the absolute vitamin value in the phenotypic test. In addition, the inhibitory effect of this amino acid on the aromatic amino acid production was confirmed by the reduction of transcriptional activity of genes constituting the tryptophan biosynthesis pathway, i.e. anthranilate phosphoribosyltransferase, anthranilate synthase, amidotransferase component, anthranilate synthase, aminase component, indole-3-glycerol phosphate synthase, phosphoribosylanthranilate isomerase tryptophan synthase alpha and beta chains.

Instead, in TH1477 the conserved activity rate moving from whey to whey with tryptophan demonstrates that the sensibility to the negative effector is reduced in this strain. Moreover, it was not recorded a decrease in tryptophan synthesis transcription, which permits to speculate that the route was expressed and the enzymes involved in their correspondent exchange reaction, but probably travel the route in the opposite direction and, hence, converting the amino acids to chorismate. Even if for a long time it was accepted the idea that anthranilate synthase could mediate the unidirectional conversion of chorismate to anthranilate, recently it has been documented its bi-functionality (176). In this strains, two enzymes reduced drastically their transcription in presence of tryptophan, namely phosphoribosylformimino-5-aminoimidazole carboxamide ribotide isomerase and methylenetetrahydrofolate dehydrogenase (NADP$^+$). The former is an enzyme involved in the Amadori rearrangement of aminoaldoses which was discovered to be structural and functionally similar to the specific genes involved in the anthranilate isomerase. It is known that such similarities may result in promiscuous activities(177) and probably it is implicated in the subtraction of these molecules favouring another metabolism. The latter is involved in the reduction of NADP$^+$ into NADPH. The same enzymes seem to explain the difference between whey and M17L fermentation: few genes appear to modify their transcript rate, and those two are more express in M17L than in whey. On the contrary, 3-dehydroquinate dehydratase I, which leads toward shikimate production, displayed a decrease in synthetic medium, together with dihydrofolate synthase. These phenomena are difficult to explain due to the folate amount measured for both the strains in the synthetic broth. All of the analysed genes were orthologous, even if three of them were overlapping only for around 50% of the sequence length, namely 3-dehydroquinate dehydratase I (1284 nt in TH1436 and 492 nt in TH1477), shikimate kinase I (1164 nt in TH1436 and 678 nt in TH1477) and prephenate dehydratase (492 nt in TH1436 and 825 nt in TH1477). The identities in the conserved part were enough high to permit the identification of their function and the determination of homology with the corresponding CDS carried from the second strain. It indicates that those sequences still conserved the functional domain, hence it can be supposed, however, that the sequence modification have conditioned enzymatic activity, for example in the case of shikimate kinase I in TH1477 it cannot be exclude that a particularly high rate of reaction is present which may lead to an additional overproduction of folate, even if it is known that in normal conditions this enzyme is undergoing a negative control from tryptophan (178).

Conclusions

Food is an indispensable part of daily life. Many common products undergo several form of processing before reached the final consumer. In several of these processes, microorganisms play an important role, guiding the food transformation into the desired end product (e.g. fermentation of olives, alcoholic beverages such as beer and wine or various fermented dairy products such as cheese and yogurt) or the controlling undesirable bacteria proliferation. Definitely, the starter culture choice strongly influences the properties of the final product. The food industry is very attentive to the optimization of strain performance toward the diversification of products properties, which mainly depends both on the organism employed and the process settings. Unique properties valorisation is a challenge task requiring an in-deep knowledge background (168). The progressive steps forward of the next generation sequencing technologies have allowed to lead food analyses and the new product development to a new level, mainly thank to an innovative approach of data integration. ‘Foodomics’ is a new, global discipline in which food, including its nutritional aspects, advanced analytical techniques and bioinformatics are combined (179).

Starting from the idea that bacterial biodiversity harbours innovative properties which can be interesting in the new product formulations, it was chosen to investigate eight strains belong to one of the most important dairy starter culture, *Streptococcus thermophilus.* Strains were isolated from typical Italian cheese-making processes and selected in order to represent the greater biodiversity, choosing technological processes, animal origins and geographic regions largely different each other.

Strain genomes were successfully sequenced using the Solexa (Illumina) technology. In this project was firstly provided a comparison between new sequenced strains and available genomes in NCBI database isolated in several region of the world, range from Canada to Australia passing through Europe. It was discovered that one of the new sequenced strains, M17PTZA496, is characterized from about 15% of additional genetic information. Two insertion islands were detected using of specific software. Further analyses on encoding sequences permit to speculate that the two insertion events were occurred independently. Moreover, duplications were verified identifying gene sequences sharing a high-scoring identity and localised in different genome areas, condition distinguishing the paralog genes. It was possible using a tool which clusters together all the sequencing achieving a matching score higher than the costumer-defined threshold. In order to discriminate whether sequences originated from species duplications or their multiplication was a strain-specific event, stringent parameters were set out and the CDS classified in relationship to their number of copies present in the other studied strains. A small amount of the discovered features are shared with one of the other strains, the one isolated from the same environment, namely MTH17CL396. The 57% of duplicate genes were found to be multi-copies only in M17PTZA496, they were recognised being homologous of CDS carried in the other strains in single copy. This class was particularly rich in RNA related genes and in some functional features, assigned to sugar and amino sugar metabolism, transcriptional regulators and transport system. Eighteen out of sixty duplicated features are encoded from flanking sequences (placed in contig69, from peg.2001 to peg.2023): this indicates that this portion of genome was duplicated in a unique event.

Then, an overall of 17 genomes were compared in order to inspect species history. Phylogenetic analyses have demonstrated that M17PTZA496 is undergoing a diversification process, detected both from SNPs and conserved genes phylogenetic reconstructions. Strains derived from nearby geographical region only partially clustered together, indicating that technological selection has strongly conditioned the species evolution. Nevertheless, functional analyses of gene content, performed on the features overall and on the unique features carried from each strain, demonstrated that M17PTZA496 has not acquired innovative capabilities ad that, alike to what previously reported (79), in general the main variations are stated on basal function categories, namely ‘Amino Acids and Derivatives’, ‘Carbohydrates’, ‘DNA Metabolism’ and ‘Membrane Transport’ which together describe almost 50% of the strain specificity,

The technological properties were deeply inspected at both genetic and phenotypic levels. Growth curve, fermentation rate and proteolytic activity were described for each strain in order to identify the profile determining a good fermentation performance. It was discovered that even if the analysed strains, with the exception of 1F8CT, showed a similar profile in the growth performance and proteolytic capability, they differ significantly in milk fermentation task. Genetic analyses have registered small differences in the *gal-lac* operon genes and in the two principal intergenic regions underling genes expression. Those variations are mainly connected with the galactose metabolising. A new, smaller, consensus sequence regulating the galactose consumption has been pinpointed in these strains. One of the strains, TH1435, possesses the sequence coding for species-specific proteases but did not express the proteolytic activity.

Other two interesting characters, enhancing the end product acceptance and its safety, were considered in this work. In both the cases, gene subsystems, fulfilled from by RAST automatic annotation, were compared founding promising pattern for the character active coding. In the first case, orthologous sequences of all the main *eps* genes categories were found in strains 1F8CT, M17PTZA496 and TH982, together with a key enzyme, glycosyltransferases. However, phenotypic tests have demonstrated that none of the strains produce EPS and that the differences in the structural organization of cells depend on variation of cell-cell anchoring system. Putative bacteriocins' sequences were investigated using a specific tool, BAGEL3, which mines CDS carrying high specific targets identifying the biocins. Almost all the detected sequences were ascribed to lantibiotic peptides, which are widely express in LAB and of major interest for the dairy production. In addition, in four out of eight strains, i.e. M17TZA496, TH1435, TH1436 and TH1477, specific immunity proteins were individuated. This was a strong indication of antimicrobial activity. Otherwise, the phenotypic assay, conducted on a total of seventeen bacteria strains commonly found in dairy products, has revealed that no one of those strains expressed antimicrobial activity.

Another character was used to describe strain specificity, the biosynthetic and amino acid interconversion systems capabilities. An overall of 168 genes distributed in fourteen pathways were studied to identify genetic variations. It was noted that as reported before (63), the amino acid biosynthetic pathways are strongly conserved at species level. In fact it was clearly determined that all the strains are carrying the same panel of genes, while only two function were occasionally lacking, namely cystathionine gamma-lyase (EC 4.4.1.1) and cystathionine beta-synthase (EC 4.2.1.22) both in M17PTZA496, MTH17CL396 and TH1477. Those genes are strictly connected because, in two consecutive reactions, covert L-serine into L-cysteine. Phenotypic data showed important differences among strains. Two strains M17PTZA496 and TH1477 are more exigent than the other, while TH1436 is the most tolerant, proliferating in the 80% of one-amino acid omitted tests. Even if these diversities reflect promising application in new product tuning, as example in relationship to flavour (180), in the past years their technological applications were rarely investigated for this species.

Streptococcus thermophilus is widely recognised as one of the best folate producers. Studied strains were tested in the vitamin production during their growth in synthetic medium, to evaluate their ability in this task comparing the results against those reported in literature.

Four out of eight strains displayed a total folate production similar to the literature findings (149) while three, namely TH982, TH985 and TH1435 show an elevate capability in this task, achiving up to 498 ng/ml. Concerning the short-tailed folates, strains showed tendencies similar to those recorded for the total folate, namely the major part of the strain reached the average amount of folate expected from this species whilst two strains, TH982 and TH1435, displayed greatest abilities, overcoming 200 ng/ml. In the analyses of folates released in the medium, profiles of total folate detected were alike to those identify in the whole broths, but with an reduction of about 200 ng/ml in the absolute values. Unexpected, the short-glutaryl-chain folate realesed in the supernatant achieved high values in M17PTZA496, up to 135 ng/ml. Indeed, its profile was unusual, after 18 hours it seems to undergo a phenomenon which leads to the release of almost all its folate content in the medium. The higher value of folate was achieved in all the cases after 18 hours of fermentation. After that point, vitamin decreased, probably because consumed from the bacteria.

In the view of the promising application of selected strains as folate providers, bacteria were tested in whey fermentations. Firstly, folate synthesis were compared in two fermentation set up, changing the tempreature from 37°C to 42°C.

Total folate production in whey at 37°C ranged between 12 and 30 ng/ml and in whey at 42°C between 10 and 25 ng/ml. According to the preceding findings (152), peaks of folate production were recorded early in this medium, and as expected it folate production did not achieves the same greatness recorded for the synthetic medium. Instead, raising the temperature did not increase its values, which, in average, achieved the best results after 6 hours at 37°C. Nonetheless, the major increase was obtained using TH1477, which duplicates the folate amount respect to the zero point, and recorded after 24 hours of fermentation. The overall comparison of two experiments revealed that temperature affects the kinetics of the folate production but not its yield.

Several compounds were supposed to influence folate synthesis in *S. thermophilus,* mainly precursors of the biosynthetic pathway (154) and carbohydrates with prebiotic properties (166). It is known (175) that three amino acids play a role in the folate upstream pathway regulation, the shikimate pathway. In the view of the interesting findings on strain amino acid metabolism, the bacterial reactions, in terms of folate production, to different amino acid concentration was analysed. Four amino acids, namely glycine, phenylalanine, tyrosine and tryptophan, were tested at three concentrations, i.e. 10mM, 20mM and 50mM. Glycine seems little affected the vitamin production. Strain TH1477 showed being the most susceptible to amino acids, in particular after 24h with 50mM of glycine. Analyses of conditions in large excesses of phenylalanine showed a general increase of folate, main expressed after 18 hours of fermentation. TH982, which has yet displayed marked capabilities in this task fermenting the synthetic medium, revealed to be the most responsive to phenylalanine and it best performed with 20mM of this amino acid. Although some evidences witness symmetries in the phenylalanine and tyrosine metabolisms, this test demonstrated that the regulation is differently influenced from these amino acids. In fact strains response and absolute values achieved in the two scenarios are quite different. Contrary to what recorded in literature (149), tyrosine did not show a marked inhibitory effect. The last amino acid tested recorded the highest value of folate, after 18 hours of TH1477 fermentation in whey enriched with 50mM of the tryptophan. The folate amount has been increased four- fold respect to the initial amount.

Despite the generalized increased usage of the next generation sequencing in food science, nowadays the associations between genomic features and phenotypic characters are still being a complex problem. The development of novel high-throughput RNA sequencing technologies provides new methods for both mapping and quantifying transcriptomes. This approach was chosen to discover the mechanism underlies the elevate ability in folate synthesis showed by TH1477 in tryptophan enriched whey, Data furnished the idea that TH1477 has a reduced sensibility to tryptophan, since gene expression changes in two out of 36 genes in comparison to the 18 influenced from its presence in TH1436. In fact, TH1436 shows a lowering of transcription rate in set of genes involved in the shikimate and tryptophan biosynthesis pathway, where tryptophan was expected playing a role in the negative regulation (175). On contrary, TH1477 has displayed the normal rate transcription for these two pathways, allowing supposing that probably both these way were active and the second probably retrace towards chorismate. Whilst for long anthranilate synthase has been considered mediating the unidirectional conversion of chorimate to anthranilate, recently it has been documented its bi-functionality (176). All of the analysed genes are homologous but in three cases the sequences overlap for around 50% of length, namely in 3-dehydroquinate dehydratase I, shikimate kinase I and prephenate dehydratase. Those sequences probably conserved the functional domains, however the sequence modification could conditioned the enzymatic activity.

This work has permit to exemplify how the new sequencing technologies will lead, in the next years, to a food revolution. 'Foodomics' will represent the main road toward a more suitable product design and process optimization. The results obtained during this work are interesting both from a scientific and applicative point of view. They witness that autochthonous strains could guide toward products which response to the new requirements of healthy food and of nutritional deficiencies' fulfilling.

Bibliography

1. **Koller M**, **Bona R**, **Braunegg G**, **Hermann C**, **Horvat P**, **Kroutil M**, **Martinz J**, **Neto J**, **Pereira L**, **Varila P**. 2005. Production of Polyhydroxyalkanoates from Agricultural Waste and Surplus Materials. Biomacromolecules 561–565.

2. **Baldasso C**, **Barros TC**, **Tessaro IC**. 2011. Concentration and purification of whey proteins by ultrafiltration. Desalination **278**:381–386.

3. **Panesar P**, **Kennedy J**, **Gandhi D**, **Bunko K**. 2007. Bioutilisation of whey for lactic acid production. Food Chem. **105**:1–14.

4. **Mollea C**, **Marmo L**, **Bosco F**. 2013. Valorisation of Cheese Whey , a By-Product from the Dairy Industry, p. 549–588. *In* Food Industry.InTech.

5. **Carvalho F**, **Prazeres AR**, **Rivas J**. 2013. Cheese whey wastewater: characterization and treatment. Sci. Total Environ. **445**:385–96.

6. **Pintado M E MAC**, **Malcatal FX**. Review : Technology , Chemistry an «! Microbiology of Whey Cheese. Food Sci Technol Int **7**:105-116

7. **Kosikowski F V.** 1979. Whey Utilization and Whey Products. J. Dairy Sci. **62**:1149–1160.

8. **Pescuma M**, **Hébert EM**, **Mozzi F**, **Font de Valdez G**. 2008. Whey fermentation by thermophilic lactic acid bacteria: evolution of carbohydrates and protein content. Food Microbiol. **25**:442–51.

9. **Guimarães PMR**, **Teixeira J a**, **Domingues L**. 2010. Fermentation of lactose to bio-ethanol by yeasts as part of integrated solutions for the valorisation of cheese whey. Biotechnol. Adv. **28**:375–84.

10. **Povolo S**, **Toffano P**, **Basaglia M**, **Casella S**. 2010. Polyhydroxyalkanoates production by engineered Cupriavidus necator from waste material containing lactose. Bioresour. Technol. **101**:7902–7907.

11. **Schirru S**, **Favaro L**, **Mangia NP**, **Basaglia M**, **Casella S**, **Comunian R**, **Fancello F**, **De Melo Franco BDG**, **De Souza Oliveira RP**, **Todorov SD**. 2014. Comparison of bacteriocins production from Enterococcus faecium strains in cheese whey and optimised commercial MRS medium. Ann. Microbiol. **64**:321–331.

12. **deWit JN**, **Klarenbeek G**. 1984. Effects of various heat treatments on structure and solubility of whey proteins. J. Dairy Sci. **67**:2701–2710.

13. **Jeli I**, **Božani R**, **Tratnik L**. 2008. Whey-based beverages- a new generation of diary products. Mljekarstvo **58**:257–274.

14. **Shiby VK**, **Radhakrishna K**, **Bawa AS**. 2013. Development of whey-fruit-based energy drink mixes using D-optimal mixture design. Int. J. Food Sci. Technol. **48**:742–748.

15. **Saxena D**, **Chakraborty SK**, **Sabikhi L**, **Singh D**. 2013. Process optimization for a nutritious low-calorie high-fiber whey-based ready-to-serve watermelon beverage. J. Food Sci. Technol. **52**:960–967.

16. **Viljanen K**, **Kylli P**, **Hubbermann EM**, **Schwarz K**, **Heinonen M**. 2005. Anthocyanin antioxidant activity and partition behavior in whey protein emulsion. J. Agric. Food Chem. **53**:2022–2027.

17. **Assadi MM**, **Abdolmaleki F**, **Mokarrame RR**. 2008. Application of whey in fermented beverage production using kefir starter culture. Nutr. Food Sci. **38**:121–127.

18. **Magalh??es KT**, **Dragone G**, **De Melo Pereira G V.**, **Oliveira JM**, **Domingues L**, **Teixeira JA**,

E Silva JBA, **Schwan RF**. 2011. Comparative study of the biochemical changes and volatile compound formations during the production of novel whey-based kefir beverages and traditional milk kefir. Food Chem. **126**:249–253.

19. **Kurmann JA**, **Rasic JL**, **Kroger M**. 1992. Encyclopedia of fermented fresh milk products: an international inventory of fermented milk, cream, buttermilk, whey, and related products. Van Nostrand Reinhold.

20. **Huth PJ**, **DiRienzo DB**, **Miller GD**. 2006. Major Scientific Advances with Dairy Foods in Nutrition and Health. J. Dairy Sci. **89**:1207–1221.

21. **Madureira AR**, **Pereira CI**, **Gomes AMP**, **Pintado ME**, **Xavier Malcata F**. 2007. Bovine whey proteins – Overview on their main biological properties. Food Res. Int. **40**:1197–1211.

22. **Pérez MD**, **Calvo M**. 1995. Interaction of beta-lactoglobulin with retinol and fatty acids and its role as a possible biological function for this protein: a review. J. Dairy Sci. **78**:978–988.

23. **Stanciuc NSTĂ**, **Râpeanu G**. 2010. An overview of bovine α-lactalbumin structure and functionality. Food Technol. **34**:82–93.

24. **Sharpe, S, Gamble, G, Sharpe N**. 1994. Cholesterol-lowering of immune blood of immune milk. Am. J. Clin. Nutr. **59**:929–934.

25. **González-Chávez S a**, **Arévalo-Gallegos S**, **Rascón-Cruz Q**. 2009. Lactoferrin: structure, function and applications. Int. J. Antimicrob. Agents **33**:301.e1–8.

26. **Seifu E**, **Buys EM**, **Donkin EF**. 2005. Significance of the lactoperoxidase system in the dairy industry and its potential applications : a review. Trend Food Sci Tech **16**:137–154.

27. **Clement M**, **Tremblay J**, **Lange M**, **Thibodeau J**, **Belhumeur P**. 2008. Purification and Identification of Bovine Cheese Whey Fatty Acids Exhibiting In Vitro Antifungal Activity. J. Dairy Sci. **91**:2535–2544.

28. **Vandenplas Y**. 2002. Oligosaccharides in infant formula. Br. J. Nutr. **87**:S293–S296.

29. **Sundekilde UK**, **Barile D**, **Meyrand M**, **Poulsen N a.**, **Larsen LB**, **Lebrilla CB**, **German JB**, **Bertram HC**. 2012. Natural variability in bovine milk oligosaccharides from Danish Jersey and Holstein-Friesian breeds. J. Agric. Food Chem. **60**:6188–6196.

30. **Eustache J-M**. 1977. Extraction of glycoproteins and sialic acid fom whey. U.S. Pat.

31. **Salcedo J**, **Barbera R**, **Matencio E**, **Alegría a.**, **Lagarda MJ**. 2013. Gangliosides and sialic acid effects upon newborn pathogenic bacteria adhesion: An in vitro study. Food Chem. **136**:726–734.

32. **Madureira AR**, **Tavares T**, **Gomes AMP**, **Pintado ME**, **Malcata FX**. 2010. Invited review : Physiological properties of bioactive peptides obtained from whey proteins. J. Dairy Sci. **93**:437–455.

33. **Briczinski EP**, **Roberts RF**. 2002. Production of an Exopolysaccharide-Containing Whey Protein Concentrate by Fermentation of Whey. J. Dairy Sci. **85**:3189–3197.

34. **Badel S**, **Bernardi T**, **Michaud P**. 2011. New perspectives for Lactobacilli exopolysaccharides. Biotechnol. Adv. **29**:54–66.

35. **Nath A**, **Verasztó B**, **Basak S**, **Koris A**. 2015. Synthesis of Lactose-Derived Nutraceuticals from Dairy Waste Whey — a Review. Food Bioprod Process **9**:16-48.

36. **Kim YS**, **Park CS**, **Oh DK**. 2006. Lactulose production from lactose and fructose by a thermostable ??-galactosidase from Sulfolobus solfataricus. Enzyme Microb. Technol. **39**:903–908.

37. **Virtanen T**, **Pihlanto A**, **Akkanen S**, **Korhonen H**. 2007. Development of antioxidant activity in milk whey during fermentation with lactic acid bacteria. J. Appl. Microbiol. **102**:106–115.

38. **Beaulieu J**, **Dupont C**, **Lemieux P**. 2007. Anti-inflammatory potential of a malleable matrix composed of fermented whey proteins and lactic acid bacteria in an atopic dermatitis model. J. Inflamm. (Lond). **4**:6.

39. **Kato-Mori Y**, **Orihashi T**, **Kanai Y**, **Sato M**, **Sera K**, **Hagiwara K**. 2010. Fermentation Metabolites from Lactobacillus gasseri and Propionibacterium freudenreichii Exert Bacteriocidal Effects in Mice Yuko. J. Med. Food **13**:1460–1467.

40. **Hernandez-Mendoza A**, **Robles VJ**, **Angulo JO**, **De La Cruz J**, **Garcia HS**. 2007. Preparation of a whey-based probiotic product with Lactobacillus reuteri and Bifidobacterium bifidum. Food Technol. Biotechnol. **45**:27–31.

41. **Maragkoudakis P**, **Vendramin V**, **Bovo B**, **Treu L**, **Corich V**, **Giacomini A**. 2015. Potential use of scotta, the by-product of the ricotta cheese manufacturing process, for the production of fermented drinks. J. Dairy Res. 1–5.

42. **Madureira AR**, **Pereira CI**, **Truszkowska K**, **Gomes AM**, **Pintado ME**, **Malcata FX**. 2005. Survival of probiotic bacteria in a whey cheese vector submitted to environmental conditions prevailing in the gastrointestinal tract. Int. Dairy J. **15**:921–927.

43. **Leroy F**, **De Vuyst L**. 2004. Lactic acid bacteria as functional starter cultures for the food fermentation industry. Trends Food Sci. Technol. **15**:67–78.

44. **Ayad EHE**, **Verheul A**, **Wouters JTM**, **Smit G**. 2002. Antimicrobial-producing wild lactococci isolated from artisanal and non-dairy origins. Int. Dairy J. **12**:145–150.

45. **Beresford TP**, **Fitszimons N a**, **Brennan NL**, **Cogan TM**. 2001. Recent Advance in Cheese Microbiology. Int. Dairy J. **11**:259–274.

46. **Garabal JI**. 2007. Biodiversity and the survival of autochthonous fermented products. Int. Microbiol. 1–3.

47. **O'Sullivan L**, **Ross R.**, **Hill C**. 2002. Potential of bacteriocin-producing lactic acid bacteria for improvements in food safety and quality. Biochimie **84**:593–604.

48. **Schnu J**. 2005. Antifungal lactic acid bacteria as biopreservatives. Trends Food Sci. Technol. **16**:70–78.

49. **Corsetti A**, **Gobbetti M**, **De Marco B**, **Balestrieri F**, **Paoletti F**, **Russi L**, **Rossi J**. 2000. Combined effect of sourdough lactic acid bacteria and additives bread firmness and staling. J. Agric. Food Chem. **48**:3044–3051.

50. **Dragone G**, **Mussatto SI**, **Oliveira JM**, **Teixeira JA**. 2009. Characterisation of volatile compounds in an alcoholic beverage produced by whey fermentation. Food Chem. **112**:929–935.

51. **Florou-paneri P**, **Christaki E**, **Bonos E**. 2013. Lactic Acid Bacteria as Source of Functional IngredientsLactic Acid Bacteria in Research & Development for Food, Health and Livestock Purpose. InTech.

52. **Collins YF**, **McSweeney PLH**, **Wilkinson MG**. 2003. Evidence of a relationship between autolysis of starter bacteria and lipolysis in Cheddar cheese during ripening. J. Dairy Res. **70**:105–113.

53. **Wisselink H.**, **Weusthuis R.**, **Eggink G**, **Hugenholtz J**, **Grobben G.** 2002. Mannitol production by lactic acid bacteria: a review. Int. Dairy J. **12**:151–161.

54. **LeBlanc JG**, **Laino JE**, **Valle MJ**, **Vannini V**, **Sinderen D Van**, **Taranto MP**, **Font G**, **del Valle**

MJ, **Vannini V**, **van Sinderen D**, **Taranto MP**, **Font de Valdez G**, **Savoy de Giori G**, **Sesma F**, **Laiño JE**, **del Valle MJ**, **Vannini V**, **van Sinderen D**, **Taranto MP**, **de Valdez GF**, **de Giori GS**, **Sesma F**. 2011. B-Group vitamin production by lactic acid bacteria – current knowledge and potential applications. J. Appl. Microbiol. **111**:1297–1309.

55. **Delorme C**. 2008. Safety assessment of dairy microorganisms : Streptococcus thermophilus. Int J Food Microbiol. **126**:274–277.

56. **Giraffa G**, **Paris a**, **Valcavi L**. 2001. Genotypic and phenotypic heterogeneity of Streptococcus thermophilus strains isolated from dairy products. J. Appl. Microbiol **91**:937–43.

57. **Iyer R**, **Tomar SK**, **Maheswari TU**, **Singh R**. 2010. Streptococcus thermophilus strains : Multifunctional lactic acid bacteria. Int. Dairy J. **20**:133–141.

58. **Erkus O**, **Okuklu B**, **Yenidunya AF**, **Harsa S**. 2014. High genetic and phenotypic variability of Streptococcus thermophilus strains isolated from artisanal Yuruk yoghurts. LWT - Food Sci. Technol. **58**:348–354.

59. **Morandi S**, **Brasca M**. 2012. Safety aspects, genetic diversity and technological characterisation of wild-type Streptococcus thermophilus strains isolated from north Italian traditional cheeses. Food Control **23**:203–209.

60. **ISTAT** 2014. http://www.istat.it/

61. **Andrighetto C**, **Borney F**, **Barmaz A**, **Stefanon B**, **Lombardi A**. 2002. Genetic diversity of Streptococcus thermophilus strains isolated from Italian traditional cheeses. Int. Dairy J. **12**:141–144.

62. **Marino M**, **Maifreni, M**, **Rondinini G**. 2003. Microbiological characterization of artisanal Montasio cheese: Analysis of its indigenous lactic acid bacteria. FEMS Microbiol. Lett. **229**:133–140.

63. **Hols P**, **Hancy F**, **Fontaine L**, **Grossiord B**, **Prozzi D**, **Leblond-bourget N**, **Decaris B**, **Bolotin A**, **Delorme C**, **Ehrlich SD**, **Guedon E**, **Monnet V**, **Renault P**, **Kleerebezem M**. 2005. New insights in the molecular biology and physiology of Streptococcus thermophilus revealed by comparative genomics. FEMS Microbiol. Rev. **29**:435–463.

64. **Dandoy D**, **Fremaux C**, **De Frahan MH**, **Horvath P**, **Boyaval P**, **Hols P**, **Fontaine L**. 2011. The fast milk acidifying phenotype of Streptococcus thermophilus can be acquired by natural transformation of the genomic island encoding the cell-envelope proteinase PrtS, p. S21. *In* Microbial Cell Factories. BioMed Central Ltd.

65. **Neviani E**, **Giraffa G**, **Brizzi A**, **Carminati D**. 1995. Amino acid requirements and peptidase activities of Streptococcus salivarius subsp. thermophilus. J. Appl. Microbiol. **79**:302–307.

66. **Smit BA**. 2004. Flavour formation from Amino Acids in fermented dairy product.PhD thesis. Wageningen University, Wageningen.

67. **Vin F De**, **Rådstro P**, **Herman L**, **Vuyst L De**, **Icrobiol APPLENM**. 2005. Molecular and Biochemical Analysis of the Galactose Phenotype of Dairy Streptococcus thermophilus Strains Reveals Four Different Fermentation Profiles Appl. Environ. Microbiol.**71**:3659–3667.

68. **Wu Q**, **Christine KW**, **Shah NP**. 2014. Towards galactose accumulation in dairy foods fermented by conventional starter cultures : Challenges and strategies. Trends Food Sci. Technol.

69. **Bourgoin F**, **Pluvinet A**, **Gintz B**, **Decaris B**, **Guédon G**. 1999. Are horizontal transfers involved in the evolution of the Streptococcus thermophilus exopolysaccharide synthesis loci? Gene **233**:151–161.

70. **Brigidi P**, **Swennen E**, **Vitali B**, **Rossi M**, **Matteuzzi D**. 2003. PCR detection of Bifidobacterium strains and Streptococcus thermophilus in feces of human subjects after oral bacteriotherapy and yogurt consumption. Int. J. Food Microbiol. **81**:203–209.

71. **Yang S-Y**, **Lü F-X**, **Lu Z-X**, **Bie X-M**, **Jiao Y**, **Sun L-J**, **Yu B**. 2008. Production of gamma-aminobutyric acid by Streptococcus salivarius subsp. thermophilus Y2 under submerged fermentation. Amino Acids **34**:473–8.

72. **Siragusa S**, **De Angelis M**, **Di Cagno R**, **Rizzello CG**, **Coda R**, **Gobbetti M**. 2007. Synthesis of gamma-aminobutyric acid by lactic acid bacteria isolated from a variety of Italian cheeses. Appl. Environ. Microbiol. **73**:7283–90.

73. **Bank TW**. 2012. Malnutrition and changing food systems, p. 13–25. *In* THe state of food and agriculture 2013.

74. **Stanton C**, **Ross RP**, **Fitzgerald GF**, **Sinderen D Van**. 2005. Fermented functional foods based on probiotics and their biogenic metabolites. Curr. Opin. Biotechnol. **16**:198–203.

75. **Iyer R**, **Tomar SK**. 2012. Folate and Prevention of Neural Tube Disease. *In* Neural Tube Defects – Role of Folate, Prevention Strategies and Genetics. InTech.

76. **Juarez M**, **Laiño JEE**, **Giori GS De**, **Leblanc JGG**, **Juarez del Valle M**, **Laiño JEE**, **Savoy de Giori G**, **Leblanc JGG**. 2014. Riboflavin producing lactic acid bacteria as a biotechnological strategy to obtain bio-enriched soymilk. Food Res. Int. **62**:1015–1019.

77. **Sybesma W**, **Burgess C**, **Starrenburg M**, **Van Sinderen D**, **Hugenholtz J**. 2004. Multivitamin production in Lactococcus lactis using metabolic engineering. Metab. Eng. **6**:109–115.

78. **Ibrahim GA**, **Sayed HSE-**, **El-shafei K**, **Sharaf OM**. 2015. Riboflavin and Folate Production in Different Media using Encapsulated Streptococcus Thermophilus and Lactobacillus Plantarum. Middle East J. Appl. Sci. **5**:663–669.

79. **Rasmussen TB**, **Danielsen M**, **Valina Garrigues C**, **Johansen E**, **Pedersen MB**, **Valina O**, **Garrigues C**, **Johansen E**, **Pedersen MB**. 2008. Streptococcus thermophilus core genome: comparative genome hybridization study of 47 strains. Appl. Environ. Microbiol. **74**:4703.

80. **Bolotin A**, **Quinquis B**, **Renault P**, **Sorokin A**, **Ehrlich SD**, **Kulakauskas S**, **Lapidus A**, **Goltsman E**, **Mazur M**, **Pusch GD**, **Fonstein M**, **Overbeek R**, **Kyprides N**, **Purnelle B**, **Prozzi D**, **Ngui K**, **Masuy D**, **Hancy F**, **Burteau S**, **Boutry M**, **Delcour J**, **Goffeau A**, **Hols P**. 2004. Complete sequence and comparative genome analysis of the dairy bacterium Streptococcus thermophilus. Nat. Biotechnol. **22**:1554–1558.

81. **Shendure J**, **Ji H**. 2008. Next-generation DNA sequencing. Nat. Biotechnol. **26**:1135–1145.

82. **Treu L**. 2012. A genomic and transcriptomic approach to characterize oenological Saccharomyces cerevisiae strains. PhD thesis. Padua University, Padua.

83. **Quail MA**, **Smith M**, **Coupland P**, **Otto TD**, **Harris SR**, **Connor TR**, **Bertoni A**, **Swerdlow HP**, **Gu Y**. 2012. A tale of three next generation sequencing platforms : comparison of Ion Torrent , Pacific Biosciences and Illumina MiSeq sequencers. BMC Genomics **13**:1.

84. **Makarova K**, **Slesarev A**, **Wolf Y**, **Sorokin A**, **Mirkin B**, **Koonin E**, **Pavlov a**, **Pavlova N**, **Karamychev V**, **Polouchine N**, **Shakhova V**, **Grigoriev I**, **Lou Y**, **Rohksar D**, **Lucas S**, **Huang K**, **Goodstein DM**, **Hawkins T**, **Dosti B**, **Plengvidhya V**, **Welker D**, **Hughes J**, **Goh Y**, **Benson A**, **Baldwin K**, **Smeianov V**, **Wechter W**, **Barabote R**, **Lorca G**, **Altermann E**, **Barrangou R**, **Ganesan B**, **Xie Y**, **Rawsthorne H**, **Tamir D**, **Parker C**, **Breidt F**, **Broadbent J**, **Hutkins R**, **Steele J**, **Unlu G**, **Saier M**, **Klaenhammer T**, **Richardson P**, **Kozyavkin S**, **Weimer B**, **Mills D**, **Lee J-H**, **Díaz-Muñiz I**, **Dosti B**, **Smeianov V**, **Wechter W**, **Barabote R**, **Lorca G**, **Altermann E**, **Barrangou R**, **Ganesan B**, **Xie Y**, **Rawsthorne H**, **Tamir D**, **Parker C**, **Breidt F**,

Broadbent J, **Hutkins R**, **O'Sullivan D**, **Steele J**, **Unlu G**, **Saier M**, **Klaenhammer T**, **Richardson P**, **Kozyavkin S**, **Weimer B**, **Mills D**. 2006. Comparative genomics of the lactic acid bacteria. Proc. Natl. Acad. Sci. U. S. A. **103**:15611–6.

85. **Delorme C**, **Bartholini C**, **Luraschi M**, **Pons N**, **Loux V**, **Almeida M**, **Guédon E**, **Gibrat J-F**, **Renault P**. 2011. Complete genome sequence of the pigmented Streptococcus thermophilus strain JIM8232. J. Bacteriol. **193**:5581–2.

86. **Sun Z**, **Chen X**, **Wang J**, **Zhao W**, **Shao Y**, **Wu L**, **Zhou Z**, **Sun T**, **Wang L**, **Meng H**, **Zhang H**, **Chen W**. 2011. Complete genome sequence of Streptococcus thermophilus strain ND03. J. Bacteriol. **193**:793–4.

87. **Kang X**, **Ling N**, **Sun G**, **Zhou Q**, **Zhang L**, **Sheng Q**. 2012. Complete genome sequence of Streptococcus thermophilus strain MN-ZLW-002. J. Bacteriol. **194**:4428–9.

88. **Wu Q**, **Tun HM**, **Leung FC**, **Shah NP**. 2014. Genomic insights into high exopolysaccharide-producing dairy starter bacterium Streptococcus thermophilus ASCC 1275 Sci Rep **4**: 17–21.

89. **Labrie SJ**, **Tremblay DM**, **Plante P**, **Wasserscheid J**, **Dewar K**, **Corbeil J**, **Moineau S**. 2015. Complete Genome Sequence of Streptococcus thermophilus SMQ-301 , a Model Strain for Phage-Host Interactions. Genome Announc. **3**:12–13.

90. **Shi Y**, **Chen Y**, **Li Z**, **Yang L**, **Chen W**, **Mu Z**. 2015. Complete Genome Sequence of Streptococcus thermophilus MN-BM- A02 , a Rare Strain with a High Acid-Producing Rate and Low Post- Acidification Ability. Genome Announc.**3**:2–3.

91. **Simpson JT**, **Wong K**, **Jackman SD**, **Schein JE**, **Jones SJM**. 2009. ABySS : A parallel assembler for short read sequence data. Genome research **19:** 1117–1123.

92. **Zerbino DR**, **Birney E**. 2008. Velvet : Algorithms for de novo short read assembly using de Bruijn graphs Genome research **18:** 821–829.

93. **Rissman AI**, **Mau B**, **Biehl BS**, **Darling AE**, **Glasner JD**, **Perna NT**. 2009. Reordering contigs of draft genomes using the Mauve Aligner. Bioinformatics **25**:2071–2073.

94. **Aziz RK**, **Bartels D**, **Best AA**, **Dejongh M**, **Disz T**, **Edwards RA**, **Formsma K**, **Gerdes S**, **Glass EM**, **Kubal M**, **Meyer F**, **Olsen GJ**, **Olson R**, **Osterman AL**, **Overbeek RA**, **Mcneil LK**, **Paarmann D**, **Paczian T**, **Parrello B**, **Pusch GD**, **Reich C**, **Stevens R**, **Vassieva O**, **Vonstein V**, **Wilke A**, **Zagnitko O**. 2008. The RAST Server : Rapid Annotations using Subsystems Technology. BMC Genomics **15**:1–15.

95. **Aziz RK**, **Devoid S**, **Disz T**, **Edwards R a.**, **Henry CS**, **Olsen GJ**, **Olson R**, **Overbeek R**, **Parrello B**, **Pusch GD**, **Stevens RL**, **Vonstein V**, **Xia F**. 2012. SEED Servers: High-Performance Access to the SEED Genomes, Annotations, and Metabolic Models. PLoS One **7**:e48053.

96. **Carver T**, **Harris SR**, **Berriman M**, **Parkhill J**, **McQuillan J a**. 2012. Artemis: an integrated platform for visualization and analysis of high-throughput sequence-based experimental data. Bioinformatics **28**:464–9.

97. **Tuimala J**. A primer to phylogenetic analysis using the PHYLIP package. CSC—Scientific Computing Ltd.: Espoo, Finland.

98. **Segata N**, **Bornigen D**, **Xochitl M**, **Hutternhower C**. 2013. PhyloPhlAn is a new method for improved phylogenetic and taxonomic placement of microbes. Nat. Commun. **18**:2304.

99. **Vernikos G**, **Parkhill J**. 2006. Genome analysis Interpolated variable order motifs for identification of horizontally acquired DNA : revisiting the Salmonella pathogenicity islands. Bioinformatics **22**:2196–2203.

100. **Bratlie MS**, **Johansen J**, **Sherman BT**, **Huang DW**, **Lempicki R a**, **Drabløs F**. 2010. Gene duplications in prokaryotes can be associated with environmental adaptation. BMC Genomics **11**:588.

101. **Tettelin H**, **Masignani V**, **Cieslewicz MJ**, **Donati C**, **Medini D**, **Ward NL**, **Angiuoli S V**, **Crabtree J**, **Jones AL**, **Durkin a S**, **Deboy RT**, **Davidsen TM**, **Mora M**, **Scarselli M**, **Margarit y Ros I**, **Peterson JD**, **Hauser CR**, **Sundaram JP**, **Nelson WC**, **Madupu R**, **Brinkac LM**, **Dodson RJ**, **Rosovitz MJ**, **Sullivan S a**, **Daugherty SC**, **Haft DH**, **Selengut J**, **Gwinn ML**, **Zhou L**, **Zafar N**, **Khouri H**, **Radune D**, **Dimitrov G**, **Watkins K**, **O'Connor KJB**, **Smith S**, **Utterback TR**, **White O**, **Rubens CE**, **Grandi G**, **Madoff LC**, **Kasper DL**, **Telford JL**, **Wessels MR**, **Rappuoli R**, **Fraser CM**. 2005. Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial "pan-genome". Proc. Natl. Acad. Sci. U. S. A. **102**:13950–5.

102. **Fu L**, **Niu B**, **Zhu Z**, **Wu S**, **Li W**. 2012. BIOINFORMATICS APPLICATIONS NOTE Sequence analysis CD-HIT : accelerated for clustering the next-generation sequencing data. Bioinformatics **28**:3150–3152.

103. **Campanaro S**, **Treu L**, **Vendramin V**, **Bovo B**, **Giacomini A**, **Corich V**. 2014. Metagenomic analysis of the microbial community in fermented grape marc reveals that Lactobacillus fabifermentans is one of the dominant species: Insights into its genome structure. Appl. Microbiol. Biotechnol. **98**:6015–6037.

104. **Saeed, A**, **Sharov V**, **White J**, **Li J**, **Liang W**, **Bhagabati N**, **Braisted J**, **Klapa M**, **Currier T**, **Thiagarajan M**, **Sturn A**, **Snuffin M**, **Rezantsev A**, **Popov D**, **Ryltsov A**, **Kostukovich E**, **Borisovsky I**, **Liu Z**, **Vinsavich A**, **Trush V**, **Quackenbush** J. 2003. TM4: a free, open-source system for microarray data management and analysis. Biotechniques **34**:374–348.

105. **R Development Core Team R**. 2008. Computational Many-Particle Physics. R Found. Stat. Comput. R Foundation for Statistical Computing.

106. **Brocchieri L**. 2014. Phylogenetics & Evolutionary Biology The GC Content of Bacterial Genomes **2**:2–4.

107. **Lassalle F**, **Périan S**, **Bataillon T**, **Nesme X**, **Duret L**, **Daubin V**. 2015. GC-Content Evolution in Bacterial Genomes: The Biased Gene Conversion Hypothesis Expands. PLOS Genet. **11**:e1004941.

108. **Makarova KS**, **Haft DH**, **Barrangou R**, **Brouns SJJ**, **Mojica FJM**, **Wolf YI**, **Yakunin AF**, **Oost J Van Der**, **Koonin E V**. 2011. Evolution and classification of the CRISPR–Cas systems. Nat. Publ. Gr. **9**:467–477.

109. **Makarova KS**, **Koonin E V.** 2007. Evolutionary Genomics of Lactic Acid Bacteria. J. Bacteriol. **189**:1199–1208.

110. **Nguyen TKC**, **Tran NP**, **Cavin J-F**. 2011. Genetic and biochemical analysis of PadR-padC promoter interactions during the phenolic acid stress response in Bacillus subtilis 168. J. Bacteriol. **193**:4180–91.

111. **Johansson P**, **Hederstedt L**. 1999. Organization of genes for tetrapyrrole biosynthesis in Gram-positive bacteria. Microbiology, **145**: 529–538.

112. **Sleator RD**, **Hill C**. 2002. Bacterial osmoadaptation: the role of osmolytes in bacterial stress and virulence. FEMS Microbiol. Rev. **26**:49–71.

113. **Serror P**, **Dervyn R**, **Ehrlich SD**, **Maguin E**. 2003. *csp* -like genes of *Lactobacillus delbrueckii* ssp. *bulgaricus* and their response to cold shock. FEMS Microbiol. Lett. **226**:323–330.

114. **Awad S**, **Hassan a N**, **Muthukumarappan K**. 2005. Application of exopolysaccharide-

producing cultures in reduced-fat Cheddar cheese: texture and melting properties. J. Dairy Sci. **88**:4204–13.

115. **Vonk RJ**, **Reckman GAR**, **Harmsen HJM**, **Priebe MG**. 2012. Probiotics and Lactose Intolerance, p. 149–160. *In* Probiotics.

116. **Vuyst L De**, **Weckx S**, **Ravyts F**, **Herman L**, **Leroy F**. 2011. New insights into the exopolysaccharide production of Streptococcus thermophilus. Int. Dairy J. **21**:586–591.

117. **Izawa N**, **Hanamizu T**, **Iizuka R**, **Sone T**, **Mizukoshi H**, **Kimura K**, **K C**. 2009. Streptococcus thermophilus produces exopolysaccharides including hyaluronic acid. J. Biosci. Bioeng. **107**:119–123.

118. **Pascuma M**, **Hebert EM**, **Mozzi F**, **Font De Valdez G**, **Pescuma M**, **Mozzi F**, **Valdez GF De**. 2007. Hydrolysis of whey proteins by Lactobacillus acidophilus , Streptococcus thermophilus and Lactobacillus delbrueckii ssp . bulgaricus grown in a chemically defined medium. J. Appl. Microbiol. **103**:1738–1746.

119. **Rossi F**, **Marzotto M**, **Cremonese S**, **Rizzotti L**, **Torriani S**. 2013. Diversity of Streptococcus thermophilus in bacteriocin production ; inhibitory spectrum and occurrence of thermophilin genes. Food Microbiol. **35**:27–33.

120. **Riley M a**. 1998. Molecular mechanisms of bacteriocin evolution. Annu. Rev. Genet. **32**:255–278.

121. **Zwietering MH**, **Jongenburger I**, **Rombouts FM**, **Van K**. 1990. Modeling of the Bacterial Growth Curve Modeling of the Bacterial Growth Curve. Appl. Environ. Microbio. **56**: 1875-1881.

122. **Baranyi J**, **Roberts T a**. 1994. A dynamic approach to predicting bacterial growth in food. Int. J. Food Microbiol. **23**:277–294.

123. **Huang L**. 2008. Growth kinetics of Listeria monocytogenes in broth and beef frankfurters--determination of lag phase duration and exponential growth rate under isothermal conditions. J. Food Sci. **73**:E235–42.

124. **Huang L**. 2010. Growth kinetics of Escherichia coli O157:H7 in mechanically-tenderized beef. Int. J. Food Microbiol. **140**:40–48.

125. **van den Bogaard PTC**, **Hols P**, **Kuipers OP**, **Kleerebezem M**, **de Vos WM**. 2004. Sugar utilisation and conservation of the gal-lac gene cluster in Streptococcus thermophilus. Syst. Appl. Microbiol. **27**:10–17.

126. **Tamura K**, **Stecher G**, **Peterson D**, **Filipski A**, **Kuman S**. 2013. MEGA: Molecular Evolutionary Genetics Analysis version 6. Mol. Biol. Evol. **30**:2725–2729.

127. **Morris LS**, **Evans J**, **Marchesi JR**. 2012. A robust plate assay for detection of extracellular microbial protease activity in metagenomic screens and pure cultures. J. Microbiol. Methods **91**:144–146.

128. **Maragkoudakis P a.**, **Nardi T**, **Bovo B**, **D'Andrea M**, **Howell KS**, **Giacomini A**, **Corich V**. 2013. Biodiversity, dynamics and ecology of bacterial community during grape marc storage for the production of grappa. Int. J. Food Microbiol. **162**:143–151.

129. **van Heel a. J**, **de Jong a.**, **Montalban-Lopez M**, **Kok J**, **Kuipers OP**. 2013. BAGEL3: automated identification of genes encoding bacteriocins and (non-)bactericidal posttranslationally modified peptides. Nucleic Acids Res. **41**:W448–W453.

130. **Letort C**, **Letort C**, **Juillard V**, **Juillard V**. 2001. Development of a minimal chemically-defined medium for the exponential growth of Streptococcus thermophilus. J. Appl. Microbiol. **91**:1023–1029.

131. **Linares DM**, **Kok J**, **Poolman B**. 2010. Genome sequences of Lactococcus lactis MG1363 (revised) and NZ9000 and comparative physiological studies. J. Bacteriol. **192**:5806–12.

132. **López S**, **Prieto M**, **Dijkstra J**, **Dhanoa MS**, **France J**. 2004. Statistical evaluation of mathematical models for microbial growth. Int. J. Food Microbiol. **96**:289–300.

133. **Vaillancourt K**, **Bédard N**, **Bart C**, **Tessier M**, **Robitaille G**, **Turgeon N**, **Frenette M**, **Moineau S**, **Vadeboncoeur C**. 2008. Role of galK and galM in galactose metabolism by Streptococcus thermophilus. Appl. Environ. Microbiol. **74**:1264–7.

134. **Ercolini D**, **Fusco V**, **Blaiotta G**, **Coppola S**. 2005. Sequence heterogeneity in the lacSZ operon of Streptococcus thermophilus and its use in PCR systems for strain differentiation **156**:161–172.

135. **Cochu a**, **Roy D**, **Vaillancourt K**, **Lemay JD**, **Casabon I**, **Frenette M**, **Moineau S**, **Vadeboncoeur C**. 2005. The doubly phosphorylated form of HPr, HPr(Ser~P)(His-P), is abundant in exponentially growing cells of *Streptococcus thermophilus* and phosphorylates the lactose transporter LacS as efficiently as HPr(His~P). Appl Env. Microbiol **71**:1364–1372.

136. **Vaughan EE**, **van den Bogaard PT**, **Catzeddu P**, **Kuipers OP**, **de Vos WM**. 2001. Activation of silent gal genes in the lac-gal regulon of Streptococcus thermophilus. J. Bacteriol. **183**:1184–94.

137. **Vaillancourt K**, **Moineau S**, **Frenette M**, **Lessard C**, **Vadeboncoeur C**. 2002. Galactose and lactose genes from the galactose-positive bacterium Streptococcus salivarius and the phylogenetically related galactose-negative bacterium Streptococcus thermophilus: Organization, sequence, transcription, and activity of the gal gene produc. J. Bacteriol. **184**:785–793.

138. **Stingele F**, **Neeser JR**, **Mollet B**, **Stingele F**, **Neeser J**. 1996. Identification and characterization of the eps (Exopolysaccharide) gene cluster from Streptococcus thermophilus Sfi6 . Identification and Characterization of the eps (Exopolysaccharide) Gene Cluster from Streptococcus thermophilus Sfi6. J. Bacteriol. **178**:1680.

139. **Arqués JL**, **Rodríguez E**, **Langa S**, **Landete JM**, **Medina M**. 2015. Antimicrobial Activity of Lactic Acid Bacteria in Dairy Products and Gut: Effect on Pathogens. Biomed Res. Int. **2015**:1–9.

140. **Willey JM**, **van der Donk W a**. 2007. Lantibiotics: peptides of diverse structure and function. Annu. Rev. Microbiol. **61**:477–501.

141. **Stoddard GW**, **Petzel JP**, **van Belkum MJ**, **Kok J**, **McKay LL**. 1992. Molecular analysis of the lactococcin A gene cluster from Lactococcus lactis subsp. lactis biovar diacetylactis WM4. Appl. Environ. Microbiol. **58**:1952–1961.

142. **Alting a. C**, **Engels WJM**, **Van Schalkwijk S**, **Exterkate F a.** 1995. Purification and characterization of cystathionine ??-lyase from Lactococcus lactis subsp. cremoris B78 and its possible role in flavor development in cheese. Appl. Environ. Microbiol. **61**:4037–4042.

143. **Garault P**, **Letort C**, **Juillard V**, **Monnet V**. 2000. Branched-chain amino acid biosynthesis is essential for optimal growth of Streptococcus thermophilus in milk. Appl Env. Microbiol **66**:5128–5133.

144. **Chopin a**. 1993. Organization and regulation of genes for amino acid biosynthesis in lactic acid bacteria. FEMS Microbiol. Rev **12**:21–37.

145. **Van Kranenburg R**, **Kleerebezem M**, **Van Hylckama Vlieg J**, **Ursing BM**, **Boekhorst J**, **Smit B a.**, **Ayad EHE**, **Smit G**, **Siezen RJ**. 2002. Flavour formation from amino acids by lactic acid

bacteria: Predictions from genome sequence analysis. Int. Dairy J. **12**:111–121.

146. **Obeid R**, **Pietrzik K**, **Jr GPO**, **Kancherla V**, **Holzgreve W**, **Wieser S**. 2015. Preventable Spina Bifida and Anencephaly in Europe. Birth Defects Res A Clin Mol Teratol **103**: 1–9

147. **Ohrvik VE**, **Witthoft CM**. 2011. Human Folate Bioavailability. Nutrients 475–490.

148. **Wouters JTM**, **Ayad EHE**, **Hugenholtz J**, **Smit G**. 2002. Microbes from raw milk for fermented dairy products. Int. Dairy J. **12**:91–109.

149. **Sybesma W**, **Starrenburg M**, **Tijsseling L**, **Hoefnagel MHN**, **Hugenholtz J**. 2003. Effects of Cultivation Conditions on Folate Production by Lactic Acid Bacteria. Metab. eng. **69**:4542–4548.

150. **D RIP**, **D SKTP**. 2011. Dietary effect of folate-rich fermented milk produced by Streptococcus thermophilus strains on hemoglobin level. Nutrition **27**:994–997.

151. **Laiño JE**, **Leblanc JG**, **Giori GS De**. 2012. Production of natural folates by lactic acid bacteria starter cultures isolated from artisanal Argentinean yogurts **588**:581–588.

152. **Laino JE**, **del Valle MJ**, **Savoy De Giori G**, **Leblanc JGJ**. 2013. Development of a high folate concentration yogurt naturally bio-enriched using selected lactic acid bacteria. LWT - Food Sci. Technol. **54**:1–5.

153. **Hugenschmidt S**, **Schwenninger SM**, **Gnehm N**, **Lacroix C**. 2010. Screening of a natural biodiversity of lactic and propionic acid bacteria for folate and vitamin B12 production in supplemented whey permeate. Int. Dairy J. **20**:852–857.

154. **Mousavi SS**, **Moeini H**, **Mohamad R**, **Dinarvand M**, **Ariff A**, **Ling FH**, **Raha R**. 2013. Effects of medium and culture conditions on folate production by Streptococcus thermophilus BAA-250. Res. Biotechnol. **4**:21–29.

155. **Maeda H**, **Dudareva N**. 2012. The Shikimate Pathway and Aromatic Amino Acid Biosynthesis in Plants. Annu. Rev. Plant Biol. **63**:73–105.

156. **de Crécy-Lagard V**, **El Yacoubi B**, **de la Garza RD**, **Noiriel A**, **Hanson AD**. 2007. Comparative genomics of bacterial and plant folate synthesis and salvage: predictions and validations. BMC Genomics **8**:245.

157. **Matsui K**, **Wang H-C**, **Hirota T**, **Hirokazu M**, **Sabu K**, **Kunihiro S**. 1982. Riboflavin production by roseoflavin-resistant strains of some bacteria. Agric. Biol. Chem. **48**:2003–2008.

158. **Burgess C**, **O'Connell-Motherway M**, **Sybesma W**, **Hugenholtz J**, **Van Sinderen D**. 2004. Ribo avin Production in. Appl. Environ. Microbiol. **70**:5769–5777.

159. **Perkins JB**, **Sloma a**, **Hermann T**, **Theriault K**, **Zachgo E**, **Erdenberger T**, **Hannett N**, **Chatterjee NP**, **Williams II V**, **Jr GR**, **Hatch R**, **Pero J**. 1999. Genetic engineering of Bacillus subtilis for the commercial production of riboflavin. J. Ind. Microbiol. Biotechnol. **22**:8–18.

160. **Russo P**, **Capozzi V**, **Arena MP**, **Spadaccino G**, **Dueñas MT**, **López P**, **Fiocco D**, **Spano G**. 2014. Riboflavin-overproducing strains of Lactobacillus fermentum for riboflavin-enriched bread. Appl. Microbiol. Biotechnol. **98**:3691–3700.

161. **Juarez del Valle M**, **Laiño JE**, **Savoy de Giori G**, **LeBlanc JG**. 2014. Riboflavin producing lactic acid bacteria as a biotechnological strategy to obtain bio-enriched soymilk. Food Res. Int. **62**:1015–1019.

162. **Laiño JE**, **de Giori GS**, **LeBlanc JG**. 2007. Folate production by lactic acid bacteria and other food-grade microorganisms, p. 329–339. *In* Communicating Current Research and Educational Topics and Trends in Applied Microbiology. A. Méndez-Vilas

163. **Horne DW**, **Patterson D**. 1988. Lactobacillus casei microbiological assay of folic acid derivatives in 96-well microtiter plates. Clin. Chem. **34**:2357–2359.

164. **Wegkamp A**. 2003. Metabolic engineering of folate production in lactic acid bacteria. PhD thesis. Wageningen University, Wageningen.

165. **Divya JB**, **Nampoothiri KM**. 2014. Folate fortification of skim milk by a probiotic Lactococcus lactis CM28 and evaluation of its stability in fermented milk on cold storage. J. Food Sci. Technol. **52**:3513–3519.

166. **Padalino M**, **Perez-Conesa D**, **López-Nicolás R**, **Frontela-Saseta C**, **Ros-Berruezo G**. 2012. Effect of fructooligosaccharides and galactooligosaccharides on the folate production of some folate-producing bacteria in media cultures or milk. Int. Dairy J. **27**:27–33.

167. **Douillard FP**, **de Vos WM**. 2014. Functional genomics of lactic acid bacteria: from food to health. Microb. Cell Fact. **13**:S8.

168. **Alkema W**, **Boekhorst J**, **Wels M**, **van Hijum S a. FT**. 2015. Microbial bioinformatics for food safety and production. Brief. Bioinform. 1–10.

169. **Leimena MM**, **Wels M**, **Bongers RS**, **Smid EJ**, **Zoetendal EG**, **Kleerebezem M**. 2012. Comparative analysis of Lactobacillus plantarum WCFS1 transcriptomes by using DNA microarray and next-Generation sequencing technologies. Appl. Environ. Microbiol. **78**:4141–4148.

170. **Bisanz JE**, **Macklaim JM**, **Gloor GB**, **Reid G**. 2014. Bacterial metatranscriptome analysis of a probiotic yogurt using an RNA-Seq approach. Int. Dairy J. **39**:284–292.

171. **Garber M**, **Grabherr MG**, **Guttman M**, **Trapnell C**. 2011. Computational methods for transcriptome annotation and quantification using RNA-seq. Nat. Methods **8**:469–477.

172. **Li B**, **Ruotti V**, **Stewart RM**, **Thomson JA**, **Dewey CN**. 2010. RNA-Seq gene expression estimation with read mapping uncertainty. Bioinformatics **26**:493–500.

173. **Creecy JP**, **Conway T**. 2014. Quantitative bacterial transcriptomics with RNA-seq. Curr. Opin. Microbiol. **23C**:133–140.

174. **Mortazavi A**, **Williams B**, **McCue K**, **Schaeffer L**, **Wold B**. 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat. Methods **5**:621 – 628.

175. **Gientka I**, **Duszkiewicz-Reinhard W**. 2009. Shikimate pathway in yeast cells: Enzymes, functioning, regulation - A review. Polish J. Food Nutr. Sci. **59**:113–118.

176. **Plach MG**, **Löffler P**, **Merkl R**, **Sterner R**. 2015. Conversion of anthranilate synthase into isochorismate synthase: implications for the evolution of chorismate-utilizing enzymes. Angew. Chem. Int. Ed. Engl. **54**:11270–4.

177. **Barona-Gómez F**, **Hodgson D a**. 2003. Occurrence of a putative ancient-like isomerase involved in histidine and tryptophan biosynthesis. EMBO Rep. **4**:296–300.

178. **Ely B**, **Pittard J**. 1979. Aromatic amino acid biosynthesis: regulation of shikimate kinase in Escherichia coli K-12. J Bacteriol **138**:933–943.

179. **García-Cañas V**, **Simó C**, **Herrero M**, **Ibáñez E**, **Cifuentes A**. 2012. Present and future challenges in food analysis: Foodomics. Anal. Chem. **84**:10150–10159.

180. **Tavaria FK**, **Dahl S**, **Carballo FJ**, **Malcata FX**. 2002. Amino Acid Catabolism and Generation of Volatiles by Lactic Acid Bacteria. J. Dairy Sci. 2462–2470.

List of abbreviations

bp	Base-pairs
CDS	Coding DNA sequence
Chr	Chromosome
DIP	Single-base indel polymorphism
GI	genome island
INDEL	Insertions and deltions of sequence
LTG	Lateral gene transfert
M17L	M17 broth added with 0.5% of sterile lactose after autoclaving
NGS	Next geenration sequencing

Media and solutions

TAE buffer (50X)

Tris base 242 g of, Acetic Acid 57.1 ml of, EDTA 100 ml of 0.5 M (pH 8.0), water to 1L.

Agarose Gel 1% (50 ml)

Agarose 0.5 g, TAE 1X buffer 50 ml, Sharpmass 2.5 µl.

PBS buffer

NaCl 137 mM, KCl 2.7 mM, Na_2HPO_4 10 mM, KH_2PO_4 2 mM, pH 7.4

M17 broth

Pancreatic digest of casein 5.0 g/l, soy peptone 5.0 g/l, beef extract 5.0 g/l, yeast extract 2.5 g/l, ascorbic acid 0.5 g/l, $MgSO_4$ 0.25 g/l, disodium-β –glycerophosphate 19.0 g/, pH 6.9

MRS broth

Beef extract 10 g/l, yeast extract 5 g/l, dextrose 20 g/l, Na Ac 5 g/l, polysorbate 80 1 g/l, KH_2PO_4 2 g/l, ammonium citrate 2 g/l $MgSO_4$ 0.1 g/l, $MnSO_4$ 0.05 g/l, pH 6.5

Baird Parker broth

Enzymatic digest of casein 10 g/l, beef extract 5 g/l, yeast extract 1 g/l, LiCl 5g/l, glycine 12 g/l, Na pyruvate 10 g/l, enriched with egg yolk 30%, potassium tellurite 0.15%, pH 7.0

Brain Heart Infusion Broth

Brain heart infusion 17.5 g/l, enzymatic digest of gelatin 10 g/l, dextrose 2 g/l NaCl 5 g/l, Na_2HPO_4 2.5 g/l, pH 7.4.

Nutrient Broth

`Lab-Lemco' powder 1 g/l, yeast extract 2.0 g/l, peptone 5.0 g/l, NaCl 5.0 g/l, pH 7.4.

CDM Chemically-define medium

Lactose 5.0 g/l, Na acetate 1.0 g/l, NH4 citrate 0.6 g/l, KH_2PO_4 3.0 g/l,K_2HPO_4 2.5 g/l, Urea 0.24 g/l, ascorbic acid 0.5 g/l, pyridoxamine HCl 0.8*10^{-3} g/l, nicotinic acid 0.1*10^{-3} g/l, riboflavine 0.05*10^{-3} g/l, Ca-pantothenate 0.1*10^{-3} g/l, thiamine HCl 0.005*10^{-3} g/l, MgCl2 $6H_20$ 0.16 g/l, $CaCl_2$ $2H_20$ 0.01g/l, aspartic acid 0.46 g/l, asparagine 0.46 g/l, glutamic acid 0.40 g/l, glutamine 0.39 g/l, lysine 0.44 g/l, arginine 0.13 g/l, histidine 0.15 g/l, proline 0.68 g/l, phenylalanine 0.28 g/l, tryptophane 0.05 g/l, methionine 0.13 g/l, alanine 0.24 g/l, valine 0.33 g/l, leucine 0.48 g/l, isoleucine 0.22 g/l, glycine 0.18 g/l, serine 0.34 g/l, threonine 0.23 g/l, cysteine 0.25 g/l, tyrosine 0.29 g/l, pH 6.4.

FACM Folic Acid Casei Medium

Charcoal treated pancreatic digest of casein 10.0 g/l, dextrose 40.0 g/l, Na Ac 40.0 g/l, dipotassium phosphate 1.0 g/l, monopotassium phosphate 1.0 g/l, DL-tryptophan 0.2 g/l, L-asparagine 0.6 g/l, L-cysteine HCL 0.5 g/l, adenine sulfate 10.0 mg/l, guanine HCL 10.0 mg/l, uracil 10.0 mg/l, xanthine 20.0 mg/l, polysorbate 80 0.1 g/l, glutathione (reduced) 5.0 mg/l, $MgSO_4$ 0.2 g/l, NaCl 20.0 mg/l, $FeSO_4$ 20.0 mg/l, M_nSO_4 15.0 mg/l, riboflavin 1.0 mg/l, p-aminobenzoic Acid 2.0 mg/l, pyridoxine HCl 4.0 mg/l, thiamine HCl 400.0 µg/l, Ca pantothenate 800.0 µg/l, nicotinic acid 800.0 µg/l, biotin 20.0 µg/l. pH 6.7.

Riboflavin Assay Medium

Dextrose 20.0 g/l, Sodium acetate 15.0 g/l, Vitamin assay casamino acids 10.0 g/l, Dipotassium phosphate 1.0 g/l, Monopotassium phosphate 1.0 g/l, L-asparagine 0.6 g/l, DL-tryptophan 0.2 g/l, L-cystine0.2 g/l, Magnesium sulfate USP 0.4 g/l,Adenine sulfate 20.0 mg/l, Guanine HCl 20.0 mg/l, Uracil 20.0 mg/l, Xanthine 20.0 mg/l, Ferrous Sulfate 20.0 mg/l, Manganese sulfate (monohydrate) 20.0 mg/l, NaCl 20.0 mg/l, Pyridoxine HCl 4.0 mg/l, Pyridoxal HCl 4.0 mg/l, p-aminobenzoic acid 2.0 mg/l, Calcium pantothenate 800.0 µg/l, folic acid 800.0 µg/l, nicotinic acid 800.0 µg/l, Thiamine HCl 400.0 µg/l, Biotin 1.0 µg/l. pH 6.8.

Bioinformatic tools and database

Here are reported sources of the bioinformatic tools and database cited in the main text.

Abyss assembly software

ABySS is a de novo sequence assembler designed for short reads and large genomes

http://www.bcgsc.ca/platform/bioinfo/software/abyss

Alien hunter

Alien hunter is an application for the prediction of putative horizontal gene transfer (HGT) events with the implementation of interpolated variable order motifs (IVOMs)

http://omictools.com/alien-hunter-tool

Artemis

Artemis is a free genome browser and annotation tool that allows visualisation of sequence features, next generation data and the results of analyses within the context of the sequence, and also its six-frame translation

http://www.sanger.ac.uk/science/tools/artemis

BAGEL3

BAGEL is a webbased bacteriocin mining tool

http://bagel.molgenrug.nl/index.php/bagel3

BLAST: Basic Local Alignment Search Tool

BLAST finds regions of similarity between biological sequences

http://blast.ncbi.nlm.nih.gov/Blast.cgi

CD-HIT

Cd-hit is a very widely used program for clustering and comparing protein or nucleotide sequences.

http://weizhongli-lab.org/cd-hit/

FASTX-Toolkit

FASTX-Toolkit is a collection of command line tools for Short-Reads FASTA/FASTQ files processing

http://hannonlab.cshl.edu/fastx_toolkit/

GeneBank

GeneBank is the NIH genetic sequence database

http://www.ncbi.nlm.nih.gov/genbank/

Glimmer: Gene Locator and Interpolated Markov ModelER

Glimmer is a system for finding genes in microbial DNA

http://www.cs.jhu.edu/~genomics/Glimmer/

MAUVE software

Mauve is a system for constructing multiple genome alignments in the presence of large-scale evolutionary events such as rearrangement and inversion.

http://darlinglab.org/mauve/mauve.html

MEGA: Molecular Evolutionary Genetics Analysis

MEGA is an user-friendly software suite for analyzing DNA and protein sequence data from species and populations

http://www.megasoftware.net/

MeV: MultiExperiment Viewer

MeV generates informative and interrelated displays of expression and annotation data from single or multiple experiments.

http://www.tm4.org/mev.html

Newbler assembler

Newbler is a software package for de novo DNA sequence assembly

http://swes.cals.arizona.edu/maier_lab/kartchner/documentation/index.php/home/docs/newbler

Pfam

Pfam database is a large collection of protein families

http://pfam.xfam.org/

RAST: Rapid Annotation using Subsystem Technology

RAST is a fully-automated service for annotating complete or nearly complete bacterial and archaeal genomes.

http://rast.nmpdr.org/

The SEED

The SEED is a platform for discovering and developing de novo annotations.

http://www.theseed.org/wiki/Main_Page

UniProt

UniProtprovide a comprehensive, high-quality and freely accessible resource of protein sequence and functional information.

http://www.uniprot.org

Velvet software

Velvet is a sequence assembler for very short reads

https://www.ebi.ac.uk/~zerbino/velvet/

Supplementary info

Are reported here for space reasons supplementary tables cited in the main text.

TABLE S1

IDENTIFIED GENES PLACED IN TWO ISLANDS PRESUMPTIVE ACQUIRED BY LATERAL GENE TRANSFER OF M17PTZA496

TABLE S2

PARALOGS OF M17PTZA496 DISTRIBUTED IN CLASSES. CLASSES IDENTIFY THE NUMBER OF STRAIN HAVING PARALOGOUS GENES IN THE CLUSTER. ALL: AT LEAST 5 OUT OF 7 OTHER STRAINS POSSESS PARALOGS IN THE CLUSTER. SEVERAL: UP TO FOUR STRAINS HAVE PARALOGS IN THE CLUSTER. WITH MTH17CL396: PARALOGS RECOGNISED ONLY IN M17PTZA496 AND MTH17CL396. SPECIFIC: ONLY M17PTZA496 HAS PARALOGS ASCRIBED IN THE CLUSTER. UNIQUE: ONLY THE M17PTZA496 PARALOGS CONSTITUTE THE CLUSTER. *: PARALOGS WITH 99% OF IDENTITY.

TABLE S3

SUMMARY OF INDICES USED TO DETERMINATE GOOD-FITTING, BIAS AND ACCURANCY OF THE THREE GROW CURVE STATISTIC MODEL. APPLIED MODEL IS HIGHLIGHTED IN GREEN.

TABLE S4

VALINE, LEUCINE AND ISOLEUCINE BIOSYNTHESIS PATHWAY RECOVERED FROM KEGG DATABASE. IN RED, GENES FOUND IN TH1436 AND TH1477.

TABLE S5

gal-lac OPERON INTERGENIC REGION COMPARISON. *galR-galM and galM-lacS* REGIONS ARE REPRESENTED IN TABLE a) AND b). THE -10 AND -35 REGIONS ARE HIGHLITED IN RED AND GREEN RESPECTIVELY.

TABLE S6

pH ACHIEVES AFTER 24 HOURS OF FERMENTATION IN WHEY. LS FISHER POST-HOC ANALYSES HAS DEFINED DIFFERENT GROUPS EXPRESSED FROM LETTERS NEAR THE VALUES. VALUE WITH THE SAME LETTER DO NOT DIFFER (P>0.05)

TABLE S7

GENES CHANGING IN EXPRESSION IN BOTH WHEY AND TRYPTOPHAN ENRICHED WHEY. IN ORANGE: HIGHER VALUE REGISTERED IN TH1477, IN GREEN: HIGHER VALUE RECORDED IN TH1436. BOLD: GENES OF FOLATE PATHWAY

TABLE S1

Protein	Sequence length	E-value	Matching
	ISLAND 1		
UDP pyrophosphate phosphatase	571	0.0	streptococci
Fe-S cluster assembly protein SufD	870	0.0	streptococci
Fe-S cluster assembly protein SufD	870	0.0	streptococci
Peptide ABC transporter substrate-binding protein	720	0.0	streptococci
Heat shock protein Hsp33	613	0.0	streptococci
Malate transporter	620	0.0	multispecie
Zinc ABC transporter permease	539	3 .00E-175	streptococci
PTS glucose transporter subunit IIABC	777	0.0	streptococci
Cytidylyltransferase	560	0.0	streptococci
Metalloprotease RseP	859	0.0	streptococci
Prolyl-tRNA ligase	1282	0.0	streptococci
Hypothetical protein	90	1 .00E-17	streptococci
ABC transporter ATP-binding protein	543	4 .00E-177	multispecies
	ISLAND 2		
Hypothetical protein BN871_AI_01260	169	2 .00E-41	Paenibacillus sp
Large-conductance mechanosensitive channel	248	3 .00E-72	streptococci
Hypothetical protein	227	2 .00E-59	Collinsella sp
Peptide ABC transporter ATP-binding protein	520	2 .00E-169	streptococci
Amino acid ABC transporter permease	669	0.0	streptococci
Histidine kinase	86	1 .00E-15	*S. salivarius*
Transporter drug/metabolite exporter family	98	5 .00E-20	*S. thermophilus*
Membrane protein	233	2 .00E-64	streptococci
Hypothetical protein	85	6 .00E-16	*Streptococcus sobrinus*
Lactoylglutathione lyase	122	3 .00E-28	multispecies
Hypothetical protein	55	1 .00E-05	*Streptococcus cristatus*
Toxin RelE	66	9 .00E-09	streptococci
Glycerol-3-phosphate acyltransferase domain protein	58	2 .00E-05	*S. pneumoniae*
Hypothetical protein	68	8 .00E-10	streptococci
Glycerol-3-phosphate acyltransferase	55	8 .00E-05	*S. pneumoniae*
Hypothetical protein BN871_HK_00030	170	2 .00E-41	Paenibacillus sp
Hypothetical protein	56	1 .00E-05	uncultured marine group
Alpha-amylase	983	0.0	streptococci
Tellurite resistance protein TehB	195	2 .00E-51	*S. thermophilus*

Exodeoxyribonuclease	528	2 .00E-171	streptococci
Hypothetical protein	57	2 .00E-05	*Bifidobacterium bifidum*
Transposase	73	7 .00E-12	streptococci
Truncated IS1193 transposase	58	5 .00E-06	streptococci
Peptidoglycan N-acetylglucosamine deacetylase	302	1 .00E-88	streptococci
Methyltransferase	614	0.0	*Streptococcus porci*
Superfamily II helicase domain protein	62	3 .00E-07	Gordonia sp. NB4-1Y
putative uncharacterized protein	84	5 .00E-14	Dialister invisus
Proton-coupled thiamine transporter YuaJ	399	1 .00E-124	*S. pneumoniae*
Hypothetical protein	134	6 .00E-31	streptococci
6-pyruvoyl-tetrahydropterin synthase	82	2 .00E-13	Burkholderia sp. SJ98

TABLE S2

All	Several	with MTH17CL396	Specific	Unique
ABC transporter	Mobile element protein	FIG01114374: hypothetical protein	UDP-glucose 4-epimerase (EC 5.1.3.2)	Conserved hypothetical protein TIGR00730*
ABC transporter (EC 3.A.1.5.1)	Mobile element protein	FIG01114697: hypothetical protein	LSU ribosomal protein L29p (L35e)	Mobile element protein*
Ribosomal RNA large subunit	Twin-arginine translocation protein	FIG01114232: hypothetical protein	DNA-directed RNA polymerase alpha subunit (EC 2.7.7.6)-	Mobile element protein*
Methyltransferase N (EC 2.1.1.-)	TatCd		Fructose-bisphosphate aldolase class II (EC 4.1.2.13)	
DNA-methyltransferase subunit M (EC 2.1.1.72)	Choline binding protein D		Choline binding protein A	
hypothetical protein BH3604	Mobile element protein		Mobile element protein	
Iron-sulfur cluster assembly protein SufB	Mobile element protein		SSU ribosomal protein S3p (S3e)	
Mobile element protein	Hypothetical protein . truncated		Peptide ABC transporter ATP binding protein . putative . truncated-	
ABC transporter	Cysteine ABC transporter . substrate-binding protein		LSU ribosomal protein L5p (L11e)	
ABC transporter oligopeptide (TC 3.A.1.5.1)	platelet activating factor . putative		LSU ribosomal protein L6p (L9e)	
	Mobile element protein		Cysteinyl-tRNA synthetase related protein	
	Mobile element protein*		Hypothetical protein	
			SSU ribosomal protein S5p (S2e)	
			LSU ribosomal protein L13p (L13Ae)	
			LSU ribosomal protein L15p (L27Ae)	
			LSU ribosomal protein L16p (L10e)	
			SSU ribosomal protein S8p (S15Ae)	
			SSU ribosomal protein S9p (S16e)	
			SSU ribosomal protein S11p (S14e)	
			LSU ribosomal protein L14p (L23e)	
			SSU ribosomal protein S13p (S18e)	
			Mobile element protein	
			LSU ribosomal protein L18p (L5e)	

Rossmann fold nucleotide-binding protein Smf
LSU ribosomal protein L24p (L26e)
Mobile element protein*
SSU ribosomal protein S19p (S15e)
SSU ribosomal protein S17p (S11e)*
Acyl carrier protein
LSU ribosomal protein L29p (L35e)
Mobile element protein*
SSU ribosomal protein S14p (S29e) . zinc-dependent
LSU ribosomal protein L36p

TABLE S3

1F8CT	Gompertz	Baranyi	Huang	***TH985***	Gompertz	Baranyi	Huang
BF	0.992	1.003	1.011	BF	0.990	0.997	0.937
AF	1.106	1.013	1.029	AF	1.052	1.059	1.066
MSE	0.324	0.028	0.065	MSE	0.065	0.067	0.474
AIC	1.151	5.828	1.439	AIC	-10.247	-8.593	11.654
M17PTZA496	Gompertz	Baranyi	Huang	***TH1435***	Gompertz	Baranyi	Huang
BF	1.010	1.011	1.030	BF	0.99	1.03	1.00
AF	1.113	1.082	1.117	AF	1.05	1.13	1.04
MSE	0.157	0.105	0.187	MSE	0.07	0.23	0.02
AIC	-2.340	4.400	9.637	AIC	-10.25	11.65	-8.59
MTH17CL396	Gompertz	Baranyi	Huang	***TH1436***	Gompertz	Baranyi	Huang
BF	0.980	0.986	0.997	BF	0.971	0.971	0.971
AF	1.203	1.208	1.061	AF	1.101	1.101	1.101
SSE	1.093	0.856	0.075	MSE	0.370	0.297	0.474
AIC	11.324	16.790	-6.036	AIC	-2.045	13.787	17.997
TH982	Gompertz	Baranyi	Huang	***TH1477***	Gompertz	Baranyi	Huang
BF	1.012	1.004	1.009	BF	0.991	0.981	0.981
AF	1.944	1.197	1.231	AF	1.051	1.060	1.082
MSE	0.589	0.065	0.086	MSE	0.187	0.256	0.267
AIC	9.583	12.554	14.834	AIC	-2.721	8.749	9.200

TABLE S4

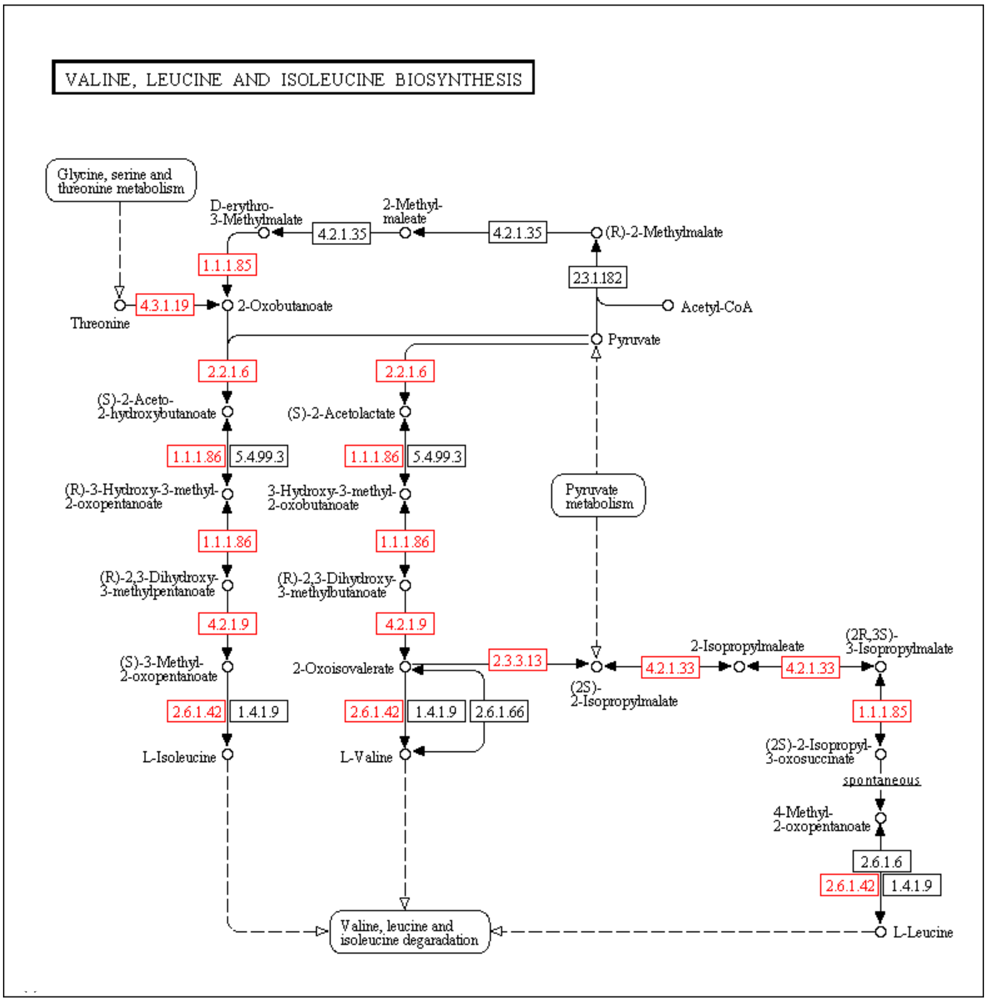

VALINE, LEUCINE AND ISOLEUCINE BIOSYNTHESIS
Glycine, serine and threonine metabolism
D-erythro-3-Methylmalate
4.2.1.35
2-Methyl-maleate
4.2.1.35
(R)-2-Methylmalate
2.3.1.182
Acetyl-CoA
1.1.1.85
Threonine
4.3.1.19
2-Oxobutanoate
Pyruvate
2.2.1.6
2.2.1.6
(S)-2-Aceto-2-hydroxybutanoate
(S)-2-Acetolactate
1.1.1.86
5.4.99.3
1.1.1.86
5.4.99.3
(R)-3-Hydroxy-3-methyl-2-oxopentanoate
3-Hydroxy-3-methyl-2-oxobutanoate
Pyruvate metabolism
1.1.1.86
1.1.1.86
(R)-2,3-Dihydroxy-3-methylpentanoate
(R)-2,3-Dihydroxy-3-methylbutanoate
4.2.1.9
4.2.1.9
(S)-3-Methyl-2-oxopentanoate
2-Oxoisovalerate
2.3.3.13
(2S)-2-Isopropylmalate
4.2.1.33
2-Isopropylmaleate
4.2.1.33
(2R,3S)-3-Isopropylmalate
2.6.1.42
1.4.1.9
2.6.1.42
1.4.1.9
2.6.1.66
1.1.1.85
L-Isoleucine
L-Valine
(2S)-2-Isopropyl-3-oxosuccinate
spontaneous
4-Methyl-2-oxopentanoate
2.6.1.6
2.6.1.42
1.4.1.9
Valine, leucine and isoleucine degaradation
L-Leucine

TABLE S5

a)

LMG 18311				A	G	T	A	T	C	C	T	C	C	T	C	A	T	A	T	T	T	C	A	G	T	A	T	A	A	C
1F8CT				.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
M17PTZA496				.	.	.	.	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
MTH17CL396				.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
TH982	*galR* ←			.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
TH985				.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
TH1435				.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
TH1436				.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
TH1477				.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.

LMG 18311	A	T	A	A	C	T	T	T	T	A	T	T	T	T	T	T	T	A	C	C	T	A	T	A	T	T	T	T	A	C
1F8CT	.	.	.	.	.	.	.	.	.	.	-	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
M17PTZA496	.	.	.	.	.	.	.	.	.	C	-	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
MTH17CL396	.	.	.	.	.	.	.	.	.	.	-	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
TH982	.	.	.	.	.	.	.	.	.	.	-	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
TH985	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
TH1435	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
TH1436	.	.	.	.	.	.	.	.	.	C	-	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
TH1477	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.

LMG 18311	T	-	A	A	A	A	A	A	A	T	A	G	T	A	A	A	A	A	T	A	T	T	G	A	T	T	T	T	C	C
1F8CT	.	-	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
M17PTZA496	.	-	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.	.	T	.
MTH17CL396	.	-	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
TH982	.	-	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
TH985	.	-	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
TH1435	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
TH1436	.	-	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.	.	.	.
TH1477	.	-	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.	.	T	.

LMG 18311	A	T	G	T	G	A	A	A	G	G	G	G	T	T	A	C	G	A	T	T	T	C	A	G	T	A	T	A	A	A
1F8CT	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.	.
M17PTZA496	.	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	A	.	.	.	.	.	.	.	.	G	.	.	.	.
MTH17CL396	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
TH982	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.	.
TH985	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
TH1435	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
TH1436	.	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	A	.	.	.	.	.	.	.	.	G	.	.	.	.
TH1477	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.

LMG 18311	C	A	A	A	A	A	G	A	A	T	A	A	G	T	G	A	G	A	T	A	C	A	T	C	C	T				
1F8CT	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.				
M17PTZA496	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	G	.	.	.				
MTH17CL396	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.				
TH982	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.			*galK* →	
TH985	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.				
TH1435	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.				
TH1436	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	G	.	.	.				
TH1477	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	G	.	.	.				

b)

LMG 18311		A	C	C	A	T	G	T	A	T	T	A	G	T	A	A	A	A	T	T	T	T
1F8CT		.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
M17PTZA496		.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	G	.	.	.
MTH17CL396		.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	G	.	.	.
TH982	*galM*	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
TH985		.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
TH1435		.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
TH1436		.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	G	.	.	.
TH1477		.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.

LMG 18311	A	G	T	A	A	A	A	A	A	C	-	A	C	T	G	A	A	A	T	T	A	T	T	G	A	C	T	G	C	A
1F8CT	.	.	.	.	.	.	.	.	.	.	-	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
M17PTZA496	.	.	.	.	.	.	.	.	.	.	T	.	.	.	A	.	.	.	C	.	.	.	.	.	.	.	A	A	T	T
MTH17CL396	.	.	.	.	.	.	.	.	.	.	-	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
TH982	.	.	.	.	.	.	.	.	.	.	-	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
TH985	.	.	.	.	.	.	.	.	.	.	-	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
TH1435	.	.	.	.	.	.	.	.	.	.	-	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
TH1436	.	.	.	.	.	.	.	.	.	.	-	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
TH1477	.	.	.	.	.	.	.	.	.	.	-	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.

LMG 18311	T	A	A	A	C	C	A	A	T	T	T	T	C	A	T	A	T	A	A	T	G	T	A	A	A	C	G	T	A	T
1F8CT	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
M17PTZA496	C	.	.	.	.	T	.	.	.	A	C	.	T	G	.	.	.	.	.	.	.	G	.	.	G	.	.	.	.	.
MTH17CL396	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
TH982	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
TH985	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
TH1435	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
TH1436	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
TH1477	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.

LMG 18311	T	C	-	-	A	A	A	T	A	A	T	A	G	G	A	G	G	T	T	T	C	C	G	A	A	
1F8CT	.	.	-	-	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	
M17PTZA496	.	.	A	A	.	.	.	.	.	.	C	.	.	.	.	.	.	.	.	C	T	.	A	T	.	
MTH17CL396	.	.	-	-	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	
TH982	.	.	-	-	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	*lacS* →
TH985	.	.	-	-	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	
TH1435	.	.	-	-	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	
TH1436	.	.	-	-	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	
TH1477	.	.	-	-	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	

TABLE S6

Strain ID	Whey	Whey+ Gly 10 mM	Whey+ Phe 10 mM	Whey+ Tyr 10 mM	Whey+ Trp 10 mM	Whey+ Gly 20 mM	Whey+ Phe 20 mM	Whey+ Tyr 20 mM	Whey+ Trp 20 mM	Whey+ Gly 50 mM	Whey+ Phe 50 mM	Whey+ Tyr 50 mM	Whey+ Trp 50 mM	p-value
1F8CT	5.33^{c} (0.041)	5.54 (0.163)	5.61 (0.588)	5.57 (0.165)	5.46 (0.143)	5.66 (0.592)	5.78 (0.226)	5.67 (0.225)	5.64 0.169)	5.71 (0.476)	5.60 (0.268)	5.68 (0.296)	5.83 (0.077)	n.s.
M17PTZ A496	4.64^{c} (0.012)	4.66^{c} (0.026)	4.63^{c} (0.016)	4.60^{c} (0.015)	4.66^{c} (0.025)	4.67^{c} (0.025)	4.94^{ab} (0.561)	4.59^{bc} (0.005)	4.71^{c} (0.009)	4.82^{abc} (0.031)	4.61^{c} (0.025)	4.60^{c} (0.005)	4.98^{a} (0.030)	<0.01
MTH17C L396	4.57^{c} (0.046)	4.59^{ab} (0.049)	4.61^{ab} (0.050)	4.58^{c} (0.015)	4.62^{ab} (0.052)	4.67^{ab} (0.099)	4.66^{ab} (0.035)	4.60^{ab} (0.027)	4.69^{ab} (0.062)	4.70^{b} (0.047)	4.68^{ab} (0.066)	4.61^{ab} (0.020)	5.40^{a} (0.192)	<0.01
TH982	4.51^{de} (0.089)	4.50^{e} (0.049)	4.52^{de} (0.040)	4.48^{e} (0.020)	4.54^{cde} (0.058)	4.51^{de} (0.015)	4.53^{de} (0.026)	4.49^{e} (0.015)	4.60^{bc} (0.026)	4.57^{cd} (0.012)	4.66^{b} (0.023)	4.50^{e} (0.009)	4.93^{a} (0.061)	<0.01
TH985	4.94 (0.456)	4.94 (0.395)	4.85 (0.359)	4.99 (0.311)	5.02 (0.405)	4.87 (0.286)	4.90 (0.353)	5.20 (0.025)	5.0 (0.244)	4.82 (0.191)	4.90 (0.272)	5.13 (0.041)	5.36 (0.072)	n.s.
TH1435	4.36^{bcd} (0.051)	4.39^{bcd} (0.046)	4.39^{bcd} (0.046)	4.36^{cd} (0.028)	4.34^{bcd} (0.006)	4.39^{b} (0.058)	4.41^{bcd} (0.049)	4.38^{d} (0.070)	4.33^{bc} (0.005)	4.41^{a} 0.026)	4.35^{bcd} (0.005)	4.33^{d} (0.005)	4.57^{a} (0.050)	<0.01
TH1436	4.37^{b} 0.097)	4.39^{b} 0.091)	4.39^{b} (0.057)	4.36^{b} (0.015)	4.41^{b} (0.098)	4.41^{b} (0.089)	4.38^{b} (0.042)	4.37^{b} (0.021)	4.44^{b} (0.067)	4.46^{b} (0.080)	4.41^{b} (0.023)	4.37^{b} (0.020)	4.69^{a} (0.071)	<0.01
TH1477	4.30^{e} (0.046)	4.32^{e} 0.036)	4.35^{cde} (0.030)	4.31^{bcd} (0.012)	4.40^{e} (0.104)	4.34^{cde} (0.045)	4.41^{bc} (0.070)	4.32^{de} (0.005)	4.45^{b} (0.071)	4.33^{cde} (0.012)	4.47^{b} (0.015)	4.33^{cde} (0.012)	4.58^{a} (0.040)	<0.01

TABLE S7

geneID	log$_2$FC whey	p-value whey	log2FC wheyTRP	p-value wheyTRP	Function
gene_1744	**1.31**	1.05E-02	0.95	1.87E-06	**5-formyltetrahydrofolate cyclo-ligase (EC 6.3.3.2)**
gene_1232	**1.41**	3.53E-03	0.74	1.66E-03	ABC transporter ATP binding protein
gene_1414	0.88	4.01E-02	**1.11**	3.18E-10	ABC-type multidrug transport system, ATPase component
gene_1702	**1.81**	1.63E-06	0.74	1.86E-04	Acetyltransferase (EC 2.3.1.-)
gene_1604	**1.39**	9.80E-03	0.95	3.52E-05	Acetyltransferase (EC 2.3.1.-)
gene_1764	0.94	3.23E-02	**2.44**	1.27E-28	Aggregation promoting factor
gene_0836	**1.09**	2.10E-02	0.98	7.52E-05	Agmatinase (EC 3.5.3.11)
gene_1305	**1.19**	2.64E-03	0.60	1.76E-03	Aldose 1-epimerase (EC 5.1.3.3)
gene_0205	**1.54**	1.71E-04	0.59	4.14E-03	Amino acid ABC transporter, glutamine-binding protein
gene_0204	**1.63**	3.27E-05	0.59	1.17E-03	Amino acid transport ATP-binding protein
gene_1345	0.91	3.73E-02	**1.16**	1.28E-04	**Anthranilate phosphoribosyltransferase (EC 2.4.2.18)**
gene_1347	0.91	1.94E-02	**1.43**	5.98E-12	**Anthranilate synthase, aminase component (EC 4.1.3.27)**
gene_1689	**1.12**	6.38E-03	0.87	6.03E-05	Argininosuccinate synthase (EC 6.3.4.5)
gene_1557	**1.03**	4.71E-03	0.66	2.64E-04	Aspartate aminotransferase (EC 2.6.1.1)
gene_0473	**1.01**	8.68E-03	0.60	2.92E-03	ATP synthase A chain (EC 3.6.3.14)
gene_0471	0.96	3.32E-02	**1.09**	8.12E-09	ATP synthase C chain (EC 3.6.3.14)
gene_0059	**1.34**	7.84E-04	0.65	3.56E-03	ATP-dependent Clp protease ATP-binding subunit ClpA
gene_1464	**1.40**	3.33E-04	0.71	1.43E-04	ATP-dependent RNA helicase YqfR
gene_1889	**1.27**	4.08E-03	0.70	3.24E-03	Chromosome (plasmid) partitioning protein ParB
gene_1171	**1.65**	5.51E-06	1.14	4.20E-10	Chromosome partition protein smc
gene_1085	**1.48**	7.16E-03	1.00	1.44E-08	Chromosome replication initiation protein dnaD
gene_1652	**1.11**	1.02E-02	0.90	1.09E-03	COG2110, Macro domain, possibly ADP-ribose binding module
gene_1415	0.89	3.58E-02	**1.04**	1.77E-08	Cysteine ABC transporter, permease protein
gene_1603	**1.26**	2.98E-02	0.87	1.25E-04	D-3-phosphoglycerate dehydrogenase (EC 1.1.1.95)
gene_0721	**1.08**	6.99E-03	0.85	7.87E-05	D-alanyl transfer protein DltB
gene_0060	**1.53**	6.58E-05	0.65	8.07E-04	D-alanyl-D-alanine carboxypeptidase (EC 3.4.16.4)
gene_0192	**1.65**	8.62E-04	0.89	2.37E-05	Dextranase precursor (EC 3.2.1.11)
gene_0352	**1.40**	2.32E-04	0.75	5.24E-05	DNA polymerase III delta prime subunit (EC 2.7.7.7)
gene_1776	**1.61**	1.26E-05	0.62	2.42E-03	DNA-directed RNA polymerase beta' subunit (EC 2.7.7.6)
gene_0363	0.86	2.04E-02	**1.63**	4.51E-19	Enoyl-[acyl-carrier-protein] reductase [FMN] (EC 1.3.1.9)
gene_1377	**1.12**	6.18E-03	0.71	8.59E-04	FIG004454: RNA binding protein
gene_0119	**1.36**	4.95E-04	0.72	1.11E-03	FIG005986: HD family hydrolase
gene_0127	**2.17**	3.44E-09	0.98	8.03E-07	FIG009886: phosphoesterase
gene_0118	**1.42**	6.10E-04	0.61	3.87E-03	FIG011178: rRNA methylase

gene_0704	1.31	4.71E-03	0.78	3.51E-04	FIG146085: 3'-to-5' oligoribonuclease A, Bacillus type
gene_0877	2.04	4.81E-08	0.89	2.58E-05	Glutamine amidotransferase, class I
gene_0805	1.33	1.61E-03	0.77	1.04E-03	GMP synthase [glutamine-hydrolyzing], amidotransferase subunit (EC 6.3.5.2)
gene_1128	1.47	1.47E-03	0.95	1.47E-06	GtrA family protein; MesH protein
gene_1365	0.83	3.18E-02	1.56	1.04E-14	Histidine triad (HIT) nucleotide-binding protein, similarity with At5g48545 and yeast YDL125C (HNT1)
gene_0343	1.48	1.12E-03	0.97	1.35E-07	Hydrolase (HAD superfamily)
gene_1136	2.00	1.15E-06	0.72	1.29E-03	Hydrolase, alpha/beta hydrolase fold family
gene_0002	1.07	6.78E-03	0.86	8.63E-06	Hypoxanthine-guanine phosphoribosyltransferase (EC 2.4.2.8)
gene_0738	0.84	3.07E-02	1.54	2.61E-17	Lead, cadmium, zinc and mercury transporting ATPase (EC 3.6.3.3) (EC 3.6.3.5); Copper-translocating P-type ATPase (EC 3.6.3.4)
gene_0269	0.76	4.48E-02	1.12	2.60E-08	Leucyl-tRNA synthetase (EC 6.1.1.4)
gene_1156	1.30	1.77E-03	0.76	2.33E-04	L-lactate dehydrogenase (EC 1.1.1.27)
gene_1855	0.81	3.54E-02	1.03	1.96E-06	LSU ribosomal protein L16p (L10e)
gene_1834	1.17	4.88E-03	0.99	8.49E-06	LSU ribosomal protein L17p
gene_1845	1.22	3.71E-03	1.00	4.05E-05	LSU ribosomal protein L18p (L5e)
gene_1843	1.13	1.12E-02	0.60	3.93E-03	LSU ribosomal protein L30p (L7e)
gene_0702	1.31	2.86E-03	0.91	3.96E-03	LSU ribosomal protein L31p
gene_1862	1.03	1.47E-02	0.72	2.28E-04	LSU ribosomal protein L3p (L3e)
gene_0642	1.05	3.55E-03	0.99	6.80E-06	Manganese transport protein MntH
gene_1738	1.04	2.50E-02	0.70	3.71E-04	Membrane-bound protease, CAAX family
gene_0688	1.08	1.05E-02	0.67	2.66E-03	Na+ driven multidrug efflux pump
gene_0275	1.24	3.06E-03	0.78	5.40E-04	NAD synthetase (EC 6.3.1.5)
gene_0207	1.65	7.43E-05	0.86	1.36E-06	Negative regulator of genetic competence MecA
gene_0274	1.44	1.20E-04	0.96	1.94E-06	Nicotinate phosphoribosyltransferase (EC 2.4.2.11)
gene_1209	1.17	1.20E-03	0.89	1.07E-05	Oligopeptide transport ATP-binding protein OppF (TC 3.A.1.5.1)
gene_0417	1.20	2.98E-03	0.84	1.17E-03	Peptide deformylase (EC 3.5.1.88)
gene_1244	2.59	1.49E-11	0.73	3.86E-03	Peptide methionine sulfoxide reductase MsrA (EC 1.8.4.11)
gene_0709	0.97	3.23E-02	1.22	5.49E-06	Peroxide stress regulator PerR, FUR family
gene_1129	1.13	4.48E-03	0.71	1.71E-04	Phosphoglucosamine mutase (EC 5.4.2.10)
gene_0540	0.85	3.34E-02	1.33	1.05E-10	Phosphomevalonate kinase (EC 2.7.4.2)
gene_0736	1.08	4.69E-02	0.53	3.31E-03	Phosphopantothenoylcysteine synthetase (EC 6.3.2.5)
gene_0027	1.30	2.51E-03	0.89	5.56E-06	Phosphoribosylamine--glycine ligase (EC 6.3.4.13)
gene_0029	1.31	5.21E-04	0.79	1.46E-04	Phosphoribosylaminoimidazole carboxylase ATPase subunit (EC 4.1.1.21)
gene_1343	0.90	3.84E-02	1.70	2.90E-06	**Phosphoribosylanthranilate isomerase (EC 5.3.1.24)**
gene_0016	1.44	3.22E-04	0.70	1.24E-03	Phosphoribosylformylglycinamidine synthase, synthetase subunit (EC

					6.3.5.3)
gene_1593	**1.76**	1.14E-06	0.69	1.93E-04	Phosphoserine phosphatase (EC 3.1.3.3)
gene_0875	**2.50**	7.07E-09	0.97	3.39E-07	Purine nucleoside phosphorylase (EC 2.4.2.1)
gene_0779	**1.24**	2.33E-03	0.62	3.85E-03	Putative deoxyribose-specific ABC transporter, ATP-binding protein
gene_0781	**1.31**	2.94E-03	0.71	4.65E-04	Putative deoxyribose-specific ABC transporter, permease protein
gene_0780	**1.13**	4.43E-03	0.66	2.04E-03	Putative deoxyribose-specific ABC transporter, permease protein
gene_1654	**1.01**	1.21E-02	0.92	4.66E-04	putative transport accessory protein
gene_0953	**1.18**	6.44E-03	0.88	1.88E-06	putative Zn-dependent protease
gene_1419	**1.15**	7.81E-03	0.81	8.55E-05	Pyruvate formate-lyase (EC 2.3.1.54)
gene_1638	**1.14**	1.12E-02	0.60	1.38E-03	Ribonuclease BN (EC 3.1.-.-)
gene_0814	**1.93**	1.76E-06	1.53	4.59E-14	Ribonuclease HII (EC 3.1.26.4)
gene_1704	**2.74**	1.91E-09	0.74	2.73E-04	Ribonucleotide reductase of class III (anaerobic), activating protein (EC 1.97.1.4)
gene_0132	**2.00**	2.20E-07	0.69	6.44E-04	Ribosomal large subunit pseudouridine synthase B (EC 4.2.1.70)
gene_1641	**1.04**	3.59E-02	0.96	2.73E-06	Ribosomal-protein-S5p-alanine acetyltransferase
gene_0130	**2.09**	5.21E-08	0.66	8.29E-04	Segregation and condensation protein A
gene_0131	**2.14**	2.23E-08	0.59	4.10E-03	Segregation and condensation protein B
gene_1853	**1.04**	6.69E-03	0.68	5.31E-04	SSU ribosomal protein S17p (S11e)
gene_1660	**1.09**	2.82E-03	0.91	5.03E-07	Thiamin pyrophosphokinase (EC 2.7.6.2)
gene_0462	0.89	2.17E-02	**1.14**	1.34E-11	Trans-2,cis-3-Decenoyl-ACP isomerase
gene_0261	**1.26**	4.40E-03	0.92	2.96E-05	Transcription antitermination protein NusG
gene_0291	**1.77**	1.29E-04	0.82	1.91E-04	Transcription elongation factor GreA
gene_0048	**1.61**	3.00E-04	0.85	5.31E-04	Transcriptional regulator SpxA2
gene_1875	**1.19**	2.90E-02	0.99	2.72E-04	transcriptional regulator, Cro/CI family
gene_0349	**1.37**	1.72E-03	0.63	3.96E-03	Translation elongation factor Tu
gene_0954	**1.13**	7.45E-03	0.81	8.99E-05	Tributyrin esterase
gene_1012	**1.81**	1.01E-06	2.05	7.03E-25	tRNA and rRNA cytosine-C5-methylases
gene_1342	0.88	3.80E-02	**1.41**	1.38E-07	**Tryptophan synthase beta chain (EC 4.2.1.20)**
gene_1880	0.81	2.66E-02	**1.00**	3.12E-09	**Tryptophanyl-tRNA synthetase (EC 6.1.1.2)**
gene_0742	**1.19**	6.99E-03	0.76	3.45E-04	Two component system sensor histidine kinase CiaH (EC 2.7.3.-)
gene_0129	**2.26**	1.24E-09	0.91	6.95E-07	Tyrosine recombinase XerD
gene_1780	**1.21**	3.23E-02	0.95	2.45E-04	Tyrosyl-tRNA synthetase (EC 6.1.1.1)
gene_0151	0.96	2.29E-02	**1.10**	9.08E-08	Urease accessory protein UreE
gene_1457	**1.55**	1.44E-05	0.67	1.08E-03	VanZF-related protein