Pathogen detection by shotgun metagenome sequencing

Whether it is salmonella in your chicken sandwich, listeria in your favourite cheese, or aspergillus on your fruits, unwanted microbes in the food production chain are a danger to human health, and are costly to monitor for the food industry. Therefore, it is important to have an efficient, fast and accurate way to detect pathogens along the entire food production chain.

by The Technical Team, BaseClear, The Netherlands. www.baseclear.com

Historically, culture-based methods are seen as the golden standard in the detection of foodborne pathogens. Although these type of tests are very affordable and simple, they have notable downsides. Culture-based methods are time consuming, require additional analysis as a follow up, and will only show you pathogens that will grow on your media of choice.

In recent years metagenomic sequencing techniques have developed significantly, becoming an accessible, faster and affordable alternative testing method. For this reason an increasing number of companies in the food industry are looking at metagenomics for pathogen detection. However, for the data analysis and interpretation, appropriate and validated bioinformatics and biostatistical tools are required.

Therefore, BaseClear has developed an innovative pathogen detection method that is comprehensive, fast and able to detect contaminants at low abundance levels. This rapid method is assembly-free (based on state-of-the-art tools Kraken2 and Bracken) and uses specific pathogen databases and downstream analyses to check for low abundant potentially pathogenic species. Their pipeline can be used on metagenomic sequencing data from a wide array of food industry samples, for example dairy, meat, plant based and pre-processed meals.

Table 1. Overview of pathogens present at low abundance in the ZymoBIOMICS mock community.

Species	Abundance
Salmonella enterica	0.01%
Enterococcus faecalis	0.001%
Clostridium perfringens	0.0001%

Detecting pathogens below 0.01% abundance

In metagenomics analysis generally low abundant species are difficult to distinguish from false positive hits. Common bioinformatics tools discard all results below 1% abundance, which means detecting low abundant pathogens is not possible. However many pathogens can act at a low infective dose, for example Salmonella spp.

BaseClear's newly developed bioinformatics method can distinguish false positives from low abundant hits. They tested this with a sequenced mock community including the low abundant species shown in Table 1.

To distinguish true positives from false positives, they use the amount of distinct genomic regions that are found in the data of your samples. Their pipeline usually finds sequencing reads matching throughout the entire genome of species that are truly present in a sample.

On the other hand, species that are not truly present but are detected by mistake only show a few genomic regions with a match. This could be due to sequencing error.

BaseClear can use this amount of distinct matching regions in the genome, or 'minimisers' to distinguish between true and false positives by setting a threshold.

This is visualised in Fig. 1, where the amount of minimisers is plotted for every hit in the data. Here we see that many species have a high amount of distinct genomic regions (red/orange), while there is a long tail of samples with a low amount of distinct regions (green).

The species expected to be truly present in the sample coincide with the species with high abundant distinct regions.

Based on testing and validation on different simulated samples, a distinct minimiser threshold system can be set. For example, this allowed BaseClear to detect both S. enterica and E. faecalis present at only 0.01 and 0.001% in the mock community.

Continued on page 20

Continued from page 19

Functional analysis to detect true pathogenic species

When a potential pathogenic species is detected, it is not always known if this species is an actual pathogen. Looking into the functional potential (genes present in their genome) gives the user an idea if the detected species might actually be able to be harmful.

Therefore, the company also implemented an additional tool (based on HUMAnN3) to detect which gene families and pathways are present in the sample after detection of the potential pathogen. Samples were simulated in silico, which contains pathogenic Salmonella enterica strains. The gene families detected were compared with the gene families known to be present in different pathogenic S. enterica subtypes.

Multiple virulence factors were found, such as:

- **Agf (Thin aggregative fimbriae (or curli):** Aids in attachment to the villi of enterocytes, also cause the bacteria to become attached to each other.
- **Lpf (Long polar fimbriae):** Extracellular matrix adhesin involved in intestinal colonisation.
- **VI antigen:** Prevents antibody-mediated opsonisation, increases resistance to host peroxide and resistance to complement activation by the alternate pathway and complement-mediated lysis.
- **CdtB:** Involves chromatin disruption, which leads to G2/M-phase growth arrest of the target cell and ultimately cell death.

If such virulence factors are detected it is very likely that the pathogen detected in your sample is actually harmful. This functional analysis thus offers an extra layer of information on top of the taxonomic profiles, and can aid in distinguishing contaminated from non-contaminated samples.

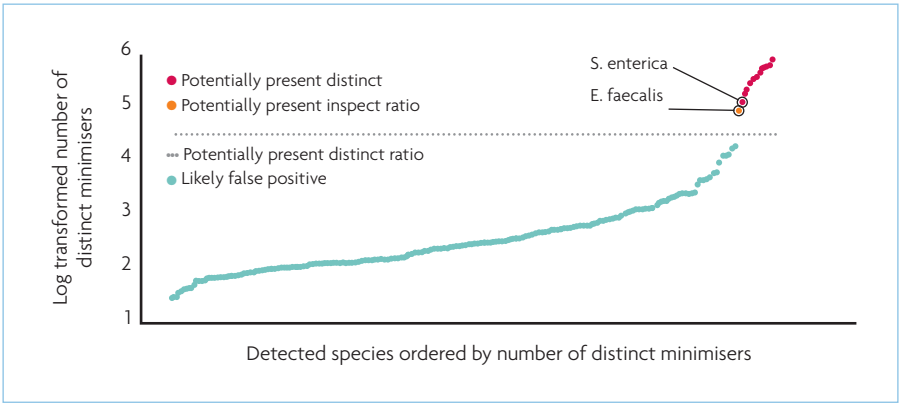

Fig. 1. The amount of distinct minimisers plotted per species detected in a mock community. The red and orange coloured data points indicate species expected and found in the sample. The grey line represents the threshold used. The green coloured data points indicate false positives that are filtered out.

Conclusions

With the development of BaseClear's pathogen detection pipeline, they can now offer false-positive adjusted species-level taxonomic analysis by shotgun metagenomics sequencing.

With this pipeline it is possible to quickly scan for food pathogens in metagenome data and detect known and unknown pathogens.

State-of-the-art tools have been combined into a pipeline with a more user friendly output and a way to better distinguish low abundant pathogens from false positives.

The advantages of this pathogen detection pipeline include:

- Dedicated databases covering pathogenic strains from different domains.
- The database is customisable according to the client needs.
- False-positive adjustment reduces the number of low abundance false positives.
- Fast detection of contamination.
- Detection of low abundant species.
- User friendly output table.
- Additional functional analysis to investigate the pathogenicity of the detected species.

Altogether, this integrated metagenome pipeline offers a reliable approach to quickly detect pathogens in the food industry environment, which has great potential for accurate risk assessment, food safety and public health.

At the same time this method replaces more laborious and less precise traditional methods of pathogen detection. ■

References are available from the author on request