Bioinformatics Practicals In Sillico

Project

Thibault Schowing
Lionel Rohner
Alain Rohrbasser
Rares Cristea

October 23, 2019

Introduction

File types used

- **FA** The files with the .fa extension store FASTA format sequences. In this project the .fa file contains the reference genome.
- **GTF** The Gene transfer format (GTF) is a file format used to hold information about gene structure. It is a tab-delimited text format based on the general feature format (GFF), but contains some additional conventions specific to gene information. A significant feature of the GTF that can be validated: given a sequence and a GTF file, one can check that the format is correct. This significantly reduces problems with the interchange of data between groups.
- **VCF** The Variant Call Format stores the gene sequence variation. By using the variant call format only the variations need to be stored along with a reference genome which make the file less redundant.
- **BAM** Binary Alignment Map (BAM) is the comprehensive raw data of genome sequencing; it consists of the lossless, compressed binary representation of the Sequence Alignment Map. BAM is the compressed binary representation of SAM (Sequence Alignment Map). BAM is in compressed BGZF format.

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 Analysis

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 [4]. 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.

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* was used to check the quality of the reads and *Trimmomatic* 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.

Genome annotation and pan-genome analysis We used the *PROKKA* pipeline [1] 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* [2] to generate a pan-genome of the six strains. The result was then visualized with *Phandango* [3] 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 [4] 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.

Results

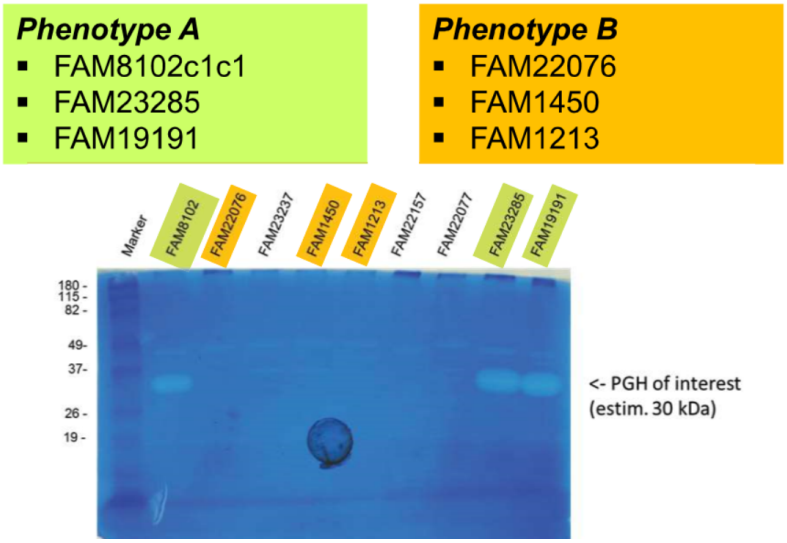

Figure 1: Phenotype A is expressing an active peptidoglycan hydrolase.

Bibliography

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

[2] 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.

[3] 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.

[4] 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.