# Bioinformatics Practicals In Sillico

**BC-7107**

Thibault Schowing
Lionel Rohner
Alain Rohrbasser
Rares Cristea

November 14, 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 mutation with these tools and learn how to design a bioinformatics 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.1 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 $\alpha$) 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 INTRO SERIEUSE SUR TOM1**

***tom1*** The deletion of Tom1 has been associated with aggregation of ribosomal proteins, which results in a temperature-sensitive phenotype of S.cerevisiae that renders them incapable of growing at temperature exceeding 20°C.

The aim of this study is to screen several Δtom1 *S.cerevisiae* strains for mutations in genes associated to tom1 (positive and negative regulators) and genes coding for riboproteins.

In this project, we present a number of potential candidates that might help deciphering the temperature-sensitive phenotype of Δtom1 *S.cerevisiae*.

## Methods

**Quality control of sequences** The high throughput sequencing method isn't infaillble, 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 [4]. 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* [5] which will help us visualize our fasta file given by the sequencer.

**Trimming and bad quality removal** After the quality control check, we used the tool *trimmomatic* [6] 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, containing the positional information.

However, in order to do that, we first had to index the reference fasta using the *bwa index* tool [7], 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 [8] 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 prefered considering the fact that our genome is a haploid yeast genome, and it doesn't need a complicated algorithm as used by the *GATK* pipeline. The *tabix* [9] 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.

**TODO Gene TOM ou autre polymorphisme trouvé à décrire dans le fichier vcf.**

# Arabidopsis Thaliana Genome Analysis

## Introduction

**GAI** Gibberellic-Acid Insensitive 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.Gibberellin is a tetracyclic diterpenoid growth factor and influence essentially the stem elongation and other plant developmental processes.If it's mutated (gai) and the plant growth better, it a gain of function gene, in contrary it's a loss of function. The cellular gai's component is in the nucleus and is described as a transcription region of DNA and bind it directly. 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 elongation1. The gai mutation affects GA reception or subsequent signal transduction and does not result in GA deficiency.Gai encodes a mutant protein that lacks a region of 17 amino acids from close to the N terminus and confers a dominant dwarf, reduced GA-response phenotype.The gai allele contains a deletion of 51-bp from within the GAI ORF. This in-frame deletion results in the absence of a 17-amino-acid residue segment situated close to the amino terminus of the predicted protein sequence.

**SPY** For spy, three independent recessive mutations at the SPINDLY (SPY) locus of Arabidopsis confer resistance to the gibberellin (GA) biosynthesis inhibitor paclobutrazol. Paclobutrazol or $\alpha$-tert-Butyl-$\beta$-(4-chlorobenzyl)-1H-1,2,4-triazole-1-ethanol, 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.

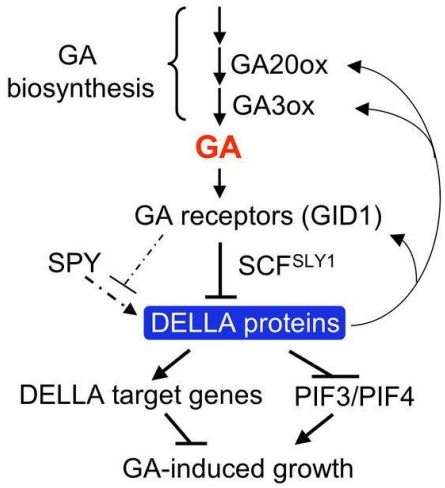



Figure 1: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3243332/figure/i1543-8120-64-1-1-f21/

# Methods

# 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 [10]. By releasing **peptidoglycan hydrolases**(PGHs), it has the ability to digest the bacterial cell wall (gram+) inducing death of surrounding bacteria but also its autolysis.

The genomic plasticity of *Lactobacillus helveticus* leads to a high variation in PGHs activity from one strain to another. In a previous study, 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 2). 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* [5] was used to check the quality of the reads and *Trimmomatic* [6] to filter out bad quality reads. *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 [11] 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* [12] to generate a pan-genome of the six strains. The result was then visualized with *Phandango* [13] allowing visualisation of phylogenetic tree, associated metadata and genomic information.

**Extraction of the genes for each phenotypes** Grep was applied to the files generated by *Roary* to extract the nine PHG's [10] labelled "Lhv_" with *PROKKA* (table 2). The set of genes

[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.

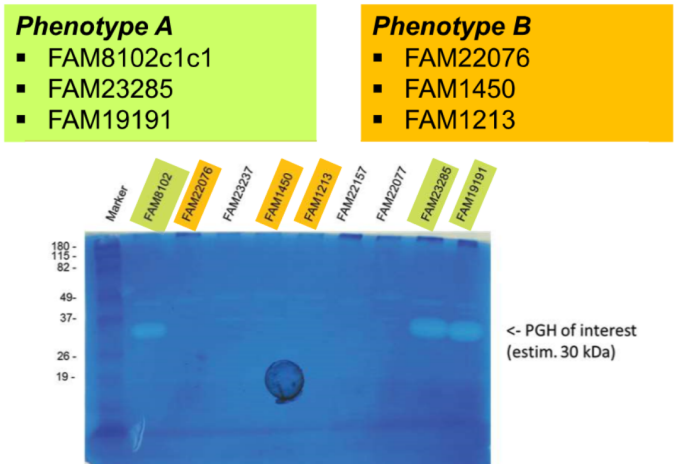
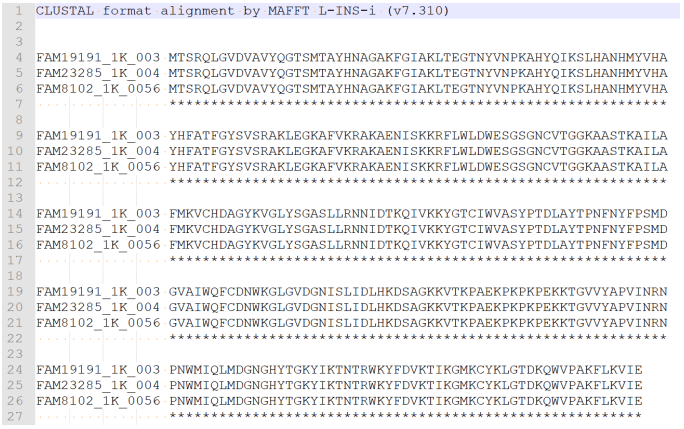



Figure 2: Phenotype A is expressing an active peptidoglycan hydrolase and phenotype B is not.

found in strains expressing phenotype A was then compared to the set of gene showing phenotype B. In table 1 we have the two PGHs present only in the three strains expressing the PGHs activity. The nucleotide sequences were then converted to amino acid sequences for further comparison.



```
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 3: Alignment of amino acid sequences of group 2372 for the three strains.

# 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 1: Genes present only in the three strains with a PGH activity.

According to figure 2, the PGH involved is approximately 30kDa thus matches with group 2372. Looking at the alignment of the amino acid sequences (Figure 3) 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 (table 2), therefore the phenotype observed in the figure 2 is not due to an absence of PGH.

Using BLASTp [14] with default parameters, the protein was searched to be a particular lysin (WP_101853908.1) encoded by the pneumococcal bacteriophage Cp-1 [15]. To look further into this sequence, we could use PHASTER [16], the PHAge Search Tool - Enhanced Release, which helps identifying and annotate prophage sequences within bacterial genomes and plasmids.

# 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] 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.

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

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

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

[8] 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.

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

[10] 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.

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

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

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