NIHR | Health Protection Research Unit in Gastrointestinal Infections at University of Liverpool

Preliminary investigation of the co-occurrence of gastrointestinal pathogens targeted by the Luminex GPP panel in genomic data

Edward Cunningham-Oakes[1], K. Marie Mcintyre[1], Alistair C. Darby[1], Nigel Cunliffe[1] and Sarah O'Brien[2]

[1]Institute of Infection, Veterinary and Ecological Sciences, University of Liverpool, Leahurst Campus, Neston, Wirral, UK; [2]School of Nature and Environmental Sciences, Newcastle University, Newcastle upon Tyne, UK

Introduction

Traditional surveillance methods tend to detect point source outbreaks of diarrhoea and vomiting, however, they are less effective at identifying low-level and intermittent contamination of the food supply, unless the organism is very rare[1]. Further, it may take up to nine weeks for infections to be confirmed by a reference laboratory, reducing recognition of 'slow-burn' outbreaks that can affect hundreds or thousands of people over a wide geographical area[2]. There is a need to address fundamental problems inherent in traditional surveillance for diarrhoeal disease. The overall aim of the INTEGRATE study[3] was to create a new, one-health paradigm for detecting and investigating clusters and outbreaks of diarrhoea and vomiting in the community, shifting from passive surveillance and management of laboratory-confirmed infection to enhanced surveillance and management of people with symptoms. This was achieved by obtaining stool samples from willing participants, who presented to GPs with GI symptoms. Using bioinformatic and statistical methodologies, we highlight congruence and disparity between traditional, molecular and sequencing-based methodologies for the detection of GI pathogens, as well as instances where pathogen co-occurrence is indicated.

Methods

Quality-control of DNA and RNA reads

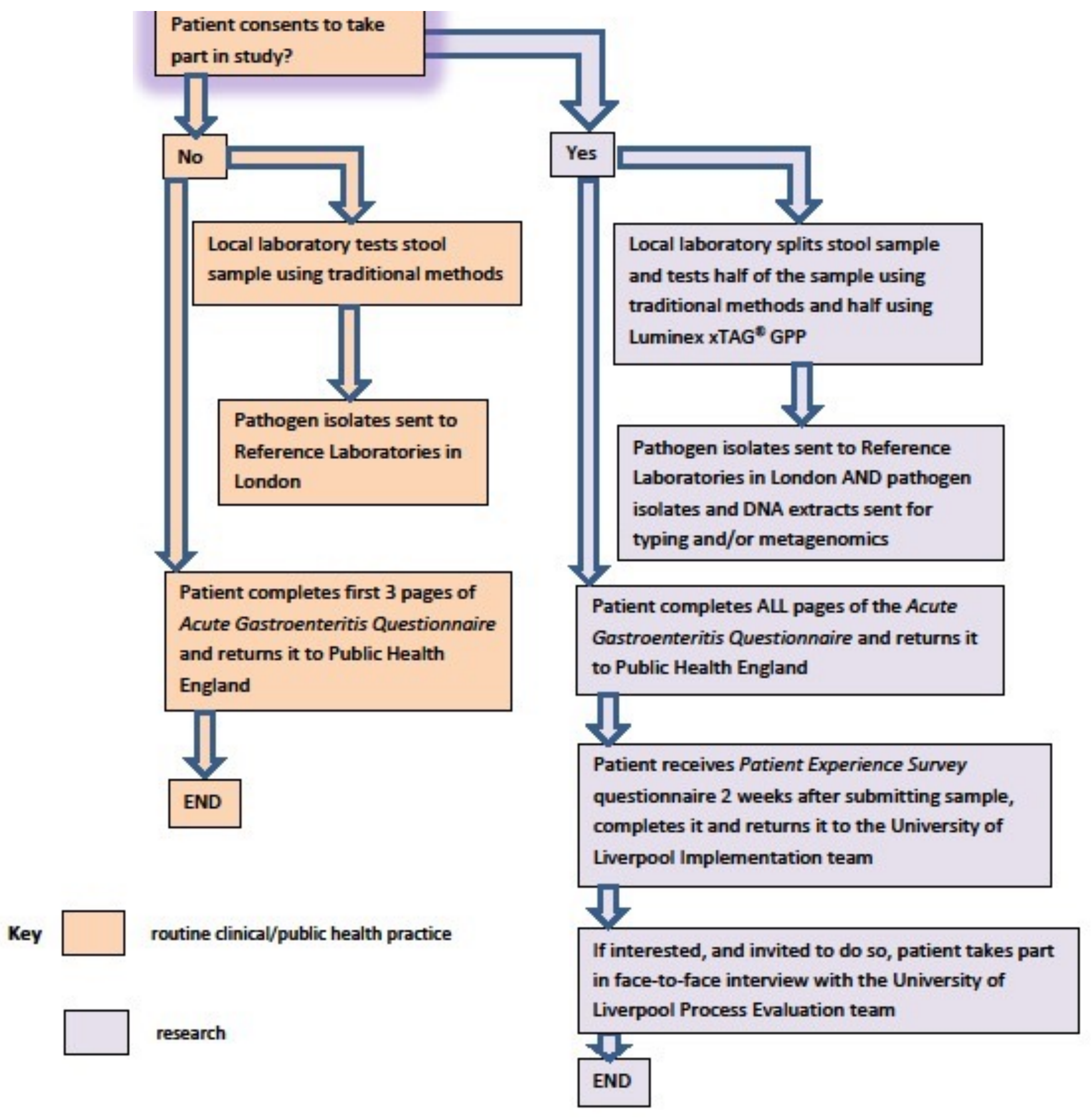
The "clean_reads" module of the MetaWRAP (v1.3.2) pipeline, was used to remove low-quality, adapter and human contamination sequences from DNA (*n* = 1022) and RNA (*n* = 1060) reads. Human contamination was removed via alignment to the hg38 reference genome.

Correlation of genomic reads assigned to taxa of interest with laboratory data

- Taxonomy was assigned to DNA and RNA reads using the k-mer based tool, Kraken2 (v2.1.2), using a confidence threshold of 0.1.
- Taxonomy was assigned using a custom database, which included the libraries archaea, bacteria, fungi, human, plant, plasmid, protozoa, UniVec_Core and viral.
- Taxa of interest were Adenovirus, Astrovirus, *Campylobacter*, *Clostridioides difficile* (*C. difficile*), *Cryptosporidium*, *Entamoeba histolytica* (*E. histolytica*), *Giardia*, Norovirus, Rotavirus (A), *Salmonella*, Sapovirus, *Shigella*, *Vibrio cholerae* and *Yersinia enterocolitica*. Adenovirus 40/41 and *Escherichia coli* (*E. coli*) results were not considered, as k-mer based methodologies do not lend themselves well to the level of specificity required to correlate to Luminex.
- For DNA reads, read counts assigned to taxonomies in each sample were then re-estimated with the average read length of that sample, using Bracken (v2.0).
- Kraken-biom (v1.0.1) was used to generate a biom file in json format from Kraken and Bracken reports from DNA and RNA samples respectively. Biom (v2.1.6) was then used to assign tabulated metadata to the biom file.

Statistically significant associations of read taxonomy and laboratory diagnostics

- Associations between read counts, traditional laboratory diagnostics, Luminex laboratory diagnostics for organisms of interest, and all associated plots, were generated using the multivariate linear regression tool, MaAsLin 2(v1.6.0). Analysis used a linear model under default settings. DNA and RNA results were stratified prior to analysis. For Astrovirus, only traditional results were available for correlation, whilst for Rotavirus, only Luminex results were available.
- A minimum threshold of 0.05 was then applied to q-values, (p-values adjusted for the False Discovery Rate), to select for coefficients with statistically significant results.
- A final matrix was then generated using extracted coefficients, before generating a heatmap using heatmaply(v1.3.0). Significance levels according to q-value were manually annotated onto the final heatmap in Inkscape (v1.1.1)



Patient consents to take part in study?

No

Local laboratory tests stool sample using traditional methods

Pathogen isolates sent to Reference Laboratories in London

Patient completes first 3 pages of *Acute Gastroenteritis Questionnaire* and returns it to Public Health England

END

Yes

Local laboratory splits stool sample and tests half of the sample using traditional methods and half using Luminex xTAG® GPP

Pathogen isolates sent to Reference Laboratories in London AND pathogen isolates and DNA extracts sent for typing and/or metagenomics

Patient completes ALL pages of the *Acute Gastroenteritis Questionnaire* and returns it to Public Health England

Patient receives *Patient Experience Survey* questionnaire 2 weeks after submitting sample, completes it and returns it to the University of Liverpool Implementation team

If interested, and invited to do so, patient takes part in face-to-face interview with the University of Liverpool Process Evaluation team

END

Key

routine clinical/public health practice

research

Figure 1: Patient recruitment flow diagram and study processes for the INTEGRATE project. ASAP: as soon as possible; xTAG GPP: Luminex xTAG Gastrointestinal Pathogen Panel. Adapted from McIntyre *et al*. 2019[3]

Results

- Significant positive correlations were seen between positives from traditional tests for Adenovirus, Astrovirus, *Cryptosporidium*, *Salmonella*, and Sapovirus, and the presence of these taxa in sequencing reads. Significant correlations were not seen for other taxa.
- Significant positive correlations were also seen between positive Luminex results and reads for Adenovirus, *Campylobacter*, *Cryptosporidium*, Rotavirus, Sapovirus and *Shigella*.
- *Entamoeba* reads were absent across all samples, including samples where *E. histolytica* positive results were observed (14 samples by Luminex, 0 by traditional methodologies).
- No cross-taxa associations between reads and lab diagnostics were more significant than the relationship between a specific taxa, and it's corresponding laboratory diagnostic.

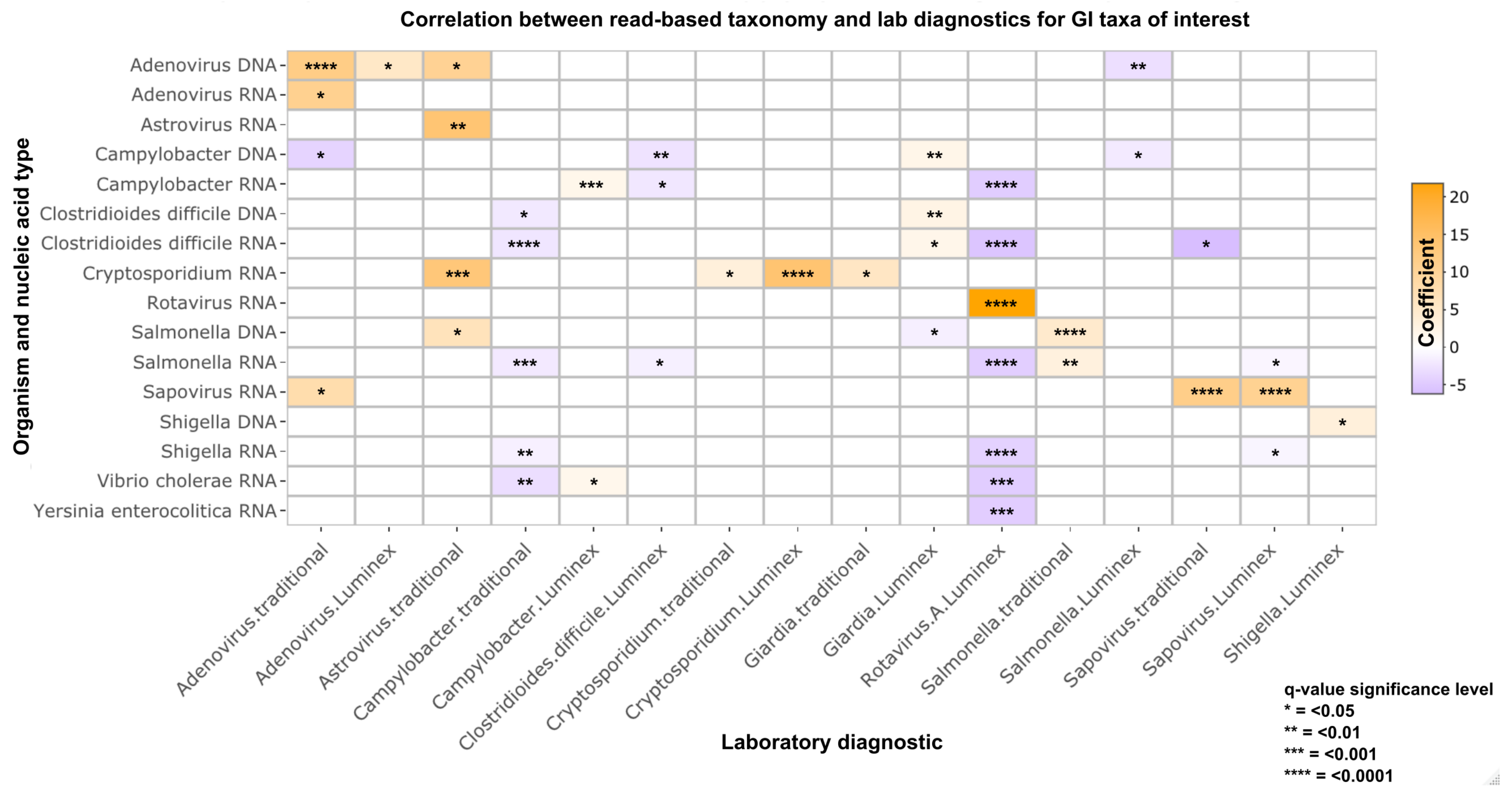


Figure 2: Heatmap of significant associations observed between read-based taxonomy and lab diagnostics for gastrointestinal pathogens of interest. The heatmap only shows read-based results or laboratory diagnostics where at least one significant correlation (q < 0.05) was observed.

Conclusions and future directions

- Statistically significant overlaps between k-mer based taxonomy and laboratory diagnostics are seen for 8/14 taxa of interest (excluding *E. coli*).
- Cross-taxa associations, such as negative associations between *C. difficile* reads and *Campylobacter* laboratory diagnostics (and vice versa) should be investigated in more detail – is this artefactual of the methodology, or an indicator of taxa dominance?
- Wider investigation into the prevalence of novel pathogens in the dataset.
- Incorporation of Adenovirus 40/41 and *E. coli* into the analysis – sequence-based gene and genomic methodologies to highlight prevalence of strains and serotypes. For *E. coli*, genes associated with phenotypes of interest (enterotoxicity[4] and enteroaggregation[5]) will be investigated.
- Absence of *Entamoeba* reads in *E. histolytica* positive samples– is this indicative of issues with Luminex, or k-mer based methodologies? Previous studies have shown poor positive predictive values for the Luminex panel when used to screen for *E. histolytica*[6].
- Are the taxa targeted by laboratory diagnostics the most important in this dataset?

Integrate

References: 1 Tam CC, Rodrigues LC, *et al.* ; IID2 Study Executive Committee. Longitudinal study of infectious intestinal disease in the UK (IID2 study): incidence in the community and presenting to general practice. *Gut*. 2012;61(1):69-77. **2** Amjad M. An Overview of the Molecular Methods in the Diagnosis of Gastrointestinal Infectious Diseases. *Int J Microbiol*. 2020 24;2020:8135724. **3** McIntyre KM, Bolton FJ, *et al.* A Fully Integrated Real-Time Detection, Diagnosis, and Control of Community Diarrheal Disease Clusters and Outbreaks (the INTEGRATE Project): Protocol for an Enhanced Surveillance System. *JMIR Res Protoc*. 2019 Sep 26;8(9):e13941. **4** Chung, S. Y., Kwon, T., *et al.* (2019). Comparative genomic analysis of enterotoxigenic *Escherichia coli* O159 strains isolated from diarrheal patients in Korea. *Gut pathogens*, 11(9). **5** Boisen N, Østerlund MT, *et al.* Redefining enteroaggregative Escherichia coli (EAEC): Genomic characterization of epidemiological EAEC strains. *PLoS Negl Trop Dis.* 2020 ;14(9):e0008613. **6** Navidad, J. F., Griswold, D. J *et al.* (2013). Evaluation of Luminex xTAG gastrointestinal pathogen analyte-specific reagents for high-throughput, simultaneous detection of bacteria, viruses, and parasites of clinical and public health importance. *Journal of clinical microbiology*, *51*(9), 3018–3024.

UNIVERSITY OF LIVERPOOL | WARWICK THE UNIVERSITY OF WARWICK | UK Health Security Agency | FUNDED BY NIHR National Institute for Health Research

The authors acknowledge the Department of Health and Social Care and the Wellcome Trust for the funding received for the initial project through the Health Innovation Challenge Fund (grant reference: HICF-T5-354). The research was funded by the National Institute for Health Research Protection Research Unit (NIHR HPRU) in Gastrointestinal Infections at University of Liverpool in partnership with the UK Health Security Agency (UKHSA), in collaboration with University of Warwick. The views expressed are those of the author(s) and not necessarily the NIHR, the Department of Health and Social Care or the UK Health Security Agency.