Sequencing, Assembling, and Annotating the *Kaistella koreensis* Genome

Timothy Hostelley, Jeffrey Newman

Department of Biology, Lycoming College, 700 College Place, Williamsport, PA 17701

**Abstract**

Advancements in DNA sequencing technology have made DNA sequencing cheaper, and more efficient. As a result, there has been an enormous increase in the number of genomes being sequenced. The sequence data can be assembled into complete genomes and annotated in order to reveal information about the organism’s physiology. The objective of this study was to sequence, assemble, and annotate the genome of the bacteria *Kaistella koreensis*.

The DNA of *Kaistella Koreensis* was isolated as a part of the Molecular Biology class in the fall of 2011. The isolated DNA was then sequenced via Ion-Torrent as a part of the GCAT-Seek consortium, and gave approximately 650,000 reads. These reads were assembled using NextGene software into 3,058 contigs. The assembled contigs were then imported into Geneious to be assembled into supercontigs. To increase the efficiency of the supercontig assembly, trimming was done, using Geneious, on the ends of each contig due to increased likelihood of errors near the ends. Several trimming lengths were tested until the ideal length was determined that would assemble the most contigs into the fewest supercontigs. This length was 50 bases trimmed from each end of the contig. The 3,058 contigs were assembled into 482 supercontigs in the Geneious program. Next, to decrease the ambiguities in the supercontig consensus sequence, bad contigs were removed from the assembly. Bad contigs were defined as contigs that were shorter, located completely within the sequences of larger contigs, and only added ambiguity to the sequence. Once the bad contigs were removed, the remaining ambiguities were manually corrected based on the assumption that bases closer to the ends are more likely to be wrong. The supercontig consensus sequences were then uploaded to the Rapid Annotation using Subsystem Technology website for initial annotation and comparison to reference genomes.

**Background**

*Kaistella koreensis* was recently reclassified as *Chryseobacterium koreense* and we disagree with this reclassification.

**Hypothesis**

The overall genome of *Kaistella koreensis* will be significantly different than those of the *Chryseobacterium* genus, disputing the reclassification.

**Methods**

- **DNA Isolation**
- **Ion-Torrent Sequencing**
  - 650,000 reads
- **NextGen Assembly**
  - 3,058 contigs
- **Geneious Assembly**
  - 483 supercontigs
- **Correct ambiguities in consensus sequence**
- **Assembly 2 with Geneious**
- **Trim ends and remove bad contigs**
- **Sequence based comparison to reference genome in same family**
- **Upload contigs to RAST**
- **RAST Annotation**

**Results**

- **Trimming Ends**
  - Ends of each contig trimmed due to increased likelihood of errors at ends
  - Ideal trimming length of 50 bases
- **Bad Contig Removal**
  - Removed bad contigs to decrease ambiguity
  - Bad contigs were defined as contigs that were shorter, located completely within the sequences of larger contigs, and only added ambiguity to the sequence
- **Geneious Assembly**
  - 2,517 of 2,884 contigs were assembled to produce 483 supercontigs
  - 367 contigs were not assembled
- **Correcting Ambiguities**
  - Ambiguities were manually corrected based on the assumption that bases closer to the ends are more likely to be wrong
- **RAST Annotation**
  - Rapid Annotation using Subsystems Technology

**Conclusions**

- *Kaistella koreensis* has a much smaller genome size than that of the *Chryseobacterium* genus
- *Kaistella koreensis* genome is most similar to that of Antarctic Ice Bacterium Flavobacteriaceae sp. 3519-10, NOT *Chryseobacterium*.

**References**


**Acknowledgments**

The author would like to thank the GCAT-SEEK consortium for funding of the genome sequencing and the Lycoming College Biology Department for funding of this project.