Metatranscriptomics improves the laboratory diagnosis of infectious intestinal disease from human diarrhoeal samples

Edward Cunningham-Oakes[1,2], Yan Li[3], Blanca Perez-Sepulveda[3], Jay C. D. Hinton[3], Charlotte A. Nelson[4], K. Marie McIntyre[5], Maya Wardeh[6,7,8], Miren Iturriza-Gómara[3,9], Christiane Hertz-Fowler[4], Sarah J. O'Brien[5], Nigel A. Cunliffe[2,3], Alistair C. Darby[1,2,4*]; on behalf of the INTEGRATE consortium

[1]Department of Infection Biology and Microbiomes, Institute of Infection, Veterinary and Ecological Sciences, University of Liverpool, Liverpool, UK.
[2]NIHR Health Protection Research Unit in Gastrointestinal Infections, Liverpool, UK.
[3]Department of Clinical Infection, Microbiology and Immunology, Institute of Infection, Veterinary and Ecological Sciences, University of Liverpool, Liverpool, UK.
[4]Centre for Genomic Research, Institute of Integrative Biology, University of Liverpool, Liverpool, UK.
[5]School of Natural and Environmental Sciences, Newcastle University, Newcastle, UK.
[6]Department of Computer Science, University of Liverpool, Liverpool, UK.
[7]Department of Livestock and One Health, Institute of Infection, Veterinary and Ecological Sciences, University of Liverpool, Liverpool, UK.
[8]Department of Mathematics, University of Liverpool, Liverpool, UK.
[9]Centre for Vaccine Innovation and Access, Program for Appropriate Technology in Health (PATH), Geneva, 1218, Switzerland.

*** Correspondence to:**
Alistair.Darby@liverpool.ac.uk

ABSTRACT
Background
Traditional laboratory-based surveillance of gastrointestinal pathogens is time-consuming, which can impact successful outbreak detection. The INTEGRATE study investigated the utility of metagenomic and metatranscriptomic sequencing for the rapid diagnosis of community-associated gastrointestinal infections.

Methods
We performed an observational study using stool samples from 1,407 patients with acute gastroenteritis, recruited via general practitioners in the UK. 1,067 stool samples were processed using i) routine clinical methods, ii) a molecular multiplex real-time polymerase chain reaction (PCR) assay, and iii) DNA and RNA sequencing. The relationship between assigned taxonomy, routine clinical diagnostics, and PCR was determined with multivariable linear regression models.

Findings
There is a strong, positive relationship between the identification of pathogens in metatranscriptomic reads, and positive results from traditional diagnostics for five out of fifteen pathogens: *Campylobacter* ($p<0{\cdot}001$), *Cryptosporidium* ($p<0{\cdot}001$), *Salmonella* ($p<0{\cdot}01$), *Rotavirus* ($p<0{\cdot}001$) and *Sapovirus* ($p<0{\cdot}001$). Metagenomic sequencing displayed this relationship for two out of fifteen pathogens: *Campylobacter* ($p<0{\cdot}001$) and *Salmonella* ($p<0{\cdot}01$).

Strong positive relationships between metatranscriptomic reads and positive PCR results were observed for six out of fourteen pathogens: Adenovirus, *Campylobacter*, *Cryptosporidium*, *Norovirus*, *Rotavirus* and *Sapovirus* ($p<0{\cdot}001$). In metagenomic data, the same relationship was observed for four out of fourteen pathogens: Adenovirus ($p<0{\cdot}001$), *Campylobacter* ($p<0{\cdot}001$), *Salmonella* ($p<0{\cdot}05$) and *Shigella* ($p<0{\cdot}01$).

A comprehensive transcriptomic profile of *Salmonella* Enteritidis was recovered from the stool of a patient with a subsequently confirmed *Salmonella* infection.

Interpretation
The metatranscriptomic strategy successfully detected a variety of gastrointestinal pathogens and provided viral and bacterial gene expression profiling directly from stool. We propose that metatranscriptomics could be considered for the future surveillance of gastrointestinal pathogens.

Funding
This publication presents independent research supported by the Health Innovation Challenge Fund (WT096200, HICF-T5-354), a parallel funding partnership between the Department of Health and Wellcome Trust. The views expressed in this publication are those of the author(s) and not necessarily those of the Department of Health or Wellcome Trust.

INTRODUCTION

The incidence of infectious intestinal disease (or acute gastroenteritis) is estimated to be 18 million cases each year in the United Kingdom (UK)[1]. About 25% of infected people experience diarrhoeal and related gastrointestinal symptoms. The current mainstay for identifying gastrointestinal pathogens in faecal specimens in the UK are conventional laboratory techniques including microscopy, culture and antigen detection; and, increasingly, molecular assays, such as nucleic acid amplification tests[2].

Although conventional and polymerase chain-reaction (PCR)-based approaches are validated for clinical laboratory use[2], both provide limited information about pathogens (focussing on a single gene or set of characteristics)[3]. In the case of bacterial culture, growth time, lack of sensitivity, and the challenge of culturing fastidious organisms contribute to diagnostic delays[3]. Current methods lack the sensitivity required to detect pathogens that are present intermittently in low numbers[4]. In contrast, PCR, which uses target sequences for organism detection, is more sensitive[3], not dependent on high viral or bacterial loads and is culture-independent. However, whilst molecular methods are more sensitive than culture-based methods they lack resolution[5] and are unable to achieve the strain-level discrimination required for outbreak monitoring[6]. Inevitably, molecular assays are limited to known sequences[5], meaning that unexpected pathogens will be missed, with associated functional insights. The adoption of whole-genome sequencing is intended to overcome some limitations of current methods but requires the isolation of a pure culture of pathogenic organisms.

The INTEGRATE study[7] compared traditional diagnostic methods with modern, sensitive molecular and genomic microbiological methods for identifying and characterising responsible pathogens[7]. Here, we present data generated by next-generation sequencing of the stool of 1,060 symptomatic patients. We provide evidence that both metagenomic (DNA) and metatranscriptomic (RNA) sequencing directly from stool detects the major community-associated GI pathogens of relevance to the UK setting.

METHODOLOGY

Patient recruitment and sample collection

Recruitment and sample collection was described previously[7]. Briefly, stool was collected from members of the public with symptoms of acute gastroenteritis via practices in the Royal College of General Practitioners Research and Surveillance Centre National Monitoring Network (RCGP RSC NMN). Patients meeting inclusion criteria were invited to submit a stool sample for microbiological analysis. Consent was obtained for this procedure, as stool sampling is usually only performed if a case is severe, or persistent. Patients who provided a stool sample were then recruited into the study.

Sample processing

Faecal samples were received by one of three clinical laboratories (Royal Liverpool and Broadgreen University Hospitals NHS Trust, Central Manchester University Hospitals NHS Foundation Trust, or Lancashire Teaching Hospitals NHS Foundation Trust), and divided into two aliquots. One part of the sample was processed according to routine clinical practice at each laboratory; the other was processed using a molecular multiplex real-time polymerase chain reaction (PCR) assay [Luminex xTAG Gastrointestinal Pathogen Panel, Luminex, I032C0324], supplemented with targets for Enteroaggregative *Escherichia coli* and *Sapovirus.* Extractions of nucleic acid from faeces were performed using QIASymphany and EasyMag automated nucleic acid extraction platforms. Further details can be found in the primary study protocol[7]. Samples that returned a positive result according to routine clinical practice were designated as "clinical positives". Those that returned a positive result by PCR were designed as "molecular positives".

Metagenomic and metatranscriptomic sequencing

Illumina fragment libraries from DNA were prepared using NEBNext DNA Ultra kits. For the generation of dual-indexed, strand-specific RNASeq libraries, RiboZero rRNA depleted RNA samples were prepared using NEBNext Ultra Directional RNA kits. For all libraries, paired-end, 150-bp sequencing was then performed on an Illumina HiSeq 4000, generating data from >280 million clusters per lane.

Quality control for second-generation sequencing reads

Modules from the MetaWRAP[8] (v1·3·2) pipeline were used to standardise metagenome analysis. The pipeline was deployed in a dedicated conda environment, using the "manual installation" guide (see https://github.com/bxlab/metaWRAP). All paired-end reads underwent quality-control using the MetaWRAP "read_qc" module to remove low-quality, adapter, and human sequence reads.

Assigning taxonomy to genomic DNA and RNA reads

DNA and RNA reads were used for taxonomic assignments with Kraken2[9] (v2·1·2), using a custom database, which included all RefSeq complete genomes and proteins for archaea, bacteria, fungi, viruses, plants, protozoa, as well as all complete RefSeq plasmid nucleotide and protein sequences, and a false-positive minimized version of the NCBI UniVec database (see the Kraken2 GitHub page for more information on the construction of custom databases available for Kraken2 https://github.com/DerrickWood/kraken2/blob/master/docs/MANUAL.markdown). A confidence threshold of 0·1 was set for read assignments, and reports were generated for downstream biom file generation. For DNA reads, read counts assigned to taxonomies in each sample were then re-estimated using the average read length of that sample, using Bracken[10] (v2·0). Kraken-biom (v1·0·1) was then used to generate biom file in json format, using initial Kraken reports for RNA samples, and Bracken reports for DNA samples. Biom (v2·1·6) was then used to assign tabulated metadata to this biom file.

Visualisation and comparison of taxa of interest in RNA and DNA

A taxonomy table was generated from the biom file in R (v4·2·2) using Phyloseq (v1·42·0). Read-assigned taxonomy counts were parsed from this table for any samples with both metagenomic (DNA)

and metatranscriptomic data (n=985). Counts were extracted for the following taxa: *Adenoviridae*, *Campylobacter*, *Clostridioides difficile* (*C. difficile*), *Cryptosporidium*, *Escherichia coli* (*E. coli*), *Norovirus*, *Rotavirus*, *Salmonella*, *Shigella*, *Sapovirus*, *Vibrio cholerae (V. cholerae*) and *Yersinia enterocolitica* (*Y. enterocolitica*). These taxa were chosen to reflect the PCR panels used during this study. RNA virus (*Astroviridae*, *Norovirus*, *Rotavirus* and *Sapovirus*) read counts were not extracted for this part of the analysis due to an absence of DNA reads. DNA and RNA counts were log-transformed and plotted against one another as a line graph using ggplot2 (v3·4·0). Visualisations were then used to assess the sensitivity of metagenomics and metatranscriptomics for the selected taxa, where we define sensitivity as the skew of datapoints towards either metagenomics (x-axis) or metatranscriptomics (y-axis). A dashed line was included in each line graph to assist in illustrating sensitivity differences.

Regression models are used to model the relationship between dependent and independent variables of a dataset a simplified mathematical form[13]. The strength of the relationship between dependent and independent variables is described differently, based on the type of linear regression model used. For the simplest form of linear regression model (univariable), an R^2 value is generated to describe this relationship. An R^2 value (ranging from 0-1) is a measure of how well the regression model describes the observed data, with a lower R^2 values indicating that a given dataset does align (or fit) with the regression model. Additionally, statistical significance (p-values) can be assigned to the model, which the highlight the reliability of the model itself[14]. For each taxa, the strength of correlation between metagenomic and metatranscriptomic data was quantified (R^2 value) and statistically assessed (p-value) using a univariable linear regression model for DNA against RNA. The assumption for this univariable regression model was that the type of (metagenomic or metatranscriptomic) sequencing used would account for the differences in reads assigned to GI pathogens.

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

More complex (multivariable) models, allow us to account for additional, clinically-relevant variables[15], making our models even more reliable. However, in these models, deriving an R^2 value can be difficult. As such, certain multivariable models used for microbiome analysis[11] instead use an arbitrary coefficient value as an alternative to the R^2 value. The statistical significance of associations between read counts and laboratory results for organisms of interest were assessed using a multivariable linear regression model in MaAsLin 2[11](v1·6·0), under default settings. The introduction of another variable into the model (laboratory results) provided a point of reference. This allowed us to determine the relationship between any sample with sequencing data, and laboratory results. As such, for this analysis, all sequenced patient samples (n=1067) were used, even if they did not contain both metagenomic and metatranscriptomic data. Our approach allowed RNA virus read counts from metatranscriptomic data to be included in this analysis. To visualise the strength of correlations between laboratory results, and pathogen-assigned sequencing reads, correlation coefficients and adjusted p-values from the model were tabulated and used to generate a heatmap with corrplot (v0·9·2). These arbitrary coefficients are generated via MaAsLin 2, which has comprehensively evaluated against methods of generating coefficients[11]. Adjusted p-values were generated using the Benjamini-Hochberg Procedure.

Visualisation of a *Salmonella* transcriptome directly from stool

Metatranscriptomic reads from a sample of a patient with a later-confirmed (culture positive) *Salmonella* spp. infection underwent quality control, alignment, and quantification using the Bacpipe RNA-seq processing pipeline (v0·6·0). The GFF annotation[12] for the *Salmonella enterica* subsp. enterica serovar Enteritidis PT4 strain P125109 (Accession: GCA_015240635.1) was used in this analysis. Coverage tracks and annotation were visualized using JBrowse (v1·16·8). This visualisation can be found here: https://s.hintonlab.com/study_74.

RESULTS

Metagenomics and metatranscriptomics show different sensitivities for GI pathogens, but all pathogens are detectable

A total of 1,067 samples were sequenced, with 985 providing both metagenomic and metatranscriptomic data. For *Campylobacter*, *Cryptosporidium*, and *Giardia* (Figure 1), metatranscriptomics showed a greater sensitivity (see Methods for definition of sensitivity) than metagenomics. In contrast, metagenomics displayed a greater sensitivity than metatranscriptomics for the *Adenoviridae*, *C. difficile*, pathogenic *E. coli, Salmonella, Shigella* and *Y. enterocolitica* (Figure 1).

The assumption for this univariable regression model was that the type of (metagenomic or metatranscriptomic) sequencing used would account for the differences in reads assigned to GI pathogens. Regression models reached significance ($p<0{\cdot}05$), for several bacterial pathogens of interest (*Campylobacter*, *C. difficile*, *E. coli*, *Salmonella*, *Shigella* and *Y. enterocolitica*; see Table 1). However, R^2 values indicated that the regression model was a poor fit for these data. In this model, 8.0-37·8% of the difference in assigned read counts was explained by the type of sequencing used. Within these results, *Salmonella* and *Shigella* performed particularly poorly, with R^2 values of 0·0798 (8.0%) and 0.113 (11·3%) respectively. *V. cholerae* did not reach significance within the model ($p = 0{\cdot}29$), and was a poor fit for the model (R^2 value = 0·0152). This is consistent with laboratory results, where no cases of *V. cholerae* were confirmed by either routine clinical methods, or by molecular diagnostics.

In contrast, for parasites (*Cryptosporidium* and *Giardia*), the confidence that these data fit the regression model was high (61.1–76.7%). Moreover, there was greater sensitivity for the detection of parasites in metatranscriptomic data.

Adenoviruses (the family *Adenoviridae*) are DNA viruses that can be pathogenic, or commensal. As such, it was expected that Adenovirus could be identified in the metagenomic reads in most samples. Accordingly, 11 out of 12 (91·7% of) samples identified as Adenovirus-positive in the laboratory, for which metagenomic data could be generated, possessed >1000 Adenovirus-assigned reads. However, 77 out of 988 (7·8%) Adenovirus-negative samples also possessed >1000 Adenovirus-assigned reads. Adenovirus was also detected in metatranscriptomic data from Adenovirus-positive stool samples. For samples where metatranscriptomics could be performed, 4 of 10 (40·0% of) samples possessed >1000 Adenovirus-assigned reads. In contrast, only 9 of 1024 (0·9% of) Adenovirus-negative samples with metatranscriptomic data possessed >1000 Adenovirus-assigned reads. The correlation between DNA and RNA for *Adenoviridae* was positive, but weaker than for other pathogens ($R^2=0{\cdot}212$).

A positive correlation was seen between the taxonomic assignments to key GI pathogens in metagenomic and metatranscriptomic reads (Figure 1).

Taxa	Adjusted R^2	Residual standard error	Significance of univariable regression model (p-value)
Adenoviridae	**0·212**	**2·147**	**$<2{\cdot}2\times10^{-16}$**
Campylobacter	**0·378**	**1·128**	**$<2{\cdot}2\times10^{-16}$**
Cryptosporidium	**0·611**	**1·124**	**$1{\cdot}7\times10^{-5}$**
C. difficile	**0·235**	**1·193**	**$<2{\cdot}2\times10^{-16}$**
E. coli	**0·343**	**1·635**	**$<2{\cdot}2\times10^{-16}$**
Giardia	**0·767**	**1·259**	**$2{\cdot}7\times10^{-8}$**
Salmonella	**0·080**	**1·712**	**$3{\cdot}3\times10^{-13}$**
Shigella	**0·113**	**1·771**	**$2{\cdot}5\times10^{-16}$**
V. cholerae	**0·015**	**1·930**	**0·29**
Y. enterocolitica	**0·474**	**1·000**	**$7{\cdot}4\times10^{-8}$**

Table 1: Correlations between DNA and RNA reads were significant for 9 of 10 major GI community pathogens in the United Kingdom.

Columns are: i) Taxa – taxonomy assigned to reads, ii) Adjusted R^2 – how well the regression model describes the relationship between DNA and RNA read counts in this dataset, iii) Residual standard error – standard deviation of around the regression line, iv) Significance of linear regression model (p-value) – our confidence in any results generated by the univariable regression model (including R^2 values).

The detection of GI pathogens in metagenomic and metatranscriptomic data is comparable to clinical laboratory results

Our analysis showed that the pathogens detected in sequencing reads match closely with results generated by laboratory diagnostics for *Adenoviridae*, *Campylobacter*, *Cryptosporidium*, *Norovirus*, *Rotavirus*, *Salmonella*, *Sapovirus* and *Shigella* (Figure 2). Moreover, major GI community pathogens in the United Kingdom could be detected in metagenomic and metatranscriptomic data, but RNA viruses could only be detected in metatranscriptomic data.

Viral pathogens

It is possible to detect RNA viruses (*Norovirus*, *Sapovirus* and *Rotavirus*) in metagenomic data. However, DNA viruses, such as the *Adenoviridae*, can be detected in both metagenomic and metatranscriptomic data. For the *Adenoviridae*, positive correlations were observed between detection in metagenomic reads, metatranscriptomic reads and Luminex results ($p<0{\cdot}001$). Positive correlations between *Rotavirus* and all laboratory tests ($p<0{\cdot}001$) were observed in metatranscriptomic data. The detection of *Norovirus* and *Sapovirus* in metatranscriptomic data displayed significant correlation ($p<0{\cdot}001$) with Luminex results. No correlations between sequencing data and corresponding laboratory results were observed for Astrovirus (*Astroviridae*).

Parasites

It was possible to detect parasites in both metagenomic and metatranscriptomic data. However, metatranscriptomics displayed a much higher sensitivity for the detection of parasites than metagenomics. Positive correlations between the detection of *Cryptosporidium* in metatranscriptomic data and laboratory data were highly significant ($p<0{\cdot}001$). This finding is consistent with our initial sensitivity (Figure 1) and univariable regression model results (see Table 1). No associations were observed between metagenomic data and laboratory data for *Cryptosporidium*. No associations were seen between sequencing data and corresponding laboratory results for *Giardia*.

Bacterial pathogens

The identification of bacterial pathogens from sequencing data is more difficult, as commensal organisms and pathogens can have extremely similar taxonomy and genomic profiles. As such, laboratory diagnostics differentiate commensal and pathogenic organisms using pathogenic genes, or phenotypes. Our results show that metagenomics and metatranscriptomics can both be used to identify bacterial pathogens with differing sensitivities. For *Campylobacter*, positive correlations were observed between sequencing and all laboratory results ($p<0{\cdot}001$). *Salmonella* displayed positive correlations between sequencing data and traditional diagnostics ($p<0{\cdot}01$) for both metagenomics and metatranscriptomics, and between metagenomic sequencing data and Luminex results ($p<0{\cdot}05$). Correlations between sequencing data and corresponding laboratory results were not significant for *C. difficile*, and *Y. enterocolitica*.

A summary of all correlations between the detection of GI pathogens in sequencing reads, and laboratory data is provided in Supplementary File 1.

A full transcriptomic profile can be generated for *Salmonella*

Metatranscriptomic analysis of stool from a patient with a laboratory-confirmed *Salmonella* infection yielded functional insights that surpassed traditional and PCR diagnostics. A high-quality gene expression transcriptomic profile was generated from 12·7 million sequence reads that mapped to the *S.* Enteritidis P125109 genome. The *S.* Enteritidis transcripts that underpin this novel gene expression data can be visualised and interrogated in a bespoke genome browser (https://s.hintonlab.com/study_74). A variety of environmentally responsive *Salmonella* genes were highly expressed, likely reflecting physicochemical stresses the bacteria had been exposed to in the stool sample. Examples include *ahpC* (oxidative stress), *hmpA* (nitrosative stress), *phoH* (phosphate starvation), *pspA* (extracytoplasmic stress), and the *rpoE* and *rpoS* transcription factor genes. A number of these genes have been previously implicated in the virulence and stress responses of *Salmonella*. The unexpected discovery that the metatranscriptomic analysis of a human stool sample can generate a comprehensive gene expression profile of a *Salmonella* pathogen is worthy of future exploitation.

Perspective

We have demonstrated that the agnostic detection of important UK GI pathogens from human stool can be achieved with metagenomic and metatranscriptomic approaches. The primary remit for the impact of this work is within GI pathogen diagnostics. This work, and these data demonstrate the potential for i) improvements within the scope of current GI pathogen diagnostics or ii) bridging gaps not addressed by current diagnostics.

Improvements within the scope of current diagnostics

Sequencing directly from stool could minimizes the time required for pathogen detection, meaning that more laborious detection methods, such as cultivation, can be appropriately tailored to confirm the presence of the suspected pathogen.

The metatranscriptomic strategy displays increased sensitivity for *Campylobacter*, *Cryptosporidium* and *Giardia*, whilst metagenomics displays increased sensitivity for other GI pathogens including Adenovirus, *C. difficile* pathogenic *E. coli*, *Salmonella*, *Shigella* and *Y. enterocolitica*. However, RNA viruses cannot be detected by metagenomic sequencing, as demonstrated by our comparison of sequencing data to laboratory-based pathogen testing (Figure 2). Therefore, direct extraction of RNA from stool represents a single sample format and cultivation-independent process for the detection of a broad range of GI pathogens, including unexpected aetiological agents. This finding is supported by previous clinical studies, which used metatranscriptomics to simultaneously measure the virome, microbiome, and host response[16], as well as to characterise disease-related microbiomes with increased sensitivity[17]. These benefits are not offered by current routine clinical diagnostic methods.

Additionally, the observation that pathogenic Adenovirus can be detected in metatranscriptomic data with a low rate of false-positive identification (0·9%) in comparison to metagenomic data (7·8%) is important. Less Adenovirus was identified in metatranscriptomic data than metagenomic data; however, an eight-fold reduction in false-positives demonstrates the potential for metatranscriptomics to distinguish between commensals and pathogens of concern. Moreover, we can be equally confident in the comparability of metagenomic and metatranscriptomic data to laboratory diagnostics for Adenovirus (see Figure 2).

These data also demonstrate increased sensitivity for the detection of parasites of concern in GI infections. Our visualisations of metagenomic and metatranscriptomic reads (Figure 1) showed that metatranscriptomic data was more sensitive for *Cryptosporidium* and *Giardia* (parasites). We also observed significant regression (R^2 = 61·1-76·7%) between differences in read counts for parasites across metagenomic and metatranscriptomic data. Finally, our multivariable model demonstrated the

strong correlation and high significance of the detection of *Cryptosporidium* in the laboratory, and in metatranscriptomic data. Improved sensitivity for the detection of parasites with metatranscriptomics has been reported previously, in a study that detected 23% more blood infections than traditional methods[18].

Additional work is needed to adapt our approach for the accurate identification of *E. coli* pathovariants from sequencing data. *Shigella* and *E. coli* pathovariants are extremely similar on a genome-wide (and taxonomic) level[19], and are distinguished using specific gene-based assays[20]. In contrast, our study drew correlations between pathogens in reads and laboratory tests based on taxonomy. Due to this, and the ubiquitous presence of *E. coli* in all samples, it was not possible to draw associations between the presence of *E. coli*, and gene-based assays used for *E. coli* pathovariant identification (see Supplementary Table 1, Supplementary File 2). This limitation may explain the less significant relationships between *Shigella* sequencing reads and laboratory tests (Figure 2), due to the misidentification of *Shigella* as *E. coli* (and vice versa) in these data.

Overall, our findings imply that using RNA as a diagnostic target is useful for the detection of low-level pathogens and reduces false-positive signals from commensals. Our approach could be used to modify the allocation of resources for reference laboratory diagnostics.

Bridging gaps not addressed by current diagnostics

These metatranscriptomic data demonstrate areas of clinical relevance that are not fulfilled by routine clinical diagnostics. Firstly, metagenomic and metatranscriptomic data permits the identification of multiple species and strains within a sample (see Supplementary Figure 1, Supplementary File 2), including novel pathogens. This type of analysis is beyond the scope of our study, but has been used to successfully identify novel pathogens from the stools of various species[21,22]. Additionally, these data also demonstrate the ability to rapidly generate gene expression profiles for pathogens of concern, without prior enrichment. Finally, these data demonstrate the ability to successfully recover metatranscriptomic data from a human diarrhoeal sample. Future studies could generate a true disease-state expression profile with appropriate methodology. From a clinical perspective, if performed on a case-by-case basis, this opens up the potential to monitor the effects of interventions[23] and to accurately determine host-pathogen interactions during genuine human infections[16].

We conclude that metatranscriptomic sequencing direct-from-stool could minimise the time required to generate a taxonomic profile from patient samples. To our knowledge, this is the largest retrospective observational study that used metatranscriptomic analysis to investigate samples from patients with acute gastrointestinal symptoms. Moreover, this is the first study to demonstrate and quantify the benefits that metatranscriptomics could bring to gastrointestinal surveillance in the United Kingdom via direct comparison to validated diagnostics for all major community pathogens. The study provides a foundation for the implementation of sequencing-based diagnostics to the clinic, with the potential of capturing a wider breadth of organisms than current diagnostics and identifying novel pathogens.

Acknowledgements

This publication presents independent research supported by the Health Innovation Challenge Fund (WT096200, HICF-T5-354), a parallel funding partnership between the Department of Health and Wellcome Trust. The views expressed in this publication are those of the author(s) and not necessarily those of the Department of Health or Wellcome Trust. E.C.-O., N.A.C. and A.C.D. are affiliated to the National Institute for Health Research (NIHR) Health Protection Research Unit in Gastrointestinal Infections at University of Liverpool, in partnership with the UK Health Security Agency (UKHSA), in collaboration with University of Warwick. E.C.-O., N.A.C. and A.C.D. are based at The University of

Liverpool. The views expressed are those of the author(s) and not necessarily those of the NIHR, the Department of Health and Social Care or the UK Health Security Agency. N.A.C. is a NIHR Senior Investigator (NIHR203756).

The authors thank the members of the INTEGRATE Consortium. The INTEGRATE Consortium investigators in the United Kingdom are Sarah J O'Brien (principal investigator), Frederick J Bolton, Rob M Christley, Helen E Clough, Nigel A Cunliffe, Susan Dawson, Elizabeth Deja, Ann E Durie, Sam Haldenby, Neil Hall, Christiane Hertz-Fowler, Debbie Howarth, Lirije Hyseni, Miren Iturriza-Gomara, Kathryn Jackson, Lucy Jones, Trevor Jones, K Marie McIntyre, Charlotte A Nelson, Lois Orton, Jane A Pulman, Alan D Radford, Danielle Reaves, Helen K Ruddock, Darlene A Snape, Debbi Stanistreet, Tamara Thiele, Maya Wardeh, David Williams, and Craig Winstanley (University of Liverpool), Kate Dodd (NIHR Clinical Research Network: North West Coast), Peter J Diggle, Alison C Hale, Barry S Rowlingson (Lancaster University), Jim Anson, Caroline E Corless, Viki Owen (Royal Liverpool and Broadgreen University Hospitals NHS Trust), Malcolm Bennett (University of Nottingham), Lorraine Bolton, John Cheesbrough, Katherine Gray, David Orr, Lorna Wilson (Lancashire Teaching Hospitals NHS Foundation Trust), Andrew R Dodgson, Ashley McEwan (Manchester University NHS Foundation Trust), Paul Cleary, Alex J Elliot, Ken H Lamden, Lorraine Lighton, Catherine M McCann, Matthieu Pegorie, Nicola Schinaia, Anjila Shah, Gillian E Smith, Roberto Vivancos, Bernard Wood (PHE), Rikesh Bhatt, Dyfrig A Hughes (Bangor University), Rob Davies (APHA); Simon de Lusignan, Filipa Ferreira, Mariya Hriskova, Sam O'Sullivan, Stacy Shinneman and Ivelina Yonova (University of Surrey/Royal College of General Practitioners).

Author contributions

We describe author contributions to the paper using the CRedit taxonomy. Conceptualization: E.C.-O. and A.C.D. Data curation: E.C.-O., B.P-S., K.M.M., C.A.N. and M.W. Formal analysis: E.C.-O. and Y.L. Funding acquisition: S.J.O'B. and N.A.C. Investigation: E.C.-O. and A.C.D. Project administration: E.C.-O., M.I.G., C.H.-F., S.J.O'B., N.A.C. and A.C.D. Resources: Y.L., B.P-S., J.C.D.H., C.A.N., C.H.-F. and A.C.D. Supervision: N.A.C. and A.C.D. Validation: E.C.-O. Visualization: E.C.-O. and Y.L. Writing – original draft: E.C.-O., J.C.D.H., N.A.C. and A.C.D. Writing – review and editing: E.C.-O., Y.L., B.P-S., J.C.D.H., K.M.M., S.J.O'B., N.A.C. and A.C.D.

Competing Interests

M.I.G. has received research grants from GSK and Merck, and has provided expert advice to GSK. M.I.G. has been an employee of GSK since January 2023, although the work presented here was completed prior to this date.

References

1. Foodborne Disease Estimates for the United Kingdom in 2018. (2020).
2. Public Health England, NHS. UK Standards for Microbiology Investigations - Gastroenteritis. (2020).
3. Foddai, A. C. G. & Grant, I. R. Methods for detection of viable foodborne pathogens: current state-of-art and future prospects. *Appl. Microbiol. Biotechnol.* **104**, 4281–4288 (2020).
4. Kanagarajah, S. *et al.* Whole genome sequencing reveals an outbreak of *Salmonella* Enteritidis associated with reptile feeder mice in the United Kingdom, 2012-2015. *Food Microbiol.* **71**, 32–38 (2018).
5. Chiu, C. Y. & Miller, S. A. Clinical metagenomics. *Nat. Rev. Genet.* **20**, 341–355 (2019).
6. Buytaers, F. E. *et al.* Application of a strain-level shotgun metagenomics approach on food samples: resolution of the source of a *Salmonella* food-borne outbreak. *Microbial Genomics,* vol. 7 (2021).
7. McIntyre, K. M. *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.* **8**, e13941 (2019).
8. Uritskiy, G. V., DiRuggiero, J. & Taylor, J. MetaWRAP—a flexible pipeline for genome-resolved metagenomic data analysis. *Microbiome* **6**, 158 (2018).
9. Wood, D. E., Lu, J. & Langmead, B. Improved metagenomic analysis with Kraken 2. *Genome Biol.* **20**, 257 (2019).
10. Lu, J., Breitwieser, F., Thielen, P. & Salzberg, S. Bracken: Estimating species abundance in metagenomics data. *PeerJ Comput. Sci.* **3**, e104 (2017).
11. Mallick, H. *et al.* Multivariable association discovery in population-scale meta-omics studies. *PLOS Comput. Biol.* **17**, e1009442 (2021).
12. Perez-Sepulveda, B. M. *et al.* Complete Genome Sequences of African *Salmonella enterica* Serovar Enteritidis Clinical Isolates Associated with Bloodstream Infection. *Microbiol. Resour. Announc.* **10**, (2021).
13. Schneider, A., Hommel, G. & Blettner, M. Linear regression analysis: part 14 of a series on evaluation of scientific publications. *Dtsch. Arzteblatt Int.* **107**, 776–782 (2010).
14. Ali, P. & Younas, A. Understanding and interpreting regression analysis. *Evid. Based Nurs.* **24**, 116 (2021).
15. McDonald, J. H. *Handbook of biological statistics*. vol. 2 (sparky house publishing Baltimore, MD, 2009).
16. Rajagopala, S. V. *et al.* Metatranscriptomics to characterize respiratory virome, microbiome, and host response directly from clinical samples. *Cell Rep. Methods* **1**, 100091 (2021).
17. Feng, Y. *et al.* Metagenomic and metatranscriptomic analysis of human prostate microbiota from patients with prostate cancer. *BMC Genomics* **20**, 146 (2019).
18. Galen, S. C., Borner, J., Williamson, J. L., Witt, C. C. & Perkins, S. L. Metatranscriptomics yields new genomic resources and sensitive detection of infections for diverse blood parasites. *Mol. Ecol. Resour.* **20**, 14–28 (2020).
19. Parks, D. H., Chuvochina, M., Reeves, P. R., Beatson, S. A. & Hugenholtz, P. Reclassification of *Shigella* species as later heterotypic synonyms of *Escherichia coli* in the Genome Taxonomy Database. *bioRxiv* 2021.09.22.461432 (2021) doi:10.1101/2021.09.22.461432.
20. Devanga Ragupathi, N. K., Muthuirulandi Sethuvel, D. P., Inbanathan, F. Y. & Veeraraghavan, B. Accurate differentiation of *Escherichia coli* and *Shigella* serogroups: challenges and strategies. *New Microbes New Infect.* **21**, 58–62 (2018).
21. Geldenhuys, M. *et al.* A metagenomic viral discovery approach identifies potential zoonotic and novel mammalian viruses in *Neoromicia* bats within South Africa. *PloS One* **13**, e0194527 (2018).
22. Vibin, J., Chamings, A., Klaassen, M., Bhatta, T. R. & Alexandersen, S. Metagenomic characterisation of avian parvoviruses and picornaviruses from Australian wild ducks. *Sci. Rep.* **10**, 12800 (2020).

23. Qin, H. *et al.* Comparative transcriptomics of multidrug-resistant *Acinetobacter baumannii* in response to antibiotic treatments. *Sci. Rep.* **8**, 3515 (2018).

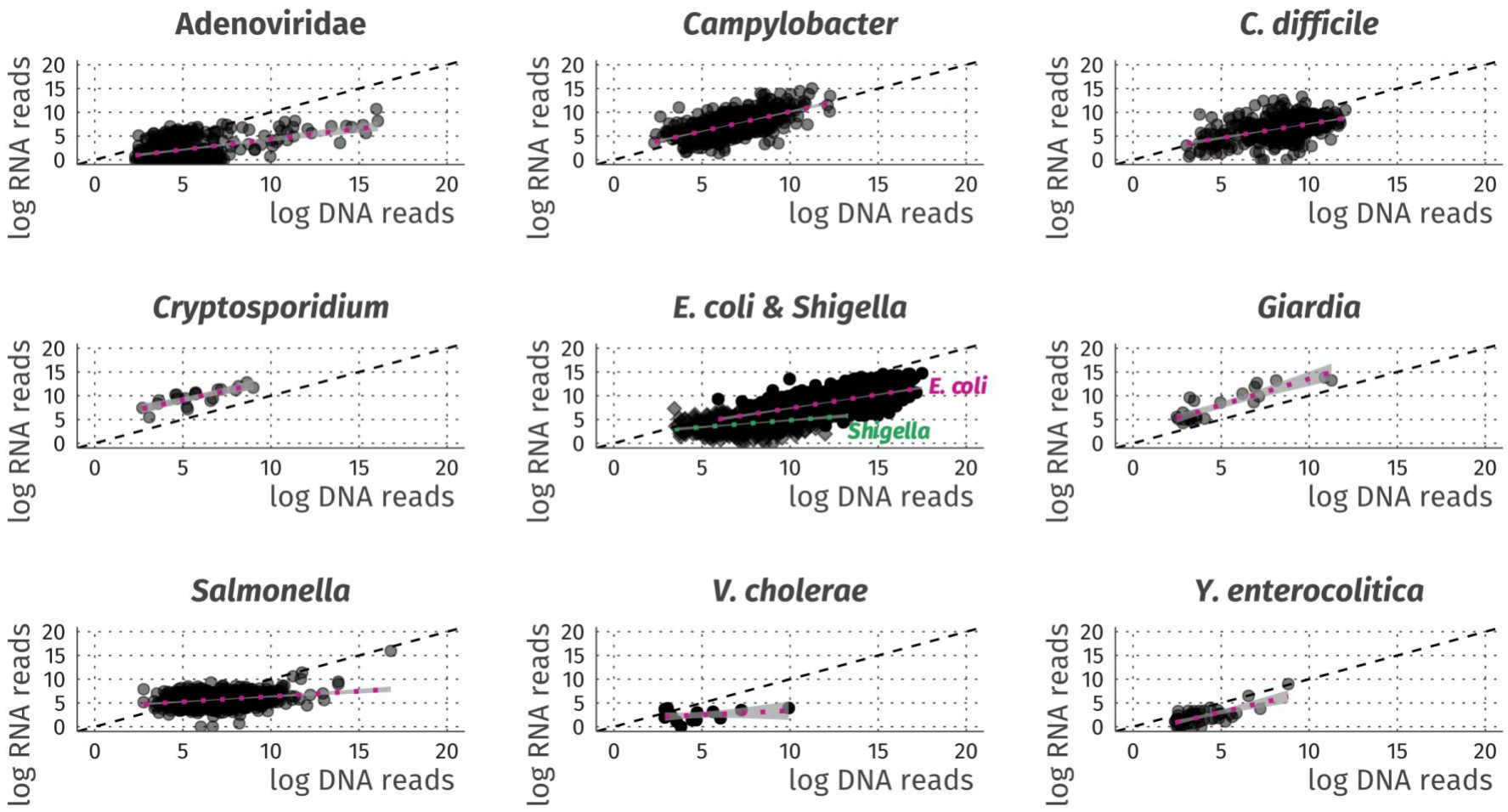

Figure 1: Visual overview and comparison of DNA (metagenomic) and RNA (metatranscriptomic) sequencing reads assigned to GI pathogens of relevance to the UK setting*. For all graphs, the dashed (black) intercept line is provided to highlight the skew of sensitivity towards either DNA, or RNA. The dashed pink line represents the results of a linear regression model, whilst the grey band surrounding this line is the 95% confidence interval. An additional dashed (green) line is present on the graph for E. coli & Shigella, in order to distinguish the univariate regression model results for these two taxa.*

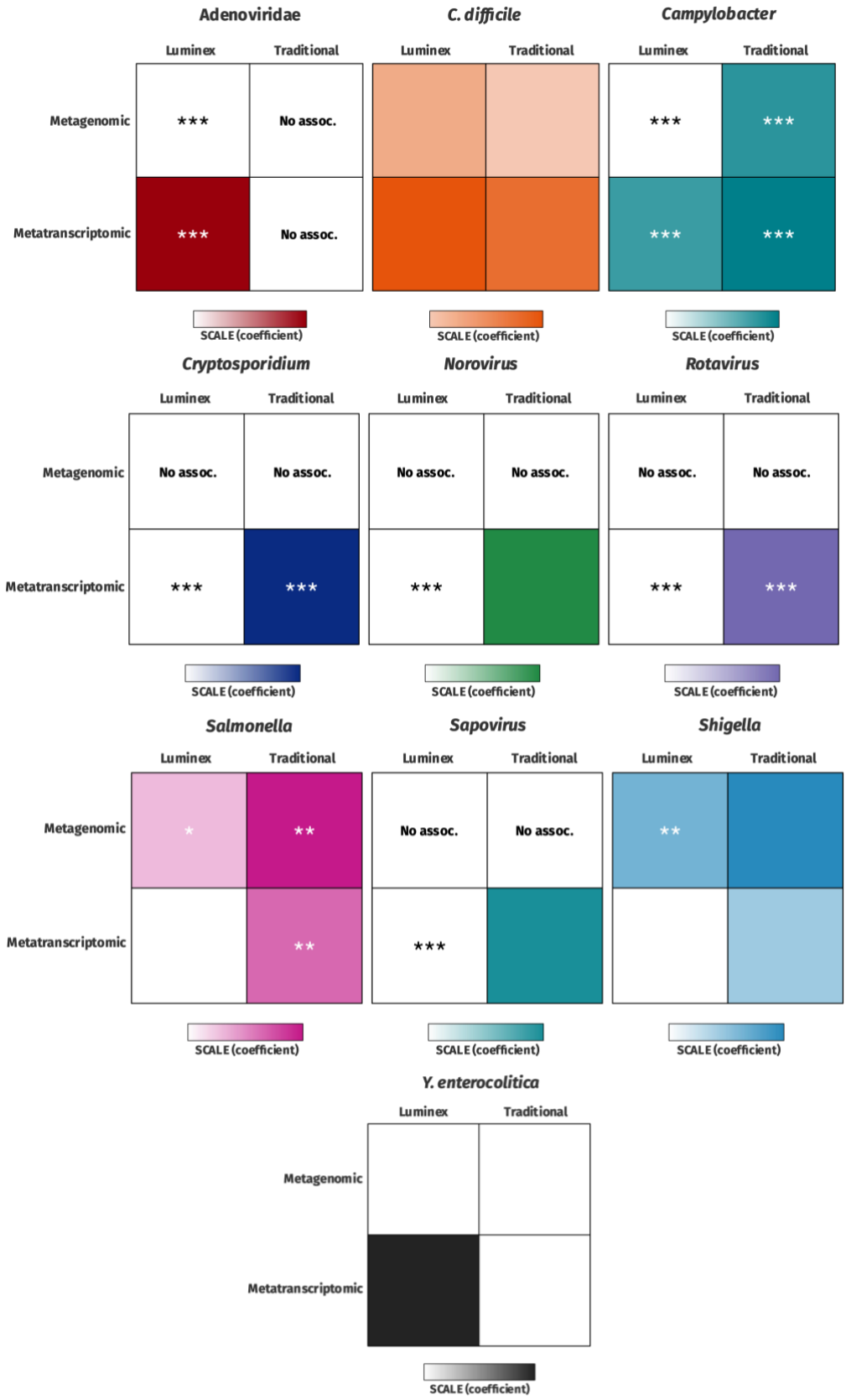

Figure 2: Statistically significant correlations were observed between sequencing data and laboratory tests for 8 out of 10 major GI community pathogens in the United Kingdom. *The darker the colour of a quadrant in a heatmap, the stronger the correlation (coefficient) between the detection of a pathogen in sequencing data (metagenomic or metatranscriptomic) and a laboratory result (Luminex or Traditional). Asterisks in quadrants indicate the statistical significance of correlations as follows: *: q < 0·05; **: q < 0·01; ***: q < 0·001*