A Proposal of Genomic Analytical Workflow in a Bacterial Pathogen Outbreak Investigation

Hoi Shan KWAN*, Chun Hang AU, Chi Keung CHENG, Man Kit Cheung, Qianli HUANG, Lei LI, Wenyan NONG & Man Chun WONG

Food Research Centre, School of Life Sciences,
The Chinese University of Hong Kong, Shatin, Hong Kong SAR, China

*Correspondence: hskwan@eservices.cuhk.edu.hk

A. Introduction

The German *Escherichia coli* O104:H4 outbreak causing serious diseases, notably, Hemolytic Uremic Syndrome (HUS) started in early May. The conventional typing methods including serotyping, multi-locus sequence typing (MLST) and pulse-field gel electrophoresis (PFGE) seemed to yield only limited information about the strain. The situation was a little chaotic. The strain could not be typed definitively as enteroaggregative *E. coli* (EAEC) or enterohemorrhagic *E. coli* (EHEC). It was suspected to be a hybrid of EAEC and EHEC. More detailed analysis was needed.

Nick Loman started a very interesting “Crowdsourcing” genomic analysis of the German *Escherichia coli* outbreak strain (https://github.com/ehec-outbreak-crowdsourced/BGI-data-analysis/wiki/). The first set of analyses was based on the Beijing Genomics Institute (BGI) genome sequencing reads. The first two sets of sequencing reads were provided by BGI and Life Technologies, both used the Ion Torrent Personal Genome Machine (PGM). Ion Torrent PGM was the machine of choice for its quick operation time. It seemed that the draft sequences were completed in only three days albeit the coverage was low. The quick turnaround time allowed “Crowdsourcing” to start very early after the strains were given to BGI and Life Technologies to sequence. The speed of sequencing was important to give an initial quick analysis of the outbreak strain to show its major characteristics. Treatment could be chosen with this information. For deadly diseases, timing is always very important. Certainly, a few days later, the more powerful machines, Illumina HiSeq and Roche 454 Titanium, provided higher coverage for better assembly which was important for high resolution analyses.

B. Questions and approaches to answer them

My understanding from the “Crowdsourcing” experience is that we have been asking several important questions about the strain responsible for the outbreak and using genome analysis to find the answers.

The questions are:

1. Is the outbreak strain isolated previously? That is, is it a known strain?
2. If not known, is it related to known strains?
3. If answer is no to above, then is it novel?
4. What are the pathogenicity features of the strain?
5. Is a typing scheme available to follow further cases?
6. If not, can we develop a typing scheme?

Answering these questions would allow us to understand the strain better and help us to:

1. trace the outbreak strain,
2. find appropriate treatment for the disease,
3. understand the pathogenicity of the outbreak strain,
4. develop preventive measures for further and future outbreaks.

The genome sequencing combined with “Crowdsourcing” seemed to work well. We could answer the questions and further use the answers for treatment and control of the outbreak. Amazed by the efficiency and effectiveness of the process, I analyzed the genome analyses reported by the laboratories contributing to “Crowdsourcing” to find out which analysis is for answering each question.

C. Synthesis: Analytical workflow in a pathogenic bacteria outbreak

My synthesis from the genome sequencing-crowdsourcing experience is shown in Figure 1. The advantages of Genome sequencing approach are apparent comparing to conventional methods:

1. Genomic approach is definitive with little uncertainty
2. Genomic approach reveals whether the outbreak strain is known, related, or novel
3. Genomic analysis provides comprehensive information on pathogenicity of the outbreak strain
4. Comparative Genomics provide information to develop typing schemes for diagnosis of the outbreak strain in patients and food, important for management of the outbreak
5. Genomics provide lots of data for further comprehensive understanding of the outbreak strain, important for prevention of further and future outbreaks

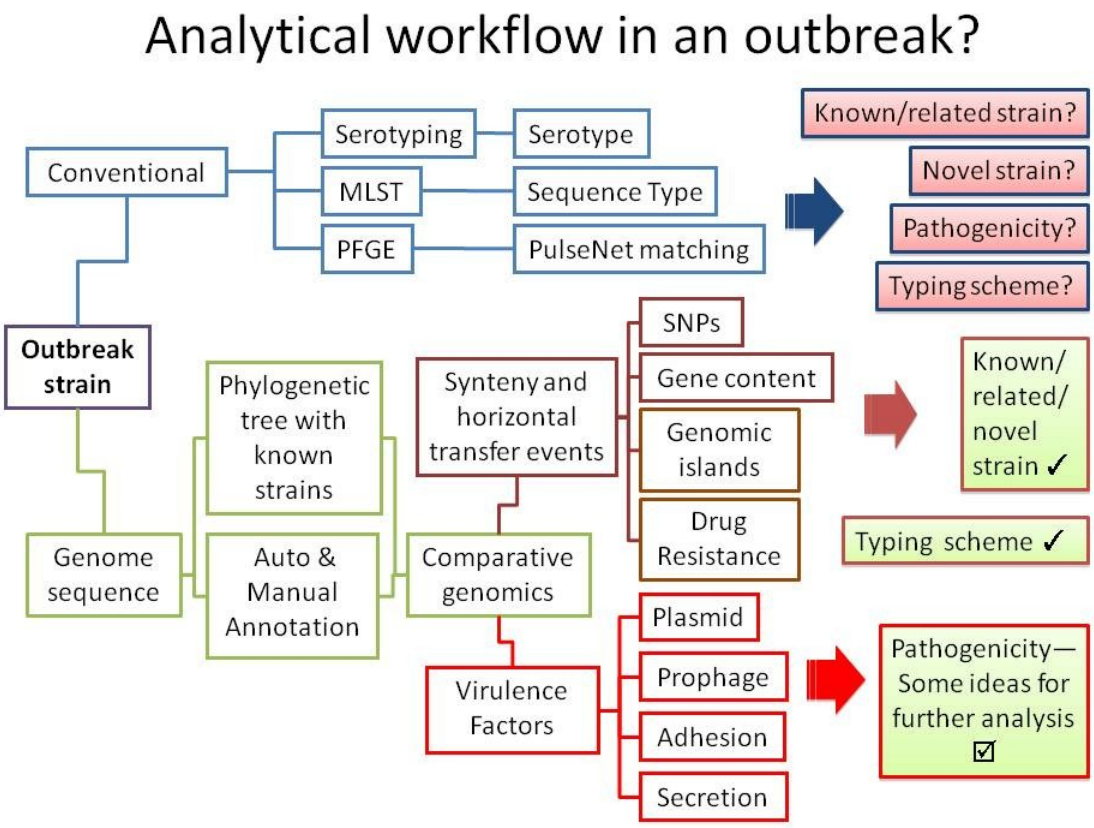

Figure 1. Putative analytical workflows in a bacterial outbreak.

1. Top group, conventional workflow: after typing and analysis, there is still great uncertainty about the relationship of outbreak strain to previously isolated and analyzed strains. It is uncertain whether it is novel. Not much is known about the pathogenicity feature of the strain. The data cannot be used to generate a typing scheme.
2. Bottom group, genome sequencing and genomic analysis: after the analysis, the strain can be shown to be a known strain, or related to a known strain, or a novel strain. The pathogenicity is well studied at the genomic/gene level, providing knowledge for further analysis.
3. The right columns of boxes show the questions and whether they can be answered. The question marks indicate much uncertainties. The tick marks indicate answers can be provided. The tick mark in the box means some information can be obtained for further analysis.

D. Concluding remarks

Genome sequencing and analyses with an appropriate workflow would be the new paradigm in bacterial outbreak investigation. It is important to have a fast genome sequencing turnaround time and later a high coverage sequencing for

further detailed analyses. This approach still costs dearly and can be prohibitive. When we can perform inexpensive, fast, and high-throughput genome sequencing, the genomic analysis paradigm can be established.