Bioinformatics Practicals In Sillico

BC-7107

Thibault Schowing
Lionel Rohner
Alain Rohrbasser
Rares Cristea

November 30, 2019

Introduction

Bioinformatics is the application of computational technology to handle the rapidly growing repository of information related to molecular biology. Bioinformatics combines different fields of study, including computer sciences, molecular biology, biotechnology, statistics and engineering. It is particularly useful for managing and analysing large sets of data, such as those generated by the fields of genomics and proteomics.

In this report we focus on the bioinformatics tools for mutant analysis through three different projects; mutations in gai and spy in *Arabidopsis Thaliana*, mutations in *Saccharomyces cerevisiae* and mutations as well as Denovo assembly in *Lactobacillus Helveticus*. We want to sort out new mutations with these tools and learn how to design a bioinformatic test. It includes the quality test, the annotation of our sequenced genomes and various analysis of these results. Thus, everything upstream of the analysis must be properly done, using several software described thereafter. Our machines are too week in order to analyse the data and performed the bioinformatics steps, thus, we will use a cluster dedicated for this lecture, in Bern Switzerland.

Yeast Genome Analysis

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 (e.g. beer, which is the epony of *S.cerevisae*) and the production of biofuel. In addition to the contribution industrial fermentation, *S.cerevisiae* has become one of the most popular model organism for eukaryotic biology, due to its simple cellular architecture, cheap maintenance cost, fast growth, non-pathogenic nature (discussed in [1]) , and homologies to human cells (e.g. ribosomes), which cannot be studied in prokaryotic model organism, such as *E.coli* [2]. In particular, the genetic analysis of *S.cerevisiae* has gained popularity in the scientific community since it was the first eukaryotic organism whose genome was fully sequenced. The haploid genome of *S.cerevisiae* consists of 16 linear chromosomes containing 6604 genes encoded within approximately 12 megabase-pairs (Mbp) [3]. The fact that genome of *S.cerevisiae* is quite small and almost completely void of intronic DNA, thus making it an ideal microorganism for the identification of mutations and single nucleotide polymorphisms.

Mating of two haploid yeast of opposite mating type (i.e. Mat a or α) gives rise to diploid cells that possess 32 chromosomes. This is a single eukaryotic organism with a division cycle of 90 minutes. Through the process of budding in which smaller daughter cells pinch, or bud, off the mother cell. *S.cerevisiae* forms colonies on agar plates in the laboratory in a few days with no special incubators required (best grows at 30°C).

TODO raccourcir intro

Target of Myb protein1 (Tom1) , is a gene involved in the ribosomal biogenesis in yeast and human respectively. In human, this gene is involved in several pathways including endocytosis, endosomal transport, intracellular protein transport, neutrophil degranulation and protein transport [4, 5]. It is located on the ch.22 (component UP000005640) and ch.4 in human and yeast respectively. As it is well conserved among the eukaryote, we can study the gene with yeast for the raisons explained above. In yeast specifically, TOM1 was first described as a gene involved in temperature sensitivity and could be supressed by STM1 [6]. TOM1 is a hect-domain, which has been identified as a conserved feature of E3 ubiquitin ligases group. It regulates transcriptional activation through effectors ADA on coactivator proteins on the DNA. The action of TOM1 is to regulate through ubiquitination the temperature sensitivity [7].

A tom1-1 mutant has been isolated, and under electron microscopy and indirect immunofluorescence microscopy, it has been shown that the large nucleus contains duplicated DNA and short spindle, and structures fragmentations. This show that the disruption of the system, impacts the nuclear transport and the cell division in the G1 phase [5, 8].

TOM1 encode for a large 380KDa proteins with a hect-domain at his C terminus (homologous to E6-AO C terminus). Site-directed mutagenesis of the conserved cysteine residue (tom1C3235A) in the hect-domain, supposed to be necessary for thioester-bond formation with ubiquitin, abolished the gene function. After a test with the over production of a myc-tagged ubiquitinRA, it shows that TOM1 is a ubiquitin ligase [8]. More recently, TOM1 has been described as a fundamental macromolecular machine. The ribosomes biogenesis is much more complex in eukaryote cells as in bacteria, and it is involved in several fundamental cellular processes, including growth and cell division [9].

Ribosomes are subunits assemblages allowing the production of proteins. Subunits are made of RNA (rRNA) and specifies proteins (r-proteins) (*Saccharomyces cerevisiae*: 40S [18S rRNA, 33

RPs]; 60S [25S, 5.8S, 5S rRNA, 46 RPs]– *Escherichia coli*: 30S [16S rRNA, 21 RPs]; 50S [23S, 5S rRNA, 34 RPs]) [10]. Recent studies have shown that defect in the biogenesis linked to a wide range of hereditary diseases like Alzheimer's and anemia [11]. Ribosomes are a mixture of almost 80 different protein and stick together through a scaffold made by the RNA as explain above. Each protein is expressed in one copy and each of these proteins are needed to assemble the ribosomes. However, the number of steps needed for the biogenesis is large and not totally known. Moreover, it is impossible for a cell to produce the exact number of the needed proteins, including the same number of copies of all the proteins needed in a ribosome [12]. It will build up an approximate number needed and then degrade the leftovers, that will be ubiquinined by TOM1 and degraded in lysosomes. We can say that TOM1 act as a quality control on this mechanism during the anabolism and division phases of the cells, leading to a week and crucial homeostasis [12].

Ribosome biogenesis is an intricate process involving many chaperons and assembly factors (>200 factors) and snoRNAs (75) [10]. Two subunits are part of the final ribosomes, the 40S has one rRNA (18S) and 33 r-proteins. The 60S comprises three rRNAs (25S, 5.8S, 5S) and 47 r-proteins subunit [10].

The assembly and maturation of the ribosomes passes from the nucleus to the cytosol. ATP-dependent RNA helicases and three AAA-type ATPases (ATPases associated with various cellular activities) are mandatory nedded to make this process occure.This suggests that the energy derived by these enzymes is required for ribosomes assembly. The absence of one of these proteins might stall ribosome biogenesis and terminate cell growth even under optimal growth conditions [9, 10].

To summarize, TOM1 is necessary to the ribosome's biogenesis and the elimination of the leftover building blocks by ubiquitination. Thus, the aim of this project is to identify the suppressor of this gene by high sequencing throughput with bioinformatics tools.

TODO on en fait référence nule part à cette figure -¿ nécessaire ?

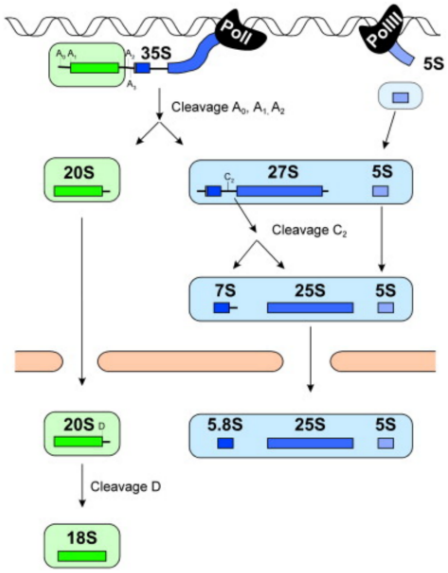

Figure 1: Simplified overview of the major steps in pre-rRNA processing.

Methods

Quality control of sequences The high throughput sequencing method is not infallible, so the data will have contaminants, badly read sequences due to the intensity of the fluorescent signal, or the quality of the reagents that have a decaying quality with the number of sequencing cycles over time [13]. These errors might add a lot of false signals, useless extrawork, and complicate the data analysis, therefore we need to get rid of them. In order to check for the quality of the data, we use the tool : *fastqc* [14] which will help us visualize our fasta file given by the sequencer.

Trimming and bad quality reads removal After the quality control check, we used the tool *trimmomatic* [15] with the given 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. The next step was checking if the quality of the data has improved after the trimming process, by using again fastqc, on the trimmomatic fastq file output.

Sequence alignment against reference Since the fasta files gives no information about the sequences position in the yeast genome, we had to align all of the reads against the fasta file of a known yeast genome, or most likely a consensus of a yeast genome (Assembly R64-1-1.92), containing the positional information.

However, in order to do that, we first had to index the reference fasta using the *bwa index* tool [16], which is a way of giving a sort of table of contents of our reference fasta file (in our case the R64-1-1.92.fa), that is used by the burrows-wheeler aligner algorithm. Subsequently, we used the *bwa mem* tool in order to align our sequenced data against the reference. Then, we have converted all the SAM [17] files containing our aligned sequences, into BAM files, a compressed binary format easier to work with.

Variant calling and annotation This has been done using only *samtools mpileup*, that took our reference file, and the aligned BAM files as an input, and gave us the binary format of the *Variant Calling Format* files (.vcf). Then we used *bcftools* to convert them into 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 *GATK* pipeline. The *tabix* [18] tool was used on the vcf files, in order to index them properly.

In order to annotate the variants given by the vcf files, we had to use SnpEff tool on the vcf files that were merged together with all their indexes, and we also kept only the variants that were found in less than all 4 strains that we had to analyse, since we had to filter through all the variants that were different from the reference. The variants interesting to us, are indeed the ones that are specific to one of the mutants, and that's why we had to filter this way.

The results were then visualized by either reading the vcf files in xcel or in IGV (Integrative Genomics Viewer).

TODO sources de l'image ?

Results

During our practical 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 ΔTOM1 *S.cerevisae* strains. The sample T5 contains only one haploid strain named YDK1364 and was used as a reference (Figure 2). 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 2), 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.

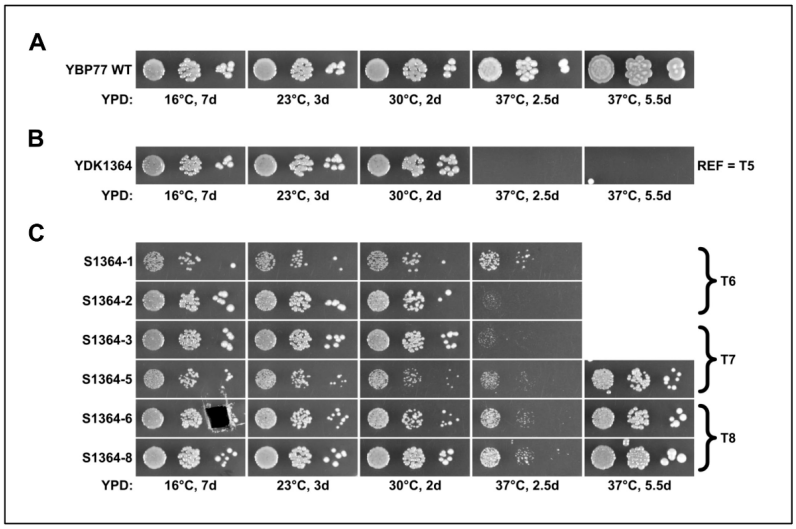

Figure 2: Growth characteristics of *S.cerevisae* strains cultured in Yeast Extract–Peptone–Dextrose medium (YPD) at different temperatures. Strain names are indicated on the left of the growth assay photographs. *S.cerevisae* cells were plated onto agar plates and were cultured for up to 7 days at increasing 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.

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. Using our pipeline described in the Variant calling and annotation of 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 [19] [20].

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 ATP-dependent RNA helicase that is involved in the biogenesis of the 40S and 60S ribosome subunits [21] [22].

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.

In sample T7, we found an already described ΔTOM1 suppressor gene named KRE6 as well as some putative candidates that might be worth investigating in more depth. In T7, we identified a nonsense mutation in the known extragenic suppressor KRE6 of ΔTom1. 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 [23, 24].

In spite of the fact that the function of the KRE6 protein is not directly involved stress responses, Sasaki and colleagues found that mutation in the KRE6 gene acts as a weak suppressor of heat sensitivity mediated by TOM1 deletion [25].

Although the authors were not able to decipher the underlying mechanism responsible for restoring heat tolerance in ΔTOM1 S.cerevisae with mutated KRE6, but they concluded that the mutation may activate unknown suppressor genes of ΔTOM1. 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 to be 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 and thus causes defects to its integrity [26].

Similar to KRE6, KRE9 also interacts with several gene coding for the small and large subunit 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 ΔTom1. 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 RBL40B [27]. 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 [28]. 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 form 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 [29]. 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 ΔTOM1 phenotype. A possible interesting scenario could be that TOM1 is an inhibitor of VTC4.

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 [30]. 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 ribosomal genes. These findings are of particular interest since deletion of TOM1 *S.cerevisae* has been associated with greatly increased levels of ribosomal proteins. Under normal conditions, the E3 ubiquitin ligase TOM1 rapidly removes excess of ribosomal proteins via proteasomal degradation [12].

In conclusion, we found one known as well as seven potential new suppressors of ΔTOM1. 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 a additional stop codon, the spontaneous mutation in T8 only affected one amino acid and consequently affects the function of the protein to a lesser extent (Table1) [25]. 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. Take 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.

Arabidopsis Thaliana Genome Analysis

TODO ALAIN: raccourcir

TODO LIONEL: CHANGER IMAGE

Introduction

The plant Arabidopsis thaliana is a genetic model worldwide used in plant biology since 1995, when it has been promoted as model for molecular genetics. The genome is entirely sequenced in 200. ATH is a diploid organism of 114,5 to 125 million base pairs within 5 chromosomes (haploid). The germination to mature seed is done about 6 weeks and easy to cultivate in restricted space and produce a lot of seed. A wide range of mutants are already available and it growth from year to another through multinational research community of academic, government and industry laboratories. The importance of ATH is crucial and invaluable resources to fight the loss of crops due to plants diseases.

GAI Gibberellic-Acid Insensitive (GAI) is a gene in Arabidopsis thaliana in chromosome 1 which is involved in the regulation of plant growth. Precisely, it mediated the input signals and module the growth by decreasing the responsiveness to gibberellin [31]. Gibberellin is a tetracyclic diterpenoid growth factor and influence essentially the stem elongation and other plant developmental processes [32]. The main mutation involved a deletion of a 17 amino acid segment. The gai allele contains a deletion of 51-bp from within the GAI ORF ,from close to the N terminus and confers a dominant dwarf phenotype. The GAI (gai1-1 and gai 1-2, two mutations on the same gene) protein as normally a length of 533 AA and is normally located in the nucleus. The deleted segment is shown in yellow for DELLA, the common one [31, 33]. If it is mutated (gai) and the plant growth better, it is a gain of function gene, in contrary it is a loss of function.

The mutation in SPY (spy) is a suppressor of gai, conferring to the plant a normal phenotype. GA-deficient Arabidopsis mutants display characteristic phenotypes, including dark green leaves and a dwarf growth habit attributable to reduced stem elongation [31]. The gai mutation affects GA reception or subsequent signal transduction and does not result in GA deficiency [32].

SPY For spy, three independent recessive mutations at the SPINDLY (SPY) locus of Arabidopsis confer resistance to the gibberellin (GA) biosynthesis inhibitor paclobutrazol. Paclobutrazol is a plant growth retardant. It is an antagonist of the plant hormone gibberellin. It works by inhibiting gibberellin biosynthesis by inhibiting endoplasmic reticulum monooxygenases. Relative to wild type, spy mutants exhibit longer hypocotyls, leaves that are a lighter green colour, increased stem elongation, early flowering, parthenocarpy, and partial male sterility. All of these phenotypes are also observed when wild-type Arabidopsis plants are repeatedly treated with gibberellin A3 (GA3). The spy-1 allele is partially epistatic to the ga1-2 mutation, which causes GA deficiency. In addition, the spy-1 mutation can simultaneously suppress the effects of the ga1-2 mutation and paclobutrazol treatment, which inhibit different steps in the GA biosynthesis pathway. This observation suggests that spy-1 activates a basal level of GA signal transduction that is independent of GA [33].

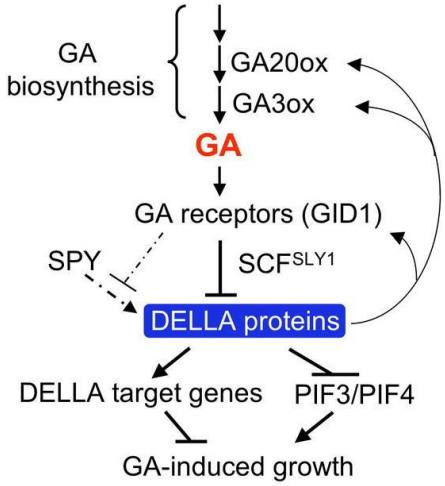

Figure 3: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3243332/figure/i1543-8120-64-1-1-f21/

Methods

Trimming, Read Group Informations, Aligning and MarkDuplicates After the quality control check with *fastqc* [14], *Trimmomatic* [15] 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.

Haplotype Caller and Base Quality Score Recalibration Since the genome of the *Arabidopsis thaliana* is diploid, in order to analyse its sequences we had to use the Haplotype Caller (GATK4), in order 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, since they bring no new information to our analysis, and also to keep only the variants that are found in less than 2 strains. Eventually, we got a nice annotated and filtered vcf file. The analysis was further done by using GeneSearch (PhenoSystems), and since the whole bam files were a bit too big to handle, we have used the shortened versions of the bam files, provided by the professors, containing only the information about the GAI gene on the first chromosome, and the SPY gene on the third chromosome.

Results

TODO Rares

Discussion

Lactobacillus Heleveticus Genome Assembly

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 helps to kill pathogens and reduce spoilage micro-organisms. *Lactobacillus helveticus* is a thermophilic lactic acid bacterium (LAB) used in the dairy industry as a starter or an adjunct culture for cheese manufacture. By releasing **peptidoglycan hydrolases**(PGHs), it has the ability to digest the bacterial cell wall (gram+) inducing death of surrounding bacteria but also its own autolysis.

The genomic plasticity of *Lactobacillus helveticus* leads to a high variation in PGHs activity from one strain to another. In a previous study [34], 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 *Lactobacillus helveticus* of which six were sequenced (see 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 work was to detect potential genomic differences involved in the two different phenotypes by sequencing, assembling and compare the genome of the six strains using a previously annotated reference genome of *Lactobacillus helveticus* (NC_010080). A potential candidate present only in the strains expressing a PGHs activity suggests that it might have been acquired by a viral insertion.

Methods

Sequencing and genome assembly The six *Lactobacillus helveticus* strains **FAM8102c1c1**, **FAM23285**, **FAM19191**, **FAM22076**, **FAM1450**, **FAM1213** were sequenced by Illumina sequencing. The following tasks were performed using the cluster provided by the University of Bern. *FastQC* [14] was used to check the quality of the reads and *Trimmomatic* [15] to filter out bad quality reads. Only the ones with a minimal length of 100 and an average quality of 8 were kept. *SOAPdenovo* as well as *Spades* 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 with a maximum number of contigs set to 1000. The best genome assemblies with the bigger N50 and a approximate genome size of 20Mbp (Genome size of *Lactobacillus helveticus*) were then chosen[1].

Genome annotation and pan-genome analysis We used the *PROKKA* pipeline [35] to annotate the genome of the six best assemblies and the reference genome for *Lactobacillus helveticus* NC_010080. *PROKKA* is an automated pipeline that annotates prokaryotic genomes. It locates open reading frames ans 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 are then used by *Roary* [36] to generate a pan-genome of the six strains. The result was then visualized with *Phandango* [37] allowing visualisation of phylogenetic tree, associated metadata and genomic information.

[1]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.

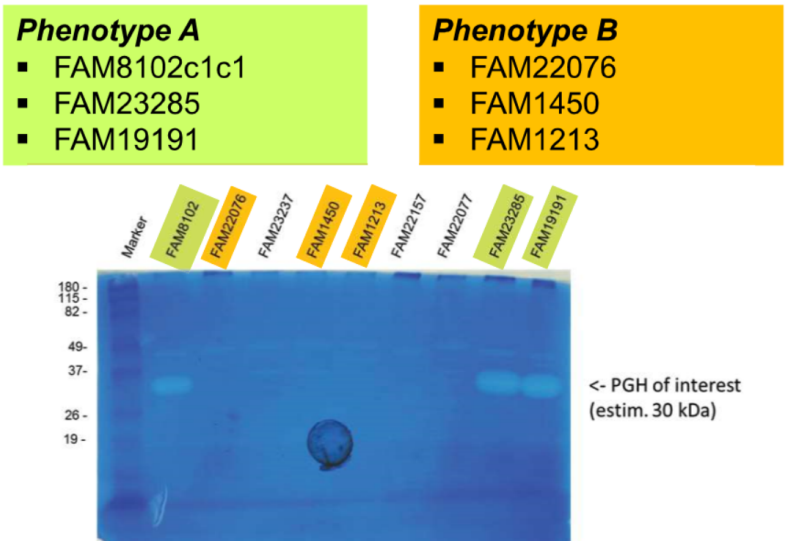

Figure 4: Phenotype A is expressing an active peptidoglycan hydrolase and phenotype B is not.

Extraction of the genes for each phenotypes Grep was applied to the files generated by *Roary* to extract the nine PHG's [34] labelled "Lhv_" with *PROKKA* (table 3). The set of genes found in strains expressing phenotype A was then compared to the set of gene showing phenotype B. In table 2 we have the two PGHs present only in the three strains expressing the PGHs activity.

Results

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

Table 2: Genes present only in the three strains with a PGH activity.

According to figure 4, the PGH involved is approximately 30kDa thus matches with group 2372. Looking at the alignment of the amino acid sequences (Figure 5) we see that the sequences are identical thus showing a great conservation between the three strains.

Discussion We can see that PGHs are present in all strains (Supplementary table 3), therefore the phenotype observed in the figure 4 is not due to an absence of PGH. With ROARY we identified

```
CLUSTAL format alignment by MAFFT L-INS-i (v7.310)


FAM19191_1K_003 MTSRQLGVDVAVYQGTSMTAYHNAGAKFGIAKLTEGTNYVNPKAHYQIKSLHANHMYVHA
FAM23285_1K_004 MTSRQLGVDVAVYQGTSMTAYHNAGAKFGIAKLTEGTNYVNPKAHYQIKSLHANHMYVHA
FAM8102_1K_0056 MTSRQLGVDVAVYQGTSMTAYHNAGAKFGIAKLTEGTNYVNPKAHYQIKSLHANHMYVHA
                ************************************************************

FAM19191_1K_003 YHFATFGYSVSRAKLEGKAFVKRAKAENISKKRFLWLDWESGSGNCVTGGKAASTKAILA
FAM23285_1K_004 YHFATFGYSVSRAKLEGKAFVKRAKAENISKKRFLWLDWESGSGNCVTGGKAASTKAILA
FAM8102_1K_0056 YHFATFGYSVSRAKLEGKAFVKRAKAENISKKRFLWLDWESGSGNCVTGGKAASTKAILA
                ************************************************************

FAM19191_1K_003 FMKVCHDAGYKVGLYSGASLLRNNIDTKQIVKKYGTCIWVASYPTDLAYTPNFNYFPSMD
FAM23285_1K_004 FMKVCHDAGYKVGLYSGASLLRNNIDTKQIVKKYGTCIWVASYPTDLAYTPNFNYFPSMD
FAM8102_1K_0056 FMKVCHDAGYKVGLYSGASLLRNNIDTKQIVKKYGTCIWVASYPTDLAYTPNFNYFPSMD
                ************************************************************

FAM19191_1K_003 GVAIWQFCDNWKGLGVDGNISLIDLHKDSAGKKVTKPAEKPKPKPEKKTGVVYAPVINRN
FAM23285_1K_004 GVAIWQFCDNWKGLGVDGNISLIDLHKDSAGKKVTKPAEKPKPKPEKKTGVVYAPVINRN
FAM8102_1K_0056 GVAIWQFCDNWKGLGVDGNISLIDLHKDSAGKKVTKPAEKPKPKPEKKTGVVYAPVINRN
                ************************************************************

FAM19191_1K_003 PNWMIQLMDGNGHYTGKYIKTNTRWKYFDVKTIKGMKCYKLGTDKQWVPAKFLKVIE
FAM23285_1K_004 PNWMIQLMDGNGHYTGKYIKTNTRWKYFDVKTIKGMKCYKLGTDKQWVPAKFLKVIE
FAM8102_1K_0056 PNWMIQLMDGNGHYTGKYIKTNTRWKYFDVKTIKGMKCYKLGTDKQWVPAKFLKVIE
                ************************************************************
```

Figure 5: Alignment of amino acid sequences of group 2372 for the three strains.

a PGH only present in the three strains exhibiting phenotype A with a matching molecular weight.

Using BLASTp [38] with default parameters, the protein was searched to be a particular lysin (WP_101853908.1) encoded by the pneumococcal bacteriophage Cp-1 [39]. To look further into this sequence, we tried PHASTER [40], the PHAge Search Tool - Enhanced Release, which helps identifying and annotate 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, shown in figure 6, confirms that the PGH comes from a phage. The locus is shown in supplementary figure 7.

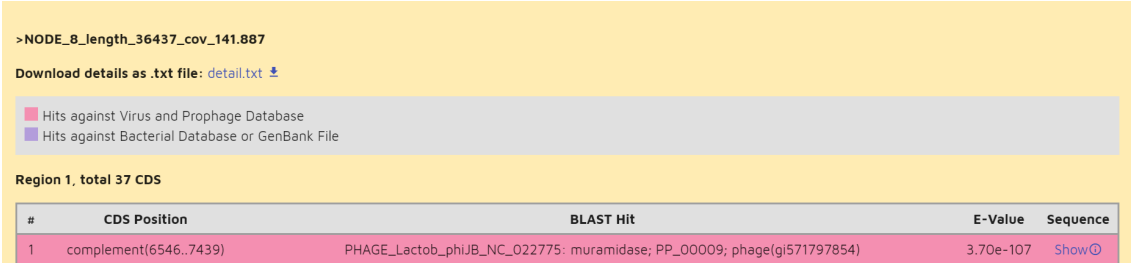

Figure 6: Partial result with the assembly of **FAM19191** run in PHASTER

Many other genes sequences of other PGHs, pseudogenes or hypothetical proteins were found in some or all of the six different strains. It would be interesting to pursue further analysis of the transcription and expression of these sequences, to assess more precisely the differences between the strains.

References

[1] R. Pérez-Torrado and A. Querol, "Opportunistic Strains of Saccharomyces cerevisiae: A Potential Risk Sold in Food Products," *Frontiers in Microbiology*, vol. 6, Jan. 2016.

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

[3] I. Belda, J. Ruiz, A. Santos, N. Van Wyk, and I. S. Pretorius, "Saccharomyces cerevisiae," *Trends in Genetics*, p. S0168952519301829, Oct. 2019.

[4] E. Seroussi, D. Kedra, M. Kost-Alimova, A.-C. Sandberg-Nordqvist, I. Fransson, J. F. Jacobs, Y. Fu, H.-Q. Pan, B. A. Roe, S. Imreh, and J. P. Dumanski, "TOM1genes Map to Human Chromosome 22q13.1 and Mouse Chromosome 8c1 and Encode Proteins Similar to the Endosomal Proteins HGS and STAM," *Genomics*, vol. 57, pp. 380–388, May 1999.

[5] L.-F. Seet, N. Liu, B. J. Hanson, and W. Hong, "Endofin Recruits TOM1 to Endosomes," *Journal of Biological Chemistry*, vol. 279, pp. 4670–4679, Feb. 2004.

[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] A. Saleh, M. Collart, J. A. Martens, J. Genereaux, S. Allard, J. Cote', and C. J. Brandl, "TOM1p, a yeast hect-domain protein which mediates transcriptional regulation through the ADA/SAGA coactivator complexes," *Journal of Molecular Biology*, vol. 282, pp. 933–946, Oct. 1998.

[8] 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.

[9] J. D. Dinman, "The Eukaryotic Ribosome: Current Status and Challenges," *Journal of Biological Chemistry*, vol. 284, pp. 11761–11765, May 2009.

[10] D. Kressler, E. Hurt, and J. Bassler, "Driving ribosome assembly," *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research*, vol. 1803, pp. 673–683, June 2010.

[11] K. Makioka, T. Yamazaki, M. Takatama, M. Ikeda, S. Murayama, K. Okamoto, and Y. Ikeda, "Immunolocalization of Tom1 in relation to protein degradation systems in Alzheimer's disease," *Journal of the Neurological Sciences*, vol. 365, pp. 101–107, June 2016.

[12] 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.

[13] I. Abnizova, R. t. Boekhorst, and Y. L. Orlov, "Computational Errors and Biases in Short Read Next Generation Sequencing," *Journal of Proteomics & Bioinformatics*, vol. 10, no. 1, 2017.

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

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

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

[17] 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.

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

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

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] 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.

[29] 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.

[30] 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.

[31] 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.

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

[33] 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.

[34] 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.

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

[36] 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.

[37] 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.

[38] 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.

[39] 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.

[40] 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.

Part I
Appendix

A Supplementary figures Yeast Genome Analysis

B Supplementary figures Arabidopsis Thaliana Genome Analysis

C Supplementary figures Lactobacillus Helveticus Genome Assembly

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

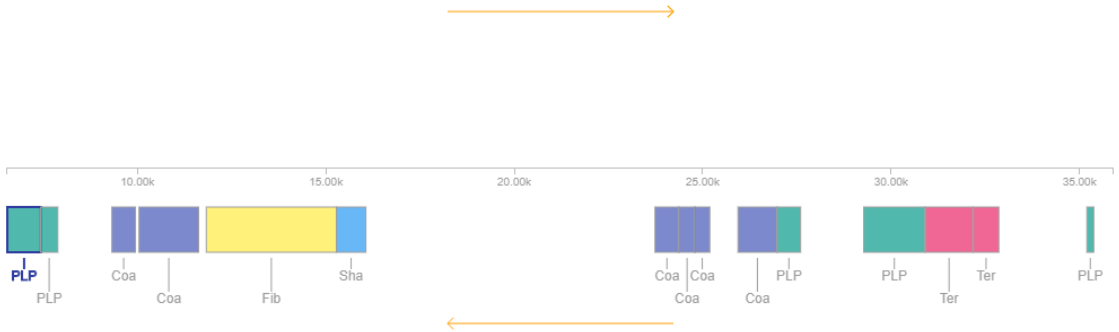

Figure 7: Node 8 of **FAM19191** assembly showing annotated locus. The highlited first one represents our muramidase.