Margus Veanes

Uppsala University, 2019

Genome Analysis

Paper 1 Wiki

Paper Overview:

Zhang et al. analyzed a problematic strain of Enterococcus faecium known as the E745 strain. Normally commensal in the intestines, the E745 strain of Enterococcus faecium has evolved antibiotic resistance as well as the ability to translocate from the intestines to the bloodstream. This is unusual for an otherwise commensal strain, as the working environments of the intestines versus the serum are completely different: one being heavily rich in nutrients and the other being nearly entirely nutrient poor.

Given its antibiotic resistance, Zhang et al. were curious to examine the genetic features of this specific strain as well as its expression profile in order to highlight potential genetic sources for its ability to survive in serum as well as to provide potential areas of interest for further research into non-antibiotic based treatment approaches for E745 infections.

Goals and Hypothesis:

The goal herein was to follow the workflow of the original paper, using their underlying data to attempt to independently reconstruct the results they observed. Firstly, the genome of the E. faecium strain must be established and annotated. Subsequently, once a successful genomic assembly has been established, the goal becomes determining the expression profile of the strain.

I hypothesize that using the original data I will be able to independently verify the original observed results, coming to similar conclusions as were originally determined. I expect there to be some variability in terms of my results as I will be using independently determined parameters for the software used. Additionally, certain software steps apply stochastically procedures which are also expected to introduce some variability in terms of final outcomes. Generally, I expect to generate a highly similar genomic assembly, with the transcriptomics side of the analysis being what may differ more compared to the original results.

Approach:

In order to replicate the approach that the original paper took, I broke my approach into two parts: *genomics and transcriptomics*.

For the genomics part of the analysis, I began by assembling the genome for the problematic E745 strain. This was done using both long read and short read data provided by PacBio and Illumina sources respectively. Subsequently, the genomic assembly was checked for quality, was annotated for genetic and structural features, and was finally compared against a related genome for synteny.

For the transcriptomics part of the analysis, I began by mapping their RNA expression data from both control (BH) conditions as well as treatment (Serum) conditions back to the initial annotated genome assembly. Subsequent counting was then performed on the mapped read data and finally an analysis of differential expression was employed in order to identify expression of genes of interest across treatments.

Further details regarding the software and results of both parts of my approaches are provided below.

Genomics:

-Genome Assembly via *CANU*:

CANU software was used to assemble the E. faecium strain using PacBio long read data. This generated a '.contigs.fasta' file which served as the primary assembly containing both unique and repetitive elements.

-Assembly Improvement via *FastQC, BWA, SamTools, and Pilon:*

Prior to improving the CANU assembly, the Illumina short reads were checked for quality using FastQC software. The resulting read quality analysis indicated that the Illumina reads were of high enough quality to use without trimming as the reads were all in the green region of the software's output graph.

Subsequently, BWA was used to map the quality checked Illumina reads to the original CANU assembly. This generated a '.sam' filetype which needed to be converted to '.bam' format using the SamTools software.

Finally, using the converted '.bam' file, Pilon software was employed to generate a consensus genome from the Illumina reads mapped to the PacBio genome. This generated a higher quality consensus genome in a '.fasta' format which was used for all subsequent analysis in the transcriptomics part of the analysis.

-Assembly Quality Assessment via *QUAST*:

Furthermore, the consensus genome was evaluated via QUAST software in order to ensure its quality for use in subsequent transcriptomics steps. Among the results, the mummerplot graph indicated a successful assembly as the assembled genome plotted successfully against a reference genome of E. faecium and a neither broken nor problematic alignment was observed over the expected 2.8m genomic length.

**Critical note, during initial assembly using CANU, the software required a 'genomeSize=##' input parameter. A mistake was made, as the input provided was* *'genomeSize=2.8m' rather than 'genomeSize=3.1m'. Given that the core circular genome of the E. faecium is ~2.75m base pairs in length, and its additional plasmids are an additional ~300k in length, using my erroneous input parameters only the core genome was assembled.*

Due to time limitations and late realization of the original mistake, analysis proceeded using the assembly containing only the core genome. This was determined to be acceptable considering the context of the paper, as one of the original goals were to identify potential genes of interest for further research into possible means of treatment. Given that bacterial evolution rate is high due to their ability to horizontally transfer genes mainly via the exchange of plasmids, analysis into only the core genome still provides an interesting perspective as bacterial populations are expected to be slower to evolve resistance to treatments targeting genes within the core genome (barring these targets being located within transposable elements)

-Genome annotation via *PROKKA:*

The resulting Pilon assembly was annotated using Prokka software. This resulted in a '.gbk' GeneBank filetype which needed to be converted to '.gtf' format for use with subsequent software. This was accomplished via a python script which functioned to convert from GeneBank to gtf format. The resulting '.gtf' file contained the annotation information for the genomic assembly.

-Synteny Analysis via *ACT:*

In order to examine the genomic assembly and confirm its expected similarity to a closely related genome, a similar strain of E. faecium was sourced from NCBI and aligned to our annotated genome assembly and compared using ACT visualization. What was seen indicated highly similar regions across a large part of the genome (blue areas in ACT visualization) with some regions having been moved around from their expected location, possibly relating back to what makes the E745 strain unique in its pathogenicity.

Transcriptomics:

-Illumina read quality control via *FastQC:*

Illumina based RNA reads were initially checked for quality via FastQC, the resulting graph indicated the reads to be of high quality, as the reads appeared in the green region of the software's output graph.

-Read trimming via *Trimmomatic:*

Trimming was performed on untrimmed RNA reads for Serum treatment data via the Trimmomatic software. This ensured the removal of unnecessary elements within the read data like adapters and cleaned the data for subsequent mapping and counting analysis.

-Read mapping via *BWA* and *SamTools:*

The cleaned RNA reads were subsequently mapped onto the consensus genome assembly (the PacBio + Illumina genome assembly) using BWA one again. One again the resulting '.sam' filetype was converted to '.bam' filetype via Samtools software for use in subsequent counting analysis.

-Read counting via *HTseq:*

With the RNA reads from both treatments as well as all replicates individually mapped to the consensus genome, these respective '.bam' files of reads mapped to the genome were fed into the HTseq software in order to generate gene counts using as a reference the earlier generated '.gtf' file which served as the annotation for our genome.

-Differential expression analysis via *DESeq2:*

In order to make sense of the count data generated above, finally the DESeq2 library was employed in R's markdown in order to perform a differential expression analysis on our RNA expression count results across all replicates for both treatments provided by HTseq. DEseq2 analyzed the provided HTseq count data and performed a differential expression analysis across the treatments and their replicates. Subsequently, using the 'pheatmap' library useful visual graphs were generated for visualizing the top genes of interest in terms of the log2fold change across treatments. Furthermore, R was used to generate a CSV file of all counts for visual reference of the raw count data that was generated by HTseq.

Conclusions and Remarks:

The project undertaken above, the main purpose being to independently replicate the methods and ideally also the results observed in the original paper, has taught me of the complex intricacies related to such research. Not only does one need to become intimately aware of each software intended to be used, but one must also be careful in guiding the software in terms of any requested analysis.

A dogma in biological research is 'crap in crap out,' a saying which relates to the philosophy that results can only be of the same quality or worse as the input provided. Essentially, the results of any analysis do not only depend on the input data quality, but also on the correct curation of the data. In this case, curation related to the correct use of the analysis software and in my case also lead to the initial misstep of generating an assembly only for the core genome of the bacterium in question.

Given that my assembly contained only the core genome of ~2.75m bp in length and not the associated plasmids which were also analyzed in the original paper, my results did differ somewhat in terms of the genes that were identified to be differentially expressed across treatments. However, even considering my differing assembly size compared to the paper, certain results still ended up being highly similar to what was identified in the original paper. Namely, my differential expression analysis indicated a significant upregulation in expression of purine biosynthesis genes in treatment (serum) compared to control (BH) conditions. This was mirrored in the paper which noted, *"The pruine metabolism genes… were found to be required for growth in serum… [and] were among those that were significantly upregulated during growth in serum compared to growth in rich medium" [1]*

Furthermore, my results also indicated that across treatments there appeared to be significant downregulation in serum compared to BH, of genes related to fatty acid biosynthesis (*acpA_1)* as well as a downregulation of genes related to sugar import (*lacS)*. This lends further credibility to my observed results, as logic would dictate a downregulation of nutrient related genes in a nutrient poor environment like serum.

Given that my results closely mirrored a central takeaway from the original paper coupled to the logic behind certain observed results, I consider my approach to have been largely successful in independently following the original paper's approach as well as successful in verifying certain core results in terms of what was observed to be differentially expressed in the original paper and in my results.

References:

[1] Zhang, Xinglin, et al. “RNA-Seq and Tn-Seq Reveal Fitness Determinants of Vancomycin-Resistant Enterococcus Faecium during Growth in Human Serum.” *BMC Genomics*, vol. 18, no. 1, 2017, pp. 893–12., doi:10.1186/s12864-017-4299-9. Accessed 25 May 2019.