Group C: Report DNA/RNA Sequencing Course Bread and Cheese

Max Suter, David Wittwer, (Audrey Minden)

November 16, 2015

Abstract

This is the final report of our work during the 2015 *DNA/RNA Sequencing Course*, which is part of the joint master's programme in Bioinformatics of the Universities of Bern and Fribourg.

We present the result of two tasks: (1) finding novel regulators/targets of TORC1 in genomes of yeast mutants, and (2) building and annotating a de-novo assembly of the genomes of five strains of Lactobacillus Paracasei. For both tasks, we're using free-of-charge and open source software tools which are well known in their respective field. (Quality control, assembling, SNP calling, annotation, visualization).

Using results from the first task, we are able to identify a candidate mutant, which shows interesting SNP mutations that are not yet known to be interferring with the TORC1 pathway.

Our main result from the second task is the newly assembled and annotated genomes of five strains of Paracasei.

Filename	Total Sequences	Sequence length	%GC
M1_S237_R1_001.fastq.gz	3493039	35-151	37
M1_S237_R2_001.fastq.gz	3493039	35-151	38
M14_S196_R1_001.fastq.gz	5619726	35-151	37
M14_S196_R2_001.fastq.gz	5619726	35-151	37
M16_S224_R1_001.fastq.gz	5311850	35-151	37
M16_S224_R2_001.fastq.gz	5311850	35-151	37
M18_S257_R1_001.fastq.gz	5483798	35-151	37
M18_S257_R2_001.fastq.gz	5483798	35-151	37
M21_S250_R1_001.fastq.gz	4089470	35-151	36
M21_S250_R2_001.fastq.gz	4089470	35-151	36
M24_S240_R1_001.fastq.gz	4029845	35-151	36
M24_S240_R2_001.fastq.gz	4029845	35-151	36

Table 1: Yeast Raw Data

Part I
Identification of novel regulators and/or targets of TORC1 in yeast mutants

1 Introduction

What is the goal?

Identify mutations that confer a growth phenotype after rapamycine treatment.

Asses, whether the involved proteins are already known to be involved in the TORC1 pathway.

If there are new candidates, we were trying to characterize their function as new regulators or targets of the TORC1 pathway.

2 Data and Methods

Paired-end DNA libraries were prepared in advance and handed over to us in gzipped FASTQ file format. Table 1 shows some basic information about the data. As a reference genome, we used *R64-1-1.79.fa*, which was provided by the course authorities.

Data processing was done in multiple stages using free-of-charge and open-sourse software tools. The main steps were executed manually on the High Performance Computing Cluster of the University of Bern.

Quality Control After receiving the raw data, we used the FastQC [1] tool to assess the quality of the reads. The quality of the data was below our expectations. Especially the reverse reads (*_R2_*-files) were of bad quality at high positions in the reads.

Filtering In order to remove bad quality data and trim the ends of the reads, we used Sickle [3] with the option for paired end sequence trimming on all our raw data files. The output was three files for each strain under consideration: a trimmed forward file, a trimmed reverse file, and a file containing trimmed single reads. All of which were in FASTQ-format.

Indexing, mapping, and sorting After indexing the reference genome using the BWA [4] tool, we were able to map each pair of files (R1 and R2) onto it. By further processing of the data with Samtools [5], which includes format conversion, sorting, and indexing, we produced a binary compressed BAM file, which is the binary version of a SAM file. A SAM file is a tab-delimited text file that contains sequence alignment data. Both formats are described on the SAM Tools web site [5].

SNP calling and identification of SNP effects In order to find potential interesting mutations within our strains of yeast ("variants"), we were looking for SNPs ("Single Nucleotide Polymorphisms"). Using SAM Tools [5], a combined VCF file was generated. By using the SNPeff tool, which is a genetic variant annotation and effect prediction toolbox [2], we tried to identify interesting point mutations and their effects. Table 2 shows the main output of this step.

Visualization IGV

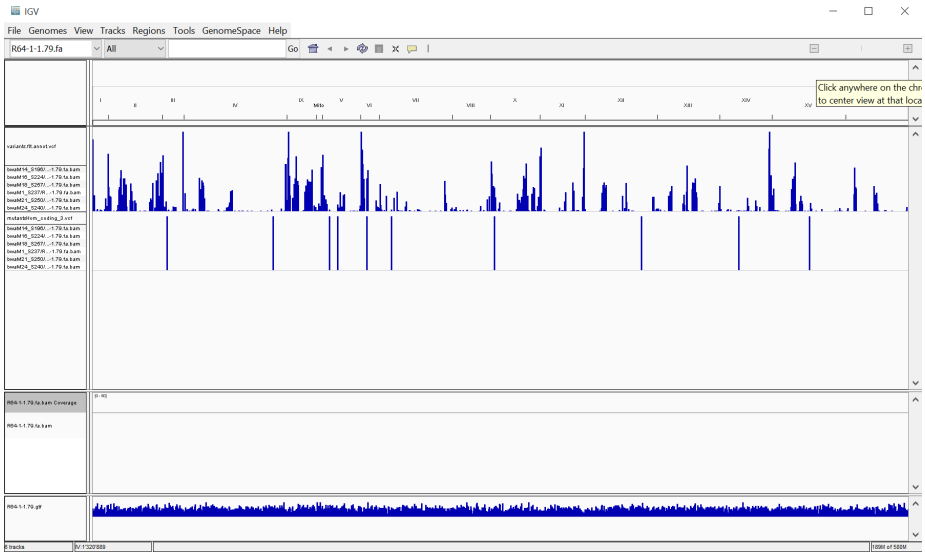

Figure 1: SNP visualization using IGV

3 Results

We identified the following SNPs to be interesting. Table xy shows these observed mutations.

Table 2: SNPs

3.1 Mutant I

Already known to be interferring w/ the TORC1 pathway. (Refernece)

3.2 Mutant II

Already known to be interferring w/ the TORC1 pathway. (Refernece)

3.3 Mutant III

Shows some significant differences. ...interesting.

4 Analysis and Discussion

Part II
Paracasei

1 Introduction

(Very) Short introduction to cheese making. Collaboration w/ Agroscope. Sequencing of different starters that they have in their library.

What is the goal of our work?

Which genes are associated with the growth phenotype?

What is the biochemical relation between VSC and the growth phenotype?

2 Data and Methods

We had the data from ...?

We used the following Pipeline...

Although the overall quality of the data was acceptable, we used sickle to correct and filter...

3 Results

4 Analysis and Discussion

A Appendix Some picture of data processing pipeline

B Appendix Second picture of data processing pipeline

References

[1] Babraham Bioinformatics. *FastQC - A quality control tool for high throughput sequence data*. Nov. 16, 2015. URL: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

[2] P. Cingolani et al. "A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3". In: *Fly* 6.2 (2012), pp. 80–92.

[3] Fass JN Joshi NA. *Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files*. Nov. 16, 2015. URL: https://github.com/najoshi/sickle.

[4] Durbin R Li H. *BWA: Fast and accurate long-read alignment with Burrows-Wheeler transform*. Nov. 16, 2015. URL: https://github.com/lh3/bwa.

[5] Wysoker A. Fennell T. Ruan J. Homer N. Marth G. Abecasis G. Durbin R. Li H. Handsaker B. and 1000 Genome Project Data Processing Subgroup. *The Sequence alignment/map (SAM) format and SAMtools*. Nov. 16, 2015. URL: http://www.htslib.org/.