Identification of ΔTOM1 suppressors in *S. Cerevisiae*

Thibault Schowing, Lionel Rohner, Alain Rohrbasser, Rares Cristea

Introduction

The budding yeast *Saccharomyces cerevisiae* (*S.cerevisiae*) is a single-celled lower eukaryote belonging to the kingdom of fungi. Ever since its discovery, *S.cerevisae* has nourished human advancements in the field of fermented food products, alcoholic beverages, and nowadays, the production of biofuel. Beyond its contributions to industrial fermentation, *S.cerevisiae* has become one of the most popular model organism in eukaryotic biology, due to its simple cellular architecture, cheap maintenance cost, fast growth, and homologies to human cells [1]. Since the genome of *S.cerevisiae* is fully sequenced, relatively small and virtually free of intronic DNA, this microorganism is ideally suited for genetic analyses.

In this project, we focused on identifying novel mutations that may suppress the phenotypic alteration caused by the deletion of the gene Trigger of mitosis, abbreviated *TOM1*. *TOM1* is a gene located on chromosome 4 and codes for a Hect-domain E3 ubiquitin ligase [2]. Utsugi and colleagues found that mutations in the Hect-domain of *TOM1*, which is necessary for thioester-bond formation with ubiquitin, or deletion of the entire gene rendered *S.cerevisiae* unable to grow at high temperatures [2]. The authors found that *S.cerevisiae* with mutated *TOM1* exhibited an abnormally large nucleus containing duplicated DNA, fragmented nucleoli, accumulation of poly(A)+RNA in the nucleus, and cell cycle arrest at G2/M, suggesting that the disruption of *TOM1* had a severe impact on nuclear transport and cell division. Later it was shown that TOM1 was targeting DIA2 and CDC6 for proteasomal degradation during G1 and G2/M phases [3], confirming that TOM1 is involved in the regulation of the cell cycle.

Furthermore, TOM1 has been found to play an integral role in the regulation of proteostasis by ubiquitinating highly basic histone proteins [4] and various unassembled ribosomal proteins of the large and small subunit (RPL and RPS, respectively) [5]. This demonstrates that the deletion of *TOM1* (termed Δ*TOM1*) has a highly pleiotropic effect on the phenotype of *S.cerevisiae* under heat stress. To date, several extragenic suppressors of Δ*TOM1* have been identified that restore the ability of *S.cerevisiae* to grow at high temperatures, such as deletion of *DIA2* [3], overexpression of *STM1* [6], as well as mutations in genes that are linked to down-regulation of the cAMP/PKA pathway [7].The aim of this study was to identify potential new extragenic suppressors of Δ*TOM1* in six strains of *S.cerevisiae* that were derived from YDK1364. While YDK1364 is unable to grow under heat stress due to the deletion of *TOM1*, the tested strains partially regained the ability to grow at high temperatures due to the accumulation of random mutations. The six strains were sequenced and potential mutations associated with the revertant phenotype were identified using a bioinformatic pipeline. We found one known suppressor of Δ*TOM1*, as well as some new candidates that were mostly associated with the *RPL* and *RPS* gene family.

Methods

Sequencing and quality control of raw data Yeast samples were sequenced by Illumina Sequencing Technology (Illumina, SanDiego, CA, USA). As high-throughput sequencing is not infallible and sequencing errors cannot be avoided, we used FastQC to perform various quality checks of raw sequence data to ensure good quality and avoid biased data. Quality checks of FastQC, include sequence quality, GC content, failed base call content (N content), and k-mer content.

Quality control of reads After the quality control check, *trimmomatic* [8] was used with the parameters of MINLEN : 130 to trim down the bad quality ends of the reads, keeping at least 130bp of the trimmed read, and the parameter SLIDINGWINDOW:4:15, thus removing the reads that have an average base pair quality score lower than 15. To check the quality of the data after the trimming process, FastQC was used again on the filtered fasta files.

Sequence alignment against reference The fasta files containing the filtered reads were mapped to the reference genome Assembly R64-1-1.92. In a first step, the reference genome was indexed by the *bwa* (Burrows-Wheeler Aligner) index tool [9]. Subsequently, the *bwa mem* tool was used to map the sequenced reads to the reference genome. The output files of the *bwa* tool were then converted into SAM files [10].

Variant calling and annotation First, *SAMtools view* [10] was used to convert the SAM files into BAM files, which is a compressed binary format. Next, the BAM files were sorted using *SAMtools sort*. This has been done using *SAMtools mpileup*, that took our reference file, and the aligned BAM files as an input. Then the output file, a binary format of the *Variant Calling Format* files (.bcf) was directly piped into *BCFtools* to convert them into compressed vcf.gz files. This method was preferred considering the fact that our genome is a haploid yeast genome, and it does not need a complicated algorithm as used by the GATK4 pipeline. The *Tabix* [11] tool was used on the vcf files, in order to index them properly. Annotations of the variants was done by SnpEff tool on the merged vcf files. The results were then visualized by either reading the vcf files in a spreadsheet or in IGV (Integrative Genomics Viewer).

Results

We screened the *S.cerevisae* samples T5, T6, T7 and T8 for interesting mutations that could potentially be responsible for reverting the heat-sensitive phenotype of Δ*TOM1 S.cerevisae* strains. The sample T5 contains only one haploid strain named YDK1364 and was used as a reference (Figure 1). All strains found in T6 to T8 arose from the strain YDK1364 found in T5, but as mutations emerged in strains S1364-1 to S1364-8, these strains became more resistant to high-temperature stress (i.e. 37°C). Therefore, we excluded all mutations that co-occurred in T5 and any of the samples of interest T6 to T8. Since each of the samples is composed of two pooled haploid *S.cerevisae* strains (Figure 1), we excluded all homozygous mutations from further analysis, since it is highly unlikely that two strains have the same mutation suppressing the ΔTOM1 heat sensitivity phenotype.
In this first screening approach, we included only mutations with a Phred-scaled quality score of over 200. Furthermore, we focussed on mutations that interact genetically or physically with TOM1 or genes that encode for ribosomal proteins of the large (RPL) or small (RPS) subunit, because they have observed to accumulate and form detergent-insoluble protein complexes under heat stress. Using our pipeline described in the Methods section, we identified nine mutations that may explain why samples T6, T7, and T8 are less susceptible to high temperatures compared to our reference T5.

Discussion

In sample T6, we found two potential high-quality mutation candidates, namely an inframe insertion in *YGR160W* and frameshift mutation in *PMT1*. On closer inspection, we found that *YGR160W* was flagged as a dubious gene, which is unlikely to encode a functional protein, thus in spite of the quality we discarded it from further research. In contrast, *PMT1* codes for an O-mannosyltransferase that is involved in ER quality control among other things [12, 13]. As a result of the enormous global genetic interaction network that has been created by M. Constanzo and colleagues, *PMT1* has been shown to genetically interact with *HAS1*, which codes for an

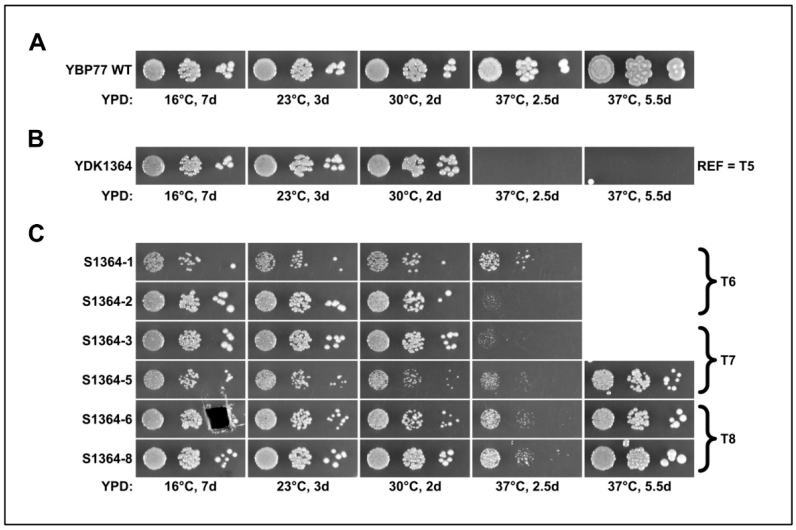

Figure 1: Growth characteristics of *S.cerevisae* strains cultured in Yeast Extract–Peptone–Dextrose medium (YPD) at different temperatures. Strain names are indicated on the left of the growth assay photographs. *S.cerevisae* cells were plated onto agar plates and were cultured for up to 7 days at increasing temperature starting from 16°C to 37°C. (A) Wildtype yeast grows well at all tested temperatures. (B) Depicted is the growth behaviour of the ΔTOM1 strain YDK1364, which is highly sensitive to heat. (C) Growth of mutated ΔTOM1 yeast strains derived from YDK1364 that exhibit lower heat susceptibility due to accumulation of mutations that counteract the ΔTOM1 phenotype. Depicted are representative growth assays of the different yeast strains. Adapted from course BC.7107, UNIFR, Benjamin Pillet.

ATP-dependent RNA helicase that is involved in the biogenesis of the 40S and 60S ribosome subunits [14, 15].

In sample T7, we found an already described Δ*TOM1* suppressor gene named *KRE6* as well as some putative candidates that might be worth investigating in more depth. We identified a nonsense mutation in the known extragenic suppressor *KRE6* of Δ*TOM1*. The premature stop codon is inserted at nucleotide position 1431 of 2161, which strongly suggests that translation of the *KRE6* transcript results in the formation of a truncated protein. *KRE6* codes for a type II membrane protein involved in the synthesis of β-(1,6)-glucan, which is an essential constituent of the fungal cell wall [16, 17]. In spite of the fact that the function of the KRE6 protein is not directly involved in stress responses, Sasaki and colleagues found that mutation in the *KRE6* gene acts as a weak suppressor of heat sensitivity mediated by *TOM1* deletion [7].

Although the authors were not able to decipher the underlying mechanism responsible for restoring heat tolerance in Δ*TOM1 S.cerevisae* with mutated *KRE6*, they concluded that the mutation may activate unknown suppressor genes of Δ*TOM1*. According to the Biological General Repository for Interaction Datasets BioGRID, *KRE6* has been shown to exhibit genetic interaction with *TOM1* itself as well as multiple genes coding for mitochondrial and cytoplasmic ribosomal proteins of the large subunit (Table 1). Moreover, we found a missense mutation in gene *KRE9*, which is also involved in the synthesis of β-(1,6)-glucan like KRE6 protein. The deletion of *KRE9* has long been known to have a deleterious effect on the growth of *S.cerevisae* by altering the composition of its cell wall [18]. Similar to *KRE6*, *KRE9* also interacts with several genes coding for the small and large subunits of ribosomes (Table 1). It is also conceivable that *KRE9* mutation alone or in combination with the mutation in *KRE6* has an impact on the transcription of ribosomal genes and thus may passively counteract the stress response associated with the accumulation of protein related to Δ*TOM1*. Another gene related to ribosome biogenesis was also found to be mutated in sample T7, namely *ISC1*, which has been reported to interact genetically with *RPL40B* [19].

Sample	Gene Name	Mutation Type	Chromosome	Interactors
T6	YGR160W	Inframe insertion	VII	Unknown
T6	PMT1	Frameshift	IV	HAS1
T7	KRE6	Stop gained	XVI	TOM1, RPL1B, RPL34B
T7	KRE9	Missense	X	MRPL17, MRPL25, MRPL38, RPL10, RPL11B, RPL15A, RPL1B, RPL24A, RPL2A, RPL3
T7	ISC1	Missense	V	RPL40B
T7	FLO9	Inframe insertion	I	IMG2, YAR1
T7	VTC4	Missense	X	TOM1 (Physical)
T8	KRE6	Missense	XVI	TOM1, RPL1B, RPL34B
T8	ROT1	Missense	XIII	RPL4B, RPS25A

Table 1: Table depicting all heterozygous mutations found in the sample T6 to T8, which might be potential suppressors of the ΔTOM1 phenotype. Annotation of the mutation type was done by SnpEff. With the exception of VTC4, only genetic interactions are shown. All intergenic and synonymous mutations were excluded from further analysis.

While the protein ISC1 is not directly linked to the regulation of ribosome biosynthesis, RPL40B is involved in the maturation of the 60S ribosomal subunit [20]. Interestingly, the null mutation of *ISC1* has been associated with heat sensitivity, which implies that the mutation observed in sample T7 does not yield a non-functional protein. Therefore, we concluded that the mutation in *ISC1* is most likely not responsible for counteracting the heat-sensitive phenotype resulting from the deletion of *TOM1*. Last but not least, we also identified a mutation in a direct physical interactor of the TOM1 protein in sample T7 named VTC4, which is a component of the vacuolar transporter chaperone (VTC) complex [21]. Unfortunately, the nature of the interaction between VTC4 and TOM1 is not described and thus it is not possible to evaluate whether a mutation in this gene has an impact on the Δ*TOM1* phenotype.

In T8, we observed a missense mutation in *KRE6*. In addition, we discovered another missense mutation in a gene named *ROT1*, which codes for a chaperone involved in protein folding [22]. Similar to the other gene candidates, *ROT1* also interacts genetically with ribosomal genes, namely *RPL4B* and *RPS25A*.

Interestingly, most of the mutations we found in samples T6, T7, and T8 were indirectly linked to genes coding for ribosomal proteins. These findings are of particular interest since deletion of *TOM1* has been associated with an accumulation of ribosomal proteins. Under normal conditions, the E3 ubiquitin ligase TOM1 rapidly removes excess of ribosomal proteins via proteasomal degradation [5]. In conclusion, we found one known as well as seven potential new suppressors of Δ*TOM1*. The known suppressor *KRE6*, has been found to be mutated in T7 and T8. Due to the heterozygosity of the mutation, we conclude that *KRE6* may explain the partially restored capability of *S.cerevisae* to grow at 37°C in one of the two strains found in each sample. However, it is important to note that while the mutation in T7 introduces an additional stop codon, the spontaneous mutation in T8 only changed one amino acid and consequently affects the function of the protein to a lesser extent (Table 1) [7]. Since most of the mutated genes found in our samples were associated with biogenesis or regulation of ribosomes, it would be interesting to investigate whether these mutations suppress the accumulation of ribosomal proteins in the absence of *TOM1*. Taken together, our data provide the basis for further investigations aimed at clarifying whether accumulation of ribosomal proteins may be causative of the heat sensitivity of yeast lacking *TOM1*.

Identification of *gai* phenotype revertant mutations in *Arabidopsis thaliana*

Thibault Schowing, Lionel Rohner, Alain Rohrbasser, Rares Cristea

Introduction

Gibberellic Acid (GA) is a growth factor that influences essentially the stem elongation and other plant developmental processes [23]. The study of plants deficient in the biosynthesis of this hormone has been essential to agriculture. Selective breeding of these mutant plants was one of the key factors leading to the green revolution since 1960, thus increasing the yield of the crops. Hence, the study of plant's response to GA is still essential for many agricultural applications [24]. Gibberellic-Acid Insensitive (*GAI*) is a gene in *Arabidopsis thaliana* (*A. thaliana*) in chromosome 1 which is involved in the regulation of plant growth. More precisely, it modulates plant growth by decreasing the responsiveness to GA [25]. (*GAI*) encodes for a protein containing a DELLA domain, which interacts with a receptor bound GA, and a functional GRAS-domain. When GA binds to its receptor GID1, they form a complex with DELLA and recruit an utiquitin ligase named SCFSLY1. DELLA is then ubiquinated and degraded in the proteasome (Figure 2) [26]. The CS63 strain contains a 51bp deletion in the DELLA-domain of *GAI*. The deletion acts as a gain-of-function mutation in DELLA, thus resulting in a dwarf phenotype due to its reduced sensitivity to GA-signaling [25, 27].

SPINDLY (*SPY*) is a gene coding for a peptide sequence that is thought to interact and activate DELLA, and thus negatively regulates GA signaling pathway [28]. Three independent recessive mutations at the *SPY* locus of *A. thaliana* gives resistance to the paclobutrazol inhibitor molecule in the GA biosynthesis pathways. This confers to the spy mutants an other phenotype also observed when wild-type *A. thaliana* plants are constantly treated with GA. The *spy-1* allele[1] is partially epistatic to the loss-of-function mutation of *gai*, that causes GA defect. Furthermore, the *spy-1* mutation can together suppress the repercussion of the loss-of-function mutation of *gai* and paclobutrazol therapy, which inhibit diverse steps in the GA biosynthesis pathway [27, 29].

The aim of this project was to screen two mutant strains of *A. thaliana* that were generated by random γ-ray mutagenesis of CS63 for mutations in gene *GAI* and *SPY*. As opposed to C63, Gar12 and Gar13 exhibit normal growth and we hypothesized that random mutation in *GAI* and/or *SPY* may be causative for their phenotypic reversion.

Methods

Trimming, Read Group Informations, Aligning and MarkDuplicates After the quality control check with *FastQC* [30], *Trimmomatic* [8] was used to filter out bad quality reads. Only the ones with a minimal length of 140 and an average quality of 15 were kept. The *bwa* (Burrows-Wheeler Aligner) tool was then used to index the reference genome and to align all the reads against this reference while adding the Read Group Information at the same time. The resulting SAM files were then sorted and converted into BAM files through SortSam (Picard) and the duplicate reads were marked using MarkDuplicates (Picard). This step, takes a sorted BAM file and add information about reads that might come from the same DNA fragment, in order to avoid counting the information given by one fragment more than one time.

[1] Mutant alleles are written in lower-case.

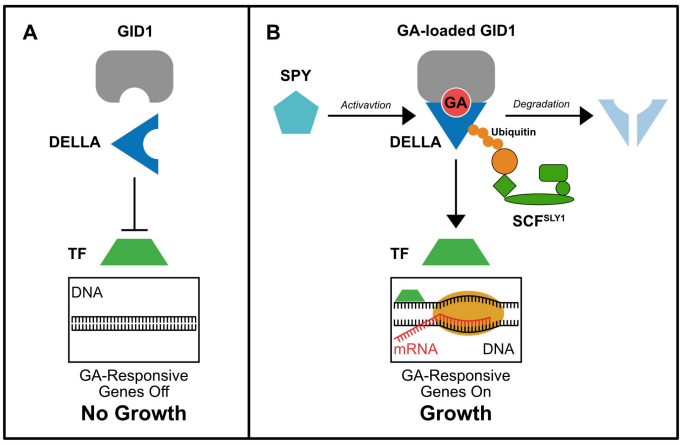

Figure 2: (A) In the abscence of GA, DELLA is not able to interact with GID1. Free DELLA proteins sequester transcription factors (TF) that activate GA-responsive genes, thus preventing growth. (B) In the presence of GA, GID1 and GA form a ligand-receptor complex that favors binding of DELLA. Once bound to GA-loaded GID1, DELLA is ubiquitinated by SCF^{SLY1} and eventually degraded, thus promoting growth. Recent findings suggest that SPY can counteract the activation of GA-responsive genes by activating DELLA by fucosylation. Adapted from [26].

Haplotype Caller and Base Quality Score Recalibration Since the genome of the *A. thaliana* is diploid, the Haplotype Caller (GATK4)was used to have a first variant call. The VCF file was then split in two, one containing only the INDELs information, and one containing the SNPs, with SelectVariants (GATK4) because the filtration methods differ. Then the new VCF files were filtered to get rid of all the false variants, through VariantFiltration (GATK4), using different parameters for INDELs and for SNPs. Then the BQSR, a data pre-processing step that detects systematic errors made by the sequencer when it estimates the quality score of each base call, was performed in two iteration to recalibrate the original bam files. A final variant call was made to generate a single vcf file containing both SNPs and INDELs.

Annotation For the annotation step, SnpEff was used in order to annotate the VCF file, to get rid of all the synonymous and intergenic variants and also to keep only the variants that are found in less than 2 strains. The analysis was further done by using GeneSearch (PhenoSystems), and since the whole BAM files were too big to handle, a shortened version of the BAM files, containing only the information about the *GAI* gene on the first chromosome, and the *SPY* gene on the third chromosome was used.

Results

We found that Gar13 had the original mutation in the DELLA-domain of the *GAI* gene, whereas Gar12 had a deletion at the same location, but affected 52 nucleotides instead of 51 like in the original deletion found in C63 (Table 2). Furthermore, we found one additional frameshift deletion in the functional GRAS-domain of strain Gar13. In *SPY* only one single missense mutation that was present in both strains was found.

Discussion Due to the additional nucleotide affected by the large deletion in Gar12, the deletion leads to a frameshift that affects all codons upstream of the DELLA-domain and thus substantially affects the structural integrity of DELLA. This means that although the *GAI* gene exhibits a gain-of-function mutation, which renders the protein resistant to proteasomal degradation, the deletion changes a substantial number of amino acids in the functional GRAS-domain. Therefore

Strains	Gene	Chromosome	Position	Mutation type
Gar12	*GAI*	I	5'149'424	3bp inframe deletion (ATC)
Gar12	*GAI*	I	5'149'496	52bp frameshift Deletion
Gar12	*SPY*	III	3'637'438	Missense mutation: T>C
Gar13	*GAI*	I	5'149'424	3bp inframe Deletion (ATC)
Gar13	*GAI*	I	5'149'495	51bp inframe Deletion (DELLA)
Gar13	*GAI*	I	5'149'624	1bp frameshift deletion (G)
Gar13	*SPY*	III	3'637'438	Missense mutation: T>C

Table 2: Table showing all non-synonymous mutations found in *A. thaliana* strains Gar12 and Gar13.

the repressive effect of DELLA is subverted and the plant is able to thrive (Figure 3).

In the Gar13 strain, the gain-of-function deletion of 51bp of the DELLA region is present, meaning that the phenotypic reversion is not given by the presence of the DELLA domain, but rather by another mutation in the functional domain. We assume that the only mutation that could be able to abrogate the gain-of-function mutation in DELLA is a single nucleotide, which induces a frameshift in the reading frame and thus probably affects the functional GRAS-domain (Table 2 and Figure 3). Similar to Gar12, this strain is able to grow normally, due to inactivation of the DELLA protein (Figure 3).

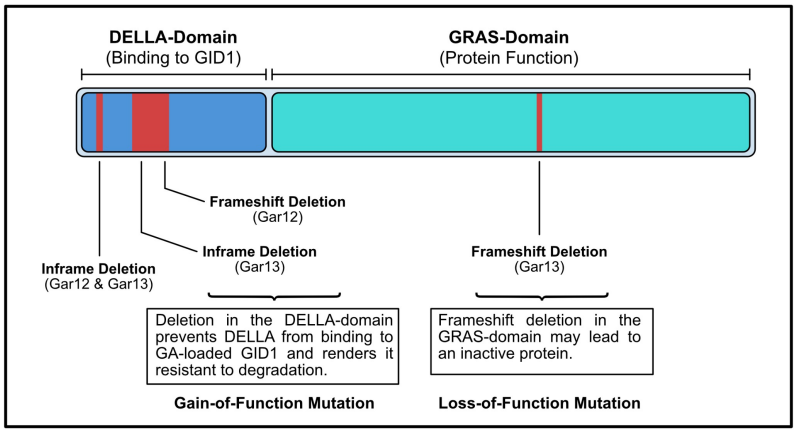

Figure 3: Illustrative summary of the project. The DELLA protein is shown with its main domains. Regions highlighted in red represent deletions that were found in Gar12 and/or Gar13.

The other 3bp deletion, which was discovered downstream of the primary deletion, is probably redundant as the 51bp deletion prevents DELLA from binding to GA-loaded GID1 regardless of other mutations. Furthermore, this variant is found in both Gar12 and Gar13, which suggests that the mutation did not arise from random mutagenesis, but was probably already present in C63. Since loss-of-function mutations in *SPY* are known to suppress gain-of-function mutations in *GAI*, we also searched for potential non-synonymous mutations in *SPY*. However, in both strains, we only found one missense variant in the *SPY* gene, changing a valine to alanine in the amino acid chain. This mutation is unlikely to have an impact on the function of SPY since the amino acid substitution (V to A) is a conservative replacement with a low impact on biochemical properties. Moreover, similar to the 3bp inframe deletion in *GAI*, this mutation was also found in both Gar12 and Gar13 and was therefore presumably present in the original plant C63. Taken together, Gar12 and Gar13 do not exhibit the dwarf phenotype due to mutations in the functional GRAS-domain of GAI rather than loss-of-function mutations in *SPY* (Figure 3).

Lactobacillus heleveticus Genome De Novo Assembly

Thibault Schowing, Lionel Rohner, Alain Rohrbasser, Rares Cristea

Introduction

The diverse bacteria involved in cheese production are essential for the texture and taste development but also, during the ripening process, the microbial changes help to kill pathogens and reduce spoilage microorganisms. *Lactobacillus helveticus* (*L. helveticus*)is a thermophilic lactic acid bacterium (LAB) used in the dairy industry as a starter as well as an adjunct culture for cheese manufacture [31]. During cheese ripening LAB undergo autolysis, which is a lytic event of the bacterial cell caused by its own intracellular enzymes, named autolysins or peptidoglycan hydrolases (PGHs) [32]. Autolysis of the bacteria leads to the release of these enzymes, which have the ability to digest the cell wall peptidoglycan of surrounding Gram+ bacteria and thus reduce microbial spoilage of the cheese [33]. The genomic plasticity of *L. helveticus* leads to a high variation in PGHs activity from one strain to another. In a previous study [33], nine genes coding PGHs were annotated and the activity of a PGH with an estimated size of 30kDa was tested by zymography in nine strains of *L. helveticus* of which six were sequenced (Figure 4). Two phenotypes were shown: phenotype A exhibits PGH activity (strains **FAM8102c1c1**, **FAM23285** and **FAM19191**) and phenotype B does not (strains **FAM22016**, **FAM1450** and **FAM1213**).

The aim of this project was to perform a de novo assembly and annotation of the genomes of the above mentioned *L. helveticus* strains, in order to detect potential genomic differences between phenotype A and phenotype B.

Methods

Sequencing and genome assembly The six *L. helveticus* strains **FAM8102c1c1**, **FAM23285**, **FAM19191**, **FAM22076**, **FAM1450**, **FAM1213** were sequenced by Illumina sequencing (Illumina, SanDiego, CA, USA). The following tasks were performed using the cluster provided by the University of Bern. *FastQC* [30] was used to check the quality of the reads and *Trimmomatic* [8] to filter out bad quality reads. Only the ones with a minimal length of 100 and an average quality of 8 were kept. *SOAPdenovo*[2] as well as *SPAdes* [34] were used to perform the genome assembly with the reads of each strains. For *SOAPdenovo* the k-mer sizes were set to 95, 85, 75 and 65. For *SPAdes* k-mere sizes were set to 21, 33, 55, 77 and 99 (default values). The four assemblies of SOAPdenovo and the assembly of *SPAdes* were compared using Abyss [35] with a maximum number of contigs set to 1000. The best genome assemblies with the bigger N50 and an approximate genome size of 20Mbp (Genome size of *L. helveticus*) were chosen[3].

Genome annotation and pan-genome analysis We used the *PROKKA* pipeline [36] to annotate the genome of the six best assemblies and the reference genome of *L. helveticus* NC_010080. *PROKKA* is an automated pipeline that annotates prokaryotic genomes. It locates open reading frames and RNA regions on contigs and translates it to protein sequences, searching for protein homologues in public databases. The resulting standards .gff files containing the annotated genome for each strain were then used by *Roary* [37] to generate a pan-genome of the six strains. The result was then visualized with *Phandango* [38] allowing visualisation of phylogenetic tree, associated metadata, and genomic information.

[2]Cited as requested by the authors but the content is not accessible.
[3]Due to the temporary unavailability of the cluster, this operation has been performed by L. Falquet and the results were provided to the students afterwards.

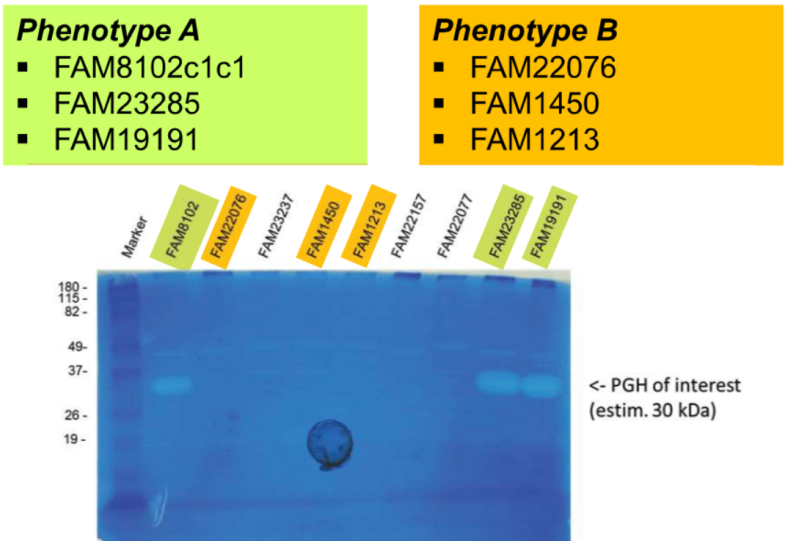

Figure 4: Result of a zymogram of the different *L. helveticus* strains of interest. The zymogram reveals that strains with phenotype A (green) express an active PGH with a size of about 30 kDa. In contrast, strains with phenotype B (yellow) appear to have lost the activity of the enzyme or possibly do not express it at all.

Extraction of the genes for each phenotypes Grep was used on the files generated by *Roary* to extract the nine PHG's [33] labelled "Lhv_" with *PROKKA* (Appendix 1). The set of genes found in strains expressing phenotype A was compared to the set of gene showing phenotype B. In Table 3 we have the two PGHs present only in the three strains expressing the PGHs activity.

Results

Gene	**Annotation**	**Avg group size [nt]**	**FAM19191_ 1K**	**FAM23285_ 1K**	**FAM8102_ 1K**
group_2348	Lhv_2053 Lysin (L.crispatus) pseudo-gene in L.helveticus	1121 bp 41 kDa	FAM19191_ 1K_00069	FAM23285_ 1K_00060	FAM8102_ 1K_00069
group_2372	Lhv_2053 Lysin (L.crispatus) pseudo-gene in L.helveticus	893 bp 33 kDa	FAM19191_ 1K_00397	FAM23285_ 1K_00499	FAM8102_ 1K_00565

Table 3: Genes present only in the three strains with a PGH activity. To convert nt to kDa the Bioline DNA to Protein Converter was used.

According to the zymography (Figure 4), the PGH involved is approximately 30kDa thus matches with group 2372. Looking at the alignment of the amino acid sequences in the output files, we saw that the sequences are identical, suggesting that it is conserved among the three strains.

Discussion We can see that PGHs are present in all strains (Appendix 1), therefore the phenotype is not due to an absence of PGH (Figure 4). With *Roary* we identified a *PGH* pseudogene that was present in the three strains exhibiting phenotype A with a matching molecular weight. The fact that it is labelled as *pseudogene* can mean that it might have lost part of its functionality and further investigation would be needed to assess the activity of the protein.

Using BLASTp [39] with default parameters, the protein was searched to be a particular lysin (WP_101853908.1) encoded by the pneumococcal bacteriophage Cp-1 [40]. To look further into this sequence, we used PHASTER [41], the PHAge Search Tool - Enhanced Release, which helps identifying and annotating prophage sequences within bacterial genomes and plasmids. The research was made only for **FAM19191**, as the sequence is identical in the three strains. The result indicating a highly conserved muramidase sequence (Figure 5A) confirms that the PGH oroginates from a phage. The locus is also shown (Figure 5B).

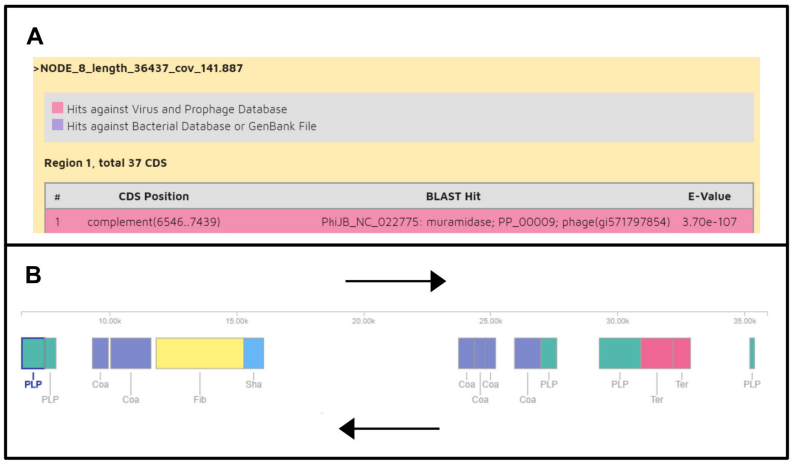

Figure 5: (A)Partial result with the assembly of **FAM19191** run in PHASTER. (B)Node 8 of **FAM19191** assembly showing annotated locus. The highlited locus represents the phage-derived muramidase.

Many gene sequences of other PGHs, pseudogenes and hypothetical proteins were found in some or all of the six different strains (Appendix 1). It would be interesting to pursue further analysis of the transcription and expression of these sequences, to further unravel the differences between the strains. The increased production of PGHs can be of major importance, especially in cheese ripening. The PGHs present in all strains and the ones present only in a small subset are most certainly located on the core-genome and the pan-genome, respectively. However, some of them might be absent only in one or two of all the strains and might be labelled as part of the "soft-core" or the "shell" genome [42, 43] as we can see in the summary statistics produced by *Roary*.

References

[1] D. Botstein, S. A. Chervitz, and J. M. Cherry, "Yeast as a Model Organism," p. 4, 2011.

[2] T. Utsugi, A. Hirata, Y. Sekiguchi, T. Sasaki, A. Toh-e, and Y. Kikuchi, "Yeast tom1 mutant exhibits pleiotropic defects in nuclear division, maintenance of nuclear structure and nucleocytoplasmic transport at high temperatures," *Gene*, vol. 234, pp. 285–295, July 1999.

[3] D.-H. Kim, W. Zhang, and D. M. Koepp, "The Hect Domain E3 Ligase Tom1 and the F-box Protein Dia2 Control Cdc6 Degradation in G $_1$ Phase," *Journal of Biological Chemistry*, vol. 287, pp. 44212–44220, Dec. 2012.

[4] R. K. Singh, M.-H. M. Kabbaj, J. Paik, and A. Gunjan, "Histone levels are regulated by phosphorylation and ubiquitylation-dependent proteolysis," *Nature Cell Biology*, vol. 11, pp. 925–933, Aug. 2009.

[5] M.-K. Sung, T. R. Porras-Yakushi, J. M. Reitsma, F. M. Huber, M. J. Sweredoski, A. Hoelz, S. Hess, and R. J. Deshaies, "A conserved quality-control pathway that mediates degradation of unassembled ribosomal proteins," *eLife*, vol. 5, p. e19105, Aug. 2016.

[6] T. Utsugi, A. Toh-e, and Y. Kikuchi, "A high dose of the STM1 gene suppresses the temperature sensitivity of the tom1 and htrl mutants in Saccharomyces cerevisiae," *Biochimica et Biophysica Acta*, p. 4, 1995.

[7] T. Sasaki, A. Toh-e, and Y. Kikuchi, "Extragenic suppressors that rescue defects in the heat stress response of the budding yeast mutant tom1," *MGG - Molecular and General Genetics*, vol. 262, pp. 940–948, Jan. 2000.

[8] A. M. Bolger, M. Lohse, and B. Usadel, "Trimmomatic: a flexible trimmer for Illumina sequence data," *Bioinformatics*, vol. 30, pp. 2114–2120, Aug. 2014.

[9] H. Li and R. Durbin, "Fast and accurate long-read alignment with Burrows–Wheeler transform," *Bioinformatics*, vol. 26, pp. 589–595, Mar. 2010.

[10] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, and 1000 Genome Project Data Processing Subgroup, "The Sequence Alignment/Map format and SAMtools," *Bioinformatics*, vol. 25, pp. 2078–2079, Aug. 2009.

[11] H. Li, "Tabix: fast retrieval of sequence features from generic TAB-delimited files," *Bioinformatics*, vol. 27, pp. 718–719, Mar. 2011.

[12] S. Strahl, "PMTI, the gene for a key enzyme of protein 0-glycosylation in Saccharomyces cerevisiae," p. 5, 1993.

[13] V. Goder and A. Melero, "Protein O-mannosyltransferases participate in ER protein quality control," *Journal of Cell Science*, vol. 124, pp. 144–153, Jan. 2011.

[14] M. Costanzo, B. VanderSluis, E. N. Koch, A. Baryshnikova, C. Pons, G. Tan, W. Wang, M. Usaj, J. Hanchard, S. D. Lee, V. Pelechano, E. B. Styles, M. Billmann, J. van Leeuwen, N. van Dyk, Z.-Y. Lin, E. Kuzmin, J. Nelson, J. S. Piotrowski, T. Srikumar, S. Bahr, Y. Chen, R. Deshpande, C. F. Kurat, S. C. Li, Z. Li, M. M. Usaj, H. Okada, N. Pascoe, B.-J. San Luis, S. Sharifpoor, E. Shuteriqi, S. W. Simpkins, J. Snider, H. G. Suresh, Y. Tan, H. Zhu, N. Malod-Dognin, V. Janjic, N. Przulj, O. G. Troyanskaya, I. Stagljar, T. Xia, Y. Ohya, A.-C. Gingras, B. Raught, M. Boutros, L. M. Steinmetz, C. L. Moore, A. P. Rosebrock, A. A. Caudy, C. L. Myers, B. Andrews, and C. Boone, "A global genetic interaction network maps a wiring diagram of cellular function," *Science*, vol. 353, pp. aaf1420–aaf1420, Sept. 2016.

[15] J. A. Dembowski, B. Kuo, and J. L. Woolford, "Has1 regulates consecutive maturation and processing steps for assembly of 60s ribosomal subunits," *Nucleic Acids Research*, vol. 41, pp. 7889–7904, Sept. 2013.

[16] T. Kurita, Y. Noda, T. Takagi, M. Osumi, and K. Yoda, "Kre6 Protein Essential for Yeast Cell Wall β-1,6-Glucan Synthesis Accumulates at Sites of Polarized Growth," *Journal of Biological Chemistry*, vol. 286, pp. 7429–7438, Mar. 2011.

[17] T. Roemer, S. Delaney, and H. Bussey, "SKN1 and KRE6 define a pair of functional homologs encoding putative membrane proteins involved in beta-glucan synthesis.," *Molecular and Cellular Biology*, vol. 13, no. 7, pp. 4039–4048, 1993.

[18] J. L. Brown and H. Bussey, "The Yeast KRE9 Gene Encodes an 0 Glycoprotein Involved in Cell Surface ,-Glucan Assembly," *MOL. CELL. BIOL.*, vol. 13, p. 11, 1993.

[19] S. Hoppins, S. R. Collins, A. Cassidy-Stone, E. Hummel, R. M. DeVay, L. L. Lackner, B. Westermann, M. Schuldiner, J. S. Weissman, and J. Nunnari, "A mitochondrial-focused genetic interaction map reveals a scaffold-like complex required for inner membrane organization in mitochondria," *The Journal of Cell Biology*, vol. 195, pp. 323–340, Oct. 2011.

[20] A. Fernández-Pevida, O. Rodríguez-Galán, A. Díaz-Quintana, D. Kressler, and J. de la Cruz, "Yeast Ribosomal Protein L40 Assembles Late into Precursor 60 S Ribosomes and Is Required for Their Cytoplasmic Maturation," *Journal of Biological Chemistry*, vol. 287, pp. 38390–38407, Nov. 2012.

[21] O. Muller, "Role of the Vtc proteins in V-ATPase stability and membrane trafficking," *Journal of Cell Science*, vol. 116, pp. 1107–1115, Mar. 2003.

[22] M. Takeuchi, Y. Kimata, and K. Kohno, "Saccharomyces cerevisiae Rot1 Is an Essential Molecular Chaperone in the Endoplasmic Reticulum," *Molecular Biology of the Cell*, vol. 19, pp. 3514–3525, Aug. 2008.

[23] R. Hooley, "Gibberellins: perception, transduction and responses," p. 27.

[24] R. L. Ordonio and M. Matsuoka, "New path towards a better rice architecture," *Cell Research*, vol. 27, pp. 1189–1190, Oct. 2017.

[25] J. Peng, P. Carol, D. E. Richards, K. E. King, R. J. Cowling, G. P. Murphy, and N. P. Harberd, "The Arabidopsis GAI gene defines a signaling pathway that negatively regulates gibberellin responses," *Genes & Development*, vol. 11, pp. 3194–3205, Dec. 1997.

[26] A. L. Hauvermale, T. Ariizumi, and C. M. Steber, "Gibberellin Signaling: A Theme and Variations on DELLA Repression," *Plant Physiology*, vol. 160, pp. 83–92, Sept. 2012.

[27] S. Lee, "Gibberellin regulates Arabidopsis seed germination via RGL2, a GAI/RGA-like gene whose expression is up-regulated following imbibition," *Genes & Development*, vol. 16, pp. 646–658, Mar. 2002.

[28] R. Zentella, N. Sui, B. Barnhill, W.-P. Hsieh, J. Hu, J. Shabanowitz, M. Boyce, N. E. Olszewski, P. Zhou, D. F. Hunt, and T.-p. Sun, "The Arabidopsis O-fucosyltransferase SPINDLY activates nuclear growth repressor DELLA," *Nature Chemical Biology*, vol. 13, pp. 479–485, May 2017.

[29] S. E. Jacobsen and N. E. Olszewski', "Mutations at the SPINDLY Locus of Arabidopsis Alter Gibberellin Signal Transduction," p. 10.

[30] S. Andrews, F. Krueger, A. Segonds-Pichon, L. Biggins, C. Krueger, and S. Wingett, "FastQC." Babraham Institute, Jan. 2012.

[31] J. Marcelino, "Lactic Acid Bacteria as Starter-Cultures for Cheese Processing: Past, Present and Future Developments," in *Lactic Acid Bacteria - R & D for Food, Health and Livestock Purposes* (J. M. Kongo, ed.), InTech, Jan. 2013.

[32] J. Delcour, T. Ferain, M. Deghorain, E. Palumbo, and P. Hols, "The biosynthesis and functionality of the cell-wall of lactic acid bacteria," *Antonie Van Leeuwenhoek*, vol. 76, pp. 159–184, Nov. 1999.

[33] I. Jebava, M. Plockova, S. Lortal, and F. Valence, "The nine peptidoglycan hydrolases genes in Lactobacillus helveticus are ubiquitous and early transcribed," *International Journal of Food Microbiology*, vol. 148, pp. 1–7, July 2011.

[34] S. Nurk, A. Bankevich, D. Antipov, A. Gurevich, A. Korobeynikov, A. Lapidus, A. Prjibelsky, A. Pyshkin, A. Sirotkin, Y. Sirotkin, R. Stepanauskas, J. McLean, R. Lasken, S. R. Clingenpeel, T. Woyke, G. Tesler, M. A. Alekseyev, and P. A. Pevzner, "Assembling Genomes and

Mini-metagenomes from Highly Chimeric Reads," in *Research in Computational Molecular Biology* (D. Hutchison, T. Kanade, J. Kittler, J. M. Kleinberg, F. Mattern, J. C. Mitchell, M. Naor, O. Nierstrasz, C. Pandu Rangan, B. Steffen, M. Sudan, D. Terzopoulos, D. Tygar, M. Y. Vardi, G. Weikum, M. Deng, R. Jiang, F. Sun, and X. Zhang, eds.), vol. 7821, pp. 158–170, Berlin, Heidelberg: Springer Berlin Heidelberg, 2013.

[35] J. T. Simpson, K. Wong, S. D. Jackman, J. E. Schein, S. J. Jones, and I. Birol, "ABySS: A parallel assembler for short read sequence data," *Genome Research*, vol. 19, pp. 1117–1123, June 2009.

[36] T. Seemann, "Prokka: rapid prokaryotic genome annotation," *Bioinformatics (Oxford, England)*, vol. 30, pp. 2068–2069, July 2014.

[37] A. J. Page, C. A. Cummins, M. Hunt, V. K. Wong, S. Reuter, M. T. G. Holden, M. Fookes, D. Falush, J. A. Keane, and J. Parkhill, "Roary: rapid large-scale prokaryote pan genome analysis," *Bioinformatics*, vol. 31, pp. 3691–3693, Nov. 2015.

[38] J. Hadfield, N. J. Croucher, R. J. Goater, K. Abudahab, D. M. Aanensen, and S. R. Harris, "Phandango: an interactive viewer for bacterial population genomics," *Bioinformatics*, vol. 34, pp. 292–293, Jan. 2018.

[39] S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman, "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs," *Nucleic Acids Research*, vol. 25, pp. 3389–3402, Sept. 1997.

[40] A. C. Martín, R. López, and P. García, "Pneumococcal Bacteriophage Cp-1 Encodes Its Own Protease Essential for Phage Maturation," *Journal of Virology*, vol. 72, pp. 3491–3494, Apr. 1998.

[41] D. Arndt, J. R. Grant, A. Marcu, T. Sajed, A. Pon, Y. Liang, and D. S. Wishart, "PHASTER: a better, faster version of the PHAST phage search tool," *Nucleic Acids Research*, vol. 44, pp. W16–W21, July 2016.

[42] R. S. Kaas, C. Friis, D. W. Ussery, and F. M. Aarestrup, "Estimating variation within the genes and inferring the phylogeny of 186 sequenced diverse Escherichia coli genomes," *BMC Genomics*, vol. 13, no. 1, p. 577, 2012.

[43] R. C. Inglin, L. Meile, and M. J. A. Stevens, "Clustering of Pan- and Core-genome of Lactobacillus provides Novel Evolutionary Insights for Differentiation," *BMC Genomics*, vol. 19, p. 284, Dec. 2018.

Appendix

Gene	Annotation	FAM1213 1K	FAM1450 1K	FAM19191 1K	FAM22076 1K	FAM23285 1K	FAM8102 1K
group_1103	Lhv_0549 N-acetylmuramidase	FAM1213_ 1K_01187	FAM1450_ 1K_00785	FAM19191_ 1K_01147	FAM22076_ 1K_00934	FAM23285_ 1K_01072	FAM8102_ 1K_01185
group_1218	Lhv_1433 Lysin	FAM1213_ 1K_01833	FAM1450_ 1K_00044	FAM19191_ 1K_01884	FAM22076_ 1K_01582	FAM23285_ 1K_01903	FAM8102_ 1K_01986
group_3457	Lhv_0649 Lysozyme	FAM1213_ 1K_00895	FAM1450_ 1K_00838	FAM19191_ 1K_01232	FAM22076_ 1K_00917	FAM23285_ 1K_01191	FAM8102_ 1K_01268
group_852	Lhv_1295 Enterolysin M23 family peptidase	FAM1213_ 1K_00043	FAM1450_ 1K_01113	FAM19191_ 1K_00150	FAM22076_ 1K_00164	FAM23285_ 1K_00217	FAM8102_ 1K_00225
group_862	Lhv_1059 LysM peptidoglycan binding domain-containing protein	FAM1213_ 1K_00147	FAM1450_ 1K_00238	FAM19191_ 1K_00248	FAM22076_ 1K_00274	FAM23285_ 1K_00308	FAM8102_ 1K_00381
group_993	Lhv_1433 Lysin	FAM1213_ 1K_00691	FAM1450_ 1K_01203	FAM19191_ 1K_01800	FAM22076_ 1K_00088	FAM23285_ 1K_01748	FAM8102_ 1K_01891
group_995	Lhv_0191 Amidase	FAM1213_ 1K_00700	FAM1450_ 1K_00303	FAM19191_ 1K_00506	FAM22076_ 1K_00064	FAM23285_ 1K_00566	FAM8102_ 1K_00638
group_1862	Lhv_2053 Lysin (L.crispatus) pseudogene in L.helveticus		FAM1450_ 1K_00045	FAM19191_ 1K_01885	FAM22076_ 1K_01583	FAM23285_ 1K_01904	FAM8102_ 1K_01987
group_1899	Lhv_2053 Lysin (L.crispatus) pseudogene in L.helveticus		FAM1450_ 1K_00267	FAM19191_ 1K_00615	FAM22076_ 1K_00716	FAM23285_ 1K_00607	FAM8102_ 1K_00746
group_1344	Lhv_1307 Enterolysin M23 family peptidase			FAM19191_ 1K_00162	FAM22076_ 1K_00152	FAM23285_ 1K_00229	FAM8102_ 1K_00237
group_1345	Lhv_0190 N-acetylmuramidase			FAM19191_ 1K_00507	FAM22076_ 1K_00063	FAM23285_ 1K_00565	FAM8102_ 1K_00639

Table 1: PGHs in common between all strains. Extracted from the files generated by *Roary* and labeled "Lhv_" by *PROKKA*.