nature research

Corresponding author(s): M. Stephen Trent

☐ Initial submission ☒ Revised version ☐ Final submission

Life Sciences Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form is intended for publication with all accepted life science papers and provides structure for consistency and transparency in reporting. Every life science submission will use this form; some list items might not apply to an individual manuscript, but all fields must be completed for clarity.

For further information on the points included in this form, see Reporting Life Sciences Research. For further information on Nature Research policies, including our data availability policy, see Authors & Referees and the Editorial Policy Checklist.

Experimental design

1. Sample size

Describe how sample size was determined.

For genomic studies of isolate populations from humans, sample availability, volunteer clinical outcomes, and read coverage of the reference genome, all influenced the number of samples in the study. First, a group of volunteers that had varying clinical outcomes during the human infection trial were determined. These patients included placebo treated and prophylactic treated patients, patients with and without recrudescent infections, and patients with and without severe disease that required early antibiotic intervention. Then available isolate population samples from these volunteers were prepared for genomic sequencing and sequenced. After sequencing and read mapping, the samples with robust coverage of the reference genome were advanced to variant analysis. We defined robust coverage in line with field standards. Samples that had at least 95% of genes covered with at least 25 fold coverage across the entire gene length, and at least 25 average fold coverage across the genome were considered to have robust coverage. Only samples with robust coverage were advanced and included in the manuscript. To our knowledge this work produced the deepest sequencing of C. jejuni isolate populations to date.

For genomic studies of isolate populations from non-human primates, a similar procedure was followed. All isolate populations harvested from days animals had diarrhea were sequenced and those that met the coverage analysis described above advanced to variant analysis and are included in the manuscript.

For transcriptomic studies, availability of RNA-later preserved samples and robust read coverage of the reference genome were the determining factors for sample selection. After PCR and Illumina MiSeq based screenings, 3 infected diarrhea samples from 3 different patients were chosen based on the feasibility of sequencing the samples enough for appropriate read coverage.

2. Data exclusions

Describe any data exclusions.

For genomic samples, exclusion criteria were dependent on genome coverage. Genomic samples with less than 95% of genes with at least 25 fold coverage across their entire length were excluded. As we wanted to be sure the homopolynucleotide tracts in the C. jejuni genome were accurately counted, we produced the deepest genomic sequencing of campy isolates ever published to our knowledge (average fold coverage >1000). Overall, 5 human isolate populations and 5 non-human primate isolate population samples were excluded from the variant analysis due to poor coverage. Overall, the excluded samples had very poor coverage of the genome (approximately < 5 percent of genes with at least 25 fold coverage).

For transcriptomic studies, the feasibility of obtaining robust read coverage of the reference genome excluded many samples. 19 infected diarrhea samples preserved in RNA-later were screened by PCR (to determine relative C. jejuni loads between samples) and/or Ilumina MiSeq (to quantify mappable RNA-seq reads). These screenings identified which samples would require the fewest sequencing reads for appropriate reference sequence coverage to yield robust statistically

significant differential gene expression analysis. Based on these initial screenings alone, three were chosen for the required additional sequencing and are used in the manuscript. All three samples were at least grade 3 stools produced by different volunteers and are considered biological replicates of infected diarrhea populations. Rejected samples would have been cost-prohibitive to sequence to an acceptable coverage.

3. Replication

Describe whether the experimental findings were reliably reproduced.

When repeat analyses/experiments were appropriate to preform they were successful in reproducing the results presented in the manuscript.

4. Randomization

Describe how samples/organisms/participants were allocated into experimental groups.

Samples were allocated into groups as determined by the clinical outcomes as discussed in the manuscript. Groups of data used in comparison analyses were determined as described in the manuscript. Briefly, we present full data sets first before using appropriate groups of these data in comparisons or figures. For instance, we include all genome variants called in all samples (down to 1 percent frequency of occurrence in the individual sample population) in supplemental data before highlighting the most common variants with the largest change in frequency between pre and post infection in the main text. Similarly, we include the entire statistically significant differentially expressed genes between transcriptomics samples before highlighting special genes of interest.

5. Blinding

Describe whether the investigators were blinded to group allocation during data collection and/or analysis.

Those involved with genetic and transcriptional analyses were and are still blind to patient identification information. Those involved with genetic and transcriptional analyses were not blind to clinical outcomes (severe disease, prophylactic vs placebo treatment, recrudescence status) of patients along with their corresponding samples, as these outcomes determined the groups for comparison analyses. There was similarly no blinding of non-human primate isolate populations, as only those from diarrheal samples were relevant for analysis.

Note: all studies involving animals and/or human research participants must disclose whether blinding and randomization were used.

6. Statistical parameters

For all figures and tables that use statistical methods, confirm that the following items are present in relevant figure legends (or in the Methods section if additional space is needed).

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.)
☐	☒ A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ A statement indicating how many times each experiment was replicated
☐	☒ The statistical test(s) used and whether they are one- or two-sided (note: only common tests should be described solely by name; more complex techniques should be described in the Methods section)
☐	☒ A description of any assumptions or corrections, such as an adjustment for multiple comparisons
☐	☒ The test results (e.g. *P* values) given as exact values whenever possible and with confidence intervals noted
☐	☒ A clear description of statistics including central tendency (e.g. median, mean) and variation (e.g. standard deviation, interquartile range)
☐	☒ Clearly defined error bars

See the web collection on statistics for biologists for further resources and guidance.

▸ Software

Policy information about availability of computer code

7. Software

Describe the software used to analyze the data in this study.

CLC Genomics Workbench (Through version 9.5) along with the Microbial Genomics module (Trial version), Prism, and Microsoft Excel (2010 Mac) were used for data analysis when appropriate.

For manuscripts utilizing custom algorithms or software that are central to the paper but not yet described in the published literature, software must be made available to editors and reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). *Nature Methods* guidance for providing algorithms and software for publication provides further information on this topic.

Materials and reagents

Policy information about availability of materials

8. Materials availability

Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a for-profit company.	Material availability is only restricted by the amount of sample collected during the human and non-human primate infection models, minus the amount of material used to prepare nucleic acids for sequencing.

9. Antibodies

Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species).	Not applicable

10. Eukaryotic cell lines

a. State the source of each eukaryotic cell line used.	Not applicable
b. Describe the method of cell line authentication used.	Not applicable
c. Report whether the cell lines were tested for mycoplasma contamination.	Not applicable
d. If any of the cell lines used are listed in the database of commonly misidentified cell lines maintained by ICLAC, provide a scientific rationale for their use.	Not applicable

Animals and human research participants

Policy information about studies involving animals; when reporting animal research, follow the ARRIVE guidelines

11. Description of research animals

Provide details on animals and/or animal-derived materials used in the study.	Research animals were not directly examined in this study. Bacterial isolate populations were derived from samples produced from an independent study approved by U.S. Navy Medical Research included in the main text disclaimer.

Policy information about studies involving human research participants

12. Description of human research participants

Describe the covariate-relevant population characteristics of the human research participants.	Human subjects were not directly examined in this study. Bacterial isolate populations were derived from samples produced by volunteers that participated in the ClinicalTrials.gov Identifier NCT02280044 clinical trial and included in the main text disclaimer. These volunteers were considered healthy adults from the mid-Atlantic region of North America.