Bioinformatics Practicals In Sillico

BC-7107

Thibault Schowing
Lionel Rohner
Alain Rohrbasser
Rares Cristea

October 24, 2019

Introduction

Yeast Genome Analysis

Introduction

Biological introduction The budding Yeast Saccharomyces cerevisiae is a common organism used for genetics manipulation. This organism is well conserved among the eukaryote and can be used correlate with human pathways. With a genome with 16 chromosomes (haploid, Mat a or α) or 32 chromosomes (diploid). 99% of the genome is without introns, make this organism handy to manipulate. 12 million bases pair and contains between 5 800 to 6 572 genes [TODO REF]. The homology with human is estimate to 23%, which is a good candidate for preliminary studies regarding human pathways. The short mating time and growth is also short. Thus, the identification of potential mutant is grandly enhanced. 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. Due to the microscopic size (5 microM, between bacteria and human cell size) and simple growth environment, yeasts are inexpensive and easy to grow in silico. Saccharomyces cerevisiae is also no-pathogen, and forms colonies on agar plates in the laboratory in a few days with no special incubators required (best grow at 30 deg).

tom1

Methods

Arabidopsis Thaliana Genome Analysis

Introduction

Biological introduction

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 [1]. 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 1). 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* [2] was used to check the quality of the reads and *Trimmomatic* [3] to filter out bad quality reads. *SOAPdenovo* [4] as well as *Spades* [5] 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.

Genome annotation and pan-genome analysis We used the *PROKKA* pipeline [6] 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* [7] to generate a pan-genome of the six strains. The result was then visualized with *Phandango* [8] allowing visualisation of phylogenetic tree, associated metadata and genomic information.

Extraction of the genes for each phenotypes Grep [9] was applied to the files generated by *Roary* to extract the nine PHG's [1] labelled "Lhv_" with *PROKKA*. The set of genes found in strains expressing phenotype A was then compared to the set of gene showing phenotype B. In

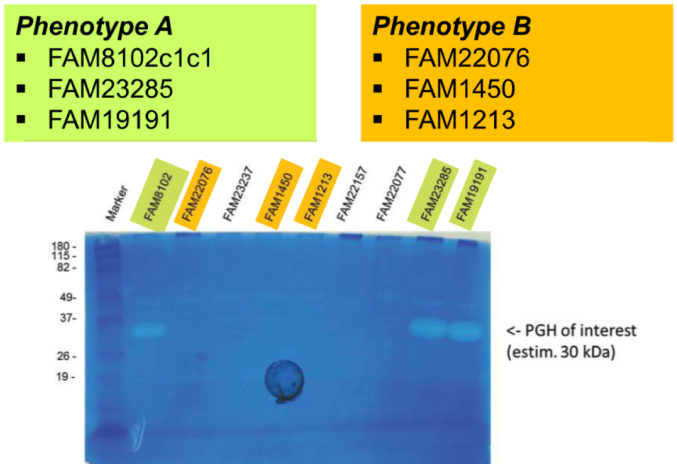

Figure 1: Phenotype A is expressing an active peptidoglycan hydrolase and phenotype B is not.

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.

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 1, the protein involved is approximately 30kDa thus matches with group 2372. Using BLASTp [10], the protein was searched to be a particular lysin (WP_101853908.1) encoded by the pneumococcal bacteriophage Cp-1 [11].

References

[1] 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.

[2] "Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data."

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

[4] "SOAPdenovo: short-read assembly."

[5] "SPAdes 3.0.0 Manual."

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

[7] 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.

[8] 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.

[9] "grep(1) - Linux manual page."

[10] 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.

[11] 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.