Tracking a Hospital Outbreak of Carbapenem-Resistant *Klebsiella pneumoniae* with Whole-Genome Sequencing

Dreycey Albin (da39)

April 16, 2019

Scientific question

The scientific question and goal for this paper focuses on *using available genomics and epidemiological information to reconstruct a transmission tree of a carbapenemase resistant Klebsiella pneumoniae outbreak in a NIH clincal center.*

Hypothesis tested

The hypothesis tested in this research is that *their algorithm, couple with unique genomic and epidemiolgical data, will be able to reconstruct the transmission tree for the Klebsiella pneumoniae outbreak.* They tested this hypothesis by utilizing the samples from multiple patients, from the groin and throat, and genomically sequencing the samples. Thereafter, a parsimonous tree was made by connecting patients based on the genetic similiairity between the samples. The most likely tree was selected based on the patient trace data, with information on where the patients were located during their stay in the NIH clinic.

Methods/Results

0.1 Initial Testing

The methods seem to span multiple types of analysis, ranging from experimental validation and verification, to bioinformatic methods to build contigs, track single nucleotide varients, and to reconstruct the most likely transmission tree. First the strain of patient zero was indicated to be from the epidemic strain ST 258 *K. pneumoniae* lineage. Next, the genetic heterogeniety of the samples was seen by testing multiple sites of the index paitent, as they found 7 different varients by testing 4 seperate body sites.

0.2 Construction Putative Transmission Map

The transmisssion map was constructed using a combination of both epidemiological and genomic data. The samples collected from the patients were seqeunced on a Roche 454 instrument. From there gsAssembler v.2.3 was used to assemble the contigs, resulting in a N50 contig size of 154,336 bp. These contigs were used to form pseudo contigs in Mauve, and thereafter the high confidence SNVs were extracted by aligning the pseudo chromosomes to a reference (again using Mauve).

Next, the transmission map was constructed. Patients were connected using edges to related patient samples, based on genetic distance. The most parsimonous tree was the one that spanned all the patients and also minimized the total genetic distance for all of the samples. This was the classic problem of the minimum spanning tree. The authors utilized the widely used Edmonds algorithm to meet this criteria. The way the authors did this was pretty interesting, but overall was, in my own opinion, a general heuristic to calculate distances. simplifying the distance equation for the distance matrix gave the following equation (when patients overlapped, or else it was set to max(E)):

$$D_{AB} = G_{AB} + 0.09 * (\frac{E_{AB}}{max(E)})$$

This yielded a non-symmetric distance matrix, in which Edmond's algorithm could be applied. Note that the epidemiological information on the right term of the above equation was down-weighted. The authors did this so that epidemiological data would only play a role when the distance based on genetic inference was equally likely, $G_{AB} = G_{BA}$.

Key implications of the results

The results of this study were crucial in understanding how the resistant K. pneumoniae spread throughout the NIH clincal wards. From the given information, they were able to quantitatively show that the index patient had spread the infection to several other patients. This included finding that the transmission was not always correlated with symptomatic response times, as in the case of 1 to 3 to 2. Likewise, they saw that there were 3 different resulting clusters from this information, where the transmission routes were not always completely identified, such as the case with patient 4, where the transmission process is not completely understood.