Whole Genome Sequencing (WGS) and Bioinformatics analysis of Escherichia coli isolates from Nepal

(Bid Reference: RFP/NEP/2023/005; Unit Name: WHE)

Background Information and Technical Details to Achieve the Required Objectives

Child Health Research Foundation (CHRF) is a non-profit, non-government organization, located in Dhaka, Bangladesh. With a mission to prevent infections and save lives, the foundation has been conducting semi-national infectious disease surveillance in Bangladesh for the last several decades, specifically focusing on pediatric infectious diseases. To achieve its goals, the foundation has established microbiology laboratories in four hospitals across Bangladesh, including two of the largest pediatric hospitals in the country; all four laboratories serve as high-performing WHO-sentinel sites. CHRF's surveillance platform uses traditional laboratory techniques like culture, polymerase chain reaction, and serology. In 2016, the foundation enhanced its surveillance network by adding pathogen genomics capacity.

CHRF has a state-of-the art genomics center that can undertake sequencing under three different modalities. The first modality - unbiased-RNA metagenomics - has primarily been used for investigating infectious outbreaks of unknown etiology. We have used this technique to uncover a Chikungunya meningitis outbreak in 2017 [1], and find the circulating viral genotype during the highly invasive dengue outbreak in 2020 [2].

The second modality at CHRF and one most relevant to the proposal is whole genome sequencing for genomic epidemiological studies on bacterial pathogens. Our group was the first to describe the population structure (using phylogenetic analysis and sequence typing) and antimicrobial resistance of Salmonella Typhi in Bangladesh [3]. We also discovered a novel mechanism of azithromycin resistance in Salmonella Typhi and Paratyphi A [4, 5]. Genomic analysis from our team also demonstrated the intercontinental spread of different genotypes of Salmonella Typhi and their antimicrobial resistance [6]. More recently, our team sequenced ~600 Salmonella Paratyphi A from Bangladesh [7]. We also developed the first open-access tool for genotyping Salmonella Paratyphi A, which was beta tested by Sanger Institute and has now been published in Nature Communications [7].

The third modality is amplicon-based sequencing that is used to investigate outbreaks from known viral pathogens. In this approach, we utilize a library of primers designed to cover the entire viral genome to amplify viral genetic materal and subject all these fragments to next-generation sequencing. This is the strategy we have used to track SARS-COV-2 evolution in

Bangladesh during the COVID-19 pandemic [8], and are currently using it to track respiratory syncytial virus and parvoviruses.

Complementing the capacity for genome sequencing, CHRF has developed world leading expertise in bioinformatics analysis of genomic data. Our bioinformatics team has worked extensively on data generated by different sequencing projects and developed in-house computational expertise required for analyzing sequencing data. We also have built the computational infrastructure needed for securely storing and processing genomic data for our genomic epidemiological studies.

Proposed solution

The Tricycle project conducted by WHO is utilizing a One Health approach to investigating Escherichia coli infections in Nepal. The proposed project will focus on ESBL E. coli and obtain isolates from three main areas of concern: human (hospital and community), food, and the environment. Sequencing whole genomics from 100 E. coli isolates from this project will inform the public health community about the transmission of this important pathogen across these different environmental niches and allow for design of evidence-based interventions.

To fulfill the objectives of this project (RFP/NEP/2023/005), we propose to generate whole-genome sequence (WGS) data of 100 E. coli isolates (extracted DNA) using 150-bp paired-end sequencing technology on our Illumina Nextseq2000 platform. We aim to provide at least 1 million reads for each sample to ensure and over 50x genome coverage (considering 5.5 Mbp genome size).

The data obtained will then allow us to achieve the following objectives derived from the RFP:

1) Explore the genetic diversity among the isolates from different niches

2) Compare commensal and pathogenic E. coli

3) Identify potential transmission events

4) Acquisition/emergence of AMR and other virulence genes, bacterial sequence types (STs), and plasmids in Nepal.

Our proposed bioinformatic solutions for each of these four stated objectives is listed below:

1. Whole-genome mapping of the 100 WGS of E. coli isolates, followed by variant calling, SNP filtering, consensus alignment, and generation of maximum-likelihood phylogenetic trees. This analysis will uncover the genomic diversity and evolution of the strains across different isolation sources and environmental niches (Objective 1).

2. Comparison of genomes of commensal and pathogenic E. coli isolates from humans by overlaying this information on the phylogenetic tree. (Objective 2). The phylogenetic tree will also help us identify potential transmission events (Objective 3). These would be pathogenic isolates that cluster with environmental isolates rather than other pathogenic isolates or vice versa.

3. De-novo assembly and annotation of all WGS data to ease the downstream analyses. Assembled contigs will be required to analyze and validate the screening for AMR genes, virulence factors, and plasmids. (Objectives 3 and 4)

4. Screening for presence of AMR genes in all E. coli WGS data using Resfinder and NCBI database. Resfinder is a well-known AMR database, especially for food-borne pathogens. It is regularly maintained by Technical University of Denmark (DTU) and has been cited in over 4200 articles to date [9]. The recently updated NCBI AMR database will also be utilized [10]. Both these databases include all known ESBL, carbapenem and colistin resistance genes which would be of interest for this project (Objective 3 and 4)

5. Identification of virulence genes in E. coli WGS data using the virulence factor database (VFDB) and NCBI database. The VFDB is maintained by the Chinese Academy of Medical Sciences in Beijing, China, and has been cited in almost 1000 articles to date [11]. The NCBI also maintains a reference database of virulence factors that will be used [10]. (Objective 3)

6. Identify the multi-locus sequence type (MLST) of the isolates, using 7-gene loci sequences from WGS data. In case of any novel or, incomplete ST type, the WGS data will be checked on Enterobase (https://enterobase.warwick.ac.uk/species/ecoli/allele_st_search). This database is regularly maintained by the University of Warwick and is a well-known repository of publicly available Escherichia genomes [12]. (Objective 3)

7. Detect presence of different plasmids using Plasmidfinder and PLSDB database. Plasmidfinder is replicon based-typing system maintained by the Technical University of Denmark (DTU) and has been cited in over 2700 articles [13]. PLSDB is another recent comprehensive database, maintained by the University of California Merced, USA [14]. Both these tools will be used to identify the plasmids present in each sample via their origin of replication. (Objective 3)

The tools and bioinformatics pipelines proposed are already set up on our servers and are being regularly used for ongoing projects at our organization.

Proposed Approach/Methodology

Following the proposed solution above, we will execute the analyses using the methodology described below:

Sequencing and generating raw Fastq data

The extracted DNA will be measured for its quality and quantity with NanoDrop (Thermo Fisher Scientific, MA, USA) and Qubit (Thermo Fisher Scientific, MA, USA). The 260/280 nm ratio must be 1.8–2.0 on NanoDrop and a minimum of 30 ng/µl concentration on Qubit. The minimum required volume for sequencing would be 30 µl. The extracted DNA will then be run on an agarose gel to check for intactness of the DNA. In case of an unsatisfactory DNA quality/quantity check, the WHO Nepal team will be notified. Either samples can be removed, or fresh genomic DNA can be shipped to us. In case WHO Nepal team requests WGS for unsatisfactory DNA samples, CHRF will follow all steps to maximize the chance of success, but for bad sequence quality obtained for the unsatisfactory DNA samples, the CHRF team will not repeat sequencing and will not be held responsible.

Once the DNA quality/quantity check is complete, sequencing libraries will be prepared using the NEBNext® Ultra™ DNA Library Prep Kit for Illumina (New England Biolabs, MA, USA) following the manufacturer's instructions. All DNA libraries will be sequenced on Illumina Nextseq 2000 platform (150-bp paired-end) at CHRF. All sequence data will be saved in fastq data format.

Quality check, genome assembly, annotation of the genomes

All raw Illumina fastq reads of all 100 E. coli isolates will be quality-checked using FastQC and trimmed using Trimmomatic (if necessary) [15, 16]. All trimmed sequencing data will then be de-novo assembled using Unicycler (default options, with –min_fasta_length 200) [17]. The assembled contigs will be annotated using Prokka (default options, with --gcode 11 --mincontiglen 200) [18].

Genome mapping, variant calling, and generating a phylogenetic tree

All trimmed raw fastq reads of 100 E. coli isolates will be mapped against the NCBI reference genome, Escherichia coli str. K-12 substr. MG1655 (genbank accession: NC_000913.3) using BWA-MEM [19]. Variant calling and SNP filtering will be done using SAMtools and BCFtools [20]. Only the homozygous, unambiguous SNPs with a Phred-quality score of >20 will be selected using a customized python script, developed at CHRF [7]. SNPs will be discarded in case of any strand bias $p < 0.001$, mapping bias $p < 0.001$, or tail bias $p < 0.001$ (using vcfutils.pl script, from SAMtools). SNPs located in phage or repeat regions (as described within genbank

features of NC_000913.3), or recombinant regions (as detected by Gubbins [21]) will be discarded as well, using a customized python script developed at CHRF [7]. Using these high-quality SNPs, a consensus genome alignment will be built using SAMtools [20]. Based on this alignment, a maximum likelihood phylogenetic tree will be generated using RAxML with 100 bootstrap pseudo-analysis [22]. A related Escherichia sp. reference genome will be used as an outgroup in the tree and go through the same mapping pipeline. Tree visualization and annotation will be done using iTol [23].

Screening for presence of AMR and virulence genes, and plasmids

All de-novo assembled contigs will be screened with ResFinder, Pointfinder, and AMRFinderPlus (default options) to detect the presence of AMR genes (including details on ESBL genes) and mutations [9, 10, 24]. To find the presence of virulence genes, we will scan all assembled contigs with the VFDB and NCBI AMRFinderPlus database [10, 11]. In case of finding plasmids, we will use the fastq raw data and assembled contigs with Plasmidfinder and PLSDB database to detect the plasmid incompatibility types [13, 14]. All results will be curated and combined using a customized python script for individual tools, developed at CHRF.

Identify sequence types (STs) and eBURST analyses

To detect the sequence type (ST) of the sequenced E. coli genomes, we will use multi locus sequence typing/MLST (https://github.com/tseemann/mlst). This tool uses the PubMLST database (https://pubmlst.org/) to detect the ST of an isolate. In case of any novel or, incomplete ST type, the genome sequences will be checked on Enterobase (https://enterobase.warwick.ac.uk/species/ecoli/allele_st_search).

Using the MLST profiles of all 100 isolates, an MLST-based eBURST analysis will be done using the goeBURST algorithm on Phyloviz [25].

Proposed Timeline

The following timeline is set based on timeline of deliverables.