

International Journal of Advanced Biochemistry Research



ISSN Print: 2617-4693
 ISSN Online: 2617-4707
 IJABR 2017; 1(1): 22-37
 www.biochemjournal.com
 Received: 05-01-2017
 Accepted: 16-03-2017

Eze Uchenna Nwabunwanne
 Biotechnology Program,
 Department of Mathematics
 and Natural Sciences, BRAC
 University, Dhaka,
 Bangladesh

Ifekthar Bin Naser
 Biotechnology Program,
 Department of Mathematics
 and Natural Sciences, BRAC
 University, Dhaka,
 Bangladesh

Olunuga Omotola
 College of Life Sciences and
 Technology, Huazhong
 Agricultural University,
 Hongshan, Wuhan, P.R.,
 China

Ubenyi Stanley Mary
 Department of Medical
 Biochemistry, Faculty of Basic
 Medical Sciences, University of
 Nigeria, Enugu State, Nigeria

Ugwu Chukwuebuka V
 Department of Medical
 Biochemistry, Faculty of Basic
 Medical Sciences, University of
 Nigeria, Enugu State, Nigeria

Correspondence
Eze Uchenna Nwabunwanne
 Biotechnology Program,
 Department of Mathematics
 and Natural Sciences, BRAC
 University, Dhaka,
 Bangladesh

Comparative analysis of CRISPR-Cas systems of *Yersinia pestis* and *Escherichia coli* strains

Eze Uchenna Nwabunwanne, Ifekthar Bin Naser, Olunuga Omotola, Ubenyi Stanley Mary and Ugwu Chukwuebuka V

DOI: <https://doi.org/10.33545/26174693.2017.v1.i1a.94>

Abstract

The CRISPR-Cas system primarily refers to the clustered regularly interspaced short palindromic repeats (CRISPR) and its related protein enzymes (Cas) that confer adaptive protection against bacteriophages and other exogenous elements. The CRISPR-spacers that constitute the CRISPR-Cas systems have been reportedly found in Prokaryotes and Archea. This study analyzed and compared genome assemblies of *E. coli* and *Yersinia pestis* with the aid of bioinformatics tools (MinCED, MUSCLE alignment tool, nucleotide Basic Local Alignment Search tool (nBLAST), and prokaryotic database of the National Centre for Biotechnology Information (NCBI). The CRISPR-spacer analyses of meta-genomic sequences of 162 *Escherichia coli* and 121 *Yersinia pestis* strains showed the presence of 4 CRISPR classes (I, II, III, and IV) in *E. coli* and 3 CRISPR classes (I, II, and III) in *Y. pestis*, respectively. The result of MinCED-CRISPR analysis of CRISPR size of genome assemblies of *E. coli* and *Y. pestis* strains showed that *E. coli* CRISPR class I had the highest percentage value (51.9%) compared with *E. coli* class II (42%), *Y. pestis* CRISPR class I (40%), *Y. pestis* CRISPR class II (36%), *Y. pestis* CRISPR class III (24%), *E. coli* class III (4.8%) and *E. coli* class IV (1%). It further revealed that no CRISPR sequences were found in class IV of *Y. pestis* strains. The result of CRISPR-spacer contents also showed that the *E. coli* CRISPR class I showed the highest significant percentage value (54.6%) than *Y. pestis* CRISPR class I (36.93%), *Y. pestis* CRISPR class II (41.13%), *Y. pestis* CRISPR class III (21.93%), *E. coli* CRISPR class II (42.18%), *E. coli* class III (2.86%) and *E. coli* class IV (0.33%). It was observed that there was absence of CRISPR class IV in the genomic sequences of *Y. pestis* strains analyzed. The percentage distribution of exogenous spacers among *E. coli* CRISPR classes revealed that CRISPR class I showed 65.77% homologous sequence match with plasmids spacers, 51.96% bacteriophages spacers, 68.26% bacterial spacers, and 20.18% unknown target spacers, while CRISPR II had 79.82% homologous sequence match with unknown self-targets spacers, 49% bacteriophages spacers, 35% plasmids and 29.9% from other bacterial spacers available in NCBI databank. *E. coli* CRISPR class III showed no homologous spacers sequence match with unknown targets and bacteriophages but showed homologous sequence match with plasmid (0.19%) and other bacteria (1.83%). *E. coli* CRISPR class IV showed absence of homologous spacers matches against bacteriophages spacers from the NCBI databank. The spacer distribution among *Y. pestis* CRISPR classes revealed that the CRISPR class I showed a highest percentage value of homologous sequence match with 47.62% bacteriophages spacers, 38.7% plasmids, and 38.44% from other bacterial spacers, while the *Y. pestis* CRISPR class II showed spacers similarity with plasmids (25.45%) and bacteriophages (25.71%) and *Y. pestis* CRISPR III also showed percentage similarity sequence match with plasmids spacers (35.84%), bacteriophages spacers (26.67%) and 22.06% bacterial spacers available in NCBI databank. However, there was no homologous spacer sequence match contributed by unknown mobile genetic targets in all the *Y. pestis* CRISPR classes compared with the *E. coli* CRISPR classes. The evolutionary relationships of 22 representative strains of each bacterium were carefully selected on temporary relationship over a specific period [n < 10 years] and analyzed. The phylogenetic tree of *E. coli* revealed that 22 *E. coli* strains showed a common ancestral origin from *E. coli* BIDMC_74, and the *E. coli* BIDMC_74 strain was more closely related to *E. coli* strain IH57218 than *E. coli* str. HVH 50 and *E. coli* str. 122262 NODE_1 respectively. Phylogenetic analysis of all 22 *Y. pestis* showed a close relationship with one another, suggesting a common evolutionary relationship among them. The results obtained from this study give credence to show that the *E. coli* showed a significant CRISPR diversity than *Y. pestis*, in terms of its CRISPR class size, spacer's contents, exogenous homologous sequence matches and phylogenetic relationships among its strains. This characteristic feature showed by *E. coli* strains could be attributed to increased homologous spacer acquisition from exogenous plasmids, bacteriophages, bacteria, and unknown targeting elements. The findings further suggest that the increased CRISPR diversity observed in the *E. coli* could be associated with increased exposure of its strains to these exogenous elements than *Y. pestis* strains due to bacteriophages infection, co-evolution and conjugation with exogenous mobile genetic elements (MGE).

Keywords: CRISPR-Cas, spacers, bacteriophages, MinCED tool, CRISPR class, mobile genetic elements (MGE), *Escherichia coli*, *Yersinia pestis*

1. Introduction

Bacteria dominate many natural habitats, including unfavorable ones, even though they are frequently attacked by predatory viruses such as bacteriophages.

Bacteriophages are bacteria-infecting viruses that are often considered the most abundant and diverse organisms in the biosphere due to their presence in nature ^[1]. Bacteriophages have a basic or complex particulate structure, and their genomes may encode a few to hundreds of genes, and they are made up of either genetic material (DNA or RNA) enveloped by a protein coat ^[2]. Bacteriophages may exist as a single organism or in combination with bacteria in nature and there are more than 1,031 bacteriophages in the globe ^[3, 4]. Bacteriophages are regarded as one of the most abundant living organisms in the water, and often infect 70% of bacteria in the ecosystem ^[5]. Importantly, phage infects bacteria to propagate and multiply in the host organism, as well as reproduce within the host by integrating its genome into the host organism's cytoplasm ^[5]. Following the injection of their genetic material into the bacterium, they multiply and integrate their genetic material into the host organism, making them a useful molecular tool for genome editing and genetic manipulation ^[6, 7]. The infiltration of bacteriophages into the host organism as well as the integration of its genetic material led to the evolution of a protective and adaptive immunity by host organism ^[7]. The evolution of the CRISPR system in the host organism provides adaptive immunity to the bacteria and other microbes against bacteriophages' invasion into the host ^[8]. CRISPRs have been shown experimentally to protect *Streptococcus thermophilus* and *Staphylococcus epidermidis* against phages and plasmids, respectively ^[9]. Also, it has been shown that CRISPR sequences are found, in approximately 50% of sequenced bacterial genomes and nearly 90% of sequenced archaea ^[10]. CRISPR constitute a family of DNA sequences found in genomes of prokaryotic organisms usually bacteria and archaea, and their sequences are derived from DNA fragments of bacteriophages that had previously infected the prokaryotes ^[11]. Studies have shown that CRISPR provides innate adaptive immunity to several bacteria and archaea, through the use of a microbial protective enzyme, *Cas* endonucleases (nucleases) that are involved in the anti-phage defense system. CRISPRs have guide sequences used to detect and destroy DNA sequences from similar bacteriophages during subsequent infections ^[12, 13]. These CRISPR sequences play a key role in the antiviral (anti-phage) defense system of prokaryotes, hence confers a form of acquired immunity to the host organism ^[14]. For instance, Cas 9 (CRISPR-associated protein 9) is an enzyme that uses CRISPR sequences as a guide to recognize and cleave specific strands of DNA which are complementary to the CRISPR sequence found in the bacteriophages and other exogenous elements ^[15]. The Cas9 enzymes together with CRISPR sequences form the basis of a technology known as CRISPR-Cas9 tool which can be used to edit genes within organisms ^[16]. Several studies have shown that the repeat nucleotide sequences, otherwise termed "Clustered Regularly Interspaced Short Palindromic Repeats regions (CRISPRs)" are characterized by short, and perfectly conserved elements consisting of 20 to 40 nucleotide base pairs separated by spacer sequences ^[17]. In several bacteria anti-phage defenses, it has been shown that CRISPR-Cas systems could be used to detect and destroy foreign nucleic acids and exogenous elements through the use of CRISPR RNAs (crRNAs) and Cas nucleases ^[18, 19]. In addition, these CRISPRs have several functions that are directly involved in DNA rearrangement and replication, host cell defenses, and

guide RNA (gRNA) control, and most bacterial cells have developed numerous means to survive by fending off phage viruses and other exogenous elements during virulent conditions through the recognition of their sequences and breakdown of their genetic materials by the use of their CRISPR-cas enzymes ^[20]. Bacteria also resist phages, due to presence of phage-inducible chromosomal islands (PICIs), in such a way to defend against phage infection and invasion of the bacterial chromosome upon phage infection, replicates, and interferes with phage reproduction ^[21]. Notably, the CRISPR-Cas system has been broadly classified into two broad CRISPR classes, six distinct CAS subtypes (I–VI), and dozens of cas enzymes from different CAS genes, with diverse mechanisms of actions ^[22, 23]. For instance, the Class 1 system (subtypes: I, III, and IV) encode multi-subunit effector complexes, and the Class 2 system (subtypes: II, V, and VI) relies upon a single subunit to destroy nucleic acid invaders ^[24]. Several notable research attempts have explicitly elucidated and analyzed the functional mechanism of the CRISPR-cas adaptive immunity in the host organism and discovered that the CRISPR arrays are separated by different short sequences known as "spacers" ^[25]. However, the functional mechanisms have showed that the CRISPR-Cas immunity occurs in three main steps namely: adaptation, CRISPR RNA (crRNAs) biogenesis, and interference. During the adaptation stage, short (30–40 nucleotides) invader-derived sequences called "spacers" are captured and integrated into CRISPR loci in between partially palindromic DNA repeats of similar length ^[25]. During crRNA biogenesis, the repeat-spacer array is transcribed into a long precursor crRNA, which is further processed to liberate mature crRNAs that specifies a single target. During interference, crRNAs combine with one or more CAS proteins to form an effector complex that recognizes and degrades nucleic acids (*Protospacers*) that are complementary to the crRNA ^[26]. The crRNA biogenesis and interference stages constitute the defense phase of CRISPR-Cas immunity, and all CRISPR-Cas systems adhere to this general pathway ^[27]. The three basic stages appear to be common to all CRISPR systems, CRISPR loci, and the proteins that mediate each stage of adaptive immunity are remarkably diverse ^[28].

1.1 Bacteriophages: Literature review

In 1896, Ernest Hanbury Hankin reported that some particles in the waters of the Ganges and Yamuna rivers in India had a marked antibacterial action against cholera and it could pass through a very fine porcelain filter ^[29]. In 1915, Frederick Twort, a British bacteriologist and superintendent at the Brown Institute of London discovered a small agent that infected and killed bacteria ^[30]. He believed the agent must be one of the following: *a stage in the life cycle of the bacteria, the enzyme produced by the bacteria themselves, or a virus that grew on and destroyed the bacteria* ^[31]. However, Twort's research was interrupted by the onset of World War I, due to shortage of funding, and his works led to the discovery of antibiotics ^[32]. Independently, French-Canadian Microbiologist Félix d'Hérelle, who worked at the Pasteur Institute in Paris, announced on September 3, 1917, that he had discovered "*an invisible, antagonistic microbe of the dysentery bacillus*" ^[33]. Even though, d'Hérelle thought there was no question as to the nature of his discovery, but reported his observation: "*In a flash, I had understood: what caused my clear spots was an invisible microbe - a*

virus parasitic on bacteria [34]." d'Hérelle named the virus as "bacteriophages", implying a bacteria-eater (Greek phage in meaning "to devour") [35]. He also recorded a dramatic account of a man suffering from dysentery who was restored to good health by the bacteriophages [36]. Notably, d'Herelle was known to have conducted much research into bacteriophages and later introduced the concept of phage therapy [37]. The classification of viruses in the early 1940s by Lwoff, Horne and Touiner recognized and took into account of the viral nature of bacteriophages under the electron microscope, and classified them into groups based on their morphologies and type of nucleic acid [38]. Later,

Lwoff, Horne, and Tournier published a list of the order of the classification of prokaryotic viruses (Table 1).

Morphological and Structural Classification of Bacteriophages

The phages were classified into six groups based on their morphologies and types of nucleic acid by the International Committee on Taxonomy of Viruses (ICTV) (as shown in Table 1). Some of the major classes of these Bacteriophages include *Urovirales* for tailed phages: *Microviridae* (icosahedron), *Inoviridae* (filamentous), and Φ X-type phages [39].

Table 1: ICTV classification of prokaryotic (bacterial and archaeal) viruses

Order	Family	Morphology	Nucleic acid	Examples
Belfryvirales	<i>Turriviridae</i>	Enveloped, isometric	Linear dsDNA	-
Caudovirales	<i>Ackermannviridae</i>	Non enveloped, contractile tail	Linear dsDNA	-
	<i>Myoviridae</i>	Non enveloped, contractile tail	Linear dsDNA	<i>Bacteriophages: T4, Mu, P1, P2</i>
	<i>Siphoviridae</i>	Non enveloped, non contractile tail (long)	Linear dsDNA	<i>Lambda phage(λ) Phage T5, Bacteriophage HK97, Enterobacteria phage(N15)</i>
	<i>Podoviridae</i>	Non enveloped, non contractile tail (short)	Linear dsDNA	<i>T7 (Bacteriophage T7) T3 (Bacteriophage T3) Bacillus phage phi 29(Φ29), Enterobacteria phage 22(P22)</i>
Halopanivirales	<i>Sphaerolipoviridae</i>	Enveloped, isometric	Linear dsDNA	-
Haloruvirales	<i>Pleolipoviridae</i>	Enveloped, pleomorphic	Circular ssDNA, circular dsDNA, or linear dsDNA	-
Kalamavirales	<i>Tectiviridae</i>	Non enveloped, isometric	Linear dsDNA	-
Levivirales	<i>Leviviridae</i>	Non enveloped, isometric	Linear ssRNA	<i>Bacteriophages MS2, Bacteriophages Qβ</i>
Ligamenvirales	<i>Lipothrixviridae</i>	Enveloped, rod-shaped	Linear dsDNA	Acidianus filamentous virus 1
	<i>Rudiviridae</i>	Non enveloped, rod-shaped	Linear dsDNA	Sulfolobus islandicus rod-shaped virus 1
Mindivirales	<i>Cystoviridae</i>	Enveloped, spherical	Segmented dsRNA	<i>Pseudomonas virus phi (Φ6)</i>
Petitvirales	<i>Microviridae</i>	Non enveloped, isometric	Circular ssDNA	<i>Bacteriophages Phi ΦX174</i>
Tubulavirales	<i>Inoviridae</i>	Non enveloped, filamentous	Circular ssDNA	<i>M13 bacteriophages</i>
Vinavirales	<i>Corticoviridae</i>	Non enveloped, isometric	Circular dsDNA	<i>Pseudomonas virus PM12</i>
Unassigned	<i>Ampullaviridae</i>	Enveloped, bottle-shaped	Linear dsDNA	-
	<i>Bicaudaviridae</i>	Non enveloped, lemon-shaped	Circular dsDNA	-
	<i>Clavaviridae</i>	Non enveloped, rod-shaped	Circular dsDNA	-
	<i>Finnlakeviridae</i>		dsDNA	<i>FLiP(Flavobacterium virus)</i>
	<i>Fuselloviridae</i>	Non enveloped, lemon-shaped	Circular dsDNA	-
	<i>Globuloviridae</i>	Enveloped, isometric	Linear dsDNA	-
	<i>Guttaviridae</i>	Non enveloped, ovoid	Circular dsDNA	-
	<i>Plasmaviridae</i>	Enveloped, pleomorphic	Circular dsDNA	-
	<i>Portogloboviridae</i>	Enveloped, isometric	Circular dsDNA	-
	<i>Spiraviridae</i>	Non enveloped, rod-shaped	Circular ssDNA	-
	<i>Tristromaviridae</i>	Enveloped, rod-shaped	Linear dsDNA	-

Phages are divided into several categories depending on their shape, sizes, and genetic make-up of single nucleic acid (DNA or RNA) enclosed by a protein capsid coat [40]. The great majority of phages usually have protein tails that allow them to recognize a receptor on the surface of the host bacterium [41]. In 1967, Bradley recognized six basic phages, and further classified them into three groups as tailed,

filamentous and icosahedral phages with either single-stranded DNA (ssDNA) or single-stranded RNA (ssRNA) [42]. Presently, newly discovered bacteriophages are classified into these groups, based on their size and shapes namely: Icosahedron bacteriophages, Filamentous bacteriophages, and Complex bacteriophages (as shown in Figure 1).

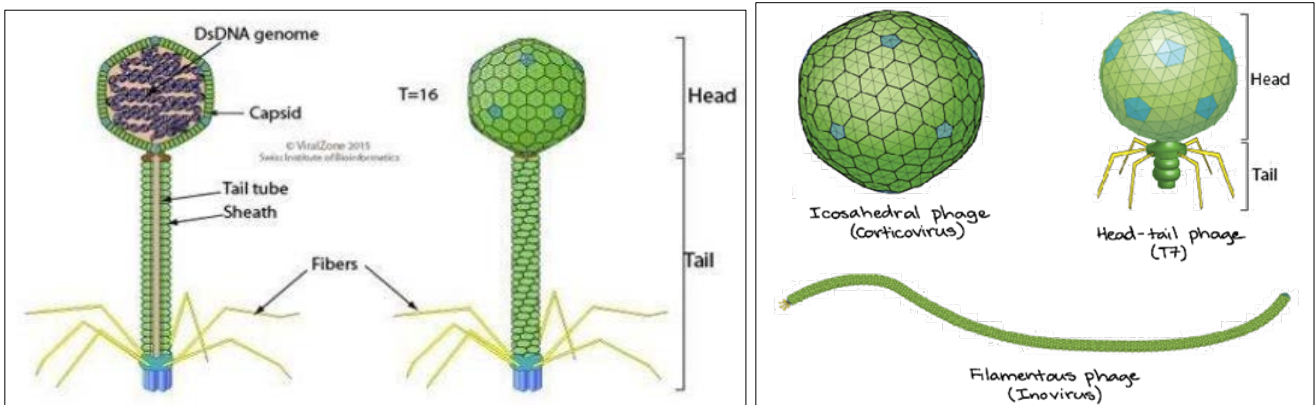


Fig 1: Structure of T4 bacteriophage, Corticovirus (Icosahedron) and Inovirus (Filamentous).

1. **Icosahedron bacteriophages:** this type of bacteriophage has a spherical shape, with twenty triangular facets, and the smallest size of 25nm in diameter.
2. **Filamentous bacteriophages:** They are made up of long tubes formed by capsid protein and assembled into the helical structure with about 900nm in diameter size.
3. **Complex bacteriophages:** They have icosahedral heads attached to helical tails fibers and base plates.

1.2 Bacteriophage: life Cycle

The Life cycle of Bacteriophages is classified into two groups namely: lytic (virulent) and lysogenic (temperate) bacteriophages, according to their biological cycles [43]. Bacteriophages are infectious to bacteria, due to the presence of receptors on their surfaces which enable bacteria to bind onto the phage as well as identify their specific host organisms [43]. Bacteriophages bind to their host’s cell receptors by adsorption and inject their genetic material into the host cell. The significant distinction between the lytic and lysogenic phage cycles lies in the fate of bacteria [44]. In the lytic cycle, lytic phages take over the bacterial replication machinery by reproducing a new phage virus (progeny) within its host. As a result of the reproduction of multiple copies of these new phages that form a critical mass which triggers lysing of the bacteria cell wall, and the release of new phage viruses, and beginning of another lytic cycle [45]. The formation of critical mass (or burst size) mostly depends on several factors including

specific phage characteristics, bacteria-infected by the phage, and the environment within which the phage-bacteria interaction occurs [46]. Contrastingly, lysogenic phages involve the integration of the genetic material of lysogenic phages into the host which results in the formation of an entity called “prophage” [47]. The prophage is involved in the vertical transmission of genetic information of the phage virus to newly formed bacteria daughter cells by cell division, and expression of viral genes and proteins [48]. Less commonly, the genetic material of lysogenic phage does not integrate itself into the host bacterial chromosome, but remains in the intracellular as a separate plasmid, until it becomes transferred to new bacterial cells. Under exceptional circumstances, it has been reported that environmental stress may likely induce a transition from a lysogenic cycle into a lytic cycle [49]. As a result of their ability to cause bacterial cell lysis, lytic phages are often used in phage therapy, because of their unique characteristics, while lysogenic phages confer anti-microbial resistance [50]. Notably, the CRISPRs consist of a diverse family of DNA repeats that all share a common architecture. Each CRISPR locus consists of a series of short repeat sequences typically containing 23–47 base pairs long separated by unique spacer sequences of a similar length (as shown in Figure 2) [50]. The repeat sequences within a CRISPR locus are conserved, and other repeat sequences in different CRISPR loci can vary in both sequence and length [51].

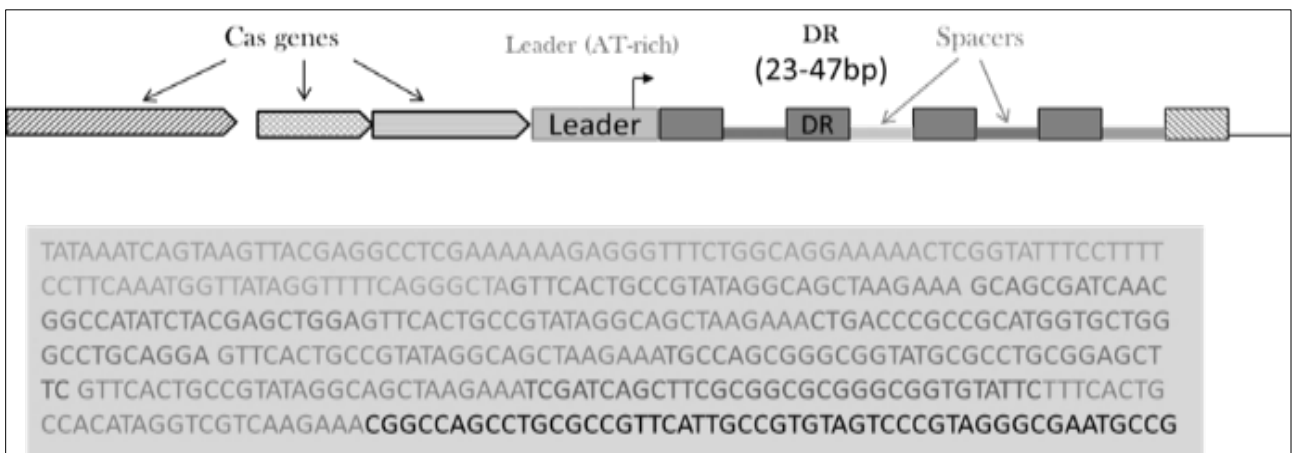


Fig 2: Structural description and components of CRISPR-cas system

The CRISPR repeat clusters were initially numbered 1 to 12, whereas the CAS systems were first designated after a representative organism, using a three-letter code, and each CAS gene was assigned a number according to its position in the CAS gene cluster (e.g., cse1, cse2) [41]. The CAS genes in the other systems were named using a similar strategy, while some of the CAS gene families were later determined to be orthologous and renamed using a “clusters of orthologous groups” classification scheme [42]. The diversity of CAS genes and their association with different CRISPR repeat clusters had also made it difficult to arrive at common nomenclatures that are easy to understand, but these pioneering phylogenetic studies were critical in establishing a basis for biochemical and mechanistic investigation [43]. The CAS gene and CRISPR repeat phylogenies are now combined in a novel categorization method [43]. Three major forms of CRISPR/Cas systems have been identified using this method, and each of these major types has been described. In addition to the leader sequence, comparative research studies have identified a range of CAS genes, which are often located close to a CRISPR locus. Initially, four CAS genes were found in genomes containing CRISPRs, but as genome sequences grew larger and more sophisticated search methods were developed, a total of 45 gene families linked with CRISPRs have been revealed [44]. Six of these CAS genes (*cas1–cas6*) are widely conserved and are considered core CAS genes, but only *cas1* and *cas2* are universally conserved in genomes that contain CRISPR loci [45]. *Cas1* is a hallmark of this immune system, and phylogenetic analysis of *cas1* sequences suggests several distinct versions of CRISPR systems exist [46]. Each of these different phenotypes is defined by a unique composition and conserved arrangement of CAS genes [47]. Remarkably, this CAS gene-based classification appears to correlate well with a CRISPR repeat-based classification, suggesting that

the *Cas* proteins interact with specific sets of CRISPR loci [47]. CRISPR-Cas systems have been classified into three major types, namely type I, type II, and type III, and 12 subtypes, given their genetic content, structural and functional differences [48]. The core defining feature of CRISPR-Cas types and subtypes are the CAS genes and the proteins they encode, which are highly genetically and functionally diverse, illustrating the many biochemical functions that they carry throughout the different steps of CRISPR-mediated immunity. Noteworthy, the RNA recognition motif is widespread in many Cas proteins, and most of the Cas families of proteins carry functional domains that interact with nucleic acids, such as DNA binding, RNA binding, helicase, and nuclease motifs [49]. Genetically, *cas1* and *cas2* universally occur across types and subtypes, whereas *cas3*, *cas9*, and *cas10* have been defined as the signature genes for type I, type II, and type III, respectively. Phylogenetic analysis has shown that type II systems have solely been identified in bacteria, thus far, and there is a bias for type I systems in bacteria and type III systems in archaea and hyper thermophiles [48, 49]. The molecular mechanism of CRISPR-Cas system is categorized into three functional phases (as shown in Figure 3) and they include; immune specificity adaptation, crRNA expression and maturation, and target interference. At one end of the CRISPR locus, fragments of foreign DNA are integrated during the adaption step. Although, *cas1* and *cas2* are known to be required for adaptation, and the process of spacer acquisition is yet unclear [48, 52]. The CRISPR array is transcribed into a lengthy precursor crRNA during the expression stage, which is then cleaved in the repetitions by specialized CAS proteins or RNase III, and occasionally trimmed to generate mature crRNAs. Subsequently, the CAS protein complex is loaded with these short guide RNAs [49].

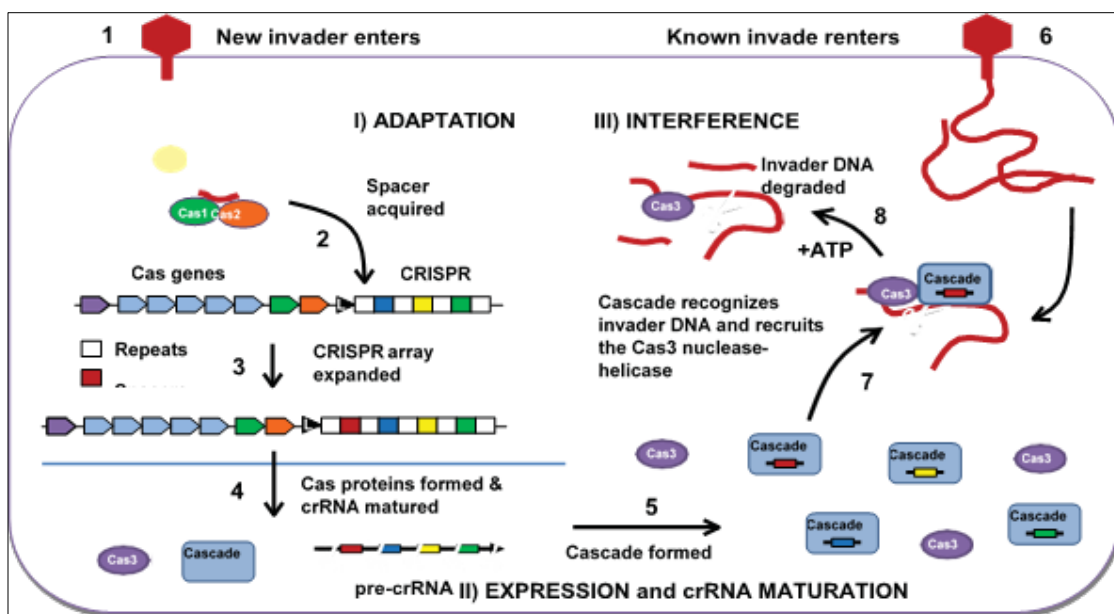


Fig 3: Three Functional Steps of CRISPR-Cas Mechanism

Yersinia pestis

Yersinia pestis is a causative agent of plague and exhibits multi-host and multi-vector pathogenic characteristics [50]. Over 200 species of this organism live in wild rodents as host organisms and over 80 species of fleas as vectors. The

disease caused by *Y. pestis*, is zoonotic infection which is regarded as one of the most devastating infections in human history and the disease is transmitted to humans from natural rodent reservoirs, usually through the bite of an infected flea [50]. Different hosts and vectors have their

specific ecological landscape and different levels of susceptibility to the organism. The classification of *Yersinia pestis* is shown in Table 2 below.

Table 2: Scientific classification of *Yersinia pestis*

Domain	Bacteria
Phylum	Proteobacteria
Class	Gammaproteobacteria
Order	Enterobacterales
Family	Yersiniaceae
Genus	<i>Yersinia</i>
Species	<i>Y. pestis</i>

In addition, the survival of the bacteria in the soil is likely to contribute to the long-term persistence of *Y. pestis*. During its expansion and adaptation to new niches, *Y. pestis* undergoes genetic variations, some of which may help overcome natural selective forces. These variations may be used as markers to reconstruct the historical spread of the plague. Because of its importance in human history, many investigations have aimed at deciphering the evolution of this major pathogen [50, 51]. In 1894, Alexandre Yersin identified this organism based on its biochemical characteristics and further employed a molecular typing technique to classify *Y. pestis* strains. Based on these characteristics, it further constituted a strong phylogenetic signal that has been used to identify and classify typical human pathogenic *Y. pestis* (subspecies *pestis* in Russian nomenclature) into three *biovars*, *bv. Antiqua*, *bv. Medievalis* and *Orientalis bv.* [50].

E. coli

Escherichia coli, commonly known as *E. coli* is a Gram-negative, facultative anaerobe and rod-shaped coliform bacteria found in the lower intestine of warm-blooded animals (endotherms) [48, 50]. Notably, the scientific classification of *E. coli* bacterium is shown in Table 3 below.

Table 3: Scientific classification of *Escherichia coli*

Domain	Bacteria
Phylum	Proteobacteria
Class	Gammaproteobacteria
Order	Enterobacterales
Family	Enterobacteriaceae
Genus	<i>Escherichia</i>
Species	<i>E. coli</i>

Although most *E. coli* strains are innocuous and categorized into different serotypes (EPEC, ETEC, and others) that may cause acute food poisoning and food contamination episodes that results to food-borne diseases. *E. coli* is the most studied prokaryotic model organism and a key species in biotechnology and microbiology, where it has served as the host organism for the majority of recombinant DNA research. Several pathogenic *E. coli* groups cause disease in humans and animals, such as diarrheagenic *E. coli* and extra-intestinal pathogenic *E. coli* (ExPEC), which cause sickness outside of the GI tract. Diarrheagenic *E. coli* that cause human sickness have been categorized based on particular sets of virulence genes they contain and the features of the disease they produce [47, 48]. These patho types include enteropathogenic *E. coli* (EPEC), enterotoxigenic *E. coli* (ETEC), enteroinvasive *E. coli* (EIEC),

enteroaggregative *E. coli* (EAEC), Shiga toxin-producing *E. coli* (STEC), diffusely adherent *E. coli* (DEAC), and adherent invasive *E. coli* (AIEC) that have been linked to Crohn's disease [34]. There are hybrid patho types, such as enteroaggregative hemorrhagic *E. coli* (EHEC), which have both STEC and EAEC-associated virulence genes. In 2011, EAHEC serotype O104:H4, an EAEC that acquired the phage carrying the Shiga toxin gene of STEC, produced a major epidemic that resulted in sickness in over 3800 people and 54 fatalities [48]. Certain *E. coli* serotypes, such as STEC O157:H7 and O103:H21 are frequently linked with certain patho types, such as enterohemorrhagic *E. coli* (EHEC). As a result, pathogenic *E. coli* is a genetically diverse family of bacteria that is continuously evolving [51, 52].

Aim of the study

The aim of this study is to analyze and compare CRISPR-cas systems of *Yersinia pestis* and *Escherichia coli* strains using bioinformatics tools.

Specific objectives

1. To identify and analyze CRISPR and spacer compositions of meta-genomic sequences of *Yersinia pestis* and *Escherichia coli* strains.
2. To determine and compare the pattern of exogenous spacer distributions of *Yersinia pestis* and *Escherichia coli* strains.
3. To conduct phylogenetic analyses of *Yersinia pestis* and *Escherichia coli* strains.

2. Materials and Method

In this study, NCBI databank, nBLAST tool, MUSCLE alignment tool, MinCED-CRISPR tools and online phylogeny server were used to conduct the CRISPR-cas analysis of *E. coli* and *Y. pestis* strains.

2.1 Bioinformatics tools

2.1.1 National Center for Biotechnology Information (NCBI) prokaryotic database/server

NCBI prokaryotic database served as a major source of data for genomic "FASTA" nucleotide sequences of *Escherichia coli* and *Yersinia pestis* strains. The first task and procedure undertaken were to specifically search and download complete genome assemblies of available "FASTA" nucleotide sequences of *E. coli* and *Yersinia pestis* strains separately from NCBI web-based prokaryotic database. The web-based prokaryotic database of the National Center for Biotechnology Information (NCBI) contains repositories of genomic nucleotide sequences, and annotated information of Bacterial strains. NCBI online prokaryotic repository is a collection or assembly of genomes, genes, taxonomic classification of bacterial strains, and archaea. Bacterial genome sequencing started by an approach made on genome analysis through sequencing and assembly of unselected pieces of DNA to get the complete nucleotide sequence of the genome from the whole chromosome in the year 1995 which led to a promising breakthrough in microbiology and infectious disease research. In addition, the National Centre for Biotechnology Information advances science and health by providing access to biomedical and genomic information. NCBI has a multi-disciplinary research group that consists of computer scientists, molecular biologists, mathematicians, and biochemists, research physicians concentrating on basic and applied research in

computational molecular biology. NCBI prokaryotic database contains a total number of 12,000 bacterial genome sequences.

2.1.2 MINCED (Mining CRISPRs in Environmental datasets) - Linux software tool

Min CED is a program used to find Clustered Regularly Interspaced Short Palindromic Repeats (CRISPRs) in full genomes or environmental datasets such as assembled contigs from meta-genomes. It identifies CRISPRs in raw short read data, in the size range of 100-200bp and this software package can be sourced from the Crass portal (<https://github.com/ctskennerton/Crass>). The MinCED software tool runs from the command-line and was derived from CRT (<http://www.room220.com/crt/>). Clustered Regularly Interspaced Palindromic Repeats (CRISPRs) are a novel type of direct repeat found in a wide range of bacteria and archaea. CRISPRs work by defending their hosts against invading extra-chromosomal elements such as viruses. The CRISPR arrays are identified using MinCED (Mining CRISPRs in Environmental Datasets), a derivative of the CRISPR Recognition Tool that is more conservative in repeat calling and allows more flexible user outputs. Custom code determines the orientation of the repeats, generates the consensus repeat sequences, and returns the number of repeats by indicating the size of the array. After the identification of CRISPR loci, the types and subtypes are assigned by using the presence or absence of genes, detecting multiple systems in a genome, and identifying the missing repeats and CAS proteins it determines the completeness of the system.

2.1.3 Multiple sequence comparison by log expectation (MusCLE) tool

MusCLE (Multiple Sequence Comparison by Log Expectation) is a program for creating multiple alignments of amino acid or nucleotide sequences. The MUSCLE web-based tool is used for multiple sequence alignment (MSA) of three or more biological sequences, generally a protein, DNA, or RNA. Multiple Sequence Alignment program has an algorithm for finding regions of similarity between biological sequences through comparing nucleotide or protein sequence from the databases and calculates the statistical significance, one of the most widely used bioinformatics programs for sequence searching. In many cases, the input set of query sequences are assumed to have an evolutionary relationship by which they share a linkage and are descended from a common ancestor. From the resulting MSA, sequence homology can be inferred, and phylogenetic analysis can be conducted to assess the sequences' shared evolutionary origins. It provides a range of options to the user for a better choice of optimizing accuracy, speed, or some compromise between the two. Default parameters are those that give the best average accuracy. Some published tests show that MUSCLE can achieve both better average accuracy and better speed than CLUSTALW or T-Coffee, depending on the chosen options. MUSCLE enables high-throughput applications to achieve average accuracy comparable to the most accurate tools previously available, which we expect to be increasingly important for in sequence data. MUSCLE has been integrated into DNASTAR's Lasergene software, Geneious, and MacVector and is available in Sequencher, MEGA, and UGENE as a plug-in. MUSCLE is also available as a

web service via the European Molecular Biology Laboratory (EMBL)-European Bioinformatics Institute (EBI).

2.1.4 Basic local alignment search tool (nBLAST)

The BLAST (basic local alignment search tool) is an algorithm and program for comparing primary biological sequence information, such as the amino-acid sequences of proteins or the nucleotides of DNA and /or RNA sequences. The 'BLAST' searches compare a subject protein or nucleotide sequence (called a query) with a library or database of sequences and identify database sequences that resemble the query sequence above a certain threshold. For example, following the discovery of a previously unknown gene in the mouse, a scientist will typically perform a BLAST search of the human genome to see if humans carry a similar gene; BLAST will identify sequences in the human genome that resemble the mouse gene based on similarity of sequence. This program, given a DNA query, returns the most similar DNA sequences from the DNA database that the user specifies. BLAST can be used for several purposes. These include identifying species, locating domains, establishing phylogeny, DNA mapping, and comparison.

Uses of nucleotide BLAST Tool

- By identifying species with the use of BLAST, you can correctly identify a species or find homologous species. This can be useful, for example, when you are working with a DNA sequence from an unknown species.
- **Locating domains:** When working with a protein sequence you can input it into BLAST, to locate known domains within the sequence of interest.
- **Establishing phylogeny:** Using the results received through BLAST you can create a phylogenetic tree using the BLAST web page. Phylogenies based on BLAST alone are less reliable than other purpose-built computational phylogenetic methods, so should only be relied upon for "first pass" phylogenetic analyses.
- **DNA mapping:** When working with a known species and looking to sequence a gene at an unknown location, BLAST can compare the chromosomal position of the sequence of interest, to relevant sequences in the database(s). NCBI has a "Magic-BLAST" tool built around BLAST for this purpose.
- **Comparison:** When working with genes, BLAST can locate common genes in two related species and can be used to map annotations from one organism to another.

2.1.5 Phylogeny tool

Phylogeny.fr is a web-designed program that operates on a web interface (www.phylogeny.fr) used to conduct phylogenetic studies of biological sample data. It has a high-performance platform that transparently chains programs relevant to phylogenetic analysis in a comprehensive and flexible pipeline. Although phylogenetic aficionados will be able to find most of their favorite tools and run sophisticated analyses. The primary aim of Phylogeny.fr web- designed program is to assist biologists with no experience in phylogeny in robustly analyzing their data. The Phylogeny.fr platform offers a *phylogeny pipeline* that can be executed through three main modes: The "*One Click mode*" targets users that do not wish to deal with program and parameter selection. By default, the pipeline is already

set up to run and connect programs recognized for their accuracy and speed (MUSCLE for multiple alignments and PhyML for phylogeny) to reconstruct a robust phylogenetic tree from a set of sequences. In the "Advanced mode", the Phylogeny.fr server proposes the succession of the same programs, but users can choose the steps to perform (multiple sequence alignment, phylogenetic reconstruction, tree drawing) and the options of each program. The "A la carte mode" offers the possibility of running and testing more alignment and phylogeny programs: MUSCLE, Clustal W, T-Coffee, Phy ML, Bio NJ, TNT. Alternatively, users can run the different programs on its web interface separately.

3. Results

3.1 Preliminary analysis of *E. coli* meta-genomic sequences

Table 4 showed the CRISPR size, Spacer compositions of *Escherichia coli* and 4 CRISPR classes (CRISPR I, II, III, and IV) were present, and randomly distributed in 162 *Escherichia coli* strains.

Table 4: CRISPR and Spacers compositions of *Escherichia coli*

CRISPR Class	Number of CRISPR	Number of Spacers
CRISPR class I	162	1,512
CRISPR class II	132	1,167
CRISPR class III	15	79
CRISPR class IV	3	9
Total	312	2,767

From Table 4, the result values obtained have shown that 162 *E. coli* strains had a CRISPR and spacer size comprising 312 CRISPR classes and 2,767 spacers accordingly. A total number of 312 CRISPR sequences were found in *E. coli* and distributed unevenly in 4 CRISPR classes of *E. coli* strains as follows: Class I (162), Class II (132), Class III (15), and Class IV (3). The total number of 2,767 spacers sequences were associated with 4 CRISPR classes of *E. coli* and distributed as follows: CRISPR class I (1,512 spacers), CRISPR Class II (1,167 spacers), CRISPR class III (79 spacers), and CRISPR class IV (9 spacers).

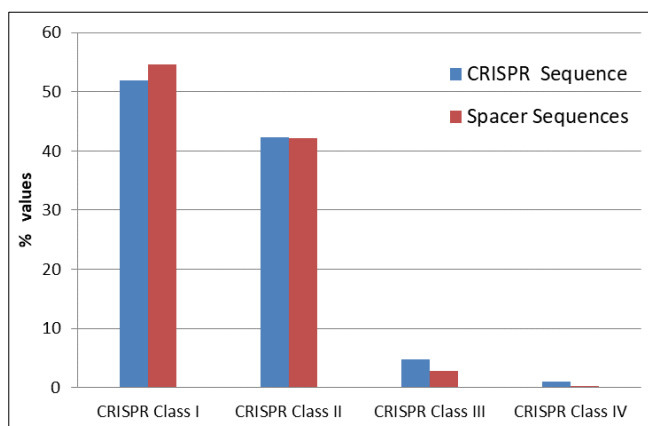


Fig 4: Percentage CRISPR and spacers compositions of *E. coli*

The percentage distribution chart of CRISPR size and spacer content of *E. coli* as shown in Figure 4 illustrates the

distribution of percentage values of CRISPR and spacer sequences among CRISPR Classes identified in 162 *E. coli* strains. The result indicated that the percentage distribution values of CRISPR and spacers among 4 CRISPR classes of *E. coli* showed that the CRISPR class I had the highest percentage value of 54.64%, followed by 42.18% CRISPR II (42.18%), CRISPR III (2.86%) and CRISPR IV (0.33%). The least percentage values were recorded mainly by CRISPR class IV and CRISPR class III, respectively. The significance of these preliminary findings explicitly indicated that 162 *E. coli* strains had 4 major CRISPR classes, with a high preponderance of CRISPRs present in class I and class II and thus suggests presence of CRISPR class I and its subtypes, and spacer content. This further explains the possibility of high CRISPR diversity in Class I and II compared with CRISPR class III and class IV.

3.1.2 Percentage distribution of exogenous spacers in *E. coli* CRISPR classes

The result presented in Table 5 showed the distribution of homologous spacer sequences derived from Bacteriophages, Plasmids, Unknown targeting elements and other bacterial strains, respectively.

Table 5: Spacer distribution of *Escherichia coli* CRISPR classes

CRISPR Class	Bacteriophages	Plasmids	Other Bacteria	Unknown targets
Class I	53	5,448	671	22
Class II	49	2,947	294	87
Class III	0	16	18	0
Class IV	0	0	0	0
Total	102	8,411	983	109

Table 5 above showed the exact values of homologous spacers matches of *E. coli* CRISPR compared to other exogenous spacers. It further showed that the total number of 102 bacteriophages spacers, 8,411 plasmid spacers, 109 spacers from unknown targets, and 983 spacers identified from bacterial strains were found in the NCBI databank.

Results interpretation

The result indicated that the CRISPR class I had a total number of 6,194 spacers, with significant portions of homologous spacers matches from bacteriophages (53 spacers), plasmids (5,448 spacers), other bacterial organisms (671 spacers), and unknown targets (22 spacers) accordingly. The CRISPR class II had a total number of 3,377 spacers, with significant distributions from bacteriophages (49 spacers), plasmids (2,947 spacers), other bacterial organisms (294 spacers), and unknown targets (87 spacers). The result further indicated that the CRISPR class III had the total number of 34 spacers which showed resemblance with plasmids (16 spacers) and other bacterial strains (18 spacers). However, no homologous spacer sequence match with these exogenous elements was identified in CRISPR class IV. Figure 5 below showed the summary of the percentage distribution chart of exogenous spacers sequence match between the spacer sequences of *E. coli* strains and exogenous elements (bacteriophages, plasmids, unknown targets, and other bacterial strains).

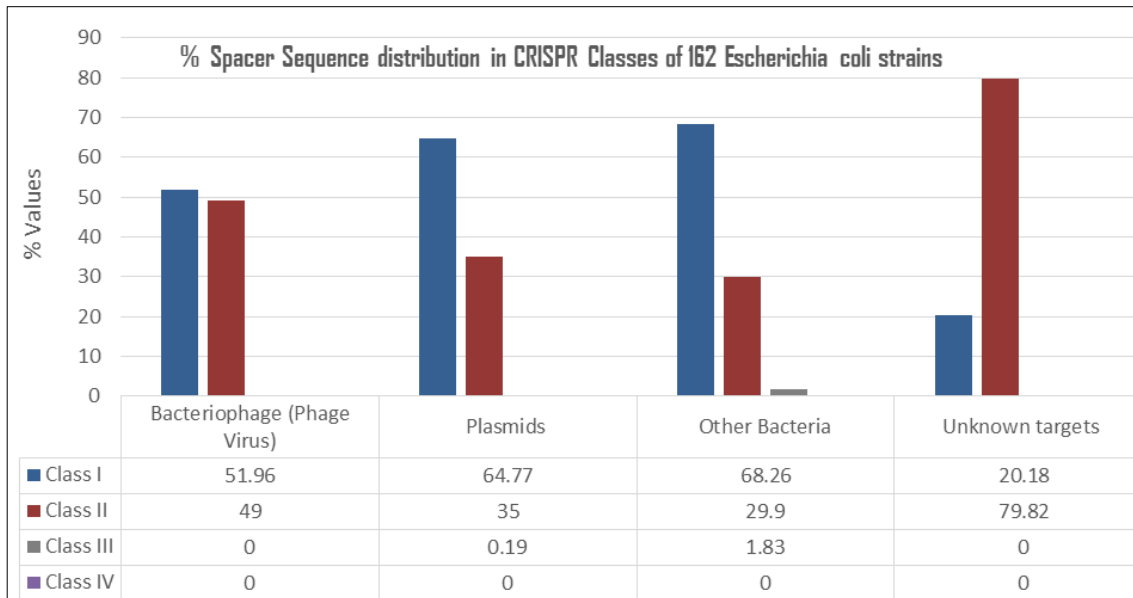


Fig 5: Percentage Spacer distribution of E coli CRISPR classes

Results Interpretation

1. **% Bacteriophages:** It further revealed that the *E. coli* CRISPR I had the highest percentage value of 52% homologous sequence matches with spacers from bacteriophages, and the *E. coli* CRISPR class II recorded 49% homologous spacers matches. However, it was observed that no homologous spacer matches were recorded in the *E. coli* CRISPR class III and IV respectively.
2. **% Plasmid:** From the above distribution chart, it showed that CRISPR class I showed the highest percentage value of 68.26% homologous sequence matches with plasmids spacers while Class II had 35%, CRISPR class III (0.19%) and CRISPR class IV showed no homologous sequence match with plasmids' spacers available in the NCBI repositories.
3. **% Exogenous bacteria:** The result indicated that *E. coli* CRISPR Class I showed a significant percentage proportion (68.26%) of homologous sequence matches with exogenous bacterial spacers, while CRISPR class II and CRISPR class III showed 29.9% and 1.83% homologous sequence matches with exogenous bacterial spacers from NCBI repository. It was also observed that CRISPR class IV showed no percentage homologous sequence matches with spacers from exogenous bacteria.
4. **% Unknown targets:** *E. coli* CRISPR class II showed 79.8% homologous sequence matches with spacers from unknown targeting elements, while CRISPR class I showed 20% homologous sequence matches with spacers from unknown targets. It was further observed that there were no homologous sequence matches recorded by *E. coli* CRISPR class III and IV with spacers from unknown targets.

The significance of this result showed that a significant variation in the patterns of spacer distribution and acquisitions by the *E. coli* CRISPR Classes. However, it was observed that *E. coli* CRISPR class I showed highest

percentage values of homologous sequence matches with spacers from bacteria, bacteriophages and plasmids, and lesser percentage homologous spacers from unknown targets. In addition, the pattern of the percentage spacer distribution of exogenous spacers in the CRISPR class II varied significantly. However, CRISPR class II significant percentage proportions of homologous spacers were derived from unknown targets. It was observed that there was a non-significant percentage proportions of homologous spacers sequence match found in the CRISPR class III as compared with CRISPR class I and class II. However, the non-significant proportions of homologous spacers sequence matches found in CRISPR class III were mainly derived from plasmid and exogenous bacterial plasmids. The CRISPR class IV showed no homologous sequence match with exogenous spacers available in the NCBI repositories. The implication of these findings thus further suggests that the *E. coli* bacterial strains could have acquired significant portions of their spacers' sequences from these exogenous elements, with highest exogenous spacers acquired by CRISPR class I, followed by CRISPR class II and CRISPR class III. It was observed that CRISPR diversity of *E. coli* strain could have been contributed by significant proportion of exogenous spacers acquired from homologous spacers often associated with plasmid, bacteria, unknown targets, and bacteriophages. The CRISPR diversity of *E. coli* might have been significantly due to incessant exposure of numerous *E. coli* bacterial strains, to infectious attack by bacteriophages, conjugation, and co-evolution of *E. coli* strains with other exogenous elements in the environment.

3.2 Preliminary analysis of *Yersinia pestis* metagenomic sequences

Table 6 showed the CRISPR and Spacer compositions of *Yersinia pestis*. The result showed that 3 CRISPR classes namely CRISPR I, II and III were present while CRISPR-spacer of Class IV was not found in 121 *Yersinia pestis* strains. The preliminary result of the meta-genomic CRISPR analysis of 121 *Yersinia pestis* were tabulated as shown in Table 6.

Table 6: CRISPR and spacer compositions of *Yersinia pestis* classes

CRISPR classes	N Number of CRISPR	Number of Spacers
CRISPR class I	121	554
CRISPR class II	110	617
CRISPR class III	71	329
CRISPR class IV	0	0
Total	302	1,500

From Table 6, the result values of 121 *Yersinia pestis* strains showed 302 CRISPR size and 1,500 spacer contents accordingly. A total number of 302 CRISPR sequences were distributed in 3 CRISPR classes of *Y. pestis* strains as follows: Class I (121), Class II (110), Class III (71), and Class IV (0). Total numbers of 2,767 spacers' sequences were associated with 3 CRISPR classes of *Yersinia pestis*, and their distributions are as follows: CRISPR class II (617 spacers), CRISPR Class I (554 spacers) and CRISPR class III (329 spacers). It is showed that there was absent of CRISPR class IV which further suggests absence of CRISPR and spacer sequences.

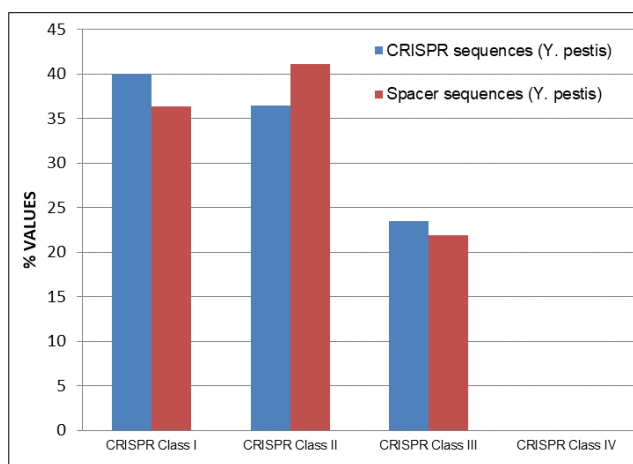


Fig 6: Percentage CRISPR and spacer compositions of *Yersinia pestis* classes

Figure 6 showed the percentage distribution chart of CRISPR size and spacer content of *Y. pestis*. The result indicated that the percentage distribution values of CRISPR and spacers in 3 CRISPR classes of *Yersinia pestis* showed that the CRISPR class I had the highest percentage value of 40%, followed by CRISPR class II (36.9%), CRISPR class III (21.93%) and CRISPR class IV (0%). The least percentage values were recorded mainly by CRISPR class IV and CRISPR class III, respectively. The significance of these findings further shown that a high percentage proportions of CRISPR sequences composition was found in CRISPR class I than CRISPR classes II and III respectively. It further indicated that the CRISPR class I had the least CRISPR size than CRISPR class I and class II, while

CRISPR class III. The findings from the above results have remarkably showed a significant difference in CRISPR sizes, regarding the number of CRISPR classes, CRISPR sequences and spacer contents compared with *E. coli* CRISPR-cas system.

3.2 Percentage distribution of exogenous spacers in *Yersinia pestis* CRISPR classes.

Table 7 showed the exact distribution of homologous spacer sequences from Bacteriophages, Plasmids, Unknown targeting elements and other bacterial strains, respectively. The results of homologous spacers sequences were identified from bacteriophages, plasmids, bacterial strains, and unknown targets. The preliminary data were computed and tabulated as shown in Table 7 below.

Table 7: Spacer distribution of *Yersinia pestis* CRISPR classes

CRISPR Class	Bacteriophages	Plasmids	Other Bacteria	Unknown targets
Class I	50	149	1654	0
Class II	27	98	1699	0
Class III	28	138	949	0
Class IV	0	0	0	0
Total	105	385	4,302	0

Table 7 above showed the exact values of homologous spacers matches between *Y. pestis* spacers and other spacers from exogenous elements. It further indicated that a total number of 105 bacteriophages spacers, 385 plasmid spacers and 4,302 bacterial spacers, while no homologous spacer sequence matches were found with any unknown targets.

Results interpretation

The result indicated that the CRISPR class I had a total number of 1,853 spacers, with significant portions of homologous spacers matches identified from bacteriophages (50 spacers), plasmids (149 spacers), other bacterial organisms (1,654 spacers), and no spacer sequence were identified with unknown targets accordingly. The CRISPR class II had a total number of 1,824 spacers, with significant distributions from bacteriophages (27 spacers), plasmids (98 spacers), other bacterial organisms (1699 spacers), and no spacers from unknown targets. CRISPR class III had a total number of 1,115 spacers, with some portions of its homologous spacers from plasmids (138 spacers), bacteriophages (28 spacers) and other bacterial strains (949 spacers). However, no homologous spacer sequence match with any unknown elements was identified in CRISPR class IV.

Figure 7 below showed the summary of the percentage distribution chart of exogenous spacers sequence match between the spacer sequences of *Y. pestis* strains and exogenous elements (bacteriophages, plasmids, unknown targets, and other bacterial strains).

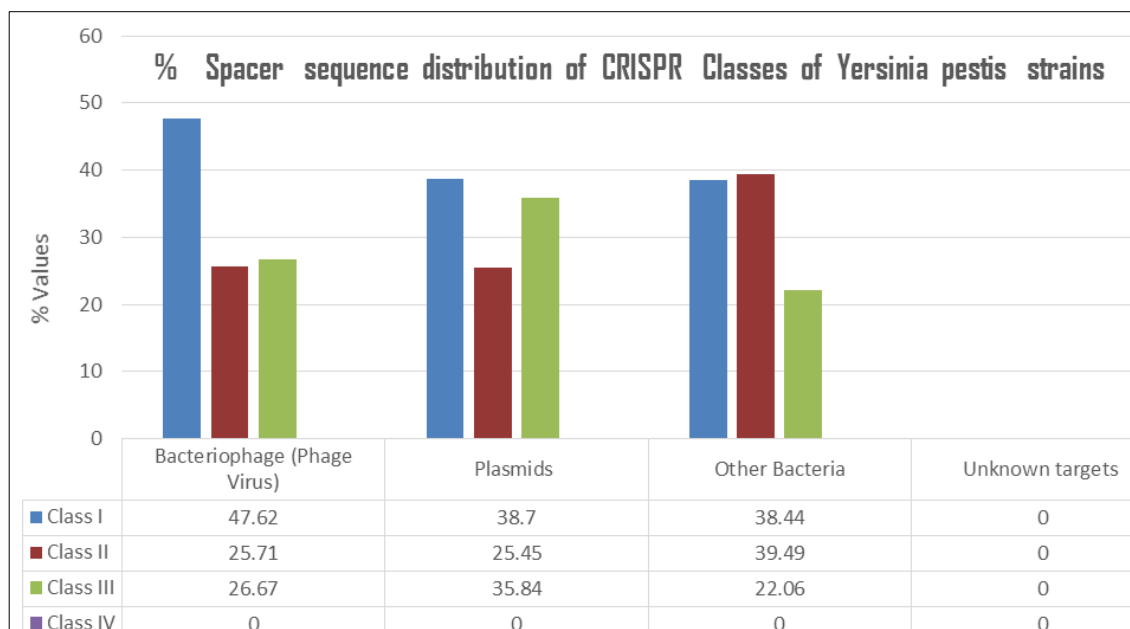


Fig 7: Percentage spacer distribution of exogenous spacers in *Y. pestis* CRISPR classes

Results Interpretation

1. **% Bacteriophages:** It further revealed that the CRISPR I had the highest percentage value (47.6%) of homologous sequence matches with spacers from bacteriophages, and the CRISPR class III (26.7%) and CRISPR class II (25.7%) homologous spacers matches. However, it was observed that no homologous spacer matches were recorded in the *Yersinia pestis* CRISPR class IV respectively.
2. **% Plasmid:** From the above distribution chart, it showed that CRISPR class I showed the highest percentage value (38.7%) homologous sequence matches with plasmids spacers compared with the CRISPR Class III (35.8%) and CRISPR class II (25.5%). CRISPR class IV showed no homologous sequence match with plasmids' spacers available in the NCBI repositories.
3. **% Exogenous bacteria:** The result indicated that CRISPR Class II showed a significant percentage proportion (39.49%) of homologous sequence matches with exogenous bacterial spacers, while CRISPR class I and CRISPR class III showed 38% and 22% homologous sequence matches with exogenous bacterial spacers from NCBI repository.
4. **% Unknown targets:** there was no homologous spacer sequence match found in any of the *Y. pestis* CRISPR classes.

The significance of this result showed significant variations in the pattern of spacer distribution and acquisitions by *Y. pestis* CRISPR Classes. However, it was observed that *Y. pestis* CRISPR class I showed highest percentage values of homologous sequence matches with spacers from bacteria, bacteriophages, and plasmids. In addition, the pattern of the percentage spacer distribution of exogenous spacers in the

CRISPR class II varied significantly. However, CRISPR class II contributed significant percentage proportions of homologous spacers derived mainly from exogenous bacteria. It was observed that CRISPR class III spacers derived some proportions of its spacers from plasmids, bacteriophages, and other bacterial plasmids, with absence of homologous spacers from unknown targets.

The finding suggests that *Y. pestis* bacterial strains might have acquired significant portions of their spacers' sequences from these exogenous elements excluding unknown targets. The significant proportion of spacers acquired by the bacteria was contributed by CRISPR class I, and CRISPR class II, while CRISPR class III had a least percentage spacer distribution. A comparison of the CRISPR size and spacer content of *Y. pestis* have clearly showed that *Y. pestis* CRISPR diversity is lesser compared with *E. coli* CRISPR systems, in terms of its CRISPR size, spacer sequence content and CRISPR class. The spacer distributions of *E. coli* strains were widely contributed by exogenous plasmids, bacteriophages, bacteria, and unknown target, whereas the percentage proportions of spacers acquired by *Y. pestis* CRISPRs are mainly derived from plasmids, bacteriophages and bacterial spacers only. This further implied that there is tendency of increased CRISPR diversity in *E. coli* strains than *Y. pestis*, despite the fact that both bacterial species are constantly exposed to infectious attack from bacteriophages and conjugation, and co-evolution of *E. coli* strains from other exogenous elements in the environment.

3.3 Phylogenetic analyses of *Escherichia coli* and *Yersinia pestis* strains

3.3.1 Phylogenetic analysis of *Escherichia coli* strains

The phylogenetic tree showing the evolutionary relationships of 22 representative strains of *Escherichia coli* as obtained from the *phylogeny*, a web-server bioinformatics tool, and the result is presented (Figure 8) below;

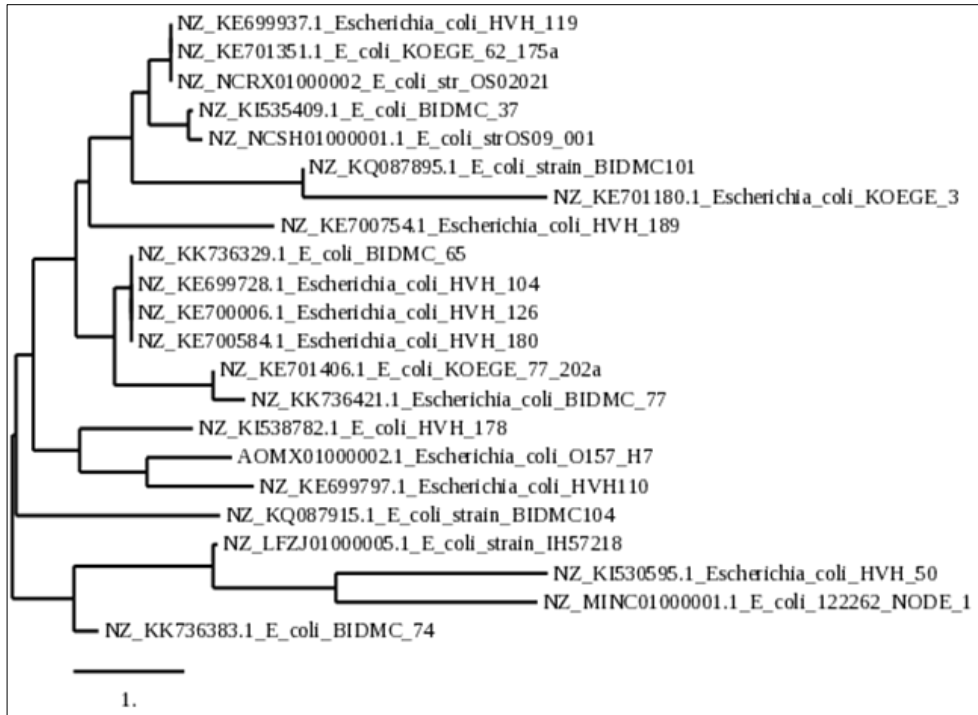


Fig 8: Phylogenetic tree of *E. coli* strains

From figure 8, the result showed the phylogenetic relationship of 22 representatives of 162 *E. coli* strains. The analysis of the result of the dendrogram clearly showed that 22 representative strains had a common ancestral origin from *E. coli* BIDMC 74 (Accessioned: NZ_KK736383.1). *E. coli* BIDMC_74 strain showed a distant relationship from three *E. coli* bacterial strains (*E. coli* HVH str. 119, *E. coli* KOEGE 62[175a], and *E. coli* strain OSO2021). In Figure 8, it further showed that 19 strains (out of 22 *E. coli* strains) had a common ancestral origin with *E. coli* BIDMC_74, despite showing a distant relationship with *E. coli* BIDMC_74 because of divergent roots and branches linking 19 strains together with their origin. *E. coli* BIMC_74 strain is closely related to *E. coli* strain IH57218 compared with two other bacterial strains (*E. coli* str. HVH 50 and *E. coli* str. 122262 NODE_1). However, *E. coli* strain_H57218 was closely related with sister taxa (*E. coli*

strains HVH_50 and *E. coli* NODE_1). It was equally found that a sister taxa comprising four closely related *E. coli* strains (*E. coli* str. BIDMC_65, *E. coli* str. HVH 104, *E. coli* 126, and *E. coli* 180) and another sister taxa slightly related 2 *E. coli* strains (*E. coli* str. KOEGE 77[202a] and *E. coli* str. BIDMC 77) showed a close evolutionary relationship due to presence of a common root linking them together in the phylogenetic tree. The phylogenetic tree showed presence of divergent branches and roots that separately linked some sister taxa (groups) of *E. coli* strains together, despite having a common ancestral origin.

3.3.2 Phylogenetic analysis of *Yersinia pestis* strains

The result showed that the phylogenetic tree of the 22 representative strains of this bacterial species were obtained, and was presented in figure 9 below.

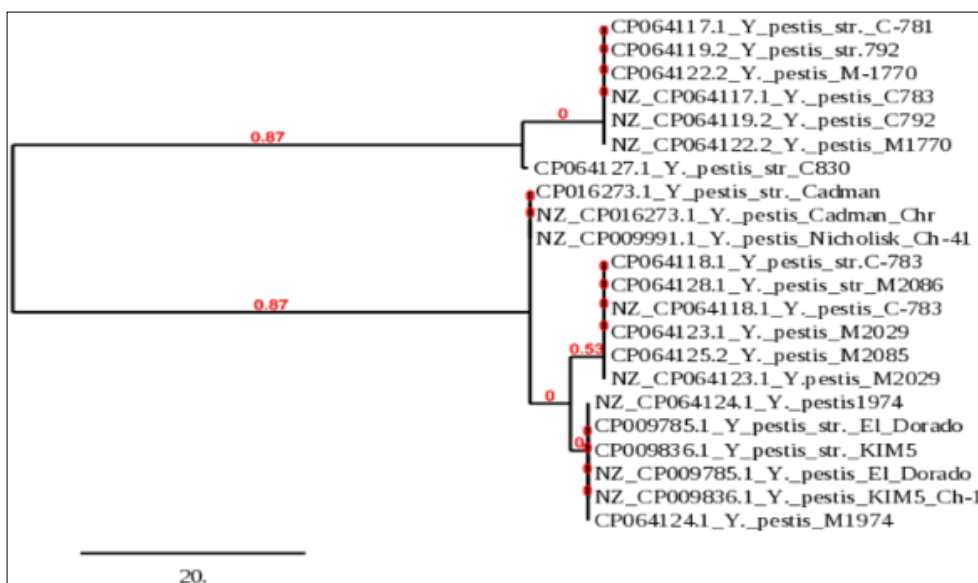


Fig 9: Phylogenetic tree of *Yersinia pestis* strains

Figure 9 showed the phylogenetic tree of 22 representatives of 121 *Yersinia pestis* strains, which were compared together, to elucidate their evolutionary relationships of *Y. pestis* strains. The result showed that the 22 *Yersinia pestis* strains were closely related and could have originated from any of these *Y. pestis* strains [*Y. pestis* str. C830, *Y. pestis* str. Cadman, *Y. pestis* str. sub species Cadman Chr_1 and *Y. pestis* str. Chr_41] showed in the phylogenetic tree. It was observed that 22 representatives were more closely related, because they had a common root of origin. Thus, the close relationship existed among different strains belonging to each group as categorized separately: group A (*Yersinia pestis* str. C.78, *Y. pestis* str. 792, *Y. pestis* str. M1770, *Y. pestis* str. C.792, *Y. pestis* str. C. 783 and *Y. pestis* str.M.170); group B (*Y. pestis* Cadman, *Y. pestis* Cadman Chr_1 and *Y. pestis* Nicholisk chr_41, group C (*Y. pestis* str. C.783, *Y. pestis* str. M2086, *Y. pestis* str. subsp. C.783, *Y. pestis* str. M2085 and *Y. pestis* str. M2029), group D (*Y.*

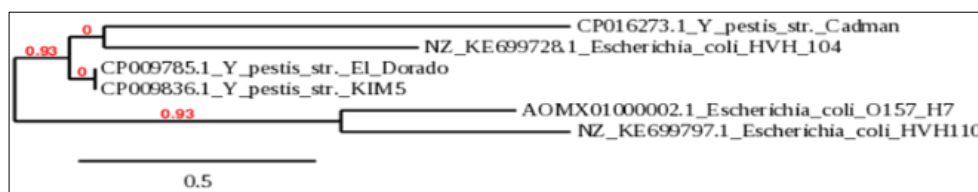


Fig 10: Phylogenetic tree of *Yersinia pestis* and *E. coli* strains

The result of comparative phylogenetic analyses of six bacterial strains comprising 3 *Yersinia pestis* and 3 *E. coli* as showed in the Figure 10. There was a close relationship existing separately within the three sister taxa (strains belonging to each group), comprising of Group A {*Y. pestis* str. Cadman, *E. coli* str. HVH_104}, Group B {*Y. pestis* str. Eldorado and *Y. pestis* str. KIM5} and Group C {*E. coli* str. O157H7 and *E. coli* HVH10}. The result of the phylogenetic tree (as shown in Figure 10) indicated a close relationship existing among the groups: A, B and C, which include these bacterial strains {*Y. pestis* str. Cadman [CP016273.1], *Y. pestis* str. El Dorado [CP009785.1], *Y. pestis* str. KIM5 [CP009836.1], *Escherichia coli* str. HVH 104 [NZ_KE699728.1], *E. coli* O157_H7 [AOMX01000002.1] and *E. coli* str. HVH110 [NZ_KE699797.1]} because they had same root of origin. However, it is observed that *Escherichia coli* strain HVH_104 and *Y. pestis* strain Cadman (Group A) is more closely related to Group B { *Y. pestis* str. El_Dorado and *Y. pestis* str. KIM5} than Group C {*E. coli* str. O157H7 and *E. coli* HVH10} respectively. Although, the result of this study revealed that all these bacterial strains shared close evolutionary origin, despite a slight phylogenetic distance between group A and group C.

4. Discussion

The results findings from this research study have compared the CRISPR –CAS systems of *Yersinia pestis* and *Escherichia coli* strains. Significantly, the result of MinCED-CRISPR analysis showed that *E. coli* strain had 4 CRISPR classes (I, II, III and IV), and subsequent analysis of its CRISPR size and spacer content of 162 *E. coli* nucleotide genomes further revealed that a total of 312 CRISPR and 2,767 spacer sequences were present in 162 *E. coli* meta-genomic data sequences, with distribution of CRISPR Class I (162 CRISPR sequences), Class II (132 sequences), Class III (15 sequences), while the CRISPR

pestis str. 1974, *Y. pestis* str. El Dorado, *Y. pestis* str. KIM5, *Y. pestis* KIM5 subsp. Chr 1 and *Y. pestis* str. M1974).

It is shown that there is a slight evolutionary relationship between *Y. pestis* strains belonging to group C and group D with a phylogenetic distance { $d= 0.53$ }, while groups: A, B and D are more closely related respectively. Since the 22 representative strains showed a close relationship with one another as compared to *E. coli* strains, it was observed that there were few branches, except groups C and D though both groups shared same root of origin. It can further be elucidated that spacer sequences of CRISPR class I genes of *Y. pestis* strains showed a significant close relationship with one another than *E. coli* strains. Thus, it suggests that there is little or no significant CRISPR diversity in CRISPR class I spacers of 22 *Yersinia pestis* strains which differ from *E. coli* strains.

3.3.3 Comparative phylogenetic analysis of *Yersinia pestis* and *Escherichia coli* species

Class IV had only 3 CRISPR sequences. The percentage CRISPR size and spacer content showed that the CRISPR class I had the highest percentage value of 54.64%, followed by 42.18% CRISPR II (42.18%), CRISPR III (2.86%) and CRISPR IV (0.33%). The least percentage values were recorded mainly by CRISPR class IV and CRISPR class III, respectively. The significance of these preliminary findings explicitly indicated that 162 *E. coli* strains had 4 major CRISPR classes, with a high percentage composition of CRISPR sequences present in class I and class II, which further suggests that the CRISPR diversity of *E. coli* is largely contributed by CRISPR class I and II and their spacer contents. This finding therefore suggests that the CRISPR diversity of *E. coli* is largely contributed by Class I and II, when compared with CRISPR class III and class IV. Contrarily, the result of the MinCED CRISPR analysis that determined CRISPR size and spacer content of 121 meta-genomic sequences of *Y. pestis*, however, identified mainly 3 CRISPR classes comprising Class I, II and Class III. The result further indicated that a total number of 302 CRISPR sequences were distributed in 3 CRISPR classes of *Y. pestis* strains as follows: Class I (121), Class II (110), Class III (71), and Class IV (0). The percentage values of CRISPR and spacer sequence contents of *Y. pestis* strains also revealed that the CRISPR class II had the highest CRISPR size (41.13%), while CRISPR class I (36.9%), CRISPR class III (21.93%) and CRISPR class IV (0%). The least percentage value was recorded by CRISPR class III, while no CRISPR sequences and spacers were present in CRISPR class IV, respectively. The total number of 2,767 spacers' sequences associated with 3 CRISPR Classes of *Yersinia pestis*, and their distributions showed that the CRISPR class II had 617 spacers, CRISPR Class I (554 spacers) and CRISPR class III (329 spacers), whereas CRISPR class IV showed absence of CRISPR and spacer sequences. The significance of this findings showed that *Y. pestis* had a

lesser CRISPR size and spacer sequence content compared with *E. coli* strains. The results of percentage distribution of exogenous spacers from plasmids (Protospacers), bacteriophages (phage virus), bacterial species, and non-prophagic chromosomal regions (non-targets) found in *Yersinia pestis* and *Escherichia coli* strains. The result showed that *E. coli* and *Y. pestis* had a total number of 9,605 and 4,792 homologous spacers sequences from exogenous elements were found in the NCBI data bank, excluding other CRISPR spacers. The overall percentage proportions of homologous spacers from exogenous elements, in relation to 162 *E. coli* strain indicated the percentage values of spacers sequences acquired from plasmids (87.56%), bacteriophages (1.06%), bacterial species (10.2%), and unknown targets (1.13%) in all CRISPR classes identified. It further showed that the percentage distribution of these spacers among *E. coli* CRISPR classes revealed that a significant percentage proportion of bacteriophages (51.96%) was found in CRISPR class I, 49% (CRISPR class II) while CRISPR III and IV had no spacers from bacteriophages. The percentage distribution of plasmids spacers also showed that 64.77% (CRISPR class I), 35% (CRISPR class II), 0.19% (CRISPR class III) while CRISPR IV had no spacers from plasmids. The percentage distribution of bacterial spacers: 68.26% (CRISPR class I), 29.9% (CRISPR class II), 1.83% (CRISPR class III), and no spacers were found in CRISPR class IV. In addition, it was found that the percentage distribution of unknown targets' spacers showed that 79.82% (CRISPR class II), 20.18% (CRISPR class I), and no spacers (CRISPR class III and IV). However, the total percentage contribution of homologous spacers from exogenous elements, in relation to 121 *Y. pestis* spacers indicated that 2.19% spacers were acquired from bacteriophages, 8% from plasmids and 89.8% bacteria. However, it was observed that there was no spacer sequence contributed by unknown targeting elements. The percentage spacer distributions across the 3 CRISPR classes showed as following:

% Bacteriophages: CRISPR I had the highest percentage value (47.6%) of homologous sequence matches with spacers from bacteriophages, CRISPR class III (26.7%) and CRISPR class II (25.7%). However, it was also observed that no homologous spacer matches were found in the *E. coli* CRISPR class III and IV respectively.

% Plasmid: CRISPR class I showed the highest percentage value (38.7%) homologous sequence, while CRISPR Class III (35.8%) and CRISPR class II (25.5%). Similarly, CRISPR class IV showed no homologous sequence match with plasmids' spacers available in the NCBI repositories. The percentage spacer distribution of *Exogenous bacteria* indicated that the CRISPR Class II showed a significant percentage proportion (39.49%) of homologous sequence matches with exogenous bacterial spacers, while CRISPR class I and CRISPR class III showed 38% and 22% homologous sequence matches with exogenous bacterial spacers from NCBI repository. Also, it was showed that unknown target elements showed no homologous spacer sequence match with all *Y. pestis* CRISPR classes.

The phylogenetic analysis of the dendrogram clearly showed that 22 representative strains had a common ancestral origin from *E. coli* BIDMC 74 (*Accessionid:* NZ_KK736383.1). The phylogenetic tree showed presence of divergent

branches and roots that separately linked some sister taxa (groups) of *E. coli* strains together, despite having a common ancestral origin. This however showed that there is increasingly huge diversity in the evolutionary relationships of CRISPR spacers of 22 *E. coli* strains.

Analysis of the phylogenetic tree of 22 representatives of 121 *Yersinia pestis* strains further elucidated that there was a close evolutionary relationship among strains, because each strain was more closely related to one another. However, it can be showed that all *Y. pestis* strains had a closer linkage with *Y. pestis* strains [*Y. pestis* str. C830, *Y. pestis* str. Cadman, *Y. pestis* str. sub species Cadman Chr_1 and *Y. pestis* str. Chr_41]. More so, it was observed that there was a slight evolutionary relationship between *Y. pestis* strains belonging to group C and group D with a phylogenetic distance { $d=0.53$ }, while groups: A, B and D are more closely related respectively. It showed that there is an evolutionary relationship between *Y. pestis* strains belonging to group A and group D, while group C differs slightly from group A and D respectively. Despite having the presence of common ancestral origin from *Yersinia pestis* strain Cadman, few branching and phylogenetic distance, groups C and D shared the same root of origin which further implies that CRISPR class I genes of *Y. pestis* strains from these two groups had a common ancestral origin. It indicates that there is no wide diversity in CRISPR spacers of 22 *Yersinia pestis* strains. The comparative phylogenetic study of the combined strains showed a close relationship within each of the three sister taxa comprising of Group A {*Y. pestis* str. Cadman, *E. coli* str. HVH_104}, Group B {*Y. pestis* str. El_Dorado and *Y. pestis* str. KIM5} and Group C {*E. coli* str. 0157H7 and *E. coli* HVH10}. However, it is observed that *Escherichia coli* strain HVH_104 and *Y. pestis* strain Cadman (Group A) is more closely related to Group B {*Y. pestis* str. El_Dorado and *Y. pestis* str. KIM5} than Group C {*E. coli* str. 0157H7 and *E. coli* HVH10} respectively. Although, the result obtained from this study further revealed that all the six bacterial strains shared a close evolutionary relationship, in spite of the slight phylogenetic distance { $d=0.93$ } between group A and group C. The findings from the phylogenetic analyses revealed have shown that there is an existing common evolutionary relationship between the CRISPR-Cas of *Yersinia pestis* and *Escherichia coli* strains.

5. Conclusion

The comparison of overall spacer sequence distribution of the CRISPR classes of 162 *E. coli* strain and 121 *Y. pestis* strains further showed that a significant percentage proportion of homologous spacers were significantly contributed by plasmids (87.56%) in *E. coli* than *Y. pestis* strains (8%). Contrarily, exogenous bacterial spacers also contributed largely to CRISPR diversity of *Y. pestis* (89%) than *E. coli* (10.2%). The percentage distribution of bacteriophages spacers in *E. coli* (2.19%) and *Y. pestis* (1.06%) is not significantly different between two bacterial species. It is equally observed that the unknown target elements contributed a little percentage proportion (1.1%) in *E. coli* and contributed no spacers in *Y. pestis* strains. The significance of these results showed that there is significant variations in the pattern of spacer distribution and acquisitions in both *E. coli* and *Y. pestis* CRISPR Classes. However, *E. coli* CRISPR showed a wide CRISPR diversity than *Y. pestis*, in terms of its CRISPR size, spacer contents

and phylogenetic relationship. In conclusion, the findings from the phylogenetic studies have indicated that the CRISPR spacers of the two different bacterial species, *Yersinia pestis* and *E. coli* strains are closely related to each other, and the significant difference associated with the evolutionary relationships of these bacterial strains may be due to the pattern of CRISPR spacer acquisition by each bacterial strain from exogenous mobile elements. The findings from this study suggest the spacer sequences are derived from DNA fragments of bacteriophages and other exogenous elements that had previously infected the prokaryote and could be used to detect and destroy DNA from similar bacteriophages during subsequent infections. The findings thus showed that the increased CRISPR diversity in *E. coli* strains could be associated with increased exposure of its strains to bacteriophages' attacks, conjugation and co-evolution with other mobile exogenous elements like plasmid vectors, unknown targets and numerous bacterial strains in the environment and their host organisms.

6. References

1. Abedon ST, Kuhl SJ, Blasdel BG, *et al.* Phage treatment of human infections. *Bacteriophage*. 2011;1:66-85.
2. Barrangou R. The roles of CRISPR-Cas systems in adaptive immunity and beyond, *Current Opinion in Immunology*. 2015;32:36-41.
3. Barrangou R, Fremaux C, Deveau H, Richards M, Boyaval P, Moineau S, *et al.* CRISPR provides acquired resistance against viruses in prokaryotes. *Science*. 2007;315(5819):1709-1712.
4. Bhaya D, Davison M, Barrangou R. CRISPR-Cas systems in bacteria and archaea: versatile small RNAs for adaptive defense and regulation. *Annu. Rev. Genet.* 2011;45:273-297.
5. Bolotin A, Quinquis B, Sorokin A, Ehrlich SD. Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extra chromosomal origin. *Microbiology*. 2005;151:2551-2561.
6. Brabban AD, Hite E, Callaway TR. Evolution of foodborne pathogens via temperate bacteriophage-mediated gene transfer. *Food borne Pathog. Dis.* 2005;2:287-303.
7. Brüssow H, Canchaya C, Hardt WD. Phages and the evolution of bacterial pathogens: from genomic rearrangements to lysogenic conversion. *Microbiology and molecular biology reviews: MMBR*. 2004;68(3):560-602. <https://doi.org/10.1128/MMBR.68.3.560-602.2004>.
8. Bult CJ, White O, Olsen GJ, Zhou L, Fleischmann RD, Sutton GG, *et al.* Complete genome sequence of the *methanogenic archaeon, Methanococcus jannaschii*. *Science*. 1996;273:1058-1073.
9. Carlton RM. Phage therapy: past history and future prospects. *Arch Immunol Ther Exp*. 1999;47:267-274.
10. Chanishvili N. Phage therapy-history from Twort and d'Herelle through Soviet experience to current approaches. *Adv Virus Res*. 2012;83:3-40.
11. Chanishvili N. Bacteriophages as therapeutic and prophylactic means: summary of the Soviet and Post-Soviet experiences. *Curr Drug Deliv*. 2016;13:309-323.
12. Clokie MR, Millard AD, Letarov AV, Heaphy S. Phages in nature. *Bacteriophage*. 2011;1(1):31-45. <https://doi.org/10.4161/bact.1.1.14942>
13. DeBoy RT, Mongodin EF, Emerson JB, Nelson KE. Chromosome evolution in the Thermotogales: large-scale inversions and strain diversification of CRISPR sequences. *J Bacteriol*. 2006;188:2364-2374.
14. Delbrück M. The growth of bacteriophage and lysis of the host, *J Gen Physiol*. 1940;23:643-660.
15. d'Herelle F. Bacteriophage as a treatment in acute medical and surgical infections. *Bull N Y Acad Med*. 1931;7:329-348.
16. East-Seletsky A, O'Connell M, Knight S, *et al.* Two distinct RNase activities of CRISPR-C2c2 enable guide-RNA processing and RNA detection. *Nature*. 2016;538:270-273. <https://doi.org/10.1038/nature19802>.
17. Godde JS, Bickerton A. The repetitive DNA elements called CRISPRs and their associated genes: evidence of horizontal transfer among prokaryotes. *J Mol. Evol.* 2006;62:718-29.
18. Grissa I, Vergnaud G, Pourcel C. The CRISPRdb database and tools to display CRISPRs and to generate dictionaries of spacers and repeats, *BMC Bioinformatics*. 2007;8:172. DOI: 10.1186/1471-2105-8-172.
19. Haft DH, Selengut J, Mongodin EF, Nelson KE. A Guild of 45 CRISPR-Associated (Cas) Protein Families and Multiple CRISPR/Cas Subtypes Exist in Prokaryotic Genomes. *PLoS Comput Biol*. 2005;1:e60.
20. Hale CR, Majumdar S, Elmore J, Pfister N, Compton M, *et al.* Essential features and rational design of CRISPR RNAs that function with the Cas RAMP module complex to cleave RNAs. *Mol. Cell*. 2012;45:292-302.
21. Hille F, Richter H, Wong SP, Bratovič M, Ressel S, Charpentier E. The Biology of CRISPR-Cas: Backward and Forward Cell. 2018;172(6):1239-1259.
22. Iftexhar Bin Naser, Mozammel Hoque M, Ausrafuggaman Nahid M, Tokee Tareq M, Kamruzzaman Rocky M, Shah Faruque M. Analysis of the CRISPR-Cas system in bacteriophages active on epidemic strains of *Vibrio cholera* in Bangladesh, *Scientific Reports*. 2017;7:14880. DOI: 10.1038/s41598-017-14839-2.
23. Jansen, R, van Embden, JD, Gaastra, W, Schouls LM. Identification of a novel family of sequence repeats among prokaryotes. *Omics*. 2002;6:23-33.
24. Jansen R, Embden JD, Gaastra W, Schouls LM. Identification of genes are associated with DNA repeats in prokaryotes. *Mol. Microbiol*. 2002;43:1565-1575.
25. Kunin V, Sorek R, Hugenholtz P. Evolutionary conservation of sequence and secondary structures in CRISPR repeats. *Genome Biol*. 2007;8:R61.
26. Labrie SJ, Samson JE, Moineau S. Bacteriophage resistance mechanisms, *Nat. Rev. Microbiol*. 2010;8:317-327.
27. Louwen R, Staals RHJ, Endtz HP, van Baarlen P, van der Oost J. The role of CRISPR-Cas systems in virulence of pathogenic bacteria, *Microbiol. Mol. Biol. Rev.* 2014;78:74-88. DOI: 10.1128/MMBR.00039-13.
28. Makarova KS, Grishin NV, Shabalina SA, Wolf YI, Koonin EV. A putative RNA-interference-based immune system in prokaryotes: computational analysis

- of the predicted enzymatic machinery, functional analogies with eukaryotic RNAi, and hypothetical mechanisms of action. *Biol Direct*. 2006;1:7.
29. Makarova KS, Wolf YI, van der Oost J, Koonin EV. Prokaryotic homologs of Argonaute proteins are predicted to function as key components of a novel system of defense against mobile genetic elements. *Biol. Direct*. 2009;4:29.
 30. Makarova KS, Haft DH, Barrangou R, Brouns SJ, Charpentier E, *et al.* Evolution and classification of the CRISPR-Cas systems. *Nat. Rev. Microbiol*. 2011;9:467-77.
 31. Marraffini LA, Sontheimer EJ. CRISPR interference limits horizontal gene transfer in staphylococci by targeting DNA. *Science*. 2008;322(5909):1843-1845.
 32. Marraffini LA, Sontheimer EJ. CRISPR interference: RNA-directed adaptive immunity in bacteria and archaea. *Nat. Rev. Genet*, 2010, 181-190.
 33. Moineau S, Stanley M, Kelly Hughes. Bacteriophage In Brenner's Encyclopedia of genetics (Second Edition), Academic Press, 2013, 280-283pp. ISBN 9780080961569, <https://doi.org/10.1016/B978-0-12-374984-0.00131-5>.
 34. Mojica FJM, Díez-Villasenor C, García-Martínez J, Almendros C. Short motif sequences determine the targets of the prokaryotic CRISPR defence system. *Microbiology*. 2009;155:733-740.
 35. Mojica FJ, Díez-Villasenor C, Soria E, Juez G. Biological significance of a family of regularly spaced repeats in the genomes of Archaea, Bacteria and mitochondria. *Mol Microbiol*. 2000;36:244-246.
 36. Ochman H, Selander RK. Standard reference strains of *Escherichia coli* from natural populations. *J Bacteriol*. 1984;157:690-693.
 37. Pougach K, Semenova E, Bogdanova E, Datsenko KA, Djordjevic M, *et al.* Transcription, processing and function of CRISPR cassettes in *Escherichia coli*. *Mol. Microbiol*. 2010;77:1367-79.
 38. Pourcel C, Salvignol G, Vergnaud G. CRISPR elements in *Yersinia pestis* acquire new repeats by preferential uptake of bacteriophage DNA and provide additional tools for evolutionary studies. *Microbiology*. 2005;151:653-663.
 39. Pul U, Wurm R, Arslan Z, Geissen R, Hofmann N, Wagner R. Identification and characterization of *E. coli* CRISPR-*cas* promoters and their silencing by H-NS. *Mol. Microbiol*. 2010;75:1495-512.
 40. Rakhuba DV, Kolomiets EI, Dey ES, *et al.* Bacteriophage receptors, mechanisms of phage adsorption and penetration into host cell. *Pol J Microbiol*. 2010;59:145-155.
 41. Rohde C, Wittmann J, Kutter E. Bacteriophages: A therapy concept against multi-drug-resistant bacteria. *Surg Infect (Larchmt)*. 2018;19:737-744.
 42. Rotem SC, Martin L, Blake W. CRISPR-Mediated Adaptive Immune Systems in Bacteria and Archaea. *Annu. Rev. Biochem*. 2013;82:237-66.
 43. Rousseau C, Gonnet M, Le Romancer M, Nicolas J. CRISPI: a CRISPR interactive database. *Bioinformatics*. 2009;25:3317-18.
 44. Salmond GP, Fineran PC. A century of the phage: past, present and future. *Nat Rev Microbiol*. 2015;13:777-786.
 45. Suttle CA. Marine viruses--major players in the global ecosystem. *Nat Rev Microbiol*. 2007;5:801-812.
 46. Suttle, Curtis A. Viruses in the sea *Nature*. 2005;437(7057):356-361. DOI: 10.1038/nature 04160.
 47. Weber-Dąbrowska, B, Jończyk-Matysiak, E, Żaczek M, *et al.* Bacteriophage procurement for therapeutic purposes. *Front Microbiol*. 2016;7:1177.
 48. Weinbauer MG. Ecology of prokaryotic viruses. *FEMS Microbiol Rev*. 2004;28:127-181.
 49. Wiedenheft B, Sternberg SH, Doudna JA. RNA -guided genetic silencing systems in bacteria and archaea. *Nature*. 2012;482:331-338.
 50. Wommack KE, Colwell RR. Virioplankton: Viruses in Aquatic Ecosystems. *Microbiology and Molecular Biology Reviews*. 2000;64(1):69-114. DOI: 10.1128/MMBR. 64.1.69-114.2000.
 51. Yosef I, Goren MG, Qimron U. Proteins and DNA elements essential for the CRISPR adaptation process in *Escherichia coli*. *Nucleic Acids Res*. 2012;40:5569-76.
 52. Young R. Phage lysis: do we have the whole story yet? *Curr Opin Microbiol*. 2013;16:790-797.