Τμήμα Βιοχημείας και Βιοτεχνολογίας
Πανεπιστήμιο Θεσσαλίας

Ανάλυση γονιδιωμάτων πυρήνα στην *E. coli* με βιοπληροφορικές μεθόδους

Core genome analysis of *E.coli* strains with Bioinformatics methods

Καπετάνος Δημήτριος

ΛΑΡΙΣΑ 2022

Η παρούσα πτυχιακή εργασία εκπονήθηκε στο εργαστήριο Βιοπληροφορικής, του τμήματος Βιοχημείας και Βιοτεχνολογίας, Σχολής Επιστημών Υγείας του Πανεπιστημίου Θεσσαλίας (ΠΘ)

Τριμελής Συμβουλευτική Επιτροπή:

Αμούτζιας Γρηγόριος, Αναπληρωτής Καθηγητής Βιοπληροφορικής με έμφαση στη Μικροβιολογία, Τμήμα Βιοχημείας και Βιοτεχνολογίας, Σχολή Επιστημών Υγείας, Πανεπιστήμιο Θεσσαλίας (Επιβλέπων)

Professor Barry Campbell, Infection Biology and Microbiomes, University of Liverpool Faculty of Health and Life Sciences

Παπουτσοπούλου Σταματία, Επίκουρη Καθηγήτρια Μοριακής Ανοσοβιολογίας, Τμήμα Βιοχημείας και Βιοτεχνολογίας, Σχολή Επιστημών Υγείας, Πανεπιστήμιο Θεσσαλίας

ΕΥΧΑΡΙΣΤΙΕΣ

Η παρούσα διπλωματική εργασία εκπονήθηκε στο εργαστήριο Βιοπληροφορικής του τμήματος Βιοχημείας & Βιοτεχνολογίας του Πανεπιστημίου Θεσσαλίας υπό την επίβλεψη του Αναπληρωτή Καθηγητή κ. Γρηγορίου Αμούτζια. Θα ήθελα να τον ευχαριστήσω για την ευκαιρία που μου έδωσε να πραγματοποιήσω την πτυχιακή μου εργασία υπό την επίβλεψή του καθώς και για την καθοδήγησή του. Επίσης, θα ήθελα να ευχαριστήσω τα υπόλοιπα μέλη της τριμελούς επιτροπής, τον Καθηγητή Barry Campbell, από το Πανεπιστήμιο του Liverpool και την Επίκουρη Καθηγήτρια Σταματία Παπουτσοπούλου για την καθοδήγησή τους.

Στη συνέχεια, θα ήθελα να ευχαριστήσω τον υποψήφιο διδάκτορα κ. Νικολαΐδη Μάριο για τη καθοδήγησή του στην υλοποίηση της εργασίας. Τέλος θα ήθελα να ευχαριστήσω την οικογένεια και τους φίλους μου για τη στήριξη και τη βοήθεια που προσέφεραν καθ’ όλα τα χρόνια φοίτησής μου.

ΠΕΡΙΛΗΨΗ

Τα βακτήρια του γένους *Escherichia* συναντώνται κυρίως στο γαστρεντερικό σύστημα των θηλαστικών. Αν και στην πλειοψηφία τους είναι συμβιωτικά και μη παθογόνα, ορισμένα στελέχη έχουν εξελιχθεί σε παθογόνα και η μελέτη τους έχει ιδιαίτερο κλινικό ενδιαφέρον. Χρησιμοποιήθηκαν 240 γονιδιώματα στελεχών του είδους *E.coli* και στη συνέχεια πραγματοποιήθηκε φυλογενωμική ανάλυση για αναγνώριση των πρωτεϊνών πυρήνα μεταξύ των στελεχών και την εύρεση γονιδίων μοναδικών στα στελέχη *E.coli* που ανήκουν στον παθότυπο AIEC. Από την ανάλυση δεν βρέθηκε κάποιο χαρακτηριστικό γονίδιο υπογραφή (fingerprint) που είναι μοναδικό στον παθότυπο Adherent Invasive *E. coli* (AIEC).

Λέξεις-κλειδιά: φυλογενωμική ανάλυση,*E. coli,* παθότυπος, πρωτεΐνες πυρήνας

ABSTRACT

The bacteria of genus *Escherichia* colonize the intestine of mammals. Though, most of them are symbiotic and non pathogenic, some strains have evolved into pathogens and they are of high clinical interest. 240 genomes of different *E. coli* strains were used for a phylogenomic analysis, in order to identify the core genome in the group and also identify genes that are exclusive in *E. coli* strains that belong to the Adherent Invasive *E. coli* (AIEC) pathotype. After the analysis no fingerprint genes that are unique to the AIEC pathotype were identified.

Key-words : phylogenomic analysis, *E. coli,* pathotype, core proteome

Περιεχόμενα

Content (Tables)

Content (Figures)

1.INTRODUCTION

1.1.*The genus Escherichia*

The genus *Escherichia* is part if the family of *Enterobacteriaceae,* which consists of Gram-negative, rod-shaped bacteria that usually colonize the gastrointestinal system of diverse vertebrate hosts. Historically, biochemical analyses were used to identify new species of the genus. Consequently, it was impossible to differentiate some species before the invention of genomic and genotypic methods. For instance, using DNA-DNA hybridization, 16S rDNA sequence analyses and identification of virulence genes, the reclassification of a group of strains as a new species with the name *E. alberti* (Huys et al. 2003). Another new species of *Escherichia, E. marmotae* was classified as a novel *Escherichia* species using the 16S rRNA gene and core genome sequences to analyze strains from fecal samples of the Himalayan marmot (Marmota himalayana) (Liu et al. 2015, 2019). There are four recognized species of the genus: *E. coli/Shigella, E. fergusonii, E. albertii and E. marmotae.*

For many years *Shigella* was believed to belong to a different species in the genus, based on its biochemical properties, the lack of motility and the different pathology. Specifically, Shigella causes invasive intestinal infections, with most common symptoms being bloody or mucous diarrhea and ulcer. However, identification methods based on 16S rRNA, core genome and MLST fail to distinguish *Shigella* from *E. coli (Devanga Ragupathi et al. 2018).*

1.2. Methods of separation of microorganisms

For many years, the identification and separation of microorganisms in different species was based in phenotypic, biochemical and morphological characteristics. After the advance in modern technologies in the last decades, new methods like DNA-DNA hybridization and especially the 16S ribosomal RNA. Another method that was developed is MLST (multi-locus sequence typing) that uses many genetic loci. With the utilization of the new technologies and methods more accurate classifications of microorganisms were possible and older classifications were revised (Yu et al. 2020). By utilizing genotypic and genomic methods the separation of closely related microorganisms is more effective.

1.3. *Escherichia coli*

The species *Escherichia coli* consists of thousands of strains of gram-negative, potentially anaerobic, mostly symbiotic, non pathogenic bacteria that are important for the health of the intestine. A great number of strains colonize and survive in the gut of hosts like humans and many animals like cattle, reptiles or birds and is part of the normal microbiota of the gut (Gordon and Cowling 2003). The different strains of the *E. coli* have developed a huge genetic variation and as a result many of these species have evolved to survive in different environments and hosts. This genetic variation also contributed to the survival of some species outside of their hosts. In non-host environments, some species have been found to survive in soil, wastewater or natural water (Tenaillon et al. 2010).

Some *E. coli* strains have evolved into pathogens by utilizing virulence factors, that cause disease in host organisms. These virulence factors include toxins, siderophores, cell-adhesion or invasion proteins, enzymes or secretion systems. Most of the *E. coli* strains cause infections of the intestine with the main symptom being diarrhea, but also cause infections of other tissues and organs causing urinary tract infections, meningitis or septicemia (Bekal et al. 2003).

1.3.1. *Phylogroups of E. coli*

The strains of *E. coli* are divided in seven phylogroups termed A, B1, B2, C, D, E and F. These phylogroups are related with the properties and the lifestyles of the microorganisms like the host organism, the environment they survive and also the classification in the above phylogroups is important for epidemiological studies. Clermont and colleagues since 2000 have developed several PCR assays that allow a fast and easy assignment of *E. coli* strains in phylogroups (Beghain, Clermont et al. 2009).

1.3.2. E. coli serotypes

E. coli strains are divided in serotypes following the Kauffman-White classification model which was developed for bacteria of the genus *Salmonella.* The model is based on the identification of the antigen that exist in the Gram-negative bacteria. The antigens used for the serotyping are surface O-polysaccharide antigens and the flagellar H-antigens. There are 186 identified O-groups and 53 H-groups in *E. coli,* which combined can give many different serotypes (Fratamico, DebRoy et al. 2016).

The serotype that most often causes diarrhea in humans is the O157:H7. Occasionally, different pathogen strains are responsible for epidemic outbreaks, for

instance the outbreak in Germany in May 2011, caused by a Shiga-toxin producing strain with serotype O104:H4 (Frank C, Werber D, Cramer JP, et al. 2011). Usually, young children, elderly people, travelers and immunosuppressed patients in hospitals are most susceptible to pathogenic *E. coli* strains.

1.3.3. *E. coli pathotypes*

The various pathogenic *E. coli* strains belong to different pathotypes, with each pathotype causing a different type of infection and symptoms. The *E. coli* pathotypes include: enterotoxigenic *E. coli* (ETEC), enterohaemorrhagic *E. coli* (EHEC), enteropathogenic *E. coli* (EPEC), enteroaggregative *E. coli* (EAEC) (Kai et al. 2010), diffuse-adherent *E. coli* (DAEC) (Scaletsky et al. 2002), enteroinvasive *E. coli* (EIEC), adherent-invasive *E. coli* (AIEC) (Barnich and Darfeuille-Michaud 2007), extra intestinal pathogenic *E. coli* (ExPEC). The extra-intestinal pathogenic ExPEC is related to three other pathotypes : uropathogenic *E. coli* (UPEC) (Terlizzi et al. 2017), meningitic *E. coli* (NMEC) and septicemia causing bloodborne *E. coli* (BBEC) (Mokady et al. 2005). There is also a group of strains that produce the Shiga toxin, that belong to the STEC pathotype (Kai et al. 2010) and the avian pathogenic *E. coli* (APEC) in which belong strains that cause infections in birds (Yu et al. 2020).

Most of the pathotypes that cause infections of the intestine, cause diarrhea as the main symptom, whereas the EHEC pathotype causes hemorrhagic diarrhea. The invasive strains of the AIEC and EIEC pathotypes have different infection and pathogenic mechanisms since they must invade inside the host's cells of the intestinal epithelium, unlike strains of other pathotypes that remain out of the host's cells. The UPEC pathotype strains cause urinary tract infections, that may also infect the kidneys if the infection is not restricted in the early stages. Both UPEC and NMEC pathotypes can cause septicemia. NMEC strains have the ability to cross the blood-brain barrier and cause meningitis (Croxen,Finlay 2010).

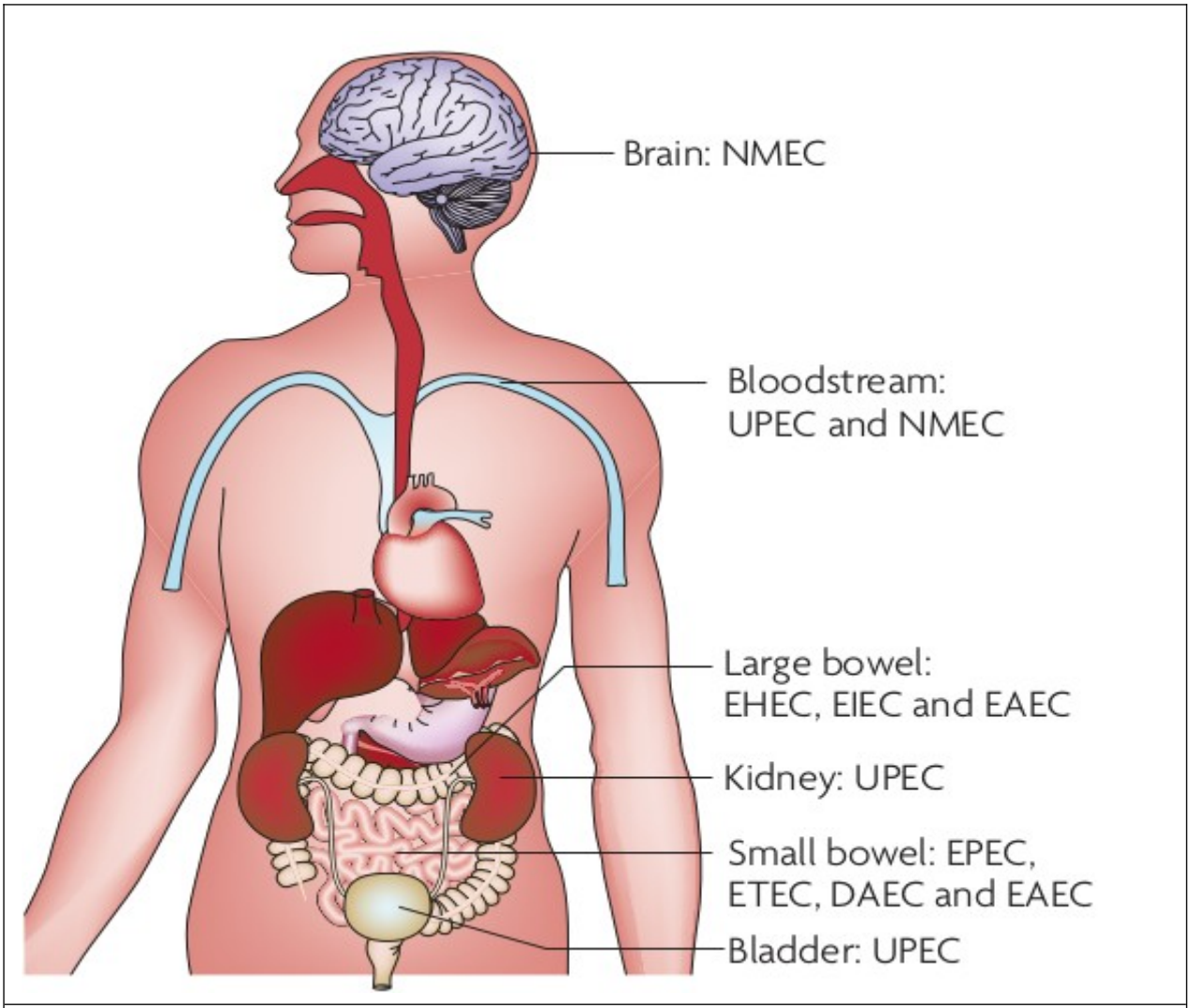


Figure 1: Different E. coli pathotypes and the tissues and organs they infect. E. coli strains that colonize the small bowel, are related to the EPEC, ETEC, DAEC and EAEC causing diarrhea, while EHEC and EIEC are related to strains that colonize the large bowel and EAEC strains colonize both small and large bowel. (Croxen,Finlay 2010).

1.3.4. AIEC and Crohn's disease

The AIEC strains are related to the Crohn's disease (CD) and other Inflammatory bowel diseases, since they are found in the intestine of Crohn's disease patients. The most characteristic symptom of the CD are the chronic wounds on the intestine and the extensive inflammation. It is estimated that in Europe and North America, 20 patients for every 10,000 people are affected (Molodecky et al. 2012). The symptoms that are observed at the Peyer's patches and colon in the early phases of the disease include lymphoid aggregates, ulcer, microabscesses, granulomas, and lymphangitis, stenosis of the intestinal tract and on the long term it is possible that the patients may also develop cancer of the colon (Darfeuille-Michaud, 2002). Some times CD can also cause diarrhea that can also be hemorrhagic, though this last symptom is mostly attributed to other enteropathogenic pathotypes of *E. coli.* Apart from strains that infect humans, AIEC strains have also been found to infect other hosts like cattle, mice, dogs and even some bird species (Dogan B, Suzuki H, Herlekar D, et al. 2014).

It is believed that the Crohn's disease can be caused by perturbations of the microorganisms that colonize the intestine. In Crohn's disease patients an unusual increase in *E. coli* bacteria of the intestine is observed, among them are several AIEC strains that are believed to have a role in the development of the disease (Miquel, Peyretaillade et al. 2010). Patients with CD have an improvement in symptoms when bacterial population decreases after intestinal lavage or treatment with antibacterial drugs.

In the developed countries, modern lifestyle can be the main cause of CD. Nutrition habits that include many preservatives, artificial additives, alcohol and limited consumption of fresh fruits and vegetables have as a result decreased intestinal motility. Other factors that contribute to the onset of CD are the underexposure to microorganisms during early age, sedentary lifestyle and lack of exercise. The frequent use of antibiotics may also have an impact on the onset of CD, since it can harm symbiotic bacteria on the bowel and can also promote resistant strains (Molodecky et al. 2012).

An *E. coli* strain is classified as AIEC if it has all the following characteristics: Ability to adhere and invade the cells of the intestinal epithelium, ability to invade, survive and multiply inside macrophages and also displays the above phenotypic characteristics during *in vitro* assays in cell lines (Camprubí-Font, Lopez-Siles et al. 2018).

Treatment of AIEC strains is very challenging, since every patient of CD may have several different strains, which are genetically different and some of them also posses

genes that provide resistance to antibiotics. Consequently it is not easy to use antibacterial drugs against all of the different strains, while at the same time the treatment may damage useful commensal microbiota (Tyakht, Manolov, Kanygina et al 2018).

To this date, no molecular markers have been found that can be used to identify the AIEC strains. This is explained by the genetic variation and different evolution paths that each different strain followed. AIEC strains have been found that classify in four different phylogroups, though the majority of strains belong to the B2 phylogroup (O'Brien, Bringer, Holt, et al. 2017). Consequently, molecular markers might not be enough to identify all the strains that belong to the AIEC pathotype. The only effective way of identifying AIEC strains is through *in vitro* invasion assays, in which bacteria interact with the host cells and the main phenotypic characteristics of the pathotype can be observed.

1.3.5. Infection Process and mechanisms

Adhesion of the bacteria in the epithelial cells of the intestine is the first step of the pathogenesis of many intestinal pathogenic bacteria. Adhesion allows the bacteria to create colonies and resist mechanical cleansing of the microorganisms in the gut. A main characteristic of AIEC strains is the invasion of the epithelial cells of the intestine and macrophages, where bacteria survive without damaging the host cell. The invasion and survival inside host's cells makes their destruction from defensive mechanisms even harder. Through infected macrophages the infection can spread in many cells of the intestine's mucosa. Infected macrophages secrete the TNFa cytokine, which further enhances inflammation in the gut (see Figure 3). This way, the patients develop a chronic inflammation and many of the symptoms of CD (Dogan B, Suzuki H, Herlekar D, et al. 2014).

So far, studies in AIEC strains have revealed information related to its pathogenicity, some of the mechanism involved and genes that are related to the pathotype. Sequencing of nine genomes from *E. coli* strains including eight of the AIEC pathotype and comparison with 38 other available *E. coli* and *Shigella* showed that AIEC strains have an independent evolution path and have not evolved from a common ancestor (Dogan B, Suzuki H, Herlekar D, et al. 2014). The above information explains the high genetic variation between AIEC strains and the existence of different virulence proteins that are used in the mechanisms of invasion.

Also, no genes where found that are unique to the AIEC pathotype and do not exist in other *E. coli* species. Usually AIEC strains have more genes coding for virulence proteins in total compared to other pathogenic *E. coli*. It is believed that the pathotype is

the outcome of many different genes that contribute to virulence and also allow them to survive in the conditions inside host cells. Genes related to the AIEC pathotype are often coding for a protein of metabolism or iron acquisition (Dogan B, Suzuki H, Herlekar D, et al. 2014).

It is believed that AIEC strains, at the first stages of CD exploit the weakness of the host in antimicrobial defense and they use the carcinoembryonic antigen-related cell adhesion molecule (CEACAM6) receptors, which are overexpressed during inflammation to adhere using the FimH adhesin (Dogan B, Suzuki H, Herlekar D, et al. 2014). The flagella of AIEC strains apart from its role in motility, also participates in biofilm formation and adherence and invasion of Caco-2 cells (Zhou M, Yang Y et al. 2015). Caco-2 is a cell line of human cancer cells often used as model for *in vitro* assays of the intestinal epithelium.

An important gene in AIEC strains is the *IbeA,* which codes for a protein that takes part in the invasion process to host cells. Moreover, this protein also has a role in the survival of the bacteria inside macrophages and under water stress (Cieza, Hu et al. 2015). The product of *neuC* gene is also crucial as it favors the expression of capsular polysaccharide that inhibits phagocytosis (Barnich N, Boudeau J et al. 2003). The *lpfA* is found in most AIEC strains, it is related with increased invasion activity of Caco-2 epithelial cells (Dogan B, Suzuki H, Herlekar D, et al. 2014).

A secretion system type 6 is present in AIEC strains known as T6SS (Desilets et al. 2016). It participates in many important activities related to the invasion of AIEC strains in host cells and their survival inside them. T6SS interacts with the microtubule network of host cells to start the invasion and promotes replication within macrophages (Sana, Baumann et al. 2015). Also, it is linked to the suppression of the host immunity, possibly by inhibiting the secretion of cytokines from the host organism that are part of defense mechanisms against intracellular pathogens (Chen, Yang et al. 2017). Last but not least, T6SS promotes proliferation of bacteria after the invasion of macrophages (Eshraghi, Kim et al. 2016).

The pduC gene codes for the large subunit of propanediol dehydratase and seems to be related to the AIEC pathotype and the usage of alternative substrates. The *pdu* operon is often found in enteropathogenic bacteria that colonize the intestine like *Salmonella* and *Listeria*, but it mostly exists in AIEC than other *E. coli*. This operon participates in the metabolic path of fucose in anaerobial conditions (see Figure 2). This way, AIEC strains are capable of using a new nutrient when other common options are not

available within the macrophages and other cells. To conclude, the *pdu* operon allows AIEC to have a competitive advantage against other bacteria that cannot find nutrients inside host cells (Dogan B, Suzuki H, Herlekar D, et al. 2014).

1.3.6. Similarities of AIEC to other pathotypes

Part of the T6SS secretion system are the genes *chuA* and *yersiniabactin,* which participate in iron uptake. Their expression levels are increased in AIEC and other pathotypes like EHEC and ExPEC. These two genes are expressed mostly in bacteria that invade host cells, while it is not known if they are expressed in non-invasive bacteria. Another gene that is common between AIEC and ExPEC is the *ibeA* (Dogan B, Suzuki H, Herlekar D, et al. 2014).

Generally, the existing data leads to the conclusion that AIEC strains have more similar characteristics with pathogen strains of *E. coli* that belong to the extra-intestinal pathotypes than strain that colonize and infect the intestine. This is also supported by the fact that AIEC and ExPEC have similar genes that code for virulence proteins (Desilets et al. 2016).

AIEC strains use a number of genes to overcome mechanical removal of microorganisms in the gut, overcome mucosal defensive mechanisms and to survive within macrophages. UPEC have also to survive from several defensive mechanisms to adhere and invade the urinary tract (O'Brien, Bringer, Holt, et al. 2017).

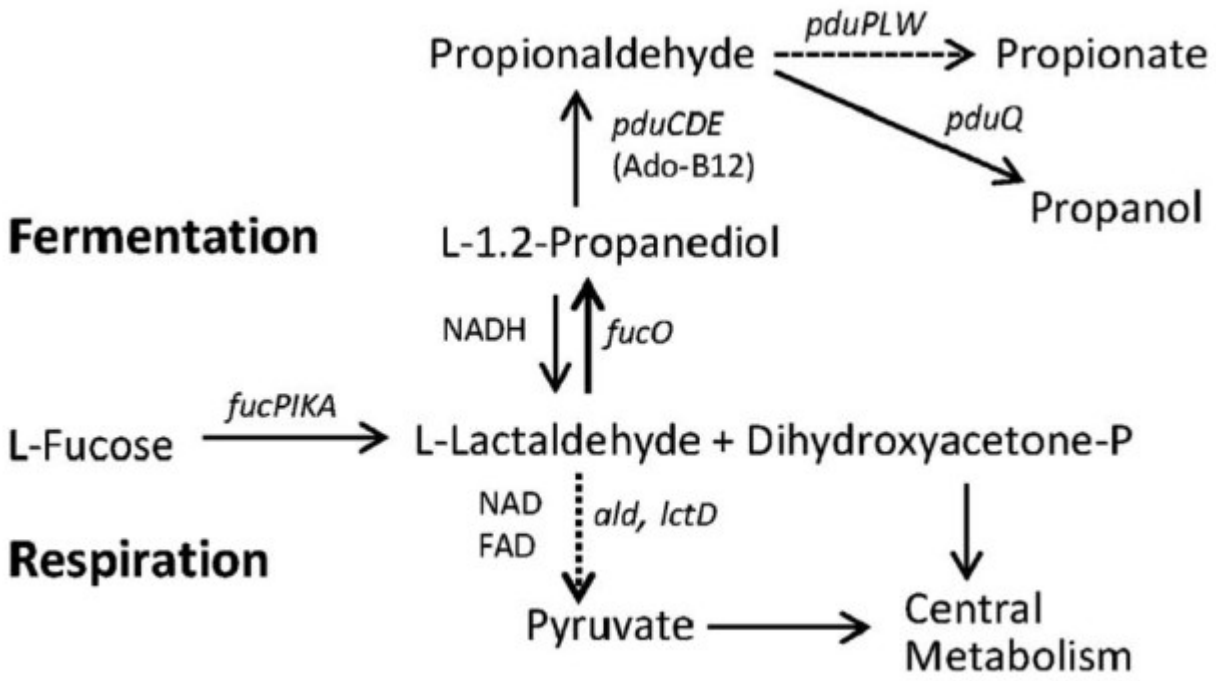


Figure 2 : The pathway of L-Fucose and Propanodiol metabolism in *E.coli* (Dogan B, Suzuki H, Herlekar D, et al. 2014).

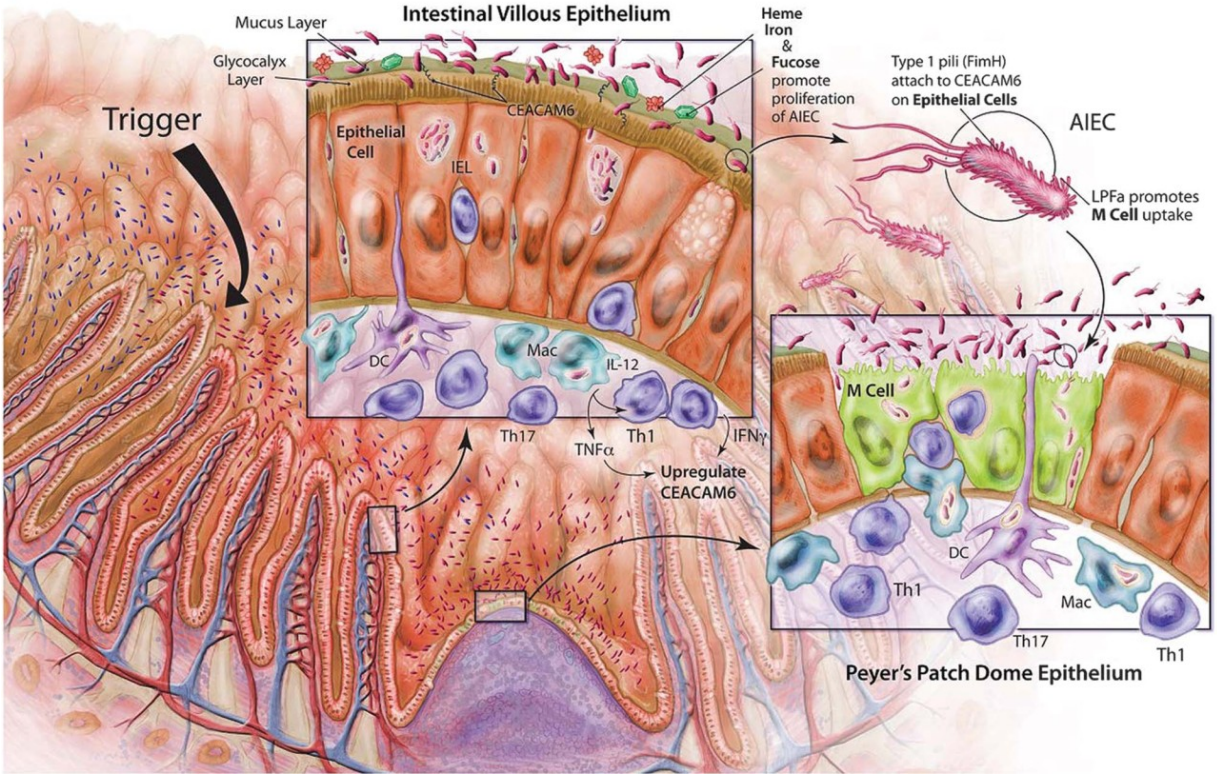


Figure 3. *The mechanisms that AIEC are believed to use in order to promote intestinal inflammation. AIEC exploit changes in the mucosal environment caused by inflammation to be able to adhere, invade and multiply. The ability of AIEC to alternative nutrients and the enhanced iron uptake is crucial to survive inside host cells. The overexpression of CEACAMG receptors in inflammation helps AIEC to adhere using fimH adhesin, in order to allow the invasion process to begin. Bacteria use Peyer's patches to invade macrophages. Within macrophages TNFa cytokin is secreted that further promotes the intestinal infection of AIEC strains (Dogan B, Suzuki H, Herlekar D, et al. 2014).*

1.4 Pan-genome and core genome

1.4.1. Pan-genome

The progress in new sequencing technologies during the last years had as a result the dramatic decrease in cost of whole genome sequencing of microorganisms and the availability of a very large amount of data in public databases, that can be used in genomic analyses. Genomic analyses that include thousands of genomes of different organisms are very common today. In a genomic analysis of a group of genomes that belong in the same species or genus, the total genes that exists in the organisms is called the pan-genome.

If a pan-genome increases in size for every new genome that is added, it is characterized as open pan-genome, while a closed pan-genome remains the same when new genomes are added (Bosi et al. 2015).

1.4.2. Core-genome

The core-genome consists of all the orthologs that are common in all the organisms of the same species or genus in an analysis. It is believed that the genes of the core genome are essential for the basic metabolic processes and other basic cell functions of the organisms (Bosi et al. 2015). These are also the roles of core proteins in *Escherichia coli.* As the number of genomes increases in an analyses, the core genome becomes smaller. A large number of genomes will result in less core proteins that are present in all of the organisms. Usually, this is due to sequencing errors of the included strains.

1.4.3. Identification of core genome

Usually, in order to identify the core genome, a group of ortholog genes from the total organisms in an analysis is used (Vernikos et al. 2015). Several software have been developed for finding ortholog genes, some of them are: InParanoid, OMA, OrthoMCL, OrthoFinder (Emms and Kelly 2019). These software use reciprocal BLAST, while OrthoFinder can alternatively use the DIAMOND algorithm. DIAMOND is faster than BLAST, though it is less sensitive, which makes the use of most strict settings necessary in order to ensure the best reciprocal hits (Hernández-Salmerón and Moreno-Hagelsieb, 2020;Nikolaidis et al., 2020).

Reciprocal BLAST is a simple method to find ortholog genes without being seriously affected by false positive results. The BLAST algorithm designed to find ortholog genes in

two or more proteomes is Blastp. The procedure includes the comparison of the entire proteome of two organisms. Two proteins from different genomes are orthologs of they are reciprocal best BLAST hits. In the first phase, all the proteins from an organism are used as query sequence in order to be compared with the proteins of the second organism and identify homologous proteins. Vise versa, the second proteome is used as a query for comparison with the proteins of the first organism. Consequently, two proteins will be orthologues if they are the best BLAST hits in both comparisons (Hernández-Salmerón and Moreno-Hagelsieb, 2020).

In this analysis, the identification of ortholog genes was based in python 3 scripts that use reciprocal BLAST with strict criteria, in order to identify the core genome of a group of organisms (Nikolaidis et al., 2020). The method uses one proteome as reference proteome, that is used to search for ortholog proteins in the rest of the other proteomes. Since this method compares only the reference proteome to the proteomes of the other organisms in the analysis it is much faster than methods which compare every proteome with all the other proteomes (all against all). After the reciprocal BLAST is completed between the reference proteome and the proteomes of the rest of the organisms, for each comparison the standard deviation and average are calculated. Afterwards, all proteins that have two standard deviations or higher difference from the average are excluded, because they are not appropriate as orthologues. When the ortholog proteins are found a table is generated, that includes in every row the proteins of the reference proteome and in every column the rest of the organisms with their ortholog proteins, if they exist. From this table, if all the proteomes of the reference proteome without orthologues in every organism get excluded, the rest of the proteins are the core proteome in the analysis.

For the phylogenomic analysis, the software MUSCLE (Edgar 2004) was used, that performs multiple alignment which is filtered by Gblocks (Castresana, 2000). With the final alignment a phylogenomic tree with distances is calculated using the BioNJ algorithm (Gouy et al., 2010).

The goal of this analysis was to study AIEC pathotype strains, in order to identify the core and fingerprint proteins. As fingerprints, we denote these proteins that are present in all the AIEC members and absent in all other *E. coli*. Also, the AIEC strains of the analysis where classified in serotypes and phylogroups and their similarities with strains of other pathotypes were studied.

2. MATERIALS AND METHODS

2.1. Software

2.1.1. Linux Ubuntu 22.04

Linux is a free, open source and high security operating system. The graphical environment used is GNOME and the terminal environment is the BASH (Bourne Again Shell). BASH is appropriate for the management of large amount of data and automated procedures (“Enterprise Open Source and Linux,” n.d.).

2.1.2. Python

Python is an object-oriented programming language that is widely used. Many different libraries are available for use in various applications (“Welcome to Python.org,”n.d.). Biopython is especially used for bioinformatic research (Cock, Antao et al. 2009).

2.1.3. Perl

Perl is an object-oriented programming language, compatible with most operating systems that is constantly developed for 30 years. It is appropriate for many applications and can be used for text and files manipulation. Apart from object-oriented it also supports procedural and functional programming (www.perl.org, about Perl).

2.1.4. Treedyn

Treedyn is a program capable of visualizing, manipulating and processing data of phylogenetic and phylogenomic trees. It is written in Tcl/Tk programming language and it is free and compatible with many operating systems including Linux, Windows and Mac OSX. It provides the ability to modify and add metadata from files to a genomic tree and the option of exporting and saving the projects in files of various formats (Chevenet et al., 2006).

2.1.5. Software of core genome identification

In order to identify the core genome/proteome in the analyses a series of procedures (pipeline) was used which was originally implemented in bacteria of the genus *Pseudomonas* (Nikolaidis et al., 2020). The basic principle of the method is the finding of orthologue genes, by using a reference proteome to perform reciprocal BLAST in pairs with the rest of the other proteomes of the analysis, requiring significantly less times than methods which compare all the organisms with each other (all against all).

2.1.6. RAPT

RAPT (Read Assembly and Annotation Pipeline Tool) is an NCBI tool that consists of a series of procedures that allow the *de novo* assembly of prokaryotic genomic reads, from Illumina sequencing. It consists of three basic tools which are: the genome assembler SKESA (Souvorov, Agarwala, Lipman 2018), the taxonomic assignment tool ANI (Ciufo, Kannan et al. 2018) and the Prokaryotic Genome Annotation Pipeline PGAP (Li, O'Neill et al. 2021).

2.1.7. Ectyper

Because the serotype of most species was not available in databanks, the program Ectyper was used in conda environment to identify the serotype of the different *E.coli* and *Shigella* strains of the analysis in serotypes according to the Kauffman-White classification model. Ectyper uses fasta or fastq format files (Bessonov, Laing et al. 2021).

2.1.8. ClermonTyping

ClermonTyping is an *in silico* PCR assay that uses several tools to classify *E.coli* strains in one of the various phylogroups (A, B1, B2, C, D και F). In order to identify the phylogenetic group it uses fasta format files containing the genome sequences of the strains (Beghain, Clermont et al. 2009).

2.2. Data downloading

From NCBI taxonony database, taxonomy identifications were saved for species of the *Escherichia* group. The taxonomy identifications were used to download the files of each strain from NCBI Assembly database. At the first stage, proteomes were obtained from strains with available assembly at the chromosome or complete genome level.

Proteomes that are not fully assembled, may affect the results of the core and fingerprint analyses.

For each strain of the analysis, the four files were downloaded which include the protein sequence in FASTA format and the genomic sequences, the feature tables and the GBFF files which contain information for every strain like the serotype, the area where the strain was found, the host organism or the environment that the species survives. Some of the GBFF files also have information about the pathotype of the strain.

3. RESULTS AND DISCUSION

In the first stage of the analysis, a total of 2074 proteomes of *E.coli* strains*, 80 Shigella* strains, one *E. fergusonii strain and one E. albertii* were used for the first genomic analysis and the construction of the phylogenomic tree. The visualization and processing of the phylogenomic was performed using Treedyn which was the best available program for the handling of a very large genomic tree consisted of thousands of organisms.

After the pipeline was completed using as reference the *E.coli* K-12 strain (assembly accession GCF_000005845.2), the first genomic tree was created, that included many strains clustered together in the tree, that were similar. The core proteome for the 2156 strains consisted of 218 orthologs. Because in the first tree there were many strains almost identical and redundant, 211 strains were selected from the total of 2156, for a second genomic analysis. In these 211 strains, were included organisms with specific attributes of interest, including the pathotype and the serotype. Also, from eight articles of studies related to the AIEC pathotype, 29 more strains were selected for enrichment of the analysis, as displayed in Table 1. Most of these strains were found on the intestine of Crohn's disease patients. Some of these strains that were included in the first phylogenomic tree turned out to belong to the AIEC pathotype according to the information of the articles, but were not included in the 211 selected for the second phase. Five of the 29 strains had no available proteome in NCBI Assembly. Subsequently, the RAPT software of NCBI was used to assemble these proteomes.

A second pipeline was executed for the 240 strains, including the 211 of the first phase and the 29 more of the enrichment. The reference strain used for this second analysis was once more the *E.coli* K-12. This second pipeline also produced a phylogenomic tree Figure 4 . The core genome in this second genomic analysis was consisted of 438 genes.

Table 1. The strains that were used for the enrichment of the second pipeline and phylogenomic tree. The strains CDEC were isolated from CD patients, but they have not been identified as AIEC.

Assembly ID	pathotype	status	PMID
ERR2265581	unknown	New strain	29426864
ERR2265582	AIEC	New strain	29426864
ERR2265584	unknown	New strain	29426864
ERR2265585	AIEC	New strain	29426864
ERR2265586	AIEC	New strain	29426864
GCA_000148605.1	AIEC	New strain	21075930
GCA_000179795.1	AIEC	New strain	25230163
GCA_000183345.1	AIEC	New strain	21108814
GCA_000264095.1	AIEC	New strain	25230163
GCA_000264115.1	AIEC	New strain	25230163
GCA_000264135.1	AIEC	New strain	25230163
GCA_000264175.1	AIEC	New strain	25230163
GCA_000264195.1	AIEC	New strain	25230163
GCA_000264215.1	AIEC	New strain	25230163
GCA_000264235.1	AIEC	New strain	25230163
GCF_000284495.1	AIEC	Existed in the first tree, LF82 strain	25230163
GCA_001630825.1	CDEC	New strain	28724357
GCA_001630835.1	CDEC	New strain	28724357
GCA_001630845.1	CDEC	New strain	28724357
GCA_001630855.1	CDEC	New strain	28724357
GCA_001630905.1	CDEC	New strain	28724357
GCA_001630915.1	CDEC	New strain	28724357
GCA_001630925.1	CDEC	New strain	28724357
GCA_001630965.1	CDEC	New strain	28724357
GCA_001630975.1	CDEC	New strain	28724357
GCA_001631005.1	CDEC	New strain	28724357
GCA_003258455.1	AIEC	New strain	30425690
GCF_000285375.1	AIEC	Existed in the first tree	21705601
GCF_017901015.1	AIEC	Existed in the first tree	34710159

GCF_021398975.1	AIEC	Existed in the first tree	34710159
GCF_021398995.1	AIEC	Existed in the first tree	34710159

In order to identify proteins that are unique to the AIEC pathotype, a third pipeline was executed, this time using as reference proteome the LF82 with assembly accession number GCF_000284495.1. The LF82 strain is the reference strain for studies related to the AIEC pathotype, so it is appropriate to be used as a reference for proteins related to the pathotype. The core proteome in this third analysis was increased to 441 genes. After the BLAST was completed all the proteins of the LF82 strain and their orthologues were compared to identify fingerprint proteins that only exist in the all of the AIEC strains.

No fingerprint proteins for the AIEC pathotype were found. This was to a great extent not surprising, since no studies so far revealed genes that define the pathotype and are not found in non AIEC strains. Nevertheless, we performed this analysis by utilizing a significantly higher number of complete genomes than other studies. It is believed that the pathotype does not depend in one or more genes but to a group of several genes that may be different in the various AIEC strains that may have evolved from independent evolutionary paths. Another consequence of the large genetic variation of the AIEC strains is the identification of many different serotypes. The AIEC strains of the analysis were classified in phylogroups using ClermonTyping *in silico* PCR assay and the AIEC was scattered in three phylogroups which were B1, B2 and D.

In the phylogenomic tree, the AIEC strains displayed a large diversity and various characteristics. Almost every AIEC strain in the analysis displayed a unique serotype compared to the other strains of the pathotype, for a total of 15 different serotypes. Only three pairs of strain share a common serotype, including a pair of strains of the serotype O1:H7, two strains sharing the O2:H7 serotype and the O83:H1 serotype which is the serotype of the reference strain LF82 and one more strain in the analysis. The serotypes of AIEC and other strains are displayed in Table 2.

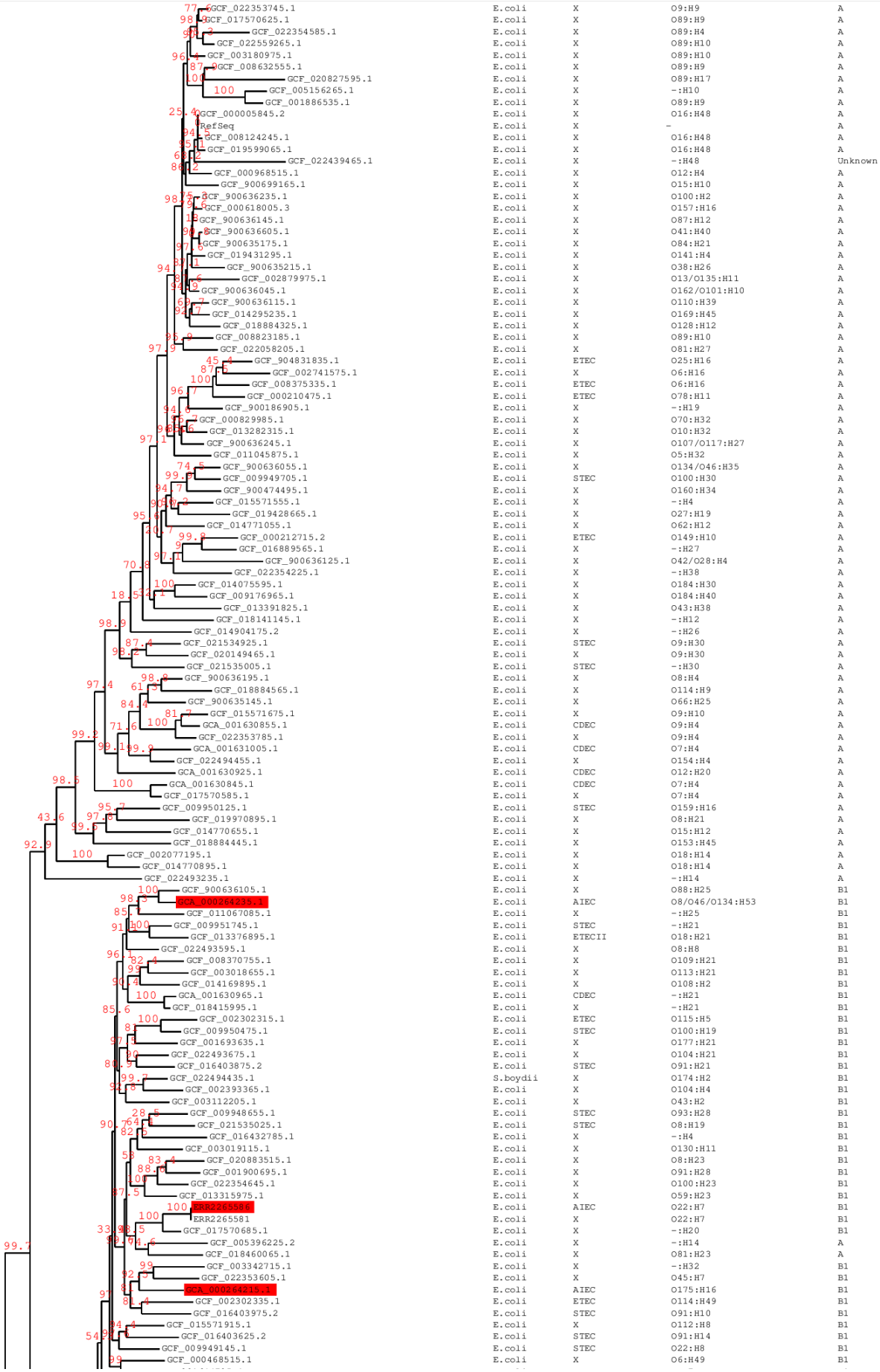
The phylogenomic tree consists of two different branches where AIEC strains clustered which is another evidence that supports the independent emergence and evolutionary history of the AIEC strains. It is believed that some of the virulence genes were transferred horizontally from other bacteria that may be either *E. coli* and Shigella or

even bacteria of other genera that have the ability to invade the cells of the intestine, like the Salmonella genus.

In the fourth branch of the tree there are 4 AIEC strains clustered together with one strain of the pathotype ExPEC, one UPEC strain and one APEC strain. Since UPEC and ExPEC pathotypes are extra-intestinal types, AIEC pathotype appears to be closely related to extra-intestinal pathotype and share several common attributes, including their virulence proteins, their serotype an the phylogroup (Desilets et al. 2016). Also, most of the enteropathogenic strains in the phylogenomic tree were distant from AIEC and extra-intestinal pathotype. Enteropathogenic strains clustered in the first two branches of the tree where few AIEC where present and also were classified in A, B1, E and rarely in B2 phylogroups. All the above is strong evidence that AIEC strains have different virulence genes and pathogenicity compared to the enteropathogenic strains.

Table 2. The AIEC strains that were used in the analysis, the pathotype, phylogroup and serotype of each strain is also displayed. Also some extra-intestinal pathotype strain for comparison with AIEC.

Assembly ID	pathotype	phylogroup	serotype
GCF_000284495.1	AIEC	B2	O83:H1
GCA_000183345.1	AIEC	B2	O83:H1
GCA_00264235.1	AIEC	B1	O8/O46/O134:H53
ERR2265586	AIEC	B1	O22:H7
GCA_000264215.1	AIEC	B1	O175:H16
GCA_000264115.1	AIEC	A	O21:H33
GCA_003258455.1	AIEC	B2	O2/O50:H7
GCA_000264175.1	AIEC	B2	O1:H7
GCF_000285375.1	AIEC	B2	O1:H7
GCA_000148605.1	AIEC	B2	O18:H7
GCF_017901015.1	AIEC	B2	O2:H6
GCA_000179795.1	AIEC	B2	O2:H6
GCA_000264095.1	AIEC	B2	O8:H10
ERR2265582	AIEC	B2	O46:H31
ERR2265585	AIEC	B2	O6:H1
GCF_021398995.1	AIEC	B2	O25:H4
GCA_000264195.1	AIEC	D	O1:H6
GCA_000264135.1	AIEC	D	O77/O17/O44/O106/ O73:H18
GCF_017356665.1	UPEC	B2	O2/O50:H7
GCF_002844685.1	ExPEC	B2	O2/O50:H7
GCF_001021615.1	APEC	B2	unknown
GCF_001693315.1	UPEC	B2	O6:H1
GCF_003856995.1	UPEC	B2	O25:H4


GCF_022353745.1 E.coli X O9:H9 A
GCF_017570625.1 E.coli X O89:H9 A
GCF_022354585.1 E.coli X O89:H4 A
GCF_022559265.1 E.coli X O89:H10 A
GCF_003180975.1 E.coli X O89:H10 A
GCF_008632555.1 E.coli X O89:H9 A
GCF_020827595.1 E.coli X O89:H17 A
GCF_005156265.1 E.coli X -:H10 A
GCF_001886535.1 E.coli X O89:H9 A
GCF_000005845.2 E.coli X O16:H48 A
RefSeq E.coli X - A
GCF_008124245.1 E.coli X O16:H48 A
GCF_019599065.1 E.coli X O16:H48 A
GCF_022439465.1 E.coli X -:H48 Unknown
GCF_000968515.1 E.coli X O12:H4 A
GCF_900699165.1 E.coli X O15:H10 A
GCF_900636235.1 E.coli X O100:H2 A
GCF_000618005.3 E.coli X O157:H16 A
GCF_900636145.1 E.coli X O87:H12 A
GCF_900636605.1 E.coli X O41:H40 A
GCF_900635175.1 E.coli X O84:H21 A
GCF_019431295.1 E.coli X O141:H4 A
GCF_900635215.1 E.coli X O38:H26 A
GCF_002879975.1 E.coli X O13/O135:H11 A
GCF_900636045.1 E.coli X O162/O101:H10 A
GCF_900636115.1 E.coli X O110:H39 A
GCF_014295235.1 E.coli X O169:H45 A
GCF_018884325.1 E.coli X O128:H12 A
GCF_008823185.1 E.coli X O89:H10 A
GCF_022058205.1 E.coli X O81:H27 A
GCF_904831835.1 E.coli ETEC O25:H16 A
GCF_002741575.1 E.coli X O6:H16 A
GCF_008375335.1 E.coli ETEC O6:H16 A
GCF_000210475.1 E.coli ETEC O78:H11 A
GCF_900186905.1 E.coli X -:H19 A
GCF_000829985.1 E.coli X O70:H32 A
GCF_013282315.1 E.coli X O10:H32 A
GCF_900636245.1 E.coli X O107/O117:H27 A
GCF_011045875.1 E.coli X O5:H32 A
GCF_900636055.1 E.coli X O134/O46:H35 A
GCF_009949705.1 E.coli STEC O100:H30 A
GCF_900474495.1 E.coli X O160:H34 A
GCF_015571555.1 E.coli X -:H4 A
GCF_019428665.1 E.coli X O27:H19 A
GCF_014771055.1 E.coli X O62:H12 A
GCF_000212715.2 E.coli ETEC O149:H10 A
GCF_016889565.1 E.coli X -:H27 A
GCF_900636125.1 E.coli X O42/O28:H4 A
GCF_022354225.1 E.coli X -:H38 A
GCF_014075595.1 E.coli X O184:H30 A
GCF_009176965.1 E.coli X O184:H40 A
GCF_013391825.1 E.coli X O43:H38 A
GCF_018141145.1 E.coli X -:H12 A
GCF_014904175.2 E.coli X -:H26 A
GCF_021534925.1 E.coli STEC O9:H30 A
GCF_020149465.1 E.coli X O9:H30 A
GCF_021535005.1 E.coli STEC -:H30 A
GCF_900636195.1 E.coli X O8:H4 A
GCF_018884565.1 E.coli X O114:H9 A
GCF_900635145.1 E.coli X O66:H25 A
GCF_015571675.1 E.coli X O9:H10 A
GCA_001630855.1 E.coli CDEC O9:H4 A
GCF_022353785.1 E.coli X O9:H4 A
GCA_001631005.1 E.coli CDEC O7:H4 A
GCF_022494455.1 E.coli X O154:H4 A
GCA_001630925.1 E.coli CDEC O12:H20 A
GCA_001630845.1 E.coli CDEC O7:H4 A
GCF_017570585.1 E.coli X O7:H4 A
GCF_009950125.1 E.coli STEC O159:H16 A
GCF_019970895.1 E.coli X O8:H21 A
GCF_014770655.1 E.coli X O15:H12 A
GCF_018884445.1 E.coli X O153:H45 A
GCF_002077195.1 E.coli X O18:H14 A
GCF_014770895.1 E.coli X O18:H14 A
GCF_022493235.1 E.coli X -:H14 A
GCF_900636105.1 E.coli X O88:H25 B1
GCA_000264235.1 E.coli AIEC O8/O46/O134:H53 B1
GCF_011067085.1 E.coli X -:H25 B1
GCF_009951745.1 E.coli STEC -:H21 B1
GCF_013376895.1 E.coli ETECII O18:H21 B1
GCF_022493595.1 E.coli X O8:H8 B1
GCF_008370755.1 E.coli X O109:H21 B1
GCF_003018655.1 E.coli X O113:H21 B1
GCF_014169895.1 E.coli X O108:H2 B1
GCA_001630965.1 E.coli CDEC -:H21 B1
GCF_018415995.1 E.coli X -:H21 B1
GCF_002302315.1 E.coli ETEC O115:H5 B1
GCF_009950475.1 E.coli STEC O100:H19 B1
GCF_001693635.1 E.coli X O177:H21 B1
GCF_022493675.1 E.coli X O104:H21 B1
GCF_016403875.2 E.coli STEC O91:H21 B1
GCF_022494435.1 S.boydii X O174:H2 B1
GCF_002393365.1 E.coli X O104:H4 B1
GCF_003112205.1 E.coli X O43:H2 B1
GCF_009948655.1 E.coli STEC O93:H28 B1
GCF_021535025.1 E.coli STEC O8:H19 B1
GCF_016432785.1 E.coli X -:H4 B1
GCF_003019115.1 E.coli X O130:H11 B1
GCF_020883515.1 E.coli X O8:H23 B1
GCF_001900695.1 E.coli X O91:H28 B1
GCF_022354645.1 E.coli X O100:H23 B1
GCF_013315975.1 E.coli X O59:H23 B1
ERR2265586 E.coli AIEC O22:H7 B1
ERR2265581 E.coli X O22:H7 B1
GCF_017570685.1 E.coli X -:H20 B1
GCF_005396225.2 E.coli X -:H14 A
GCF_018460065.1 E.coli X O81:H23 A
GCF_003342715.1 E.coli X -:H32 B1
GCF_022353605.1 E.coli X O45:H7 B1
GCA_000264215.1 E.coli AIEC O175:H16 B1
GCF_002302335.1 E.coli ETEC O114:H49 B1
GCF_016403975.2 E.coli STEC O91:H10 B1
GCF_015571915.1 E.coli X O112:H8 B1
GCF_016403625.2 E.coli STEC O91:H14 B1
GCF_009949145.1 E.coli STEC O22:H8 B1
GCF_000468515.1 E.coli X O6:H49 B1

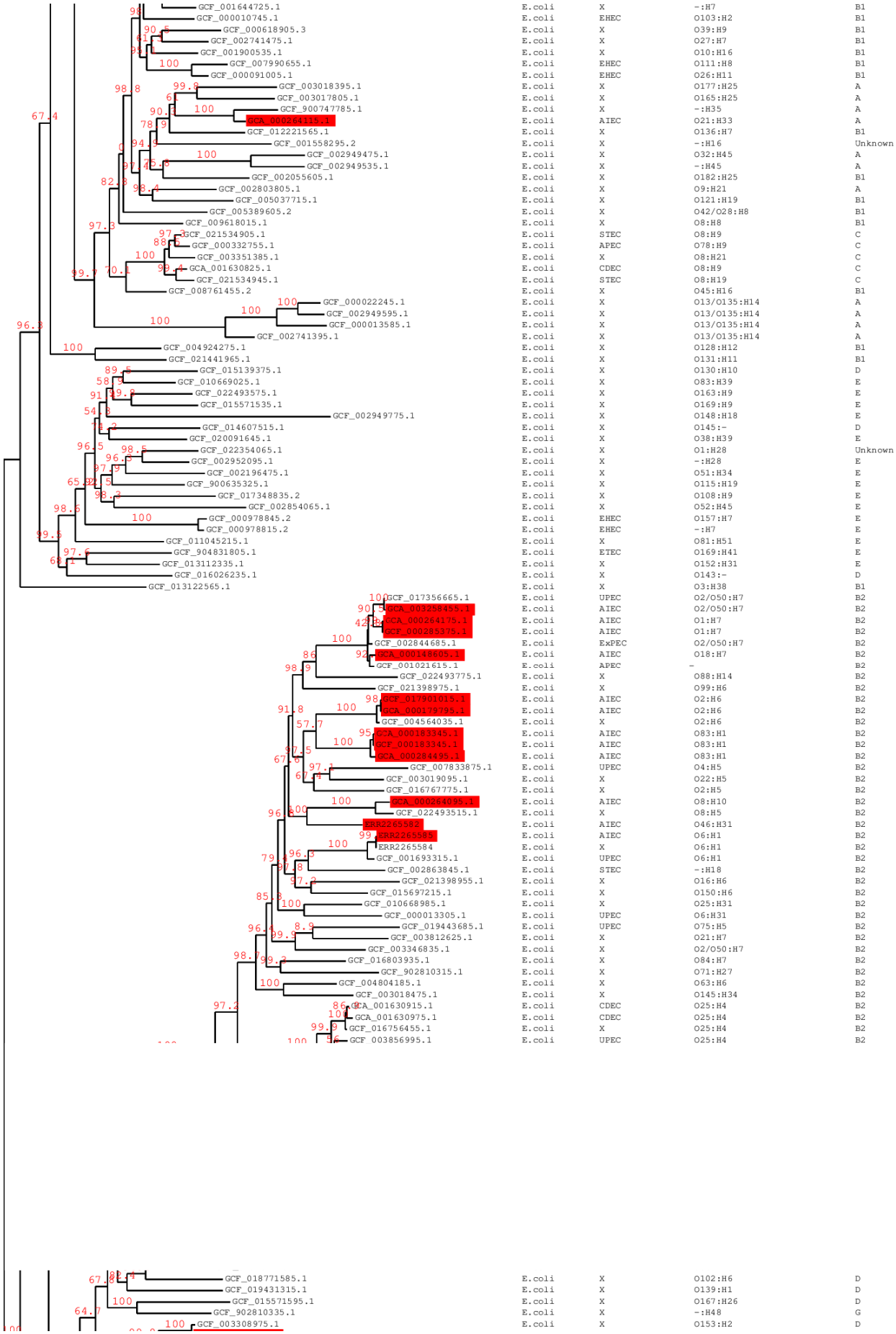

GCF_001644725.1 E.coli X -:H7 B1
GCF_000010745.1 E.coli EHEC O103:H2 B1
GCF_000618905.3 E.coli X O39:H9 B1
GCF_002741475.1 E.coli X O27:H7 B1
GCF_001900535.1 E.coli X O10:H16 B1
GCF_007990655.1 E.coli EHEC O111:H8 B1
GCF_000091005.1 E.coli EHEC O26:H11 B1
GCF_003018395.1 E.coli X O177:H25 A
GCF_003017805.1 E.coli X O165:H25 A
GCF_900747785.1 E.coli X -:H35 A
GCA_000264115.1 E.coli AIEC O21:H33 A
GCF_012221565.1 E.coli X O136:H7 B1
GCF_001558295.2 E.coli X -:H16 Unknown
GCF_002949475.1 E.coli X O32:H45 A
GCF_002949535.1 E.coli X -:H45 A
GCF_002055605.1 E.coli X O182:H25 B1
GCF_002803805.1 E.coli X O9:H21 A
GCF_005037715.1 E.coli X O121:H19 B1
GCF_005389605.2 E.coli X O42/O28:H8 B1
GCF_009618015.1 E.coli X O8:H8 B1
GCF_021534905.1 E.coli STEC O8:H9 C
GCF_000332755.1 E.coli APEC O78:H9 C
GCF_003351385.1 E.coli X O8:H21 C
GCA_001630825.1 E.coli CDEC O8:H9 C
GCF_021534945.1 E.coli STEC O8:H19 C
GCF_008761455.2 E.coli X O45:H16 B1
GCF_000022245.1 E.coli X O13/O135:H14 A
GCF_002949595.1 E.coli X O13/O135:H14 A
GCF_000013585.1 E.coli X O13/O135:H14 A
GCF_002741395.1 E.coli X O13/O135:H14 A
GCF_004924275.1 E.coli X O128:H12 B1
GCF_021441965.1 E.coli X O131:H11 B1
GCF_015139375.1 E.coli X O130:H10 D
GCF_010669025.1 E.coli X O83:H39 E
GCF_022493575.1 E.coli X O163:H9 E
GCF_015571535.1 E.coli X O169:H9 E
GCF_002949775.1 E.coli X O148:H18 E
GCF_014607515.1 E.coli X O145:- D
GCF_020091645.1 E.coli X O38:H39 E
GCF_022354065.1 E.coli X O1:H28 Unknown
GCF_002952095.1 E.coli X -:H28 E
GCF_002196475.1 E.coli X O51:H34 E
GCF_900635325.1 E.coli X O115:H19 E
GCF_017348835.2 E.coli X O108:H9 E
GCF_002854065.1 E.coli X O52:H45 E
GCF_000978845.2 E.coli EHEC O157:H7 E
GCF_000978815.2 E.coli EHEC -:H7 E
GCF_011045215.1 E.coli X O81:H51 E
GCF_904831805.1 E.coli ETEC O169:H41 E
GCF_013112335.1 E.coli X O152:H31 E
GCF_016026235.1 E.coli X O143:- D
GCF_013122565.1 E.coli X O3:H38 B1
GCF_017356665.1 E.coli UPEC O2/O50:H7 B2
GCA_003258455.1 E.coli AIEC O2/O50:H7 B2
GCA_000264175.1 E.coli AIEC O1:H7 B2
GCF_000285375.1 E.coli AIEC O1:H7 B2
GCF_002844685.1 E.coli ExPEC O2/O50:H7 B2
GCA_000148605.1 E.coli AIEC O18:H7 B2
GCF_001021615.1 E.coli APEC - B2
GCF_022493775.1 E.coli X O88:H14 B2
GCF_021398975.1 E.coli X O99:H6 B2
GCF_017901015.1 E.coli AIEC O2:H6 B2
GCA_000179795.1 E.coli AIEC O2:H6 B2
GCF_004564035.1 E.coli X O2:H6 B2
GCA_000183345.1 E.coli AIEC O83:H1 B2
GCF_000183345.1 E.coli AIEC O83:H1 B2
GCA_000284495.1 E.coli AIEC O83:H1 B2
GCF_007833875.1 E.coli UPEC O4:H5 B2
GCF_003019095.1 E.coli X O22:H5 B2
GCF_016767775.1 E.coli X O2:H5 B2
GCA_000264095.1 E.coli AIEC O8:H10 B2
GCF_022493515.1 E.coli X O8:H5 B2
ERR2265582 E.coli AIEC O46:H31 B2
ERR2265585 E.coli AIEC O6:H1 B2
ERR2265584 E.coli X O6:H1 B2
GCF_001693315.1 E.coli UPEC O6:H1 B2
GCF_002863845.1 E.coli STEC -:H18 B2
GCF_021398955.1 E.coli X O16:H6 B2
GCF_015697215.1 E.coli X O150:H6 B2
GCF_010668985.1 E.coli X O25:H31 B2
GCF_000013305.1 E.coli UPEC O6:H31 B2
GCF_019443685.1 E.coli UPEC O75:H5 B2
GCF_003812625.1 E.coli X O21:H7 B2
GCF_003346835.1 E.coli X O2/O50:H7 B2
GCF_016803935.1 E.coli X O84:H7 B2
GCF_902810315.1 E.coli X O71:H27 B2
GCF_004804185.1 E.coli X O63:H6 B2
GCF_003018475.1 E.coli X O145:H34 B2
GCA_001630915.1 E.coli CDEC O25:H4 B2
GCA_001630975.1 E.coli CDEC O25:H4 B2
GCF_016756455.1 E.coli X O25:H4 B2
GCF_003856995.1 E.coli UPEC O25:H4 B2
GCF_018771585.1 E.coli X O102:H6 D
GCF_019431315.1 E.coli X O139:H1 D
GCF_015571595.1 E.coli X O167:H26 D
GCF_902810335.1 E.coli X -:H48 G
GCF_003308975.1 E.coli X O153:H2 D

Figure 4. The phylogenomic tree of 240 species of E.coli including one Shigella, which was clustered in the same branch with E.coli strains. AIEC strains are clustered in two different branches of the tree where extra-intestinal pathotype species also exist.

4. CONCLUSIONS

The majority of AIEC strains were phylogenetically distant from enteropathogenic strains and the pathotypes ETEC, STEC, EHEC. The AIEC pathotype was related more to the ExPEC and UPEC pathotypes. It is possible that more strains that were included in this analysis will be defined as AIEC in the future. Some of those strains were isolated from the intestine of CD patients and it is not yet known if they are related with any kind of pathogenicity, subsequently in the analysis are classified as CDEC (Crohn's Disease *Escherichia coli*).

Provided that most of the AIEC strains that were used have not yet an available genome at the chromosome or complete genome assembly level, it is possible that some of the results related to the AIEC strains could theoretically be inaccurate. Consequently, many of the results of this analysis should be re-tested in the future when more complete genome of more AIEC strains become available, however, the lack of fingerprints at this point is not expected to change.

5. REFERENCES

Gordon, D.M., and Cowling, A. 2003. The distribution and genetic structure of Escherichia coli in Australian vertebrates: Host and geographic effects. Microbiology, 149(12):3575–3586. doi:10.1099/mic.0.26486-0. PMID:14663089.

Tenaillon, O., Skurnik, D., Picard, B., and Denamur, E. 2010.The population genetics of commensal Escherichia coli. Nat.Rev. Microbiol. 8(3): 207–217. doi:10.1038/nrmicro2298. PMID:20157339.

Bekal, S., Brousseau, R., Masson, L., Prefontaine, G.,Fairbrother, J., and Harel, J. 2003. Rapid identification ofEscherichia coli pathotypes by virulence gene detection with DNA microarrays. J. Clin. Microbiol. 41(5): 2113–2125.doi:10.1128/JCM.41.5.2113-2125.2003. PMID:12734257.

Huys, G., Cnockaert, M., Janda, J.M., and Swings, J. 2003. Escherichia albertii sp. nov., a diarrhoeagenic species isolated from stool specimens of Bangladeshi children. Int. J. Syst. Evol. Microbiol. 53(3): 807–810. doi:10.1099/ijs.0.02475-0. PMID:12807204.

Vernikos, G., Medini, D., Riley, D.R., Tettelin, H., 2015. Ten years of pan-genome analyses. Curr. Opin. Microbiol. 23, 148–154. https://doi.org/10.1016/j.mib.2014.11.016

Bosi, E., Fani, R., Fondi, M., 2015. Defining Orthologs and Pangenome Size Metrics, in: Mengoni, A.,Galardini, M., Fondi, M. (Eds.), Bacterial Pangenomics, Methods in Molecular Biology. Springer New York, New York, NY, pp. 191–202. https://doi.org/10.1007/978-1-4939-1720-4_13

Rasko, D.A., Rosovitz, M.J., Myers, G.S.A., Mongodin, E.F.,Fricke, W.F., Gajer, P., et al. 2008. The pangenome structure of Escherichia coli: Comparative genomic analysis of E. coli commensal and pathogenic isolates. J. Bacteriol. 190(20): 6881–6893. doi:10.1128/JB.00619-08. PMID:18676672.

Vernikos, G., Medini, D., Riley, D.R., Tettelin, H., 2015. Ten years of pan-genome analyses. Curr. Opin.Microbiol. 23, 148–154. https://doi.org/10.1016/j.mib.2014.11.016

Hernández-Salmerón, J.E., Moreno-Hagelsieb, G., 2020. Progress in quickly finding orthologs as reciprocal best hits: comparing blast, last, diamond and MMseqs2. BMC Genomics 21, 741.
https://doi.org/10.1186/s12864-020-07132-6

Nikolaidis, M., Mossialos, D., Oliver, S.G., Amoutzias, G.D., 2020. Comparative Analysis of the Core Proteomes among the Pseudomonas Major Evolutionary Groups Reveals Species-Specific Ad-
aptations for Pseudomonas aeruginosa and Pseudomonas chlororaphis. Diversity 12, 289.
https://doi.org/10.3390/d12080289

Castresana, J., 2000. Selection of Conserved Blocks from Multiple Alignments for Their Use in Phylogenetic Analysis. Molecular Biology and Evolution 17,540–552. https://doi.org/10.1093/oxfordjournals.molbev.a026334

Edgar, R.C., 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nu-
cleic Acids Research 32, 1792–1797. https://doi.org/10.1093/nar/gkh340

Gouy, M., Guindon, S., Gascuel, O., 2010. SeaView Version 4: A Multiplatform Graphical User Interface
for Sequence Alignment and Phylogenetic Tree Building. Molecular Biology and Evolution 27,
221–224. https://doi.org/10.1093/molbev/msp259

Arlette Darfeuille-Michaud, Adherent-invasive Escherichia coli: a putative new E. coli pathotype associated with Crohn's disease, International Journal of Medical Microbiology, Volume 292, Issues 3–4, 2002, Pages 185-193, ISSN 1438-4221, https://doi.org/10.1078/1438-4221-00201.

Liu, S., Feng, J., Pu, J., Xu, X., Lu, S., Yang, J., et al. 2019.Genomic and molecular characterisation of Escherichia marmotae from wild rodents in Qinghai-Tibet plateau as a potential pathogen. Sci. Rep. 9(1): 10619. doi:10.1038/s41598-019-46831-3. PMID:30626917.

Liu, S., Jin, D., Lan, R., Wang, Y., Meng, Q., Dai, H., et al.2015. Escherichia marmotae sp. nov., isolated from faeces of Marmota himalayana. Int. J. Syst. Evol. Microbiol. 65(7): 2130–2134. doi:10.1099/ijs.0.000228. PMID:25851592.

Devanga Ragupathi, N.K., Muthuirulandi Sethuvel, D.P.,Inbanathan, F.Y., and Veeraraghavan, B. 2018. Accurate differentiation of Escherichia coli and Shigella serogroups: challenges and strategies. ,New Microbes New Infect. 21: 58–62. doi:10.1016/j.nmni.2017.09.003. PMID:29204286.

Bekal, S., Brousseau, R., Masson, L., Prefontaine, G.,Fairbrother, J., and Harel, J. 2003. Rapid identification of Escherichia coli pathotypes by virulence gene detection with DNA microarrays. J. Clin. Microbiol. 41(5): 2113–2125. doi:10.1128/JCM.41.5.2113-2125.2003. PMID:12734257.

Frank C, Werber D, Cramer JP, et al.Epidemic profile of Shiga-toxin–producing Escherichia coli O104:H4 outbreak in Germany-preliminary report. N Engl J Med 2011. DOI: 10.1056/NEJMoa1106483.

Kai, A., Konishi, N., and Obata, H. 2010. [DiarrheagenicEscherichia coli.] Nihon Rinsho. Jpn. J. Clin. Med. 68(1): 203–207. [In Japanese.] PMID:20942038.

Barnich, N., and Darfeuille-Michaud, A. 2007. Adherent-invasive Escherichia coli and Crohn's disease. Curr. Opin. Gastroenterol. 23(1): 16–20. doi:10.1097/MOG.0b013e3280105a38. PMID: 17133079.

Terlizzi, M.E., Gribaudo, G., and Maffei, M.E. 2017. UroPathogenic Escherichia coli (UPEC) infections: Virulence factors, bladder responses, antibiotic, and non-antibiotic antimicrobial strategies. Front. Microbiol. 8: 1566. doi:10.3389/fmicb.2017.01566. PMID:28861072.

Mokady, D., Gophna, U., and Ron, E.Z. 2005. Virulence factors of septicemic Escherichia coli strains. Int. J. Med. Microbiol. 295(6–7): 455–462. doi:10.1016/j.ijmm.2005.07.007. PMID:6238019.

Scaletsky, I.C.A., Fabbricotti, S.H., Carvalho, R.L.B., Nunes, C.R.,Maranhão, H.S., Morais, M.B., and Fagundes-Neto, U. 2002.Diffusely adherent Escherichia coli as a cause of acute diarrhea
in young children in northeast Brazil: A case-control study.J. Clin. Microbiol. 40(2): 645–648. doi:10.1128/JCM.40.2.645-648.2002. PMID:11825986.

Dogan B, Suzuki H, Herlekar D, et al. Inflammation-associated adherent-invasive Escherichia coli are enriched in pathways for use of propanediol and iron and M-cell translocation. Inflamm Bowel Dis 2014;20:1919–32.

Molodecky NA, Soon IS, Rabi DM, Ghali WA, Ferris M, Chernoff G, et al. Increasing Incidence and Prevalence of the Inflammatory Bowel Diseases With Time, Based on Systematic Review. Gastroenterology. 2012;142:46–54.e42.

Tyakht, A.V., Manolov, A.I., Kanygina, A.V.*et al.*Genetic diversity of *Escherichia coli* in gut microbiota of patients with Crohn's disease discovered using metagenomic and genomic analyses. *BMC Genomics* 19, 968 (2018). https://doi.org/10.1186/s12864-018-5306-5

Yu D, Banting G, Neumann NF. A review of the taxonomy, genetics, and biology of the genus *Escherichia* and the type species *Escherichia coli*. Can J Microbiol. 2021 Aug;67(8): 553-571. Doi: 10.1139/cjm-2020-0508. Epub 2021 Mar 31. PMID: 33789061.

Enterprise Open Source and Linux [WWW Document], n.d. . Ubuntu. URL https://ubuntu.com/(accessed 2.8.21).

Welcome to Python.org [WWW Document], n.d. . Python.org. URL https://www.python.org/ (accessed 2.9.21).

Chevenet, F., Brun, C., Bañuls, A.-L., Jacq, B., Christen, R., 2006. TreeDyn: towards dynamic graphics and annotations for analyses of trees. BMC Bioinformatics 7, 439. https://doi.org/10.1186/1471-
2105-7-439

Miquel S, Peyretaillade E, Claret L, de Vallée A, Dossat C, Vacherie B, et al.Complete genome sequence of Crohn's disease-associated adherent invasive E. coli strain LF82. PLoS One. 2010;5. https://doi.org/10.1371/journal. Pone.0012714.

Desilets M, Deng X, Rao C, Ensminger AW, Krause DO, Sherman PM, Gray-Owen SD. Genome- based Definition of an Inflammatory Bowel Disease-associated Adherent-Invasive Escherichia coli Pathovar. Inflamm Bowel Dis. 2016 Jan;22(1):1-12. doi: 10.1097/MIB.0000000000000574. Erratum in: Inflamm Bowel Dis. 2016 Feb;22(2):E10. Deng, Xiangding [corrected to Deng, Xianding]. PMID: 26444104.

Beghain J, Bridier-Nahmias A, Le Nagard H, Denamur E, Clermont O. ClermonTyping: an easy-to-use and accurate in silico method for Escherichia genus strain phylotyping. Microb Genom. 2018 Jul;4(7):e000192. doi: 10.1099/mgen.0.000192. Epub 2018 Jun 19. PMID: 29916797; PMCID: PMC6113867.

Camprubí-Font, C., Lopez-Siles, M., Ferrer-Guixeras, M. *et al.* Comparative genomics reveals new single-nucleotide polymorphisms that can assist in identification of adherent-invasive *Escherichia coli*. *Sci Rep* 8, 2695 (2018). https://doi.org/10.1038/s41598-018-20843-x

Souvorov A, Agarwala R, Lipman DJ. SKESA: strategic k-mer extension for scrupulous assemblies. Genome Biol. 2018 Oct 4;19(1):153. doi: 10.1186/s13059-018-1540-z. PMID: 30286803; PMCID: PMC6172800.

Ciufo S, Kannan S, Sharma S, Badretdin A, Clark K, Turner S, Brover S, Schoch CL, Kimchi A, DiCuccio M. Using average nucleotide identity to improve taxonomic assignments in prokaryotic genomes at the NCBI. Int J Syst Evol Microbiol. 2018 Jul;68(7):2386-2392. doi: 10.1099/ijsem.0.002809. Epub 2018 May 24. PMID: 29792589; PMCID: PMC6978984.

Li W, O'Neill KR, Haft DH, DiCuccio M, Chetvernin V, Badretdin A, Coulouris G, Chitsaz F, Derbyshire MK, Durkin AS, Gonzales NR, Gwadz M, Lanczycki CJ, Song JS, Thanki N, Wang J, Yamashita RA, Yang M, Zheng C, Marchler-Bauer A, Thibaud-Nissen F. RefSeq: expanding the Prokaryotic Genome Annotation Pipeline reach with protein family model curation. Nucleic Acids Res. 2021 Jan 8;49(D1):D1020-D1028. doi: 10.1093/nar/gkaa1105. PMID: 33270901; PMCID: PMC7779008.

Fratamico PM, DebRoy C, Liu Y, Needleman DS, Baranzoni GM, Feng P. Advances in Molecular Serotyping and Subtyping of Escherichia coli. Front Microbiol. 2016 May 3;7:644. doi: 10.3389/fmicb.2016.00644. PMID: 27199968; PMCID: PMC4853403.

Croxen, M., Finlay, B. Molecular mechanisms of *Escherichia coli* pathogenicity. *Nat Rev Microbiol* 8, 26–38 (2010). https://doi.org/10.1038/nrmicro2265

Bessonov K, Laing C, Robertson J, Yong I, Ziebell K, Gannon VPJ, Nichani A, Arya G, Nash JHE, Christianson S. ECTyper: *in silico Escherichia coli* serotype and species prediction from raw and assembled whole-genome sequence data. Microb Genom. 2021 Dec;7(12):000728. doi: 10.1099/mgen.0.000728. PMID: 34860150; PMCID: PMC8767331.

M. Desilets, X. Deng, C. Rao, A.W. Ensminger, D.O. Krause, et al., Genome-based definition of an inflammatory bowel disease-associated adherent-invasive Escherichia coli pathovar, Inflamm. Bowel Dis. 22 (2016) 1–12.

H. Chen, D. Yang, F. Han, J. Tan, L. Zhang, et al., The Bacterial T6SS effector EvpP prevents NLRP3 inflammasome activation by inhibiting the Ca2+-dependent MAPK-Jnk pathway, Cell Host Microbe 21 (2017) 47–58.

R.J. Cieza, J. Hu, B.N. Ross, E. Sbrana, A.G. Torres, The IbeA invasin of adherent-invasive Escherichia coli mediates interaction with intestinal epithelia and macrophages, Infect. Immun. 83 (2015) 1904–1918.

A. Eshraghi, J. Kim, A.C. Walls, H.E. Ledvina, C.N. Miller, et al., Secreted effectors encoded within and outside of the Francisella Pathogenicity Island promote intramacrophage growth, Cell Host Microbe 20 (2016) 573–583.

T.G. Sana, C. Baumann, A. Merdes, C. Soscia, T. Rattei, et al., Internalization of Pseudomonas aeruginosa strain PAO1 into epithelial cells is promoted by interaction of a T6SS effector with the microtubule network, mBio. 6 (2015) e00712.

Peter J. A. Cock, Tiago Antao, Jeffrey T. Chang, Brad A. Chapman, Cymon J. Cox, Andrew Dalke, Iddo Friedberg, Thomas Hamelryck, Frank Kauff, Bartek Wilczynski, Michiel J. L. de

Hoon, Biopython: freely available Python tools for computational molecular biology and bioinformatics, *Bioinformatics*, Volume 25, Issue 11, 1 June 2009, Pages 1422–1423, https://doi.org/10.1093/bioinformatics/btp163

Camprubí-Font C, Lopez-Siles M, Ferrer-Guixeras M, Niubó-Carulla L, Abellà-Ametller C, Garcia-Gil LJ, Martinez-Medina M. Comparative genomics reveals new single-nucleotide polymorphisms that can assist in identification of adherent-invasive Escherichia coli. Sci Rep. 2018 Feb 9;8(1):2695. doi: 10.1038/s41598-018-20843-x. PMID: 29426864; PMCID: PMC5807354.

Krause DO, Little AC, Dowd SE, Bernstein CN. Complete genome sequence of adherent invasive Escherichia coli UM146 isolated from Ileal Crohn's disease biopsy tissue. J Bacteriol. 2011 Jan;193(2):583. doi: 10.1128/JB.01290-10. Epub 2010 Nov 12. PMID: 21075930; PMCID: PMC3019814.

Nash JH, Villegas A, Kropinski AM, Aguilar-Valenzuela R, Konczy P, Mascarenhas M, Ziebell K, Torres AG, Karmali MA, Coombes BK. Genome sequence of adherent-invasive Escherichia coli and comparative genomic analysis with other E. coli pathotypes. BMC Genomics. 2010 Nov 25;11:667. doi: 10.1186/1471-2164-11-667. PMID: 21108814; PMCID: PMC3091784.

Rakitina DV, Manolov AI, Kanygina AV, Garushyants SK, Baikova JP, Alexeev DG, Ladygina VG, Kostryukova ES, Larin AK, Semashko TA, Karpova IY, Babenko VV, Ismagilova RK, Malanin SY, Gelfand MS, Ilina EN, Gorodnichev RB, Lisitsyna ES, Aleshkin GI, Scherbakov PL, Khalif IL, Shapina MV, Maev IV, Andreev DN, Govorun VM. Genome analysis of E. coli isolated from Crohn's disease patients. BMC Genomics. 2017 Jul 19;18(1):544. doi: 10.1186/s12864-017-3917-x. PMID: 28724357; PMCID: PMC5517970.

Fang X, Monk JM, Nurk S, Akseshina M, Zhu Q, Gemmell C, Gianetto-Hill C, Leung N, Szubin R, Sanders J, Beck PL, Li W, Sandborn WJ, Gray-Owen SD, Knight R, Allen-Vercoe E, Palsson BO, Smarr L. Metagenomics-Based, Strain-Level Analysis of *Escherichia coli* From a Time-Series of Microbiome Samples From a Crohn's Disease Patient. Front Microbiol. 2018 Oct 30;9:2559. doi: 10.3389/fmicb.2018.02559. PMID: 30425690; PMCID: PMC6218438.

Clarke DJ, Chaudhuri RR, Martin HM, Campbell BJ, Rhodes JM, Constantinidou C, Pallen MJ, Loman NJ, Cunningham AF, Browning DF, Henderson IR. Complete genome sequence of the Crohn's disease-associated adherent-invasive Escherichia coli strain HM605. J Bacteriol. 2011 Sep;193(17):4540. doi: 10.1128/JB.05374-11. Epub 2011 Jun 24. PMID: 21705601; PMCID: PMC3165516.

O'Brien CL, Bringer M, Holt KE, et al. Comparative genomics of Crohn's disease-associated adherent-invasive *Escherichia coli* Gut 2017;66:1382-1389.

Zhou M, Yang Y, Chen P, Hu H, Hardwidge PR, Zhu G. More than a locomotive organelle: flagella in Escherichia coli. Appl Microbiol Biotechnol. 2015 Nov;99(21):8883-90. doi: 10.1007/s00253-015-6946-x. Epub 2015 Sep 8. PMID: 26346269.

Barnich N, Boudeau J, Claret L, Darfeuille-Michaud A. Regulatory and functional co-operation of flagella and type 1 pili in adhesive and invasive abilities of AIEC strain LF82 isolated from a patient with Crohn's disease. Mol Microbiol. 2003 May;48(3):781-94. doi: 10.1046/j.1365-2958.2003.03468.x. PMID: 12694621.

Wang J, Bleich RM, Zarmer S, Zhang S, Dogan B, Simpson KW, Arthur JC. Long-read sequencing to interrogate strain-level variation among adherent-invasive Escherichia coli isolated from human intestinal tissue. PLoS One. 2021 Oct 28;16(10):e0259141. doi: 10.1371/journal.pone.0259141. PMID: 34710159; PMCID: PMC855