

GENETICS AND TRANSCRIPTOMICS OF
ADHERENT-INVASIVE ESCHERICHIA COLI
(AIEC): NEW APPROACHES TO UNCOVER
MOLECULAR MARKERS FOR ITS RAPID
IDENTIFICATION

Carla Camprubí Font

Per citar o enllaçar aquest document:
Para citar o enlazar este documento:
Use this url to cite or link to this publication:
<http://hdl.handle.net/10803/672302>



<http://creativecommons.org/licenses/by-nc-nd/4.0/deed.ca>

Aquesta obra està subjecta a una llicència Creative Commons Reconeixement-
NoComercial-SenseObraDerivada

Esta obra está bajo una licencia Creative Commons Reconocimiento-NoComercial-
SinObraDerivada

This work is licensed under a Creative Commons Attribution-NonCommercial-
NoDerivatives licence

Doctoral thesis

Genetics and transcriptomics of
adherent-invasive *Escherichia coli* (AIEC):
New approaches to uncover molecular
markers for its rapid identification

CARLA CAMPRUBÍ FONT
2019



Doctoral thesis

**Genetics and transcriptomics of adherent-invasive
Escherichia coli (AIEC): new approaches to uncover
molecular markers for its rapid identification**

Carla Camprubí Font
2019



Doctoral thesis

**Genetics and transcriptomics of adherent-invasive
Escherichia coli (AIEC): new approaches to uncover
molecular markers for its rapid identification**

Carla Camprubí Font
2019

Doctorate program in Molecular Biology, Biomedicine and Health.

Thesis supervisor

Dr. Margarita Martinez Medina

Tutor

Dr. L. Jesús García Gil

PhD candidate

Carla Camprubí Font

The present thesis contain 34 supplementary material from chapters in printed and electronic format at the end of the document

This thesis is submitted in fulfilment of the requirements to obtain the doctoral degree from the Universitat de Girona

AGRAÏMENTS

Diuen que si camines sol vas ràpid, però si vas acompanyat arribes més lluny, i al llarg d'aquesta tesi no podria haver estat més ben acompanyada. Tots hi heu aportat el vostre granet de sorra, en diferent mesura però tots igual d'importants. Com podeu imaginar, aquestes han sigut segurament les paraules més difícils d'escriure. He intentat no descuidar-me ningú, en cas que si, perdoneu.

Com es tradició voldria començar per donar les gràcies a la meva directora, la Marga, per donar-me la oportunitat de fer aquesta tesi, per la seva confiança i per ensenyar-me que hi ha mil maneres de treballar. I al tutor, en Jesús, per haver acceptat la meva incorporació al grup.

Gràcies a la millor companya de grup que podria haver tingut. Mireia, gràcies per ensenyar-me tant, per aconsellar-me sempre i per estar preparada per arremangar-te en qualsevol moment. Tota la teva ajuda ha estat fonamental. Encara ara, crec que em queda molt per aprendre de tu.

Agrair també a tots els membres de l'àrea de microbiologia i en especial a les meves companyes de despatx. A l'Ellana per la seva energia i humor (*No lo pierdas nunca, Elianita!*), a l'Eli per la seva ajuda sempre que és necessària i a l'Elena per ensenyar-me tantes coses en aquestes últimes setmanes de docència! A la Sara i l'Imma per tots els consells en els primers passos. També a la Laia M per tot el seu suport tècnic al llarg d'aquests anys i per interessar-se sempre en la recerca. Agrair també totes les aportacions de l'Ari, la Txell, la Laura, l'Anna, la Míriam, l'Ahmed, la Natàlia, l'Eva, entre d'altres. I a la Marina i a la Carla S.

Gràcies també a tots els de la Granja i als que han estat allà cada migdia per compartir xerrades i desconnexions ja sigui en els primers com en els últims anys de tesi. Així doncs, una sincera abraçada pels bioquímics: Pedro (*son muchas las tardes de viernes en cultivos, mucha suerte y ánimos en la recta final*), Àlex (te'n dec una, aquella *E. coli* amb GFP em va salvar), Santi (per tot el teu bon rotllo), Montse (encara guardem aquells súper gorros), Txell, Anna, i Adrià; pels genètics: Luís (per acompanyar-me durant tot aquell maleit curs d'anglès) i Judit; pels Biocels: Pau B (qui avisa no és traïdor, diuen. Sort que em vas avisar de que l'RNA no seria fàcil!), Irene (per estar sempre disponible per discutir fórmules i gràfics) i Iker (per la teva energia a pàdel i per la sinceritat); i per les noves incorporacions: Núria i Queralt. Finalment, i de forma molt especial a la Laura i la Sandra, no penseu pas que em descuido de vosaltres! Ara mateix crec que no em surten les paraules ideals, però espero haver-ho demostrat en cada moment en el que heu estat allà. En resum, gràcies per estar el meu costat, per preocupar-vos i pel vostre suport.

A nivell de departament, voldria agrair-li totes les recomanacions i consells a la Jess (no sé pas què hauria fet amb els macròfags sense la teva ajuda! I gràcies per deixar-nos utilitzar el lab!), a la Mercè i l'Olga (per resoldre dubtes de RNA) i a la Sílvia (per participar a l'estudi de porines).

To all the M2iSH laboratory in Clermont Ferrand (Amélie, Allison, Alexis, Michael, Jérémy, Hang, Nicolas...), thanks to welcome me so well and to teach me all the secrets behind the realization of mutants.

Als de sempre. Marina i Núria per estar sempre al peu del canó, per tots els moments que hem compartit i el que ens queda! A la Mendo i en Xevi que tot i que ara ens veiem poc, em van patir fort durant la carrera i encara ara no han aconseguit treure-se'm de sobre. Prat, perquè tot i la distància, estàs molt a prop i sempre mires cap al futur.

Per acabar, i per sobre de tot, gràcies a la família. Al meus pares, per ensenyar-me que reculls el que sembres i que les coses no sempre són fàcils però que amb constància sempre hi ha una solució. I en fi, per fer-ho possible. A en Marc, no parís, tens una gran camí! I per acabar a en Pau, per tots els moments junts, per ensenyar-me tantes i tantes coses infinites i per cuidar-me dia darrera dia.

Em sento afortunada d'haver estat envoltada de tots vosaltres! Les meves sinceres gràcies a tots!

Carla

LIST OF PUBLICATIONS

Part of the results of this PhD Thesis has been published/submitted in scientific journals:

- Camprubí-Font, C., Lopez-Siles, M., Ferrer-Guixeras, M., Niubó-Carulla, L., Abellà-Ametller, C., Garcia-Gil, L.J., and Martinez-Medina, M. (2018) **Comparative genomics reveals new single-nucleotide polymorphisms that can assist in identification of adherent-invasive *Escherichia coli***. *Sci Rep* 8 (2695): 1-11.
- Camprubí-Font, C., Lopez-Siles, M., Ewers, C., and Martinez-Medina, M. (2019) **Genetic and phenotypic features to screen for putative adherent-invasive *Escherichia coli***. *Front. Microbiol.* 10 (108): 1-11.
- Camprubí-Font, C., Ruiz del Castillo, B., Barrabés Vera, S., Martínez Martínez, L., and Martinez-Medina, M. (in submission) **Amino acid substitutions and differential gene expression of outer membrane proteins in adherent-invasive *Escherichia coli***.

PATENT

Part of the results of this PhD thesis are included in a European patent application filed at the Spanish Patent and Trademarks Office (OEPM):

- Margarita Martinez-Medina, Mireia Lopez-Siles and Carla Camprubí-Font. **New molecular markers for the adherent-invasive *Escherichia coli* (AIEC) pathotype, related methods and kits**. Universitat de Girona. (Application number: P201830112)

ACKNOWLEDGMENTS

This work has been founded through the following research projects:

- SAF2017-82261-P. Study of new factors involved in the virulence of the Adherent-invasive *Escherichia coli* (AIEC) pathotype associated with Crohn's disease: transcriptomics and the role of outer membrane vesicles. Ministerio de Economía y Competitividad. PI: Dra. Margarita Martinez Medina. (Universitat de Girona). 2018-2021.
- GdRCompetUdG2017. Programa d'impuls a la recerca per als grups de recerca amb més projecció competitiva de la Universitat de Girona. PI: Dr. L. Jesús Garcia-Gil. (Universitat de Girona). 2017-2019.
- MPCUdG2016/009. Genòmica i transcriptòmica comparativa per identificar gens implicats en el fenotip AIEC. Universitat de Girona. PI: Dra. Margarita Martinez Medina. (Universitat de Girona). 2016-2018.
- SAF2013-43284-P. Research of genetic elements for the identification of the Adherent-invasive *Escherichia coli* (AIEC) based on comparative genomics and the analysis of their relevance in the AIEC pathogenicity. Ministerio de Economía y

Competitividad and co-funded by the European Regional Development Fund. PI: Dra. Margarita Martinez Medina. (Universitat de Girona). 2014-2016.

- SAF2010-15896. Adherent Invasive *Escherichia coli* (AIEC): distribution among intestinal diseases other than Crohn's disease and genes involved in its pathogenicity. Ministerio de Ciencia e Innovación. PI: Dr. L. Jesús Garcia Gil (Universitat de Girona). 2011 -2013.

During the PhD period, Carla Camprubí Font was awarded with:

- IF pre-doctoral grant from the Universitat de Girona. (IF-UdG 2015)

LIST OF ABBREVIATIONS

ABBREVIATION	DESCRIPTION
AB	Antibiotic
AdiA	Arginine decarboxylase A
AdiC	Arginine/agmatine antiporter C
AIEC	Adherent-invasive <i>E. coli</i>
AJ	Adherens junctions
AJC	Apical junctional complex
AMP	Antimicrobial peptides
APEC	Avian pathogenic <i>E. coli</i>
ATCC	American type culture collection
ATG16L1	Autophagy-related protein 16L1
ATG5	Autophagy-related protein 5
Bfp	Bundle-forming pili
BMDM	Bone marrow-derived macrophages
CD	Crohn's disease
CdtB	Cytotoxic distending toxin subunit B
CEACAM6	Carcinoembryonic antigen related cell adhesion molecule 6
CHI3L1	Chitinase 3-like 1 receptor
ChiA	Chitinase A
cJUN	Jun-related antigen, isoform C
CNF1	Cytotoxic necrotizing factor 1
CNF2	Cytotoxic necrotizing factor 2
ColV	Colicin V
CRC	Colorectal cancer
Ct	Comparative threshold cycle
CvaC	Colicin C precursor
DAEC	Diffusely adherent <i>E. coli</i>
DC	Dendritic cell
DEC	Diarrhoeagenic <i>E. coli</i>
DEGs	Differentially expressed genes
DNA	Deoxyribonucleic acid
EAEC	Enterohemorrhagic <i>E. coli</i>
EAST1	Enteraggregative heat-stable enterotoxin
EHEC	Enteraggregative <i>E. coli</i>
E-Hly	Enterohaemolysin
EIEC	Enteroinvasive <i>E. coli</i>
EltA	Enterotoxin A
EPEC	Enteropathogenic <i>E. coli</i>
ER	Endoplasmic reticulum
ETEC	Enterotoxigenic <i>E. coli</i>
ExPEC	Extraintestinal pathogenic <i>E. coli</i>
FBS	Fetal bovine serum
FhuD	Ferrichrome-binding periplasmic protein
FC	Fold-change
FimH	Type 1 fimbrial adhesin FimH
FocG	Minor fimbrial subunit G
FPKM	Fragments per kilobase of transcript per million mapped reads
FyuA	Ferric yersiniabactin uptake receptor A
GadA	Glutamate decarboxylase alpha
GadB	Glutamate decarboxylase beta
GC	Globet cell
GP2	Glycoprotein 2
Gp96	Endoplasmic reticulum-localised stress response chaperone
Gsp	Glutathionylspermidine synthase

ABBREVIATION	DESCRIPTION
Hcp	Hydroxylamine reductase
HM	Heptyl mannose derivatives
HMDM	Human monocyte-derived macrophages
IBD	Inflammatory bowel disease
I-CD	Ileal Crohn's disease
IECs	Intestinal epithelial cells
IEL	Intraepithelial lymphocyte
IESC	Intraepithelial stem cell
IFN γ	Interferon gamma
IL-1 β	Interleukin 1 beta
IL-8	Interleukin 8
IpaC	Invasion plasmid-coded protein antigen
IpaH	Invasion plasmid antigen H
IPEC	Intestinal pathogenic <i>E. coli</i>
IRGM	Immunity related GTPase M
IRP1	Iron regulatory protein 1
IRP2	Iron regulatory protein 2
Iss	Increased serum survival protein
I_ADH	Adhesion index
I_INV	Invasion index
I_REPL	Replication index
Lamp-1	Lysosomal-associated membrane protein-1
LB	Luria-Bertani
LpfA	Long polar fimbriae A
LT enterotoxin	Heat-labile enterotoxin
MAP	<i>Mycobacterium avium</i> subspecies <i>paratuberculosis</i>
M-cells	Microfold cells
MH	Mueller-Hinton
MIR130A	MicroRNA 130a
MIR30C	MicroRNA 30c
miRNAs	MicroRNAs
MLST	Multilocus sequence typing
MNEC	Meningitis-associated <i>E. coli</i>
MOI	Multiplicity of infection
NGS	Next-generation sequencing
NF- κ B	Nuclear factor kappa B
NLRs	Nucleotide-binding oligomerization domain (NOD)-like receptors
NOD2	Nucleotide-binding oligomerization domain 2
NT	Neutrophil
NTEC	Necrotoxic <i>E. coli</i>
OCG	Orthologous clusters of genes
OMPs	Outer membrane proteins
OmpA	Outer membrane protein A
OmpC	Outer membrane protein C
OmpF	Outer membrane protein F
OMVs	Outer membrane vesicles
p38	Protein kinase p38
PBS	Phosphate-buffered saline
PC	Paneth cell
PCR	Polymerase chain reaction
PduC	Propanediol dehydratase
PFGE	Pulsed-field gel electrophoresis
PIAS3	Protein inhibitor of activated STAT 3
PMA	Phorbol 12-myristate 13-acetate
PP	Paracellular pathway

ABBREVIATION	DESCRIPTION
PPDT	Particle-based photodynamic therapy
PRRs	Pattern-recognition receptors
RatA	Ribosome association toxin A
RNA	Ribonucleic acid
RNA-Seq	RNA sequencing
ROS	Reactive oxygen species
RTA	Relative abundance of transcripts
RT-qPCR	Reverse transcription quantitative polymerase chain reaction
ShET1	<i>Shigella</i> enterotoxin subunit 1
ShET2	<i>Shigella</i> enterotoxin subunit 2
SNPs	Single-nucleotide polymorphisms
ST	Sequence type
ST enterotoxin	Heat-stable enterotoxin
STAT1	Signal transducer and activator of transcription 1
STEC	Shiga toxin <i>E. coli</i>
Stx	Shiga-like toxin
Stx1	Shiga toxin 1
Stx2	Shiga toxin 2
SUMO	Small ubiquitin-like modifier
TEER	Trans epithelial electrical resistance
TJ	Tight junction
TLRs	Toll-like receptors
TNF	Tumour necrosis factor
TP	Transcellular pathway
Tsh	Temperature-sensitive hemagglutinin autotransporter
UC	Ulcerative colitis
UPGMA	Unweighted pair group method with arithmetic mean
Usp	Uropathogenic specific protein
UTR	Untranslated region
UPEC	Uropathogenic <i>E. coli</i>
Vat	Vacuolating autotransporter toxin
VGs	Virulence genes
WT	Wild type

LIST OF FIGURES

INTRODUCTION

- Figure 1.** Transmission electron micrographs of AIEC LF82 strain invading intestinal epithelial cells and replicating within macrophages. 4
- Figure 2.** Worldwide incidence of CD from 1990 to 2016. 5
- Figure 3.** Schematic representation of the epithelial barrier system in CD. 8
- Figure 4.** AIEC mechanisms of pathogenicity. 23

MATERIALS & METHODS

CHAPTER 2.2: Construction of isogenic mutants to study the role in pathogenicity of three genes related to AIEC pathotype

- Figure 5.** Principle of the first (50 bp homology) and second (Three-step PCR, 500 bp homology) approaches used to construct isogenic mutants. 58

RESULTS & DISCUSSION

CHAPTER 1.1: Virulence gene carriage and adhesin variants of AIEC and commensals isolated from humans and animals

- Figure 6.** Distribution of virulence genes prevalence according to origin of isolation. 66
- Figure 7.** Distribution of virulence genes prevalence according to pathotype. 67
- Figure 8.** Reticulate tree representing FimH (A) and ChiA (B) variants. 72

CHAPTER 1.2: Amino acid substitutions and differential gene expression of outer membrane proteins in AIEC

- Figure 9.** Reticulated trees representing the distribution of strains carrying specific amino acid substitutions in OMPs. 78
- Figure 10.** Protein OMPs expression profiles in AIEC and non-AIEC strains growing in MH broth. 80
- Figure 11.** Paired tests evaluating the OMPs expression difference between supernatant and cell-associated fractions of infected I-407 cultures in each strain collection (AIEC and non-AIEC). 81
- Figure 12.** Differential OMPs expression between AIEC and non-AIEC strains from each fraction of the infected I-407 cultures (SN and INV). 82
- Figure 13.** Schematic representation of gene expression levels of OMPs according to pathotype and fraction analysed. 86

CHAPTER 2.1: Identification by comparative genomics of new single nucleotide polymorphisms to distinguish between AIEC and non-AIEC strains

- Figure 14.** Genome similarities among the six strains and within each pair by analysis of orthologous clusters of genes (OCG) analysis. 89
- Figure 15.** Adhesion (a) and invasion (b) abilities of the strains according to specific nucleotide variants of SNPs. 95
- Figure 16.** Classification algorithm for AIEC identification. 97

CHAPTER 2.2: Construction of isogenic mutants to study the role in pathogenicity of three genes related to AIEC pathotype

- Figure 17.** Agarose gel electrophoresis (1%) of PCR products obtained in each PCR reaction. 105
- Figure 18.** Phenotypic characteristics of the mutant in comparison to the wild type strain. 106
- Figure 19.** Example of electropherogram. 111

CHAPTER 3.1: RNA-Seq analysis of the transcriptome during growth in cell culture media and during intestinal epithelial cell infection of AIEC in comparison with non-AIEC strains

- Figure 20.** Volcano plots of the $-\log_{10}(\text{p-values})$ versus the $\log_2 \text{FC}$. 113
- Figure 21.** Differentially expressed genes during AIEC growth in suspension (SN) and during IECs infection (INV), relative to its non-AIEC counterpart. 115
- Figure 22.** Predicted function of the differentially expressed genes in each comparison distributed in seven functional categories. 116
- Figure 23.** Correlation between the \log_2 fold-change (FC) values obtained by RNA-seq and those obtained by RT-qPCR. 118

GENERAL DISCUSSION

- Figure 24.** Review of cell lines used for AIEC identification. 133

SUPPLEMENTAL MATERIALS

METHODS 2.1: Identification by comparative genomics of new single nucleotide polymorphisms to distinguish between AIEC and non-AIEC strains

- Figure S1.** Consensus UPGMA dendrogram generated from the Pearson correlation coefficients of XbaI PFGE profiles of the three pair of strains selected for genome sequencing. 154

RESULTS 1.1: Virulence gene carriage and adhesin variants of AIEC and commensals isolated from humans and animals

- Figure S2.** Venn diagram depicting the virulence genes statistically more prevalent in human or animal-isolated strains (A) or in AIEC or non-AIEC strains (B). 155

RESULTS 1.2: Amino acid substitutions and differential gene expression of outer membrane proteins in AIEC

- Figure S3.** Prevalence of previously detected mutations in OmpA according to pathotype. 155
- Figure S4.** Correlations between OMPs expression and AIEC phenotypic characteristics (adhesiveness and invasiveness) in the SN fraction. 156
- Figure S5.** Correlations between OMPs expression and AIEC phenotypic characteristics (adhesiveness and invasiveness) in the INV fraction. 157

RESULTS 2.1: Identification by comparative genomics of new single nucleotide polymorphisms to distinguish between AIEC and non-AIEC strains

- Figure S6.** Whole genome map comparison of AIEC/non-AIEC strains with MAUVE 2.3. 158

RESULTS 2.2: Construction of isogenic mutants to study the role in pathogenicity of three genes related to AIEC pathotype

- Figure S7.** Schematic view of the regions where the primers were designed for 4.3 mutant. ●
- Figure S8.** Schematic view of the regions where the primers were designed for 4.4 mutant. ●

RESULTS 3.1: RNA-Seq analysis of the transcriptome during growth in cell culture media and during intestinal epithelial cell infection of AIEC in comparison with non-AIEC strains

- Figure S9.** Denaturing agarose gel showing RNA integrity of two SN and INV samples isolated with two different kits. 159

(●) Supplementary materials that are not included in the printed version of this thesis for space reasons but can be found at the end of the document in electronic format (CD-ROM attached).

LIST OF TABLES

INTRODUCTION

Table 1. General characteristics and main virulence genes of human ExPEC, DEC and AIEC pathotypes.	3
Table 2. Studies where AIEC prevalence has been analysed in ileal and colonic samples from subjects with IBD and controls.	15
Table 3. Summary of virulence factors related to AIEC virulence.	17
Table 4. Review of studies in which the prevalence of particular VGs has been examined according to the AIEC pathotype and origin of isolation.	31
Table 5. Summary of the comparative genomics studies conducted in AIEC to date.	34

MATERIALS & METHODS

CHAPTER 1.1: Virulence gene carriage and adhesin variants of AIEC and commensals isolated from humans and animals

Table 6. Strain collection used to study virulence gene carriage, gene sequence variants and the combination of <i>pic</i> gene and ampicillin resistance according to host, disease and pathotype.	42
--	----

CHAPTER 1.2: Amino acid substitutions and differential gene expression of outer membrane proteins in AIEC

Table 7. Primers and probes used to amplify, sequence and to analyse differential expression of <i>ompA</i> , <i>ompC</i> and <i>ompF</i> genes.	46
---	----

CHAPTER 2.1: Identification by comparative genomics of new single nucleotide polymorphisms to distinguish between AIEC and non-AIEC strains

Table 8. Characteristics of the three sequenced AIEC/non-AIEC strain pairs.	52
Table 9. Primers and PCR conditions used to amplify fragments of the genes in which the Confirmed SNPs were located.	55

CHAPTER 2.2: Construction of isogenic mutants to study the role in pathogenicity of three genes related to AIEC pathotype

Table 10. Characteristics of the strains selected for the construction of isogenic mutants (AIEC) and the candidate non-AIEC (ECG) strains with which isogenic mutants would be compared.	57
Table 11. Primers used for the construction of isogenic mutants.	59

RESULTS & DISCUSSION

CHAPTER 1.1: Virulence gene carriage and adhesin variants of AIEC and commensals isolated from humans and animals

Table 12. Frequency of amino acid substitutions for FimH and ChiA proteins in relation to pathotype and phylogroup.	71
Table 13. Binary logistic regression model evaluating the prevalence of the <i>pic</i> gene and ampicillin resistance as a putative model for AIEC identification.	73

CHAPTER 1.2: Amino acid substitutions and differential gene expression of outer membrane proteins in AIEC

Table 14. Variable positions in OMPs sequence related to pathotype or AIEC phenotypic characteristics.	79
---	----

CHAPTER 2.1: Identification by comparative genomics of new single nucleotide polymorphisms to distinguish between AIEC and non-AIEC strains

Table 15. Assembly features of the AIEC/non-AIEC sequenced genomes.	87
Table 16. Number of SNPs present in the sequenced AIEC/non-AIEC strain pairs.	90

Table 17. Location of the Confirmed SNPs, nucleotide variants and gene functions. 92

Table 18. Prevalence of genes encompassing SNPs in a collection of AIEC/non-AIEC strains and the frequency of particular nucleotide variants with respect to AIEC phenotype and phylogroup origin of the strains. 93

Table 19. Binary logistic regression model for the SNPs associated with the AIEC pathotype. 96

Table 20. Accuracy of the algorithm in strain collections from diverse geographic origins. 99

CHAPTER 3.1: RNA-Seq analysis of the transcriptome during growth in cell culture media and during intestinal epithelial cell infection of AIEC in comparison with non-AIEC strains

Table 21. Experimental approaches followed to optimise the total RNA extraction during AIEC17 growth in cell culture media and during infection of intestinal epithelial cells (I-407). 109

Table 22. Example of qualitative and quantitative data at each step of the RNA extraction procedure of AIEC/non-AIEC samples for both fractions (Supernatant (SN) and Invasion (INV)). 110

Table 23. General characteristics of AIEC and non-AIEC transcriptomes. 112

Table 24. Relative gene expression values assessed by RNA-Seq and Fluidigm/RT-qPCR. 117

GENERAL DISCUSSION

Table 25. Genetic elements more frequently found in strains from the AIEC pathotype and suggested as putative AIEC molecular markers. 128

Table 26. Comparison of the principal experimental conditions of the protocols used to assess bacterial invasion to intestinal epithelial cells and survival and replication inside macrophages. 132

SUPPLEMENTAL MATERIALS

METHODS 1.1: Virulence gene carriage and adhesin variants of AIEC and commensals isolated from humans and animals

Table S1. Information of the patients from whom the UC and CRC strains were isolated. 160

Table S2. Information about the strains analysed in chapter 1.1. ●

Table S3. Distribution of the phylogenetic origin of the strains according to pathotype (A) or origin of isolation (B) in each group of study. 161

METHODS 1.2: Amino acid substitutions and differential gene expression of outer membrane proteins in AIEC

Table S4. Information about the strains analysed in chapter 1.2. ●

METHODS 2.1: Identification by comparative genomics of new single nucleotide polymorphisms to distinguish between AIEC and non-AIEC strains

Table S5. Information on the patients from whom the studied strains in chapter 2.1 were isolated. ●

Table S6. Information about the strains analysed in chapter 2.1. ●

Table S7. Information about the external strains analysed for algorithm validation in chapter 2.1. ●

METHODS 3.1: RNA-Seq analysis of the transcriptome during growth in cell culture media and during intestinal epithelial cell infection of AIEC in comparison with non-AIEC strains

Table S8. Primers used in the study to amplify and analyse differential expression of the selected genes. 162

RESULTS 1.1: Virulence gene carriage and adhesin variants of AIEC and commensals isolated from humans and animals

- **Table S9.** Prevalence of virulence-associated genes according to origin of isolation (all strains, AIEC or non-AIEC) considering all strains or B2 strains.
- **Table S10.** Prevalence of virulence-associated genes according to phylogroup in strains isolated from humans and/or animals and AIEC/non-AIEC.
- **Table S11.** Prevalence of virulence-associated genes according to pathotype in all or B2 strains isolated from humans and/or animals.
- 163 **Table S12.** Distribution of FimH amino acid substitutions among the strain collection.
- 164 **Table S13.** Distribution of ChiA amino acid substitutions among the strain collection.

RESULTS 1.2: Amino acid substitutions and differential gene expression of outer membrane proteins in AIEC

- 165 **Table S14.** Distribution of OmpA amino acid substitutions among the strain collection.
- 166 **Table S15.** Distribution of OmpC amino acid substitutions among the strain collection.
- 167 **Table S16.** Distribution of OmpF amino acid substitutions among the strain collection.
- **Table S17.** Prevalence of amino acid mutations in a collection of AIEC/non-AIEC strains regarding AIEC phenotype and phylogroup origin of the strains.
- 168 **Table S18.** OMPs gene expression according to the phylogenetic origin of the strains.

RESULTS 2.1: Identification by comparative genomics of new single nucleotide polymorphisms to distinguish between AIEC and non-AIEC strains

- 169 **Table S19.** Distribution of amino acid substitutions in three genes previously associated with AIEC pathogenesis in our AIEC/non-AIEC strain pairs.
- **Table S20.** Number of orthologous clusters of genes for each strain and specific clusters present in each strain compared to its counterpart.
- **Table S21.** Total SNPs recovered by comparative genomics between AIEC17 and ECG28 strains.
- **Table S22.** Total SNPs recovered by comparative genomics between AIEC01 and ECG11 strains.
- **Table S23.** Total SNPs recovered by comparative genomics between AIEC07 and ECG04 strains.

RESULTS 3.1: RNA-Seq analysis of the transcriptome during growth in cell culture media and during intestinal epithelial cell infection of AIEC in comparison with non-AIEC strains

- **Table S24.** Differentially expressed genes between AIEC and non-AIEC strains in two conditions (SN and INV).
- 170 **Table S25.** Level of expression of genes previously associated with AIEC pathogenesis in AIEC and non-AIEC strains analysed in this study.

(●) Supplementary materials that are not included in the printed version of this thesis for space reasons but can be found at the end of the document in electronic format (CD-ROM attached).

TABLE OF CONTENTS

Table of abbreviations	i
List of figures	v
List of tables	vii
Table of contents	xi
Summary	xv
Resum	xix
Resumen	xxiii
INTRODUCTION	1
1. <i>Escherichia coli</i> in human gut	1
1.1. Pathogenic groups of <i>E. coli</i>	1
1.2. Adherent invasive <i>E. coli</i> (AIEC) definition	2
2. AIEC and Crohn's disease	4
2.1. Crohn's disease	4
2.2. AIEC and Crohn's disease pathogenesis	6
2.3. Crohn's disease therapies and their effects on AIEC	10
2.4. AIEC prevalence in Crohn's disease and other gastrointestinal disorders	14
3. AIEC pathogenicity and virulence factors involved	17
3.1. AIEC and intestinal epithelial barrier	18
3.2. AIEC and immune cells	23
3.3. AIEC and autophagy	26
3.4. AIEC and biofilm formation	27
3.5. Other pathogenic mechanisms	27
4. AIEC genetic markers	29
4.1. PCR-based gene prevalence	30
4.2. Pathoadaptative mutations	32
4.3. Comparative genomics	33
4.4. Gene expression and comparative transcriptomics	35
AIMS & SCOPE OF THE THESIS	37
MATERIALS & METHODS	41
CHAPTER 1.1: Virulence gene carriage and adhesin variants of AIEC and commensals isolated from humans and animals	41
1.1 <i>E. coli</i> strain collection	41
1.2 Adhesion and invasion assays	42
1.3 Survival and replication within macrophages	43
1.4 Virulence genotyping by PCR	44
1.5 Gene sequencing and sequence analysis	44
1.6 Antibiotic resistance	44
1.7 Statistical analysis	45
CHAPTER 1.2: Amino acid substitutions and differential gene expression of outer membrane proteins in AIEC	45
2.1 <i>E.coli</i> strain collection	45

2.2 Amplification and gene sequencing	45
2.3 Sequence analysis	46
2.4 OMPs isolation and separation by SDS-PAGE	47
2.5 Infection and RNA extraction	47
2.6 Gene expression quantification by RT-qPCR	48
2.7 Statistical analysis	49
CHAPTER 2.1: Identification by comparative genomics of new single nucleotide polymorphisms to distinguish between AIEC and non-AIEC strains	49
3.1 <i>E. coli</i> strain selection and characterisation	49
3.2 Genomic DNA extraction and sequencing	53
3.3 <i>De novo</i> genome assembly	54
3.4 Comparative genomics of strain pairs (gene structure, gene contents and SNPs)	54
3.5 Distribution of SNPs among a collection of strains	56
3.6 Algorithm validation in external strain collections	56
3.7 Statistical analysis	56
CHAPTER 2.2: Construction of isogenic mutants to study the role in pathogenicity of three genes related to AIEC pathotype	57
4.1 Bacterial strains	57
4.2 Plasmid transformation in <i>E.coli</i>	57
4.3 Construction of PCR product for gene disruption	58
4.4 Electrocompetent cells and gene disruption	60
4.5 Phenotypic characterisation	60
CHAPTER 3.1: RNA-Seq analysis of the transcriptome during growth in cell culture media and during intestinal epithelial cell infection of AIEC in comparison with non-AIEC strains	61
5.1 Bacterial strains, cell line and growth conditions	61
5.2 Infection of intestinal epithelial cells	61
5.3 RNA extraction and purification	61
5.4 Sequencing and RNA-seq analysis	62
5.5 Gene expression validation	63
5.6 Statistical analysis	64
	65
RESULTS & DISCUSSION	
CHAPTER 1.1: Virulence gene carriage and adhesin variants of AIEC and commensals isolated from humans and animals	65
Results	65
1.1 Virulence gene repertoires	65
1.1.1 Animal vs Human <i>E. coli</i> strains	65
1.1.2 AIEC vs non-AIEC strains	66
1.1.3 Crohn's disease vs Controls	69
1.2 FimH and ChiA amino acid substitutions	70
1.3 Test for rapid AIEC identification	72
Discussion	73

CHAPTER 1.2: Amino acid substitutions and differential gene expression of outer membrane proteins in AIEC	77
Results	77
2.1 OMPs sequence variants	77
2.2 Distribution of amino acid substitutions in OMPs	78
2.3 OMPs protein expression	79
2.4 OMPs gene expression	80
Discussion	83
CHAPTER 2.1: Identification by comparative genomics of new single nucleotide polymorphisms to distinguish between AIEC and non-AIEC strains	87
Results	87
3.1 Characteristics of strain pairs	87
3.2 Comparative genomics of AIEC/non-AIEC strain pairs	88
3.2.1 Evaluation of gene content dissimilarities	88
3.2.2 Detection of AIEC-associated SNPs	89
3.3 Distribution of SNPs in an AIEC/non-AIEC strain collection	90
3.4 SNPs in relation to adhesion and invasion capacity	94
3.5 Usefulness of SNPs as molecular signatures for AIEC screening	94
3.6 Validation of the tool in external strain collections	97
Discussion	99
CHAPTER 2.2: Construction of isogenic mutants to study the role in pathogenicity of three genes related to AIEC pathotype	103
Results	103
4.1 Construction of isogenic mutants	103
4.2. Phenotypic characterization of LF82Δ3.16	104
Discussion	106
CHAPTER 3.1: RNA-Seq analysis of the transcriptome during growth in cell culture media and during intestinal epithelial cell infection of AIEC in comparison with non-AIEC strains	108
Results	108
5.1 Protocol optimization	108
5.2 RNA-seq analysis to detect differential expressed genes	111
5.2.1 Overview of transcriptomics analysis	111
5.2.2 Differentially expressed genes during AIEC growth in supernatants and during IECs infection, relative to its non-AIEC counterpart	112
5.2.3 RNA-seq validation	115
Discussion	119
GENERAL DISCUSSION	123
1. Approaches followed to decipher AIEC genetics	124
2. Putative biomarkers to assist AIEC identification	127
3. Possible reasons why the search for AIEC molecular markers is challenging	129
4. Is the AIEC phenotype an acquired trait of <i>E. coli</i> strains from the gut?	133

5. Concluding remarks and future directions	135
CONCLUSIONS	137
REFERENCES	141
SUPPLEMENTAL MATERIALS	155

SUMMARY

High prevalence of the adherent-invasive *Escherichia coli* (AIEC) pathotype has been reported in the gut of adult and pediatric Crohn's disease patients by several independent studies. This pathotype is characterised by its ability to adhere to and invade intestinal epithelial cells, as well as, to survive and replicate inside macrophages without triggering host-cell death. Although many approaches have been conducted in order to identify the genetic basis of AIEC phenotype so far, an AIEC molecular biomarker is still missing. At present its identification relies on phenotypic traits undergoing cell-culture infection assays, which are extremely time consuming and hard to standardise. The finding of molecular tools or rapid tests to easily identify the AIEC pathotype would definitely be of interest for scientists studying the epidemiology of the pathotype and clinicians that aim to detect which patients are colonised by AIEC to apply personalised treatments. Therefore, in this thesis we principally aimed to better define the genetic characteristics of AIEC pathotype and to find putative genetic/phenotypic markers for its rapid identification. Three different approaches have been followed to achieve this purpose (chapter 1, 2 and 3).

The prevalence of 61 previously described virulence genes (VGs), point mutations in AIEC-associated genes (*fimH*, *chiA*, *ompA*, *ompC* and *ompF*) and differences in *ompA*, *ompC* and *ompF* gene expression has been assessed in a collection of AIEC/non-AIEC strains isolated from animals and/or humans (**chapter 1.1 and 1.2**). Animal strains were enriched in 12 VGs while 7 VGs were more predominant in human strains. The prevalence of 15 VGs was higher in AIEC than in non-AIEC strains, but only *pic* gene was still differentially distributed when analyzing human and animal strains separately. Among human strains, three additional VGs (*papGII/III*, *iss* and *vat*) were more prevalent in AIEC than in non-AIEC strains. Nevertheless, 25 of the studied genes also reported different prevalence regarding the phylogenetic origin of the strains. Besides, no differences in pathoadaptative mutations in any of the genes studied were suitable as molecular markers but some might be implicated in AIEC virulence (FimH-A119V, OmpA-A200V, OmpC-V220I and D232A, OmpF-E51V and M60K). Similarly, *ompA*, *ompC* and *ompF* gene expression during infection may be contributing to AIEC pathogenicity by enhancing intestinal epithelial cells adherence and intracellular persistence. Indeed, AIEC gene expression levels increased while growing in the supernatant of infected cell cultures and decreased while adhering and invading intestinal epithelial cells in comparison to non-AIEC. In these chapters, despite

no specific and widely distributed AIEC feature has been found, the combination of the *pic* gene prevalence and ampicillin resistance presented 75% of accuracy in AIEC screening.

Differences in gene content and single nucleotide polymorphisms (SNPs) in three pairs of strains have been studied by comparative genomics (**chapter 2.1**). In contrast with previous studies, each strain pair consisted of one AIEC and one non-AIEC that are considered clones with respect to their pulsed-field gel electrophoresis patterns. Three SNPs positions (E3-E4_4.3(2), E3-E4_4.4 and E5-E6_3.16=3.22(2)) presented differential distribution of nucleotide variants according to pathotype and four associated with increased adhesion and/or invasion indices (E3-E4_4.3(2), E3-E4_4.4 and E5-E6_3.16=3.22(2) and (3)). However, their implication in the AIEC phenotype could not be demonstrated either because isogenic mutants were not obtained or the deletion of the gene did not result in any perceivable phenotypic effect (**chapter 2.2**). Interestingly, our data revealed three SNPs that can be implemented in AIEC identification. Although this method does not correctly classify all *E. coli* strains, its accuracy in Spanish (Girona and Mallorca) isolates is very high (81%), and no comparable molecular tools currently exist.

In the third part of this work (**chapter 3.1**), a comparative transcriptome analysis of two AIEC/non-AIEC strain pairs during intestinal epithelial cells infection has been conducted. First, a protocol to extract and purify intracellular bacterial RNA has been optimised by adjusting sample quantity or washing steps to avoid kit saturation and ensure proper RNA quality. Finally, comparative analysis evidenced the presence of strain-specific differentially expressed genes rather than key genes associated with the AIEC pathotype, since no common differentially expressed gene was found between AIEC strains.

Although what constitutes an AIEC strain remains an enigma, the results of this work provide meaningful information that contributes to our understanding of AIEC genomics. Gene prevalence and amino acid substitutions results confirm the high genetic variability of AIEC strains and suggest that many of the genetic features described to date are in fact associated with phylogroup origin of the strains rather than with AIEC phenotype. Additionally, this work further reinforces the idea that no particular VG is related to AIEC phenotype. Despite diverse virulence factors could drive to the same phenotype, the presence of an AIEC-specific marker cannot be fully discarded. Herein, two putative molecular markers resulting from a combination of genetic and/or phenotypic features have been presented and these could assist in AIEC screening at least in our strain collection. Finally, we present for the first time two studies analysing AIEC gene

expression using an *in vitro* assay that simulates bacterial adhesion to and invasion of intestinal epithelial cells. In fact, gene expression results provide new insights to better describe genes putatively involved in AIEC virulence.

RESUM

El patotip *Escherichia coli* adherent-invasiu es troba de forma més freqüent en l'intestí de pacients adults i pediàtrics amb malaltia de Crohn que en controls de diferents països. Aquest patotip es caracteritza per la seva capacitat d'adherir-se i envair cèl·lules de l'epiteli intestinal i de sobreviure i replicar-se dins de macròfags sense induir la mort de la cèl·lula hoste. Tot i que fins al moment s'han dut a terme diverses aproximacions amb l'objectiu d'identificar les bases genètiques del fenotip AIEC, encara no existeix un biomarcador molecular específic del patotip. Actualment, la identificació d'aquest es basa en detectar trets fenotípics mitjançant assajos d'infecció de cultius cel·lulars, els quals són extremadament lents i difícils d'estandarditzar. La troballa d'eines moleculars o proves ràpides per identificar fàcilment el patotip AIEC seria sens dubte d'interès per als científics que estudien l'epidemiologia del patotip i els clínics que pretenen detectar quins pacients són colonitzats per AIEC per tal d'aplicar tractaments personalitzats. Per aquest motiu, en aquesta tesi, es tracta principalment de definir millor les característiques genètiques del patotip AIEC i de trobar possibles marcadors genètics/fenotípics per a la seva ràpida identificació. Per assolir aquest objectiu s'han seguit tres enfocaments diferents (capítol 1, 2 i 3).

S'ha estudiat la prevalença de 61 gens de virulència prèviament descrits, mutacions puntuals en gens associats a AIEC (*fimH*, *chiA*, *ompA*, *ompC* i *ompF*) i diferències en l'expressió gènica d'*ompA*, *ompC* i *ompF* en una col·lecció de soques AIEC/no-AIEC aïllades d'animals i/o humans (capítol 1.1 i 1.2). Dotze dels gens estudiats van ser més freqüents en soques aïllades d'animals mentre que 7 ho van ser en soques d'humans. Pel que fa al patotip, 15 gens eren més prevalents en AIEC que en no-AIEC, però només el gen *pic* es va mantenir diferencialment distribuït quan es van analitzar les soques separades per origen. En les soques d'humans, tres gens addicionals (*papGII/III*, *iss* i *vat*) eren més freqüents en les soques AIEC que en les no-AIEC. No obstant, 25 dels gens estudiats també presentaven diferències de prevalença segons l'origen filogenètic de la soca. A més, cap de les mutacions puntuals identificades en els gens d'estudi van ser adequades com a marcadors moleculars però podrien estar relacionades amb la virulència de les AIEC (FimH-A119V, OmpA-A200V, OmpC-V220I i D232A, OmpF-E51V i M60K). De la mateixa manera, l'expressió dels gens *ompA*, *ompC* i *ompF* durant la infecció pot contribuir a la patogenicitat de l'AIEC mitjançant la millora de l'adhesió a les cèl·lules epitelials intestinals i la persistència intracel·lular. De fet, els nivells d'expressió d'aquests gens en les soques AIEC van

augmentar mentre creixien en el sobrenedant de cultius cel·lulars infectats i es van reduir mentre adherien i/o envaïen les cèl·lules epitelials intestinals en comparació amb les no-AIEC. En aquests subcapítols, tot i que no s'ha trobat cap característica específica i àmpliament distribuïda del patotip AIEC, la combinació de la prevalença del gen *pil* i la resistència a ampicil·lina presenten un 75% de precisió en la detecció de AIEC.

Mitjançant genòmica comparativa s'han analitzat diferències del contingut genèic i polimorfismes d'un sol nucleòtid en tres parells de soques (**capítol 2.1**). A diferència d'estudis previs, cada parell de soca consistia en una soca AIEC i una no-AIEC que són considerades clons en base al seu patró de camp pulsant. Tres posicions amb variació a nivell de nucleòtid (E3-E4_4.3(2), E3-E4_4.4 i E5-E6_3.16=3.22(2)) van presentar una distribució diferencial segons patotip i quatre es van associar a una major capacitat d'adhesió i/o invasió (E3-E4_4.3(2), E3-E4_4.4 i E5-E6_3.16=3.22(2) i (3)). De totes maneres, la seva implicació en el fenotip AIEC no es va poder demostrar ja que no es van poder obtenir els mutants isogènics de dos gens i la delecció del gen 3.16=3.22 no va presentar cap efecte observable en el fenotip (**capítol 2.2**). De forma interessant, els nostres resultats van revelar tres posicions que es podrien utilitzar per la identificació de les AIEC. Tot i que aquesta aproximació no classifica correctament totes les *E. coli*, la seva precisió en soques aïllades d'Espanya (Girona i Mallorca) és molt elevada (81%), i actualment no hi ha cap eina molecular comparable.

A la tercera part d'aquesta tesi (**capítol 3.1**) s'ha dut a terme un anàlisi transcriptòmic de dues parelles AIEC/no-AIEC durant la infecció de cèl·lules de l'epiteli intestinal. En primer lloc, es va optimitzar un protocol d'extracció i purificació del àcid ribonuclèic (RNA) ajustant la quantitat de mostra processada o realitzant passos de neteja per així evitar la saturació del kit i assegurar una bona qualitat del RNA. Finalment, els resultats de l'estudi de transcriptòmica comparativa evidencien la presència de gens diferencialment expressats específics de soca enlloc de gens clau associats amb el patotip AIEC, ja que no s'han trobat gens diferencialment expressats comuns en les soques AIEC.

Encara que el que constitueix una soca AIEC segueix sent un enigma, els resultats d'aquest treball proporcionen informació significativa que contribueix a la nostra comprensió de la genòmica del patotip AIEC. Els resultats de prevalença de gens i substitucions d'aminoàcids confirmen l'elevada variació genètica de les soques AIEC i suggereixen que moltes de les característiques genètiques descrites fins ara estan associades a l'origen filogenètic de les soques i no al fenotip AIEC. Addicionalment, aquest treball reforça

encara més la idea que no hi ha cap gen de virulència particularment relacionat amb el fenotip AIEC. Tot i que és possible que diversos factors de virulència conduïxin al mateix fenotip, la presència d'un marcador específic d'AIEC no es pot descartar completament. En aquest cas, s'han presentat dos possibles marcadors moleculars resultants d'una combinació de característiques genètiques i/o fenotípiques que podrien ajudar en la detecció d'AIEC, almenys dins la nostra col·lecció de soques. Finalment, es presenten per primera vegada dos estudis que analitzen l'expressió gènica de soques AIEC mitjançant un assaig *in vitro* que simula l'adhesió i la invasió bacteriana de cèl·lules epitelials intestinals. De fet, els resultats d'expressió gènica proporcionen noves idees per descriure millor els gens implicats de forma putativa en la virulència del patotip AIEC.

RESUMEN

El patotipo adherente-invasivo de *Escherichia coli* (AIEC) se aísla más frecuentemente del intestino de los pacientes adultos y pediátricos con enfermedad de Crohn que en controles de varios países. Este patotipo se caracteriza por su capacidad para adherirse e invadir las células epiteliales intestinales, así como para replicarse y sobrevivir dentro de los macrófagos sin provocar la muerte de las células hospedadoras. Aunque se han realizado varias aproximaciones para identificar las bases genéticas del fenotipo AIEC, hasta el momento aún no se ha determinado un biomarcador molecular específico para éstas. En la actualidad, su identificación se basa en rasgos fenotípicos determinados mediante ensayos de infección de cultivos celulares, que requieren mucho tiempo y son difíciles de estandarizar. El hallazgo de herramientas moleculares o pruebas rápidas para identificar fácilmente el patotipo AIEC sería de particular interés para los científicos que estudian la epidemiología del patotipo y para los clínicos que buscan detectar qué pacientes son colonizados por AIEC para aplicar tratamientos personalizados. Por lo tanto, en esta tesis, nuestro objetivo principal fue definir mejor las características genéticas del patotipo AIEC y encontrar posibles marcadores genéticos/fenotípicos para su rápida identificación. Para lograr este propósito se han seguido tres enfoques diferentes (capítulo 1, 2 y 3).

Se ha evaluado la prevalencia de 61 genes de virulencia ya descritos, las mutaciones puntuales en los genes asociados a AIEC (*fimH*, *chiA*, *ompA*, *ompC* y *ompF*) y las diferencias en la expresión de los genes *ompA*, *ompC* y *ompF* en una colección de cepas AIEC/no-AIEC aisladas de animales y/o humanos (**capítulos 1.1 y 1.2**). Doce genes estaban enriquecidos en las cepas de animales, mientras que 7 genes fueron más predominantes en las cepas de humanos. La prevalencia de 15 genes fue mayor en las cepas de AIEC que en las no-AIEC, pero solo el gen *pil* se presentó de forma diferencial al analizar las cepas de animales y humanos por separado. Entre las cepas de humanos, tres genes adicionales (*papGII/III*, *iss* y *vat*) fueron más prevalentes en las cepas AIEC que en las no-AIEC. Sin embargo, 25 de los genes estudiados también presentaron una prevalencia diferente con respecto al origen filogenético. Además, no hubo diferencias en las mutaciones patoadaptativas en ninguno de los genes estudiados como marcadores moleculares, pero algunos podrían estar implicados en la virulencia asociada a AIEC (FimH-A119V, OmpA-A200V, OmpC-V220I y D232A, OmpF-E51V y M60K). De manera similar, la expresión de los genes *ompA*, *ompC* y *ompF* durante la infección puede contribuir a la patogenicidad de la AIEC al aumentar la adherencia de las células epiteliales intestinales y la persistencia

intracelular. De hecho, los niveles de expresión de los genes en AIEC aumentaron mientras crecían en el sobrenadante de los cultivos de células infectadas y disminuyeron al adherirse e invadir las células epiteliales intestinales en comparación con las no-AIEC. En estos subcapítulos, a pesar de que no se ha encontrado una característica de AIEC específica y ampliamente distribuida, la combinación de la prevalencia del gen *pic* y la resistencia a la ampicilina presentó un 75% de precisión en la detección de AIEC.

Las diferencias en el contenido de los genes y los polimorfismos de nucleótido único (SNP) en tres pares de cepas se han estudiado mediante genómica comparativa (**capítulo 2.1**). A diferencia de estudios previos, cada par de cepas consistió en una AIEC y una no-AIEC que se consideran clones con respecto a sus patrones de electroforesis en gel de campo pulsado. Tres posiciones (E3-E4_4.3 (2), E3-E4_4.4 y E5-E6_3.16 = 3.22 (2)) presentaron una distribución diferencial de variantes de nucleótidos según el patotipo y cuatro fueron asociadas con un aumento de los índices de adhesión y/o invasión (E3-E4_4.3 (2), E3-E4_4.4 y E5-E6_3.16 = 3.22 (2) y (3)). Sin embargo, su implicación en el fenotipo AIEC no se pudo demostrar ya sea porque no se obtuvieron mutantes isogénicos o porque la eliminación del gen no produjo ningún efecto fenotípico perceptible (**capítulo 2.2**). Curiosamente, nuestros datos revelan tres SNP que se pueden implementar para la identificación AIEC. Aunque este método no clasifica correctamente todas las cepas de *E. coli*, su precisión en cepas aisladas de España (Girona y Mallorca) es muy alta (81%), y actualmente no existen herramientas moleculares comparables.

En la tercera parte de este trabajo (**capítulo 3.1**), se ha realizado un análisis del transcriptoma de dos pares de cepas AIEC/no-AIEC durante la infección de células epiteliales intestinales. Primero, se ha optimizado un protocolo para extraer y purificar el ARN bacteriano intracelular, ajustando la cantidad de muestra o los pasos de lavado para evitar la saturación del kit y garantizar la calidad adecuada del ARN. Finalmente, mediante nuestro análisis comparativo del transcriptoma, se evidenció la presencia de genes diferencialmente expresados específicos de la cepa, en lugar de genes clave asociados con el patotipo AIEC, ya que no se encontró un gen diferencialmente expresado común entre las cepas AIEC.

Aunque lo que constituye una cepa AIEC sigue siendo un enigma, los resultados de este trabajo proporcionan información significativa que contribuye a nuestra comprensión de su genómica. Los resultados de la prevalencia de los genes y de las sustituciones de aminoácidos confirman la alta variabilidad genética de las cepas AIEC y sugieren que

muchas de las características genéticas descritas hasta la fecha se asocian con el origen del filogrupo de las cepas en lugar de con el fenotipo AIEC. Además, este trabajo refuerza aún más la idea de que ningún gen de virulencia en particular está relacionado con el fenotipo AIEC. A pesar de que diversos factores de virulencia podrían conducir al mismo fenotipo, la presencia de un marcador específico de AIEC no se puede descartar por completo. En este documento, se han presentado dos marcadores moleculares putativos resultantes de una combinación de características genéticas y/o fenotípicas, y estos podrían ayudar en la selección de cepas AIEC al menos en nuestra colección de cepas. Finalmente, presentamos por primera vez dos estudios que analizan la expresión génica de AIEC utilizando un ensayo *in vitro* que simula la adhesión bacteriana y la invasión de células epiteliales intestinales. De hecho, los resultados de la expresión génica proporcionan nuevos conocimientos para describir mejor los genes implicados en la virulencia de la AIEC.

• INTRODUCTION •

1. *Escherichia coli* in human gut

The adult human intestinal tract is a complex environment colonised by up to 100 trillion microorganisms from highly diverse species and strains¹. In a healthy context, the intestinal epithelium provides a selective permeable barrier with closely interconnected cells by different cell junctions (tight, adherent junctions and desmosomes) which normally offers an adequate environment for the survival of microbes but at the same time operates to confine the microbial population². Moreover, to avoid uncontrolled inflammation in the gut, the immune system acquires tolerance to commensal bacteria, but at the same time rapidly reacts to fight against pathogens². This host-bacterial consensus has been established since commensal microbes can also promote human health by different manners. They are responsible for protective (e.g. pathogen displacement), metabolic (e.g. synthesis of essential vitamins) and structural (e.g. promotion of epithelial cell differentiation) functions (for review³).

One of the species found in the gut is *Escherichia coli*, a facultative anaerobic gram negative from the Enterobacteriaceae family that normally interacts with the host in a mutualistic manner. *E.coli* colonises the gastrointestinal tract of human infants shortly after birth and is a lifelong coloniser of adults^{4,5}. Normally, it persists as a harmless commensal in the mucous layer of the cecum and colon⁶. Gut commensal *E. coli* strains are highly diverse in terms of phylogenetic origin with some lineages acquiring various combinations of genetic information that enable them to exploit different niches⁶⁻⁸. They frequently express adhesins (P fimbriae and type 1 fimbriae), capsular antigens (K1 and K5), the toxin α - hemolysin, as well as the siderophore system aerobactin, which are believed to lead to persistence in the gut⁶.

1.1. Pathogenic groups of *E. coli*

E. coli is a versatile bacterial species which comprises harmless commensal as well as different pathogenic strains. The latter may have acquired different sets of virulence genes (VGs) via horizontal transfer of DNA on plasmids, transposons, bacteriophages and

pathogenicity islands allowing them to adapt to a pathogenic lifestyle and cause a broad spectrum of diseases. Pathogenic *E. coli* strains are grouped in pathotypes according to its clinical spectrum and virulence factors^{5,9-11}. The extraintestinal pathogenic *E. coli* (ExPEC) group comprises those strains causing infections in the urinary tract, sepsis or meningitis, such as uropathogenic *E. coli* (UPEC) and meningitis-associated *E. coli* (MNEC) (Table 1). Interestingly, ExPEC are believed to originate in the gut where they do not appear to cause any form of diarrhoeal disease¹². In fact, they belong to the normal flora of many healthy individuals and typical ExPEC strains resemble commensal isolates with regard to the prevalence of virulence- or fitness-associated genes and phylogroup allocation. Therefore, an unambiguous distinction of ExPEC and commensals is not easy to accomplish^{6,7}. Nonetheless, some commensals might be distinguished from extraintestinal pathogenic variants because of their plasmid content¹³.

Distinct from commensal and ExPEC strains, the other group, diarrhoeagenic *E. coli* (DEC) encloses those that cause intestinal infections. These strains carry specific surface adhesins which enhance their ability to colonise the gastrointestinal tract but they rarely translocate the intestinal epithelium^{5,14}. Seven well-defined pathotypes are found in this group (Table 1): enterohemorrhagic *E. coli* (EAEC), enteropathogenic *E. coli* (EPEC), enterotoxigenic *E. coli* (ETEC), enteroaggregative *E. coli* (EHEC), enteroinvasive *E. coli* (EIEC), diffusely adherent *E. coli* (DAEC), and necrotoxic *E. coli* (NTEC). Lately, a new pathotype called adherent-invasive *E. coli* (AIEC) has been proposed to be associated with Crohn's disease (CD).

1.2. Adherent invasive *E. coli* (AIEC) definition

The most studied strain from the AIEC pathotype (LF82) was isolated for the first time two decades ago from a French CD patient with ileal affectation¹⁵. Then it was proposed as a new pathotype as its pathogenic traits were different from the rest of the previously described DEC pathotypes¹⁶ (Table 1). Instead they presented similar virulence traits to ExPEC strains¹⁷⁻²⁰. The current AIEC definition is based on the accomplishment of several factors²¹. Thus, to be considered AIEC, an *E. coli* strain has to: (1) adhere to intestinal epithelial cells (IECs) (Caco-2 and/or undifferentiated I-407 cell line) (≥ 1 bacteria/cell)¹⁵, (2) invade IECs (I-407 and HEP-2 cell lines) (Figure 1) through the involvement of host cell actin polymerisation and microtubule recruitment ($\geq 0.1\%$ of the original inoculum)¹⁶, (3) survive and replicate within macrophages (J774-A1 cell line) (Figure 1) without inducing

cell death ($\geq 100\%$ of intracellular bacteria post 24h of infection)²² and (4) do not present any known invasive determinants²¹.

Table 1. General characteristics and main VGs of human ExPEC, DEC and AIEC pathotypes. Adapted from^{5,9-11}.

Pathotype	Characteristics	Main VGs related with adhesion and toxin secretion
UPEC	The most common cause of urinary tract infections worldwide	P (pap) fimbriae, type 1 fimbriae (FimH), α -haemolysin and HlyA toxin
MNEC	Responsible for gram-negative neonatal meningitis and sepsis	S Fimbriae, OmpA, IbeA and AslA adhesins
EAEC	Autoaggregative adhesion mechanism, causes persistent diarrhoea in both developing and industrialised countries	Fimbriae GCCPQ, enteroaggregative heat-stable (EAST1) and enterotoxins (Pet, Pic and ShET1)
EPEC	Causes fatal infant diarrhoea	Intimin adhesin and fimbriae Bfp
ETEC	Non-invasive but secrete endotoxins that promote infant diarrhoea in developing countries and traveller's diarrhoea	Fimbrial adhesion (CF, K88, K99, 987P and F17), LT enterotoxin and ST enterotoxin
EHEC	Causes bloody diarrhoea, non-bloody diarrhoea and haemolytic uremic syndrome	Intimin adhesin (eae), Shiga-like toxin (Stx) and enterohaemolysin (E-Hly)
EIEC	Intracellular pathogen that causes inflammatory colitis and sometimes dysentery	Similar to <i>Shigella</i> spp (ShET1 and ShET2). Most VGs encoded in a plasmid
DAEC	Causes urinary tract infections and it is implicated in children diarrhoea. It presents a diffuse pattern of adherence	Afimbrial adhesions from the Afa/Dr family
NTEC	Causes neonatal enteritis and may contribute to urinary tract infections	Fimbriae P and cytotoxic necrotizing factor (CNF1 and CNF2)
AIEC	Associated with CD able to colonise and invade the intestinal barrier as well as to survive and replicate inside macrophages	For instance, FimH, LpfA, OmpA and ChiA. For more information see introduction section 3.

Even though the above mentioned AIEC description is the only published, several studies have identified AIEC strains based on: the accomplishment of only one feature, assessment of features on different cell lines or identification with incomplete assays. While Desilets et al.²³ classified *E.coli* strains as AIEC by only assessing their ability to replicate within J774 macrophages, Dogan et al.²⁴ and Negroni et al.²⁵ used the Caco-2 cell line or RAW264.7 cell lines to examine strain invasion or intramacrophage replication respectively. Finally, there

are few studies that include the evaluation of cytoskeleton involvement once characterising AIEC strains^{16,17,26}.

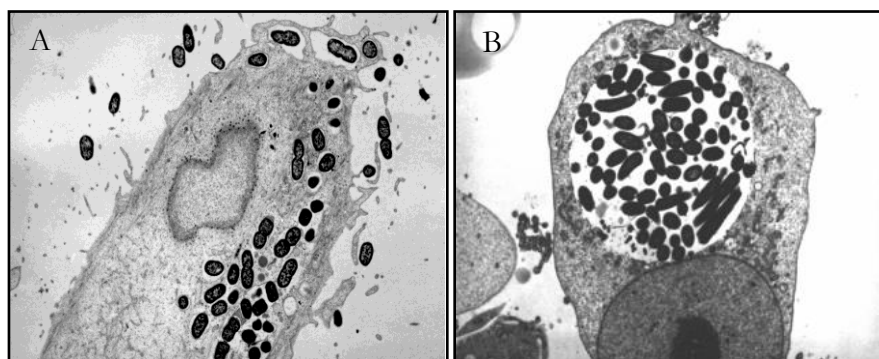


Figure 1. Transmission electron micrographs of AIEC LF82 strain invading intestinal epithelial cells and replicating within macrophages. A: Bacteria that has invaded Hep-2 cell monolayer after a 5h infection period¹⁶. B: Intracellular bacteria in a large vacuole formed inside J774 macrophage after 24h of gentamicin treatment²².

2. AIEC and Crohn's disease

2.1. Crohn's disease

CD is one of the most common subtypes of the chronic inflammatory bowel diseases (IBD) together with ulcerative colitis (UC). During the 20th century, CD was considered a disease of western countries (USA, Europe and Oceania). However, in the 21st century, CD incidence is increasing worldwide, in particular in those countries adopting westernised lifestyle (i.e. Asia, South America and Africa)²⁷⁻³⁰ (Figure 2). Taking into account that CD incidence/prevalence is subject to variation between and within geographic regions, estimated incidences in Europe ranged from 0.0 (Greenland) to 15.4 (Italy) per 100,000 person-years whilst the prevalence values varied from 1.5 (Romania) to 322.0 (Germany) per 100,000 person (period of study since 1990 until 2016)²⁷. Despite CD presents a low mortality (1.39 standardised mortality rate; 95% confidence interval 1.30-1.49)³¹, it can result in significant long-term morbidity and important challenges to health-care systems²⁹.

CD may affect the entire gastrointestinal tract and it is normally presented as a patchy inflammation characterised by intestinal barrier disruption and increased permeability due to altered transcellular and paracellular barrier function (partly mediated by TNF). Moreover, it is characterised by the presence of abscesses, fissures and granulomas^{32,33}. Elevated numbers of proinflammatory cytokines are reported in patients suffering from CD, being Th1 cytokines the most predominant in this disease³⁴. No cure for CD exists yet, as a result treatment consists mainly in the suppression of the immune system. Therapies

depend on the disease location, activity and severity. It involves mainly anti-inflammatory drugs (i.e. Infliximab and Adalimumab), corticosteroids (i.e. prednisone, methyl-prednisone and budesonide), immunosuppressant drugs (i.e. azathioprine and mercaptopurine), antibiotics (i.e. metronidazole and ciprofloxacin), biological therapy (TNF α inhibitors), fecal transplantation or nutritional interventions and surgery^{35,36}.

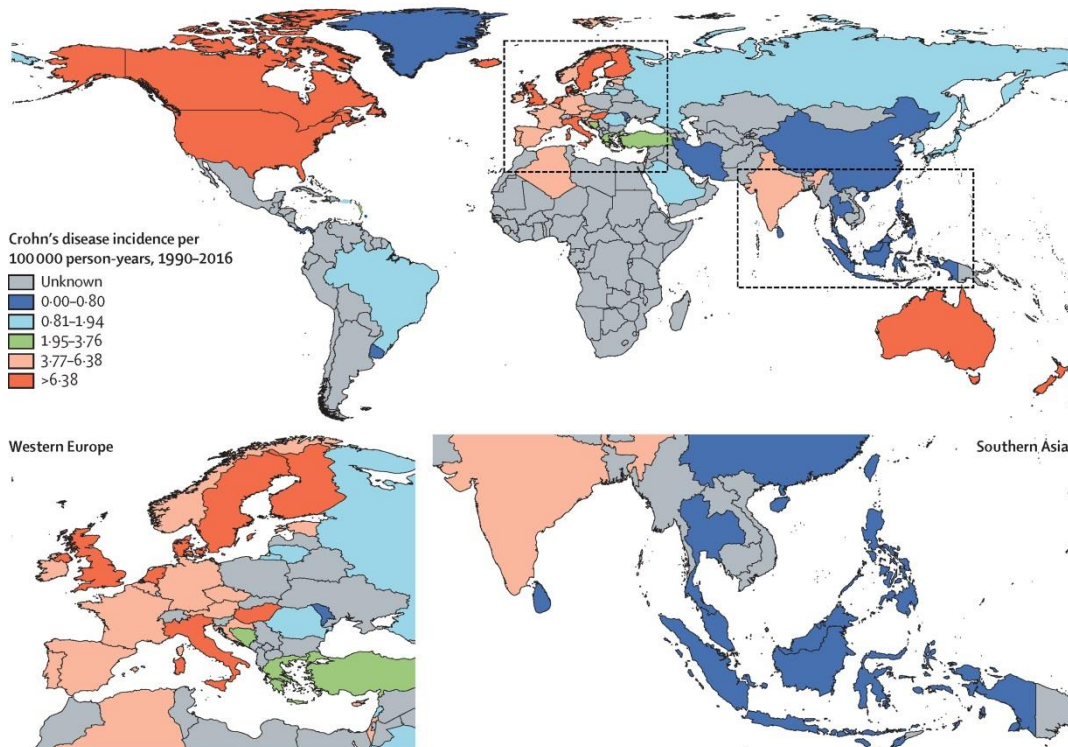


Figure 2. Worldwide incidence of CD from 1990 to 2016. Incidence values were represented in low (dark and light blue) to intermediate (green) to high (pink and orange) occurrence of disease. Adapted from Ng et al.²⁷.

CD aetiology is not fully understood but multiple factors such as host immunity, host genetics and environmental factors (i.e smoking³⁷ or high fat diet³⁸⁻⁴¹) that modify the gut microbiota are thought to play a role^{30,42}. A state of microbial imbalance, named as dysbiosis, has been reported in CD patients. This condition frequently involves a reduction of beneficial bacteria (Firmicutes and Bacteroidetes) and an increase on pro-inflammatory species (Proteobacteria, particularly *E. coli*)^{17,21,43-45}, leading to a ‘pathogenic’ community.

Apart from that, some microbial species have been reported to potentially contribute to epithelial barrier disruption and inflammation observed in CD⁴⁶. Scarce investigation exists on the contribution of *Helicobacter pylori*, *Clostridium difficile* and *Campylobacter species*. Whilst *H. pylori* has a protective effect as it is capable to reduce the production of cytokines and it has been suspected that after its eradication IBD appears⁴⁶, *Clostridium difficile* or *Campylobacter species* presence is considered a risk factor of disease exacerbations^{47,48}. Even though,

Mycobacterium avium subspecies *paratuberculosis* (MAP) and AIEC are the ones attracting the most attention.

In terms of pathophysiology, MAP causes a CD-like disease that produces chronic diarrheal disease in ruminants, primates and rabbits (Johne's disease)⁴⁹. Initially it was thought to be zoonotic but there exist several pathways for human to be infected (mainly through food inquire)⁵⁰. In particular, several studies have analysed its presence in CD patients. Distinguishing results were obtained due to the treatment received, whereas in early works MAP has been detected in CD patients but not in controls^{51,52}, in 2008 higher values of MAP were reported in non-IBD patients than in CD-patients receiving antibiotics or anti-inflammatories⁵³. Later on, in a study conducted with naïve pediatric CD patients, Kirkwood et al.⁵⁴ pointed out that MAP occurrence may be reduced under IBD treatment. Beyond that, MAP have been related with CD not only because it presents an indirect pathogenic effect by impairing the ability of monocyte-derived macrophages to kill phagocytosed *E. coli*⁴² but also because genetic defects described in CD patients (NOD2 and ATG16L1) show immune system ineffectiveness on MAP recognition^{55,56}. Moreover, the need of higher intestinal permeability to promote MAP access into inner layers it has been suggested along with its capacity to invade already inflamed tissues and promote granuloma formation⁵⁷. However, therapies directed at MAP do not produce a cure for the disease^{58,59}. Therefore, despite MAP has been hypothesised to be clinically relevant in CD, there is still great controversy due to inconsistent epidemiological results and lack of studies demonstrating disease amelioration after MAP clearance, to fully support this hypothesis^{52,60}.

Finally, the association of *E. coli* presence in IBD has been well established^{21,43,44,61-73}. In fact, increased antibodies titres directed against *E. coli* OmpC have been observed in CD patients (37-55%) in contrast with controls (less than 5%). Moreover, since several reports have shown that in subjects with CD the abundance of *E. coli* with an adherent phenotype is markedly increased^{17,24,43,63,74,75} (see introduction section 2.4), colonisation of the AIEC pathotype has been proposed to contribute to CD pathogenesis by triggering intestinal inflammation (see introduction section 2.2).

2.2. AIEC and Crohn's disease pathogenesis

To prevent unwanted guests from entering and interacting with immune cells in the lamina propria, the digestive tract is equipped with diverse specific and unspecific protective

mechanisms that collectively build a complex and effective mucosal barrier². The intestinal mucus layer is the first line of defence and it is in charge of protecting the epithelia from bacteria but also to keep the mucosal surface hydrated^{76,77}. The mucosa is predominantly composed of the mucin MUC2 which is highly secreted by goblet cells found both in the small and large intestine but also presents antimicrobial peptides (AMP). These are small cationic peptides that exhibit broad-spectrum antibiotic activity and its production is mainly taking part in the Paneth cells of the small intestine⁷⁸. The goblet cells together with Paneth cells, the microfold cells (M-cells) and primarily absorptive enterocytes constitute the intestinal epithelial layer. The latter border the majority of the intestinal lumen and present microvilli to increase the cell surface and facilitate nutrient absorption, at the same time as it limits the bacterial intracellular access to the base of the microvilli⁷⁹. Across the different cell types within the gastrointestinal tract, there are the initiators of immune responses, known as pattern-recognition receptors (PRRs). The two main families are the intracellular nucleotide-binding oligomerization domain (NOD)-like receptors (NLRs) and the membrane bound toll-like receptors (TLRs). These are dependent on ligand binding which once detected activates a signal cascade that leads to immune system responses, i.e. secretion of inflammatory cytokines and AMP for bacterial clearance. The specialised M-cells located in the Peyer patches of the intestine are in charge of delivering luminal antigens from the gastrointestinal tract to the immune cells as they can efficiently mediate transcytosis from the apical surface (lumen) to the basolateral surface (lamina propria). Therefore they contribute to the development of the host immunity. Finally, if bacteria subvert the mucus and the epithelia, the autophagy process and the action of the phagocytes found in the lamina propria are the ones responsible for preventing bacterial replication and persistence.

In patients suffering from CD the intestinal barrier function is compromised at different levels leading to a dysregulated mucosal immune response (for review^{2,80}). The protective mechanisms may be affected with alterations of PRRs, impaired AMP production, disrupted mucus layer, autophagy alterations or high permeability of the epithelial barrier. As a result, pathogenic bacteria (e.g. AIEC) come closely to epithelium, increased immune reactions occur and in turn cause chronic inflammation culminating in disease (Figure 3).

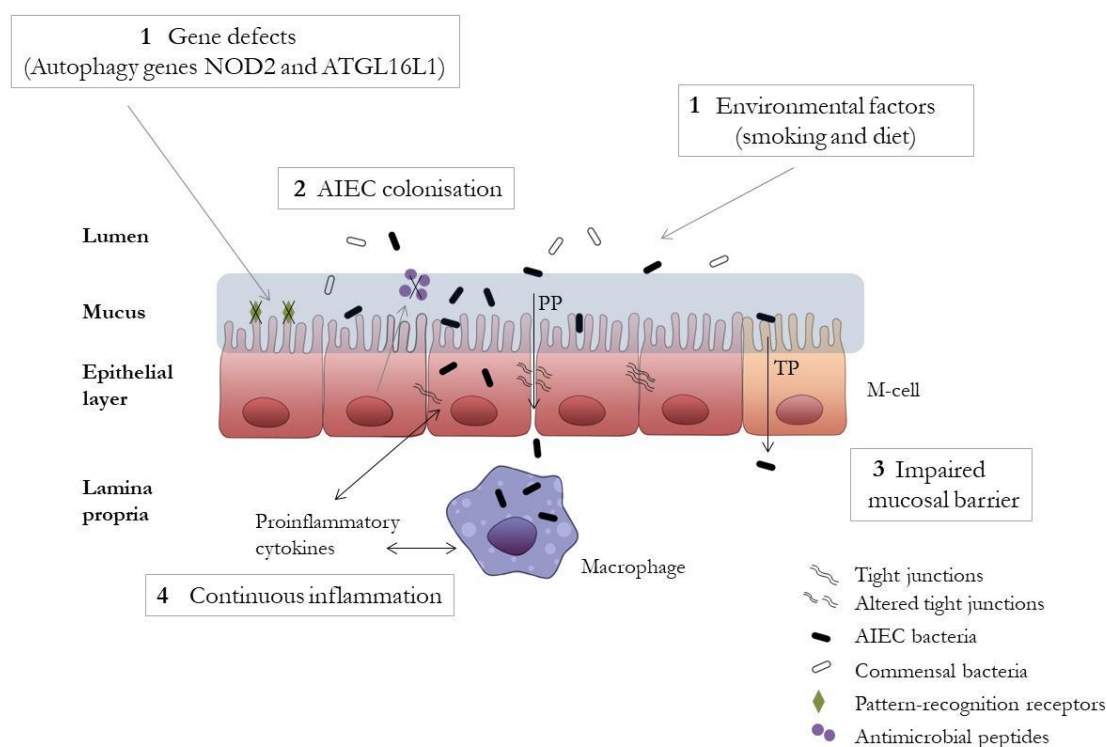


Figure 3. Schematic representation of the epithelial barrier system in CD. The intestinal barrier is composed of several layers providing protection against microbial invasion. But it is impaired due to (1) genetic defects and environmental factors, which produce alterations of pattern-recognition receptors and impaired antimicrobial peptides production leading to disordered innate immunity and dysbiosis. In particular conditions, AIEC may colonise (2) and due to impaired mucosal barrier (3) invade by transcellular (TP) or paracellular pathway (PP) reaching the lamina propria. Finally inadequate clearance of intracellular bacteria ends up with continuous inflammation (4) by secretion of proinflammatory cytokines and granuloma formation.

AIEC strains present several virulence mechanisms that can evade the protective mechanisms of the epithelium barrier by potentially contributing to its disruption, what can end up with the development of similar histopathological features as observed in CD patients. First, AIEC is able to adhere to, invade the intestinal mucosa and translocate across the human intestinal barrier, thus promoting mucosa colonization and tissue damage. In fact, studies on AIEC colonisation have demonstrated that this bacteria can cause microscopic erosion similar to those seen in the Peyer's patches of CD patients at early stage^{81,82}. Second, increased levels of particular receptors (i.e. CEACAM6, Gp96 and CHI3L1) have been observed in CD patients, which can also present some genetic deficiencies (i.e. NOD2 or ATG16L1) that compromise the function of macrophages and the autophagy process. In both cases AIEC colonisation is facilitated since it can employ the overexpressed receptors as a binding site for adhesion or invasion^{81,83–85} and defects in the immune system have been reported to contribute to unrestrained AIEC intracellular replication and persistent infection inside IECs^{86–88} (details about these processes can be found in introduction section 3). Third, AIEC is able to survive within macrophages while

preventing cell apoptosis^{89,90}, what implies continuous secretion of cytokines and chronic cell activation²². In this case, there is a constant inflammatory response and it can also promote the formation of granulomas, being both cases common histopathological features of CD⁹¹. Forth, AIEC is also capable to modulate host autophagy via the inhibition of ATG5 and ATG16L1 expression, which had already been reported to be diminished in ileal samples of CD patients⁹². In addition, dysregulation of apical junctional complex (AJC) has been observed in IBD patients. This might be explained in part by AIEC infection as it may also stimulate higher epithelial permeability and decrease transepithelial electrical resistance (TEER) by inducing a re-distribution of tight junctional proteins, as has been demonstrated in LF82 in an *in vitro* model⁹³. This leads to a disruption of barrier function which, together with inflammation, can prompt the loss of microbiome diversity and promote AIEC expansion in susceptible mice models (C57BL6, CEABAC10, *eif2ak4*^{-/-} and T5KO mice)⁹⁴⁻⁹⁶. Altogether these observations reinforce the link between AIEC properties and CD clinical manifestations.

Although substantial research has been conducted to elucidate the molecular mechanisms of AIEC virulence and its relation with the disease pathogenesis, it is still unclear how it contributes to the disease (i.e. whether AIEC bacteria trigger intestinal inflammation, thus leading to the disease, in a non-compromised host, or whether they colonise the gut mucosa as a consequence of pre-existing inflammatory context). Nonetheless, the assumption to consider AIEC as a pathobiont has gained plausibility. Referring as pathobiont bacteria that can be present in the normal microbiota and that can turn to pathogenic under specific conditions. Indeed, *in vivo* studies found that AIEC does not colonise mice spontaneously, it needs a particular context⁹⁵⁻⁹⁸. For instance, after an antibiotic treatment, AIEC is able to reside and persist in the ileum of a wild-type (WT) and NOD2KO mice⁹⁷. However, it is only able to worsen the disease severity after an inflammatory environment caused by the presence of another microorganism (in this case, *Salmonella enterica* serovar Typhimurium strain or *Citrobacter rodentium*)⁹⁸, thus suggesting that its presence alone could not result in an inflammatory disease. Moreover, without using pre-treatment with antibiotics, it has been observed that AIEC infection promotes colitis in susceptible genetic mice (TLR5 deficient model) or in mice with autophagy defects as well as, changes in microbiota composition that might eventuate in chronic inflammation^{95,96}. Bearing in mind that AIEC are present in healthy subjects (Table 2) without causing inflammation together with the fact that the AIEC capacity to induce damage in a particular host is dependent on both the microbial composition of the gastrointestinal tract

and exposure of the host to other risk factors, the most trusted hypothesis so far is that AIEC can associate with CD in a context of susceptibility and thus it could be considered a pathobiont rather than a truly pathogen^{99–103}.

2.3. Crohn's disease therapies and their effects on AIEC

Treatments commonly used in CD patients may modulate AIEC virulence but its exact impact on it has been scarcely studied. Infliximab has been demonstrated to restore paracellular permeability and restrict AIEC (HM427 strain) translocation¹⁰⁴. Similarly, AIEC LF82 strain overnight pre-conditioned with the amino-6-mercaptopurine riboside impairs its ability to adhere to and invade IECs (HT29 cell line) as well as to replicate inside macrophages (human monocyte-derived macrophages)¹⁰⁵. Therein, the fact that it is the strain that has been in contact with the drug rather than the cells indicates that the drug exclusively targets bacterial cell processes such as inhibition of fimbrial genes, cellular production or bacterial motility. In terms of AIEC phagocytosis, treatment with glucocorticoids may impair AIEC (CD2-a strain) replication at different levels, (I) as it downregulates genes involved in the recruitment of THP-1 macrophages in the affected zone (i.e. no formation of granulomas) and (II) immune system activation (i.e. less cytokines are produced and inflammation is reduced). Moreover, FimH receptors may be also downregulated in CaCo-2 cells thus reducing AIEC adhesion¹⁰⁶. Nonetheless, contrary to the effect reported in THP-1 macrophages, the same study tested the glucocorticoids treatment in AIEC-infected bone marrow-derived macrophages from NOD2-knockout mice and no AIEC phagocytosis impairment was achieved. They suggested that this unexpected result may be due to the fact that Nod2^{-/-} macrophages showed reduced phagocytosis activity in comparison to WT macrophages. Hence, untangling CD treatments effect on AIEC virulence in different environments may provide further knowledge on how to treat CD patients in order to prevent disease exacerbation or complication.

Another option would be to develop therapies directed against AIEC. To date, therapies to avoid direct AIEC-host interactions have already been assessed. For instance, (I) probiotics administration^{107–113}, (II) FimH antagonists^{114–120} or, (III) dietary modifications^{39,40,121–125}. In contrast to the therapies mentioned above, these treatments are designed to treat patients colonised by AIEC.

Probiotics may specifically prevent the colonization of AIEC and subsequently modify its pathogenic behaviour. In 2003, it was described that *E. coli* Nissle 1917 highly adhered to IECs (I-407 cell line) without inducing cell cytotoxicity. As a consequence it enabled the formation of non-pathogenic biofilm which prevented pathogenic microorganisms from accessing the cell surface. Indeed, its presence provided an inhibitory effect for AIEC colonisation¹⁰⁷. This observation was in concordance with Huebner et al.¹⁰⁸, who reported reduced LF82 adhesion and production of cytokines in IECs but this case in Caco-2 cell line. However, in cultured ileum tissue from CD and healthy subjects, no reduction in adhered-LF82 quantity after co-incubation with *E. coli* Nissle 1917 was observed¹⁰⁹. Inhibitory effects on AIEC adherence and invasiveness was also exhibited in cell culture assays with the addition of *Lactobacillus casei* DN-114 001¹¹⁰. Later on, other probiotics were recommended because either they lessened LF82 adherence as well as proinflammatory response and colitis in CEABAC10 mice (*Saccharomyces cerevisiae* CNCM-I-3856¹¹¹) or they reduced AIEC survival in an environment mimicking the human intestinal tract (two *Lactobacillus* species¹¹²). Since dose-dependent and time-dependent inhibitory effects have been exposed¹¹⁰ and disappointing results have been reported in clinical trials using probiotics¹¹³, the therapeutic value of the probiotic treatment in CD needs more attention.

FimH antagonists represent another way to interfere with AIEC adhesion. They are designed to saturate the carbohydrate recognition domain of FimH. As a consequence, AIEC bacteria are unable to adhere to IECs. Nowadays, three classes of antagonists have been successfully developed¹¹⁴. The first developed were the monovalent heptyl mannose (HM) derivatives¹¹⁵. These prevented and disrupted *E. coli* adhesion to IECs but *in vivo* failed to reduce AIEC levels in the gut¹¹⁶. Once this was modified (non-hydrolysable methylene group in place of the anomeric oxygen atom), it was administered orally with equal dose as before (10 mg/kg) and diminished AIEC levels both in faeces and in the mucosa were shown^{116,117}. The second class consists on thiazolylaminomannosides or TazMans. They are less soluble and also effective against *E. coli* strains but presented low stability in acid media¹¹⁵. Second generation TazMans were synthesised by substituting the anomeric nitrogen atom¹¹⁸. Nonetheless they conserved the high potency for FimH and were resistant to acid-promoted anomerization and glycosidase hydrolysis but they were less effective than the monovalent one and remains *in vivo* assessment. Finally, the third class was formed by HM-based glycopolymers. *In vitro*, these prevented AIEC attachment to IECs more efficiently than the monovalent HM¹¹⁹. These observations are thought to occur due to cross-linking interactions with the glycopolymers rather than due to greatest

intrinsic affinity for FimH targets. Despite the necessity of improving these polymers is evident, so far they constitute a step further toward an *E. coli* antiadhesive treatment. Interestingly, with the same purpose as FimH antagonists, it has been studied the impact of the bovine lactoferrin protein¹²⁰. This contains highly mannosylated glycans and as FimH binds to the mannose residues of cellular receptors⁸⁴, lactoferrin binds to FimH blocking its interaction with epithelial cells. As a result, it inhibits AIEC invasion and reduces intestinal inflammation as assessed *in vitro* in Caco-2 cells¹²⁰. The need of further studies analysing the AIEC inhibitory effect of particular proteins is indubitable. All these anti-adhesive molecules could provide an advance on the treatment of AIEC-carrier CD patients, yet they might also have an undesired impact on other *E. coli* found in the gut. As a consequence, additional analyses are needed to assess their pharmacokinetic behaviour in the gut environment and, most importantly, their impact on the commensal bacteria of the gut microbiota.

Regarding diet supplementation, since high fat diet enhanced AIEC colonisation^{39,40}, the modification of dietary factors may help to low, or even alter the natural course of CD. Deficiency of vitamin D has been described in CD patients and this insufficiency contributes to a microenvironment that is conducive to promote the virulence of AIEC¹²¹. The lack of vitamin D, which is a modulator of the immune system, results in an altered epithelial barrier function due to an upregulation of proinflammatory cytokines and a reduced expression of antimicrobial peptides. Indeed, it is suggested that this pathotype takes advantage of these dysregulated immune system in several aspects (see section 1.3 and 2). Thus, it has been suggested that supplementation of CD patients' diet with vitamin D, which also stimulates toll-like receptors and favours macrophage function, could impair AIEC virulence¹²². Furthermore, Denizot et al.¹²³ stated that supplementation with methyl donor molecules could be a therapeutic strategy to decrease CEACAM6 (carcinoembryonic antigen-related cell adhesion molecule 6) expression in patients with CD via a mechanism involving hypermethylation of the CEACAM6 promoter. Methylation is a process in which methyl groups are added to the DNA molecule and modulate the expression of genes encoded there. In the same way, methyl-methane sulfonate treatment decreased by almost 50% the adhesive properties of LF82 and there was a strong reduction of N-acetylneuraminidase lyases involved in sialic acid metabolism once the bacteria was treated with the methyl-methane sulfonate¹²⁴. Downregulation of acetylneuraminidase lyases by using inhibitors resulted in impaired ability to form biofilms underlining the importance of sialic acid metabolism to prevent cell-cell interactions. In addition, advantages of krill oil (rich in

omega3 fatty acids) in improving intestinal inflammation, cell survival and impairing the capacity of AIEC to adhere to and to invade intestinal cells (Caco-2, HT29 and RAW264.7) have also been revealed¹²⁵. Considering the lack of adverse effects by diet modification, it is believed that it could open new prospects for its possible use as human intestinal inflammation treatment.

Additionally, the development of treatments to target intramacrophage AIEC isolates are necessary as suggested by Tawfik et al.¹²⁶. Different approaches for killing AIEC bacteria intracellularly may also be plausible: (I) subadministration of antibiotic (AB) regimens and, (II) therapy with bacteriophages.

Although scarce studies have assessed the impact of the treatment with AB that are able to penetrate macrophages on CD patients, some AB (azithromycin, ciprofloxacin, rifampicin, rifaximin, sulfamethoxazole, tetracycline and trimethoprim) have been reported to be effective against *E. coli* within macrophages¹²⁷ or to reduce virulence gene expression (i.e. FimH), motility, and adhesion of CD-associated AIEC independently of its antimicrobial activity¹²⁸. A meta-analysis study conducted by Rahimi et al.¹²⁹ indicated that those active-CD patients receiving metronidazole, ciprofloxacin and cotrimoxazole were more likely (2.3-fold) to improve clinically than those from the placebo group. In contrast, one study conducted in 2012 stated that CD patients undergoing antibacterial therapy (metronidazole, ciprofloxacin or clarithromycin alone or in combination) are 1.35-fold more probable to experience clinical remission than those patients without antibacterial therapy¹³⁰. Therefore, it is important to identify the exact target for the antibiotics and to study AIEC resistance to AB. Resistance to some macrophage-penetrating antimicrobials (ciprofloxacin, clarithromycin, tetracycline, trimethoprim/sulfamethoxazole, and rifampin) was reported in 61% of IBD-isolated AIEC strains²⁴ and later on similar values were reported in *E. coli* isolated from granulomatous colitis in boxer dogs (42%)¹³¹. Noticeably, Brown et al.¹³² demonstrated that membrane pore-forming colicins efficiently killed LF82 at any step of its pathogenesis (forming biofilm, adhering and invading IECs and even during intramacrophage condition). Altogether suggests that particular AB may be of interest to treat this disorder but stratification of the CD population carrying AIEC would be crucial. Meanwhile, whether they are of actual benefit in CD patients remains to be determined.

Considering that there is increasing incidence of multidrug resistance bacteria, AB treatment is becoming less effective. Thus, the potential effect in CD of bacteriophage therapy, which consists on viruses infecting specific bacteria, should be under

consideration. Recently it has been demonstrated that the treatment with bacteriophages reduces the risk of dysbiosis, as well as, it decreases AIEC colonisation *in vivo*. At the same time as colitis symptoms were diminished, the number of AIEC in faeces of LF82-colonised CEABAC10 transgenic and conventional mice were seen reduced after the administration of a three bacteriophages cocktail (from the *Myoviridae* family of viruses)¹³³. Nonetheless, no clinical study on CD patients has been addressed yet.

New approaches are under constant investigation and emerging. Such is the case of the particle-based photodynamic therapy (PPDT), which involves light and a conjunction of photosensitizing chemical substances and molecular oxygen to provoke cell death. There exists only one study¹³⁴ taking it into consideration against AIEC bacteria. Therein, gold nanorods coated with indocyanine green-loaded silica shell were found to completely inactivate AIEC grown in LB after an illumination at 810 nm for one hour. Two main inconveniences lay behind this approach: (I) bacteria distant from the particles will not be affected by the PPDT and (II) there is still no *in vitro* and *in vivo* experiments assessing its effects.

2.4. AIEC prevalence in Crohn's disease and other gastrointestinal disorders

Since many studies have demonstrated the overgrowth of intracellular *E. coli* in CD patients in comparison with controls^{21,43,44,61-73} and according to disease activity^{17,44,135}, interest rose to determine whether these corresponded to the AIEC pathotype or not. The first study addressing the prevalence of AIEC was conducted in patients with CD or UC as well as controls²¹. Samples were taken from ileum and colon and it was found that 27.0% of CD patients that undergone surgical resection of the terminal ileum presented AIEC strains. Instead, only 6.2% of CD-colonic samples, 3.7% of UC patients and 8.2% of controls harboured invasive strains. Later on, many independent groups reported similar results, AIEC were mainly isolated from CD patients ($43.72 \pm 20.19\%$) than from healthy controls ($13.02 \pm 14.90\%$), both in adults^{17,24,43,63,74,75} and children^{25,136} from different countries (Table 2). Although Darfeuille-Michaud et al.²¹ described similar AIEC prevalence between colonic CD samples and controls, and suggested an association between AIEC and ileal CD, others showed also higher AIEC prevalence when comparing colonic CD with controls¹⁷. Regardless of AIEC prevalence in disease location, higher prevalence in the ileum than in the colon of CD patients has been described^{21,43}. Moreover, AIEC are also more abundant in CD patients than in controls despite little evidence exists due to the

laborious protocols required. In adults, AIEC represented 4.84% of total *E. coli* population while in controls 3.68%⁴³. More drastic values were found in paediatric patients¹³⁶, where 10.06% of CD-*E. coli* population were AIEC while for controls AIEC represented only a 1.62%.

Table 2. Studies where AIEC prevalence has been analysed in ileal and colonic samples from subjects with IBD and controls. CD patients have been classified according to anatomical location of the inflammation.

Study	Sample type	% of patients with AIEC (N total patients)				
		I-CD	IC-CD	C-CD	UC	Controls
Darfeuille-Michaud et al., 2004 ²¹	Ileum	27.0% (63)	-	-	-	6.2% (16)
	Colon	-	-	3.7% (27)	0% (8)	2.0% (102)
Baumgart et al., 2007 ¹⁷	Ileum	38.5% (13)	-	37.5% (8)	-	14.3% (7)
Sasaki et al., 2007 ⁶³	Undetailed	53.4% (15) ^b	-	-	13.4% (12)	8.4% (12)
Martinez-Medina et al., 2009 ^{43,137}	Ileum	66.7% (9)	-	50.0% (2)	-	17.6% (17)
	Colon	58.3% (12)	-	25.0% (4)	-	15.8% (19)
Raso et al., 2011 ⁷⁵	Colon	-	62.5% (8)	-	0% (6)	0% (4)
Negroni et al., 2012 ^{25 a}	Colon and ileum	5.9% (17) ^b	-	-	10.0% (10)	0% (23)
Dogan et al., 2013 ²⁴	Ileum	25.0% (32)	-	-	-	17.9% (32)
Conte et al., 2014 ^{136 a}	Ileum	-	75.0% (4)	-	-	50.0% (4)
Céspedes et al., 2017 ⁷⁴	Colon and ileum	57.1% (7)	22.2% (9)	62.5%(8)	-	0% (18)

^aStudy conducted in paediatric patients. ^bUndetailed anatomical localisation of the inflammation. I-CD: Ileal Crohn's disease. IC-CD: Ileocolonic Crohn's disease. C-CD: Colonic Crohn's disease. UC: Ulcerative colitis.

Limitations on determining AIEC prevalence exist due to the methodological approaches adopted. On one hand, frequency values are highly variable as most of the studies analysed the phenotype of less than 50 *E. coli* colonies per patient^{17,21,24,25,63,74,75} while others did so in a collection of 70-150 colonies^{43,136}. Additionally, different samples types have been examined (biopsies from ileum^{17,21,24,25,43,74,136} or colon^{21,25,43,63,75}) and those were taken from patients in variable state of activity, severity and disease phenotype (I-CD, IC-CD or C-CD) depending on each study. On the other hand, discrepancies in AIEC definition between studies have also been found. Diverse cell lines have been used to assess AIEC phenotype. For instance, adhesion and invasion have been assessed on I-407^{21,43}, Hep-2^{21,25,136} or Caco-2^{17,21,24,25,63,74,75,136} cell lines and intramacrophage replication on J774^{17,21,24,43,63,136}, RAW264.7^{25,74} or U937⁷⁵. Moreover, divergent thresholds have been applied to determine bacterial adherence and invasiveness. While most of the studies categorised as adherent a strain with an adhesion index of 1 bacteria/cell, Conte et al.¹³⁶ established that adherent bacteria had to adhere in more than 40% of the cells. In terms of invasion, the AIEC description determined an invasion index higher than 0.1% but Sasaki et al.⁶³ considered invasive those strains with more than 1%. Finally, while studies on AIEC

frequency have been placed in Europe or USA, information on AIEC prevalence in countries where CD incidence is increasing rests unexplored.

Controversial results have been found for the prevalence of AIEC in other intestinal disorders (Table 2). Despite *E. coli* abundance has been related with disease status in UC patients⁶⁶ and adherent *E. coli* has been detected in the colon of UC patients^{64,67,68,138}, AIEC population needs more investigation in larger UC cohorts. Certainly, some independent studies with more than 10 UC patients have suggested its implication in UC^{25,63}, whereas others which included less than 8 UC patients do not^{21,75}. On the other hand, no study exists analysing the prevalence of AIEC in patients with colorectal cancer (CRC) yet, but *E. coli* has been found more frequently in CRC patients than controls^{69,139–141} and strains with similar traits as AIEC have been isolated. In 2014, Raisch et al.¹⁴² found that although strains isolated from CRC patients poorly adhered to intestinal epithelial cells (I-407 cell line), they were able to promote high biofilm formation and to induce increased expression of CEACAM6 receptor to a similar level of LF82 (reference AIEC strain). Moreover, one study found that more than 50% of the *E. coli* isolated from CD (54%), UC (54%) and CRC (60%) are *afaC*-positive while only 28% of isolates from controls were positive. Of interest, the presence of *afaC* correlated with the capacity to adhere and invade IECs⁵⁰. Given that CEACAM6 implication with cellular adhesion, invasion and metastasis of tumour cells has been established¹⁴³ and *E. coli* from CRC patients presented some characteristics as those isolated from CD, it is suggested that *E. coli* found in cancer patients may belong to the AIEC pathotype and may contribute to carcinogenesis. To end, although AIEC could contribute to the symptomatology of irritable bowel syndrome (as patients with this disorder also present dysbiosis and inflammation), similar AIEC prevalence was reported recently between patients with irritable bowel syndrome (33%) and healthy controls¹⁴⁴.

In addition, AIEC do not only colonise the human intestine tract but also AIEC strains have been found in animals. Simpson et al.²⁶ isolated AIEC members from dogs with granulomatous colitis; a disease that highly resembles UC and CD in humans. Others also found AIEC in cows with bovine mastitis¹⁴⁵ and in cats, swine and dogs suffering from enteritis¹⁴⁶. Overall, they support the idea of AIEC being disease specific rather than host specific.

3. AIEC pathogenicity and virulence factors involved

Nowadays there are some evidences on the virulence mechanisms that the AIEC pathotype have developed to promote its adhesion, invasion and intramacrophage survival, albeit not unique to AIEC. These features are achieved by the evasion of host defence processes and disruption of epithelial barrier as well as by taking advantage of CD restrictions. AIEC virulence factors will be presented in the following sections and the main ones are summarised in Table 3.

Noticeably, in most cases only one AIEC strain (principally the prototype strain LF82) has been analysed and the majority of studies associated with bacterial invasion of IEC or intramacrophage survival have been performed with cells that resemble absorptive enterocytes (Caco-2, Hep2, and I-407) or in murine derived cell lines (J774), respectively (Table 3). Thereby further studies on other strains from this pathotype and the use of more accurate *in vitro* approaches or mice models would be required to complete our understanding.

Table 3. Summary of virulence factors related to AIEC virulence. In each case the process in which the virulence factor is involved, the AIEC strain and the cell line or mice model in which its function was assessed by isogenic mutants are depicted.

Virulence factor	Process	Studied AIEC	<i>in vitro/ in vivo</i>	References
AfaC	-IEC interaction	HM385	-Caco-2, Raji-B, J774, I-407 and Hep2 cell lines	Prorok-Hammon et al. ⁵⁰
ArlA, ArlC	-Mucus layer crossover	NRG857c	-C57BL/6 mice-isolated Paneth cells -CD-1 mice	McPhee et al. ¹⁴⁷
ChiA	-IEC interaction	LF82	-Caco-2 and SW480 cell lines -C57B1/6 mice	Low et al. ⁸⁵
DsbA	-Intramacrophage survival	LF82	-J774 cell line	Bringer et al. ¹⁴⁸
Flagella (FliC)	-Mucus layer crossover -IEC interaction	LF82, LF15, LF16, LF31, LF50 and LF65	-Hep2, I-407 and Caco-2 cell lines	Barnich et al. ¹⁴⁹ Sevrin et al. ¹⁵⁰
GipA	-M-cell interaction -Intramacrophage survival	LF82	-T84 and Caco-2 cell lines -HMDM and BMDM	Vazeille et al. ¹⁵¹
Hfq	-Intramacrophage survival	LF82	-J774 cell line	Simonsen et al. ¹⁵²
HtrA	-Intramacrophage survival	LF82	-J774 cell line	Bringer et al. ⁸⁹
IbeA	-IEC and M-cell interaction -Intramacrophage survival	NRG857c	-Caco-2, M-like and THP-1 cell lines	Cieza et al. ¹⁵³

Table 3. Continuation.

Virulence factor	Process	Studied AIEC	<i>in vitro/in vivo</i>	References
LpfA	-M-cell interaction	LF82 CUMT8	-Murine-isolated Peyer patches and Caco-2 cell line -Human-isolated Peyer patches from CD and non-IBD patients -M cells, generated by coculture of Caco2-cl1 cells and Raji-B lymphocytes	Chassaing et al. ⁸¹ Dogan et al. ¹⁵⁴
NlpI	-IEC interaction	LF82	-I-407 and J774 cell lines	Barnich et al. ¹⁵⁵
Type-1 pili (FimH)	-IEC and M-cell interaction	LF82, LF31, LF71, LF73 and LF100	-Human-isolated enterocytes from CD and non-IBD patients -Murine-isolated Peyer patches and Caco-2 cell line -Human-isolated Peyer patches from CD and non-IBD patients -CEABAC10 transgenic mice	Boudeau et al. ¹⁵⁶ Barnich et al. ⁸³ Carvalho et al. ⁸⁴ Chassaing et al. ⁸¹
Vat	-Mucus layer crossover	LF82	-mucin cells extracted from mouse intestine -CEACAM6 transgenic mice	Gibold et al. ¹⁵⁷
YfgL, OmpA and OmpC	-IEC interaction	LF82	-I-407 cell line	Rolhion et al. ¹⁵⁸ Rolhion et al. ¹⁵⁹ Rolhion et al. ¹⁶⁰

HMDM: human monocyte-derived macrophages. BMDM: bone marrow-derived macrophages.

3.1 AIEC and intestinal epithelial barrier

The most well-defined AIEC characteristic is its ability to adhere to and invade IECs but before getting in contact with them it first needs to cross the mucus layer. Virulence factors involved in early stages of mucosal invasion have been described in AIEC strains to achieve this crossover while evading host defence AMP (Figure 4.1). The **AIEC-Vat protease** enhances the degradation of mucins and therefore facilitates the spread of bacteria through the mucus layer in a murine model¹⁵⁷. Likewise the **FliC** protein, involved in flagella polymerisation, has been depicted to promote motility in AIEC^{149,150}. Although it is present in most enteric bacteria, its expression seems to be induced by the presence of mucus in AIEC strains but not in commensals¹⁵⁰. Moreover, alteration on the antimicrobial effect of host defence AMP can be produced by the presence of *arlA* and *arlC* genes in the plasmid of some AIEC strains¹⁴⁷. The first encodes for a Mig-14 family protein implicated in defensin resistance whereas the latter encodes for an OmpT family outer membrane protease. Altogether, expression of these VGs enhances AIEC fitness and gives a selective advantage to other strains in the gut.

Once AIEC are in contact with IECs, it is when adhesion and invasion occurs. It was reported that this process takes place through an actin microfilaments and microtubules-dependent macropinocytosis-like process in Hep-2 cells, where villi is elongated and finally engulfs the bacterium^{16,161}. Recently, a distinct mechanism was reported for AIEC (HM427 strain) internalisation in a cell line of colonic origin¹⁰⁴. Therein AIEC uptake via lipid rafts was proven, as it was seen decreased once Caco-2 monolayers were treated with an inhibitor of lipid raft (m β cd).

Several factors such as protein-receptor interactions may induce AIEC internalisation (Figure 4.2). Adhesion through **type 1 pili (FimH adhesin)** is one of the most studied virulence mechanism in AIEC pathotype. It specifically binds to oligomannose glycans on the surface of host cells, and it has been hypothesised that AIEC prefers those exposed on early apoptotic cells to promote its invasion¹⁶². FimH is expressed in the surface of AIEC isolates and interacts with the CEACAM6 receptor^{83,149,163}, which is found overexpressed on CD ileal enterocytes and its expression can be even exacerbated after AIEC infection by the induction of proinflammatory cytokines secretion¹⁶⁴. Moreover, Sevrin et al.¹⁵⁰ described higher expression of type I pili in AIEC strains than in non-AIEC strains evidencing again its role in AIEC phenotype. Even though, some polymorphisms in FimH sequence have been reported to confer higher adhesion ability as well^{138,165}. Decreased levels of the protease meprin, which may degrade type-1 pili, have been measured in CD patients, thus enhancing AIEC colonisation¹⁶⁶. In relation with type I pili, **flagella** is also required for adhesion and invasion by a direct and an indirect manner which affects type I pili expression maintenance¹⁴⁹ and transport of peptides through its machinery^{149,150}. Moreover, the chitinase 3-like 1 (CHI3L1) and the endoplasmic reticulum (ER)-localised stress response chaperone (Gp96) receptors have been described to be highly expressed in ileal and colonic IECs of CD patients respectively, due to gut inflammation. **ChiA**⁸⁵ and **OmpA**¹⁵⁹ proteins, found in AIEC, bind to them respectively to promote adhesion and invasion. The ChiA is a chitinase protein present in the bacterial membrane. Of note, increased TNF α secretion induces CHI3L1 expression and, as a result, provides higher AIEC affinity to enter the cells under inflammatory conditions. By transcomplementation analysis, the authors found that LF82 Δ *chiA*/*chiA*_{LF82} strain was able to colonise equally as the LF82-WT both in *in vitro* and *in vivo* assays while LF82 Δ *chiA*/*chiA*_{K-12} did not, suggesting that the difference on five amino acid positions between the ChiA sequence of LF82 and the commensal strain could explain the variable levels of bacterial adhesiveness and invasiveness⁸⁵. On the other hand, OmpA, found in the outer membrane of the bacteria, is

the main component of the outer membrane vesicles (**OMVs**). These can deliver virulence factors into host cells that contribute to the invasion process when in contact with IECs through the interaction of OmpA-Gp96 without the need of bacteria mobility, thus being independent of type 1 pilus and flagellum expression¹⁵⁸. In relation to OMVs, the deletion of *yfgL* gene resulted in a decreased ability to invade intestinal epithelial cells which was concomitant with a decrease in OMVs release¹⁵⁸. Furthermore, in 2010, Rolhion et al.¹⁵⁹ stated that internalisation of AIEC in the ileal mucosa of CD patients takes place in part by means of the interaction of OmpA protein and ER-localised stress response protein Gp96 rather than the quantity of OMVs released. High expression of this protein indicates ER stress in the host cells and as a consequence IEC, goblet cell and Paneth cell dysfunction. Another outer membrane protein was assessed to determine its role as an invasin. Indeed, it has been depicted that *ompC* gene expression, at high osmolarity condition, is increased favouring AIEC adhesion and invasion¹⁶⁰. Nevertheless, at the same time, the authors indicated that OmpC has an indirect role in AIEC phenotype, as its gene expression may be regulated via the σ^E -regulatory pathway and RpoE factor can bypass the effect of OmpC¹⁶⁰.

Other virulence factors in AIEC may be involved in its invasion capacities, however in those cases the exact mode of action is still under research. Such is the case of **IbeA**, which may be part of a secondary/complementary pathway. This assumption was made since the IbeA isogenic mutant (NRG857c Δ IbeA) invaded in much fewer levels than the wild-type but no differences on bacterial persistence in the intestine of a mice model were described¹⁵³. Surface-exposed lipoproteins (i.e. **NlpI**) can act as adhesins and its involvement on AIEC phenotype has been assessed by the construction of isogenic mutants^{155,158}. Although direct interaction with IECs is pointed out for the NlpI, in the NlpI mutant decrease on type I pili and flagella synthesis is also reported thus suggesting that NlpI may be involved in a two-component signal transduction pathway¹⁵⁵. When accounts for pili synthesis, **AfaC** protein has also described to play a role in bacterial IECs adhesion and invasion but scarce investigation on its mechanism has been conducted⁵⁰.

Mechanisms independent from those mentioned help dictate its translocation through AJC modifications or M-cells internalisation (Figure 4.2 and 4.3). AIEC interaction with IECs results in direct internalisation of AJC proteins and indirectly by activation of proinflammatory cytokines¹⁶¹. The structural units of adherens junctions (AJ) and tight junctions (TJ) are responsible for the regulation of barrier integrity and permeability. These

are interconnected with scaffolding proteins that cooperate with the cytoskeleton of the cell. As a consequence, once altered, cell structures are modified and permeability is increased. Dysregulation of TJ proteins has been observed in IBD patients¹⁶⁷⁻¹⁷⁰ and AIEC has been described to disrupt AJ and TJ complexes while reducing TEER^{93,171}. By *in vitro* studies^{93,171}, AIEC LF82 strain and other invasive *E. coli* strain have shown to use the **E-cadherin** displacement, the main component of AJ, as a potential mechanism to decrease epithelial barrier function. Moreover, AIEC-induced cytokines release by macrophages indirectly modifies the expression of several components of the AJC. For instance, the **zonula occludens-1**^{93,171} and the **occludin**¹⁶⁸ which are down-regulated resulting in increased appearance of gaps between cells. In contrast, increased **claudin-2** expression was reported in ileal biopsies of CD patients in quiescent phase compared to that in control biopsies¹⁷². Given that CEACAM6 was abnormally expressed in quiescent phase before the development of inflammation and significantly increased under inflammatory stimuli in the acute phase compared to the quiescent phase of CD⁸³, the authors hypothesised that this alteration in claudin-2 expression may occur due to the AIEC-CEACAM6 interaction¹⁷². Interestingly, the results of the studies assessing claudin-2 *in vivo* are similar to those from zonula occludens-1 since epithelial integrity is decreased too^{169,172}. In addition, AIEC have evolved to use M-cells as a gateway to invade the epithelium without the loss of the monolayer integrity⁴². Therein, AIEC translocation occurs by means of **type-1 pili** interaction with glycoprotein 2 (GP2) and via the binding of **LpfA** to a receptor that has not been identified yet^{81,156}. Recently, Vaizeille et al.¹⁵¹ stated the **GipA** factor as an AIEC virulence factor for its invasion of Peyer patches and dissemination to mesenteric lymph nodes by performing isogenic mutants of this gene. No reduction on IECs adhesion was seen neither on invasion. However, there was an impaired induction of the *hpf* operon, suggesting that the GipA factor may regulate LpfA expression. Therefore a GipA-mutant was prevented of the translocation across M cells. The absence of **IbeA** in the NRG857c AIEC strain also caused a significant reduction in intracellular AIEC in M-like cells¹⁵³. Nonetheless, after transcomplementation assays the wild-type levels were not achieved indicating that IbeA is not the sole invasion determinant of AIEC. To date, whether IbeA interacts with a potential receptor or not is still not elucidated.

During AIEC invasion several mechanisms are activated, mainly due to cytokines production and constant inflammation. AIEC interaction with IECs induces the secretion of cytokines (i.e. IL-8, TNF α and IFN γ) which in turn enhances the transmigration of immune cells and the reduction of the epithelial barrier resistance and lead to general

mucosa permeability defects^{164,171}. Even though, AIEC presents the capacity to block STAT1 activation after IFN γ stimulation in IECs, as assessed in LF82, NRG857c and UM146 strains¹⁷³. By obstructing this pathway, inflammatory responses to microbial infections do not occur at the same extent. For instance less immune cells will go to the site of infection and the transcription of IFN γ -dependent genes will be prevented. It has been hypothesised that AIEC secretes a factor (heat-resistant and proteinase K-sensitive), which could be the responsible for the subversion of the IFN γ -STAT1 pathway without the need to be internalised¹⁷³. Another intriguing mechanism that has been presented is the cell-to-cell communication pathway via exosomes, small membrane vesicles that can be released from different cell types¹⁷⁴. Exosomes have been involved in immune regulation processes such antigen presentation, T-cell activation and immune suppression¹⁷⁵. In 2016, it was reported that upon T84 cell line AIEC infection, high amounts of exosomes are release and at the same time NF- κ B activation and IL-8 production occurs without affecting integrity of the monolayer¹⁷⁶. Those exosomes have been stated to trigger proinflammatory responses not only in IECs but also in naïve macrophages (see introduction section 3.2). Furthermore, in the study of Mazzarella et al.¹⁶⁴, it was described that the host system for DNA repairing is blocked, accumulating high methylation levels and as a consequence the silencing of genes ending up in the IECs' apoptosis induction. Thus, apoptosis could also be a bacterial strategy to escape from infected cells and to invade deeper mucosal layers.

Special attention has been devoted to identify virulence-associated factors in AIEC adhesion and invasion processes. Overall it is suggested that AIEC is equipped with many proteins that are mediating host-bacteria interactions which are necessary to access the underlying macrophages present in the lamina propria.

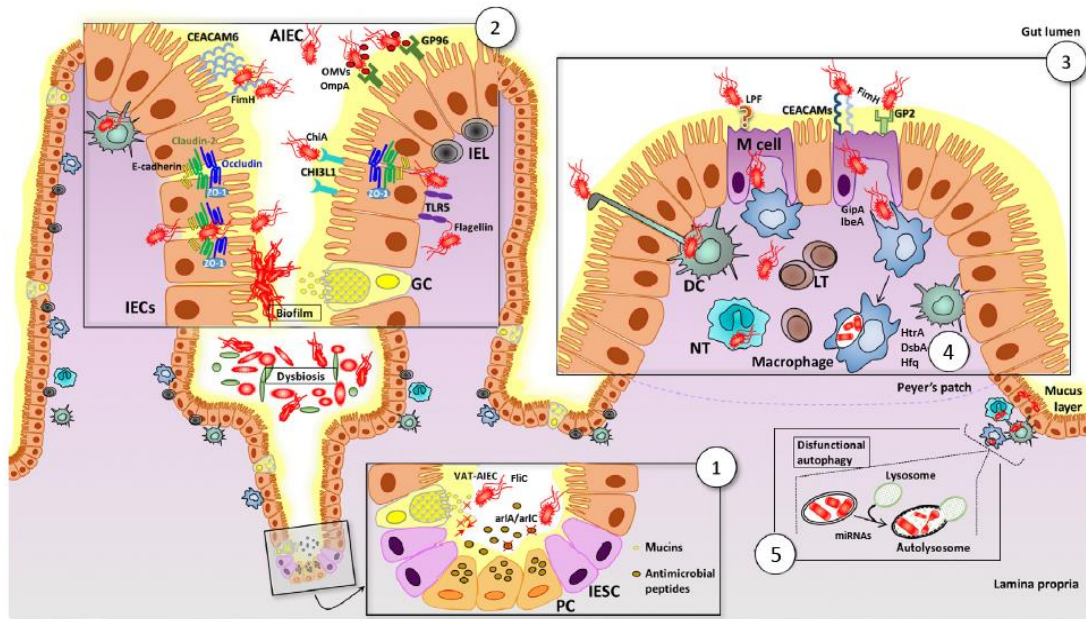


Figure 4. AIEC mechanisms of pathogenicity. 1: AIEC systems to cross the mucus layer and evade antimicrobial peptides. 2: AIEC mechanisms to adhere and invade intestinal epithelial cells, as well as, to enhance epithelial permeability. 3: AIEC interacting with M-cells of the Peyer patches. 4: AIEC dealing with immune cells. 5: AIEC and autophagy. Adapted from Palmela et al.¹⁰⁰. DC: dendritic cell. GC: globet cell. IECs: intestinal epithelial cells. IEL: intraepithelial lymphocyte. IESC: intraepithelial stem cell. LT: lymphocyte T. M cell: microfold cell. NT: neutrophil. PC: Paneth cell.

3.2 AIEC and immune cells

The immune cells located in the lamina propria are in charge of killing the pathogens. Nonetheless AIEC have been described to be able to survive and replicate within murine macrophages inside vacuoles with phagolysosomal traits, and also within human THP-1 macrophages and monocyte-derived macrophages (HMDMs) while preventing cell apoptosis^{22,86,89,90}. The ability to survive and replicate inside host macrophages implies supporting environments with low pH, low nutrient content, and high oxidative and nitrosative stress conditions. This AIEC capacity was demonstrated for the first time by Glasser et al.²² and it has been extensively confirmed by others^{127,177,178}.

Currently, the mechanism of how AIEC resist killing processes and adapt to the phagolysosome environment is still poorly understood. It seems that after AIEC internalisation in murine macrophages (J774 cell line), LF82 forms large vacuoles which may be composed by early endosomal antigen 1 rather than escaping from the vacuole, which was followed by acquisition of Rab7 GTPase and lysosomal-associated membrane protein-1 (Lamp-1), all features of late endosomes^{127,177}. Thus providing them with an early maturation and a higher acid pH and protease content¹⁷⁷. Although unexpected, it has been reported that once the pH was neutralised, intracellular replication of LF82 was inhibited¹⁷⁷

indicating that the acidic pH may induce the expression of particular genes involved in the AIEC persistence. At this step, AIEC induces the secretion of large amounts of pro-inflammatory cytokines, particularly TNF α , without inducing host cell death^{22,127,177}. In fact, evidences have been provided: (I) once levels of TNF α produced by macrophages are low, intramacrophagic bacteria is seen reduced⁹⁰ and (II) the amount of TNF α secreted correlates with the quantity of intracellular bacteria⁹⁰. In line with this observations, correlation between TNF α macrophage secreted levels and AIEC chronic activation of NF- κ B was previously reported to contribute to survival of the infected macrophages too¹⁷⁹. As mentioned before, this increase in cytokines production also provokes damages to the intestinal epithelia favouring AIEC colonisation (introduction section 3.1).

Hypothesis on how AIEC induces the inflammatory response and impair macrophages' death have been stated. First, it seems that AIEC prompt exosomes release in IECs and macrophages which in turn modulate immune response¹⁷⁶. Indeed, when exosomes isolated from AIEC-infected IECs were used to stimulate THP-1 macrophages, it responded with an amplified production of mitogen-activated protein kinase p38, cJUN N-terminal K and pro-inflammatory cytokines. Remarkably, AIEC intracellular replication was further promoted by the exosomes released from AIEC-infected IECs or macrophages. Hence, insinuating that AIEC may profit exosomes to amplify their own replication inside host cells. Nevertheless, in contrast to what is seen in other enteric pathogens¹⁸⁰, no production of IL-1 β as a result of AIEC inability to activate the inflammasome^{67,181} (a machinery by which caspases promote cleavage of pro-interleukin IL-1 β to mature active forms) has been reported, supporting the theory that AIEC-mediated pathogenesis is due to low persistent activation of gut immunity. Failure to produce IL-1 β either by a defective immune system or by contribution of bacterial metabolism might compromise bacterial clearance; even so a necessity to evaluate this trend is forthcoming. Moreover, Dunne et al.¹⁸² assessed AIEC persistence mechanisms in RAW264.7 and bone-marrow dendritic cells and concluded that inhibition of macrophages' apoptosis by AIEC may be crucial for its virulence. In absence of infection, spontaneous cell death is prevented by ubiquitination of caspase-3 to maintain basal level of caspase activity. Likewise, in AIEC-infected cells high levels of caspase-3 protein targeted for proteosomal degradation have been shown¹⁸². This process occurred simultaneously as the S-nitrosylation of caspase-3, a mechanism that inhibits caspase-3 activity and avoids undesired apoptosis^{183,184}. In this occasion it takes place independently of ubiquitination indicating that it might be carried out by AIEC mechanisms to ensure its intracellular persistence.

Some proteins described to have an essential role in the intramacrophage replication of pathogens such *Salmonella*¹⁸⁵ have also been key factors for AIEC intracellular survival (Figure 4.4). The first virulence factor related with AIEC equipment for intramacrophagic replication was **HtrA**, also known as DegP⁸⁹. This was presented in 2005 by assessment of LF82Δ*btra* mutant replication capacity in J774 murine cells. The same approach was conducted to assess the function of **DsbA**¹⁴⁸, **Hfq**¹⁵² and **IbeA**¹⁵³ although in the latter THP-1 human cells were used. All mutants presented a reduced capacity for survival, nonetheless each protein developed a different function. The stress-related protein (HtrA) provides resistance to acidic pH conditions⁸⁹ while the RNA binding protein (Hfq) a part from reducing sensitivity to low pH values, provides tolerance to reactive oxygen and nitrogen species¹⁵². Furthermore, survival in harsh conditions is favoured by DsbA protein which is responsible for the formation of intramolecular disulphide bonds¹⁴⁸. Finally, IbeA invasin, apart from its involvement in AIEC invasion, has also been related with bacterial survival in human macrophages¹⁵³. Despite the mechanism has not been elucidated, it is plausible that IbeA could play a role in tolerance to reactive oxygen species as it belongs to the FAD-dependent oxidoreductases family¹⁸⁶. Interestingly, it seems that the presence of the genes encoding for these proteins may not be a distinguishable trait across pathogenic and non-pathogenic *E. coli* strains yet its expression may be induced in specific conditions. For HtrA⁸⁹ and DsbA¹⁴⁸ higher expression was reported once the bacteria was grown inside macrophages or in a medium mimicking the conditions expected to encounter in a macrophage respectively. Similar observations were achieved recently in LF82 **GipA** mutant. Vazeille et al.¹⁵¹ reported, on one hand, higher *gipA* expression once AIEC is located inside bone marrow-derived macrophages or HMDMs. And, on the other, the mutant showed reduced replication capacity inside these types of macrophages and proinflammatory response is dampened.

Despite several factors have been suggested to be involved in intramacrophage survival, there are no functional studies looking for AIEC virulence factors related with its ability to restrict macrophage apoptosis. Two genes encoding for thioredoxins homologous to mammalian thioredoxins, proteins that are known to S-nitrosylate caspase-3, were found in LF82 genome¹⁸. While S-nitrosylation could be carried out directly by a bacterial virulence factor, it is more plausible to occur through an indirect effect. The perturbation of host cell pathways such as X-linked inhibitor of apoptosis pathway has been proposed¹⁸².

AIEC have been described to replicate not only within macrophages but also inside dendritic cells¹⁸² and neutrophils¹⁸⁷. In the first case, it seems that in order to avoid cell apoptosis AIEC follows the same inhibitory mechanism as explained for macrophages (S-nitrosylation)¹⁸². Nonetheless, in the case of neutrophils, the process of NETosis and subsequent autophagy is induced in infected cells indicating a different behaviour than in macrophages. NETosis involves externalisation of web-like structures rich in nuclear granular content, increase production of ROS and IL-8 and dysbiosis^{187,188}. The fact that AIEC enhance autophagic cell death may be used as a safeguard mechanism to control the number of polymorphonuclear lymphocytes and therefore limit polymorphonuclear lymphocyte-mediated chronic inflammation¹⁸⁷.

Finally, the involvement of AIEC in the formation of cell aggregates similar to epithelioid granulomas *in vitro* has also been pronounced⁹¹. Conversely, AIEC virulence factors related to its replication in neutrophils or its role in the formation of granulomas remain unidentified.

3.3 AIEC and autophagy

Bacteria within the macrophages or IECs are rapidly targeted to follow a lysosomal degradative pathway involved in identification and elimination of intracellular pathogens, known as autophagy. Some CD patients have mutations in innate response genes (ATGL16L1, IRGM and NOD2) which might impair autophagic responses⁸⁶. Hence, given that autophagy restricts the AIEC intracellular replication, gene defects could contribute to the overgrowth of AIEC. Indeed, Lapaquette et al.⁸⁶ reported AIEC intracellular persistence in a human cell line (THP-1) once the expression of ATG16L1 was decreased with a specific siRNA. However, apart from host genetic defects, AIEC have developed mechanisms to abrogate autophagy (Figure 4.5). It has been reported that intracellular LF82 activates NF- κ B^{92,179}, leading to the increased expression of the microRNAs (miRNAs), **MIR30C** and **MIR130A**, in T84 cells and in mouse enterocytes⁹². Given that miRNAs are small non-coding RNAs that post-transcriptionally regulate gene expression, predominantly through imperfect base pairing with the 3'-untranslated region (UTR) of target mRNAs¹⁸⁹, this upregulation diminishes the levels of ATG5 and ATG16L1, hence inhibiting autophagy and enhancing the inflammatory response. In addition, AIEC (LF82 strain) impairs host autophagy by modifying SUMOylation, a reversible post-translational modification that occurs in eukaryotes and when it is increased results in increased autophagy. This process has been described to take place upon

adhesion of LF82 to IECs partly via the inhibition of PIAS3 expression due to increased levels of miR-18¹⁹⁰.

3.4 AIEC and biofilm formation

As suspected by Martinez-Medina et al.¹⁹¹ who alluded stronger *in vitro* biofilm formation abilities for a group of AIEC strains in comparison with non-AIEC, a study conducted in 2013 by Chassaing et al.¹⁹² confirmed the ability of LF82 strain to form biofilms on IECs using cell culture and animal models. The σ^E -regulatory pathway already seen involved in AIEC adhesion and invasion processes (introduction section 3.1) was also responsible for the regulation of genetic determinants involved in biofilm formation. Later on, **WaaWVL** factors, were found to be regulated by this pathway and, apart from being involved in AIEC lipopolysaccharide structure and composition, they turn out to be essential for AIEC biofilm production and intestinal mucosa colonisation¹⁹³. Taking into account that biofilm formation can be a way to persistently colonise the intestinal mucosa and AIEC is able to produce it, this feature could also be considered part of AIEC pathogenesis.

3.5 Other pathogenic mechanisms

Up to this point, the AIEC virulence mechanisms and related virulence factors linked with AIEC main phenotypic characteristics have been presented. In this section, AIEC VGs that may indirectly modulate AIEC key features or that have been referred as potential virulence factors that could promote AIEC colonisation to the intestine will be detailed.

Some factors play a role on the AIEC colonisation of the host and virulence, via metabolic adaptation and global regulation. For example, the *pduC* gene, which encodes for the large subunit of propanediol dehydratase, is involved in fucose metabolism. Under anaerobic conditions, propanediol is produced when fucose, a component of mucin, is metabolised¹⁹⁴. Indeed, in fucose containing media the *pduC* gene presented increased expression than in normal media¹⁵⁴. This observation together with the fact that strains harbouring *pduC* gene presented higher Caco-2 invasion and J774 replication levels than those strains that are *pduC*-negative¹⁵⁴, suggests that having this gene could confer a competitive advantage for growth, invasion and perseverance within the intestinal mucosa. Similarly, Delmas et al.¹⁹⁵ suggested that AIEC strains present a competitive advantage respect to non-AIEC in an environment with bile salts, since LF82 presents high *eut* genes mRNA expression (*eutB*, *eutC*, *eutD*, *eutE*, *eutH* and *eutL*) and is able to utilise ethanolamine as a sole source of nitrogen whereas K-12 presents lower expression than LF82 and non-AIEC grow less in

minimal medium with bile salts and ethanolamine. Nonetheless, they indicated that the mucosal-associated LF82 does not differ in mice treated with a bile acid sequestrant in comparison to mice without this treatment and an *in vivo* competitive assay assessing the impact of bile salts in AIEC versus non-AIEC colonisation has not been performed yet.

Moreover, genes related with iron processes have also been revealed, such as *chuA* (heme acquisition), *fyuA* (yersiniabactin siderophore system), *sitA* (iron and manganese uptake system), and *irp1* and *irp2* genes (iron transport). In the inflamed intestine AIEC proliferation might be promoted by the iron derived from haemoglobin and serum¹⁹⁶. Despite the specific function has not been examined via isogenic mutants or murine models, strains presenting the *chuA* gene showed increased intramacrophage replication¹⁵⁴. Similarly occurred with the *fyuA* gene, which has been reported to be more prevalent in AIEC strains from IBD patients compared to those from non-IBD controls¹³⁶, and common with ExPEC strains able to survive and grow within intestinal tissues¹⁹⁷. Finally, as suggested by genomic studies^{18,19,197}, *sitA*, *irp1* and *irp2* genes related with iron acquisition could also impact on the cellular immune response and promote AIEC virulence, but together with *fyuA*, no functional studies have been carried out yet.

Additionally, the *nrdR* gene known to transcribe for a regulator of ribonuclease reductases have been described to promote AIEC (LF82) colonisation of the gut mucosa of a FVB/N CEABAC10 transgenic mice indirectly via the interference of bacterial motility and chemotaxis¹⁹⁸.

In contrast to the factors that have been outlined, the **Fis** protein contributes negatively to the AIEC virulence¹⁹⁹. This is a histone-like protein that operates as global regulators to control the expression of numerous genes, such as type-1 pili. In this case, low levels of *fis* gene expression were observed during AIEC invasion in comparison with gene expression in cell culture medium, suggesting a putative role of this gene. Therefore the LF82 mutant was created in the same study and its function was confirmed. It presented increased adherence levels since the mutant is unable to regulate the phase variation of type 1 pilus expression via the Fis protein and as a consequence, the population of AIEC LF82 bacteria associated with intestinal epithelial cells was predominantly in the ON phase¹⁹⁹. Other factors, such as the transcriptional activator **FliH2C2**, and sigma factor **FliA** have also been related with type-1 pili synthesis and flagella coordination²⁰⁰. Drastic decrease on the ability to adhere and invade IECs of LF82 was reported once the strain lacked these regulators. Nonetheless, this capacity was restored with the FliA transcomplementation,

what indicated that it compensated the FlhD2C2 function and suggested it as significant factor in AIEC pathogenicity. Finally it was deciphered that this protein controlled type-1 pili biogenesis, at least in part, via a c-di-GMP-dependent pathway using the YhjH as an intermediary²⁰⁰. Regardless of this information, the pathway connecting c-di-GMP levels to phase variation type-1 pili remains unidentified.

Factors that confer stress resistance to *E. coli* have been described (*rpoS*, *gadA*, *gadB*, *adiA* and *adiC* genes and lysine protein)^{201–203}. These may act with an indirect manner as their function on macrophage has not been assessed but they interact in conditions found inside the phagolysosomes. Even though, only **RpoS** has been tested in AIEC; deletion of **RpoS** from NRG857c strain has been observed to increase the sensitivity to oxidative stress and adherence of this clinical isolate, thus hampering AIEC to resist high levels of reactive oxygen species (ROS)²⁰³. However inactivation of *rpoS* and *tnaB* genes led to a reduced ability to adhere to Caco-2 cells in comparison to the wild-type strain²⁰³. The relation between RpoS and TnaB has still not been shed in light, but RpoS interaction with Hfq to enhance AIEC intramacrophage replication has been discarded¹⁵².

In addition, it remains to be determined whether **type VI secretion system** or the **KpsMTII protein**, related to capsule synthesis, contribute to AIEC pathogenesis. Nonetheless, up to now both have been found in the genome of several AIEC^{18–20,23,26}. Overall these findings suggest that AIEC use different subsets of genes during invasion and persistence.

4. AIEC genetic markers

AIEC strains isolated to date are clonally diverse and belong to distinct serotypes. Even though AIEC strains belonging to the A, B1, and D phylogroups have been isolated, they primarily fall into the B2 phylogroup^{17,23,43,68,69,74,136,154,171,204,205}, which contains the most virulence factors^{71,206}. In terms of VGs, AIEC resemble ExPEC, which are mostly non-invasive and the majority of them do not act like AIEC^{17–20}. To date, thirty-three AIEC genomes are public, they belong to various phylotypes (2 strains belong to A, 3 to B1, 24 to B2, 2 to D and 1 to F phylogroup) and have been isolated from different geographic regions (France, Canada, United Kingdom, USA and Australia)^{18,19,23,154,178,207,208}.

Several genes involved in AIEC pathogenicity have been depicted (introduction section 3). Nonetheless, most of them cannot be considered virulence factors per se, due to its ubiquitous expression in *E. coli* regardless of the pathogenic behaviour. Moreover, due to

its high genotypic variability, AIEC identification is currently challenging. Thus, as mentioned in introduction section 1.2, the only way to identify an AIEC strain is by assessing bacterial infection in cell culture assays which are non-standardised and highly time-consuming²¹.

Studies that have been carried to unravel AIEC characteristics either by PCR-based analysis^{20,21,43,74,81,136,146,209}, studying gene mutations^{23,74,85,138,159,165,178}, genome subtraction or comparative genomics^{17-19,23,101,154,178,210,211} or transcriptomics^{101,195} will be presented in the following sections. Altogether they showed that this pathobiont does not harbour an exclusive molecular signature, suggesting that AIEC strains are non-clonal and may have evolved from commensal *E. coli* by different mechanisms to favour their implantation in genetically susceptible CD patients. Putative AIEC molecular markers have been presented so far although they are either widely distributed or only present in a subgroup of AIEC strains. Therefore, no gene or sequence exclusive to the AIEC pathotype has yet been identified.

4.1 PCR-based gene prevalence

On account of the absence of known invasive VGs as firstly assessed by Darfeuille-Michaud et al.²¹, several studies have examined the prevalence of genes already reported to have a role in bacterial virulence with PCR reaction in order to decipher if they are associated with AIEC pathotype or the disease origin of isolation (Table 4).

In a study conducted in our research group⁴³, similar VGs distribution was found between AIEC and non-AIEC strains isolated from human intestine. When human extraintestinal strains were included, the *sfa/focDE* adhesin gene was more frequently found in non-AIEC (46%) than in AIEC (22%) strains²⁰. Conversely, the serum resistance associate gene (*malX*) and the group II capsule antigen gene (*kpsMTII*) were more common in AIEC (71% and 71% respectively) than non-AIEC (47% and 52% respectively) in a collection with human and animal extraintestinal and intestinal strains (IPEC)¹⁴⁶. Lately, five genes have been associated with the AIEC pathotype. The presence of *lpfA* and *gipA* was a specific trait of AIEC pathotype but not consistent across all AIEC strains. While 0.0% of non-AIEC strains harboured them, 31% of AIEC were PCR-positive for both genes¹⁵¹. In addition, the gene encoding for an outer membrane hemin receptor (*chuA*) was more prevalent in AIEC (93%) in comparison with non-AIEC (59%) strains isolated from Spanish and Chilean subjects⁷⁴. Otherwise, in *E. coli* strains isolated from irritable bowel disease patients

and healthy controls, *ibeA* and *colV* genes, typically associated with ExPEC, emerged as factors related with the pathotype. These genes were significantly more prevalent in AIEC strains (37% and 42% respectively) than in non-AIEC strains (3% and 16% respectively)¹⁴⁴. Despite the fact that particular genes have been related with AIEC phenotype, none of them have been confirmed by others perhaps due to strain collection diversity (i.e. origin of isolation or strain phylogenetic origin).

Table 4. Review of studies in which the prevalence of particular VGs has been examined according to the AIEC pathotype and origin of isolation. Genes associated with pathotype or origin of isolation are highlighted in bold.

Study	AIEC	Non-AIEC	Origin of isolation	Genes studied
Darfeuille-Michaud et al., 2004 ²¹	26	0	From ileal CD and controls plus colonic CD, UC and controls. Adults.	<i>afaD</i> , <i>eae</i> , <i>ipaC</i> , <i>tia</i> .
Martinez-Medina et al., 2009a ⁴³	22	38	From ileal and colonic CD and controls. Adults.	<i>afa</i> / <i>draBC</i> , <i>bfpA</i> , <i>cdtB</i> , <i>cnf1</i> , <i>eae</i> , <i>eltA</i> , <i>est</i> , <i>fimAvMT78</i> , <i>fimH</i> , <i>hlyA</i> , <i>ibeA</i> , <i>ipaH</i> , <i>iucD</i> , <i>neuC</i> , <i>papC</i> , <i>pCDV432</i> , <i>sfa</i> / <i>focDE</i> , <i>stx1</i> , <i>stx2</i> .
Martinez-Medina et al., 2009b ²⁰	27	59	From human extraintestinal infections and from the intestinal mucosa of CD, UC or controls. Adults.	<i>afa</i> / <i>draBC</i> , <i>bfpA</i> , <i>bmaE</i> , <i>cdtB</i> , <i>cnf1</i> , <i>cvaC</i> , <i>eae</i> , <i>eltA</i> , <i>est</i> , <i>fimA</i> , <i>fimAvMT78</i> , <i>fimH</i> , <i>focG</i> , <i>gafD</i> , <i>hlyA</i> , <i>ibeA</i> , <i>ipaH</i> , <i>iroN</i> , <i>iucD</i> , <i>kpsMII</i> , <i>kpsMIII</i> , <i>malX</i> , <i>neuC</i> , <i>papC</i> , <i>papGI</i> , <i>papGII</i> , <i>papGIII</i> alleles, <i>pCDV432</i> , <i>sat</i> , <i>sfa</i>/<i>focDE</i> , <i>sfaS</i> , <i>stx1</i> , <i>stx2</i> , <i>traT</i> , <i>usp</i> .
Martinez-Medina et al., 2011 ¹⁴⁶	49	134	From animal extraintestinal and intestinal infections and from the intestinal mucosa of CD, UC or controls.	<i>afa</i> / <i>draBC</i> , <i>astA</i> , <i>bmaE</i> , <i>chuA</i> , <i>cnf</i> , <i>csgA</i> , <i>cvaB</i> , <i>cvaC</i> , <i>eaI</i> , <i>eitA</i> , <i>eitC</i> , <i>etsB</i> , <i>etsC</i> , <i>fimC</i> , <i>focG</i> , <i>fyuA</i> , <i>gafD</i> , <i>gimB</i> , <i>hlyA</i> , <i>hlyF</i> , <i>bra</i> , <i>ibeA</i> , <i>iba</i> , <i>ireA</i> , <i>iroN</i> , <i>irp2</i> , <i>iss</i> , <i>iucD</i> , <i>iutA</i> , <i>kpsMTII</i> , <i>malX</i> , <i>mat</i> , <i>neuC</i> , <i>nfaE</i> , <i>ompA</i> , <i>ompT</i> , <i>papC</i> , <i>papEF</i> , <i>papGI</i> , <i>papGII</i> , <i>papGII/III</i> , <i>papGIII</i> , <i>pic</i> , <i>pks</i> , <i>sat</i> , <i>sfa</i> / <i>foc</i> , <i>sfaS</i> , <i>sitA</i> , <i>sitD</i> (<i>chr.</i>), <i>sitD</i> (<i>epis.</i>), <i>tia</i> , <i>traT</i> , <i>tsb</i> , <i>vat</i> .
Chassaing et al., 2011 ⁸¹		249	From ileal CD and controls. Adults	<i>lpfA</i>
Conte et al., 2014 ¹³⁶	27	0	From ileal CD and controls. Paediatrics.	<i>afa</i> / <i>draBC</i> , <i>aggR</i> , <i>cnf1</i> , <i>cvaC</i> , <i>fimH</i> , <i>focG</i> , <i>fyuA</i> , <i>gafD</i> , <i>hlyA</i> , <i>ibeA</i> , <i>iutA</i> , <i>kpsMT1</i> , <i>kpsMT5</i> , <i>kpsMTII</i> , <i>kpsMTIII</i> , <i>nfaE</i> , <i>pAA</i> , <i>PAI*</i> , <i>papA</i> , <i>papC</i> , <i>papEF</i> , <i>papG</i> alleles, <i>sfa</i> / <i>focDE</i> , <i>traT</i> .
Vazeille et al., 2016 ¹⁵¹	35	103	From ileal CD and controls. Adults	<i>lpfA</i>+<i>gipA</i>
Céspedes et al., 2017 ⁷⁴	15	37	From CD and controls. Adults	<i>afa</i> / <i>draBC</i> , <i>anfA</i> , <i>cdtB</i> , <i>chuA</i> , <i>cnf1</i> , <i>cvaC</i> , <i>eaaA</i> , <i>eatA</i> , <i>ecNA144</i> , <i>espC</i> , <i>espP</i> , <i>fhuD</i> , <i>fimAvMT78</i> , <i>fimH</i> , <i>gipA</i> , <i>hlyA</i> , <i>ibeA</i> , <i>irp2</i> , <i>neuC</i> , <i>papC</i> , <i>pet</i> , <i>pic</i> , <i>ratA</i> , <i>sat</i> , <i>sepA</i> , <i>sfa</i> / <i>focDE</i> , <i>sigA</i> , <i>tsb</i> , <i>vat</i> .
Dogan et al., 2018 ¹⁴⁴	19	57	From irritable bowel disease and controls. Adults	<i>afaC</i> , <i>chuA</i> , <i>cnf1</i> , <i>colIV</i> , <i>focG</i> , <i>fyuA</i> , <i>gsp</i> , <i>hcp</i> , <i>ibeA</i> , <i>iss</i> , <i>kpsMII</i> , <i>lpfA</i> , <i>malX</i> , <i>papC</i> , <i>pduC</i> , <i>pmt1</i> , <i>ratA</i> , <i>sfaDE</i> , <i>traC</i> .

CD: Crohn's disease patients, UC: Ulcerative colitis patients. *Pathogenicity island described in a virulent uropathogen.

Several studies have attempted to identify virulence factors associated with the disease origin of isolation. In the study conducted by Martinez-Medina et al.⁴³, the distribution of

18 VGs (Table 4) did not determine whether the AIEC strain was isolated from CD or controls. However, differences were found for other genes (*lpfA*⁸¹, *fyuA*¹³⁶, and *ibeA*¹³⁶ genes). Indeed, CD patients exhibited more *lpfA*-positive AIEC strains than controls⁸¹. Besides, in paediatric individuals, none of the AIEC strains isolated from controls (N=5) had the *fyuA* and *ibeA* genes whereas 70% of the AIEC strains isolated from CD patients (N=22) were *fyuA* and *ibeA*-positive¹³⁶. All in all, it evidences that equilibrated strain collections should be studied to further inspect AIEC specific characteristics.

4.2 Pathoadaptative mutations

Particular variants or point mutations of some genes have been related to AIEC virulence and their putative implication on AIEC pathotype identification has been examined. Particularly in the adhesin of type I pilus (FimH)^{23,74,138,165,178}, the outer membrane protein A (OmpA)¹⁵⁹ and the chitinase ChiA⁸⁵, all involved in bacterial adherence to IECs.

For FimH, previous studies have found some polymorphisms conferring higher adhesion ability but they have not detected a variant more prevalent in AIEC than in non-AIEC isolates^{23,74,138,165,178}, yet one has hypothesised that gene expression might explain the phenotype¹⁶⁵. In turn, some controversy exists regarding the mutations more frequent in IBD patients. Iebba et al.¹³⁸ found G66S and V27A variants more associated in CD patients and A242V, V163A and T74I variants common in UC. Instead, Dreux et al.¹⁶⁵ found no particular amino acid substitution associated to the origin of isolation of the strains but reported higher number of mutations in those strains isolated from IBD patients than from controls.

Regarding OmpA, five amino acid variants (V114I, F131V, D132Y, T228N and A276G) were described when AIEC reference strain LF82 and the commensal K-12 protein sequence was compared. In this study, Rolhion et al.¹⁵⁹ suggested that the amino acid substitutions present in the LF82 protein sequence favours invasion. Likewise occurred for ChiA, five amino acid changes (Q362K, E370K, V378A, V388E and E548V) were found located in a chitin binding domain of AIEC strain LF82 in comparison with K-12⁸⁵. These differences in the amino acid sequence were thought to be responsible for the ability of the strain to adhere and invade IECs, as well as, to be a putative AIEC identification marker. Nevertheless, in contrast with FimH, which has been studied in collections with more than 14 *E. coli* strains, the *ompA* and *chiA* genes sequences have only been assessed in LF82 and K-12 strains and no additional data on OmpA or ChiA sequence variants has been

published in other AIEC/non-AIEC strains. As a result, whether they are an AIEC-specific genetic marker remains unexplored.

At that point, given that neither prevalence nor point mutations of the already described VGs could uncover the basis of AIEC phenotype, to seek to identify new genetic elements and to apply novel techniques was required.

4.3 Comparative genomics

Approaches distinct from the ones mentioned above were also used to identify and characterise sequences that could explain the AIEC phenotype. These consisted on genome subtraction^{17,211} and comparative genomics (Table 5)^{18,19,23,101,154,178,210}.

The genome subtraction procedure consists on the characterisation of genetic fragments that are present in a particular bacterial genome and absent in a reference bacterial genome by PCR and cloning techniques. By genomic subtraction against the commensal *E. coli* K-12 reference strain, Baumgart et al.¹⁷ uncovered 115 genetic segments specific of 3 AIEC strains from different phylogroups (A: 541-15, B1: 541-1 and B2: LF82). More than 50% of these fragments encoded for hypothetical proteins or novel proteins of unknown functions. Sequences highly homologous to elements described in UPEC, APEC and other pathogenic *Enterobacteriaceae* (*ratA*, *hcp*, *pMT1* and *ColV*) were also described. Those have been related with bacterial colonisation (*ratA* and *hcp*) or plasmid-related (*pMT1* and *ColV*). Thus, the presence of the genes encoding for this proteins were screened in a collection of 22 strains but the presence of them did not correlate with strain pathogen-like behaviour in cultured cells; *pMT1* was only present in LF82 strain, *hcp* in 12/22 strains, *colV* in 8/22 and *ratA* was restricted to adherent and invasive strains isolated from I-CD (5/22). In concordance with these results, 27 out of the 58 genomic fragments found in a highly adhesive and invasive strain (HM229) were shared with other UPEC strains (CFT073 and UTI189) but not with EHEC strain (EDL933)²¹¹.

In 2010 the first AIEC genomes were sequenced and since then many comparative genomics studies have been conducted in attempts to elucidate the characteristics of the AIEC genome and to identify a genetic biomarker (Table 5). However, no gene or sequence exclusive to the AIEC pathotype has been identified yet.

Table 5. Summary of the comparative genomics studies conducted in AIEC to date. The strain collection examined according to pathotype and phylogroup is depicted. AIEC origin of isolation and study observations are also presented.

Study	AIEC	Non-AIEC	Phylogroup	AIEC Origin of isolation	Observations
Miquel et al., 2010 ¹⁸	1	21*	AIEC: B2 Commensals: 4A, 2B1, 1B2 ExPEC: 2B1, 6B2, 3D, 3E	From an I-CD patient	1 B2-AIEC genome sequenced
Nash et al., 2010 ¹⁹	2	10*	AIEC: B2 Commensals: 2A ExPEC: 7B2, 1E	From I-CD patients	1 B2-AIEC genome sequenced
Dogan et al., 2014 ¹⁵⁴	24	25	14 strains from A phylogroup, 16 B1, 10 B2 and 9 D	From I-CD patients and controls	8 AIEC genomes sequenced (4 from CD, 2 murine and 2 dogs) and 1 non-AIEC strain from CD ^β
Desilets et al., 2015 ²³	14 ^α	6	AIEC: A:1; B1:1; B2:10; D:1; F:1. non-AIEC: A:2; B1:2; B2:2.	From CD and UC patients ²¹²	11 genomes sequenced principally B2 (1A and 2 unreported)
Zhang et al., 2015 ¹⁰¹	13	11	AIEC: 1A, 1B1, 4B2, 1D, 5 Unknown. non-AIEC: 3A, 8 Unknown	From CD and UC patients and non-CD subjects.	AIEC genomes previously sequenced ^{18,19,207,208}
Deshpande et al., 2015 ²¹⁰	4	1307*	All B2	From CD patients	AIEC genomes previously sequenced ^{18,19,207,208}
O'Brien et al., 2015 ¹⁷⁸	11	30	All B2, ST95	From IBD patients and controls	10 B2-AIEC and 28 B2-non-AIEC genomes sequenced

CD: Crohn's disease. UC: Ulcerative colitis. *Include commensals and ExPEC. ^βHuman AIEC: 1A, 1B1, 1B2 and 1D; Murine AIEC: 1B1 and 1 B2; Dog AIEC: 2 B2; Human non-AIEC A phylogroup. ^αApart from LF82, UM146 and NRG857c the other strains were only assessed for intramacrophage replication in J774 cells.

Difficulties in discovering AIEC-specific traits have been probably due to the fact that the first studies compared AIEC and non-AIEC strains phylogenetically distant and the differences between these strains are related to their phylogenetic origin rather than the AIEC phenotype^{18,19,154}. In fact, by comparing the proteins encoded in AIEC genomes with other *E. coli* and *Shigella spp.* proteins, Dogan et al.¹⁵⁴ created a list of VGs that tend to be present in AIEC and absent in non-AIEC strains. They finally pointed out that *lpfA* and *pduC* genes were differentially distributed in a collection of 49 *E. coli* from both pathotypes, being more prevalent in AIEC strains (71% and 50% respectively) but also found in some non-AIEC (20% in both cases). Besides, 166 genetic segments were found across 13 AIEC strains, albeit not present in all, and absent in the 11 non-AIEC genomes studied¹⁰¹. Thereby, providing support to the concept that AIEC pathotype is heterogeneous. Recently, by comparing AIEC and non-AIEC strains of the same phylogroup, Desilets et al.²³ found three genomic regions present in all B2-phylogroup AIEC strains and absent

from AIEC strains of other phylogroups and commensal strains of any phylogenetic origin (including B2). However, whether these regions are specific to B2-AIEC strains only or also present in other pathogenic groups that share the same phylogenetic origin, such as B2 ExPEC strains is unknown. Deshpande et al.²¹⁰ described 29 diagnostic single-nucleotide polymorphisms (SNPs) that cause either synonymous or non-synonymous amino acid changes as a signature sequence that differentiates a group of B2-pathogenic strains comprising 4 AIEC, 3 ExPEC, 47 UPEC and 1 APEC strains from other *E. coli* strains present in the NCBI database. Nevertheless, no specific characteristic that discriminates the AIEC pathotype was found. Finally, O'Brien et al.¹⁷⁸ reduced gene content variability by conducting genomic analysis of a set of B2-phylogroup *E. coli* strains with identical sequence type (ST95), thereby decreasing the likelihood of detecting differential genetic elements delimited by the phylogenetic background. Nonetheless, the evaluation of gene prevalence and base composition of core genes did not result in the identification of an AIEC-specific biomarker or even in the identification of a marker common to most of the AIEC strains examined in that study. Taken together, these studies demonstrate that gene content is mostly associated with the phylogenetic origin of individual strains rather than with their AIEC pathotype.

4.4 Gene expression and comparative transcriptomics

Far from previous studies which reported either variances in the invasiveness ability of LF82 strain if blocked the expression of particular genes or different gene expression levels between conditions^{154,157,159,160,165,199,213}, there are scarce scientific works aiming at determining AIEC/non-AIEC differences based on its gene expression.

On one hand, differential expression between one AIEC and one non-AIEC strain has only been described for the *htrA* gene. Bringer et al.⁸⁹ assessed the transcriptional activation of the *htrA* gene in bacteria (LF82 and K-12) infecting macrophages (J774 cell line) by β -galactosidase activity. In this case, the intracellular AIEC strain showed $38.3 \pm 5.9\%$ -fold increase on *htrA* gene promoter expression while the intracellular K-12 had less than 10%-fold in comparison with bacteria grown in RPMI medium.

On the other, only two comparative transcriptomics studies have been published so far: one looked at differential gene expression depending on condition¹⁹⁵ and the other focused on AIEC/non-AIEC differences¹⁰¹. The first one detected higher expression of genes involved in ethanolamine utilization (*eutB*, *eutC*, *eutD*, *eutE*, *eutH* and *eutL*) once LF82 was

grown with the presence of bile salts (cholic acid sodium salt and deoxycholic acid sodium salt), and these were also overexpressed in LF82 in comparison to K-12 after incubation with bile salts¹⁹⁵. However, the expression of these genes have not been further analysed in clinical non-AIEC isolates, thus it cannot be considered an AIEC-specific trait, yet it might be strain-specific or specific of the two types of bile salts used. On the other hand, comparative transcriptomics analysis between LF82 and HS (commensal) in exponential and stationary phases grown in LB medium at 37°C has been carried out¹⁰¹. Therein, six genes were detected overexpressed in LF82 in comparison with HS in both time points. Two were related with bacteriophage infection, one in inorganic ion transport and metabolism and the others have unknown function. Moreover, from all the genes differentially expressed between these two strains, those transcripts that shared homology with 6 to 9 out of 13 AIEC strains encoded for an excisionase, CRISPR/Cas system and proteins involved in the propanediol metabolism. In the latter category, a gene that was previously associated with AIEC¹⁵⁴ (*pdu* gene) was stand out. Further characterisation of these proteins and its distribution and gene expression in AIEC strains is necessary to define its role in AIEC phenotype. Furthermore, this study presented some limitations: (I) the strain growth conditions assessed did not mimicked the environment found within the human intestine and (II) only coding regions found differentially expressed in LF82 vs HS were used to search homologies across a collection of 13 AIEC and 11 non-AIEC strains. Even though, so far, studies analysing AIEC gene expression during IECs adhesion, IECs invasion or intramacrophage replication and considering other AIEC strains (apart from LF82) are still missing.

• AIMS & SCOPE OF THE THESIS •

The involvement of the AIEC pathotype in CD pathogenesis has been extensively supported, as many researchers have reported higher AIEC prevalence in CD patients than controls (introduction section 2.4) and mechanisms of pathogenicity have been linked with CD physiopathology (introduction section 2.2). In CD, the therapeutic armamentarium remains limited and non-curative, hence the necessity to better understand AIEC pathotype as a putative instigator or the propagator of the disease is certain. Nonetheless, AIEC identification is currently challenging, as it relies on phenotypic assays based on infected cell cultures which are highly time-consuming, laborious and non-standardisable. To address this issue, AIEC molecular mechanisms and VGs have been studied but a specific and widely distributed AIEC genetic marker is still missing. The finding of molecular tools or rapid tests to easily identify the AIEC pathotype would definitely be of interest for scientists studying the epidemiology of the pathotype and clinicians that aim to detect which patients are colonised by AIEC to apply personalised treatments.

The **main goal** of this thesis is to gain insight into AIEC genetics and transcriptomics in order to look for signature traits that could assist in a rapid AIEC identification. Different approaches have been followed to address this purpose, which have been organised in three chapters with the following objectives:

Chapter 1 (AIEC characterisation based on known VGs): Given that AIEC is a genetically heterogeneous pathotype, the first approach consisted in the examination of VGs already described in other bacterial pathogens or related to AIEC virulence to further inspect AIEC specific characteristics. Previous studies analyzing the VGs distribution in intestinal AIEC/non-AIEC strains have focused on human-isolated strains^{20,21,43,74,81,136,146,209}, but AIEC have been isolated both from human and animals^{26,145,146}. Moreover, although FimH, ChiA, OmpA and OmpC have been involved in the interaction of AIEC with intestinal epithelial cells^{23,74,85,138,159,165,178}, scarce data exist about ChiA and outer membrane proteins (OMPs) sequence variants in a collection of AIEC strains and no study of OMPs gene expression during infection exists. Our hypothesis was that host origin of isolation may influence VGs carriage among AIEC strains and that combination of genetic and

phenotypic traits, as well as, point mutations or differential gene expression of genes could associate with AIEC virulence and/or could be used as AIEC molecular markers. Two specific objectives have been proposed.

1.1 To study the VGs carriage, to examine sequence variants of FimH and ChiA, and to determine if the combination of this genetic data with the antimicrobial resistance profile could be used to screen for putative AIEC strains in a collection of AIEC and non-AIEC strains isolated from humans and/or animals.

1.2 To determine if particular mutations or differential gene expression of OMPs are associated with AIEC virulence.

Chapter 2 (AIEC comparative genomics): AIEC genomic studies demonstrate that gene content is mostly associated with the phylogenetic origin of individual strains rather than with their AIEC phenotype. In this study, strains highly similar genetically but with divergent pathotype have been studied. Our hypothesis was that comparing genetically close strain pairs increases the likelihood of finding specific genetic elements characteristic of the AIEC phenotype. Moreover, from the results achieved therein, a second step was taken to assess gene putative implication in AIEC phenotype. Therefore, two specific objectives have been proposed.

2.1 To identify differences in gene content and new single nucleotide polymorphisms (SNPs) to distinguish between AIEC and non-AIEC strains by comparative genomics.

2.2 To generate isogenic mutants in order to study the role in AIEC pathogenicity of three genes encompassing SNPs associated with the AIEC pathotype.

Chapter 3 (AIEC comparative transcriptomics): Multiple studies have been conducted to identify AIEC candidate genes (introduction section 4) but only one study on AIEC transcriptomics compared to non-AIEC has been performed¹⁰¹ (introduction section 4.4) so far. In this case, the AIEC reference strain LF82 was compared against the non-invasive HS strain during exponential and stationary growth in LB medium. Given that no AIEC-specific genetic marker has been discovered, we have hypothesised that differences in the expression of genes could determine the AIEC phenotype. As gene expression highly depends on the environmental conditions, we suspected that genes expressed in a model of

infection would facilitate the finding of genes related with the AIEC phenotype. Therefore, in this chapter the following specific objective has been proposed.

3.1 To optimise a protocol for bacterial RNA extraction during eukaryotic cell infection and subsequent sequencing and to study the transcriptome of AIEC during growth in cell culture media and during intestinal epithelial cell infection in comparison with non-AIEC strains using RNA-Seq.

● MATERIALS & METHODS ●

Chapter 1.1: Virulence gene carriage and adhesin variants of AIEC and commensals isolated from humans and animals

1.1 *E. coli* strain collection

The *E. coli* collection used in this study was composed by three groups of strains: (I) strains previously isolated from the intestinal mucosa of CD patients and controls under the approval of the Ethics Committee of Clinical Investigation of the Hospital Josep Trueta of Girona on May 22, 2006⁴³ (II) strains previously isolated from animals suffering from enteritis under routine microbiological diagnostic procedures¹⁴⁶, and (III) strains newly isolated from CRC and UC patients (Table S1). Biopsies from CRC and UC patients were taken from the ileum and/or colon with sterile forceps, immediately placed in sterile tubes without any buffer, and maintained at 4°C for *E. coli* isolation. The study protocols for CRC and UC strains were approved by the local Ethics Committees (CEIC-Institut d'Assistència Sanitària, in April 2009 and January 2012; and CEIC-Hospital Universitari de Girona Doctor Josep Trueta, in May 2006). All subjects gave written informed consent in accordance with the Declaration of Helsinki. Additionally, the AIEC reference strain LF82, which was a kind gift from Prof. Darfeuille-Michaud (Université d'Auvergne, France), was also included. Information about the strains examined in each section (VGs prevalence, FimH and ChiA sequence variants and AB resistance) can be found in Table 6 and Table S2. The phylogenetic distribution of the strains studied in each section according to pathotype and origin of isolation is presented in Table S3.

Table 6. Strain collection used to study virulence gene carriage, gene sequence variants and the combination of *pic* gene and ampicillin resistance according to host, disease and pathotype.

Section	Host	Disease	AIEC	non-AIEC	B2-AIEC	B2-non-AIEC
Virulence gene prevalence	Human	CD	16	18	12	6
		C	6	19	3	8
	Animal	Enteritis	26	19	21	8
FimH and ChiA sequence variants	Human	CD	16	15	12	6
		C	6	12	3	7
		UC	7	0	ND	ND
		CRC	2	0	ND	ND
<i>pic</i> prevalence and ampicillin resistance	Human	CD	16	15	12	6
		C	6	12	3	7

CD: Crohn's disease; C: control; UC: ulcerative colitis; CRC: colorectal cancer; ND: not determined.

1.2 Adhesion and invasion assays

Adhesion and invasion assays were performed for isolates obtained from CRC and UC, whereas isolates from C, CD and animals were previously assessed^{143,146}. The Intestine-407 epithelial cell line (American Type Culture Collection (ATCC) CCL-6) was used for the adhesion and invasion assays. Cell culture, adhesion, and invasion assays were performed in triplicate as described previously¹⁶. Briefly, two 24-well plates containing 4×10^5 cells/well that had been incubated for 20 h were infected at a multiplicity of infection (MOI) of 10. Duplicate plates, one for the adhesion assay and one for the invasion assay, were incubated for 3 h at 37°C in 5% CO₂.

For the bacterial adhesion assays, the cell monolayers were washed five times with phosphate-buffered saline (PBS) and then lysed with 1% Triton X-100 (Sigma-Aldrich, St Louis, MO, USA). Adherent bacteria were quantified by plating them on Luria-Bertani (LB) agar (Liofilchem Srl, Italy). Plating was performed over a maximum period of 30 min to avoid bacterial lysis by Triton X-100. Adhesion ability (I_ADH) was determined by calculating the mean number of bacteria per cell. Isolates were considered adherent when I_ADH \geq 1. LF82 and K-12 strains have been used as positive and negative control respectively.

For the bacterial invasion assays, the monolayers were washed twice with PBS after 3 h of infection, and fresh cell culture medium containing 100 µg/ml gentamicin was added and left for 1 h to kill extracellular bacteria. After cell lysis with 1% Triton X-100, the number

of intracellular bacteria was determined by plating. Invasive ability was expressed as the percentage of the initial inoculum that became intracellular: $I_INV (\%) = (\text{intracellular bacteria} / 4 \times 10^6 \text{ bacteria inoculated}) \times 100$. Isolates were considered invasive when $I_INV \geq 0.1\%$.

1.3 Survival and replication within macrophages

The replication capacity of AIEC isolated from CRC and UC, as well as, non-AIEC strains isolated from CD and controls subjects was assessed in this study. The capacity of AIEC strains isolated from CD, controls or animals were previously assessed^{43,146}. For survival and replication assays, the murine macrophage-like J774A.1 cell line (ATCC TIB-67) was used and the ability of individual *E. coli* isolates to survive and replicate inside the macrophages was determined as described previously²².

J774 cell culture was performed, and the ability of individual *E. coli* isolates to survive and replicate inside the macrophages was determined as described previously²². Briefly, J774 macrophages were seeded at 2×10^5 cells per well in two 24-well plates. The plates were incubated for 20 h in complete medium (RPMI 1640 (Lonza, Switzerland) supplemented with 10% heat-inactivated FBS (Gibco BRL) and 1% L-glutamine (Gibco BRL)). After incubation, the medium was replaced with fresh medium, and bacteria were seeded at a multiplicity of infection of 100. To promote internalization of the bacteria by the macrophages, the plates were centrifuged at 900 rpm for 10 min and incubated for an additional 10 min at 37°C in 5% CO₂. Bacteria that were not phagocytosed were killed by inclusion of gentamicin (100 µg/mL) in the medium. After 40 min of incubation, one plate was washed twice with PBS, and 0.5 mL of 1% Triton X-100 (Sigma-Aldrich) was added to each well for 5 min to lyse the eukaryotic cells. To determine the number of intracellular bacteria recovered, samples were diluted and plated onto LB agar plates. The medium of the second plate was replaced with fresh cell culture medium containing 20 µg/mL gentamicin and incubated for 23 h. Then, the monolayer was washed and treated with 1% Triton X-100, and the cell suspension was diluted and plated as described above. LF82 and K-12 strains have been used as positive and negative control respectively.

Intracellular bacteria were quantified in the same manner as described for the invasion assays after 1 and 24 h of infection. The results are expressed as the mean percentage of bacteria recovered at 1 h and 24 h postinfection: $I_REPL (\%) = (\text{CFU mL}^{-1} \text{ at 24 h} / \text{CFU}$

mL⁻¹ at 1 h) x 100. Strains with an I_ADH higher than 1 bacterium/cell, I_INV of 0.1% and an I_REPL of 100% or higher were classified as AIEC strains in the present study.

1.4 Virulence genotyping by PCR

Fifty-four VGs from different groups, including adhesins, toxins, invasins, iron scavenging involved genes and genes involved in capsule formation and stress resistance, were amplified by PCR as defined previously¹⁴⁶. In addition, *lpfA* genes have also been studied in human-isolated strains. PCR primers for *lpfA141* and *lpfA154* genes were extracted from Chassaing et al.⁸¹ and PCR conditions were applied as explained therein. All genetic elements studied (either genes or alleles) were referred as VGs in this work.

1.5 Gene sequencing and sequence analysis

For *fimH* gene, PCR primers and program conditions were applied as described elsewhere¹³⁸. To sequence *chiA* gene, a set of four primers were designed in the present study. Two independent PCRs were performed in order to amplify the whole gene (2694 bp). The first PCR was carried with ChiA-84F (5'-TCATATTGAAGGGTTCTCG-3') and ChiA1711R (5'-TCCAGTCAACAAAAACACGC-3') leading to an amplicon of 1795 bp. The second PCR was carried with ChiA897F (5'-TAATAATGGCGGTGCTGTGA-3') and ChiA+12R (5'-TCGCCAACACATTTATTGC-3'), what resulted in an amplicon of 1818 bp. Primers ChiA897F and ChiA1711R were used to sequence a fragment of approximately 550 bp in the middle of the gene in which previously described mutations were located. PCR products were purified by ExoSap (Thermo Fisher) following manufacturer's instructions and sequenced by Sanger method (Macrogen, Netherlands). Sequences were cleaned and aligned with Bioedit software²¹⁴ using K-12 gene sequence as a reference (*fimH* gene ID: 948847; *chiA* gene ID: 947837) and uploaded in GenBank (MH730201 - MH730304). Nucleotide sequences were translated using EMBOSS Transeq²¹⁵. Reticulate trees were constructed with PopART software²¹⁶ using the median joining algorithm, considering only the variable DNA positions that caused non-synonymous amino acid changes.

1.6 Antibiotic resistance

The collection of strains isolated from human was screened against 30 antimicrobial agents using the Vitek®2 system (Biomérieux), the Sensititre standard susceptibility plate COMPAN1F (TREK Diagnostic Systems) or the macrodilution test following the Clinical

and Laboratory Standards Institute (CLSI) standards. Minimum Inhibitory Concentrations (MICs) were interpreted according to National Committee for Clinical Laboratory Standards (NCCLS) guidelines²¹⁷.

1.7 Statistical analysis

The significance of frequency values, for prevalence of VGs was measured by Pearson's χ^2 test using SPSS 23.0 software according to phenotype and phylogroup. In terms of differences in the frequency of particular mutations in the FimH or ChiA protein sequence, Pearson's χ^2 test was used only for those variable positions harboured by more than three strains. For quantitative variables (adhesion and invasion index), the Mann-Whitney non-parametric test was applied. Binary Logistic Regression was employed to depict a predictive model to classify AIEC strains. All data about VGs prevalence, amino acid variants and AB resistance were included in the model. In all cases, a p-value ≤ 0.05 was considered statistically significant.

Chapter 1.2: Amino acid substitutions and differential gene expression of outer membrane proteins in AIEC

2.1 *E.coli* strain collection

A collection of 13 AIEC strains isolated from CD patients and controls in a previous study⁴³ was analysed together with 30 non-AIEC strains that were isolated from the same group of subjects but that did not present the adherent-invasive phenotype (Table S4). The AIEC LF82 reference strain was included in the analyses¹⁶.

2.2 Amplification and gene sequencing

Strains were grown in LB broth overnight at 37°C. Total DNA was extracted by NucleoSpin® Tissue (Macherey-Nagel GmbH & Co. KG) kit following manufacturer's instructions. All genes (*ompA*, *ompC* and *ompF*) were amplified in a PCR reaction containing 1x Buffer II, 2 mM of MgCl₂, 0.2 mM of dNTPs, 0.5 mM of the corresponding primers (Table 7), 1 U/reaction of AmpliTaq Gold polymerase (Thermo Fisher, USA) and 1 μ L of DNA Template at 20 ng/ μ L in a final volume of 20 μ L. Amplification PCR program consisted in 1 cycle at 95°C for 10 min, 35 cycles of 45 sec at 95°C, 45 sec at the primer annealing temperature (Table 7) and 1 min at 72°C, finally, one cycle at 72°C during 10 min. The presence of only one band was checked by running the product on a 1.5%

agarose gel. Then, PCR products were cleaned by ExoSap (Thermo Fisher Scientific, USA) and sequenced with the Sanger method in both directions using the same primers as stated for amplification by Macrogen service (Korea). Consensus sequences were deposited under the accession number MH754762 - MH754812 (*ompA*), MH754813 - MH754863 (*ompC*) and MH754864 - MH754913 (*ompF*) in the GenBank database.

Table 7. Primers and probes used to amplify, sequence and to analyse differential expression of *ompA*, *ompC* and *ompF* genes.

Primer	Sequence (5'→3')	PCR fragment size (bp)	Annealing temperature (°C)	Reference
PCR amplification and Sanger sequencing				
<i>ompA</i> -F	TAAGCYTGC GGCTAGAGTTAC	1000	58	This study
<i>ompA</i> -R	ACCGTGT TATCTCGTTGGAG			
<i>ompC</i> -F	GCAGGCCCTTTGTTTCGATA	1236	58	Ruiz del Castillo et al., 2013 ²¹⁸
<i>ompC</i> -R	GCCGACTGATTAATGAGGGTTA			
<i>ompF</i> -F	GCAGTGGCAGGTGTCATAAA	1158	60	Ruiz del Castillo et al., 2013 ²¹⁸
<i>ompF</i> -R	TCGGCATTTAACAAAGAGGTG			
RT-qPCR				
<i>ompA</i> -F89-107	CTGGTGCTAAACTGGGCTG	126	56	This study
<i>ompA</i> -R179-200	TTAACCTGGTAACCACCAAAAAG			
<i>ompC</i> -F766-786	CTGAGCAGCCAGGTAGATGTT	369	58	This study
<i>ompC</i> -R423-440	CATGCAGCAGCGTGGTAA			
<i>ompF</i> -F468-487	CGGCGTTGCTACCTATCGTA	305	56	This study
<i>ompF</i> -R731-751	CTGCCAGGTAGATGTTGTTTCG			
<i>gapdb</i> -F752-773	CAACTTACGAGCAGATCGAAGC	170	57	Modified from Viveiros et al. 2007 ²¹⁹
<i>gapdb</i> -R902-923	AGTTTCACGAAGTTGTCGTTCA			
<i>16S E. coli 395F</i>	CATGCCGCGTGTATGAAGAA	105	60	Huijsdens <i>et al.</i> 2002 ²²⁰
<i>16S E. coli 490R</i>	CGGGTAAACGTC AATGAGCAAA			
<i>16S E. coli 437PR</i>	TATTAAC TTTACTCCCTTCCTCCCCGCTGAA			

2.3 Sequence analysis

The consensus sequence for each strain was aligned with the corresponding gene sequence of the LF82 reference strain extracted from the database (Accession id: CU651637.1) using BioEdit²¹⁴. To identify nonsynonymous point mutations, DNA sequences were translated to amino acid using EMBOSS Transeq (EMBL-EBI)²¹⁵.

Phylogenetic analyses were represented with an amino-acid based reticular tree constructed with the Popart software (version 1.7) using the median-joining algorithm for each gene²¹⁶. In all cases, gene sequences from other AIEC, commensal, ExPEC (UPEC, MNEC and APEC) and IPEC (EAEC, EHEC, ETEC, EIEC, DAEC and STEC) strains retrieved

from the GenBank were also included: AIEC UM146 strain (CP002167.1); AIEC NRG857c strain (CP001855.1); Commensal HS strain (CP000802.1); Commensal K-12 strain (CP012868.1); Commensal ED1a strain (CU928162.2); UPEC CFT073 strain (NC_004431.1); UPEC 536 strain (CP000247.1); UPEC UMN026 strain (CU928161.2); MNEC S88 strain (CU928163.2); APEC APEC01 strain (CP000468.1); EAEC 042 strain (NC_017626.1); EHEC EDL933 strain (AE005174.2); EHEC 0154 Sakai strain (BA000007.2); ETEC E24377A strain (CP000800.1); EIEC CFSAN0299787 strain (CP011416.1); DAEC SaT040 strain (CP014495.1); STEC ST540 strain (CP007265.1).

2.4 OMPs isolation and separation by SDS-PAGE

OMPs were isolated as described previously²²¹ with some modifications. Overnight culture in Mueller-Hinton (MH) broth, each strain was harvested by centrifugation and resuspended in 1 mL Tris-Mg buffer (10 mM Tris-HCl, 5 mM MgCl₂, pH 7.3). Cells were sonicated at 15% amplitude using a 1/8" diameter tapered horn for 5 cycles as previously described. Unbroken cells were eliminated by centrifugation at 5000 rpm and 4°C for 5 min, and cell envelopes were recovered by centrifugation at 17000 rpm and 4°C for 30 min. Membranes were solubilised in 2% sodium lauroyl sarcosinate for 30 min at room temperature, and centrifuged at 17000 rpm and 4°C for 30 min. Pellet was washed in 1 mL Tris-Mg buffer, centrifuged as above and finally solubilised in 40 µL Tris-Mg buffer. Protein concentration was quantified with the Quick Start Bradford 1xDye Reagent (BioRad, USA).

Separation analysis of OMPs was performed in urea-SDS-PAGE. Resolving gel was 10% acrylamide-0.27% N,N'-Methylenebisacrylamide, 6 M Urea, 375 mM Tris-HCl pH8.8, 0.2% SDS, 0.2% TEMED, 0.075% Amonium persulfate. Stacking gel was 5% acrylamide-0.13% N,N'-Methylenebisacrylamide, 6 M Urea, 125 mM Tris-HCl pH 6.8, 0.2% SDS, 0.06% TEMED, 0.1% Amonium persulfate. 15 µg of protein were loaded and the gel was stained with Coomassie as in Hernández-Allés et al.²²¹

2.5 Infection and RNA extraction

Intestinal epithelial cells (I-407 cell line; ATCC CCL-6) were grown in a T25 flask to a density of 6.6×10^6 total IECs at 37°C with 5% of CO₂ during 20 hours. Each flask was infected at a MOI of 100. After 4 hours, the supernatant containing bacteria growing in suspension (SN) was separated from the infected cells (INV) which contained the adherent and invasive bacteria. The monolayer (INV) was washed two times with EMEM (Lonza,

Switzerland) + 10% FBS inactivated (Gibco, USA) and cells were collected with scrapper. Then, both fractions (SN and INV) cells were centrifuged at 3000 x g for 10 min at 4°C. The supernatant of each fraction was discarded and the pellet washed with PBS (Lonza, Switzerland). Total RNA extraction was performed with TRIzol Max Bacterial Isolation kit (Invitrogen, USA) with some modifications. After the addition of chloroform incubation at room temperature of 15 min was done and an overnight precipitation at -20°C was performed after isopropanol addition. Subsequently, a DNase I - RNase-free treatment (Thermo Scientific, USA) was used to eliminate any possible DNA contamination in the sample.

2.6 Gene expression quantification by RT-qPCR

Total RNA (2 µg) was reverse transcribed using random hexamer primers with High-Capacity cDNA Reverse Transcription Kit with RNase inhibitor (Thermo Fisher, USA). RT-qPCR were performed with primers (Table 7) designed using Primer3 (version 0.4.0) based on the gene sequences of the strain collection that were obtained in this study. All primers were further analysed with NetPrimer to select the optimal primer pair. The amplification reactions were carried out in a total volume of 20 µL containing: 10 µL of Power SYBR™ Green PCR Master Mix 2x (Applied Biosystems, USA), 300 nM for each primer and Rnase free water up to the final volume. All quantitative PCR were performed using a 7500 Real Time PCR system (Applied Biosystems, Foster City, CA, USA). Thermal cycling conditions consisted of an initial step at 50 °C for 30 min, a PCR activation step at 95 °C for 10 min to denature DNA and activate Ampli-Taq Gold polymerase, and a further denaturation step of 40 cycles (95°C for 15 sec) followed by an annealing and extension step at 60 °C for 1 min. Data was collected and analysed with the 7500 SDS system software version 1.4 (Applied Biosystems, Foster City, CA, USA).

Samples were quantified in triplicate. Relative abundance of transcripts (RTA) for each gene of interest was determined by applying the comparative threshold cycle (Ct) method²²². Differences in expression levels were normalised against the *E. coli* house-keeping gene *gapA* measured in the same sample (ΔCt) and compared with LF82 INV gene expression by the equation $RTA = \frac{Efficiency^{(Ct\ target\ gene\ reference\ strain - Ct\ target\ gene\ sample)}}{Efficiency^{(Ct\ constitutive\ gene\ reference\ strain - Ct\ constitutive\ gene\ sample)}}$ ²²². The efficiency was calculated based on the standard curve ($E = 10^{-1/slope}$). Additionally, 16S rRNA copy number was also evaluated in all samples in duplicate. Reactions were carried out in a total volume of 20 µl, and reaction contained 1x Taqman

universal PCR master mixture (Applied Biosystems, Foster City, CA, USA), 300 nM each *E. coli*-specific oligonucleotide primer and 100 nM fluorescence-labelled *E. coli*-specific probe (Table 7). The same thermal cycle as explained before was performed. The value obtained was used to normalise the RTA to the bacterial quantity in each sample, which was subsequently normalised by the RTA/16S LF82 INV sample value to obtain a ratio; therefore, a ratio similar to 1 indicates a similar expression level between the corresponding sample and the reference strain (LF82 INV).

2.7 Statistical analysis

Differences in the amino acid present in each variable position between pathotype and phylogroup origin, were calculated using the χ^2 test. For phylogroup analysis, the atypical strain was not contemplated. To compare the mean adhesion and invasion indices between more than two amino acid variants the non-parametric Kruskal-Wallis test was used while the Mann-Whitney U-test was performed to analyse pairwise comparisons. Differences in OMPs protein expression and OMPs protein profiles were determined using the χ^2 test. To analyse differential expression according to pathotype, the Mann-Whitney U-test for two independent samples was applied. Gene expression versus adhesion and invasion capacities was calculated using Spearman's correlation. Finally, paired analysis according to condition was performed with Wilcoxon signed-rank test. A p-value ≤ 0.05 was considered statistically significant in all cases.

Chapter 2.1: Identification by comparative genomics of new single nucleotide polymorphisms to distinguish between AIEC and non-AIEC strains

3.1 *E. coli* strain selection and characterisation

Three *E. coli* strain pairs isolated in a previous study⁴³, each consisting of one AIEC and one non-AIEC of identical pulsotype and belonging to a distinct phylogroup (B1, B2 or D), were selected. The selection criterion for AIEC strains was based on: i) possessing different phylogenetic origins, ii) displaying high adhesion (>15 bacteria/I-407 cell) and invasion (>0.266% of inoculum surviving after 1 h of gentamicin treatment) indices in Intestine-407 (I-407; ATCC CCL-6) cells and iii) possessing an ExPEC-like genotype. Non-AIEC strains with pulsotypes identical to those of each selected AIEC were searched in the *E. coli* collection obtained for each patient in the same previous study⁴³. The study from which the AIEC strains were obtained was approved on May 22, 2006 by the Ethics

Committee of Clinical Investigation of the Hospital Josep Trueta of Girona. AIEC07/ECG04 pair was isolated from the ileum of a control patient while AIEC17/ECG28 and AIEC01/ECG11 pairs were isolated from the colon and ileum of an I-CD patient, respectively. Information on the patients from whom all the strains used in this study were isolated is presented in Table S5. The main characteristics of the sequenced strains are shown in Table 8.

Multilocus sequence typing (MLST) was performed *in silico* by querying the sequences of 7 housekeeping genes (*adk*, *fumC*, *gyrB*, *icd*, *mdh*, *purA*, and *recA*) extracted from the *E. coli* MLST Database (University of Warwick) against each genome. Each allele is identified by a numeric marker. The combination of the 7 numeric markers for each strain was collected and used to obtain the Sequence Types (ST) (Table 8).

Strain clonality was checked by pulsed-field gel electrophoresis (PFGE) as described elsewhere (CDCPulseNetUSA, 2004). Agarose-embedded DNA was digested with 0.2 U/ μ L XbaI (Takara Bio) according to the manufacturer's instructions. The XbaI-digested genomic DNA was analysed on a 1% agarose gel in 0.5X Tris-boric acid-EDTA buffer at 14°C using the CHEF-DR III System (Bio-Rad). The gel was run for 19 h at 6 V/cm, with initial and final switch times of 2.2 sec and 54.2 sec, respectively. The gel was stained with ethidium bromide (1 μ g/mL), and TIFF images were normalised and calibrated using GelComparII software (Applied Maths). Curve-based dendrograms were created using Pearson correlation coefficients, applying 0.5% optimization and 0.5% of curve smoothing and the UPGMA clustering method. The dendrogram of the strains is shown in Figure S1.

Phenotypic characterisation of the selected strains was performed to determine adhesion and invasion of the Intestine-407 epithelial cell line (ATCC CCL-6) as detailed in the methods section 1.2. In addition host cell cytoskeleton involvement was evaluated as described by Baumgart et al.¹⁷. I-407 cells were seeded at a density of 4×10^5 cells/well; after 24 hours, the monolayers were incubated with cytochalasin D (0.5 μ g/mL) or colchicine (1 μ g/mL) for 30 min to depolymerise microfilaments and microtubules, respectively. The monolayers were then manipulated as described for the invasion assays. Finally, the inhibitory effect was determined and presented as the percentage of reduction of invasion indices.

Survival and replication within two macrophage cell lines (J774A.1 (ATCC TIB-67) and THP-1 from mouse and human, respectively) was also assessed. Assays corresponding to

J774 were conducted as detailed in the methods section 1.3. The human THP-1 cell line (ATCC TIB-202) was maintained in RPMI 1640 medium (Lonza, Verviers, Belgium) supplemented with 10% (vol/vol) foetal bovine serum (Linus) in an atmosphere containing 5% CO₂ at 37°C. THP-1 cells were seeded in two 24-well plates at a density of 5x10⁵ cells per mL and were grown in complete medium containing 20 ng/mL of phorbol 12-myristate 13-acetate (PMA; Sigma-Aldrich) for 24 hours to promote monocytic differentiation. After incubation, the medium was replaced with fresh medium (RPMI + 10% heat-inactivated FBS), and bacteria were seeded at a MOI of 100. Similar to the J774 monolayers, the THP-1 plates were centrifuged at 900 rpm for 10 min and incubated for an additional 10 min at 37°C in 5% CO₂. The cell monolayers were washed twice with PBS, and fresh cell culture medium containing 100 µg/mL gentamicin was added to kill extracellular bacteria. After 40 min of incubation, one plate was washed twice with PBS, and 0.5 mL of 1% Triton X-100 (Sigma-Aldrich) was added to each well for 5 min to lyse the eukaryotic cells. To determine the number of intracellular bacteria recovered, samples were diluted and plated onto LB agar plates. The medium of the second plate was replaced with fresh cell culture medium containing 20 µg/mL gentamicin and incubated for 23 hours. Then, the monolayer was washed and treated with 1% Triton X-100, and the cell suspension was diluted and plated as described above.

Intracellular bacteria were quantified in the same manner as described for the invasion assays after 1 and 24 hours of infection. The results are expressed as the mean percentage of bacteria recovered at 1 h and 24 hours postinfection: I_REPL (%) = (CFU mL⁻¹ at 24 h / CFU mL⁻¹ at 1 h) x 100.

Strains with an adhesion index ≥ 1, an invasion index ≥ 0.1 (that was reduced by 90% to 99.9% when the microfilament inhibitor cytochalasin D and the microtubule inhibitor colchicine were added), and a replication index ≥ 100% in J774 and THP-1 were classified as AIEC strains in the present study.

A selection of genes previously associated with the AIEC phenotype either due to their higher prevalence in the pathotype (genes *lpfA*¹⁵⁴^{81,154}, *gipA*²⁰⁹, *cbuA*¹⁵⁴, *fyuA*²²³, *afaC*⁵⁰, *pduC*¹⁵⁴ and *ibeA*¹⁵³) or due to the presence of amino acid variants relevant to the pathotype (FimH¹⁶⁵, OmpA¹⁵⁹ and ChiA⁸⁵) was examined using BLASTn²²⁴ and ClustalW alignment²²⁵.

Table 8. Characteristics of the three sequenced AIEC/non-AIEC strain pairs.

Strain	Phylogroup	Serotype	ST ^a	Virulence genes	Adhesion index ^b	Invasion index ^c	Intramacrophage replication index in J774 ^d	Intramacrophage replication index in THP-1 ^d
AIEC17	D	ONT:HNT	569	<i>fimC</i> , <i>mat</i> , <i>ompA</i> , <i>ea/I</i> , <i>sitA</i> , <i>sitD_ch</i> , <i>irp2</i> , <i>fyuA</i> , <i>chuA</i> , <i>vat</i> , <i>ibeA</i> , <i>kpsMTII</i> , <i>neuC</i> , <i>traT</i> , <i>csgA</i> , <i>fimH</i> , <i>chiA</i> , <i>gipA</i> , <i>pduC</i>	21.6±17.5	0.266±0.055	1053±75	213±60
ECG28					0.6±0.3	0.005±0.001	774±129	228±68
AIEC01	B2	O6:H1	73	<i>foeG</i> , <i>mat</i> , <i>iba</i> , <i>ompA</i> , <i>pic</i> , <i>sitA</i> , <i>sitD_ch</i> , <i>irp2</i> , <i>incD</i> , <i>iutA</i> , <i>fyuA</i> , <i>vat</i> , <i>pks</i> , <i>kpsMTII</i> , <i>traT</i> , <i>fimH</i> , <i>chiA</i> , <i>gipA</i> , <i>chuA</i>	15.9±9.3	0.284±0.106	1567±1060	173±99
ECG11					1.1±0.8	0.004±0.002	716±315	74±29
AIEC07	B1	O22:H7	3232	<i>fimC</i> , <i>mat</i> , <i>csgA</i> , <i>ompA</i> , <i>fimH</i> , <i>lpfA₁₅₄</i> , <i>gipA</i> , <i>chuA</i> , <i>fyuA</i>	20.0±13.4	0.565±0.392	1693±297	189±71
ECG04					1.8±0.7	0.036±0.029	527±194	77±17

^a Sequence type. ^b Number of bacteria per I-407 cell. ^c Percentage of intracellular bacteria after 1 h gentamicin treatment relative to the inoculum. The percentages of reduction of invasion for AIEC17, AIEC01 and AIEC07 were 99.8%, 99.4% and 99.8%, respectively, in the presence of cytochalasin D and 90.4%, 99% and 95%, respectively, in the presence of colchicine.

^d Percentage of intracellular bacteria present at 24 h post-infection relative to the number of intracellular bacteria present after 1 h of gentamicin treatment. Results of control strains LF82 and K-12 strains respectively: φ 25.66±15.7 and 0.70±0.02. ψ 2.26±1.349 and 0.019±0.020. ϕ 777±304.8 and 11±5. γ 121±59 and 10±7. Strains with an adhesion index \geq 1, an invasion index \geq 0.1% and an intramacrophage replication index \geq 100% were classified as AIEC.

3.2 Genomic DNA extraction and sequencing

Genomic DNA was extracted from bacterial cells cultured overnight in LB culture broth using the Wizard® Genomic DNA Purification kit (Promega) according to the manufacturer's instructions; samples were treated with RNase A provided with the kit. DNA purity was determined using a NanoDrop ND-100 spectrophotometer (NanoDrop Technologies), DNA quantity was measured using a Qubit® 2.0 Fluorometer (Life Technologies), and DNA integrity and RNA elimination were examined on agarose gels. Unique bands of approximately 23 Kb were identified in agarose gels. The 260/280 ratio of the DNA preparations ranged from 1.8 to 2, and the quantity of DNA ranged from 15 to 30 µg, indicating sufficient quality of the genomic DNA for genome sequencing. Two sequencing platforms, Illumina HiSeq and PacBio Biosciences were used.

For Illumina sequencing, DNA samples were converted into sequencing libraries using the Illumina TruSeq DNA sample preparation kit at EA Quintiles. Briefly, 1 µg of genomic DNA was fragmented to ~200 bp using a Covaris E210 ultrasonicator. The fragmented DNA was then blunted, and a single "A-tail" was added to the 3' end of each fragment to facilitate ligation of sequencing adapters containing a single T base overhang. The adapter-ligated DNA was amplified by the polymerase chain reaction to increase the amount of sequencing-ready DNA in the library. The final DNA libraries were analysed for size distribution and quality using an Agilent Bioanalyser (DNA 1000 kit, Agilent # 5067-1504), quantitated using Picogreen (Life Tech # P11496), and normalised to a concentration of 2 nM. Equal volumes of the normalised DNA libraries were pooled, and the pooled DNA was used to prepare a flow cell using the Illumina TruSeq Paired-End Cluster Kit V3 (Illumina # PE-401-3001). The pools were denatured using fresh 0.1 N NaOH and diluted to 20 pM in chilled hybridization buffer. The pools were further diluted to 9 pM, and an aliquot of each was placed in an Illumina cBot instrument to produce clusters through bridge amplification. Sequencing was conducted on an Illumina HiSeq 2000 using 100-base paired-end sequencing plus a 7-base index cycle.

For the PacBio sequencing, the DNA libraries were prepared following PacBio guidelines and sequenced on SMRT cells using Pacific Biosciences RS sequencing technology (Pacific Biosciences, Menlo Park, CA, USA) at EA Quintiles. Ten micrograms of genomic DNA were purified using the PowerClean® DNA Clean-Up Kit (MO BIO Laboratories) and then sheared to 2 kb using a Covaris® Adaptive Focused Acoustics instrument. The

sheared DNA was purified using magnetic beads and verified on a Bioanalyser. Library preparation was performed using the Pacific Biosciences DNA Template Prep Kit 2.0 (3 Kb - 10 Kb). Size selection and library purification were performed using 0.6X AMPure beads (Beckman-Coulter Genomics). Each library was bound to C2 DNA polymerase, loaded into a SMRT cell, and sequencing was observed using two 45-min movies for each cell. Quality analysis of the raw data was performed with PRINSEQ.

3.3 *De novo* genome assembly

Draft genomes were assembled *de novo* (combining both platforms) using SPAdes software (ABL)²²⁶ and annotated using the BG7 bacterial genome annotation pipeline²²⁷. To assign gene function, the hit with the lowest E-value obtained after analysis against the UniProt database was chosen. The draft genomes have been deposited in the European Nucleotide Archive under the accession numbers ERS1456453 (AIEC17), ERS1456454 (ECG28), ERS1456455 (AIEC01), ERS1456456 (ECG11), ERS1456457 (AIEC07) and ERS1456458 (ECG04).

3.4 Comparative genomics of strain pairs (gene structure, gene contents and SNPs)

Mauve 2.3²²⁸ was used to identify structural rearrangements and inversions throughout the strain's genome. CRISPRFinder (<http://crispr.i2bc.paris-saclay.fr/>)²²⁹ was used to study CRISPR. To find gene content differences, BLASTP comparison (E-value cutoff 1e-5) and Markov clustering (inflation factor 2.0) were performed by ORTHOVENN²³⁰ using protein sequences. A correction with local BLASTn²²⁴ was performed to recruit genes located at the beginning or end of contigs.

The Harvest suite for rapid-core genome alignment²³¹ was used to detect SNPs between strains of the same pair. To maximise sensitivity, the best annotated AIEC genome at the moment of the analysis, UM146 (NC_017632.1)²⁰⁷, was used as a reference genome. Therefore, only those genes homologous to UM146 were considered for the analysis. For the purpose of this study, SNPs chosen for examination were those that caused non-synonymous amino acid changes in coding regions and were not present in highly variable regions or at the ends of contigs. These SNPs are referred to as “Selected SNPs”.

The Selected SNPs were validated by Sanger resequencing. Using Primer3, primers were designed to flank at least 100 bp up- and downstream of the SNP position to achieve good

sequence quality for assigning nucleotides at the position of the SNP. Care was taken to design primer sets that target the conserved regions of several AIEC and non-AIEC strains. All primers were further analysed with NetPrimer to select the optimal primer pairs. The validated polymorphisms are referred to as “Confirmed SNPs”. The primers and the PCR conditions used are presented in Table 9.

Table 9. Primers and PCR conditions used to amplify fragments of the genes in which the Confirmed SNPs were located.

Gene ID	Primer Forward (5' to 3')	Primer Reverse(5' to 3')	Annealing temperature (°C)
E1-E2_3.4	TCCTCAATGAATCGCAGTCTC*	TCAAAAAGATTGCCCGCTTAC	57
E1-E2_3.6	CTCATCAGCCGGACATACG*	CACCTGTTTTTACATTTTATCTTCTG	56
E1-E2_3.7	GGTAACCCATTTGGCCTTG	CAACACTTCGCTGACAAAACG*	57
E1-E2_5	CGCTATAACGGCGAACTGAT	TCAGTGGTCCGGTATCAAAA*	56
E3-E4_4.2	GCCAGTAACTCTTCGCCATT*	TCAGGACAGCGACAAAAGC	57
E3-E4_4.3	GTTTTCTCCTTTGCCGAACA*	TGATGGTGATAATGCTGCTCA	57
E3-E4_4.4	ATATTCAGCCTGTCCGCAAT	CGCATCATCACTTCCATCTG*	57
E3-E4_4.5	GCGTTGCCTGATGATACTGA*	CGTCGGGGACATCTGACTTA	57
E3-E4_4.7	GGAAGAGCTGGAGACAATGC	CACTACCGCCACTCTCCTGT*	57
E5-E6_3.1	CCCTGTTTGCTGTACTGCTG	CTGCTCACAGGCGTCAAATA*	56
E5-E6_3.12	GAAAAAGTCGCCCATGAGAC*	CGCAACACCAGAGGGTTAAT	57
E5-E6_3.16=3.22	CATCACTCCGGTCAGCAC*	ATTGCAGAAAAGCGAGAGGT	56
E5-E6_3.17	TTTTCACWCGAAGGTCGATG	GATGTGCTGCTGTGCTGYTT*	56

PCR program: 1 cycle at 95°C for 5 min, 30 cycles of 15 sec at 95°C and 45 sec at the primer annealing temperature, finally, one cycle at 72°C during 10 min. All primers were used at 0.2µM; PCR Buffer II at 1x; MgCl₂ at 1.5mM; dNTPs at 200µM and AmpliTaq Gold polymerase 1.25units/reaction. *Indicate the primer used for sequencing.

Of note, we found some SNPs with ambiguous nucleotide peaks (called “SNPs with overlapping peaks”) that represent a mixture of two nucleotides at a given position. The possible cause of these ambiguous nucleotide peaks was analysed *in silico* using a combination of BLASTn²²⁴ gene searches and inspection of reads through Tablet²³². The next step in the selection of SNPs was the identification of strain-specific SNPs; these SNPs were discarded from the analysis. To determine the strain specificity of the Confirmed SNPs, we aligned the sequences containing the variable position in the six strains sequenced in this study as well as in 3 AIEC (UM146, LF82, NRG857c), 9 ExPEC (CFIT073, 536, UMN026, S88, APEC01, 042, EDL933, O157 Sakai, E24377A) and 3 commensal (HS, K12 MG1655, ED1a) strains and further determined the distribution of SNPs amongst the strains. We analysed the distribution within the strain collection of SNPs that displayed variability in the base under study among the strains and that occurred within genes that were widely distributed among the majority of the strains.

Genes harbouring Confirmed SNPs were classified using the Gene Ontology Consortium²³³ and Pfam databases²³⁴.

3.5 Distribution of SNPs among a collection of strains

To analyse the putative application of the Confirmed SNPs as biomarkers for the specific identification of AIEC, a total of 16 SNPs present in 9 genes were screened in a wider strain collection of 22 AIEC and 28 non-AIEC strains⁴³ (Table S5 and Table S6) by Sanger sequencing.

3.6 Algorithm validation in external strain collections

The SNPs included in the algorithm presented (E3-E4_4.4, E5-E6_3.16=3.22(2) and E5-E6_3.12) were further screened by PCR and Sanger sequencing in a larger strain collection. Primers and PCR conditions are indicated in Table 9. Apart from the strains assessed in methods section 3.5, this collection included 60 AIEC and 29 non-AIEC strains mainly isolated from CD patients and controls from distinct geographical origin (Spain (Mallorca)⁷⁴, Chile⁷⁴, France (unpublished) and Australia¹⁷⁸)(Table S7), which were a kind gift from Dr. Roberto Mauricio Vidal, Dr. Nicolas Barnich and Dr. Claire O'Brien. Most of these strains were phenotypically characterised in previous studies. In exception, the adhesion and invasion indices of 25/33 Australian strains were measured in this study according the methodology explained in method section 1.2 and 1.3. In addition, strains causing extraintestinal diseases were also included; these were previously isolated from American patients with meningitis²³⁵, and Spanish patients with sepsis²⁰ or urinary tract infection²³⁶ (Table S7). Phenotypic characterisation of these strains was performed in Martinez-Medina et al.²⁰.

3.7 Statistical analysis

The differences in the distribution of nucleotides present in each polymorphic site between pathotype and phylogroups were calculated using the χ^2 test. The non-parametric Kruskal-Wallis test was used to compare the mean adhesion and invasion indices among more than two nucleotide variants, and the Mann-Whitney U-test was performed to analyse pairwise comparisons. Binary logistic regression was employed as a model to predict AIEC pathotype according to the nucleotide present in a particular SNP position. To establish the usefulness of the algorithm for AIEC identification, the specificity, sensitivity and accuracy values were measured as follows: Sensitivity (%) = (true positives / (true positives + false negatives)) x 100, Specificity (%) = (true negatives / (true negatives + false positives)) x 100; and, Accuracy (%) = ((true positives + true negatives) / (total of cases)) x 100. A p-value ≤ 0.05 was considered statistically significant in all cases.

Chapter 2.2: Construction of isogenic mutants to study the role in pathogenicity of three genes related to AIEC pathotype

4.1 Bacterial strains

We selected one strain from our collection for each gene that presented the nucleotide associated with AIEC pathotype and that exhibited high invasiveness (>0.482 %). For E3E4_4.3 gene deletion the AIEC23 strain (presenting thymidine in the SNP position) was chosen, for E3E4_4.4 the AIEC25 strain (presenting adenine) and for E5E6_3.16 the LF82 strain (having guanine and adenine in the second and third SNP position) (Table 10). In addition, one non-AIEC strain from the same phylogroup as each of the AIEC strain selected but with different nucleotide in a particular gene was included.

Table 10. Characteristics of the strains selected for the construction of isogenic mutants (AIEC) and the candidate non-AIEC (ECG) strains with which isogenic mutants would be compared. Nucleotide in bold indicate the gene targeted to be deleted in this particular strain.

Strain	Phylogroup	Adhesion ^a	Invasion ^b	Replication in J774 ^c	Replication in THP-1 ^d	SNP 4.3 ^e	SNP 4.4 ^e	SNP 3.16 ^e
AIEC23	A	9.73	0.568	2362	160	T	R	G
K-12	A	0.7	0.019	11	10	C	-	-
AIEC25	B2	2.77	0.482	776	142	T	A	-
ECG08/ECG49	B2/B2	0.3/0.3	0.004/0.008	49/78	25/54	C	G	C
LF82	B2	25.66	2.261	777	195	T	-	G
ECG08	B2	0.3	0.004	49	25	C	G	C

^aNumber of bacteria per 1407 cell. ^bPercentage of intracellular bacteria after 1h gentamicin treatment. ^cPercentage of intracellular bacteria present at 24h post-infection relative to the number of intracellular bacteria present after 1h of gentamicin treatment. ^dPercentage of intracellular bacteria present at 16h post-infection relative to the number of intracellular bacteria present after 1h of gentamicin treatment. ^eNucleotides present in each SNP position are indicated. T: Thymidine. R: Adenine or Guanine. G: Guanine. -: no gene.

4.2 Plasmid transformation in *E.coli*

The construction of isogenic mutants of the three different genes was performed following the red recombinase system described by Datsenko et al.²³⁷ and Chaverroche et al.²³⁸. The *E.coli* DH5 α strain was used for the propagation of the plasmid pkD46. This was isolated with Nucleospin plasmid DNA purification (Macherey-Nagel) and transformed to the strains of study. LF82, AIEC23 and AIEC25 strains were cultured at 37°C with aeration and agitation in an Erlenmeyer with 10 mL of SOC medium and 200 μ L of MgCl₂ (1M) with a starting DO620 equal to 0.1 until the DO620 was 0.4-0.6. A serial washes with 8 mL of cold distillate H₂O, 8 mL cold 10% glycerol and 1 mL cold 10% glycerol were performed. Between each step 5 min of centrifugation at 4°C at 7000 rcf was executed. Finally the pellet was resuspended in 100 μ L of cold 10% glycerol and prepared for

electroporation (2.5kV) with 5 μ L of pkD46 plasmid. Two hours incubation at 30°C with agitation was followed previous to spreading in LB plates with gentamicin (100 μ g/mL). Colonies grown after overnight incubation at 30°C were considered gentamicin resistance and hence, that acquired the pkD46 plasmid. This plasmid encodes for the red proteins under the control of a promoter inducible by L-arabinose. This plasmid was suicided after the PCR product was electroporated at 42°C.

4.3 Construction of PCR product for gene disruption

In order to replace the gene of interest for an antibiotic resistance gene, a PCR fragment was created by two strategies (Figure 5). The first approach consisted in performing one PCR using Platinum Taq high fidelity DNA polymerase (Invitrogen, USA) in order to create a PCR product with the kanamycin gene plus a 50bp of homologous to the adjacent region of the gene selected to delete. This PCR was performed using the primers indicated in the PCR2 of the Table 2 and the plasmid pkD4 as template. Then PCR product was purified before electroporation with NucleoSpin Gel and PCR Clean-up (Macherey-Nagel, Germany).

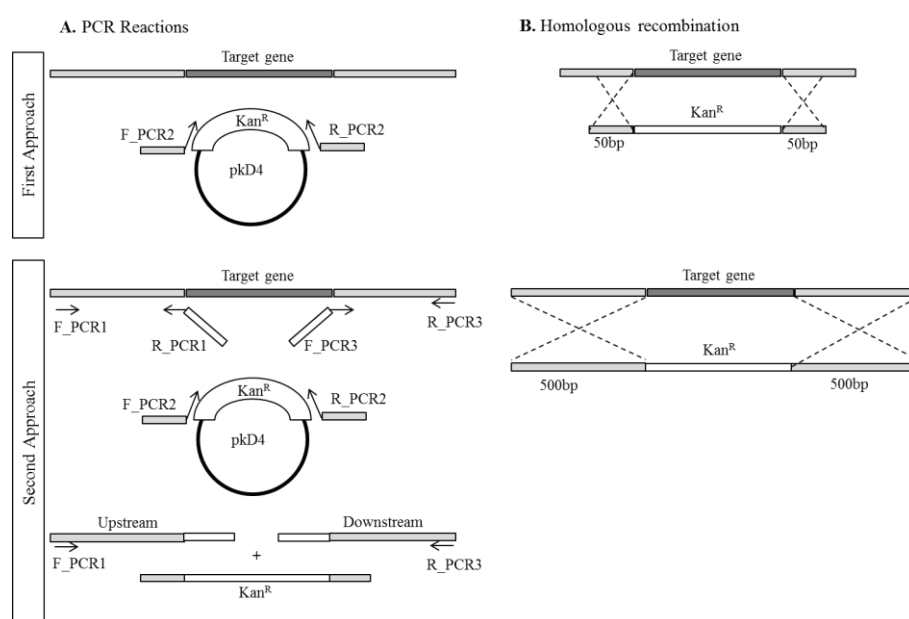


Figure 5. Principle of the first (50 bp homology) and second (Three-step PCR, 500 bp homology) approaches used to construct isogenic mutants. A: Structure of the PCR reactions required. **B.** Result of lineal PCR product and homologous recombination site.

The second approach consisted on the electroporation of a PCR fragment that presented the kanamycin cassette flanked by 500 bp homologous to the adjacent regions of the gene to delete in each extreme. This PCR product is created by a three-step PCR using Platinum

Taq high fidelity DNA polymerase (Invitrogen). The first step consisted of amplifying independently the upstream (PCR1) and downstream (PCR3) regions of the target gene and the kanamycin cassette, using the primers indicated in Table 2. In each pair (PCR1 and PCR3), one primer contained a 19 bp region homologous to the extremities of the kan gene at its 5' end which will facilitate the amplification of the PCR1 and PCR3 product together with PCR2 which harbours the kan gene. Therefore, by mixing PCR products of PCR1, PCR2 and PCR3 in PCR 4 (Table 11), a complete linear DNA was obtained. Finally, a third PCR using the primers PCR4 again (Table 11) was performed in order to yield higher PCR product quantity. All PCR products were purified with Nucleospin PCR and gel clean-up (Macherey-Nagel, Germany) to avoid the amplification of incomplete or erroneous products before proceeding to the following step.

Table 11. Primers used for the construction of isoagenic mutants. Nucleotides in underlined indicate the 5' end of the kanamycin gene and these in bold indicate the 3' gene end.

	Primer Forward (5' to 3')	Primer Reverse (5' to 3')	PCR product length
4.3 (AIEC23)	PCR1* MI_4.3gene_PCR1_F2 GGAAGAGCTGGAGACAATGC	MI_4.3gene_PCR1_R2 <u>CGAAGCAGCTCCAGCCTACGTTCCG</u> CAAAGGAGAAAACCTGGTTGCCACC	585 bp
	PCR2* MI_4.3gene_F GTGCCGGACCATCTGATAGTCGGTG GCAACCAGGTTTTCTCCTTTGCCGAAC <u>GTAGGCTGGAGCTGCTTCG</u>	MI_4.3gene_R CGGTGCAGGGCTGCCCCACCAGG GGCGGTCGTGATTGTCTGTCGGGAA GTGTCATATGAATATCCTCCTTAG	1579 bp
	PCR3* MI_4.3gene_PCR3_F2 CTAAGGAGGATATTCATATGACACTT CCCAGACGAATCACGACCCGCC	MI_4.3gene_PCR3_R2 AAATGCCTGCCTCAATATGC	605 bp
	PCR4* MI_4.3gene_PCR1_F2 GGAAGAGCTGGAGACAATGC	MI_4.3gene_PCR3_R2 AAATGCCTGCCTCAATATGC	2769 bp
	Verification# Verif_D4.3gene_F GAAGTGATTAACCGCGCCCT	Verif_D4.3gene_R GTAATGCAACCGGTTACCCTC	Mutant = 1833 bp WT = 696 bp
4.4 (AIEC25)	PCR1* MI_4.4gene_PCR1_F2 GCACTCATGACAGTGCTTCC	MI_4.4gene_PCR1_R2 <u>CGAAGCAGCTCCAGCCTACTATTTA</u> ACCTTCTGAGACGCATTTTCATGA	457 bp
	PCR2* MI_4.4gene_F AAATGACCTCCTCCGTGGTTGTCATG AAAATGCGTCTCAGGAAGGTTAAATA <u>GTAGGCTGGAGCTGCTTCG</u>	MI_4.4gene_R GACTGAATACITTTGCATCAGCCCCT GATGGCGTAACGACAGGTATTCACT GACATATGAATATCCTCCTTAG	1579 bp
	PCR3* MI_4.4gene_PCR3_F2 CTAAGGAGGATATTCATATGATCAGT GAATACCTGTCGTTACGCCATCAGG	MI_4.4gene_PCR3_R2 TTCCCCATAAAACCACTCTGC	714 bp
	PCR4* MI_4.4gene_PCR1_F2 GCACTCATGACAGTGCTTCC	MI_4.4gene_PCR3_R2 TTCCCCATAAAACCACTCTGC	2750 bp
	Verification# Verif_D4.4gene_F GTTGCACGCACGGTTCTG	Verif_D4.4gene_R CTGGCATCGATCTCCTCCAT	Mutant = 1933 bp WT = 1752 bp
3.16 (LF82)	PCR2* MI_3.16gene_F TAAGGGCACCAGAAATGGTGCCTTTT TTATTGCAGAAAAGCGAGAGGTAATT <u>GTAGGCTGGAGCTGCTTCG</u>	MI_3.16gene_R ACGTTTTGCTTTTCAGCTGGATTGTG CAGTTCGTACCGGTTTTCTGTGCC GCATATGAATATCCTCCTTAG	1579 bp
	Verification# Verif_D3.16gene_F CCATTCCCGGTAGCTACAGT	Verif_D3.16gene_R CATCCATGCTGTAACGTCCG	Mutant = 1833 bp WT = 1054 bp

The primers used in this study were designed using the Primer 3 software taking as a template the sequence of each gene in its corresponding strain. *PCR1-4 were carried out in 1x high fidelity PCR buffer, 2 mM MgSO₄, 0.2 mM dNTP mix, 0.2 μM Forward primer, 0.2 μM Reverse primer, 1 U/reaction of Platinum Taq DNA polymerase High Fidelity and 1 μL of DNA Template with a final volume of 50 μL. The PCR program for PCR product synthesis was: 1x cycle at 94°C for 2 min, 30x cycles at 94°C for 15 sec, at 56°C for 30 sec and at 68°C for 2 min; and 1 cycle at 72°C for 10 min. #The verification reactions were carried with 1x FIREPol MasterMix with 12.5 mM MgCl₂, 0.25 μM of Forward primer, 0.25 μM of Reverse primer and 2 μL of DNA Template in a final volume of 20 μL. The PCR verification program was: 1 cycle at 95°C for 1 min, 30 cycles at 95°C for 30 sec, at 56°C for 40 sec and at 72°C for 2 min, and 1 cycle at 72°C for 10 min.

4.4 Electrocompetent cells and gene disruption

Strains harbouring pKD46 plasmid were grown at 30°C with agitation in an Erlenmeyer with 20 mL of SOC, 200 µL of gentamicin (10 mg/mL) and 200 µL of L-arabinose (1M) at an initial DO620 of 0.1 until DO620 reached 0.4-0.6. L-arabinose will induce the expression of the red recombinase genes encoded in plasmid pKD46, which will promote chromosomal recombination and protect lineal DNA from degradation in bacteria. The following wash steps consisted on 10 mL cold H₂O distillate, 8 mL of cold 10% glycerol solution and 800 mL of cold 10% glycerol solution. Between each step the bacteria was centrifuged for 10 min at 5500 rpm at 4°C. Finally, bacteria pellet was resuspended in 200 µL of cold 10% glycerol solution (for 4 transformations) and frozen at -80°C for at least 30 min before usage. Electroporation was carried out using 50 µL of electrocompetent cells plus 5 µL of the linear DNA (60-170 ng/µL) desired at 2.50kW. Afterwards, bacteria were left to grow at 30°C o/n with agitation and were spread in LB plates with kanamycin (50 µg/mL). Electroporations were conducted with 100-400 ng of PCR product at 2.50 kW with time constant oscillating from 3.60 to 5.40 ms. To favour gene disruption by homologous recombination, after electroporation, bacteria were left to grow at 30°C o/n with agitation and then, they were spread in LB plates with kanamycin (50 µg/mL) and left in 37°C o/n.

4.5 Phenotypic characterisation

The ability of the LF82Δ3.16 mutant to adhere to and to invade intestinal epithelial cells using I407 cells (ATCC CCL-6), as well as, to survive and replicate inside macrophages in J774 and THP-1 cell lines (ATCC TIB-67 and ATCC TIB-202, respectively) was analysed. All assays were performed as previously described (methods section 1.2, 1.3 and 3.1) on 24-well plates in triplicate.

Chapter 3.1: RNA-Seq analysis of the transcriptome during growth in cell culture media and during intestinal epithelial cell infection of AIEC in comparison with non-AIEC strains

5.1 Bacterial strains, cell line and growth conditions

Two AIEC strains and their corresponding non-AIEC pairs (identical pulsotype and phylogroup) isolated in a previous study⁴³ were used (two pairs of the ones studied in methods section 3.1). The AIEC07/ECG04 pair was isolated from the ileum of a control patient and belongs to the B1 phylogroup. The AIEC17/ECG28 pair was obtained from the colon of a CD patient and it is from the D phylogroup. The intestine-407 epithelial cell line (I-407; ATCC CCL-6) was maintained in an atmosphere containing 5% CO₂ at 37°C with a culture EMEM medium (Lonza, Switzerland) supplemented with 10% FBS (Gibco, USA), 1% minimum essential medium vitamins (Gibco, USA), 1% glutamine (Gibco, USA), 1% antibiotic-antimycotic (Gibco, USA) and 1% minimum essential medium non-essential amino acids (Gibco, USA).

5.2 Infection of intestinal epithelial cells

I-407 cells were seeded to 2×10^7 total cells in a T75 flask and incubated for 20 hours. Infection of cells was performed with bacteria at exponential growth (optical density at 620nm = 0.625) at MOI 100 using the cell culture medium composed by EMEM (Lonza, Switzerland) and 10% of heat-deactivated FBS (Gibco, USA). After 4 hours of incubation at 37°C 5% CO₂, the supernatant (SN fraction) was recovered and kept on ice while the infected eukaryotic cells (INV fraction) were washed with the same medium twice and collected with a scrapper. Both fractions were centrifuged at 3000 x g during 10 min at 4°C. The pellet was washed with 500 µL of PBS prior to RNA extraction (Lonza, Switzerland). Before obtaining this final protocol, the procedure was tested by modifying several variables (for more information see results section 5.1): initial cell count (1×10^8 total cells in a T182 flask), the MOI (10), the starting material of the extraction protocol (5, 50 or 100% of a T75 or T182 flask) and/or the step following sample centrifugation (no pellet wash step was performed).

5.3 RNA extraction and purification

Each section (SN and INV) were treated equally. First, an extraction of total RNA with TRIzol Max Bacterial Isolation kit with Max Bacterial Enhancement reagent (Invitrogen,

USA) followed by a DNase I treatment (Thermo Scientific, USA) was carried out. Previous to protocol optimization, the kit used to extract RNA was the RiboPure-Bacteria Ambion RNA (Thermo Scientific, USA). Procedures were conducted following manufacturer's instructions, in exception of TRIzol where, after the addition of chloroform, incubation at room temperature for 15 min was done and after isopropanol addition an overnight precipitation at -20°C was conducted. Secondly, 25 µg of total RNA in a maximum of 30 µL was used as starting material for the MICROBEnrich kit (Thermo Scientific, USA). This removes by Oligo MagBeads the mRNA containing polyA tail, 18S and 28S eukaryote ribosomal RNA. The next kit, Ribo-Zero Magnetic Gold kit (epidemiology) (Illumina, USA), was performed with 2.5 µg of RNA to remove all rRNA molecules, tRNAs and mitochondrial RNAs, ending up having the prokaryotic messenger RNA. RNA quality was tested after each kit using denaturing agarose gel or Total Prokaryotic RNA program of the Agilent RNA 6000 Nano Chip Bioanalyzer. Most of the samples were quantified with Nanodrop spectrophotometer and Qubit Fluorometer with RNA HS Assay kit (Thermo Scientific, USA).

5.4 Sequencing and RNA-seq analysis

The TruSeq Stranded mRNA method (Illumina, USA) was applied for the cDNA synthesis. All samples were sequenced by Illumina Miseq but the sequencing depth varied according to the fraction: 10M reads for samples mainly composed by bacteria (SN) in duplicate and 200M reads for samples from infected cells (INV).

Sequence reads were analysed by FastQC²³⁹, trimmed accordingly and mapped to the UM146 AIEC (NC_017632) reference genome with the TopHat²⁴⁰. Transcripts were assembled using Cufflinks²⁴¹, merged and compared to UM146 by Cuffmerge²⁴¹. Finally, normalization and differential expression analysis were done using Cuffdiff²⁴¹. Gene expression levels were presented as fragments per kilobase of transcript per million mapped reads (FPKM). Four comparisons were performed, in which AIEC transcripts were compared with its non-AIEC transcripts for each condition, and each pair. Those genes with $p \leq 0.05$ were considered differentially expressed genes (DEGs). FPKM values for genes previously associated with AIEC invasion (*fis*, *lpfA*, *fucO*, *fucA*, *fimH*, *ompA* and *ompC* genes)^{18,154,159,160,165} were searched in the Cuffdiff output documents.

5.5 Gene expression validation

Gene expression levels of a subset of selected genes (all overexpressed genes in AIEC in all comparisons; XLOC_003172 and XLOC_000601 genes have not been studied because no FPKM value was obtained for the non-AIEC strain) were validated in the samples sequenced by RNA-seq using the 48x48 microfluidic array IFC chip on BioMark™ system (Fluidigm, USA). In all cases, 1.5µg of mRNA processed with the MICROBEnrich kit were reverse transcribed using the High-Capacity cDNA Reverse Transcription Kit (Applied Biosystems, USA) and the RNase Inhibitor (Life Technologies, USA) in a final volume reaction of 20 µL. Samples corresponding to the SN condition were equally mixed before cDNA synthesis. The Fluidigm loading kit-10 chip package (BMK-M10-48.48) was used following manufacturer's instructions and as recommended. Prior to amplification, a cycle of 10 min at 95°C followed by 16 cycles of 15 sec at 95°C and 4 min at 60°C was performed. Exceptionally, XLOC_000912 and XLOC_000511 samples were assessed by a 7500 Real Time PCR system (Applied Biosystems, Foster City, CA, USA). The amplification reactions were carried out in a total volume of 20 µL containing: 10 µL of Power SYBR™ Green PCR Master Mix 2x (Applied Biosystems, USA), 300 nM for each primer and RNase free water up to the final volume. Thermal cycling conditions consisted of an initial step at 50°C during 30min, a PCR activation step at 95°C for 10min to denature DNA and activate Ampli-Taq Gold polymerase, and a further denaturation step of 40 cycles (95°C for 15 sec) followed by an annealing and extension step at 60°C for 1 min. Data was collected and analysed with the 7500 SDS system software version 1.4 (Applied Biosystems, Foster City, CA, USA). Throughout the manuscript, we refer as RT-qPCR validation both RT-qPCR and Fluidigm analysis.

For each primer pair standard curves with a five-fold dilutions series (1/4, 1/20, 1/100, 1/500, 1/2500) of template was used to determine amplification efficiency with the equation $E=10^{(-1/\text{slope})}$. As a reference the respective AIEC mRNA was used. To normalise data, *gapdh* was selected as housekeeping gene. The mRNA abundances for each candidate gene were calculated as: Relative Transcript Abundance RTA = $E^{\Delta C_t(\text{control-sample})}$ (Target gene) / $E^{\Delta C_t(\text{control-sample})}$ (Reference gene)²²². The AIEC strain from the pair where the gene has been found differentially distributed has been used as the control sample. All experiments were carried out in triplicates. In addition, 16S rRNA copy number was evaluated as explained in methods section 2.5 and used to regularise RTA value according to the bacterial RNA quantified. Then, the ratio (fold-change) of (RTA/16S

sample)/(RTA/16S control) was performed and values were expressed as the log₂ fold-change. Again, the control sample corresponded to the AIEC strain of the pair. A negative value indicates over-expression of the gene in the AIEC sample and a positive value refers to AIEC under-expressed genes.

Primer3 0.4.0 software²⁴² was applied for the gene-specific primers design, the parameters that were taken into account are: primer size (15-30 nts), GC content (30-80%), amplicon size (<150 base pairs) and primer melting temperature (50-60°C). Secondary structures were assessed with NetPrimer (PREMIER Biosoft International, Palo Alto, CA). Finally, *in silico* PCR amplification was performed to test the specificity of the primers in three data bases (bacteria: <http://insilico.ehu.es/PCR/>; human: UCSC *In-Silico* PCR and BIOTECH *In Silico* PCR). Additionally, a BLASTN search to check its specificity with *E. coli* was performed. See primer characteristics in Table S8.

5.6 Statistical analysis

Lineal logistic regression was performed to compare RT-qPCR and RNA-Seq log₂ values. Pearson or Spearman's correlations were used for parametric and non-parametric data respectively.

● RESULTS & DISCUSSION ●

Chapter 1.1: Virulence gene carriage and adhesin variants of AIEC and commensals isolated from humans and animals

Results

1.1 Virulence gene repertoires

Prevalence of 54 VGs were assessed in a collection of *E. coli* strains (N=104) according to host origin (animal or human), pathotype (AIEC and non-AIEC) and group of subjects (CD and controls).

1.1.1 Animal vs Human *E. coli* strains

Regarding host origin, 19 out of the 54 studied VGs presented differential distribution between strains isolated from animals and humans (Figure 6A, Figure S2A and Table S9). Twelve genes (*focG*, *bra*, *papGII/III*, *sfa/foc*, *ireA*, *iroN*, *cnf*, *hlyA*, *malX*, *pic*, *pks*, and *eal*) were more frequent in animal-isolated strains (29-84%) than in human-isolated strains (7-42%) ($p \leq 0.025$), and 7 genes (*traT*, *iba*, *papGII*, *iucD*, *iutA*, *neuC*, and *sat*) were more prevalent in strains isolated from humans (present in 15-68% of total human strains) than in those from animals (0-40%) ($p \leq 0.010$). Considering only those strains of the AIEC pathotype, 17 VGs were still associated with origin of isolation. Of those, 11 were more frequent in strains isolated from animals and 6 in human-strains ($p \leq 0.046$) (Figure 6B and Table S9). In non-AIEC strains the prevalence of VGs was more similar when analyzing data by origin of isolation. In this case, 10 out of 54 genes were differentially distributed; six were overrepresented in animal strains and four in human strains ($p \leq 0.038$) (Figure 6C and Table S9).

The distribution of virulence-associated genes was examined according to phylogroup. Considering the whole collection of strains, 55.6% (30 genes) of the VGs studied was associated with the phylogenetic origin of the strains (Table S10). Most of the studied genes (29/30) were mainly related with B2 and/or D phylogroups, except for *csgA* gene, which was more frequent in

A and B1 phylogroups. Of note, 17 of the 19 genes associated with either human or animal hosts were differentially distributed depending on the phylogenetic origin (Table S10).

Considering that the distribution of phylogroups was different between animal and human strains ($p < 0.001$) (Table S3), we selected the most abundant phylogroup (B2) to perform the comparisons, in order to avoid differences due to phylogenetic origin. Interestingly, genes previously associated with origin of isolation in the whole collection maintained its significance after selecting B2 strains only (Table S9). Concerning AIEC and non-AIEC strains, 15/17 and 5/10 VGs respectively were still differentially distributed according to origin of isolation when only B2 phylogroup strains were analysed (Table S9).

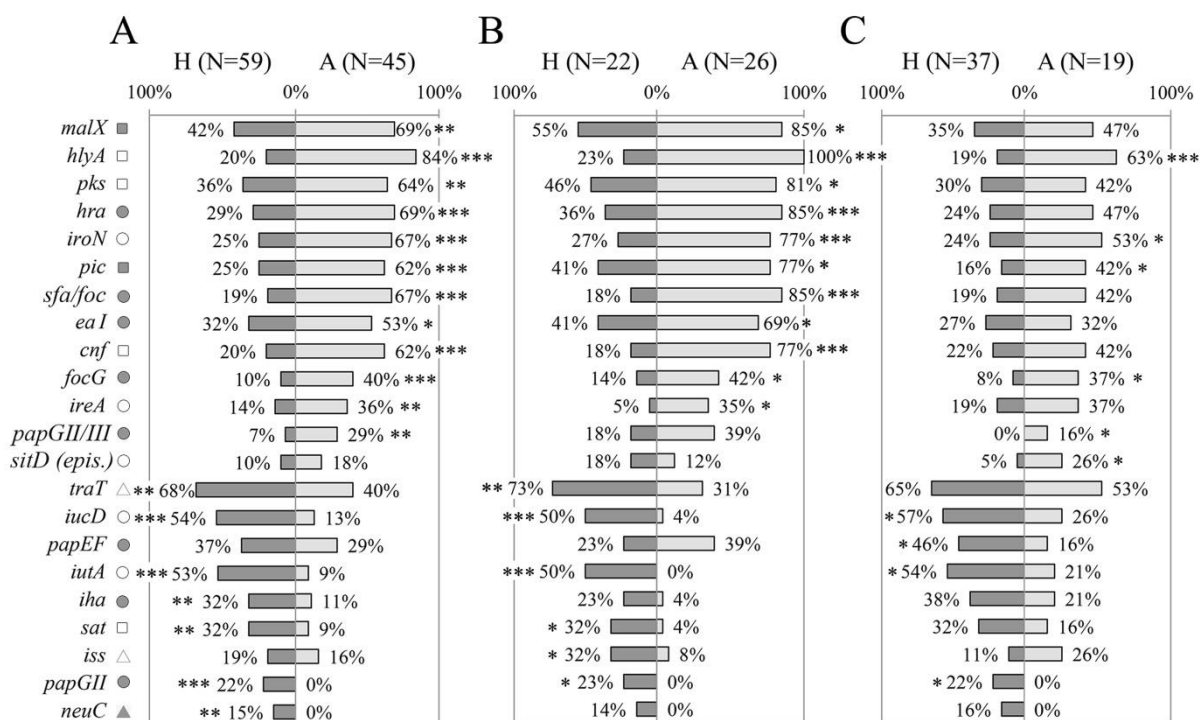


Figure 6. Distribution of virulence genes prevalence according to origin of isolation (H: strains isolated from humans. A: strains isolated from animals.). A: All *E. coli* strains. B: Only AIEC strains. C: Only non-AIEC strains. Numbers indicate gene prevalence in percentage in relation to the total of strains from each origin. Genes presenting statistically significant differences are depicted. Symbols indicate gene role in: adhesion (●), capsule formation (▲), invasion (■), iron scavenging (○), resistance (△) and toxin (□). * $p \leq 0.05$ ** $p \leq 0.01$ *** $p \leq 0.001$.

1.1.2 AIEC vs non-AIEC strains

The whole collection (N=104) was evaluated to determine whether the prevalence of VGs was different between AIEC and non-AIEC strains. AIEC strains reported significantly higher prevalence than non-AIEC strains ($p \leq 0.034$) in: four genes related to adhesion capacity (*bra*, *papGII/III*, *sfa/foc*, and *eaI*), four genes coding for toxins (*cnf*, *vat*, *blyA*, and *pks*), four genes linked

with iron processes (*fyuA*, *iroN*, *irp2*, and *sitD* (*chr.*)), and three genes related to other functions (*kpsMTII* (capsule formation); *malX* (metabolic processes) and; *pic* (invasiveness)) (Figure 7A, Figure S2B and Table S11). In contrast, three genes involved in strain adhesiveness (*csgA*, *iha*, and *sfaS*) and two in iron processes (*iucD* and *iutA*) were more frequent in non-AIEC than in AIEC strains ($p \leq 0.026$). Furthermore, higher adhesion for strains harboring *irp2*, *sitD* (*chr.*), *kpsMTII*, *vat* or *pic* (6.54 ± 7.75 ; 7.61 ± 7.84 ; 7.84 ± 8.63 ; 7.74 ± 8.11 and 8.32 ± 6.91 bacteria/cell VG-positive strains, respectively) was achieved in comparison with those that do not (2.64 ± 5.66 ; 3.13 ± 6.22 ; 3.08 ± 4.97 ; 3.69 ± 6.36 and 4.38 ± 7.39 bacteria/cell VG-negative strains, respectively) ($p \leq 0.046$). In terms of invasion, *vat*-positive strains presented higher invasion values ($0.31 \pm 0.53\%$) than *vat*-negative strains ($0.12 \pm 0.21\%$) ($p = 0.048$). Additionally, the number of VGs present in each strain's genotype was assessed according to pathotype. AIEC strains had from 4 to 30 VGs and non-AIEC carriage ranged from 4 to 33 VGs but, on average, AIEC strains tend to carry more VGs (18 ± 7 total number of genes) than non-AIEC strains (15 ± 8 total number of genes) ($p = 0.052$). No significant differences were achieved probably due to high variation in the number of VGs carried between isolates.

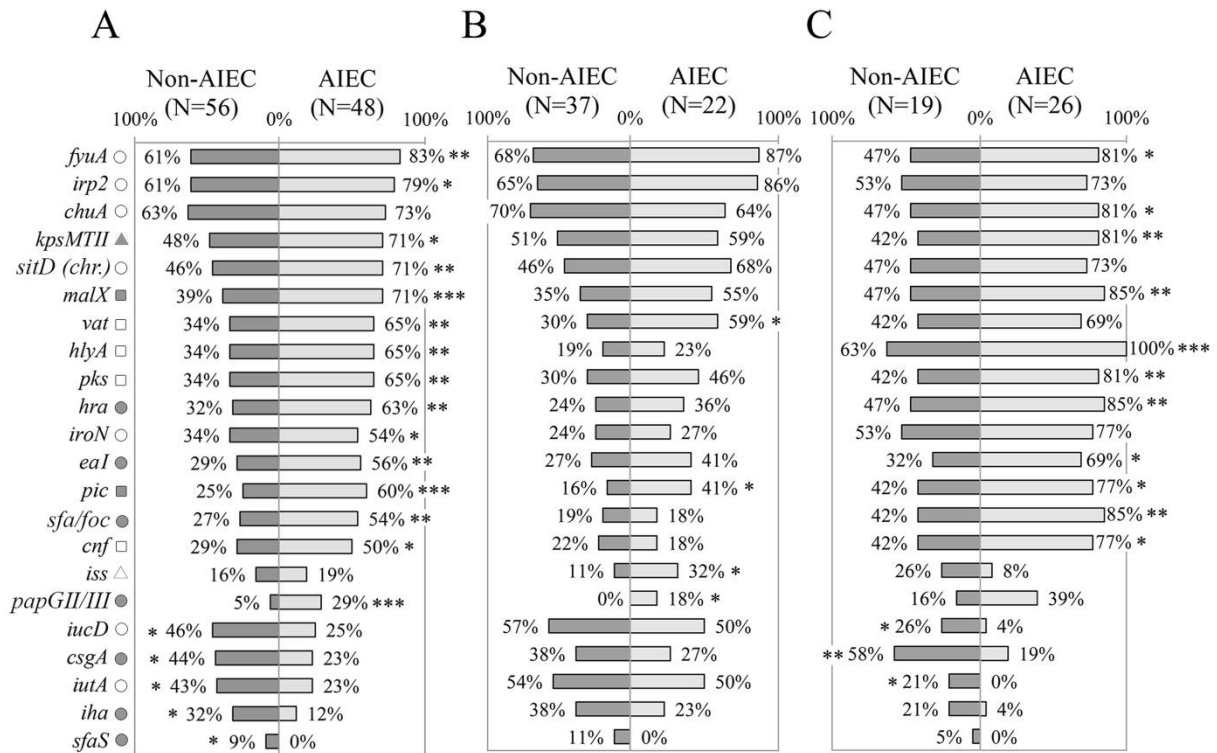


Figure 7. Distribution of virulence genes prevalence according to pathotype. A: All *E. coli* strains. B: Only animal-isolated strains. C: Only human-isolated strains. Numbers indicate gene prevalence in percentage in relation to the total of strains from each pathotype. Genes presenting statistically significant differences are depicted. Symbols indicate gene role in: adhesion (●), capsule formation (▲), invasion (■), iron scavenging (○), resistance (△) and toxin (□). * $p \leq 0.05$ ** $p \leq 0.01$ *** $p \leq 0.001$.

All the genes reported to be differentially represented according to pathotype were also associated with phylogroup, with the exception of *sfaS* (Table S10). Differential phylogroup distribution was reported between AIEC and non-AIEC strains studied in this section ($p=0.002$), as non-AIEC strains were more predominant in A, B1 and D phylogroup while AIEC mainly constituted the B2 (Table S3). Therefore, to prevent phylogroup as confounding factor, the analyses were performed only with B2 strains. Apart from three genes (*papGII/III*, *sfaS* and *pic*) that maintained its differential distribution between AIEC and non-AIEC strains, the others did not associate with pathotype (Table S11).

To unveil possible differences in gene prevalence due to isolation origin, we further evaluated the 54 VGs in each group of strains (45 from animals and 59 from humans) (Figure 7B-C and Table S11). Indeed, 13 out of the 20 genes found significant when analyzing all the strain collection, maintained the significance in strains isolated from animals but not in human strains. Only *pic* gene was more prevalent in AIEC strains irrespectively of strains' host.

Among animal strains, *csgA*, *iucD*, and *iutA* were more prevalent in non-AIEC (21.10-57.90%) than in AIEC strains (0-19.20%) ($p\leq 0.040$) while the remaining genes (*bra*, *sfa/foc*, *fyuA*, *kpsMTII*, *cnf*, *blyA*, *malX*, *pks*, *pic* and *eal*), plus an additional one (*chuA*) presented higher prevalence in AIEC (69.20-100%) than in non-AIEC strains (31.60-63.20%) ($p\leq 0.021$) (Figure 7B and Table S11). Additionally, phenotypic traits supported the difference in *bra* and *blyA* prevalence between AIEC/non-AIEC strains. In this case, higher adhesion and invasion indices were obtained for those strains harboring *bra* (2.84 ± 3.86 bacteria/cell and 1.14 ± 2.15 % for *bra*-positive strains; 2.28 ± 4.32 bacteria/cell and 0.66 ± 1.24 % for *bra*-negative strains; $p=0.050$ and $p=0.038$, respectively) or *blyA* (3.09 ± 4.18 bacteria/cell and 1.15 ± 2.05 % for *blyA*-positive strains; 0.34 ± 0.52 bacteria/cell and 0.13 ± 0.14 % for *blyA*-negative strains; $p=0.006$ and $p=0.004$, respectively). Among the animal strains, the phylogroup origin of AIEC and non-AIEC was different, being 72.4% of AIEC strains from B2 phylogroup while 68.8% of non-AIEC strains were from A phylogroup ($p=0.009$) (Table S3). If only B2 strains were selected, none of the 14 genes mentioned above (*csgA*, *iucD*, *iutA*, *bra*, *sfa/foc*, *fyuA*, *kpsMTII*, *cnf*, *blyA*, *malX*, *pks*, *pic*, *eal* and *chuA*) were found differentially distributed between pathogenic and commensal strains (Table S11).

Regarding human-isolated strains, apart from *pic* gene, three additional genes (*papGII/III*, *iss*, and *vat*) reported significantly higher prevalence in AIEC (18.20-59.10%) than in non-AIEC strains (0-29.70%) ($p<0.05$) (Figure 7C and Table S11). In addition, higher adhesion was reported for

strains harboring *pic* (8.32 ± 6.91 bacteria/cell *pic*-positive strains and 4.38 ± 7.39 bacteria/cell *pic*-negative strains; $p=0.034$) and higher adhesion ($p=0.043$) and invasion ($p=0.048$) values were obtained for *vat*-positive strains (7.74 ± 8.11 bacteria/cell and 0.31 ± 0.53 %) in comparison with *vat*-negative strains (3.69 ± 6.36 bacteria/cell and 0.12 ± 0.21 %). In this group of strains, similar phylogenetic distribution between AIEC and non-AIEC strains was observed ($p=0.072$) (Table S3). Nevertheless, 46.3% of the genes studied reported different prevalence regarding the phylogenetic origin (Table S10). Thereby for following analyses only B2 strains were selected (Table S11). In this case, none of the four genes (*papGII/III*, *iss*, *vat* and *pic*) associated with pathotype presented statistical differences, although a trend was noticeable for three of the cases (*papGII/III*, *pic* and *iss*) where the gene was more frequent in AIEC strains (*papGII/III*: 0% non-AIEC and 27% AIEC $p=0.057$; *pic*: 29% non-AIEC and 60% AIEC $p=0.092$; *iss*: 14% non-AIEC and 40% AIEC $p=0.129$; and *vat*: 71% non-AIEC and 80% AIEC $p=0.458$). In addition, the *astA* gene resulted to be more prevalent in non-AIEC (29%) than AIEC (0%) strains ($p=0.042$).

Of note, genes previously found to be more frequent in AIEC than in non-AIEC strains from human (*hpfA₁₅₄* and *cbuA*)^{74,154} reported similar percentage of PCR-positive AIEC/non-AIEC strains.

1.1.3 Crohn's disease vs Controls

Differences in the VGs carriage were also reported between CD and controls. In this case, four genes related with iron processes (*iba*, *iroN*, *iucD* and *intA*) were more frequent in strains isolated from controls than patients with CD (*iba*: 21% CD and 48% controls, $p=0.026$; *iroN*: 15% CD and 40% controls, $p=0.029$; *iucD*: 41% CD and 72% controls, $p=0.018$; *intA*: 41% CD and 68% controls, $p=0.037$). On the other hand, a gene encoding for a meningitis-associated fimbria (*mat*) was more prevalent in CD-isolated strains (100%) than controls (84%) ($p=0.028$). In addition, similar number of VGs was reported for CD strains (15 ± 7) in comparison to controls strains (17 ± 8) ($p=0.891$). Similarly occurred when comparing VGs carriage in AIEC and non-AIEC strains according to disease origin (CD–AIEC 17.3 ± 6.3 vs controls–AIEC 14.7 ± 8.9 , $p=0.169$; CD–non-AIEC 13.1 ± 7.3 vs controls–non-AIEC 17.2 ± 7.2 , $p=0.142$).

1.2 FimH and ChiA amino acid substitutions

Since it has been suggested that differences regarding phenotype may rely on variations in the protein sequence, in this study alterations in FimH and ChiA have been explored. For this analysis, strains isolated from controls, CD, UC and CRC were considered (N= 58; 31 AIEC and 27 non-AIEC strains).

Fifty-four strains presented the *fimH* gene, representing 93.9% of AIEC and 92.6% of non-AIEC strains. As shown in Figure 8A and Table S12, a total of 19 FimH amino acid substitutions were found among the strain collection which grouped the strains in 21 variants. There was no variant comprising uniquely or mainly AIEC strains. When comparing the sequence of AIEC/non-AIEC strains globally, both groups of strains presented on average two substitutions throughout the FimH sequence (AIEC: 2 ± 1 ; non-AIEC: 2 ± 1) ($p=0.915$). Individually, none of amino acid substitutions associated with the disease of isolation neither with AIEC phenotype. Only N70S and S78N were related to phylogenetic origin, as they were only found in controls or CD-isolated strains from the B2 (69.2% and 73.1% of strains respectively) and D (25.0% and 25.0% of strains respectively) phylogroup ($p<0.001$) (Table 12). Despite that, while no divergence was found for the adhesion capacity, significant difference in terms of invasion index was achieved depending on the amino acid present in the 119 position. In this case, strains presenting the amino acid A (equal to K-12) had lower invasion values ($0.223\pm 0.402\%$ of intracellular bacteria/inoculum; N=47) in comparison to strains with V ($0.401\pm 0.477\%$ of intracellular bacteria/inoculum; N=7) ($p=0.048$).

Regarding *chiA* gene, 86.2% AIEC strains (N=31) and 63% non-AIEC strains (N=27) presented this gene ($p=0.044$). Twenty-four variable amino acid positions were found, assembling the strains in a total of 16 variants (Figure 8B and Table S13). Again, similar protein sequence variants were reported among strains isolated from diverse groups of subjects (Figure 8). None of the mutations identified were associated with AIEC, neither the five mutations previously described (K362Q, K370E, A378V, E388V, V548E)⁸⁵ (Table 12). However, a subcluster of strains with *chiA* sequence identical to LF82 included a higher proportion of AIEC strains (85%) than non-AIEC strains (15%) ($p=0.027$). Nevertheless, this variant represented only the 35.5% of all AIEC strains and the 7.4% of total non-AIEC strains. Besides, the number of variable positions differed among pathotypes, being slightly higher for AIEC strains (10 ± 5) than in non-AIEC strains (8 ± 6) ($p=0.038$). Of note, most of the strains harboring an amino acid different

from K-12 strain were from the B2 phylogroup, being V335G amino acid change an exception as it was only reported in A-phylogroup strains (Table 12).

Table 12. Frequency of amino acid substitutions for FimH and ChiA proteins in relation to pathotype and phylogroup. Only amino acid substitutions existent in more than three strains were examined. *E. coli* K-12 commensal strain was used as reference.

FimH Position	PATHOTYPE			PHYLOGROUP ^a				
	AIEC (N=29)	non-AIEC (N=25)	P	A (N=8)	B1 (N=6)	B2 (N=26)	D (N=4)	P
V27A	75.9%	84.0%	NS	75.0%	100%	84.6%	50.0%	NS
N70S	31.0%	40.0%	NS	0.0%	0%	69.2%	25.0%	<0.001
S78N	41.4%	40.0%	NS	0%	0%	73.1%	25.0%	<0.001
A119V	17.2%	8.0%	NS	37.5%	0%	7.7%	0%	NS
V163A	10.3%	12.0%	NS	0%	0%	23.1%	0%	NS

ChiA Position	PATHOTYPE			PHYLOGROUP ^a				
	AIEC (N=25)	non-AIEC (N=17)	P	A (N=8)	B1 (N=1)	B2 (N=23)	D (N=1)	P
ins315_317 PET	68.0%	64.7%	NS	0%	0%	87.0%	100%	<0.001
S326N	76.0%	70.6%	NS	12.5%	0%	91.3%	100%	<0.001
V335G	12.0%	11.8%	NS	50.0%	0%	0%	0%	0.003
V335S	72.0%	58.8%	NS	0%	0%	87.0%	100%	<0.001
K362Q	72.0%	52.9%	NS	0%	0%	82.6%	100%	<0.001
K370E	72.0%	52.9%	NS	0%	0%	82.6%	100%	<0.001
A378V	72.0%	64.7%	NS	0%	0%	91.3%	100%	<0.001
E388V	72.0%	70.6%	NS	12.5%	0%	87.0%	100%	0.001
L396M	72.0%	64.7%	NS	0%	0%	91.3%	100%	<0.001
V414I	72.0%	64.7%	NS	0%	0%	91.3%	100%	<0.001
A415V	20.0%	0%	NS	0%	0%	4.3%	0%	NS
D416N	48.0%	35.3%	NS	0%	0%	65.2%	0%	0.008
D427N	64.0%	64.7%	NS	0%	0%	82.6%	100%	<0.001

^a Only strains isolated from Crohn's disease or controls were considered. Atypical strain was discarded. NS: not significant.

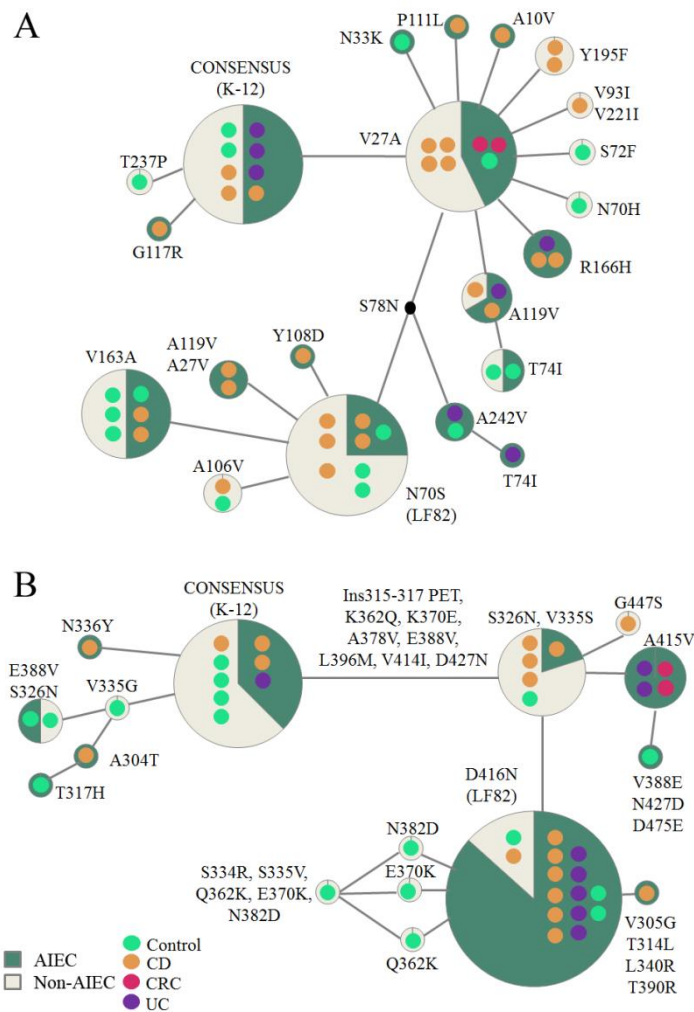


Figure 8. Reticulate tree representing FimH (A) and ChiA (B) variants. Each circle demonstrates the strains carrying specific mutations in the FimH/ChiA protein sequence. Number of strains is represented by the number of colored dots which also reflect origin of isolation. The amino acid changes indicated are derivatives of the consensus sequence (based on *E. coli* K-12 strain). AIEC pathotype proportion is indicated in green and non-AIEC in beige for each variant.

1.3 Test for rapid AIEC identification

To further establish a strategy that allows rapid identification of AIEC strains, we combined all the data of VGs carriage, amino acid variants and antibiotic resistance and performed Binary Logistic Regression to search for predictive features for AIEC screening (Table 13). In the present work, the combination of ampicillin resistance (Odds ratio=5.244; 95% CI=1.325-20.757) together with the prevalence of the *pic* gene (Odds ratio=4.854; 95% CI=1.140-20.638) uncovered a possible technique to identify AIEC strains, as it classifies strains according to the phenotype with a 75.5% of global success ($P(\text{AIEC}) = -1.974 + 1.657 \times \text{ampicillin resistance} + 1.579 \times \text{pic gene}$). For a given *E. coli* strain already isolated from human intestine that presents ampicillin

resistance and harbors the *pic* gene, the probability to be AIEC would be of 87.81%. This probability is reduced to 59.76% and 57.87% if the strain has either ampicillin resistance or the *pic* gene respectively, and it ends up to 22.07% if the strain is sensible to ampicillin and does not present the *pic* gene. Another combination resulted also significant (ampicillin resistance with *vat* gene prevalence). However, low sensitivity was achieved in this case (sensitivity 50%, specificity 77.8% and accuracy 65.3%).

Table 13. Binary logistic regression model evaluating the prevalence of the *pic* gene and ampicillin resistance as a putative model for AIEC identification.

EQUATION VALUES				
	B	p-value	Odds ratio	95% CI
<i>pic</i> gene	1.579	0.033	4.851	1.140-20.638
Ampicillin resistance	1.657	0.018	5.244	1.325-20.757
Constant	-1.974	0.020	0.139	

OBSERVED	PREDICTED			Global %
	non-AIEC	AIEC	% Correct	
Non-AIEC	18	9	66.7	75.5
AIEC	3	19	86.4	

	PROBABILITY TO BE AIEC	
	<i>pic</i> positive	<i>pic</i> negative
Ampicillin Resistant	87.81%	59.76%
Ampicillin Sensitive	57.87%	22.07%

Discussion

The AIEC pathotype has been involved in CD, and our knowledge about its distribution in other intestinal or extraintestinal diseases as well as the reservoirs and transmission paths is scarce. One reason of that is due to the fact that AIEC identification is based on phenotypic traits undergoing cell-culture infection assays, which are extremely time consuming and hard to standardise. In this work we have deeply characterised genetically and phenotypically a collection of AIEC and non-AIEC strains isolated from the intestinal mucosa of human and animals with the aim to better define the characteristics of AIEC pathotype and to find putative genetic/phenotypic markers for its rapid identification.

In our collection, higher number of VGs were associated with animal than with human strains and although the phylogenetic origin determined VGs profiles, differences between human and

animal strains were still evident when exclusively B2 strains were considered for comparison. This observation must be considered to further search for genetic traits associated with AIEC pathotype. The inclusion of animal strains in the study helped us to detect that the host origin of isolation needs to be carefully considered when drawing conclusions.

In the present work we have focused on strains isolated from humans. Four genes (*vat*, *pic*, *iss* and *papG*) differentially distributed between AIEC and non-AIEC strains have been identified. So far, there are limited studies in which the prevalence of these genes in AIEC strains has been investigated. Among these four genes, the vacuolating autotransporter toxin (*vat* gene) has been implicated in LF82 AIEC pathogenesis¹⁵⁷. It encodes for an autotransporter toxin involved in the gut mucus degradation. We found higher frequency of *vat*-positive AIEC strains similar to two previous studies: Desilets et al.²³ (9/13 AIEC and 0/6 non-AIEC) and Gibold et al.¹⁵⁷ (32/75 AIEC and 10/70 non-AIEC). On the other hand, in our work, no differences in the prevalence of *vat* according to pathotype were described once only B2 strains were considered, as occurred in O'Brien et al.¹⁷⁸. Nonetheless, higher adhesion and invasion values were reported for those strains harboring the *vat* gene, such as previously reported¹⁵⁷. The *pic* gene also encodes for a protease with toxin autotransporter activity, so it could be also involved in AIEC pathogenesis. However, so far there have only been studies relating it with *Shigella flexneri*²⁴³, and strains from de Uropathogenic *E. coli*, Enteroaggregative *E. coli* and Enteroinvasive *E. coli* pathotypes²⁴⁴. To the best of our knowledge no study previously analysed its presence in an AIEC collection. Herein, we reported occurrence of this gene in a subset of AIEC strains (41%) while it was less frequent in non-AIEC strains (16%). Moreover, higher adhesion values for *pic*-positive strains were found. This observation together with the fact that *pic* may contribute to intestinal colonization in mouse models for enteroaggregative *E. coli*²⁴⁴, suggest that the presence of *pic* might confer some bacterial virulence advantage. Isogenic mutants to confirm its implication in AIEC virulence are required. However, no differences between AIEC*pic*+ (60%) and non-AIEC*pic*+ (29%) strains was found once only B2-phylogroup strains were considered, a fact that may be attributable to the amount of strains analysed. The *iss* (increased serum survival) gene encodes for a protein responsible for serum resistance in ExPEC, such as Avian pathogenic *E. coli* strains²⁴⁵. In this case, Dogan et al.¹⁵⁴ did not describe an association with AIEC, but probably differences in the phylogenetic origin of the strain collections may influence these results. Finally, the combination of alleles *papGII-III*, encoding for adhesins of the *E. coli* pilus P, have been found in a low percentage of human strains (12%). Nonetheless, this gene is involved in adhesion processes and has been suggested to contribute to the urosepsis' pathogenesis²⁴⁶. The prevalence of *papGII* has

only been reported in AIEC strains isolated from CD pediatric patients yet in a very low frequency compared with AIEC isolated from controls¹³⁶.

Previous studies have reported differences in the prevalence of some genes according to pathotype (*pduC*, *chuA*, *lpfA*, *lpfA+gfpA* and *vat*)^{74,154,157,209}. However, in our strain collection, similar *lpfA*₁₅₄ and *chuA* gene prevalence values were reported between AIEC and non-AIEC isolates. Bearing in mind that the VG carriage is deeply associated with the phylogenetic origin⁷¹, we suspect that these discrepancies may be explained due to the diversity of the strain collection used in each study. Therefore, our results confirm the high genetic variability of AIEC strains and suggest that many of the genetic features described to date are in fact related to phylogroup origin of the strains rather than to AIEC phenotype.

Results obtained on FimH, one of the most studied virulence factor in AIEC pathotype, are in line with previous data^{23,74,165,178}, since no differences in pathoadaptative mutations were specifically associated with AIEC pathotype. Besides, although Dreux et al.¹⁶⁵ and Iebba et al.¹³⁸ indicated that N70S and S78N FimH variants could confer increased strain's capacity to adhere to the human receptor CEACAM6, no increased adhesion was observed for strains harboring these variants in our collection. Nonetheless, the strains with A119V mutation, a substitution previously reported to confer an advantage on adhesion¹³⁸, presented higher invasion indices. Actually, N70S and S78N associated with B2 and D-phylogroup strains, as it has been determined in other groups of strains^{18,23,138,165,247}. Finally, while G66S and V27A variants have been associated with CD origin of the strains and T74I, V163A and A242V variants with UC strains in a previous study¹³⁸, no particular variants were associated with disease origin in the present work.

Up to our knowledge, this is the first study examining the sequence of ChiA in a large strain collection (N=58). Until now, differences in ChiA sequence were only sought between LF82 and K-12 and five mutations (Q362K, E370K, V378A, V388E, and E548V) were described as required for the proper interaction between bacteria and epithelial cells⁸⁵. Despite we found these mutations equally distributed between AIEC and non-AIEC strains and no significant differences neither in adhesiveness nor in invasiveness between variants, it is of note that the AIEC LF82 sequence variant was mainly shared among AIEC strains (being the 85% of strains with this sequence AIEC). Unfortunately this variant is not highly frequent amongst the whole AIEC collection, only the 35.5% of AIEC strains had this gene sequence variant, so we suggest this gene is not suitable for AIEC screening. Additional studies regarding the expression of *chiA* gene

would be needed in order to decipher whether the strains harboring the same sequence express this gene differentially according to pathotype.

Scarce studies have evaluated the capacity of AIEC strains to resist the action of antibiotics^{24,127,132,248} and no one has compared antibiotic resistance between AIEC and non-AIEC strains. In this work, we have combined this feature with VGs prevalence. Despite no specific and widely distributed AIEC characteristic has been found, in this work we show that the presence of *pic* gene and ampicillin resistance are two traits that could assist in AIEC screening since AmpR *pic* + *E. coli* strains have a probability of 82% to be AIEC. This could be of use as an initial method of screening of human *E. coli* isolates. The major problem is about false-positives, so AIEC predicted strains by this method should be further tested phenotypically. It is also necessary to test the specificity of the method using genetically close pathotypes such as Extraintestinal Pathogenic *E. coli* and to test the applicability in external strain collections isolated from different geographical locations.

To sum up, this data provide deepest knowledge about AIEC VGs sets, what has revealed four VGs that could be of relevance in AIEC pathogenicity. We reinforce the idea that no particular VG is related to AIEC phenotype. Despite diverse virulence factors could drive to the same phenotype, the presence of an AIEC-specific marker cannot be discarded. Differences in gene expression or point mutations of core genes may explain the genetic basis of AIEC pathotype. Noticeably, a novel strategy to assist in AIEC identification is proposed, yet further works confirming our results in additional strain collections are necessary.

Chapter 1.2: Amino acid substitutions and differential gene expression of outer membrane proteins in AIEC

Results

2.1 OMPs sequence variants

Changes in the sequence of *ompA*, *ompC* and *ompF* genes in 43 *E. coli* isolates were studied by PCR and subsequent Sanger sequencing and compared to the OMPs gene sequence of the LF82 strain. One hundred percent amplification was achieved for the *ompA* and *ompF* genes while for the *ompC* gene was 97.6% (one non-AIEC was PCR-negative). Of the three OMPs, OmpC protein was the most variable (81% similarity), and OmpA was the most conserved (94% similarity) among all of the strains. The similarity for OmpF was 91%. The sequence of OmpA varied at 17 amino acid positions and was grouped in 13 variants (Figure 9 and Table S14). OmpC and OmpF sequences differed at 65 and 30 positions resulting in 25 and 10 variants respectively (Figure 9 and Table S15, S16).

The most common variant in OmpA was present in 15/61 strains, which shared the same OmpA sequence as K-12 (Table S14). The non-AIEC strains were predominant in this OmpA variant (N=9); nonetheless, some AIEC (N=4) and IPEC (N=2) strains also harbored the variant. The second most common variant (n=14) was found in AIEC (n=7), ExPEC (n=1) and non-AIEC (n=6) strains. Moreover, the LF82 OmpA variant was present in two AIEC strains, two non-AIEC strains and one ExPEC. For OmpC, the most common variant (12/58 strains) varied only in two positions in comparison with the LF82 sequence and comprised 6 AIEC, 5 non-AIEC and 1 ExPEC strains (Table S15). Again, the LF82 variant was shared with one AIEC and two non-AIEC strains. Finally, the OmpF protein presented the lowest number of variants, although the number of point mutations was higher than that of OmpA. In this case, 29/58 strains, including 11 AIEC, 14 non-AIEC and 4 ExPEC strains, displayed the same amino acid sequence as the LF82 strain (Table S16). The second most common variant comprised 13/58 strains: 2 AIEC, 9 non-AIEC, 1 IPEC and 1 ExPEC in which K-12 was included.

Overall, no protein variants were specifically associated with AIEC strains, and OMPs LF82 variants were also detected in non-AIEC strains.

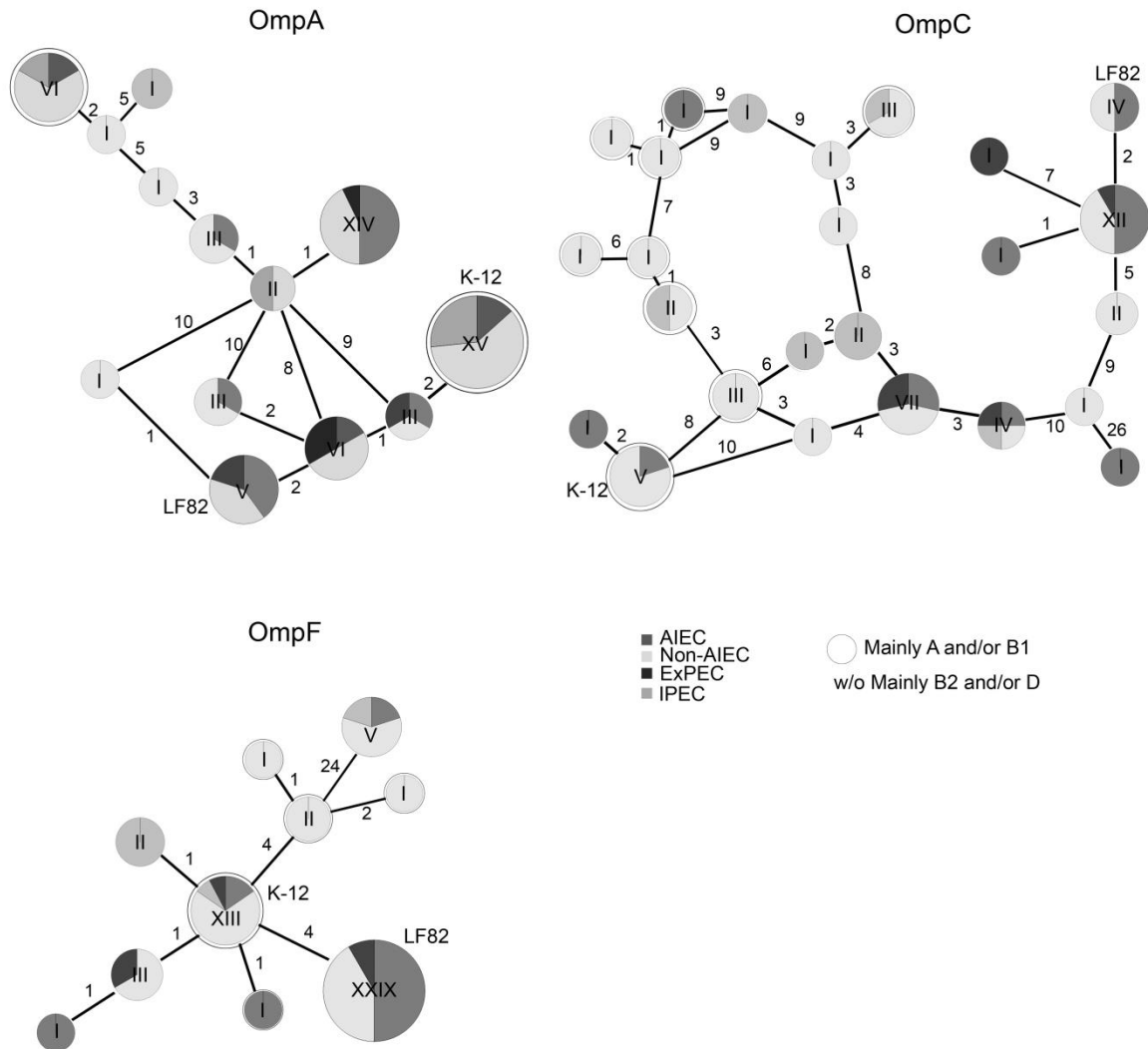


Figure 9. Reticulated trees representing the distribution of strains carrying specific amino acid substitutions in OMPs. The number of strains in each variant is pointed out by Roman numerals. The number of amino acid changes between each variant is indicated. The LF82 strain was used as a reference. ExPEC and IPEC gene sequences were retrieved from NCBI. The rounded circles are groups of strains mainly from the A and/or B1 phylogroups while the others (without circles or w/o) mainly involve B2 and/or D-phylogroup strains.

2.2 Distribution of amino acid substitutions in OMPs

Although no particular sequence variants were associated with the AIEC pathotype, the differential distribution of point mutations was analysed. In our strain collection, previously described mutations¹⁵⁹ between LF82 and K-12 strains in the OmpA protein sequence (V114I, F131V, D132Y, T228N and A276G) showed similar percentages between AIEC and non-AIEC strains ($p > 0.101$; Figure S3). In contrast, in this study, the valine (V) residue in position 200 of OmpA sequence was more frequently found in AIEC than in non-AIEC strains, and the same occurred for V220I of OmpC (Table 14, Table S17). Two OmpC positions (S89N and W231D)

were also differentially distributed among AIEC and non-AIEC strains (Table 14, Table S17). Increased adhesion indices correlated with V residue in the periplasmic position 200 of OmpA ($p=0.044$), and increased invasion indices with V in the extracellular position 220 of OmpC ($p<0.022$) (Table 14). Additional amino acid substitutions correlated with higher adhesion (OmpC extracellular position 232; OmpF periplasmic position 51; OmpF extracellular position 60) and invasion (OmpC extracellular positions 220 and 232) indices ($p<0.040$; Table 14).

Table 14. Variable positions in OMPs sequence related to pathotype or AIEC phenotypic characteristics. In bold statistically significant comparisons are indicated. For each position, the first amino acid presented is the one present in the LF82 reference strain.

Variable position	Amino acid	All strains			Adhesion index		Invasion index	
		Non-AIEC (n=30)	AIEC (n=14)	p-value	Bacteria/cell	p-value	%	p-value
OmpA 200	A	80.0%	42.9%	0.018	4.1±7.4	0.044	0.184±0.457	0.084
	V	20.0%	57.1%		7.8±7.9		0.224±0.359	
OmpC 89 220 231 232	S	31.0%	64.3%	0.041	6.4±8.1	0.372	0.314±0.589	0.109
	N	69.0%	35.7%		4.7±7.4		0.080±0.158	
	V	20.7%	57.1%	0.022	7.9±8.6	0.078	0.366±0.653	0.022
	I	79.3%	42.9%		4.2±7.0		0.087±0.169	
	W	31.0%	64.3%	0.041	6.4±8.1	0.189	0.315±0.589	0.058
	D	69.0%	35.7%		4.6±7.5		0.079±0.158	
D	62.1%	85.7%	0.108	6.5±8.0	0.017	0.229±0.474	0.002	
A	37.9%	14.3%		2.9±6.5		0.060±0.157		
OmpF 51 60	E	80.0%	100.0%	0.084	6.1±7.9	0.040	0.198±0.433	0.322
	V	20.0%	0.0%		0.2±0.4		0.021±0.019	
	M	80.0%	100.0%	0.084	6.1±7.9	0.040	0.198±0.433	0.322
	K	20.0%	0.0%		0.2±0.4		0.021±0.019	

An association with the phylogroup origin of the strains was detected in 11 variable positions of OmpA ($p<0.043$), 32 of OmpC ($p<0.046$) and 1 of OmpF ($p<0.001$) (Table S17). In most cases, A-phylogroup strains resembled B1, while B2 more closely resembled D.

2.3 OMPs protein expression

OMP's protein expression profiles of AIEC and non-AIEC strains growing in MH broth were examined by urea-SDS-PAGE (Figure 10). OmpA was expressed in the majority of both AIEC and non-AIEC strains (92.9% and 96.3% respectively), OmpF was also frequently expressed in

AIEC strains (92.9%) but was less prevalent in non-AIEC strains (70.4%), and OmpC could be detected in 64.3% of AIEC strains and only in 44.4% of non-AIEC strains. However, the decreasing tendency in OmpF and OmpC expression observed in non-AIEC strains was not significant. Considering the expression profiles of the OMPs, OmpC expression was concomitant with OmpA and/or OmpF for both groups of strains. In the absence of OmpC, OmpF was always expressed together with OmpA. OmpA was present alone in only 7% of AIEC strains and in 22% of non-AIEC strains.

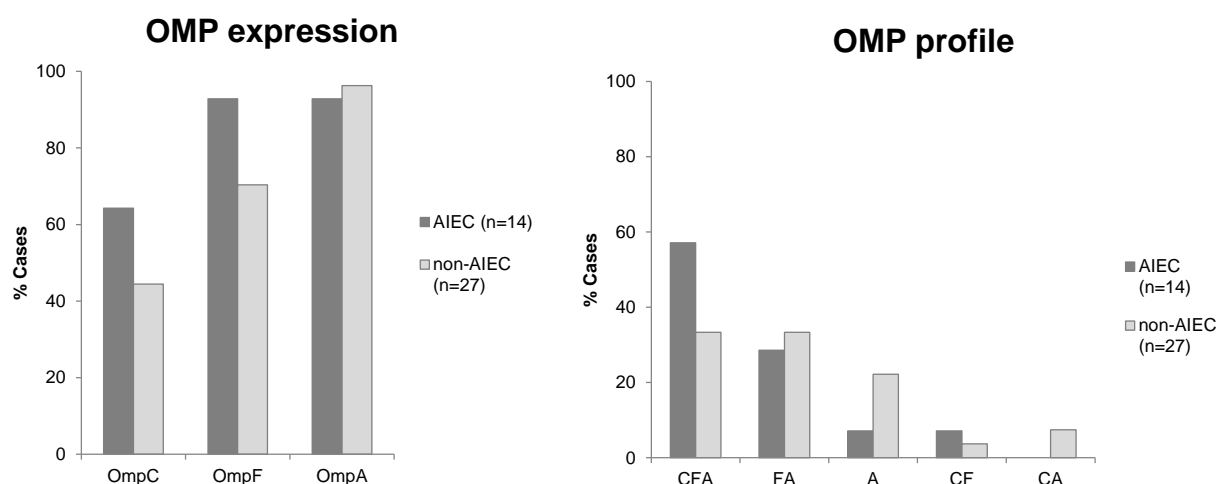


Figure 10. Protein OMPs expression profiles in AIEC and non-AIEC strains growing in MH broth. Left: Percentage of AIEC and non-AIEC strains expressing OmpC, OmpF and OmpA. Right: Percentage of OMP profiles for each strain. Less abundant profiles have been grouped as “Other”. C, OmpC; F, OmpF; A, OmpA.

2.4 OMPs gene expression

To evaluate differential expression patterns in OMPs among a collection of AIEC and non-AIEC strains during in vitro bacterial infection of intestine-407 cells by RT-qPCR. RNA was extracted from two fractions of infected cell cultures: i) from non-adhered / noninvading bacteria present in the supernatant (SN), and ii) from bacteria adhering and/or invading the intestine-407 cells (INV). RNA was thus obtained from mixed cultures (eukaryotic and prokaryotic); therefore, the greater amount of RNA was of eukaryotic origin in INV fractions. Considering that total RNA concentrations could mask the differences in bacterial quantity between strains, the 16S rRNA was measured and used as a normaliser for bacterial gene expression values. As expected, considering that AIEC strains invade and survive intracellularly, 16S rRNA gene copy numbers varied according to pathotype in the INV condition ($p=0.002$), AIEC strains presented higher values ($3.54 \times 10^8 \pm 8.60 \times 10^7$ 16S rRNA copies) than non-AIEC strains ($9.93 \times 10^7 \pm 2.12 \times 10^7$ 16S

rRNA copies). In the SN condition, minor differences in 16S rRNA gene copy numbers were identified (AIEC $1.24 \times 10^8 \pm 6.68 \times 10^7$ 16S rRNA copies; non-AIEC $3.54 \times 10^8 \pm 8.87 \times 10^7$ 16S rRNA copies; $p=0.043$).

After normalization of gene expression values, paired tests were performed to uncover differences in OMPs expression according to the condition (SN or INV) (Figure 11). AIEC strains showed significantly decreased expression of all OMPs in the fraction of adherent/invasive bacteria in comparison to the strains present in the supernatant ($p < 0.032$). All AIEC strains showed a reduction of gene expression values in the INV fraction, with the exception of AIEC17 and AIEC14-1, which presented higher expression in INV than in SN for all OMPs. Conversely, non-AIEC strains did not show differences in gene expression between conditions for any of the three genes studied ($p > 0.577$).

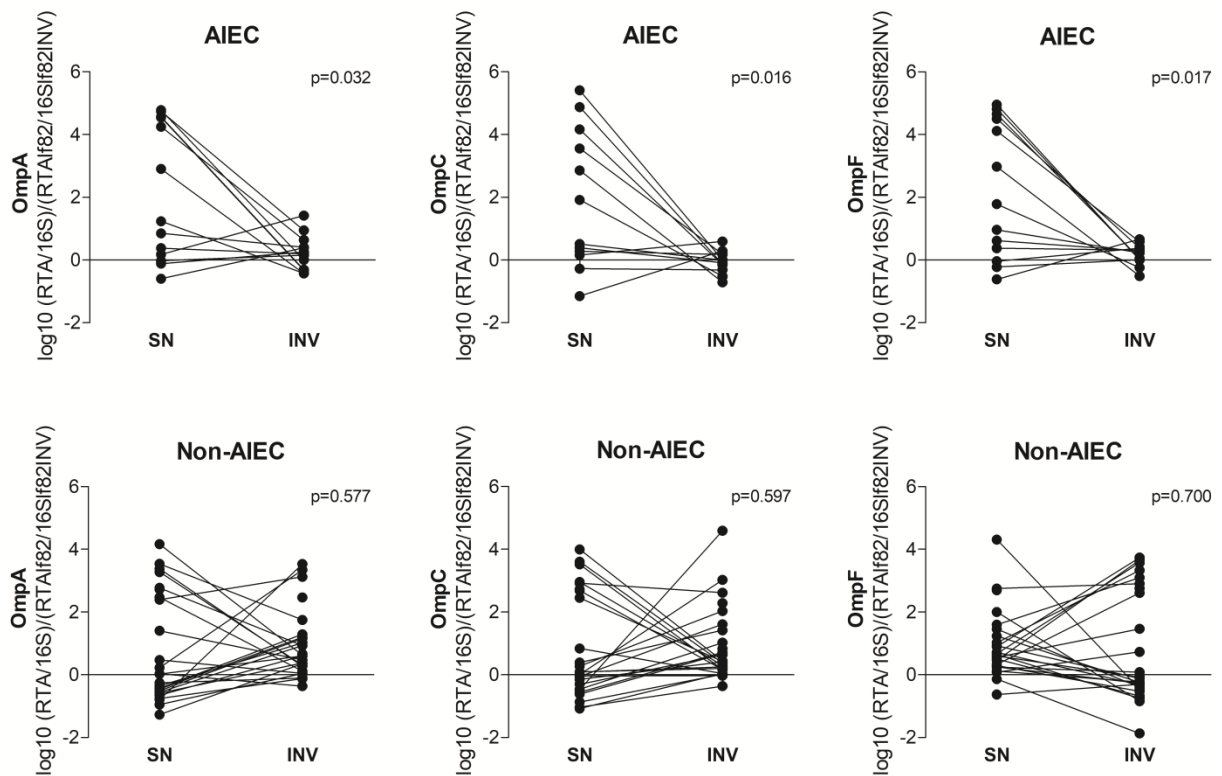


Figure 11. Paired tests evaluating the OMPs expression difference between supernatant and cell-associated fractions of infected I-407 cultures in each strain collection (AIEC and non-AIEC). Values indicate the logarithmic ratio of relative transcript abundance (RTA) of the target sample divided by its 16S rRNA copy number value and the RTA/16S of the reference strain in the INV condition (LF82_INV), being $RTA = \text{Efficiency}^{(Ct \text{ target gene reference strain} - Ct \text{ target gene sample})} / \text{Efficiency}^{(Ct \text{ constitutive gene reference strain} - Ct \text{ constitutive gene sample})}^{222}$.

In addition, OMPs gene expression differed by pathotype in both conditions analysed (SN and INV) (Figure 12). Indeed, in the SN condition, AIEC strains presented higher *ompA* and *ompF* gene expression than non-AIEC strains ($p=0.013$ and $p=0.012$, respectively), but no differences in *ompC* gene expression were observed. No correlation was evident between gene expression and adhesion ($p>0.141$), or invasion ($p>0.354$) abilities of the strains (Figure S4). Of note, a subgroup of both AIEC (AIEC04, AIEC07, AIEC09 and LF82) and non-AIEC (ECG01, ECG11, ECG12, ECG13, ECG16, ECG43, ECG57 and ECG63) strains presented higher expression values than the other isolates in their group. These groups of strains presented no specific characteristics in terms of OMPs gene sequence, phylogroup, antibiotic resistance or virulence gene profile. In contrast to what has been observed in the SN, in the INV condition, AIEC strains globally presented lower OMPs gene expression values than non-AIEC strains ($p<0.039$). Among AIEC, a negative correlation was observed between gene expression and adhesion ability of the strains but was not significant ($p>0.123$) (Figure S5). No correlation was observed for non-AIEC strains in any case (SN: adhesion $p>0.852$ and invasion $p>0.695$; INV: adhesion $p>0.478$ and invasion $p>0.621$).

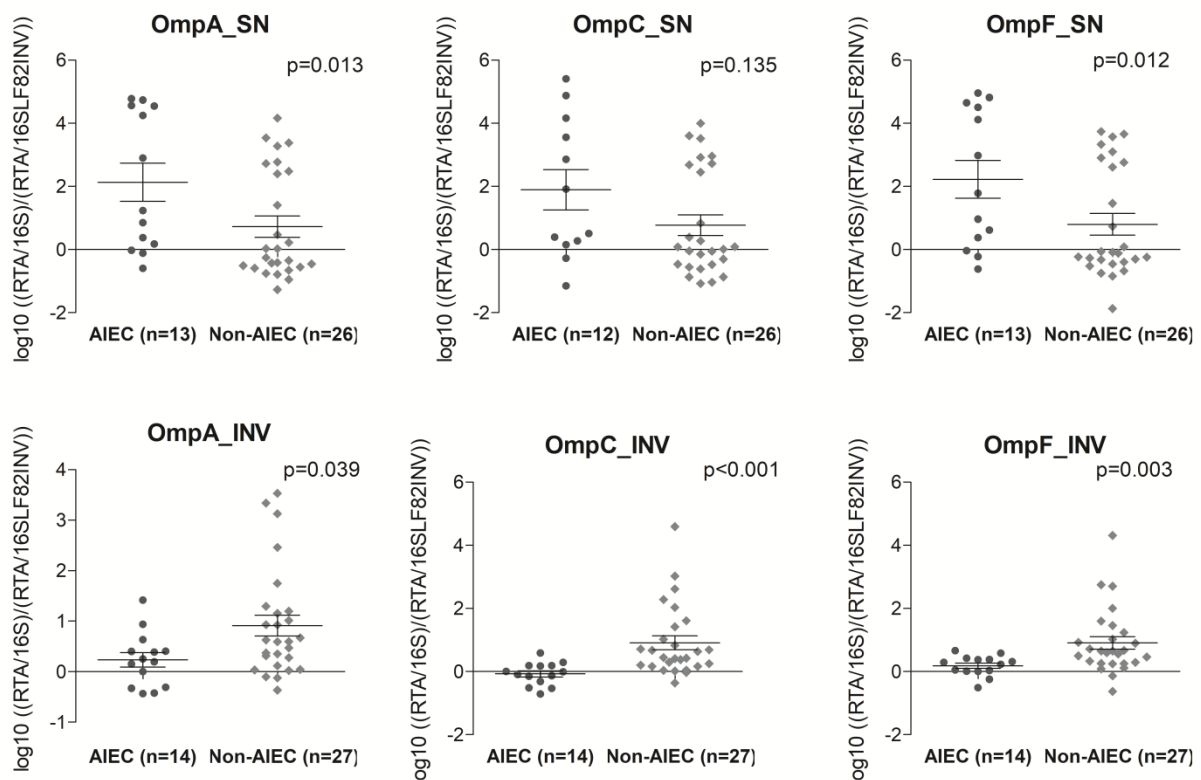


Figure 12. Differential OMPs expression between AIEC and non-AIEC strains from each fraction of the infected I-407 cultures (SN and INV). Values indicate the logarithmic ratio of relative transcript abundance (RTA) of the target sample divided by its 16S rRNA copy number value and the RTA/16S of the reference strain in the INV condition (LF82_INV), being $RTA = \text{Efficiency}^{\wedge}(\text{Ct target gene reference})$

strain – Ct target gene sample) / Efficiency \wedge (Ct constitutive gene reference strain – Ct constitutive gene sample)²²².

Despite differences in gene expression were seen according to phylogroup, there was much variability within each group, and no significant differences were observed in any of the conditions ($p > 0.161$) (Table S18).

Discussion

In gram-negative bacteria, OMPs have been shown to contribute not only to the structural integrity of the outer membrane, passive ion and solute transport but also to stress survival, bacterial virulence and resistance to antibiotics^{249–255}. Given that amino acid substitutions in these genes or differences in their expression have been implicated in the adhesion and invasion capacities of some *E. coli* pathotypes^{159,160,251,252,254,256}, in this work we sought to investigate the distribution of OMPs amino acid substitutions in a collection of AIEC and non-AIEC strains, along with the differential OMP expression of those strains in two distinct conditions (growing in suspension in a cell-culture medium and adhering/invading intestinal epithelial cells), to define whether these proteins can contribute to AIEC virulence.

Little evidence exists on the putative role of OMPs in AIEC virulence^{159,160}. It has been suggested that OmpA interacts with the overexpressed receptor Gp96 of the intestinal epithelium of CD patients to promote AIEC invasion¹⁵⁹. The authors found five OmpA amino acid variants (V114I, F131V, D132Y, T228N and A276G) in AIEC LF82 relative to *E. coli* K-12, that could be responsible for the increased invasion ability¹⁵⁹. Nonetheless, these amino acid positions were not conserved in the majority of our AIEC strains and did not correlate with the adhesion and invasion abilities of the strains. Moreover, LF82 OmpA variant was found in non-AIEC strains. In agreement with a previous study¹⁸, we found that OmpA gene variants were similar between AIEC, IPEC, ExPEC and non-AIEC strains. This suggested that (I) the five amino acid variant positions previously described are not relevant in our strain collection, (II) additional virulence genes determine the adhesion and invasion abilities of some of the strains, and (III) non-AIEC strains with the LF82 variant do not possess any invasion capacity probably because their *ompA* mRNA levels are low. Although no particular substitutions located in the *N*-terminal of OmpA were found to be associated with AIEC, the periplasmic position A200V was associated with pathotype and even bacterial adhesion. Because modifications in the periplasmic site of the protein may lead to misfolding of extracellular loops and thus also compromise protein-receptor

interactions²⁵⁷, functional studies could be conducted to assess the effect of this amino acid substitution in the AIEC phenotype.

The role of *ompC* expression in the interaction of AIEC with IECs under conditions of high osmolarity has been previously analysed by Rolhion et al.¹⁶⁰. Although reduced adhesion and invasion levels were reported in the LF82 *OmpC*-mutant, the wild-type LF82 phenotype was restored by overexpressing the RpoE (σ^E) regulatory pathway in a LF82 mutant that did not express *OmpC*. Therefore, the authors concluded that *OmpC* involvement in the ability of LF82 to adhere to and invade IECs was indirect¹⁶⁰. To our knowledge, only two studies have examined *OmpC* prevalence in AIEC strains^{18,50}, and this is the first work analyzing the *OmpC* sequence in more than one AIEC strain. Our results showed that this gene is widely present among *E. coli* strains regardless of the pathotype, as previously postulated^{18,50}. Moreover, four amino acid substitutions were differentially distributed among AIEC/non-AIEC strains (S89N, V220I and W231D) or associated with increased adhesion and/or invasion capacities (V220I and D232A), with two of the substitutions located in the extracellular region of the protein (V220I and D232A). Notably, it has been previously suggested that variations in the extracellular residues of *OmpC* may influence bacterial virulence because reduced adherence to macrophages has been reported in *Salmonella typhimurium* with an altered extracellular *OmpC* region²⁵⁸.

To date, studies examining *OmpF* amino acid substitutions have focused on analyzing the implication in antibiotic resistance^{259–261}, but information on the role of *OmpF* in bacterial pathogenicity remains poorly understood. Nonetheless, a role for *OmpF* has been pointed out in avian pathogenic *E. coli* (APEC), this protein is involved in adhesion and invasion to mouse brain microvascular endothelial cells *in vitro* and in brain, blood and lung colonization *in vivo*²⁵³. Currently, no evidence exists on the sequence variance of the *ompF* gene in AIEC, and its expression has been only studied in the AIEC LF82 strain¹⁶⁰. Here, we found no specific *OmpF* sequence variant associated with AIEC; however, two residues present in the periplasmic (E51) and extracellular (M60) domains correlated with high adhesion capacity of the strains.

Taken together, these data demonstrate that OMPs gene sequence variants are not sufficient to predict the AIEC pathotype due to their low sensitivity and specificity. However, the data contribute to increasing knowledge on AIEC genetics and reveal new putative pathoadaptative mutations that may influence adhesion and/or invasion of strains. Differences in the OMP protein sequence might be explained by the origin of the strains, as most of the variable positions

were linked with the phylogroup, which is consistent with the observations of previous studies where sequence variations of other AIEC-associated virulence genes, for instance FimH, were investigated^{138,165}. Further studies are needed to ascertain whether these amino acid residues could lead to an improved interaction between AIEC and IECs.

In an attempt to discriminate the AIEC pathotype among commensal *E. coli* isolates, OMPs expression profile was examined using overnight cultures. However, no differences were found. Thus, an analysis of OMPs expression during infection was carried out since external signals might be necessary to induce the expression of virulence factors.

This is the first study analyzing differential gene expression between a collection of AIEC and non-AIEC strains during IECs infection. We reported higher OMPs expression for AIEC strains in the SN of infected cultures than for non-AIEC strains, while in the INV condition the opposite occurred (Figure 13). According to the methodology applied in this study, the INV fraction includes all bacteria that are in contact with the epithelial cells, both adhering and invading. While non-AIEC strains showed no alteration in gene expression when growing on the SN or adhering to IECs, AIEC showed decreased gene expression in the INV fraction. We hypothesise that to increase the chance of adherence to IECs, AIEC enhance the expression of *ompA* and *ompF*, and once the bacteria have adhered to and invaded the IECs, they reduce the expression of OMPs. This may protect the bacteria from the acidic pH and from passive diffusion of oxidative residues, proteolytic molecules and antimicrobial peptides encountered in lysosomes and other solutes that might be created inside the eukaryotic cell through the OMPs channels. This hypothesis is supported by previous works, such as Lucchini et al.²⁶² who reported reduced *ompC* and *ompF* expression levels during epithelial infection by *S. flexneri*, as well as other studies that suggest that OmpC and OmpF are required for low pH survival^{263,264}. In addition, distinct from the hypothesis presented by Rolhion et al.¹⁶⁰, no correlation between OMPs gene expression and adhesion was observed in any of the conditions assessed.

Since OMPs gene expression may be regulated by, for example, the two-component OmpR-EnvZ regulatory system¹⁶⁰, the expression of other virulence factors could also be influenced by the same regulatory pathways. Therefore, it would be interesting to focus future studies not only on OMPs expression but also on coreregulated genes. Moreover, it is unclear whether the altered gene expression reported in the INV fraction was due to the adhered or the intracellular bacteria

because it included all bacteria that were in contact with IECs. Therefore, additional studies differentiating between both fractions would be of interest.

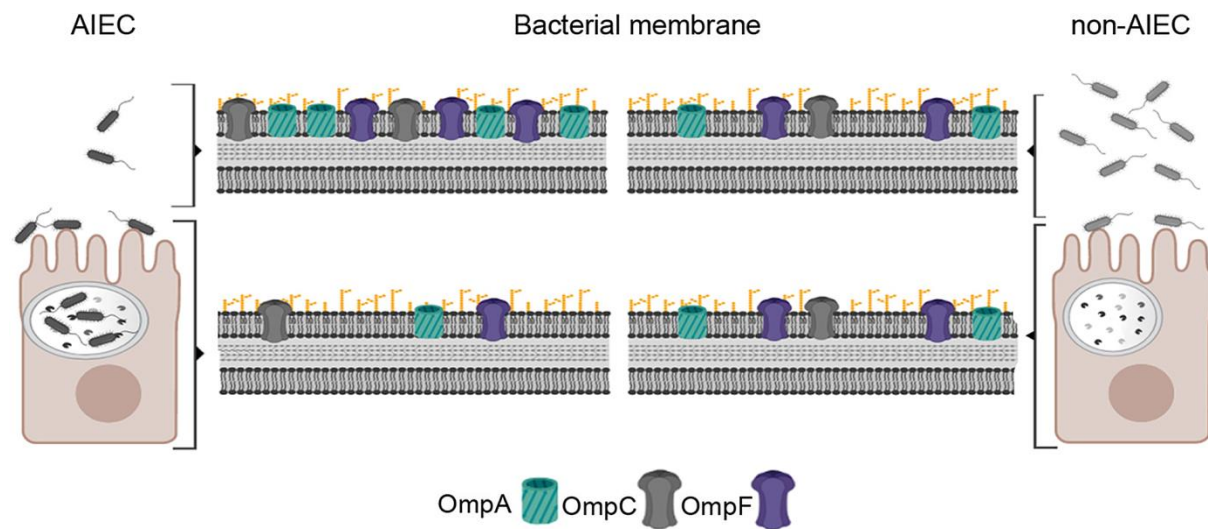


Figure 13. Schematic representation of gene expression levels of OMPs according to pathotype and fraction analysed. Protein expression has been used to represent mRNA levels. In the INV fraction, it is unclear whether the differences in OMPs mRNA levels are due to adhered or intracellular bacteria, especially for AIEC strains.

Our work provides new insights regarding OMPs sequence and expression in a wide collection of AIEC strains and adds knowledge about AIEC gene expression during IECs infection. We conclude that, although particular mutations in *ompA*, *ompC* and *ompF* gene sequences may enhance the adhesion and invasion capacity of AIEC strains, they are not crucial for the adherent-invasive phenotype. Notwithstanding, differential gene expression of these OMPs during infection may definitely contribute to AIEC pathogenicity by enhancing IECs adherence and intracellular persistence.

Chapter 2.1: Identification by comparative genomics of new single nucleotide polymorphisms to distinguish between AIEC and non-AIEC strains

Results

3.1. Characteristics of strain pairs

Three AIEC strains and three corresponding non-AIEC counterparts that shared identical PFGE patterns (Figure S1), sequence types, phylogenetic origins and virulence genes (Table 8) with the AIEC strains were selected for genome sequencing in the present study. As expected, the AIEC strains presented higher adhesion and invasion indices than did their non-AIEC counterparts. However, the non-AIEC strains also had the capacity to survive and replicate inside J774 murine macrophages but not in human THP-1 macrophages, with the exception of the ECG28 strain.

Similar total genome sizes ranging from 4,825 to 5,213 Kb were obtained for the AIEC and non-AIEC strains (Table 15), and no significant structural differences were found between the strain pairs (Figure S6). One inversion was detected in all pairwise comparisons, but the inversed regions were not homologous among the three strain pairs.

Table 15. Assembly features of the AIEC/non-AIEC sequenced genomes.

Strain	Size (kb)	Contigs (No.)	GC (%) ^a	N50 (kb) ^b	Accession No.
AIEC17	4,958	400	50.42	247	ERS1456453
ECG28	4,981	464	50.38	214	ERS1456454
AIEC01	5,213	333	50.48	186	ERS1456455
ECG11 ^c	5,212	25	50.56	555	ERS1456456
AIEC07	4,825	374	50.62	187	ERS1456457
ECG04	5,013	836	50.46	261	ERS1456458

^a GC(%), content defined as (G+C)/(A+T+G+C)x100.

^b N50, the length of the shortest contig at 50% of the assembly.

^c This strain was sequenced by PacBio (library with 10kb insert) and assembled with HGAP 3 tool.

Genes that were previously associated with AIEC were searched in the genomes of the six strains. The *gipA*, *chuA* and *fjuA* genes were present in the three AIEC strains and in their non-AIEC counterparts (Table 8). Other genes were not present in any strain (*afaC*) or present only in a single strain pair (*lpfA₁₅₄₃*, *pduC* and *ibeA*). Similar results were obtained for the analysis of the predicted amino acid sequences encoded by the *fimH*, *ompA* and *chiA* genes, for which the same

variants were found within a strain pair (Table S19). Clustered regularly interspaced palindromic repeats (CRISPR) analysis was also performed, and equal profiles were obtained for the pairs AIEC17-ECG28 and AIEC07-ECG04; no confirmed CRISPR was recognised in AIEC01 or in ECG11.

3.2. Comparative genomics of AIEC/non-AIEC strain pairs

The genomes of the three AIEC/non-AIEC strain pairs were compared within pairs to identify gene content differences and SNPs that could be implicated in the AIEC phenotype.

3.2.1 Evaluation of gene content dissimilarities

Homologous protein-encoding sequences were identified. A total of 5208 orthologous clusters of genes were obtained: 3327 (63.9%) clusters were common to the six strains (Figure 14a). These genes represented 80%, 77% and 81% of the genomes of the AIEC17-ECG28, AIEC01-ECG11 and AIEC07-ECG04 pairs, respectively. The strain pairs belonging to the B2 and D phylogroups shared a higher proportion of orthologous clusters of genes (9.2–9.6% of their genomes) than the B1 strains shared with B2 (2.5–2.6%) or D (2.2–2.3%) (Figure 14b). The high similarity between strain pairs was also evidenced in terms of gene content; because strains within a pair shared more than 99.2% of orthologous clusters of genes.

None of the clusters were shared by the three AIEC strains, not even exclusively by two of them (Figure 14a), indicating the absence of AIEC-specific genes that were present in all AIEC strains. However, within each strain pair, exclusive AIEC genes were identified (Table S20). AIEC17 contains two genes encoding uncharacterised proteins (YgiZ and YeeW), one gene encoding a cyanate transporter and a gene encoding a TraR family protein that is involved in a quorum-sensing process²⁶⁵. In the case of AIEC01, 33 of the genes present in its genome were absent from the genome of ECG11. Of those 33 genes, 20 encode proteins of unknown or generic function, six are related to transmembrane transport, four are implicated in transcriptional regulation, and three contribute to flagellum assembly. Finally, compared with ECG04, AIEC07 harbours three proteins of unknown function (one of which is an *E. coli* uropathogenic-specific protein), one protein related to intracellular iron transport (TonB), and the autotransporter UpaH. The latter protein mediates biofilm formation in the uropathogenic strain CFT073²⁶⁶. Biofilm formation is also a phenotypic trait of AIEC strains¹⁹¹.

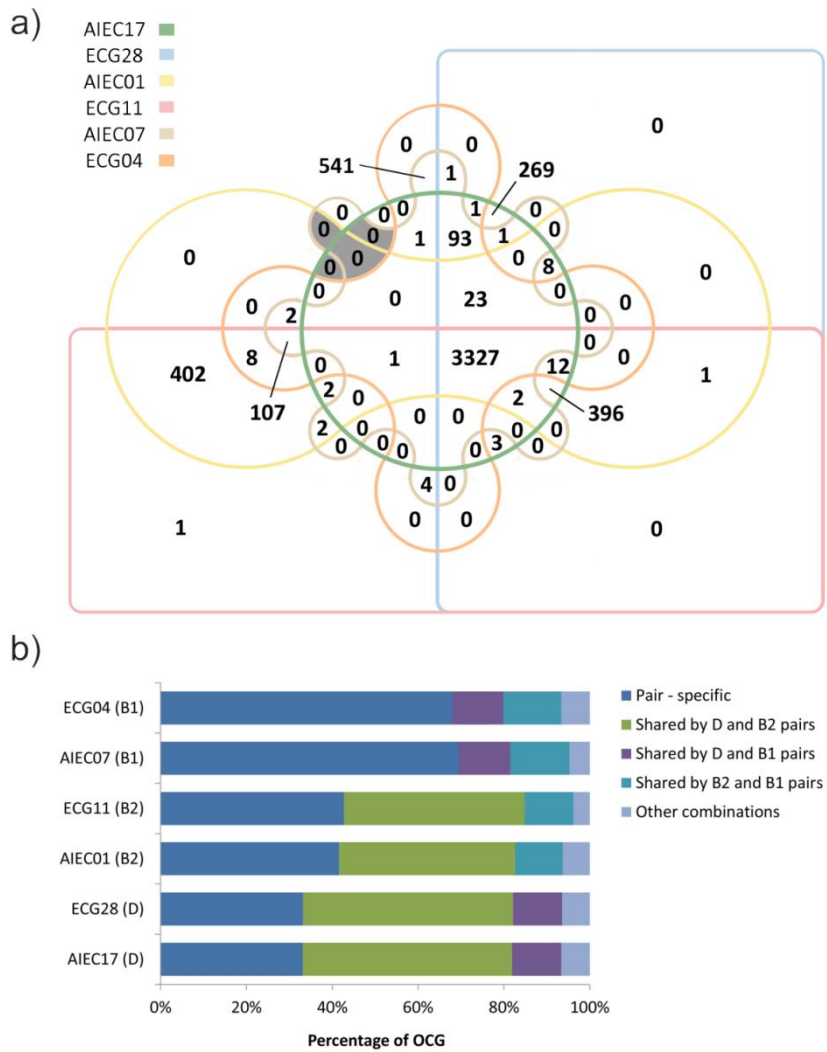


Figure 14. Genome similarities among the six strains and within each pair by analysis of orthologous clusters of genes (OCG) analysis. a: Edward’s Venn diagram indicating the number of OCG. The shadowed areas correspond to clusters exclusively shared between at least two AIEC strains. b: Percentage of OCG between phylogenetically distinct strain pairs and among AIEC strains. Percentages are calculated in relation to the number of variable OCG for each strain, those OCG that are not present in all six strains are considered variable. Other combinations include gene clusters shared by 5 strains or 3 strains from two or three different phylogroups. There were no common OCG among AIEC strains.

3.2.2 Detection of AIEC-associated SNPs

In our approach to identifying AIEC-associated SNPs, we focused our attention on the nucleotide positions that varied between the strains of a pair and were also located in homologous sequences of an AIEC reference genome (UM146 strain). A total of 286 polymorphisms were found (Table S21, S22 and S23); the majority (213) of the SNPs were

located in gene encoding sequences (Table 16). Of these SNPs, 60 were selected for resequencing, and only 20 of them were confirmed by Sanger sequencing.

Table 16. Number of SNPs present in the sequenced AIEC/non-AIEC strain pairs.

	AIEC17-ECG28	AIEC01-ECG11	AIEC07-ECG04
Total SNPs	51	126	109
Total SNPs in genes	40	91	82
Selected SNPs *¹	20 (7)	10 (5)	30 (19)
Confirmed SNPs *²	4 (4)	7 (3)	9 (4)
SNPs studied in a strain collection *³	4 (4)	4 (2)	8 (3)

* The number of genes in which the SNPs are located is indicated in parenthesis.

¹ SNPs that conform to the following criteria: (I) cause a non-synonymous amino acid change; (II) are not located at the end of a contig; and (III) were validated bioinformatically by ClustalW.

² Validated by Sanger Method.

³ Confirmed SNPs that were not strain-specific (and four strain-specific SNPs for validation).

Of note, 14 of these SNPs (70%) were found to present overlapping peaks in the Sanger chromatograms (Table 17). We hypothesised that i) strains with ambiguous bases may possess more than one copy of the relevant gene in their genomes or ii) there is intraclonal variability in the polymorphic site. To identify which of these possibilities gave rise to the ambiguous SNPs, we first performed BLASTn²²⁴ searches in the strains' genomes and the genome of the reference AIEC strain to search for duplicate genes. Next, we analysed next-generation sequencing (NGS) reads using Tablet²³² to confirm our observations. Duplicated genes showed a unique or main nucleotide in the SNP that differed from the nucleotide present in the other gene copy (this was the case for the SNPs found in the E3-E4_4.3, E3-E4_4.4 and E3-E4_4.7 genes). Genes that were not duplicated showed a single result in BLASTn and two different nucleotides that were almost equally frequent in the sequencing reads (this was the case for the SNPs found in the E5-E6_3.16=3.22 and E5-E6_3.17 genes). Therefore, we suggest that the overlapping peaks were not due to technical artefacts.

3.3 Distribution of SNPs in an AIEC/non-AIEC strain collection

We studied the variability of nucleotides within the identified SNPs in a collection of AIEC and non-AIEC strains to validate or refute the hypothesis that Confirmed SNPs represent putative molecular signatures for the specific identification of the AIEC pathotype. Sixteen SNPs were studied in 22 AIEC and 28 non-AIEC strains isolated from healthy subjects and CD patients belonging to several phylogroups (A (n=9), B1 (n=7), B2 (n=28), D (n=5) and atypical (n=1))

(Table S6). We selected only those Confirmed SNPs that were not strain-specific as assessed *in silico* because strain-specific SNPs would be uninformative for the identification of AIEC. Four SNPs considered strain-specific *in silico* (E1-E2_3.4, E1-E2_5, E1-E2_3.7, and E5-E6_3.1) were also analysed to confirm that they were strain-specific.

Interestingly, some nucleotide variants occurred more frequently in AIEC strains than in non-AIEC strains (Table 18). In particular, variants with thymidine in SNP E3-E4_4.3(2) were found only in AIEC, whereas those containing cytosine were more frequent within non-AIEC. Similar results were obtained when the analysis was restricted to B2 phylogroup strains ($p=0.037$); in this phylogroup, AIEC strains ($n=6$) were again the only ones presenting thymidine, and cytosine was also more prevalent in non-AIEC than in AIEC strains ($n=10$ and $n=7$, respectively). This gene was present in all the strains studied, and the SNP variants were not associated with the phylogroup origin of the strains but only with the AIEC phenotype. Another intriguing SNP was present in gene E3-E4_4.4 in which 42.86% of non-AIEC strains presented guanine, whereas less than 10% of the AIEC strains presented this variant. However, in this case, not all strains harboured the gene, as occurs for the LF82 strain. Finally, a guanine in SNP E5-E6_3.16=3.22(2) that presented a high variability within the strain collection was more frequently found in AIEC strains, whereas a cytosine at this position was associated with non-AIEC strains. If only B2 strains are considered, the cytosine variant is found exclusively in non-AIEC strains ($n=5$), whereas the guanine variant is specific to AIEC strains ($n=4$) ($p=0.007$). Therefore, this variant could be of interest as a biomarker for B2-phylogroup AIEC strains. However, the percentages of strains that present these two variants are low (41.7% of B2 non-AIEC and 30.8% of B2 AIEC).

Table 17. Location of the Confirmed SNPs, nucleotide variants and gene functions.

ID	Strain pair where identified	SNP location in AIEC genome (Contig; position)	Nucleotide variant*1 (AIEC/ non-AIEC)	Protein name	Protein family	Gene Ontology
E1-E2_3.4	AIEC17 vs ECG28	8: 204059	C/T	GntR family transcriptional regulator	PF00392; PF07702	GO:0003677; GO:0003700; GO:0006351
E1-E2_3.6		105: 325	C/T	Phage protein	PF06174	
E1-E2_3.7		3: 50414	T/C	Serine peptidase DegQ	PF13365; PF13180; PF00595	GO:0004252
E1-E2_5		51: 69	G/T	Vitamin B12 transporter BtuB	PF07715	GO:0009279; GO:0015235; GO:0006811; GO:0046872; GO:0046930; GO:0015288; GO:0004872
E3-E4_4.3	AIEC01 vs ECG11	3: 167, 173, 209	C/Y, Y/Y, T/K	Putative uncharacterised protein	PF06174	
E3-E4_4.4		84: 1126	R/R	dGTPase	PF00350	GO:0005525
E3-E4_4.7		80: 920, 932, 1013	S/S, S/S, Y/Y	Chemotaxis protein	PF13990	
E5-E6_3.1	AIEC07 vs ECG04	3: 6356	A/C	FimH	PF00419; PF09160	GO:0007155; GO:0009289
E5-E6_3.12		83: 442	A/G	Succinyl-CoA ligase subunit beta	PF08442	GO:0005524; GO:0000287; GO:0030145; GO:0004775; GO:0006099
E5-E6_3.16=3.22		51: 418, 544, 545, 633, 646, 650	Y/Y, S/S, R/R	Enterobacterial Ail/Lom family protein	PF06316	GO:0009279; GO:0016021
E5-E6_3.17		62:583	M/M, Y/Y, S/S	Putative prophage component	PF00877	
			R/R			

*1 A: adenine; C: cytosine; G: guanine; K: guanine or thymine; M: adenine or cytosine; R: adenine or guanine; S: guanine or cytosine; T: thymine; W: adenine or thymine; Y: cytosine or thymine.

Table 18. Prevalence of genes encompassing SNPs in a collection of AIEC/non-AIEC strains and the frequency of particular nucleotide variants with respect to AIEC phenotype and phylogroup origin of the strains. Only SNPs presenting statistically significant differences regarding pathotype are presented. Values are given in percentages with respect to the total number of AIEC or non-AIEC strains. Statistically significant differences for each variant are presented in bold type.*For phylogroup analysis, the atypical non-AIEC strain was discarded.

ID	PCR amplification		SNP base vs AIEC phenotype			SNP base vs phylogroup					
	AIEC (n=22)	non-AIEC (n=28)	Ntd (N strains)	AIEC	non-AIEC	p-value	A (n=9)	B1 (n=7)	B2 (n=28)	D (n=5)	p-value
E3-E4_4.3(2)	100	100	C (33)	45.45	82.14	<0.001	77.78	71.43	60.71	60.00	0.667
			T (9)	40.91	0.00		22.22	0.00	21.43	20.00	
			Y (8)	13.64	17.86		0.00	28.57	17.86	20.00	
E3-E4_4.4	77.28	71.43	A (5)	13.64	7.14	0.010	0.00	14.29	14.29	0.00	0.636
			G (14)	9.09	42.86		33.33	28.57	32.14	0.00	
			R (18)	54.55	21.43		33.33	14.29	42.86	40.00	
E5-E6_ 3.16=3.22(2)	90.91	89.28	C (14)	13.64	39.29	0.012	44.45	28.57	17.86	40.00	0.026
			G (8)	31.82	3.57		22.22	0.00	14.29	40.00	
			T (13)	36.36	17.86		22.22	0.00	39.29	0.00	
			S (5)	9.09	10.71		0.00	42.86	7.14	0.00	
			K (3)	0.00	10.71		0.00	0.00	10.71	0.00	
			Y (2)	0.00	7.14		11.11	14.29	0.00	0.00	

A: adenine; C: cytosine; G: guanine; K: guanine or thymine; M: adenine or cytosine; R: adenine or guanine; S: guanine or cytosine; T: thymine; W: adenine or thymine; Y: cytosine or thymine.

3.4 SNPs in relation to adhesion and invasion capacity

As expected, strains carrying SNP variants associated with the AIEC pathotype showed increased adhesion and invasion indices (Figure 15). The single exception was SNP E5-E6_3.16=3.22(2), in which increased adhesion did not reach statistical significance. In turn, strains with guanine in SNP E3-E4_4.4 showed the lowest adhesion and invasion indices. An additional polymorphism, SNP E5-E6_3.16=3.22(3), showed significant differences; strains with adenine in this SNP displayed increased invasive ability.

SNPs E5-E6_3.16=3.22(2) and E5-E6_3.16=3.22(3) are consecutive and result in the same amino acid change. To visualise their effects on amino acid sequence, we focused on the possible nucleotide combinations found across our *E. coli* strain collection. The combinations of the two changing positions lead to the possibility of 13 SNP variants that can be translated in 6 different amino acids. As expected, the guanine-adenine combination was associated with the highest invasion values ($0.63\% \pm 0.76$). Statistically significant differences according to pathotype were observed when the guanine-adenine combination (n=7) was compared with the cytosine-guanine combination (n=11) ($p=0.009$). The former leads to a basic amino acid at pH=7 (serine), whereas the latter encodes a neutral amino acid (alanine) that may affect the function of the protein.

3.5 Usefulness of SNPs as molecular signatures for AIEC screening

The use of a binary logistic regression model revealed two SNPs that are predictive of AIEC phenotype (Table 19). The SNP in E3-E4_4.4 can classify the strains as AIEC or non-AIEC with 73% global success, *E. coli* strains with a nucleotide base other than guanine at this position have a 65.2% of probability of exhibiting the AIEC phenotype, whereas those strains presenting guanine have only a 14.3% probability of exhibiting the AIEC phenotype. In the case of SNP E5-E6_3.16=3.22(2), global success was similar (68.9%), but only 35% of the AIEC strains were correctly classified.

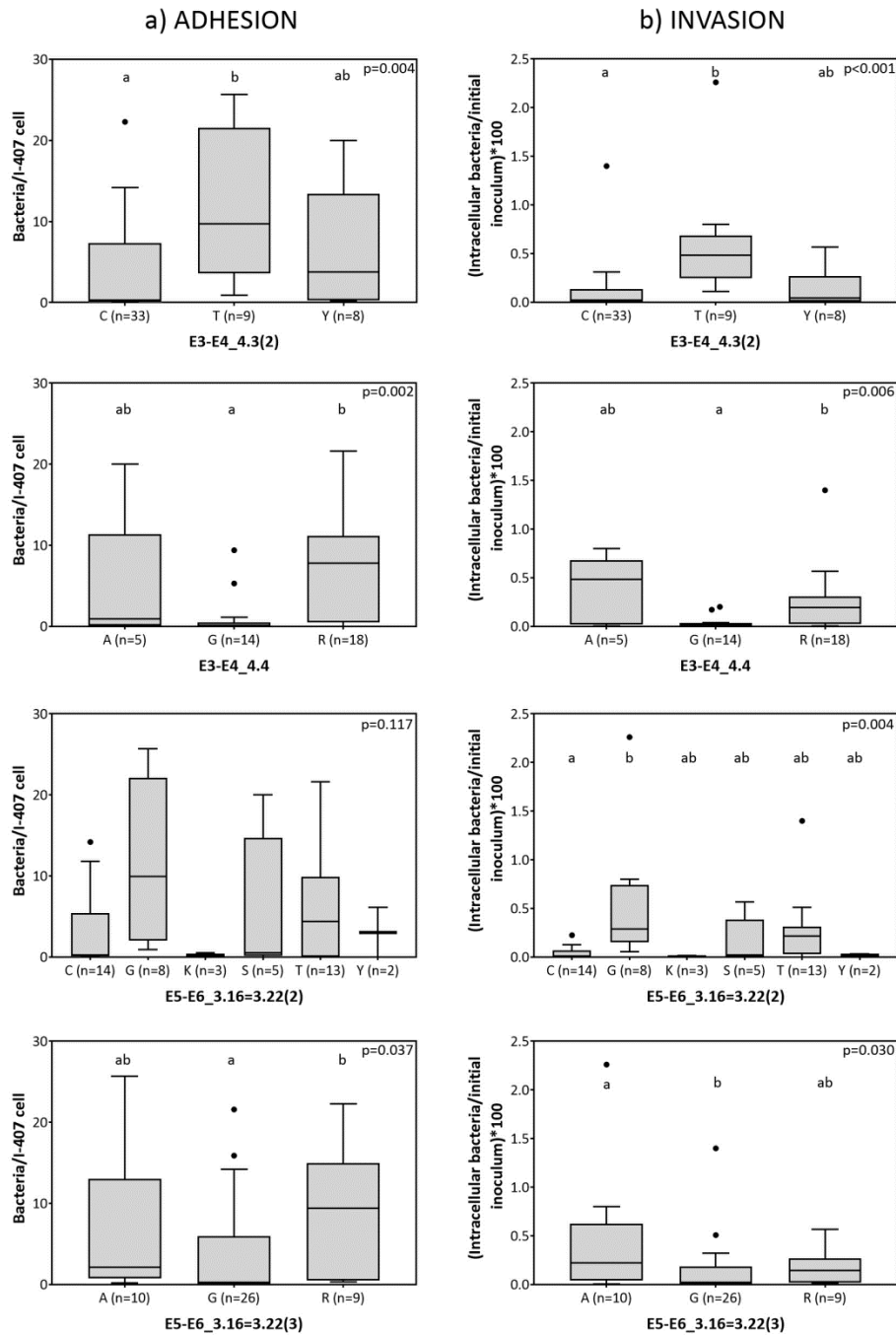


Figure 15. Adhesion (a) and invasion (b) abilities of the strains according to specific nucleotide variants of SNPs. Only SNPs associated with significant differences ($p < 0.05$ using the Mann-Whitney U-test) in the adhesion or invasion abilities of variants are shown. Homogeneous subgroups ($p > 0.05$) within each panel are indicated by the same superscripts. The median of the data is indicated by the horizontal line in each box, boxes cover the 25% and 75% quantiles, and bars show the 10% and 90% quantiles. Outliers are marked as dots.

Table 19. Binary logistic regression model for the SNPs associated with the AIEC pathotype. The equation variables, the risk of being AIEC (odds ratio), the p-value of the regression model and the percentage of successfully classified strains are indicated.

	Equation variables				Observed	Predicted			Global %
	B	p-value	Odds ratio	95% CI		Non-AIEC	AIEC	% Correct	
Not to have G in E3-E4_4.4	2.420	0.006	11.250	2.004-63.168	Non-AIEC	12	8	60.0	73.0
Constant	-1.792	0.019	0.167		AIEC	2	15	88.2	
To have G in E5-E6_3.16=3.22(2)	2.559	0.023	12.923	1.430-116.785	Non-AIEC	24	1	96.0	68.9
Constant	-0.613	0.075	0.542		AIEC	13	7	35.0	

Although the global success in AIEC prediction based on the SNP in E3-E4_4.4 was high, 40% of non-AIEC strains were misclassified as AIEC. To improve prediction specificity, we designed a classification algorithm based on the identification of the nucleotides present in three SNPs (Figure 16A). In this algorithm, the variant in SNP E3-E4_4.4 is first determined, and those strains containing a guanine are classified as non-AIEC with a percentage of success of 85.7%. Strains with another result in SNP E3-E4_4.4 (adenine, overlapping peak of adenine and guanine (R) or gene absence) are then analysed for the SNP E5-E6_3.16=3.22(2). The combined results obtained for both SNPs can classify the strains with a percentage of success that ranges from 71.4% to 100%, with the exception of isolates that present nucleotides different from guanine in both genes, which remain unclassifiable. For that reason, we included a third SNP (E5-E6_3.12). Despite not being more frequently found in AIEC considering the whole strain collection (39.3% of non-AIEC and 22.7% of AIEC presented adenine, whereas 60.7% of non-AIEC and 77.3% of AIEC presented guanine; $p=0.174$), this SNP was useful in classifying this particular group of strains (50% of non-AIEC and 0% of AIEC strains presented adenine, whereas 50% of non-AIEC and 100% of AIEC strains presented guanine; $p=0.036$). Overall, the classification algorithm displays 82.1% specificity, 86.4% sensitivity and 84.0% accuracy within our strain collection.

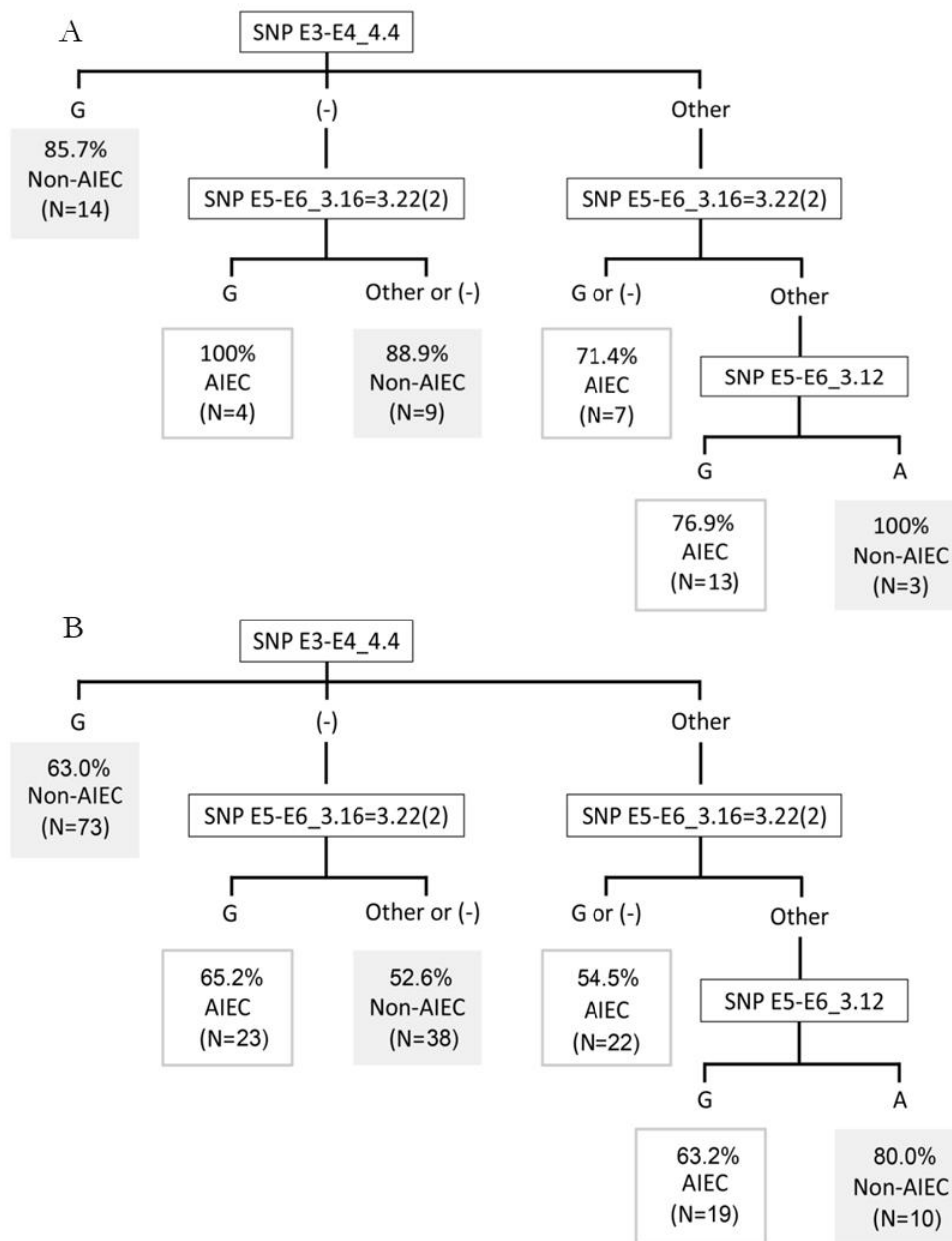


Figure 16. Classification algorithm for AIEC identification. A: assessed in Girona (Spain) strain collection. B: assessed in external strain collections (France, Chile⁷⁴, Spain (Mallorca)⁷⁴, Australia¹⁷⁸ and ExPEC-Spain^{20,236} and ExPEC-America²³⁵). Percentages represent the proportion of strains that are correctly predicted as AIEC or non-AIEC based on the result for each SNP combination. The number of total strains corresponding to each condition is indicated. (-): no amplification; other: a nucleotide different from guanine (G) or overlapping peaks.

3.6 Validation of the tool in external strain collections

Further confirmation of the validity of the tool presented in additional geographically distant strains has been performed. In total, 74/99 of the non-AIEC strains were correctly classified but only 39/86 of the AIEC were appropriately predicted, obtaining a high

probability to obtain false negatives. Therefore, in comparison to the values obtained within our strain collection (82.8% specificity, 86.4% sensitivity and 84.3% accuracy), the global accuracy was significantly reduced (61.1%), with decreased specificity (74.8%) and specially lower sensitivity (45.4%). SNPs previously found to be differentially distributed among our AIEC and non-AIEC strains (E3-E4_4.4 and E5-E6_3.16=3.22(2)) showed similar frequency values according to pathotype (SNP E3-E4_4.4, G: 64.8% of non-AIEC and 50.9% of AIEC, A: 11.3% of non-AIEC and 15.1% of AIEC, and R: 23.9% of non-AIEC and 34.0% of AIEC, $p=0.299$; SNP E5-E6_3.16=3.22(2), G: 19.2% of non-AIEC and 29.6% of AIEC, C: 28.8% of non-AIEC and 21.1% of AIEC, and others: 52.1% of non-AIEC and 49.3% of AIEC, $p=0.287$). Indeed, most AIEC strains were incorrectly classified because they presented G in SNP 4.4 or they did not have the gene where SNP 4.4 is located and then presented a result different from guanine in SNP 3.16=3.22(2) (Figure 16B). Other possible combinations of the SNPs were considered but none improved the precision of the algorithm. Nonetheless, once Spanish strains (Girona and Mallorca) (N=64) were considered the accuracy of the tool was maintained (specificity 82.9%, sensitivity 79.3% and accuracy 81.3%) (Table 20).

These results collectively suggest that although we were unable to find molecular signatures specific to AIEC and present in all AIEC, our approach identifies some genes with polymorphisms that may represent an advantage or disadvantage for the adhesion and invasion abilities of *E. coli* and may thus be involved in the AIEC phenotype. Moreover, we have designed a novel molecular strategy based on the identification of the nucleotides present in three polymorphic sites that although presented promising results in Spanish strains these were not maintained when strains from other countries were included.

Table 20. Accuracy of the algorithm in strain collections from diverse geographic origins.

	Observed	Predicted		Predictive Values	
		Non-AIEC	AIEC	% Correct	%Global
AIEC/non-AIEC Spain (Girona)^a	Non-AIEC	24	5	86.4	84.3
	AIEC	3	19	82.8	
AIEC/non-AIEC Spain (Mallorca)^b	Non-AIEC	5	1	83.3	69.2
	AIEC	4	3	57.1	
AIEC/non-AIEC France^c	Non-AIEC	0	0	0	32.3
	AIEC	23	11	32.3	
AIEC/non-AIEC Chile^b	Non-AIEC	3	0	0	33.3
	AIEC	6	0	100	
AIEC/non-AIEC Australia^d	Non-AIEC	12	8	60.0	42.4
	AIEC	11	2	15.4	
ExPEC Spain (Girona)^e	Non-AIEC	30	11	73.2	73.3
	AIEC	1	3	75.0	
AIEC/non-AIEC Spain (Girona) and ExPEC	Non-AIEC	54	16	77.1	79.2
	AIEC	4	22	84.6	
AIEC/non-AIEC Spain (Girona) and AIEC/non-AIEC Spain (Mallorca)	Non-AIEC	29	6	82.9	81.3
	AIEC	6	23	79.3	
All strains	Non-AIEC	74	25	74.8	61.1
	AIEC	47	39	45.4	

^aPresent study. It includes the K-12 strain. ^bCéspedes et al.⁷⁴. ^cUnpublished. ^dO'Brien et al.¹⁷⁸. ^eMartinez-Medina et al.²⁰, Bidet et al.²³⁵ and Blanco et al.²³⁶.

Discussion

The AIEC pathotype is of interest due to its association with gut inflammation in CD patients^{17,21,24,43,69,171,267}. At present, AIEC pathotype identification is conducted by phenotypic screening of cultured bacteria, which is an extremely time-consuming technique. Thus, the existence of a molecular tool for the specific detection of AIECs would be of great significance in facilitating, for instance, studies of AIEC distribution that define pathotype reservoirs and transmission paths.

In this study, comparative genomics between AIEC and non-AIEC strains that are considered clones with respect to their PFGE patterns has been performed for the first

time. In contrast to previous comparative genomics studies, our methodology precludes the detection of genome variations that are inherent, for example, in the phylogenetic origin of the strains. This approach may increase the chance of identifying molecular signatures that are specific to AIEC. In fact, previous difficulties in discovering the distinctive features of AIEC strains could occur because non-phylogenetically related AIEC and non-AIEC strains have usually been compared^{23,154} and most AIEC studied belonged to a particular phylogroup^{178,210}. We chose to sequence strain pairs that belong to different phylogroups (B1, B2 and D) and studied the distribution of the differences found in a collection that included isolates of different phylogenetic origin to determine whether they were universal among AIEC strains and absent from non-AIEC strains.

No significant differences in genome structure or even in gene content were found between AIEC and non-AIEC strains, confirming their close identity by PFGE. No gene was present in at least two AIEC strains and absent from all non-AIEC strains. However, small differences in gene content were observed between strains of the same pair. Nevertheless, this result should be confirmed because it could be a consequence of incomplete genome assembly (fragmented genes will not be found). In general, our results support the idea that the AIEC phenotype is not determined by the presence or absence of a particular gene, as O'Brien et al.¹⁷⁸.

Our analysis revealed no association between the presence of previously AIEC-associated genes (*lpfA₁₅₄*, *gipA*, *pduC*, *fyuA*, *afaC*, *chuA* and *ibeA*)^{50,74,81,136,153,154,209} and the AIEC phenotype, in concordance with previous observations^{154,178}. A similar situation was found with respect to the genetic variants of *fimH*, *ompA* and *chiA*. Although specific changes in the amino acid sequences of proteins encoded by these genes have been associated with higher adhesion/invasion capacity^{85,159,165}, we did not observe such differences between the isolates of our strain pairs. Therefore, differences in the sequences of these genes might not determine the pathotype of the strains. However, as Dreux et al.¹⁶⁵ suggested, gene expression should be evaluated in depth because it could also be involved in the determination of the phenotype.

In a further attempt to explain the observed phenotypic differences between pairs, we sought to identify SNPs that were differentially present in AIEC strains and their non-AIEC counterparts. The rate of occurrence of such SNPs that were further validated by Sanger sequencing was low. This result can be explained by the accumulation of small errors during library construction and sequencing caused by the imperfect fidelity of DNA

polymerases and the intrinsic error rate of the sequencing platform, as well as errors derived from the parameters used for the assembly of the reads^{268,269}. Occasionally, there is variability and the nucleotide that appears in the consensus assembled contig does not represent all the reads. Although resequencing by the Sanger method is not frequently used to validate NGS data, our results are consistent with the results of a previous work²⁷⁰. In light of the evidence, we highly recommend confirmation of NGS data, especially in SNP research. Another intriguing result was the presence of genes with intraclonal polymorphisms. Bacterial cultures grown overnight may represent genetically heterogeneous populations²⁷¹.

The distribution of the polymorphisms in our strain collection indicated the absence of a particular nucleotide from any of the SNP positions that was present in all AIEC strains and differed from the base found in that position in non-AIEC strains. However, we found SNP positions that presented differences not only in the distribution of nucleotide variants according to pathotype (E3-E4_4.3(2), E3-E4_4.4 and E5-E6_3.16=3.22(2)) but also in their association with adhesion and invasion indices (E3-E4_4.3(2), E3-E4_4.4, E5-E6_3.16=3.22(2), and E5-E6_3.16=3.22(3)). From a functional point of view, the E5-E6_3.16=3.22 gene encodes a protein of the enterobacterial Ail/Lom family. This protein family includes outer membrane proteins involved in bacterial virulence, such as OmpX (in *E.coli* and *Enterobacter cloacae*)²⁷² and PagC (in *Salmonella typhimurium*)²⁷³. These proteins play roles in cell adhesion and intramacrophage survival, respectively. The SNP-containing gene E3-E4_4.4 encodes a dGTPase that is a member of the dynamin-like protein family (PF00350), whose function in bacteria is poorly understood. Finally, E3E4_4.3 is of unknown function, yet it shares 97.2% (74.5% coverage) amino acid sequence homology with the hypothetical protein yeeT. The low sequence homology of these genes with genes in the currently available databases makes it difficult to identify the proteins encoded by these genes with confidence. Therefore, the creation of isogenic mutants will be needed to further understand the biological function of these proteins and demonstrate the effects of their possible amino acid variations.

Interestingly, two of the identified SNPs were adequate for the prediction of the AIEC phenotype as determined by the binary logistic regression model. Moreover, we present here a classification algorithm that combines three SNPs, which allows the classification of phylogenetically diverse *E. coli* isolates with a high accuracy rate. Using this algorithm, 84% of our *E. coli* strains were correctly classified as AIEC or non-AIEC. Of note, the AIEC

12–1 ti12 and non-AIEC 12–2 ti13 strains sequenced by O'Brien et al.¹⁷⁸, both strains isolated from the same patient and with the same ST (ST127), are correctly classified with the molecular tool presented here. Since the application of a molecular tool could assist in overcoming the problem of AIEC identification, we further tested the specificity and sensitivity of the tool in additional geographically distant and phylogenetically diverse AIEC strains, as well as, ExPEC strains. Unfortunately, the predictable values of the tool decreased considerably (61.1% of accuracy), indicating that this algorithm may be only suitable for Spanish strains (Girona and Mallorca) (81.3% accuracy). To our knowledge, only one reported study has searched for SNPs in the entire genome of AIEC using pathogenic and non-pathogenic *E. coli*²¹⁰. The authors reported 29 diagnostic SNPs that can differentiate a group of AIEC. However, only four AIEC strains were included in their study and the SNPs detected were also present in ExPEC, specifically in uropathogenic *E. coli* and avian pathogenic *E. coli*. Hence, no specific polymorphism that exclusively differentiates AIEC strains from non-AIEC or other pathogenic *E. coli* strains has yet been described.

In conclusion, even though no genetic element could be designated specific for AIEC classification, our data reveal three SNPs that could assist in AIEC identification. Although this tool does not correctly classify all *E. coli* strains, its accuracy is very high (84%), and no comparable molecular tools currently exist. In contrast to classical cell culture infection-based assays, this approach could represent a rapid and standardisable method for detecting AIEC from *E. coli* isolates. However, the presented tool is not universal since its accuracy was reduced to 61.1% once a larger strain collection from different geographic locations and pathotypes was screened. Further studies are needed to demonstrate or rule out the role that the variants reported in this study play in the AIEC phenotype. Taken together, the results of this study provide meaningful information that contributes to our understanding of AIEC genomics.

Chapter 2.2: Construction of isogenic mutants to study the role in pathogenicity of three genes related to AIEC pathotype

Results

4.1. Construction of isogenic mutants

The construction of isogenic mutants of the three different genes was generated by using the red recombinase system described by Datsenko et al.²³⁷ and Chaverocche et al.²³⁸. This method consists on the electroporation of a PCR fragment carrying an antibiotic resistance gene flanked by regions homologous to the target locus to a recipient strain expressing the highly proficient homologous recombination system encoded by plasmid pkD46. This plasmid, under the presence of arabinose, expresses the genes necessary to induce homologous recombination and, thus to replace the desired gene by the kanamycin resistance cassette amplified from the plasmid pkD4.

As our final goal was to test whether the genes harbouring SNPs associated with adhesion and invasion had an implication on the AIEC phenotype or not, we selected one strain from our collection for each gene that presented the nucleotide associated with AIEC pathotype (Table 10).

First of all, it was confirmed that the bacterial strains selected were not resistant to any of the two antibiotics required for the mutant construction protocol. Bacteria were grown in LB supplemented with 100 µg/mL of gentamicin or 50 µg/mL of kanamycin at 37°C without agitation overnight. No growth was perceived in any of the cultures. Then, selected strains were electroporated with the plasmid pkD46, which is temperature sensible and gentamicin resistant. Thus, colonies grown after overnight incubation at 30°C in LB supplemented with 100 µg/ml gentamicin acquired the pkD46 plasmid.

In order to replace the gene of interest for an antibiotic resistance gene, a PCR fragment was created by two strategies (methods section 4.3). In the first approach, a fragment of 1579 bp was obtained for each construction, as 100 bp were added next to the kanamycin resistant gene which was 1479 bp length (Figure 17A). The second approach was performed for 4.3 and 4.4 genes inactivation, in order to increase the chance of recombination in the desired place since we had no success with the first approach for these two genes. PCR1 and PCR3 products in agarose gel are represented in Figure 17B.

Finally, in the third step, a complete linear DNA was obtained, yielding a 2.5kb DNA fragment approximately (Figure 17C).

Strains harbouring pkD46 plasmid were prepared for electroporation and mutagenesis was carried out by electroporating the PCR fragment previously created. The replacement of the gene by the kanamycin resistance gene in each colony was verified by PCR using the verification primers depicted in Table 11. The first approach efficiently replaced the 3.16 gene and generated the 3.16 isogenic mutant. In this case, by PCR we could verify the gene replacement, as the PCR of the wild type LF82 strain resulted in a band of 1054 bp whereas the PCR fragment of the mutant was of 1833 bp (Figure 17D). The mutant PCR product was larger because the length of the kanamycin cassette is bigger than the target gene. Additionally in both cases we could observe a band of 400 bp approximately which is the result of the primer imprecision. Bioinformatical analysis revealed that the primers used for verification anneal also with the pkD46 plasmid sequence (Accession number: AY048746.1).

On the other hand, for the 4.3 and 4.4 mutants, although many colonies grew in plates, any of the isolates replaced the gene of interest. Therefore we increased the homologous region to more accurately direct the place of recombination (Second approach). However, no recombinant colony was found with this strategy. Each approach was conducted more than ten times and protocol modifications to increase recombination were performed, consisting on increased DNA quantity, temperature and time for recombination but there was no success.

4.2 Phenotypic characterization of LF82 Δ 3.16

Isogenic mutant of the AIEC reference strain LF82 regarding the 3.16 gene, which encodes for an outer membrane protein, was successfully performed. Thus, we aimed at examining whether this gene played a role in AIEC virulence properties or not. No differences were observed in the growth rates of the mutant and wild-type bacteria in LB culture medium. LF82 Δ 3.16 mutant adhesion and invasion were similar to LF82-WT abilities (Figure 18A and B). Moreover, LF82 Δ 3.16 mutant presented similar replication levels both in J774 and THP-1 cells as the LF82-WT (Figure 18C and D).

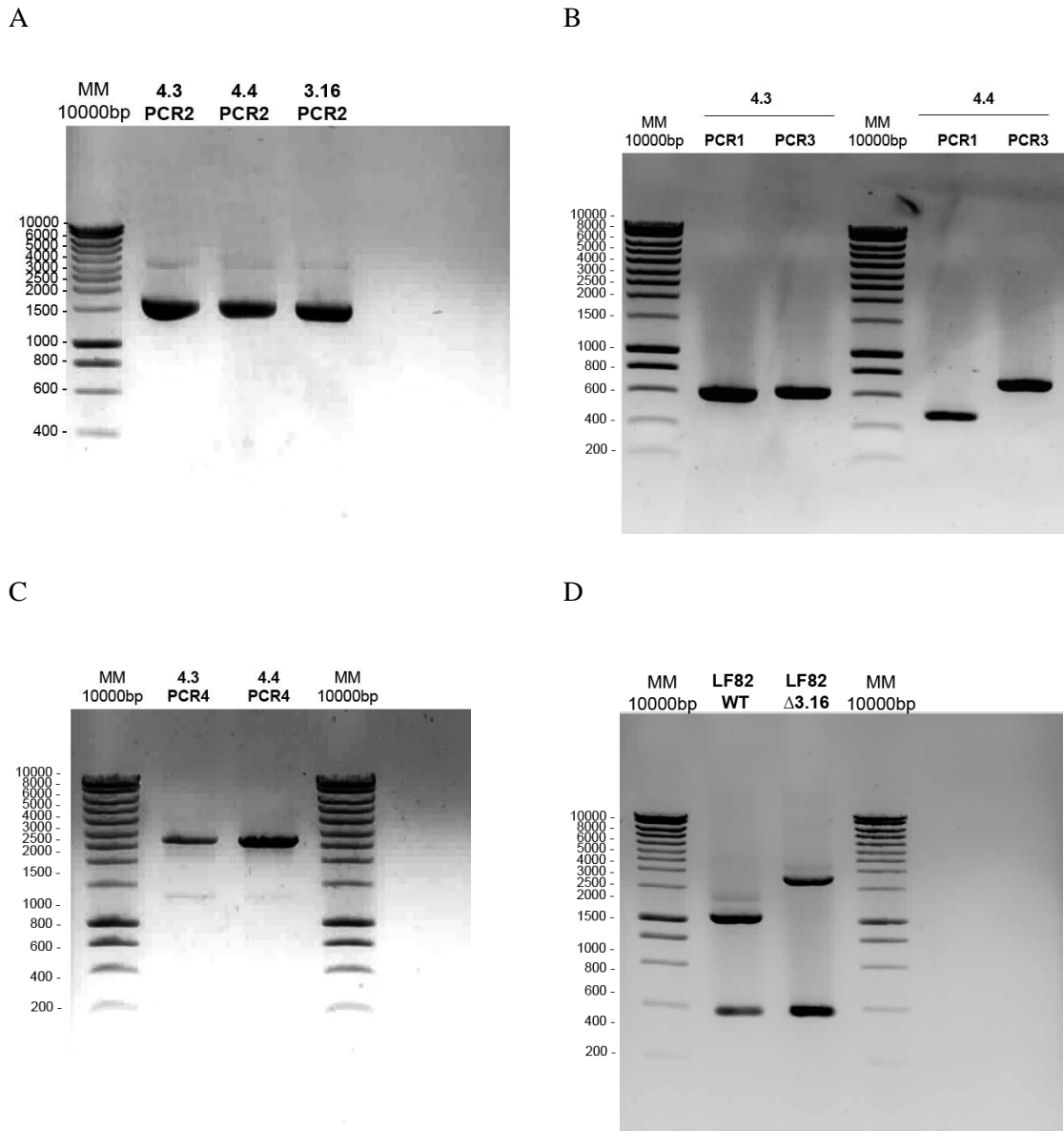


Figure 17. Agarose gel electrophoresis (1%) of PCR products obtained in each PCR reaction. A: PCR purified product resulting from PCR2. Each consists in the kanamycin resistance gene with 50 nucleotides in each extreme that are homologous to the adjacent region of the gene wanted to delete. B: Result of PCR carried with primers of PCR1 and PCR3 for each gene. C: Final PCR product obtained with PCR4 for each gene. D: Result of the PCR verification of the 3.16 isogenic mutant.

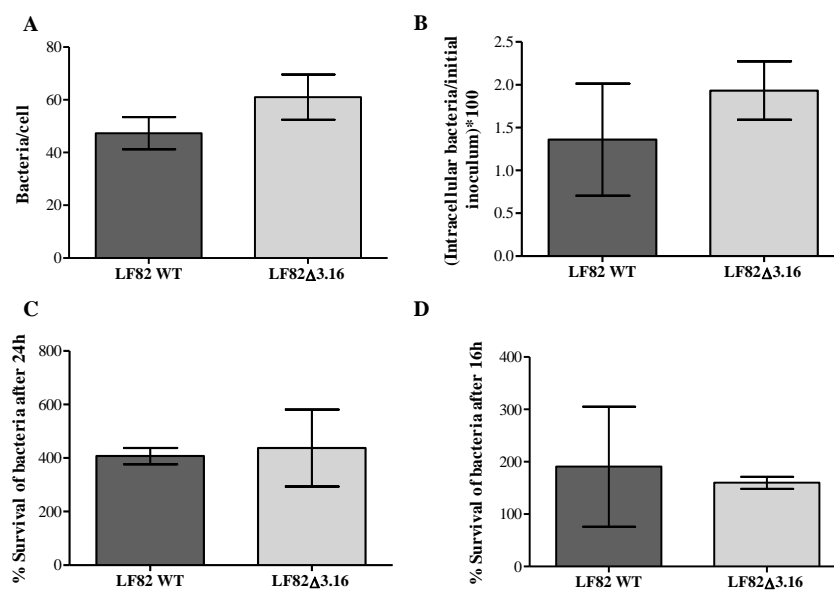


Figure 18. Phenotypic characteristics of the mutant in comparison to the wild type strain. A: Adhesion to I407 cells. B: Invasion of I407 cells. C: Replication inside J774 cells. D: Replication inside THP-1 cells.

Discussion

In this report, we have applied two approaches to generate isogenic mutants of three genes containing SNPs previously associated with AIEC pathotype. The 50 bp homology approach was sufficient to obtain the 3.16 mutant in the AIEC reference strain (LF82) but, it was not appropriate to create isogenic mutants of the 4.3 and 4.4 genes, neither the 500 bp homology approach allowed obtaining isogenic mutants of these genes. We suspected that the presence of false-positive isolates was due to erroneous recombination place of the PCR product. Failure to obtain these mutants can be explained by different hypothesis. First (I), it might be that there are multiple copies of the target gene. We previously found that in the genomes of some strains these genes are present in two copies, being each copy surrounded by different genes. As the genome sequences of the selection strains are of our disposal, blast analysis were performed and only one copy of the target gene was found and primers were designed accordingly to target this specific place (Figure S7 and S8). Nonetheless, the genomes of the selected strains are not circular and closed so we cannot discard the possibility of dual gene existence. Another possibility would be that these genes are transposable elements (II), such as transposons and its instability in the genome make impossible the deletion of the target gene. Or that, the contiguous regions of the target genes contains transposable elements, and then no homologous recombination could occur. According to the genome annotation, upstream of the 4.3 gene there is an antitoxin

system (YeeV-YeeU) (UniProt:A0A067HQT2) and downstream one uncharacterised protein (UniProt:A0A0H2V897). Similarly, an uncharacterised protein (UniProt:A0A0P0SR93) is located upstream of the 4.4 gene and the downstream englobes intergenic region. Therefore, this possibility could not be discarded and should be tested. To identify transposon insertion sites some bioinformatical tools (such as ISFinder) and sequencing techniques (such as TraDis) exist. The latter could determine gen essentiality and would be more accurate and robust than the bioinformatical tools which rely on prediction algorithms. Finally (III), although there is a tiny probability, another putative reason why no recombinant colonies were obtained could be that the deletion of the target gene causes bacterial lethality.

Regarding the implication of 3.16 gene on AIEC phenotypic features no substantial effect were observed in the LF82Δ3.16 mutant in comparison with LF82-WT. Contrary to some members of the Ail/Lom family proteins, which have been involved in *E. coli* virulence (i.e. OmpX)²⁷², these results suggested that either there is no implication of 3.16 gene on AIEC phenotype or that the function of the deleted gene is masked by other highly AIEC-related genes, perhaps the *fimH* gene which strongly modulates bacterial adhesion and invasion to I407 cells in AIEC strains¹⁵⁶. The construction of double isogenic mutants or the use of mannose-derived *fimH* antagonists during adhesion and invasion assays could provide evidence of this statement.

Bearing in mind that isogenic mutants for 4.3 and 4.4 genes were not obtained and disruption of the 3.16 gene did not result in any perceivable effect on AIEC phenotype, site-directed mutants were not conducted. Thereby, we could not provide evidence about the previously postulated (chapter 2.1) effect of the SNPs present in 4.3, 4.4 and 3.16 genes on AIEC phenotype. Nonetheless, new information about the 3.16 gene has been provided. The 3.16 gene encodes for an outer membrane protein of unknown function, this might be also related to many other processes, such as antibiotic resistance and indirect adhesion processes. For that reason, we suggest further functional studies exploring its putative role in AIEC phenotype by applying other techniques, for instance interaction with Peyer's patches, biofilm formation or invasion assays using inhibitors of adhesins. Additionally, as AIEC strains are genetically variable, the deletion of 3.16 gene in other strains in order to confirm that this gene does not affect the AIEC characteristics would be recommended.

Chapter 3.1: RNA-Seq analysis of the transcriptome during growth in cell culture media and during intestinal epithelial cell infection of AIEC in comparison with non-AIEC strains

Results

5.1 Protocol optimization

Given that the aim of this work was to sequence the bacterial transcriptome during the infection of human cells, without having to sequence both host and bacteria transcriptomes, we planned an experimental approach to enrich the bacterial mRNA from bacterial and human total RNA mixture. From each experiment two portions of the sample were extracted: (I) the supernatant (SN), mainly enriched in bacteria that are not adhering nor invading eukaryotic cells, but also some dead eukaryotic cells might be present; and (II) the experimental condition named invasion (INV), containing eukaryotic cells and the adhered and intracellular bacteria.

Optimization of the cell infection and total RNA extraction protocol was conducted with the AIEC17 strain. We tested different quantity of cells, MOI, kit of RNA extraction, elution buffer and washing steps (Table 21) in order to achieve the RNA quantity required for the next RNA isolation step (MICROBEnrich), which was 25 µg of total RNA in 30 µl. In the first approach (A), a 100% confluent T75 flask (2×10^7 I407 cells) was infected at MOI 100 and RNA extraction was carried out with the RiboPure Bacteria kit. Not enough RNA quantity was recovered for any of the fractions (SN and INV) assessed. As the cell quantity exceeded the limit threshold of the kit, the same procedure was performed but with 50% and 5% of the original sample (approach B and C, respectively). A proportional RNA quantity reduction was reported, indicating that the initial input may not be saturating the kit. In the approach D, a 5-fold cell quantity increase was performed while MOI was reduced to avoid exceeding the bacteria quantity of the kit (1×10^9 cfu). In this case, the RNA amount obtained was less than the half of the approach A, thus suggesting that the kit was oversaturated. Then, the elution buffer provided with the kit was used (approach E) as it was hypothesised that it might increase the efficiency and elution step. However, much less RNA quantity was obtained in comparison to approach A. Since the RiboPure kit was prepared for extracting RNA from pure bacterial cultures with columns, we suspected that eukaryotic cells may prevent this process. Thus, we decided to change the RNA extraction kit for the TRIzol Max Bacterial Isolation kit with Max Bacterial Enhancement reagent

which consists on phenol- and chloroform- based reagent and it can be applied in any type of sample. In this case (approach F), the Trizol kit was applied with the maximum cell quantity previously assessed and better RNA quantity was achieved but RNA degradation occurred. Again, the initial input was much higher than the limit of the kit (1×10^8 cfu), so the initial sample was reduced 5-fold and added a washing step with PBS. This washing step was added to eliminate the remaining solutions that may interfere with the extraction process, such as the cell culture medium. In this case (approach G), the RNA quantity was sufficient for to proceed to the following step and the RNA quality was adequate (main rRNA subunits were perceptible), although with higher quantity of small RNA residues in comparison with the previous kit used (Figure S9).

Table 21. Experimental approaches followed to optimise the total RNA extraction during AIEC17 growth in cell culture media and during infection of intestinal epithelial cells (I407).

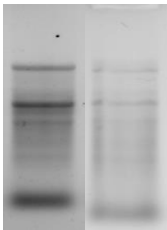
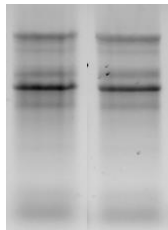
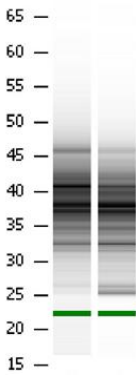
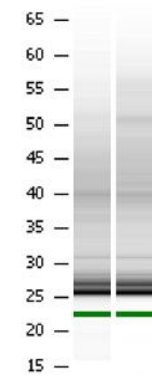
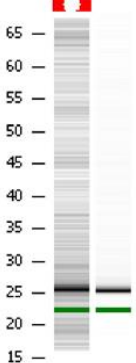
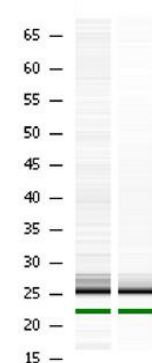
APPROACH	A	B	C	D	E	F	G
I407 cell number	2×10^7	1×10^7	1×10^6	1×10^8	2×10^7	1×10^8	2×10^7
MOI	100	100	100	10	10	100	100
Pellet wash step	No	No	No	No	No	No	With 500µl of PBS
Kit	RiboPure	RiboPure	RiboPure	RiboPure	RiboPure	Trizol	Trizol
Volume of elution	100µl with TE ^a	50 µl with TE ^a	50 µl with TE ^a	50 µl with TE ^a	100 µl with EB ^b	50µl with TE ^a	50µl with TE ^a
SN RNA Quantity / Quality	5.45 µg ^c / Good	-	-	1.91 µg / Good	3.01 µg / NA	15.75 µg / Bad	53.25 µg / Good
INV RNA Quantity / Quality	13.85 µg / Bad	6.95 µg / NA	1.06 µg / NA	2.87 µg / Good	0.41 µg / NA	21.50 µg / Bad	149.13 µg ^d / Good

^aTE: 10mM Tris-HCl + 1mM EDTA pH8. ^bEB: Elution buffer of the kit. ^cThis extraction was performed from a bacterial culture only grown in cell culture media alone. ^dAfter DNase treatment the sample volume increased to 100 µl. NA: Not analysed. Good: rRNA perceptible in the agarose gel. Bad: Undetectable bands.

The final protocol consisted in total RNA isolation with TRIzol of a T75 flask sample infected at MOI 100 (approach G), eukaryotic RNA depletion and rRNA elimination. Results of RNA concentration and quality for each step are summarised in Table 22. Two samples are shown as an example of the results of agarose gels and the Bioanalyzer. Good quality results were obtained after the first step. When eukaryotic RNA depletion was done, the gel image obtained with Bioanalyzer reported partial degradation of SN samples (Table 22 and Figure 19A) whereas INV samples showed elimination mainly of 18S rRNA

although small portion of 28S was kept (Table 22 and Figure 19B). The treatment of these samples with Ribo-Zero kit yielded total RNA amounts ranging between 64-464 ng (present in 8 µl). When checked by the Bioanalyzer Nano chip total removal of rRNAs was detected, however, some small RNAs were kept in the sample, both in the SN and the INV fractions (Figure 19C).

Table 22. Example of qualitative and quantitative data at each step of the RNA extraction procedure of AIEC/non-AIEC samples for both fractions (Supernatant (SN) and Invasion (INV)). Sample concentration determined by Qubit Fluorometer with RNA HS Assay kit (minimum and maximum values), 260/280 ratio assayed by Nanodrop and gel image for both fractions tested are shown. In each procedure step, two samples are shown, being one from an AIEC strain (left) and the other from a non-AIEC strain (right).

Procedure steps	Supernatant section (SN) (mainly bacteria)	Invasion section (INV) (eukaryotes plus the adherent-invasive bacteria)
Total RNA Extraction (TRIzol Max Bacterial RNA Isolation kit – between 50 and 400 µl of final volume)	627 - 2310 ng/µl 260/280 ratio = 2.04-2.18 	827 - 4460 ng/µl 260/280 ratio = 2.06-2.13 
Eukaryotic mRNA and rRNA depletion (MICROBEnrich kit– 30 µl of final volume)	256 – 504 ng/µl 260/280 ratio = 1.93-2.11 	115 – 252 ng/µl 260/280 ratio = 2.07-2.09 
Removal of rRNA, tRNA, and mitochondrial RNA (Ribo-Zero Magnetic Gold kit epidemiology– minimum 6 µl of final volume)	8 – 47 ng/µl* 	16 – 58 ng/µl 

*The sample labelled in red had 8 ng/µL, a concentration below the Nano chip quantitative and qualitative range (25-500 ng/µl).

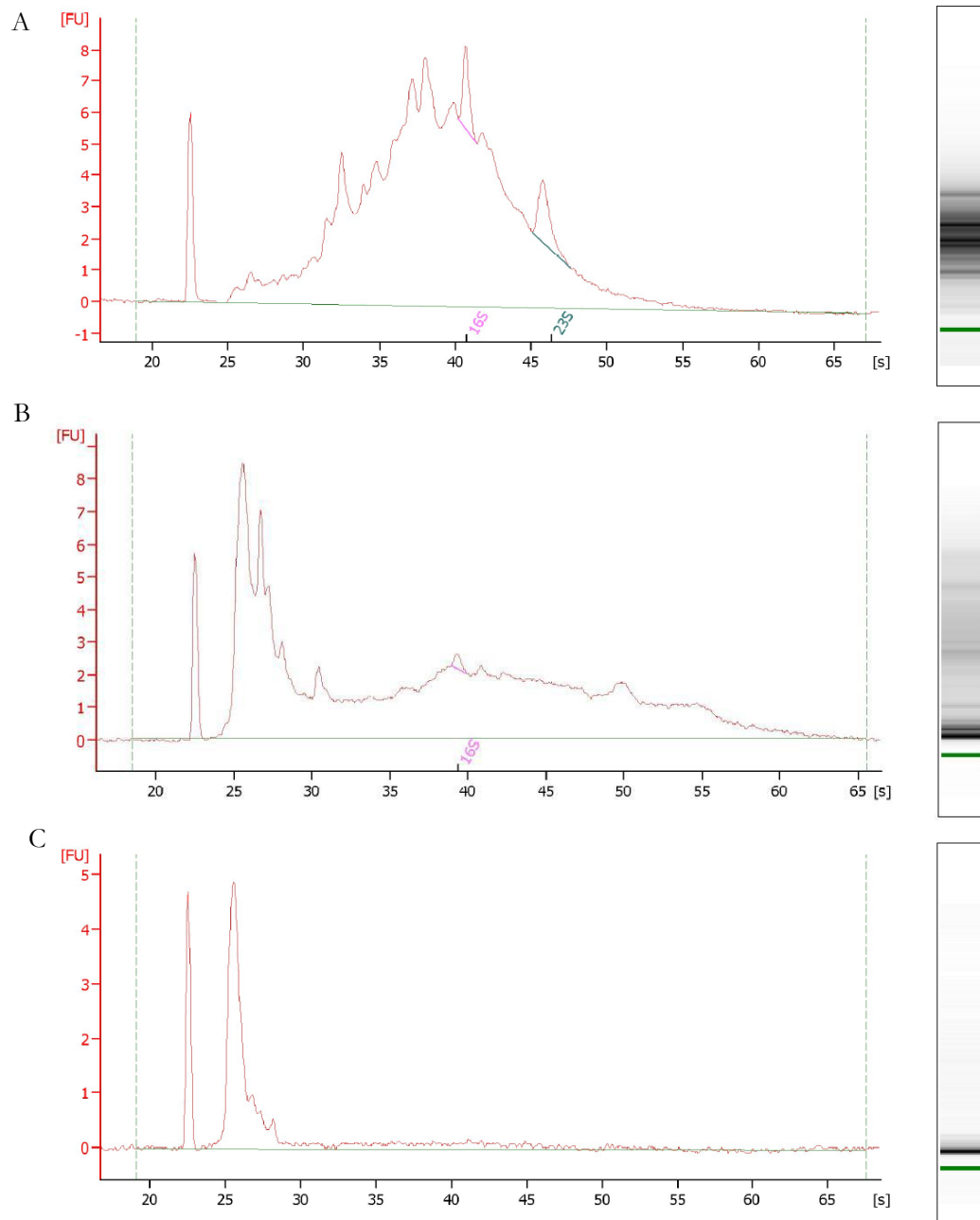


Figure 19. Example of electropherogram. A: obtained from the supernatant fraction after the MICROBEnrich kit, RIN: 3.1; B: obtained after the MICROBEnrich kit from the invasion fraction, RIN: 2.5; C: obtained from the invasion fraction after MICROBEnrich kit and rRNA depletion with Ribo-Zero Magnetic Gold kit (epidemiology), RIN: 2.6.

5.2 RNA-seq analysis to detect differential expressed genes

5.2.1 Overview of transcriptomics analysis

The global transcriptomes of two AIEC and two non-AIEC strains were determined during growth in cell culture media (SN) and during infection of IECs (INV). Samples were trimmed and for those samples corresponding to the SN fraction between 16.0 and

41.9 million reads were obtained while the INV samples ranged from 315.6 to 379.6 million reads (Table 23). The percentage of reads that mapped to the UM146 reference genome sequence oscillated from 23.20 to 80.20 % in the SN samples and from 1.50 to 8.00% in the INV ones. The bacterial reads in SN samples covered the 39.9-84.9% of the bacterial genome while the INV samples spread across the 82.8-90.9%.

Table 23. General characteristics of AIEC and non-AIEC transcriptomes.

Sample	Total Reads (Million)	Million Reads mapping UM146 genome (%)	UM146 genome coverage (%)
AIEC17.SN.1	16.2	12.7 (80.2)	66.3
AIEC17.SN.2	32.7	25.3 (80.0)	70.7
ECG28.SN.1	32.7	19.6 (61.5)	39.9
ECG28.SN.2	34.7	18.1 (54.3)	79.0
AIEC07.SN.1	33.9	9.2 (27.8)	81.7
AIEC07.SN.2	52.0	23.3 (45.0)	80.3
ECG04.SN.1	22.4	13.9 (66.0)	84.9
ECG04.SN.2	41.9	9.5 (23.2)	81.9
AIEC17.INV.1	315.6	0.9 (1.5)	90.9
ECG28.INV.1	361.9	6.0 (4.1)	82.8
AIEC07.INV.1	343.8	8.5 (8.0)	84.6
ECG04.INV.1	379.6	10.2 (6.8)	83.4

5.2.2 Differentially expressed genes during AIEC growth in supernatants and during IECs infection, relative to its non-AIEC counterpart

Preliminary differential expression analysis of the two AIEC/non-AIEC pairs in two conditions reported a total of 67 DEGs between the two pathotypes (Figure 20, Figure 21 and Table S24). Most of them (N=48) were under-expressed and 19 were over-expressed in AIEC strains. In the SN comparison between AIEC17 with ECG28, 24 genes were detected (17 under-expressed and 7 over-expressed in AIEC) while in INV, 22 were found (20 under-expressed and 2 over-expressed). In SN AIEC07-ECG04 comparison, 6 genes were reported (all under-expressed) and 15 genes were found in INV comparison (5 under-expressed and 10 over-expressed). In terms of gene expression levels between conditions, the most extreme log₂ fold change values were encountered in the INV comparisons (Figure 20). The minimum log₂ fold change value of the AIEC17-ECG28 SN comparison was -1.47 and the maximum was 3.01 while in the INV comparison it was -5.58 and 11.23 respectively. Similarly occurred for AIEC07-ECG04 where the minimum was 1.10 and -7.83 and the maximum 7.89 and 11.16 in the SN and INV comparison respectively. Indeed, this observation suggested that AIEC and non-AIEC strains are more similar during growth

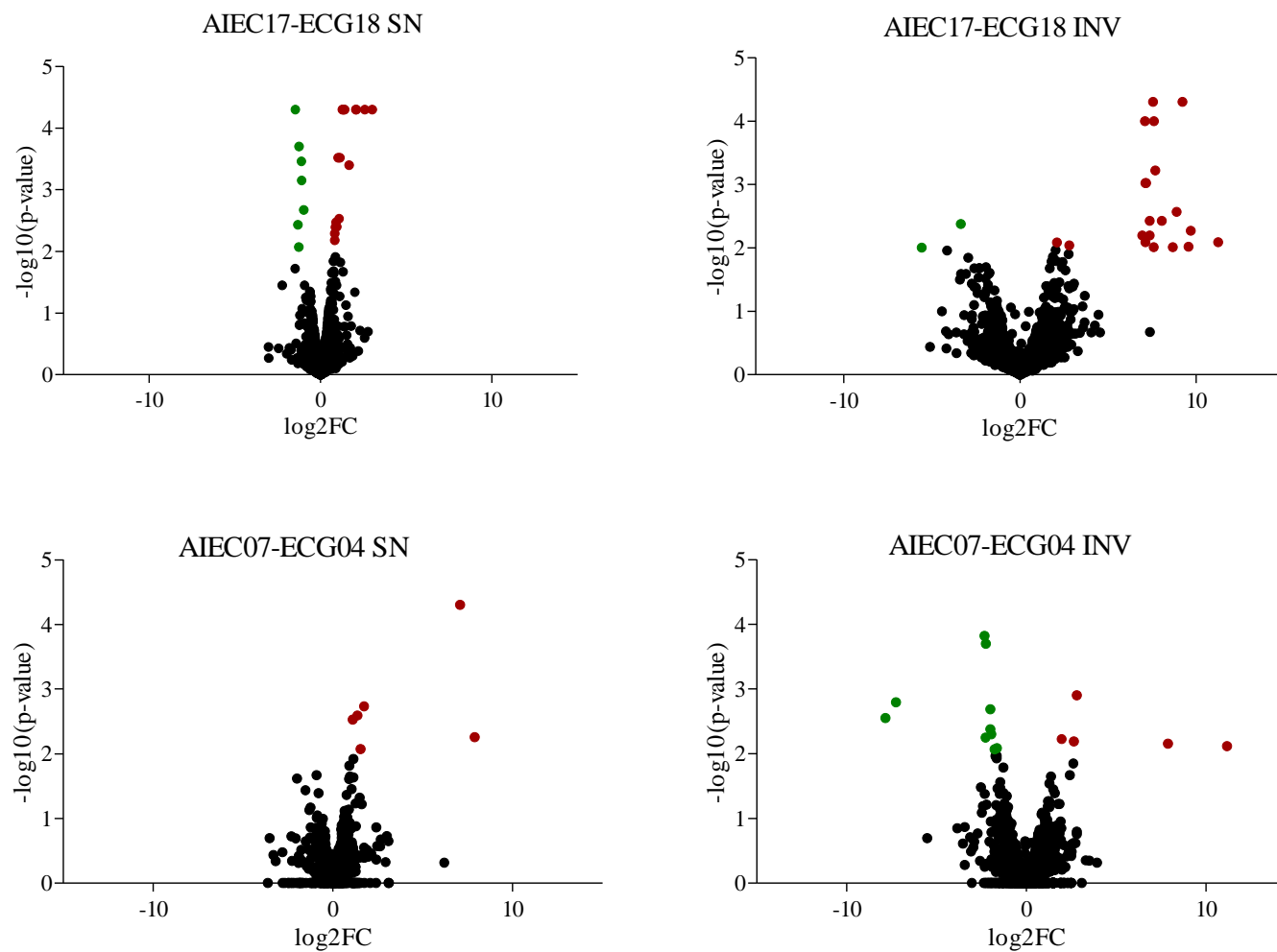


Figure 20. Volcano plots of the $-\log_{10}(\text{p-values})$ versus the \log_2 Fold-Change (FC). Results of all four comparisons are given. AIEC over-expressed genes are represented in green, those under-expressed in AIEC are coloured in red and those in black are the ones with a p-value > 0.01 .

with cell culture media than during IECs infection. In addition, for all the comparisons, a group of unknown genes (15/26) presented the highest FPKM values (>4915) (Figure 21).

Four genes were detected in two different comparisons however there was no gene differentially expressed in the same condition independently of the strain pair analysed (Figure 21). While XLOC_000511 was down-regulated in both AIEC17-ECG28 comparisons, the rest presented different patterns of expression between SN and INV. XLOC_001815 and XLOC_002831 were found in the AIEC17-ECG28 SN and AIEC07-ECG04 INV comparisons. The first was under-expressed in AIEC17 growing in suspension and over-expressed in AIEC07 during invasion, the latter was the opposite. Interestingly, XLOC_00794 was less expressed in AIEC07 in SN in relation with ECG04 SN and more expressed in AIEC07 INV in comparison with ECG04 INV. Apart from XLOC_002831 which obtained a maximum of 171 FPKMs, the other 3 genes were highly present in our samples (> 9790 FPKM).

Functional analysis suggested a representative difference in terms of gene function categories among the comparisons performed (Figure 22). Overall, the most abundant category was that including genes of unknown function (21/67), followed by those related with metabolic processes (18/67). For the AIEC17-ECG28 pair, genes differentially distributed in the SN fraction were mainly genes involved in metabolic process, 6/24 were over-expressed and 4/24 were under-expressed in the AIEC. Also, 9/24 genes of unknown function were under-expressed in the AIEC. In the INV fraction 11/22 genes were down-regulated in the AIEC, 7 genes were categorised as unknown function and 4 as genes involved in metabolic process. Moreover, one and two genes were related with adhesion and cell division respectively, both under-expressed in the AIEC, and one gene that was included in the 'regulatory functions' category was over-expressed in the AIEC. For the AIEC07-ECG04 pair, genes differentially distributed in the SN fraction were mainly genes of unknown function (5/6) and one was related with adhesion, all under-expressed in the AIEC. In the INV fraction, genes were distributed across 7 functional categories. The most common categories were: over-expressed genes of unknown function (5/15), over-expressed adhesion-related genes (3/15) and under-expressed genes associated with metabolic processes (3/15).

Moreover, following a bibliographic research, we found that 22 out of the 67 genes differentially expressed may be related with bacterial virulence (see references in Table S24).

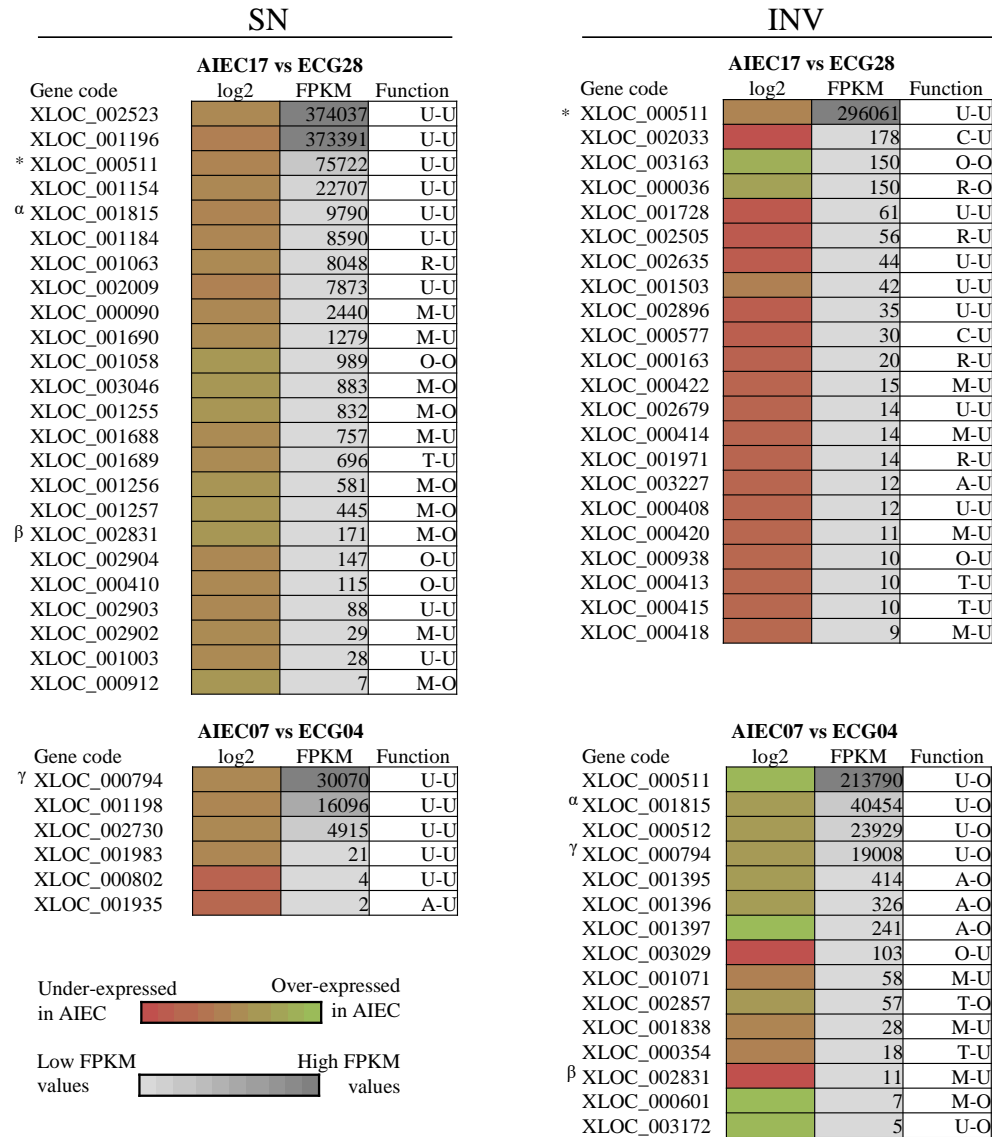


Figure 21. Differentially expressed genes during AIEC growth in suspension (SN) and during IECs infection (INV), relative to its non-AIEC counterpart. Log2 fold change, total FPKM and functional category are depicted. The red and green colours display the log2 fold change. FPKM values are indicated in grey. Symbols point out genes found in two comparisons. The first letter of the last column depict gene function: A: adhesion, C: cell division, M: metabolic, R: regulatory function, T: transport, O: others, U: unknown. The second letter indicated gene expression: O: over-expressed, U: under-expressed.

The expression of genes that were previously related with the invasive capacity of LF82 strain^{18,154,159,160,165} was checked in our samples (Table S25). For *fis*, *fucO*, *fucA*, *fimH*, *ompA* and *ompC* genes, similar gene expression values between AIEC and non-AIEC strains were obtained for all comparisons (SN and INV) assessed. In exception, no *hpfA* gene expression values were detected in any of the samples.

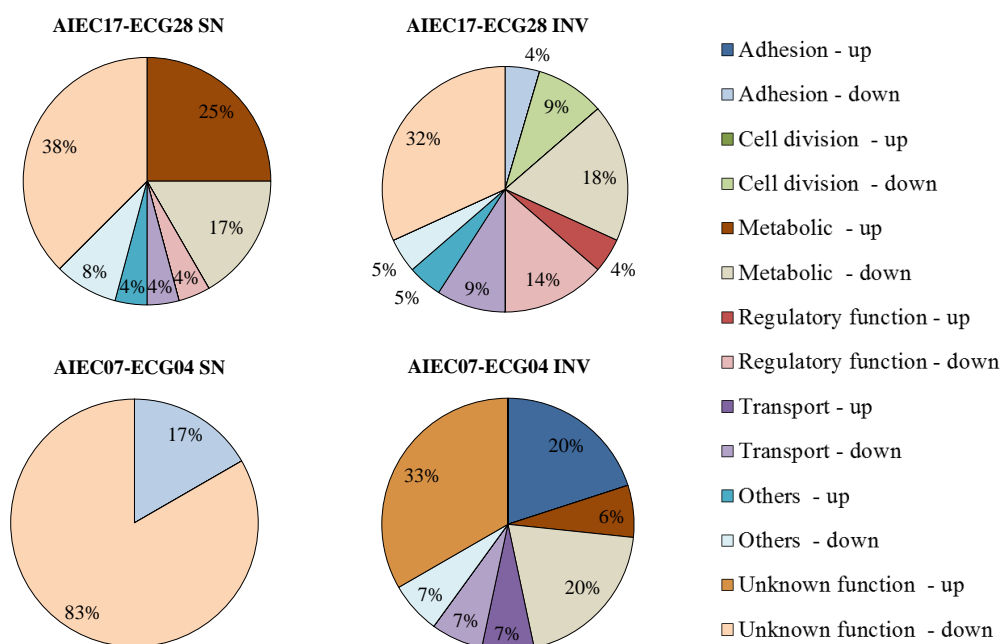


Figure 22. Predicted function of the differentially expressed genes in each comparison distributed in seven functional categories. Each category is divided according to AIEC over-expressed genes (up) and AIEC under-expressed genes (down). Others included genes involved in iron processes, antibiotic resistance, protein degradation, stress response, prophage, and toxin-antitoxin system.

5.2.3 RNA-seq validation

To validate the gene expression differences reported by RNA-seq, a RT-qPCR analysis was performed using the same samples that were sequenced for a subset of DEGs. Those genes with higher gene expression values in AIEC than in non-AIEC strains and high FPKM values were selected (17/19). Gene-specific designed primers showed efficiencies ranging from 86.92 to 103.18%, with the exception of XLOC_001397 which presented an efficiency of 59.77% and XLOC_000912 which had 79.20% (Table 24). To evaluate correlations between RNA-Seq and RT-qPCR values, the log₂ fold-change values obtained by each method for the 17 genes were compared for each condition and analysed according to Pearson or Spearman correlation (Figure 23). A significant correlation was obtained for three of the comparisons (AIEC17-ECG28 SN, AIEC17-ECG28 INV and AIEC07-ECG04 SN), indicating the validation of the RNA-Seq results. For the AIEC07-ECG04 INV comparison a tendency can be perceived.

Table 24. Relative gene expression values assessed by RNA-Seq and Fluidigm/RT-qPCR. Bold values indicate the condition in which the expression of this gene was significant by RNA-Seq. Validation values (Fluidigm) for these genes are also highlighted in bold. In each comparison the AIEC has been used as a reference, and 16S values were used to normalise fluidigm results. Negative values indicate over-expression of the gene in AIEC and positive values under-expression in AIEC.

Gene_id	RNA-Seq log2 fold-change				Fluidigm/RT-qPCR log2 fold-change				Primer Efficiency (%)
	AIEC17-ECG28, SN	AIEC17-ECG28, INV	AIEC07-ECG04, SN	AIEC07-ECG04, INV	AIEC17-ECG28, SN	AIEC17-ECG28, INV	AIEC07-ECG04, SN	AIEC07-ECG04, INV	
XLOC_001058	-1.468	-4.435	0.802	-1.298	-0.906	-1.248	0.954	-0.066	101.52
XLOC_003046	-1.314	0.064	ND	1.951	2.445	1.464	-0.568	-0.185	99.79
XLOC_000912	-1.267	ND	-1.133	ND	2.068*	NA	NA	NA	79.20
XLOC_002831	-1.253	-0.048	-0.036	7.882	-0.138	0.886	-0.258	-0.311	94.14
XLOC_001257	-1.119	-0.200	-0.164	1.182	-0.507	1.597	0.179	-0.323	97.75
XLOC_001255	-1.100	0.252	-0.206	0.573	-0.726	1.039	-0.229	-1.358	101.18
XLOC_001256	-0.975	-0.390	-0.050	1.272	-0.379	1.725	-0.186	-1.253	103.14
XLOC_003163	-0.385	-5.582	-0.905	ND	0.853	-2.715	-0.231	0.396	90.08
XLOC_000036	-0.141	-3.357	-0.240	-0.022	0.817	-1.448	-0.053	1.427	86.92
XLOC_001396	0.049	0.094	0.645	-2.334	1.061	2.509	0.700	-3.539	99.54
XLOC_001395	-0.063	0.252	0.759	-2.255	2.659	1.166	0.433	-0.159	101.14
XLOC_001397	0.113	-0.459	0.571	-2.005	1.319	1.133	0.322	-0.685#	59.77
XLOC_002857	-0.166	0.403	-0.534	-1.765	0.488	3.344	-0.100	1.075	103.18
XLOC_001815	2.094	1.521	0.367	-2.271	3.410	3.986	0.334	-2.531	101.39
XLOC_000512	0.802	2.747	0.962	-2.003	2.046	4.280	-0.140	1.231	102.48
XLOC_000794	0.613	0.693	1.364	-1.947	1.110	2.878	0.337	0.291	94.32
XLOC_000511	2.063	2.087	0.904	-1.632	NA	NA	NA	3.941*	82.12

ND: Gene not detected by RNA-seq. NA: Not assessed. *Assessed by RT-qPCR. #This sample was evaluated in a 1/4 dilution while the rest were tested in a 1/20 dilution.

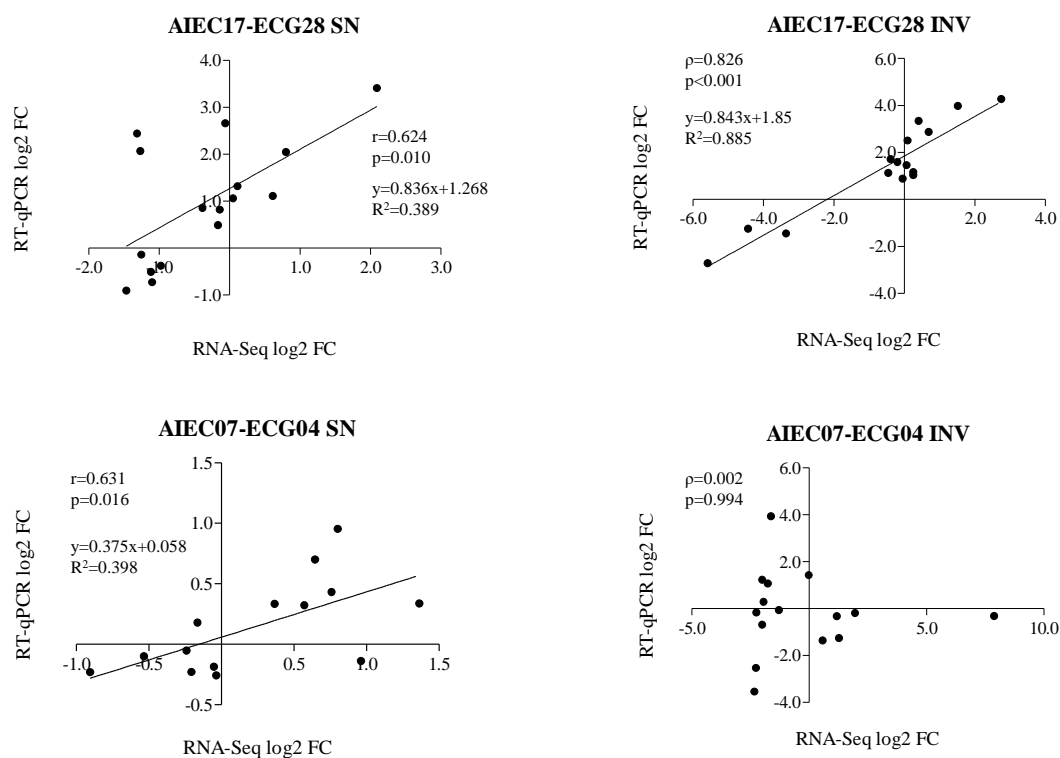


Figure 23. Correlation between the log₂ fold-change (FC) values obtained by RNA-seq and those obtained by RT-qPCR. Correlation coefficients either Pearson (r) or Spearman (ρ) and p -values are indicated. Linear regression equation values are also depicted.

When accounts for individual DEG validation, 11 out of the 17 DEGs (64.71%) assessed were also over-expressed in AIEC strains by RT-qPCR, even though the log₂ fold-change values obtained were not exactly the same as those reported by RNA-seq (Table 24). Besides, the log₂ fold-change value of each gene in all comparisons by RNA-Seq and RT-qPCR identified three genes significantly over-expressed in AIEC17 in comparison with ECG28 in the SN condition by RNA-Seq (XLOC_002831, XLOC_001255 and XLOC_001256), that were also over-expressed in AIEC07-ECG04 SN comparison and this tendency was maintained in the RT-qPCR analysis. In contrast, the rest of the genes showed divergent results once comparing the expression of the DEGs between the comparison in which it has reported a significant difference with the same comparison but for the other strain pair. Notably, comparing the RNA-seq values (over-expressed or under-expressed in AIEC) obtained between conditions (SN versus INV), generally the DEGs did not show equal tendency in SN and dissimilar from INV. Nonetheless, one DEG exhibited higher expression in AIEC in the SN fraction compared with non-AIEC strains and less expression in INV (XLOC_001255) and inversely occurred for XLOC_001397. Unfortunately these tendencies were not corroborated by RT-qPCR. This

might be explained by low primer efficiency and small number of replicates and/or differences in the methodology applied (RNA-Seq or RT-qPCR).

Discussion

Multiple studies have been conducted to identify AIEC candidate genes^{19,21,23,43,50,74,81,136,151,154,178} but only two studies on AIEC transcriptomics has been performed so far^{101,195}. In these cases, the gene expression of the AIEC reference strain LF82 was compared against itself growing with or without bile salts or it was compared with the non-invasive HS strain during exponential and stationary growth in LB medium. In this work, to avoid differences related to phylogenetic origin, AIEC/non-AIEC strain pairs with identical pulsotype have been compared and we have investigated differences in gene expression during cell culture infection, a condition closer to the real context. Therefore, a protocol for intracellular bacterial RNA extraction has been optimised and the gene expression of AIEC and non-AIEC strains during growth in suspension (SN) and IECs invasion (INV) has been investigated by RNA-Seq. The approach followed was designed with the purpose of determining key AIEC virulence genes differentially expressed that could provide further knowledge on AIEC pathogenic mechanisms and that could be further used as molecular targets for AIEC identification.

In the last years, a novel technique that provides knowledge of host-pathogen interactions have been developed (Dual RNA-seq), however it requires high sequencing depth to achieve representative host and pathogen genomes. Considering that with an infection of 20 bacteria per eukaryotic cell, the 0.25% of the total sample corresponds to bacterial mRNA and sRNA²⁷⁴, the sequencing depth must be profound (more than 400M reads). Thus, to reduce costs and to provide a more representative bacterial analysis, we optimised a protocol that physically separated the bacterial RNA from the eukaryotic RNA. Intestinal epithelial cells infection was performed in different amounts of sample and adjusted to avoid kit saturation and ensure proper RNA quality. The first step consisted in extracting total RNA. It was first started with a rapid bacterial RNA isolation kit mediated by columns and that it does not recover all 5S rRNA and tRNAs (RiboPure). Initial sample amount was modified as we exceeded the limit of the kit was 1×10^9 *E. coli* cfu. Given that it was not possible to obtain the quantity required to proceed to the second step, the Trizol method was applied. This kit included the Max Bacterial Enhancement reagent which inactivates endogenous RNases and promotes protein degradation to improve RNA quality and integrity. RNA extracted with Trizol was not as clean as the one extracted with RiboPure,

as no removal of small RNAs was performed but enough RNA quantity was obtained. In the final protocol a washing step with PBS was added to prevent the loss of RNA integrity as cell culture medium might impair the Max Enhancement reaction, and compromise RNA integrity. The second step consisted on removing eukaryotic RNA with the MICROBEnrich kit which ensures 90% mammalian RNA elimination. Finally, rRNA was depleted by the Ribo-Zero kit. In this case although the amount of RNA recovered was limited (2.5-18.5%), it is in concordance with the recovery parameters of the kit once ethanol precipitation was performed (2-8%). By this precipitation method, some small RNA molecules may precipitate and can be kept at the end of the procedure. This may explain why the last product presents a 5S rRNA high signal as detected by Agilent Bioanalyzer (Figure 19).

Samples corresponding to the fraction where bacteria are infecting IECs (INV) were sequenced twenty times more deeply than unattached bacteria (SN). As Haas et al.²⁷⁵ suggested, in our case 16-52M reads in the SN samples yielded nearly complete coverage (66-85%). With the exception of ECG28.SN.1 which did not reach the 40% of genome coverage. Similarly, in INV samples, the sequencing depth was appropriate to achieve an 83-91% of genome coverage. This was in concordance with a previous study which suggested that more than 200M reads are necessary to obtain representative bacterial transcriptomes with rRNA-depleted samples²⁷⁴.

Major differences in gene expression values were reported between AIEC and its non-AIEC counterpart during infection (INV) whereas genes expressed during AIEC growth in suspension (SN) were more similar. Noticeably, while in each comparison most of the genes were under-expressed in AIEC, for the AIEC07-ECG04 INV, 66.7% of the genes were over-expressed in AIEC. Although it is not fully-understood, we suspect that differences in the adhesive and invasive levels or the phylogenetic origin of the strains may explain this observation. Additionally, previous studies reported either differences in the invasive capacity of LF82 strain once the expression of particular genes were blocked or different gene expression levels between conditions (for instance, presence of bile salts)^{18,154,159,160,165}. Nonetheless, no comparison was performed between gene expression levels of AIEC versus non-AIEC strains. Herein, no statistically significant differences were encountered for those genes according to pathotype (Table S25) indicating that either they are not essential for AIEC invasion or that their gene expression does not differ between the strains and conditions assessed in this study. In comparison to Zhang et al.¹⁰¹

and Delmas et al.¹⁹⁵ transcriptomics studies, limited number of DEGs were described in our work. One possible explanation could be that while genetically distant strains (LF82 and HS) were analysed in the former study, our strains were really close (they shared identical pulsed field gel electrophoresis). Another possibility could be that the conditions studied altered differently the gene expression. Accordingly, the methodology applied previous to study transcriptomics is of importance.

Since the strain pairs analysed belong to different phylogroups (B1 and D), different virulence mechanisms might have been evolved and this might explain the reason why there was not a gene differentially expressed in the same condition independently of the strain pair analysed. Nonetheless, it is worth noting that some of the genes found by RNA-seq have been previously related with bacterial virulence or involved in bacterial pathogenic processes, suggesting a putative involvement in AIEC pathogenesis. Ten gene-encoding proteins involved in metabolic processes were detected, these may facilitate the bacterial immune system evasion (XLOC_000090 related with sialic acid metabolism²⁷⁶, XLOC_003046, XLOC_001255, XLOC_001256 and XLOC_001257 involved in arginine biosynthesis²⁷⁷), promote bacterial adherence (XLOC_000422, XLOC_000413, XLOC_00415, XLOC_000418 and XLOC_000414 involved in synthesis and export of colonic acid²⁷⁸) and invasion (XLOC_000090 related with sialic acid metabolism²⁷⁶). Additionally, it was suggested that two transcriptional regulators (XLOC_000036 and XLOC_00252) may have an indirect role to bacterial virulence, both encourages bacterial survival in unfavourable conditions^{279,280}. Indeed, XLOC_000036 encodes for the TdcA protein and its deletion reduces *Salmonella enterica* virulence²⁷⁹. Moreover, TdcA negatively regulates OmpA²⁸¹, a protein earlier associated with AIEC adhesiveness and invasiveness¹⁵⁹. Similarly, the inactivation of the protein permease PstA (XLOC_002857) attenuated the pathogenicity of an APEC strain²⁸². Interestingly, since AIEC adhesion and invasion is mediated by the adhesion of type-I pili (FimH)¹⁶³ and other surface structures such as ChiA⁸⁵ and OmpA¹⁵⁹, the DEGs between AIEC and non-AIEC strain which are related with fimbriae synthesis and assembly (XLOC_003227, XLOC_001396, XLOC_001395 and XLOC_001397), as well as, OmpD porin (XLOC_001935) are of relevance and may be of interest for future studies. Indeed, differences in gene expression of other surface appendages (i.e. flagellum) between one AIEC and one non-invasive strain had already been described¹⁰¹. Finally, proteins related to other processes may also be of significance. For instance, due to the zinc function described in STEC²⁸³ and EPEC²⁸⁴, the periplasmic zinc resistance-associated protein precursor (XLOC_001058) that acts as an important

component of zinc-balancing mechanism, may play a role in mediating bacterial adherence. Furthermore, genes related with protein degradation (cysteine hydrolase XLOC_00290), multidrug resistance (MdtB XLOC_000410) and carbon limitation (carbon starvation protein A XLOC_003163) may also be implicated indirectly to bacterial virulence by helping bacteria to cope with hostile environments²⁸⁵⁻²⁸⁷.

RNA-seq validation was performed by RT-qPCR using *gapdh* as a housekeeping gene. The *gapdh* gene expression was assessed in a previous study where an EHEC strain infected epithelial cells, and no variances in its expression were reported between conditions²⁸⁸. In addition, as normalization according RNA quantity is not possible in samples with mixed RNA, 16S rRNA was quantified and used to homogenise the RTA values. By RT-qPCR 64.7% of the DEGs were validated. Discordant RNA-Seq and RT-qPCR results may be explained by the different methodology applied. For instance, the fact that duplicates of SN samples were polled for the RT-qPCR analysis, sample normalisation, the sensitivity of the tool or the high false-positive rate of the Cuffdiff program²⁸⁹. On the other hand, in three of the conditions assessed correlation between RNA-seq values and RT-qPCR was obtained, suggesting a good quality of RNA-seq data. Exceptionally, no significant correlation was seen in AIEC07-ECG04 INV condition. In this case, we would recommend increasing the sample size as a tendency is perceived.

Our preliminar comparative transcriptome analysis evidenced the presence of strain-specific DEGs rather than key genes associated with the AIEC pathotype since no common DEG was found between AIEC strains. For that reason, we suggest that strains from the AIEC pathotype may use multiple approaches to promote IEC invasion. Although limitation exists in our analysis by the fact that the methodological approach embodies a different environment than within the human intestine, herein we present a protocol to ensure bacterial RNA isolation and a more realistic view of expression of virulence genes implicated in the AIEC phenotype. Nonetheless, further research should focus on the study of complex microbe-host interactions in order to direct AIEC identification and therapeutic strategies.

• GENERAL DISCUSSION •

Non-pathogenic *E. coli* are common colonisers of the mucus layer of the intestinal tract and have a mutualistic relationship with their hosts. However, some *E. coli* strains have evolved to a virulent behaviour. Among those, strains belonging to the AIEC pathovar are suggested to be of particular concern. AIEC isolates lack typical *E. coli* virulence factors but are phenotypically characterised by their ability to adhere to and invade IECs, as well as, to survive and replicate inside macrophages without inducing host-cell death²¹. By *in vitro* and *in vivo* studies, the AIEC interaction with IECs has been described to take place through its binding to host receptors which in turn promotes intestinal epithelial permeability^{81,83,84,159}. Additionally, induction of high levels of cytokines' secretion and exacerbation of intestinal inflammation in susceptible hosts due to AIEC presence has been reported⁹⁵⁻⁹⁷. Since a high prevalence of AIEC has been depicted in the mucosa of CD patients^{17,21,24,25,43,63,74,75,136} and molecular mechanisms of AIEC virulence have been related with the disease pathogenesis, AIEC has been pointed to take part in the complex multifactorial aetiology of CD (see references in introduction section 2.2.).

To further decipher AIEC role in CD (i.e. disease specificity or association with active disease), as well as to uncover the host range or AIEC reservoirs and transmission paths is of paramount importance to eventually define measures of contamination risk, prevention and/or to provide personalised treatments for AIEC carriers. One reason of the lack of information in these aspects is due to the fact that an AIEC molecular biomarker is still missing. Its identification relies on phenotypic traits undergoing cell-culture infection assays, which are extremely time consuming and hard to standardise. Therefore, in this thesis we principally aimed to better define the characteristics of AIEC pathotype and to find putative genetic/phenotypic markers for its rapid identification that could shed light on this field. Three different approaches have been followed to achieve this purpose using either all our AIEC and non-AIEC strain collection or our AIEC/non-AIEC strain pairs with identical pulsotype. The latter was used to increase the possibility to identify more narrow genetic differences specific of the AIEC phenotype. First, gene prevalence of previously described VGs and differences in gene content has been assessed. Second, we have studied amino acid substitutions in five AIEC-related genes and SNPs in the strain

pairs' genomes. Third, gene expression of three outer membrane proteins has been measured and transcriptome analysis has been conducted.

1. Approaches followed to decipher AIEC genetics

In 2004, once Darfeuille-Michaud et al.²¹ defined the AIEC pathotype, a search for unique genes that could explain its phenotype started. Several approaches have been followed to decipher AIEC genetics (gene prevalence, point mutations and gene expression) in which both known genes and novel genes have been studied.

The first studies based on PCR-based gene prevalence^{20,43} insinuated that AIEC strains did not harbour any particular genetic trait that could distinguish them from commensals and that they did not commonly present virulence genes previously described in other *E. coli* pathotypes. In line with this observation, the first genome sequencing studies^{18,19,207,208} together with the most recent genomic studies^{23,101,154,178,210} evidenced again that there was no gene strictly associated with the AIEC phenotype of the strains. However, PCR-based and genomic studies focused on gene content reported some genes to be more prevalent in AIEC than non-AIEC strains (*malX*¹⁴⁶, *kpsMTIII*¹⁴⁶, *lpfA*¹⁵⁴ and *gipA*²⁰⁹, *chuA*⁷⁴, and *pduC*¹⁵⁴). However, low difference in AIEC/non-AIEC gene prevalence was reported for these genes (19-34%) and no confirmation of the findings in other strain collections have been obtained. As a previous study pointed out¹⁷⁸, it is likely that the associations described are phylogenetic in nature and do not reflect the pathogenic potential of the strains.

In the present thesis, two approaches have been conducted to determine differences in gene content: (I) by analysing the prevalence of 61 VGs in a collection of animal and/or human-isolated strains (48 AIEC and 56 non-AIEC) and (II) by comparing the genomes of three AIEC/non-AIEC strain pairs, being each pair really close genetically. Whereas similar gene prevalence between pathotypes was mostly achieved indicating that these genes, including *lpfA*, *fimH*, *chiA*, *ompA*, *ompC* and *ompF* might not be involved in AIEC phenotype, few genes reported different prevalence between human AIEC and non-AIEC strains. Noticeably, the four genes more prevalent in human AIEC strains encoded for two toxins (Vat and Pic), one adhesin (PapGII/III) and one serum resistance protein (Iss). Obtained results were in concordance with previous studies that also found differences in *vat* prevalence^{23,157}, but in discordance with Dogan et al.¹⁵⁴ who reported comparable *iss* prevalence among AIEC and non-AIEC strains. Nevertheless, once again none of the four

genes were present in all AIEC strains, reinforcing the idea that no particular VG is related to AIEC phenotype.

Controversial results on gene frequency may be explained due to differential strain collection (origin of isolation, host and phylogenetic origin) and the amount of strains considered (Table 4). In our case it is mainly constituted by B2 strains but, for example, the collection of a previous study¹⁵⁴ is enriched in A and B1. As a consequence, the results of studies comparing unequal strains could be questioned. Such is the case of Desilets et al.²³ who reported that B2-strains harboured three genomic regions that were absent in non-AIEC strains but in the last group all were non-clinical isolates and only two B2 strains were considered. Since the AIEC pathotype is genetically highly diverse by phylogroup and invasive determinants, cross-validation of observations in a strain collection is strongly recommended.

Besides, it has been suggested that variations in the sequence of particular genes (*fimH*, *chiA* and *ompA*) may uncover AIEC virulence abilities^{83,85,159}. Therefore, one of the studies conducted in this thesis consisted on the examination of the protein sequences of FimH, ChiA, OmpA, OmpC and OmpF in a large collection of strains. Regarding FimH our results support the hypothesis of other research groups^{23,74,165,178} and for ChiA, OmpA, OmpC and OmpF only one AIEC strain has been considered so far, thus our results are the first demonstrations. In general, no relevant differences in the pathoadaptative mutations according to pathotype were reported, instead most of them related with phylogroup. Only one amino acid substitution in OmpA (A200V) and three in OmpC (S89N, V220I and W231D) associated with pathotype but these genetic traits presented low specificity and sensibility as markers for AIEC screening. Despite no particular mutations in ChiA were associated with AIEC pathotype, we found that the LF82 ChiA sequence variant was mainly shared by AIEC strains. Nonetheless it only comprised 35.5% of all AIEC strains.

The analysis of SNPs in the whole genome raised interest since it has provided a novel approach to look for AIEC genetic markers. The first study using this methodology took place in 2015, therein only B2 strains were included²¹⁰. Twenty-nine SNPs that could differentiate 4 AIEC together with 51 ExPEC strains from the commensal and other ExPEC strains were identified but no specific characteristic able to discriminate the AIEC pathotype was found²¹⁰. This observation was in concordance with O'Brien et al.¹⁷⁸ results, who analysed differences in base composition of genes among AIEC and non-AIEC

strains from the same sequence type and no clustering of AIEC strains was observed. Contrary, the comparative genomics study of our AIEC/non-AIEC strain pairs revealed three SNPs (E3-E4_4.3(2), E3-E4_4.4 and E5-E6_3.16=3.22(2)) that resulted in differential nucleotide distribution between AIEC and non-AIEC strains in a larger strain collection (22 AIEC and 28 non-AIEC). Despite they also presented association with adhesion and invasion indices, their implication with the phenotype could not be validated by isogenic-mutants. All the attempts conducted to perform 4.3 and 4.4 isogenic mutants were unsuccessful, and similar adhesiveness and invasiveness to IECs as well as replication in macrophages was reported for the LF82 wild-type and the LF82 Δ 3.16. Even though, there was no nucleotide only present in AIEC strains and absent in non-AIEC. Thus, our study corroborates absence of AIEC-specific genetic markers widely distributed across all AIEC strains. In fact, the results obtained by analysing gene prevalence and point mutations reinforce the idea that no particular VG or pathoadaptative mutation described so far is specifically linked with the AIEC pathotype, albeit diverse genetic traits could drive to the same phenotype. However, studies reinforcing this hypothesis are absent and a specific signature sequence of these strains remains to be elucidated.

In spite of the advance on the understanding of AIEC genetics, AIEC/non-AIEC differential gene expression has been scarcely studied^{89,101,195}. Indeed, the three studies earlier conducted examined only LF82 against HS or K-12 gene expression. Furthermore, they studied only one gene during intramacrophage bacterial replication⁸⁹, seven genes in the presence of bile salts¹⁹⁵ or comparative transcriptomics while growing in LB medium¹⁰¹. Our research contributed to this by studying OMPs gene expression in a collection of AIEC/non-AIEC strains and by performing a comparative transcriptomics study between two AIEC strains and their non-AIEC counterparts. Noticeably, both approaches analysed gene expression during bacterial IEC invasion. An increase of OMPs expression was reported in AIEC strains during growth in the supernatant of cell cultures while a diminution was reported during IEC infection in comparison to non-AIEC strains. Consequently, it is suggested that the expression of OMPs may participate in AIEC pathogenesis. Aside, genes-encoding for proteins involved in metabolic processes, transcriptional regulation, protein degradation, as well as, bacterial adhesion and invasion have been detected in the comparative transcriptomics analysis. In particular, it is worth mentioning that higher expression of a negative regulator of OmpA (TdcA)^{279,281} has been reported in AIEC17 in comparison to ECG28 during invasion, which will be in concordance with the observation obtained from the *ompA* gene expression (i.e. AIEC

strains showed lower *ompA* expression in the INV fraction than non-AIEC strains). Regardless of the several challenges encountered in intracellular bacterial transcriptomics (i.e. RNA purification steps and high sequencing depth), this approach can reveal novel biological insights. Future work is required to confirm the implication of the differential expression in the AIEC phenotype by performing mutants of expression and to decipher whether the differential expression is a trait common in all AIEC strains by studying the gene expression in a larger strain collection.

2. Putative biomarkers to assist AIEC identification

To date, six genetic elements have been suggested as putative AIEC molecular markers (Table 25), however they either present low sensitivity or have been studied in a reduced number of strains. The putative biomarkers presented by Dogan et al.¹⁵⁴ and Vazeille et al.¹⁵¹ were more prevalent in AIEC than in non-AIEC strains, nonetheless they were also present in non-AIEC strains (*pduC* and *lpfA*; 50% and 71% of AIEC; 20 and 20% of non-AIEC respectively), albeit in low percentage, or found only in a reduced number of AIEC strains (*lpfA+gipA*; 31% AIEC; 0% non-AIEC). As a consequence, the specificity values were still high (80-100%) but the sensitivity values were low (ranging from 31 to 71%). The opposite occurred for the *chuA* gene⁷⁴; in this case it was present in 93% of AIEC and in 59% of non-AIEC strains what ended in a high sensitivity (93%) and high probability of false-positives (i.e. low specificity). Deshpande et al.²¹⁰ discovered 29 SNPs that could differentiate a group of AIEC strains from a group of ExPEC and commensal strains (all from the B2 phylogroup) but they only studied four AIEC strains. Moreover, the three genomic regions described by Desilets et al.²³ also raised interest. Nevertheless, it should be noted that only 6 non-AIEC strains has been included and AIEC strains have been classified based only in its capacity to replicate within macrophages. Likewise, as only B2 strains were studied the general utility of this approach for any putative AIEC strain remains to be determined.

In this line, we proposed two additional markers that present either higher sensitivity or have been studied in a larger strain collections than the previous presented. On one hand, in this thesis we have deeply characterised genetically and phenotypically a collection of AIEC and non-AIEC strains isolated from the intestinal mucosa of humans. Herein, AIEC screening could be assisted by the evaluation of two traits (the presence of *pic* gene and the resistance to ampicillin). Although these traits are not specific and widely distributed across the pathotype, *E. coli* strains which have resistance to ampicillin and harbor the *pic* gene

present a probability of 82% to be AIEC. Its major problem was about false-positives, thus it could only be used as an initial screening tool and the AIEC predicted strains by this method should be further tested phenotypically. On the other hand, in contrast to previous studies seeking to find genetic markers in the genome of AIEC strains, we have compared strain pairs that could be considered clones but that differed on the phenotype. By means of this methodological approach, the combination of three point mutations (E3-E4_4.4, E5-E6_3.16=3.22(2) and E5-E6_3.12) resulted in the prediction of AIEC phenotype with a sensitivity of 82%, a specificity of 86% and an accuracy of 84%. However, the prediction values were not maintained when additional strains from different geographical locations were studied for validation (accuracy 61%). Interestingly, if only Spanish *E. coli* isolates (Girona and Mallorca) were analysed, the accuracy of the algorithm was maintained (accuracy 81%).

Table 25. Genetic elements more frequently found in strains from the AIEC pathotype and suggested as putative AIEC molecular markers.

Marker (ref)	Group of study (N)		Prevalence (%)		Sensitivity (%)	Specificity (%)	Accuracy (%)
	AIEC	non-AIEC	AIEC	non-AIEC			
<i>pduC</i> ^{154¥}	24	25	50	20	50	80	65
<i>hpfA</i> ^{154¥}	24	25	71	20	71	80	75
<i>hpfA</i> + <i>gipA</i> ¹⁵¹	35	103	31	0	31	100	83
<i>chuA</i> ⁷⁴	15	37	93	59	93	41	56
29 SNPs ^{210*}	4	1307	100	4	-	-	-
3 genomic regions ^{23#}	14	6	79	0	79	100	85
<i>pic</i> + <i>ampR</i> (this thesis)	22	27	86	33	86	67	75
SNP algorithm (this thesis - Girona and Spanish external collection [§])	29	35	-	-	79	83	81
SNP algorithm (this thesis - Girona and all external collections) [¶]	86	99	-	-	45	75	61

[¥] This strain collection was mainly formed by strains from A and B1 phylogroup (14 A, 16 B1, 10 B2 and 9D). ^{*} Only B2 strains were included. In this case, the non-AIEC group included commensal and ExPEC strains. [#] Only present in B2 AIEC strains. The strains' phylogroup were: AIEC: 1 A, 1 B1, 10 B2, 1 D and 1 F; non-AIEC: 2 A, 2 B1 and 2 B2. [§] It includes AIEC and non-AIEC strains isolated from Mallorca. [¶] It includes AIEC and non-AIEC strains from France, Australia, Chile and Spain (Mallorca), as well as, ExPEC strains from Spain and America.

No other study has performed whole-genome sequencing of strains genetically considered clones but discordant for the pathotype, only one study has focused on SNPs along the strain genome²¹⁰ and none of the studies have validated the genetic differences found in a broad strain collection as ours (Table 25). In light of these aspects, the strategy followed adds substantial novelty to the research field and better predictive value than the genetic elements previously published. However, the inclusion of external AIEC, non-AIEC and ExPEC strains helped us to detect that this algorithm achieved only good predictable values in Spanish strains. Therefore, before drawing conclusions on whether a molecular marker is adequate to identify AIEC strains, we recommend to perform additional analysis to confirm its specificity, sensitivity and accuracy. First, the obtained results should be verified in a larger set of strains including AIEC and non-AIEC strains from other geographical origins. Second, since AIEC present similar genetic traits as ExPEC strains^{17,21,43}, to determine the specificity of the method with other *E. coli* pathotypes, in particular ExPEC strains, would also be required. Finally, if the results of the previous mentioned analysis confirm the usefulness of the purposed method, to test the utility of the tool in clinical specimens (both fecal or tissue biopsies) should be considered.

3. Possible reasons why the search for AIEC molecular markers is challenging

Failure to detect a molecular property strictly associated with AIEC so far might be explained by how AIEC might be emerged:

- (I) AIEC isolates by no means represent uniform populations^{23,101,178}. This pathotype is highly diverse based on genetic and phenotypic characteristics such as virulence gene carriage or serotype. Even though most of them belong to the B2 phylogroup, they can comprise all the principal phylogenetic groups (A, B1, B2 and D)^{17,43,63,69,205}. Moreover, they present genetic similarities with ExPEC strains^{17,21,43}. Therefore, the AIEC phenotype might be driven by the combination of various virulence genes that do not necessarily need to be the same for each AIEC strain. Since different mechanisms are involved in the colonisation of the epithelium by AIEC, the hypothesis considering that there is not a key determinant in common for all the AIEC strains but that different ones can lead to the same phenotype gains plausibility. Besides, one study has recently described that the genetics of one particular AIEC strain changes during host-to-host transmissions²⁹⁰, what makes the search for biomarkers even more complex.

- (II) AIEC might present SNPs distinct from non-AIEC strains. To date, one study has focused on SNPs through the coding regions of the strain genome²¹⁰ and another has analysed differences in base composition of genes¹⁷⁸, apart from ours. No specific biomarker has been discovered so far, but since the knowledge at this level is limited, the presence of AIEC-specific mutations cannot be fully discarded. More studies focusing on synonymous SNPs throughout the genome would be of interest and studies looking for SNPs in non-coding regions are needed.
- (III) Finally, differential gene expression may determine the phenotypic characteristics of AIEC strains. So far, apart from our work, only two studies have described the transcriptome of AIEC^{101,195}. In total only three AIEC strains have been studied and the experimental designs conducted do not allow obtaining the best picture of the real expression profiles during AIEC gut colonisation. New experimental approaches directed to examine these elements in particular conditions where AIEC isolates behave distinct from other strains may help in finding molecular markers for AIEC detection that will be probably also worthy in clinical samples. Modulation of gene expression might be determined by various ways, such as DNA methylation or transposable elements. DNA methylation has been described to occur in bacteria, in a manner that clonal bacterial populations can be split by switching among alternative DNA methylation patterns²⁹¹. For instance, as studied in an UPEC strain, the Pap pilin variates the phase by a mechanism which involves methylation²⁹². Likewise, in terms of transposable elements, one study previously demonstrated that through constant macrophage exposure a commensal *E. coli* strain can evolve to pathogenic strain (i.e. being able to survive inside macrophages or escape) by the acquisition of transposable element insertion²⁹³. On the whole, epigenetics and transposable elements are unexplored in AIEC research and it should also be kept under consideration once looking for AIEC characteristic elements.

Regardless of the above mentioned study approaches, once looking for AIEC biomarkers, the first question the scientists should face is the standardisation of the current AIEC identification method. The vast majority of studies have classified an isolate as AIEC by analysing all its phenotypic characteristics *in vitro*, nonetheless some discrepancies exist on the protocols (Table 26) and the cell lines used (Figure 24). Variances in the MOI and time of infection, as well as incubation conditions occur. In terms of invasion assays, while most of the analyses are performed at MOI 10 with an infection time of 3 hours and subsequent

1-hour incubation with gentamicin (100 µg/mL), others assessed the invasive capacity with a higher MOI (20 or 100), less time of infection (30 min, 1 h or 2 h) and different antibiotic concentration (50 µg/mL or 3 mg/mL). Additionally, there is even more variability with the protocols used to determine the capacity of the strains to survive and replicate inside macrophages. In this case, the highest discrepancy is on the infection conditions; since some perform a centrifugation step to facilitate bacterial intramacrophage uptake, whereas others do not. After this time of infection, non-phagocytosed are treated with antibiotic at different concentrations and for different incubation times. The most common procedure includes a first step of 1 hour with higher antibiotic concentration (100 µg/mL) followed by a second step of 24h incubation with decreased antibiotic concentration (15, 20, 50 µg/mL). Even though, other studies perform only one incubation step which consists of 1 or 24 hour step with the same concentration of antibiotic (20, 50, 100 µg/mL or 3 mg/mL).

Moreover, the cell lines used to date (Figure 24) might not be the most appropriate considering that, for instance, I-407 and Hep-2 come from cervical carcinoma and epithelial carcinoma of unknown origin respectively and both result from HeLa contamination. In exception, Caco-2 and T84 are derived from colorectal carcinomas but it is poorly defined how applicable are them for AIEC identification based on CD pathogenesis. Similarly occurs for intramacrophage survival, the cell lines mostly used are J774 which is derived from murine origin^{17,21,24,25,43,63,74,136}. Some studies, including our work, have started to use THP-1 (human monocytes) but bacterial intramacrophage survival methodology differs among them^{153,178,294}. Additionally, previous to bacterial adhesion and invasion, bacteria need to cross the mucus layer. As a consequence, an assay examining bacteria capacity to disrupt and translocate through the mucus should also be contemplated.

Table 26. Comparison of the principal experimental conditions of the protocols used to assess bacterial invasion to intestinal epithelial cells and survival and replication inside macrophages.

Invasion assays			
MOI	Infection conditions	Incubation conditions	References
10	30 min	3 h with amikacin 100 µg/mL	74
10	1 h	2 h with gentamicin 100 µg/mL	26
10 or 20	3 h	1 h with gentamicin 100 µg/mL	16,21,43,87,120,136, 151,155,156,160,166 ,178,192 / 17,75,154
10	3 h	1 h with gentamicin 3 mg/mL	67
100	2 h	1 h with gentamicin 50 µg/mL	294
100	3 h	1 h with gentamicin 50 µg/mL	267
Survival and replication assays			
MOI	Infection conditions	Incubation conditions	References
10	20 min	Media replacement with gentamicin 100 µg/mL for 40 min and media replacement with gentamicin 50 µg/mL for 24 h	136
10	2 h	Media replacement with amikacin 100 µg/mL for 3 and 24 h	74
10	2 h	Media replacement with gentamicin 100 µg/mL for 1 h and media replacement with gentamicin 20 µg/mL for 24 h	21,178
10 or 100	Centrifugation 10 min at 1000 x g and incubation 10 min	Media replacement with gentamicin 100 µg/mL for 40 min and media replacement with gentamicin 20 µg/mL for 24 h	22,43,151
10	Centrifugation 5 min at 500 x g and incubation 30 min	Media replacement with gentamicin 100 µg/mL for 2 h and media replacement with gentamicin 15 µg/mL for 24 h	26
20	2 h	Media replacement with gentamicin 100 µg/mL for 1 h and media replacement with gentamicin 20 µg/mL for 24 h	17,154
20	2 h	Media replacement with gentamicin 100 µg/mL for 1 and 24 h	75
20	2 h	Media replacement with gentamicin 3 mg/mL for 1 and 24 h	67
100	Centrifugation 10 min at 1000 x g and incubation 10 min	Media replacement with gentamicin 20 µg/mL for 1 and 24 h	89,155
100	2 h	Media replacement with gentamicin 50 µg/mL for 1 and 24 h	294

In view of the lack of standardisation, the adhesion and invasion indices as well as the replication index of the strains are highly variable between research groups. Taking into account the indices of the LF82 AIEC strain which is commonly used as control in these procedures, the adhesion index oscillates between 4.8 and 62.8 bacteria/cell^{21,85,160,165,178,192,209,294}, the invasion index varies from 0.12 to 12.2%^{16,17,21,87,136,155,156,166,178,192,209,267,294,295} and the intramacrophage survival and replication

index ranges between 227.8 and 580.0%^{17,21,22,89,136,178,294}. This is of particular concern especially for those strains that present low indices. In this case, one strain in one laboratory may be considered AIEC while in another may be classified as non-AIEC. Therefore, there is a need to solve this issue in order to regulate AIEC strains classification. Without consistency in the actual screening method it is hard to look for AIEC genetic differences as we might be using inaccurate isolates.

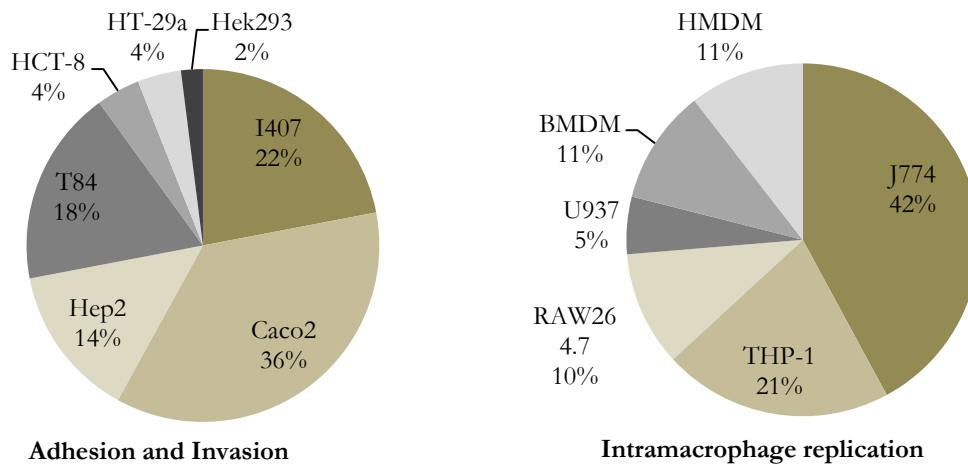


Figure 24. Review of cell lines used for AIEC identification. Analysis of the cell lines used for adhesion and invasion assays are based on 28 previously published works^{16,17,21,22,26,43,67,74,75,85,89,115,120,121,136,138,154–157,160,165,166,178,209,213,267,294,296} while for intramacrophage replication 16 studies were considered^{17,21,22,26,43,67,74,75,89,136,138,154,155,166,178,209,294}.

4. Is the AIEC phenotype an acquired trait of *E. coli* strains from the gut?

By looking at our outcomes and recent published data, it becomes believable that the AIEC phenotype is not permanent, yet we suspect that one *E. coli* can acquire the AIEC phenotype in particular conditions and inversely, one AIEC strain without specific triggers might turn to a non-AIEC strain or at some extent modify its level of virulence. One reason that could explain this hypothesis is the fact that very genetically close *E. coli* strains (identical PFGE profiles) can be classified as either AIEC or non-AIEC. Thus, indicating that these strains have evolved to a pathogenic condition by nearly imperceptible genetic, transcriptomic or epigenomic changes that may occur in particular cases. Furthermore, Elhenawy et al.²⁹⁰ have recently shown that one AIEC strain (NRG857c) evolves during host-to-host transmission in mice models, resulting in a diversified population of isolates with two predominant phenotypes: hypermotile isolates and isolates with improved acetate utilization. The first phenotype was due to the presence of an insertion sequence upstream

of the flagellar regulator *flhDC*, what resulted in hypermotile strains with enhanced IECs invasion. However, the presence of this insertion was reversible in the absence of host selection, suggesting that with the absence of particular conditions, the AIEC virulence may be altered. In the same way, Proença et al.²⁹³ observed that under continuous macrophage pressure one commensal strain evolved to increase intracellular survival due to the acquirement of a transposable element insertion. Thus, their observations reinforce the hypothesis of intra-host *E. coli* evolution to an adherent invasive phenotype and the importance of conducting experiments simulating disease conditions as much as possible, since the AIEC marker may only be detected under selective pressure conditions.

Taking all these outcomes into account, one may consider that AIEC strains come from non-AIEC strains from the gut. For that reason, in the foreseeable future, other approaches, beyond genes or SNPs prevalence, should be analysed when looking for AIEC molecular markers. These include transcriptomics, epigenetics and the study of AIEC in conditions in which they behave distinct from other pathotypes, perhaps during its interaction with host cells. Nowadays, two studies on transcriptomics^{101,195} have been conducted. One described that AIEC LF82 strain growing in contact with bile salts increases the expression of genes involved in ethanolamine utilization in comparison to K-12 and demonstrated that AIEC strains grow more after an incubation with minimum media with bile salts supplemented with ethanolamine than non-AIEC¹⁹⁵. Therefore, reinforcing the idea that AIEC strains may adapt their metabolism according to gut conditions and that experimental methods need to be carefully considered when drawing conclusions about AIEC molecular traits. Nonetheless, the gene expression analysis of other non-AIEC and AIEC strains apart from K-12 and LF82 in the presence of bile salts has not been provided, thus it is not possible to say that it is an AIEC-specific trait nor an adaptative method common among AIEC strains. Besides, Zhang et al.¹⁰¹ identified potential coding regions that could be applied as signature transcripts. Nevertheless, it is worth noting that they compared only one AIEC strain (LF82) with one commensal (HS) during growth in LB medium. Thereby, the differences found could be strain-specific or perceptible due to the phylogenetic distance of the strains rather than to the AIEC phenotype. It is against this background that we performed an RNA-seq analysis of two highly genetically similar AIEC/non-AIEC strain pairs. Of note, our work was also conducted during IECs infection to get a more reliable interpretation of the gut context. To the extent of the study, although there are some transcripts with a stimulating role in

AIEC virulence, a candidate transcript susceptible to be considered a universal and specific AIEC probe has not been detected yet.

5. Concluding remarks and future directions

Although what constitutes an AIEC strain remains an enigma, the outcomes obtained by several lines of research in this thesis provide meaningful information on AIEC genetics. Gene prevalence, amino acid substitutions and gene expression have been studied for both known and unknown genetic elements. The latter have been searched by comparative genomics and transcriptomics. The principal contribution of the present work is the finding of two putative AIEC molecular markers, at least for our strain collection, (I) the *pic* gene and ampicillin resistance method presented an accuracy of 75% and (II) the SNP algorithm classifies the Spanish strains correctly with 81% accuracy. Notably, we also present here two studies analysing AIEC gene expression using an *in vitro* assay that simulates bacterial adhesion to and invasion of intestinal epithelial cells and suggested genes putatively involved in AIEC virulence. To sum up, results presented and discussed in this thesis demonstrate that AIEC is a diverse pathotype considering gene content and point mutations, but gene expression studies insinuated that the AIEC phenotype may be determined by particular differences in gene expression.

Apart from that, there are many aspects related with the results presented here that require further exploration: (I) the *pic* prevalence and ampicillin resistance approach should be tested in a larger strain collection from other geographical locations and pathotypes; (II) functional studies to decipher the implication of point mutations in specific genes or their differential expression in AIEC pathogenesis should be performed; (III) genes found differentially expressed should be tested in a larger AIEC/non-AIEC collection; and (IV) transcriptomic studies discerning between adhered and intracellular bacteria may be of interest.

The discovery of an AIEC biomarker would significantly ease further epidemiological studies to better determine AIEC prevalence and abundance, to discover environmental and animal reservoirs and transmission pathways, as well as to facilitate clinical studies in CD patients, for example to study the variations of abundance in relation to the state of the disease or in response to treatments. This biomarker would represent a rapid and cost-effective way to identify AIEC carriers, who could be treated with AIEC-directed therapies. So far, the diversity among AIEC strains challenges the correlation of individual virulence

factors with pathotype in a way that is predictive. Moreover, AIEC pathobiont condition is gaining significance but much remains to be learned about the host-pathogen interactions that govern AIEC infection biology. As a consequence, new approaches need to be performed in order to increase the probability to find an AIEC molecular signature (these include but are not limited to SNPs in non-coding sequences, transcriptomics, metabolomics and epigenomics). Nonetheless, all these studies should be conducted using AIEC strains identified according to a standardised method, and the proposed methods should be tested in diverse strain collections from different geographical regions.

• CONCLUSIONS •

From Chapter 1.1: Virulence gene carriage and adhesin variants of AIEC and commensals isolated from humans and animals

- I. Animal *E. coli* strains present higher number of virulence genes than human-isolated strains, indicating that data obtained from human/animal isolates cannot be directly extrapolated.
- II. Virulence gens profile of strains is highly dependent on the phylogenetic origin, thus to avoid biases. Further analysis aiming at finding AIEC genetic particularities should consider AIEC and non-AIEC collections with similar phylogenetic distribution.
- III. No particular amino acid substitutions in FimH and ChiA are more prevalent in AIEC, yet mutations are mainly associated with the phylogenetic origin of the strains. Of note, the ChiA-LF82 sequence variant is mainly shared among AIEC strains but it only represents the 35.5% of AIEC strains studied. Therefore, based in our strain collection, these genes are not suitable for AIEC screening.
- IV. By combining antibiotic resistance with gene prevalence, a putative signature sequence is described which may facilitate AIEC rapid identification. Strains harbouring *pic* gene and ampicillin resistance have a probability to be AIEC of the 82%, with a global accuracy of 75.5%.

From Chapter 1.2: Amino acid substitutions and differential gene expression of outer membrane proteins in AIEC

- V. Four amino acid positions (P200 in OmpA and P89-P220-P231 in OmpC) present differential distribution between AIEC and non-AIEC strains but they report low sensitivity and specificity, so they are no suitable as molecular markers.
- VI. In addition, particular amino acid changes (OmpA-P200, OmpC-P220 and P232, and OmpF-P51 and P60) correlate with adhesion and/or invasion indices. Thus, our data reveals new putative pathoadaptative mutations that can determine better bacterial adhesiveness and invasiveness.

- VII. The expression of OMPs in AIEC strains varies depending on the condition analysed, whereas non-AIEC strains do not significantly alter their OMPs expression. While growing in suspension, AIEC increases OMPs expression and the opposite occurs in the condition where strains are in contact with IECs. Thus, our study adds knowledge on AIEC OMPs expression during IECs infection.

From Chapter 2.1: Identification by comparative genomics of new single nucleotide polymorphisms to distinguish between AIEC and non-AIEC strains

- VIII. This is the first study that provides a list of polymorphisms present in the genome of AIEC and non-AIEC strain pairs genomically clonal.
- IX. Our study corroborates the absence of AIEC-specific genetic markers widely distributed across all AIEC strains. Nonetheless, our data reveal three SNPs that can be implemented in AIEC identification. Although this tool does not correctly classify all *E. coli* strains, its accuracy is very high (84%), and no comparable molecular tools currently exist.
- X. Unfortunately, the accuracy of the algorithm presented is reduced to 62% once a larger strain collection from different geographic locations and pathotypes is screened, demonstrating that the presented tool is not universal. Nonetheless, the accuracy was maintained to 81% when the two Spanish collections (Girona and Mallorca) were analysed.

From Chapter 2.2: Construction of isogenic mutants to study the role in pathogenicity of three genes related to AIEC pathotype

- XI. Isogenic mutants for 4.3 and 4.4 genes were not obtained and disruption of the 3.16 gene does not result in any perceivable effect on AIEC phenotype. Therefore, the implication in the AIEC phenotype of the mutations found in chapter 2.1 could not be demonstrated.

From Chapter 3.1: RNA-Seq analysis of the transcriptome during growth in cell culture media and during intestinal epithelial cell infection of AIEC in comparison with non-AIEC strains

- XII. A protocol to extract and purify intracellular bacterial RNA and sequence bacterial mRNA has been optimised.

- XIII. Our comparative transcriptome analysis evidences the presence of strain-specific differentially expressed genes rather than key genes associated with the AIEC pathotype since no common gene is found among AIEC strains. However, this is a preliminary study as only two strain pairs have been assessed.
- XIV. RNA-seq and RT-qPCR fold-change values correlated indicating a good quality of RNA-seq data.
- XV. Some of the genes found by RNA-seq have been previously related with bacterial virulence or involved in bacterial pathogenic processes, what points out new molecular mechanisms putatively associated with AIEC pathogenesis still not described.

● REFERENCES ●

1. Bäckhed, F., Ley, R. E., Sonnenburg, J. L., Peterson, D. A. & Gordon, J. I. Host-bacterial mutualism in the human intestine. *Science* 307, 1915–20 (2005).
2. Antoni, L., Nuding, S., Wehkamp, J. & Stange, E. F. Intestinal barrier in inflammatory bowel disease. *World J. Gastroenterol.* 20, 1165–79 (2014).
3. Yu, L. C. H., Wang, J. T., Wei, S. C. & Ni, Y. H. Host-microbial interactions and regulation of intestinal epithelial barrier function: From physiology to pathology. *World J. Gastrointest. Pathophysiol.* 3, 27–43 (2012).
4. Penders, J. et al. Factors influencing the composition of the intestinal microbiota in early infancy. *Pediatrics* 118, 511–21 (2006).
5. Kaper, J. B., Nataro, J. P. & Mobley, H. L. Pathogenic *Escherichia coli*. *Nat. Rev. Microbiol.* 2, 123–140 (2004).
6. Leimbach, A., Hacker, J. & Dobrindt, U. in *Current topics in microbiology and immunology* 358, 3–32 (2013).
7. Tenaillon, O., Skurnik, D., Picard, B. & Denamur, E. The population genetics of commensal *Escherichia coli*. *Nat. Rev. Microbiol.* 8, (2010).
8. Gordon, D. M., O'Brien, C. L. & Pavli, P. *Escherichia coli* diversity in the lower intestinal tract of humans. *Environ. Microbiol. Rep.* 7, 642–648 (2015).
9. Croxen, M. A. et al. Recent advances in understanding enteric pathogenic *Escherichia coli*. *Clin. Microbiol. Rev.* 26, 822–80 (2013).
10. Pašić, L. et al. Two tales of prokaryotic genomic diversity: *Escherichia coli* and halophiles. *Food Technol. Biotechnol.* 52, 158–169 (2014).
11. Rahman, H. & Deka, M. Detection & characterization of necrotoxin producing *Escherichia coli* (NTEC) from patients with urinary tract infection (UTI). *Indian J. Med. Res.* 139, 632–7 (2014).
12. Dhakal, B. K., Kulesus, R. R. & Mulvey, M. A. Mechanisms and consequences of bladder cell invasion by uropathogenic *Escherichia coli*. *Eur. J. Clin. Invest.* 38, 2–11 (2008).
13. Šmajš, D. et al. Bacteriocin synthesis in uropathogenic and commensal *Escherichia coli*: colicin E1 is a potential virulence factor. *BMC Microbiol.* 10, 288 (2010).
14. Katouli, M. Population structure of gut *Escherichia coli* and its role in development of extra-intestinal infections. *Iran. J. Microbiol.* 2, 59–72 (2010).
15. Darfeuille-Michaud, A. et al. Presence of adherent *Escherichia coli* strains in ileal mucosa of patients with Crohn's disease. *Gastroenterology* 115, 1405–1413 (1998).
16. Boudeau, J., Glasser, A. L., Masseret, E., Joly, B. & Darfeuille-Michaud, A. Invasive ability of an *Escherichia coli* strain isolated from the ileal mucosa of a patient with Crohn's disease. *Infect. Immun.* 67, 4499–509 (1999).
17. Baumgart, M. et al. Culture independent analysis of ileal mucosa reveals a selective increase in invasive *Escherichia coli* of novel phylogeny relative to depletion of Clostridiales in Crohn's disease involving the ileum. *ISME J.* 1, 403–418 (2007).
18. Miquel, S. et al. Complete genome sequence of Crohn's disease-associated adherent-invasive *E. coli* strain LF82. *PLoS One* 5, 1–16 (2010).
19. Nash, J. H. et al. Genome sequence of adherent-invasive *Escherichia coli* and comparative genomic analysis with other *E. coli* pathotypes. *BMC Genomics* 11, 667 (2010).
20. Martínez-Medina, M. et al. Similarity and divergence among adherent-invasive *Escherichia coli* and extraintestinal pathogenic *E. coli* strains. *J. Clin. Microbiol.* 47, 3968–79 (2009).
21. Darfeuille-Michaud, A. et al. High prevalence of adherent-invasive *Escherichia coli* associated with ileal mucosa in Crohn's disease. *Gastroenterology* 127, 412–421 (2004).
22. Glasser, A. et al. Adherent invasive *Escherichia coli* strains from patients with Crohn's disease

- survive and replicate within macrophages without inducing host cell death. *Infect. Immun.* 69, 5529–37 (2001).
23. Desilets, M. et al. Genome-based definition of an inflammatory bowel disease-associated adherent-invasive *Escherichia coli* pathovar. *Inflamm. Bowel Dis.* 22, 1–12 (2016).
 24. Dogan, B. et al. Multidrug resistance is common in *Escherichia coli* associated with ileal Crohn's disease. *Inflamm. Bowel Dis.* 19, 141–150 (2013).
 25. Negroni, A. et al. Characterization of adherent-invasive *Escherichia coli* isolated from pediatric patients with inflammatory bowel disease. *Inflamm. Bowel Dis.* 18, 913–924 (2012).
 26. Simpson, K. W. et al. Adherent and invasive *Escherichia coli* is associated with granulomatous colitis in boxer dogs. *Infect. Immun.* 74, 4778–92 (2006).
 27. Ng, S. C. et al. Worldwide incidence and prevalence of inflammatory bowel disease in the 21st century: a systematic review of population-based studies. *Lancet (London, England)* 390, 2769–2778 (2018).
 28. M'Koma, A. E. Inflammatory bowel disease: an expanding global health problem. *Clin. Med. Insights. Gastroenterol.* 6, 33–47 (2013).
 29. Burisch, J., Jess, T., Martinato, M. & Lakatos, P. L. The burden of inflammatory bowel disease in Europe. *J. Crohn's Colitis* 7, 322–337 (2013).
 30. Ye, Y., Pang, Z., Chen, W., Ju, S. & Zhou, C. The epidemiology and risk factors of inflammatory bowel disease. *Int. J. Clin. Exp. Med.* 8, 22529–42 (2015).
 31. Duricova, D. et al. Overall and cause-specific mortality in Crohn's disease: A meta-analysis of population-based studies. *Inflamm. Bowel Dis.* 16, 347–353 (2010).
 32. Salim, S. Y. & Söderholm, J. D. Importance of disrupted intestinal barrier in inflammatory bowel diseases. *Inflamm. Bowel Dis.* 17, 362–381 (2011).
 33. Sartor, R. B. Mechanisms of disease: pathogenesis of Crohn's disease and ulcerative colitis. *Nat Clin Pr. Gastroenterol Hepatol* 3, 390–407 (2006).
 34. Strober, W. & Fuss, I. J. Proinflammatory cytokines in the pathogenesis of inflammatory bowel diseases. *Gastroenterology* 140, 1756–1767.e1 (2011).
 35. Baumgart, D. C. & Sandborn, W. J. Crohn's disease. *Lancet* 380, 1590–1605 (2012).
 36. Hansen, J. J. & Sartor, R. B. Therapeutic manipulation of the microbiome in IBD: current results and future approaches. *Curr. Treat. Options Gastroenterol.* 13, 105–20 (2015).
 37. Biedermann, L. et al. Smoking cessation induces profound changes in the composition of the intestinal microbiota in humans. *PLoS One* 8, e59260 (2013).
 38. Turnbaugh, P. J. et al. The effect of diet on the human gut microbiome: a metagenomic analysis in humanized gnotobiotic mice. *Sci. Transl. Med.* 1, 6ra14 (2009).
 39. Agus, A. et al. Western diet induces a shift in microbiota composition enhancing susceptibility to adherent-invasive *E. coli* infection and intestinal inflammation. *Sci. Rep.* 6, 1–14 (2016).
 40. Martinez-Medina, M. et al. Western diet induces dysbiosis with increased *E. coli* in CEABAC10 mice, alters host barrier function favouring AIEC colonisation. *Gut* 63, 116–24 (2014).
 41. Nickerson, K. P. & McDonald, C. Crohn's disease-associated adherent-invasive *Escherichia coli* adhesion is enhanced by exposure to the ubiquitous dietary polysaccharide maltodextrin. *PLoS One* 7, e52132 (2012).
 42. Friswell, M., Campbell, B. & Rhodes, J. The role of bacteria in the pathogenesis of inflammatory bowel disease. *Gut Liver* 4, 295–306 (2010).
 43. Martinez-Medina, M. et al. Molecular diversity of *Escherichia coli* in the human gut: new ecological evidence supporting the role of adherent-invasive *E. coli* (AIEC) in Crohn's disease. *Inflamm. Bowel Dis.* 15, 872–882 (2009).
 44. Lopez-Siles, M. et al. Mucosa-associated *Faecalibacterium prausnitzii* and *Escherichia coli* co-abundance can distinguish irritable bowel syndrome and inflammatory bowel disease phenotypes. *Int. J. Med. Microbiol.* 304, 464–475 (2014).
 45. Morgan, X. C. et al. Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment. *Genome Biol.* 13, R79 (2012).
 46. Bosca-Watts, M. M. et al. Pathogenesis of Crohn's disease: bug or no bug. *World J. Gastrointest. Pathophysiol.* 6, 1–12 (2015).
 47. Gradel, K. O. et al. Increased short- and long-term risk of inflammatory bowel disease after

- Salmonella* or *Campylobacter* gastroenteritis. *Gastroenterology* 137, 495–501 (2009).
48. Nitzan, O., Elias, M., Chazan, B., Raz, R. & Saliba, W. *Clostridium difficile* and inflammatory bowel disease: role in pathogenesis and implications in treatment. *World J. Gastroenterol.* 19, 7577–85 (2013).
 49. Greenstein, R. J. Is Crohn's disease caused by a mycobacterium? Comparisons with leprosy, tuberculosis, and Johne's disease. *Lancet. Infect. Dis.* 3, 507–14 (2003).
 50. Prorok-Hamon, M. et al. Colonic mucosa-associated diffusely adherent *afaC+* *Escherichia coli* expressing *hpfA* and *pks* are increased in inflammatory bowel disease and colon cancer. *Gut* 63, 761–70 (2014).
 51. Naser, S. A., Ghobrial, G., Romero, C. & Valentine, J. F. Culture of *Mycobacterium avium* subspecies *paratuberculosis* from the blood of patients with Crohn's disease. *Lancet* (London, England) 364, 1039–44 (2004).
 52. Mendoza, J. L. et al. High prevalence of viable *Mycobacterium avium* subspecies *paratuberculosis* in Crohn's disease. *World J. Gastroenterol.* 16, 4558 (2010).
 53. Juste, R. A. et al. On the prevalence of *M. avium* subspecies *paratuberculosis* DNA in the blood of healthy individuals and patients with inflammatory bowel disease. *PLoS One* 3, e2537 (2008).
 54. Kirkwood, C. D. et al. *Mycobacterium avium* subspecies *paratuberculosis* in children with early-onset Crohn's disease. *Inflamm. Bowel Dis.* 15, 1643–1655 (2009).
 55. Ferwerda, G. et al. *Mycobacterium paratuberculosis* is recognized by Toll-like receptors and NOD2. *J. Leukoc. Biol.* 82, 1011–1018 (2007).
 56. Gutierrez, M. G. et al. Autophagy is a defense mechanism inhibiting BCG and *Mycobacterium tuberculosis* survival in infected macrophages. *Cell* 119, 753–66 (2004).
 57. Nazareth, N. et al. Prevalence of *Mycobacterium avium* subsp. *paratuberculosis* and *Escherichia coli* in blood samples from patients with inflammatory bowel disease. *Med. Microbiol. Immunol.* 204, 681–692 (2015).
 58. Feller, M. et al. Long-term antibiotic treatment for Crohn's disease: systematic review and meta-analysis of placebo-controlled trials. *Clin. Infect. Dis.* 50, 473–480 (2010).
 59. Khan, K. J. et al. Antibiotic therapy in inflammatory bowel disease: a systematic review and meta-analysis. *Am. J. Gastroenterol.* 106, 661–673 (2011).
 60. McNeese, A. L., Markesich, D., Zayyani, N. R. & Graham, D. Y. *Mycobacterium paratuberculosis* as a cause of Crohn's disease. *Expert Rev. Gastroenterol. Hepatol.* 9, 1523–34 (2015).
 61. De la Fuente, M. et al. *Escherichia coli* isolates from inflammatory bowel diseases patients survive in macrophages and activate NLRP3 inflammasome. *Int. J. Med. Microbiol.* 304, 384–92 (2014).
 62. Schippa, S. et al. Dominant genotypes in mucosa-associated *Escherichia coli* strains from pediatric patients with inflammatory bowel disease. *Inflamm. Bowel Dis.* 15, 661–672 (2009).
 63. Martin, H. M. et al. Enhanced *Escherichia coli* adherence and invasion in Crohn's disease and colon cancer. *Gastroenterology* 127, 80–93 (2004).
 64. Elliott, T. R. et al. Quantification and characterization of mucosa-associated and intracellular *Escherichia coli* in inflammatory bowel disease. *Inflamm. Bowel Dis.* 19, 2326–2338 (2013).
 65. Kotlowski, R., Bernstein, C. N., Sepehri, S. & Krause, D. O. High prevalence of *Escherichia coli* belonging to the B2+D phylogenetic group in inflammatory bowel disease. *Gut* 56, 669–75 (2007).
 66. Sepehri, S., Kotlowski, R., Bernstein, C. N. & Krause, D. O. Phylogenetic analysis of inflammatory bowel disease associated *Escherichia coli* and the FimH virulence determinant. *Inflamm. Bowel Dis.* 15, 1737–1745 (2009).
 67. Willing, B. et al. Twin studies reveal specific imbalances in the mucosa-associated microbiota of patients with ileal Crohn's disease. *Inflamm. Bowel Dis.* 15, 653–660 (2009).
 68. Rehman, A. et al. Transcriptional activity of the dominant gut mucosal microbiota in chronic inflammatory bowel disease patients. *J. Med. Microbiol.* 59, 1114–1122 (2010).
 69. Keighley, M. R. B. et al. Influence of inflammatory bowel disease on intestinal microflora. *Gut* 19, 1099–1100 (1978).
 70. Sasaki, M. et al. Invasive *Escherichia coli* are a feature of Crohn's disease. *Lab. Investig.* 87,

- 1042–1054 (2007).
71. Thomazini, C. M., Samegima, D. A. G., Rodrigues, M. A. M., Victoria, C. R. & Rodrigues, J. High prevalence of aggregative adherent *Escherichia coli* strains in the mucosa-associated microbiota of patients with inflammatory bowel diseases. *Int. J. Med. Microbiol.* 301, 475–479 (2011).
 72. de Souza, H. L. et al. Mucosa-associated but not luminal *Escherichia coli* is augmented in Crohn's disease and ulcerative colitis. *Gut Pathog.* 4, 21 (2012).
 73. Mylonaki, M., Rayment, N. B., Rampton, D. S., Hudspith, B. N. & Brostoff, J. Molecular characterization of rectal mucosa-associated bacterial flora in inflammatory bowel disease. *Inflamm. Bowel Dis.* 11, 481–487 (2005).
 74. Céspedes, S. et al. Genetic diversity and virulence determinants of *Escherichia coli* strains isolated from patients with Crohn's disease in Spain and Chile. *Front. Microbiol.* 8, 639 (2017).
 75. Raso, T. et al. Analysis of *Escherichia coli* isolated from patients affected by Crohn's Disease. *Curr. Microbiol.* 63, 131–137 (2011).
 76. McGuckin, M. A., Eri, R., Simms, L. A., Florin, T. H. J. & Radford-Smith, G. Intestinal barrier dysfunction in inflammatory bowel diseases. *Inflamm. Bowel Dis.* 15, 100–113 (2009).
 77. Moran, A. P., Gupta, A. & Joshi, L. Sweet-talk: role of host glycosylation in bacterial pathogenesis of the gastrointestinal tract. *Gut* 60, 1412–25 (2011).
 78. Bevins, C. L. & Salzman, N. H. Paneth cells, antimicrobial peptides and maintenance of intestinal homeostasis. *Nat. Rev. Microbiol.* 9, 356–368 (2011).
 79. Crawley, S. W., Mooseker, M. S. & Tyska, M. J. Shaping the intestinal brush border. *J. Cell Biol.* 207, 441–51 (2014).
 80. Roda, G. et al. Intestinal epithelial cells in inflammatory bowel diseases. *World J. Gastroenterol.* 16, 4264 (2010).
 81. Chassaing, B. et al. Crohn disease-associated adherent-invasive *E. coli* bacteria target mouse and human Peyer's patches via long polar fimbriae. *J. Clin. Invest.* 121, 966–75 (2011).
 82. Gullberg, E. & Soderholm, J. D. Peyer's patches and M cells as potential sites of the inflammatory onset in Crohn's disease. *Ann. N. Y. Acad. Sci.* 1072, 218–232 (2006).
 83. Barnich, N. et al. CEACAM6 acts as a receptor for adherent-invasive *E. coli*, supporting ileal mucosa colonization in Crohn disease. *J. Clin. Invest.* 117, 1566–74 (2007).
 84. Carvalho, F. a et al. Crohn's disease adherent-invasive *Escherichia coli* colonize and induce strong gut inflammation in transgenic mice expressing human CEACAM. *J. Exp. Med.* 206, 2179–2189 (2009).
 85. Low, D. et al. Chitin-binding domains of *Escherichia coli* ChiA mediate interactions with intestinal epithelial cells in mice with colitis. *Gastroenterology* 145, 602–12.e9 (2013).
 86. Lapaquette, P., Bringer, M.-A. & Darfeuille-Michaud, A. Defects in autophagy favour adherent-invasive *Escherichia coli* persistence within macrophages leading to increased pro-inflammatory response. *Cell. Microbiol.* 14, 791–807 (2012).
 87. Lapaquette, P., Glasser, A. L., Huett, A., Xavier, R. J. & Darfeuille-Michaud, A. Crohn's disease-associated adherent-invasive *E. coli* are selectively favoured by impaired autophagy to replicate intracellularly. *Cell. Microbiol.* 12, 99–113 (2010).
 88. Negroni, A. et al. NOD2 induces autophagy to control AIEC bacteria infectiveness in intestinal epithelial cells. *Inflamm. Res.* 65, 803–813 (2016).
 89. Bringer, M. A., Barnich, N., Glasser, A. L., Bardot, O. & Darfeuille-Michaud, A. HtrA stress protein is involved in intramacrophagic replication of adherent and invasive *Escherichia coli* strain LF82 isolated from a patient with Crohn's disease. *Infect. Immun.* 73, 712–21 (2005).
 90. Bringer, M. A., Billard, E., Glasser, A. L., Colombel, J. F. & Darfeuille-Michaud, A. Replication of Crohn's disease-associated AIEC within macrophages is dependent on TNF- α secretion. *Lab. Investig.* 92, 411–419 (2012).
 91. Meconi, S. et al. Adherent-invasive *Escherichia coli* isolated from Crohn's disease patients induce granulomas in vitro. *Cell. Microbiol.* 9, 1252–1261 (2007).
 92. Nguyen, H. T. T. et al. Crohn's disease-associated adherent invasive *Escherichia coli* modulate levels of microRNAs in intestinal epithelial cells to reduce autophagy. *Gastroenterology* 146, 508–519 (2014).
 93. Wine, E., Ossa, J. C., Gray-Owen, S. D. & Sherman, P. M. Adherent-invasive *Escherichia coli*,

- strain LF82 disrupts apical junctional complexes in polarized epithelia. *BMC Microbiol.* 9, 180 (2009).
94. Craven, M. et al. Inflammation drives dysbiosis and bacterial invasion in murine models of ileal Crohn's Disease. *PLoS One* 7, 1–10 (2012).
 95. Bretin, A. et al. AIEC infection triggers modification of gut microbiota composition in genetically predisposed mice, contributing to intestinal inflammation. *Sci. Rep.* 8, 12301 (2018).
 96. Chassaing, B., Koren, O., Carvalho, F. A., Ley, R. E. & Gewirtz, A. T. AIEC pathobiont instigates chronic colitis in susceptible hosts by altering microbiota composition. *Gut* 63, 1069–1080 (2014).
 97. Drouet, M. et al. AIEC colonization and pathogenicity: Influence of previous antibiotic treatment and preexisting inflammation. *Inflamm. Bowel Dis.* 18, 1923–1931 (2012).
 98. Small, C. L., Xing, L., McPhee, J. B., Law, H. T. & Coombes, B. K. Acute infectious gastroenteritis potentiates a Crohn's disease pathobiont to fuel ongoing inflammation in the post-infectious period. *PLOS Pathog.* 12, e1005907 (2016).
 99. Agus, A., Massier, S., Darfeuille-michaud, A., Billard, E. & Barnich, N. Understanding host-adherent-invasive *Escherichia coli* interaction in Crohn's disease: opening up new therapeutic strategies. 2014, (2014).
 100. Palmela, C. et al. Adherent-invasive *Escherichia coli* in inflammatory bowel disease. *Gut* 67, 574–587 (2018).
 101. Zhang, Y. et al. Identification of candidate adherent-invasive *E. coli* signature transcripts by genomic/transcriptomic analysis. *PLoS One* 10, e0130902 (2015).
 102. Schippa, S. et al. A potential role of *Escherichia coli* pathobionts in the pathogenesis of pediatric inflammatory bowel disease. *Can. J. Microbiol.* 58, 426–432 (2012).
 103. Renouf, M. J., Cho, Y. H. & McPhee, J. B. Emergent behavior of IBD-associated *Escherichia coli* during disease. *Inflamm. Bowel Dis.* (2018).
 104. Yakymenko, O. et al. Infliximab restores colonic barrier to adherent-invasive *E. coli* in Crohn's disease via effects on epithelial lipid rafts. *Scand. J. Gastroenterol.* 1–8 (2018).
 105. Migliore, F., Macchi, R., Landini, P. & Paroni, M. Phagocytosis and epithelial cell invasion by Crohn's disease-associated adherent-invasive *Escherichia coli* are inhibited by the anti-inflammatory drug 6-mercaptopurine. *Front. Microbiol.* 9, 964 (2018).
 106. Olivares-Morales, M. J. et al. Glucocorticoids impair phagocytosis and inflammatory response against Crohn's disease-associated adherent-invasive *Escherichia coli*. *Front. Immunol.* 9, 1026 (2018).
 107. Boudeau, J., Glasser, a-L., Julien, S., Colombel, J.-F. & Darfeuille-Michaud, a. Inhibitory effect of probiotic *Escherichia coli* strain Nissle 1917 on adhesion to and invasion of intestinal epithelial cells by adherent-invasive *E. coli* strains isolated from patients with Crohn's disease. *Aliment. Pharmacol. Ther.* 18, 45–56 (2003).
 108. Huebner, C., Ding, Y., Petermann, I., Knapp, C. & Ferguson, L. R. The probiotic *Escherichia coli* Nissle 1917 reduces pathogen invasion and modulates cytokine expression in Caco-2 cells infected with Crohn's disease-associated *E. coli* LF82. *Appl. Environ. Microbiol.* 77, 2541–2544 (2011).
 109. Jensen, S. R., Fink, L. N., Nielsen, O. H., Brynskov, J. & Brix, S. *Ex vivo* intestinal adhesion of *Escherichia coli* LF82 in Crohn's disease. *Microb. Pathog.* 51, 426–431 (2011).
 110. Ingrassia, I., Leplingard, A. & Darfeuille-michaud, A. *Lactobacillus casei* DN-114 001 inhibits the ability of adherent-invasive *Escherichia coli* isolated from Crohn's disease patients to adhere to and to invade intestinal epithelial cells. *Appl. Environ. Microbiol.* 71, 2880–2887 (2005).
 111. Sivignon, A. et al. *Saccharomyces cerevisiae* CNCM I-3856 prevents colitis induced by AIEC bacteria in the transgenic mouse model mimicking Crohn's disease. *Inflamm. Bowel Dis.* 21, 276–286 (2015).
 112. Van den Abbeele, P. et al. Arabinoxylans, inulin and *Lactobacillus reuteri* 1063 repress the adherent-invasive *Escherichia coli* from mucus in a mucosa-comprising gut model. *NPJ biofilms microbiomes* 2, 16016 (2016).
 113. Rolfe, V. E., Fortun, P. J., Hawkey, C. J. & Bath-Hextall, F. J. Probiotics for maintenance of remission in Crohn's disease. *Cochrane Database Syst. Rev.* (2006).

114. Sivignon, A., Bouckaert, J., Bernard, J., Gouin, S. G. & Barnich, N. The potential of FimH as a novel therapeutic target for the treatment of Crohn's disease. *Expert Opin. Ther. Targets* 14728222.2017.1363184 (2017).
115. Chalopin, T. et al. Inhibition profiles of mono- and polyvalent FimH antagonists against 10 different *Escherichia coli* strains. *Org. Biomol. Chem.* (2015).
116. Sivignon, A. et al. Development of heptylmannoside-based glycoconjugate antiadhesive compounds against adherent-invasive *Escherichia coli* bacteria associated with Crohn's disease. *MBio* 6, 1–9 (2015).
117. Alvarez Dorta, D. et al. The antiadhesive strategy in Crohn's disease: orally active mannosides to decolonize pathogenic *Escherichia coli* from the gut. *ChemBioChem* 17, 936–952 (2016).
118. Chalopin, T. et al. Second generation of thiazolylmannosides, FimH antagonists for *E. coli*-induced Crohn's disease. *Org. Biomol. Chem.* 14, 3913–3925 (2016).
119. Yan, X. et al. Glycopolymers as antiadhesives of *E. coli* strains inducing inflammatory bowel diseases. *Biomacromolecules* 16, 1827–1836 (2015).
120. Bertuccini, L. et al. Lactoferrin prevents invasion and inflammatory response following *E. coli* strain LF82 infection in experimental model of Crohn's disease. *Dig. Liver Dis.* 46, 496–504 (2014).
121. Assa, A. et al. Vitamin D deficiency predisposes to adherent-invasive *Escherichia coli*-induced barrier dysfunction and experimental colonic injury. *Inflamm. Bowel Dis.* 21, 297–306 (2015).
122. Flanagan, P., Campbell, B. J. & Rhodes, J. M. P026 Vitamin D enhances macrophage function and improves killing of Crohn's associated *E. coli*. *J. Crohn's Colitis* 7, S20 (2013).
123. Denizot, J. et al. Diet-induced hypoxia responsive element demethylation increases CEACAM6 expression, favouring Crohn's disease-associated *Escherichia coli* colonisation. *Gut* 64, 428–437 (2015).
124. Di Pasquale, P. et al. Exposure of *E. coli* to DNA-methylating agents impairs biofilm formation and invasion of eukaryotic cells via down regulation of the N-Acetylneuraminatase Lyase NanA. *Front. Microbiol.* 7, 147 (2016).
125. Costanzo, M. et al. Krill oil reduces intestinal inflammation by improving epithelial integrity and impairing adherent-invasive *Escherichia coli* pathogenicity. *Dig. Liver Dis.* 48, 34–42 (2016).
126. Tawfik, A., Flanagan, P. K. & Campbell, B. J. *Escherichia coli*-host macrophage interactions in the pathogenesis of inflammatory bowel disease. *World J. Gastroenterol.* 20, 8751–8763 (2014).
127. Subramanian, S. et al. Replication of colonic Crohn's disease mucosal *Escherichia coli* isolates within macrophages and their susceptibility to antibiotics. *Antimicrob. Agents Chemother.* 52, 427–34 (2008).
128. Dogan, B., Fu, J., Zhang, S., Scherl, E. J. & Simpson, K. W. Rifaximin decreases virulence of Crohn's disease-associated *Escherichia coli* and epithelial inflammatory responses. *J. Antibiot. (Tokyo)*. (2018).
129. Rahimi, R., Nikfar, S., Rezaie, A. & Abdollahi, M. A meta-analysis of broad-spectrum antibiotic therapy in patients with active Crohn's disease. *Clin. Ther.* 28, 1983–1988 (2006).
130. Wang, S. L., Wang, Z. R. & Yang, C. Q. Meta-analysis of broad-spectrum antibiotic therapy in patients with active inflammatory bowel disease. *Exp. Ther. Med.* 4, 1051–1056 (2012).
131. Craven, M. et al. Antimicrobial resistance impacts clinical outcome of granulomatous colitis in Boxer dogs. *J. Vet. Intern. Med.* 24, 819–824 (2010).
132. Brown, C. L., Smith, K., Wall, D. M. & Walker, D. Activity of species-specific antibiotics against Crohn's disease-associated adherent-invasive *Escherichia coli*. *Inflamm. Bowel Dis.* 0, 1 (2015).
133. Galtier, M. et al. Bacteriophages targeting adherent invasive *Escherichia coli* strains as a promising new treatment for Crohn's disease. *J. Crohn's Colitis* 11, 840–847 (2017).
134. Jijie, R. et al. Particle-based photodynamic therapy based on indocyanine green modified plasmonic nanostructures for inactivation of a Crohn's disease-associated *Escherichia coli* strain. *J. Mater. Chem. B* 4, 2598–2605 (2016).
135. Schwartz, A. et al. Microbiota in pediatric inflammatory bowel disease. *J. Pediatr.* 157, 240–

- 244.e1 (2010).
136. Conte, M. P. et al. Adherent-invasive *Escherichia coli* (AIEC) in pediatric Crohn's disease patients: phenotypic and genetic pathogenic features. *BMC Res. Notes* 7, 748 (2014).
 137. Martinez-Medina, M. & Garcia-Gil, L. J. *Escherichia coli* in chronic inflammatory bowel diseases: an update on adherent invasive *Escherichia coli* pathogenicity. *World J. Gastrointest. Pathophysiol.* 5, 213–227 (2014).
 138. Iebba, V. et al. Microevolution in *fimH* gene of mucosa-associated *Escherichia coli* strains isolated from pediatric patients with inflammatory bowel disease. *Infect. Immun.* 80, 1408–17 (2012).
 139. Swidsinski, A. et al. Association between intraepithelial *Escherichia coli* and colorectal cancer. *Gastroenterology* 115, 281–6 (1998).
 140. Bonnet, M. et al. Colonization of the human gut by *E. coli* and colorectal cancer risk. *Clin. Cancer Res.* 20, 859–867 (2014).
 141. Fujita, H. et al. Quantitative analysis of bacterial DNA from *Mycobacteria* spp., *Bacteroides vulgatus*, and *Escherichia coli* in tissue samples from patients with inflammatory bowel diseases. *J. Gastroenterol.* 37, 509–516 (2002).
 142. Raisch, J. et al. Colon cancer-associated B2 *Escherichia coli* colonize gut mucosa and promote cell proliferation. *World J. Gastroenterol.* 20, 6560–6572 (2014).
 143. Blumenthal, R. D., Leon, E., Hansen, H. J. & Goldenberg, D. M. Expression patterns of CEACAM5 and CEACAM6 in primary and metastatic cancers. *BMC Cancer* 7, 2 (2007).
 144. Dogan, B. et al. Evaluation of *Escherichia coli* pathotypes associated with irritable bowel syndrome. *FEMS Microbiol. Lett.* (2018).
 145. Dogan, B. et al. Adherent and invasive *Escherichia coli* are associated with persistent bovine mastitis. *Vet. Microbiol.* 116, 270–282 (2006).
 146. Martinez-Medina, M., Garcia-Gil, J., Barnich, N., Wieler, L. H. & Ewers, C. Adherent-invasive *Escherichia coli* phenotype displayed by intestinal pathogenic *E. coli* strains from cats, dogs, and swine. *Appl. Environ. Microbiol.* 77, 5813–7 (2011).
 147. McPhee, J. B. et al. Host defense peptide resistance contributes to colonization and maximal intestinal pathology by Crohn's disease-associated adherent-invasive *Escherichia coli*. *Infect. Immun.* 82, 3383–93 (2014).
 148. Bringer, M. A., Rolhion, N., Glasser, A. L. & Darfeuille-Michaud, A. The oxidoreductase DsbA plays a key role in the ability of the Crohn's disease-associated adherent-invasive *Escherichia coli* strain LF82 to resist macrophage killing. *J. Bacteriol.* 189, 4860–71 (2007).
 149. 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.* 48, 781–794 (2003).
 150. Sevrin, G. et al. Adaptation of adherent-invasive *E. coli* to gut environment: impact on flagellum expression and bacterial colonization ability. *Gut Microbes* 1–17 (2018).
 151. Vazeille, E. et al. GipA factor supports colonization of Peyer's Patches by Crohn's disease-associated *Escherichia coli*. *Inflamm. Bowel Dis.* 22, 68–81 (2016).
 152. Simonsen, K. T. et al. A role for the RNA chaperone Hfq in controlling adherent-invasive *Escherichia coli* colonization and virulence. *PLoS One* 6, (2011).
 153. Cieza, R. J., Hu, J., Ross, B. N., Sbrana, E. & Torres, A. G. The IbeA invasin of adherent-invasive *Escherichia coli* mediates interaction with intestinal epithelia and macrophages. *Infect. Immun.* 83, 1904–18 (2015).
 154. Dogan, B. et al. Inflammation-associated adherent-invasive *Escherichia coli* are enriched in pathways for use of propanediol and iron and M-cell. *Inflamm. Bowel Dis.* 20, 1919–1932 (2014).
 155. Barnich, N., Bringer, M. A., Claret, L. & Daffeulle-Michaud, A. Involvement of lipoprotein NlpI in the virulence of adherent invasive *Escherichia coli* strain LF82 isolated from a patient with Crohn's disease. *Infect. Immun.* 72, 2484–2493 (2004).
 156. Boudeau, J., Barnich, N. & Darfeuille-Michaud, A. Type 1 pili-mediated adherence of *Escherichia coli* strain LF82 isolated from Crohn's disease is involved in bacterial invasion of intestinal epithelial cells. *Mol. Microbiol.* 39, 1272–1284 (2001).
 157. Gibold, L. et al. The Vat-AIEC protease promotes crossing of the intestinal mucus layer by Crohn's disease-associated *Escherichia coli*. *Cell. Microbiol.* 18, 617–631 (2016).

158. Rolhion, N., Barnich, N., Claret, L. & Darfeuille-Michaud, A. Strong decrease in invasive ability and outer membrane vesicle release in Crohn's disease-associated adherent-invasive *Escherichia coli* strain LF82 with the *yfgL* gene deleted. *J. Bacteriol.* 187, 2286–96 (2005).
159. Rolhion, N. et al. Abnormally expressed ER stress response chaperone Gp96 in CD favours adherent-invasive *Escherichia coli* invasion. *Gut* 59, 1355–62 (2010).
160. Rolhion, N., Carvalho, F. A. & Darfeuille-Michaud, A. OmpC and the σ E regulatory pathway are involved in adhesion and invasion of the Crohn's disease-associated *Escherichia coli* strain LF82. *Mol. Microbiol.* 63, 1684–1700 (2007).
161. Shawki, A. & McCole, D. F. Mechanisms of intestinal epithelial barrier dysfunction by adherent-invasive *Escherichia coli*. *Cmgh* 3, 41–50 (2017).
162. Dumych, T. et al. Oligomannose-rich membranes of dying intestinal epithelial cells promote host colonization by adherent-invasive *E. coli*. *Front. Microbiol.* 9, 742 (2018).
163. Barnich, N. & Darfeuille-Michaud, A. Abnormal CEACAM6 expression in Crohn's disease patients favors gut colonization and inflammation by adherent-invasive *E. coli*. *Virulence* 1, 281–282 (2010).
164. Mazzarella, G. et al. Pathogenic role of associated adherent-invasive *Escherichia coli* in Crohn's disease. *J. Cell. Physiol.* (2017).
165. Dreux, N. et al. Point mutations in FimH adhesin of Crohn's disease-associated adherent-invasive *Escherichia coli* enhance intestinal inflammatory response. *PLoS Pathog.* 9, 1–17 (2013).
166. Vazeille, E. et al. Role of meprins to protect ileal mucosa of Crohn's disease patients from colonization by adherent-invasive *E. coli*. *PLoS One* 6, e21199 (2011).
167. Oshitani, N. et al. Dislocation of tight junction proteins without F-actin disruption in inactive Crohn's disease. *Int. J. Mol. Med.* 15, 407–410 (2005).
168. Kucharzik, T., Walsh, S. V., Chen, J., Parkos, C. A. & Nusrat, A. Neutrophil transmigration in inflammatory bowel disease is associated with differential expression of epithelial intercellular junction proteins. *Am. J. Pathol.* 159, 2001–9 (2001).
169. Gassler, N. et al. Inflammatory bowel disease is associated with changes of enterocytic junctions. *Am. J. Physiol. Liver Physiol.* 281, G216–G228 (2001).
170. Laukoetter, M. G., Bruewer, M. & Nusrat, A. Regulation of the intestinal epithelial barrier by the apical junctional complex. *Curr. Opin. Gastroenterol.* 22, 85–89 (2006).
171. Sasaki, M. et al. Invasive *Escherichia coli* are a feature of Crohn's disease. *Lab. Investig.* 87, (2007).
172. Denizot, J. et al. Adherent-invasive *Escherichia coli* induce claudin-2 expression and barrier defect in CEABAC10 mice and Crohn's disease patients. *Inflamm. Bowel Dis.* 18, 294–304 (2012).
173. Ossa, J. C. et al. Adherent-invasive *Escherichia coli* blocks interferon- γ -induced signal transducer and activator of transcription (STAT)-1 in human intestinal epithelial cells. *Cell. Microbiol.* 15, 446–457 (2013).
174. Colombo, M., Raposo, G. & Théry, C. Biogenesis, secretion, and intercellular interactions of exosomes and other extracellular vesicles. *Annu. Rev. Cell Dev. Biol.* 30, 255–289 (2014).
175. Pêche, H., Heslan, M., Usal, C., Amigorena, S. & Cuturi, M. C. Presentation of donor major histocompatibility complex antigens by bone marrow dendritic cell-derived exosomes modulates allograft rejection. *Transplantation* 76, 1503–10 (2003).
176. Carrière, J., Bretin, A., Darfeuille-Michaud, A., Barnich, N. & Nguyen, H. T. T. Exosomes released from cells infected with Crohn's disease-associated adherent-invasive *Escherichia coli* activate host innate immune responses and enhance bacterial intracellular replication. *Inflamm. Bowel Dis.* 22, 516–28 (2016).
177. Bringer, M. A., Glasser, A. L., Tung, C. H., Méresse, S. & Darfeuille-Michaud, A. The Crohn's disease-associated adherent-invasive *Escherichia coli* strain LF82 replicates in mature phagolysosomes within J774 macrophages. *Cell. Microbiol.* 8, 471–484 (2006).
178. O'Brien, C. L. et al. Comparative genomics of Crohn's disease-associated adherent-invasive *Escherichia coli*. *Gut* 66(8), 1382–1389 (2017).
179. Rahman, K., Sasaki, M., Nusrat, A. & Klapproth, J. M. A. Crohn's disease-associated *Escherichia coli* survive in macrophages by suppressing NF κ B signaling. *Inflamm. Bowel Dis.* 20, 1419–1425 (2014).

180. Sahoo, M., Ceballos-Olvera, I., del Barrio, L. & Re, F. Role of the inflammasome, IL-1 β , and IL-18 in bacterial infections. *ScientificWorldJournal*. 11, 2037–50 (2011).
181. Jarry, A. et al. Subversion of human intestinal mucosa innate immunity by a Crohn's disease-associated *E. coli*. *Mucosal Immunol*. 8, 572–581 (2015).
182. Dunne, K. A. et al. Increased S-nitrosylation and proteasomal degradation of caspase-3 during infection contribute to the persistence of adherent invasive *Escherichia coli* (AIEC) in Immune Cells. *PLoS One* 8, e68386 (2013).
183. Mitchell, D. A. & Marletta, M. A. Thioredoxin catalyzes the S-nitrosation of the caspase-3 active site cysteine. *Nat. Chem. Biol.* 1, 154–158 (2005).
184. Lee, C. M., Kim, B. Y., Li, L. & Morgan, E. T. Nitric oxide-dependent proteasomal degradation of cytochrome P450 2B proteins. *J. Biol. Chem.* 283, 889–98 (2008).
185. Lewis, C et al. *Salmonella enterica* serovar *Typhimurium* HtrA: regulation of expression and role of the chaperone and protease activities during infection. *Microbiology* 155, 873–881 (2009).
186. Fléchar, M., Cortes, M. A. M., Répérant, M. & Germon, P. New role for the *ibeA* gene in H₂O₂ stress resistance of *Escherichia coli*. *J. Bacteriol.* 194, 4550–60 (2012).
187. Chargui, A. et al. Subversion of autophagy in adherent invasive *Escherichia coli*-infected neutrophils induces inflammation and cell death. *PLoS One* 7, e51727 (2012).
188. Vong, L., Yeung, C. W., Pinnell, L. J. & Sherman, P. M. Adherent-invasive *Escherichia coli* exacerbates antibiotic-associated intestinal dysbiosis and neutrophil extracellular trap activation. *Inflamm. Bowel Dis.* 22, 42–54 (2016).
189. Raisch, J., Darfeuille-Michaud, A. & Nguyen, H. T. T. Role of microRNAs in the immune system, inflammation and cancer. *World J. Gastroenterol.* 19, 2985–96 (2013).
190. Dalmaso, G. et al. Crohn's disease-associated adherent-invasive *Escherichia coli* manipulate host autophagy by impairing SUMOylation. *Cells* 8, 35 (2019).
191. Martinez-Medina, M. et al. Biofilm formation as a novel phenotypic feature of adherent-invasive *Escherichia coli* (AIEC). *BMC Microbiol.* 9, 202 (2009).
192. Chassaing, B. & Darfeuille-Michaud, A. The σ E pathway is involved in biofilm formation by Crohn's disease-associated adherent-invasive *Escherichia coli*. *J. Bacteriol.* 195, 76–84 (2013).
193. Chassaing, B. et al. Analysis of the σ E regulon in Crohn's disease-associated *Escherichia coli* revealed involvement of the *naaWVL* operon in biofilm formation. *J. Bacteriol.* 197, 1451–1465 (2015).
194. Lu, Z. et al. Evolution of an *Escherichia coli* protein with increased resistance to oxidative stress. *J. Biol. Chem.* 273, 8308–16 (1998).
195. Delmas, J. et al. Metabolic adaptation of adherent-invasive *Escherichia coli* to exposure to bile salts. *Sci. Rep.* 9, 2175 (2019).
196. Schaible, U. E. & Kaufmann, S. H. E. Iron and microbial infection. *Nat. Rev. Microbiol.* 2, 946–953 (2004).
197. Conte, M. P. et al. The adherent/invasive *Escherichia coli* strain LF82 invades and persists in human prostate cell line rwpe-1, activating a strong inflammatory response. *Infect. Immun.* 84, 3105–3113 (2016).
198. Dreux, N. et al. Ribonucleotide reductase NrdR as a novel regulator for motility and chemotaxis during adherent-invasive *Escherichia coli* infection. *Infect. Immun.* 83, 1305–1317 (2015).
199. Miquel, S. et al. Role of decreased levels of Fis histone-like protein in Crohn's disease-associated adherent invasive *Escherichia coli* LF82 bacteria interacting with intestinal epithelial cells. *J. Bacteriol.* 192, 1832–1843 (2010).
200. Claret, L. et al. The flagellar sigma factor FliA regulates adhesion and invasion of Crohn disease-associated *Escherichia coli* via a cyclic dimeric GMP-dependent pathway. *J. Biol. Chem.* 282, 33275–83 (2007).
201. Penfound, T. A., Smith, D., Elliott, J. F. & Foster, J. W. Control of acid resistance in *Escherichia coli*. 181, 3525–3535 (1999).
202. Iyer, R., Williams, C. & Miller, C. Arginine-arginine antiporter in extreme acid resistance in *Escherichia coli*. *J. Bacteriol.* 185, 6556–6561 (2003).
203. Allen, C. A., Niesel, D. W. & Torres, A. G. The effects of low-shear stress on adherent-invasive *Escherichia coli*. *Environ. Microbiol.* 10, 1512–1525 (2008).

204. Masseret, E. et al. Genetically related *Escherichia coli* strains associated with Crohn's disease. *Gut* 48, 320–325 (2001).
205. Sepehri, S., Kotlowski, R., Bernstein, C. N. & Krause, D. O. Phylogenetic analysis of inflammatory bowel disease associated *Escherichia coli* and the FimH virulence determinant. *Inflamm. Bowel Dis.* 15, 1737–1745 (2009).
206. Nowrouzian, F. L., Wold, A. E. & Adlerberth, I. *Escherichia coli* strains belonging to phylogenetic group B2 have superior capacity to persist in the intestinal microflora of infants. *J. Infect. Dis.* 191, 1078–1083 (2005).
207. Krause, D. O., Little, A. C., Dowd, S. E. & Bernstein, C. N. Complete genome sequence of adherent invasive *Escherichia coli* UM146 isolated from ileal Crohn's disease biopsy tissue. *J. Bacteriol.* 193, 583 (2011).
208. Clarke, D. J. et al. Complete genome sequence of the Crohn's disease-associated adherent-invasive *Escherichia coli* strain HM605. *J. Bacteriol.* 193, 4540–4540 (2011).
209. Vazeille, E. et al. GipA factor supports colonization of Peyer's Patches by Crohn's disease-associated *Escherichia coli*. *Inflamm. Bowel Dis.* 22, 68–81 (2016).
210. Deshpande, N. P., Wilkins, M. R., Mitchell, H. M. & Kaakoush, N. O. Novel genetic markers define a subgroup of pathogenic *Escherichia coli* strains belonging to the B2 phylogenetic group. *FEMS Microbiol. Lett.* 1–7 (2015).
211. Bronowski, C. et al. A subset of mucosa-associated *Escherichia coli* isolates from patients with colon cancer, but not Crohn's disease, share pathogenicity islands with urinary pathogenic *E. coli*. *Microbiology* 154, 571–583 (2008).
212. Sepehri, S. et al. Characterization of *Escherichia coli* isolated from gut biopsies of newly diagnosed patients with inflammatory bowel disease. *Inflamm. Bowel Dis.* 17, 1451–1463 (2011).
213. Chassaing, B., Etienne-Mesmin, L., Bonnet, R. & Darfeuille-Michaud, A. Bile salts induce long polar fimbriae expression favouring Crohn's disease-associated adherent-invasive *Escherichia coli* interaction with Peyer's patches. *Environ. Microbiol.* 15, 355–371 (2013).
214. Hall. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symp. Ser.* 41, (1999).
215. Rice, P., Longden, I. & Bleasby, A. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet.* 16, 276–7 (2000).
216. Leigh, J. W. & Bryant, D. popart: full-feature software for haplotype network construction. *Methods Ecol. Evol.* 6, 1110–1116 (2015).
217. Clinical and Laboratory Standards Institute, Wayne, P. Performance Standards for Antimicrobial Susceptibility Testing; Twenty-Fifth Informational Supplement. (2015).
218. Ruiz del Castillo, B., Moncalián, G. & Martínez-Martínez, L. Variability of *ompC* and *ompF* porin genes in multi-resistant clinical isolates of *E. coli*. (2013).
219. Viveiros, M. et al. Antibiotic stress, genetic response and altered permeability of *E. coli*. *PLoS One* 2, e365 (2007).
220. Huijsdens, X. W. et al. Quantification of bacteria adherent to gastrointestinal mucosa by real-time PCR. *J. Clin. Microbiol.* 40, 4423–7 (2002).
221. Hernández-Allés, S. et al. Porin expression in clinical isolates of *Klebsiella pneumoniae*. *Microbiology* 145, 673–679 (1999).
222. Pfaffl, M. W. A new mathematical model for relative quantification in real-time RT-PCR. *Nucleic Acids Res.* 29, e45 (2001).
223. Spurbeck, R. R. et al. *Escherichia coli* isolates that carry *vat*, *fyuA*, *chuA*, and *yfcV* efficiently colonize the urinary tract. *Infect. Immun.* 80, 4115–22 (2012).
224. Altschul, S. F. et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389–402 (1997).
225. Thompson, J. D., Higgins, D. G. & Gibson, T. J. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* 22, 4673–80 (1994).
226. Bankevich, A. et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* 19, 455–77 (2012).
227. Pareja-Tobes, P., Manrique, M., Pareja-Tobes, E., Pareja, E. & Tobes, R. BG7: A new approach for bacterial genome annotation designed for next generation sequencing data.

- PLoS One 7, e49239 (2012).
228. Darling, A. E., Mau, B. & Perna, N. T. ProgressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. *PLoS One* 5, e11147 (2010).
 229. Grissa, I., Vergnaud, G. & Pourcel, C. CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats. *Nucleic Acids Res.* 35, W52–W57 (2007).
 230. Wang, Y., Coleman-Derr, D., Chen, G. & Gu, Y. Q. OrthoVenn: A web server for genome wide comparison and annotation of orthologous clusters across multiple species. *Nucleic Acids Res.* 43, W78–W84 (2015).
 231. Treangen, T. J., Ondov, B. D., Koren, S. & Phillippy, A. M. The Harvest suite for rapid core-genome alignment and visualization of thousands of intraspecific microbial genomes. *Genome Biol.* 15, 524 (2014).
 232. Milne, I. et al. Using Tablet for visual exploration of second-generation sequencing data. *Brief. Bioinform.* 14, 193–202 (2013).
 233. Gene Ontology Consortium, T. G. O. Gene Ontology Consortium: going forward. *Nucleic Acids Res.* 43, D1049-56 (2015).
 234. Finn, R. D. et al. The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.* 44, D279-85 (2016).
 235. Bidet, P. et al. Combined multilocus sequence typing and O serogrouping distinguishes *Escherichia coli* subtypes associated with infant urosepsis and/or meningitis. *J. Infect. Dis.* 196, 297–303 (2007).
 236. Blanco, M., Blanco, J. E., Alonso, M. P. & Blanco, J. Virulence factors and O groups of *Escherichia coli* isolates from patients with acute pyelonephritis, cystitis and asymptomatic bacteriuria. *Eur. J. Epidemiol.* 12, 191–8 (1996).
 237. Datsenko, K. a & Wanner, B. L. One-step inactivation of chromosomal genes in *Escherichia coli* K-12 using PCR products. *Proc. Natl. Acad. Sci. U. S. A.* 97, 6640–6645 (2000).
 238. Chaverroche, M. K., Ghigo, J. M. & d'Enfert, C. A rapid method for efficient gene replacement in the filamentous fungus *Aspergillus nidulans*. *Nucleic Acids Res.* 28, E97 (2000).
 239. Bioinformatics, B. Fast QC. (2015).
 240. Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* 25, 1105–11 (2009).
 241. Trapnell, C. et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.* 7, 562–578 (2012).
 242. Untergasser, A. et al. Primer3--new capabilities and interfaces. *Nucleic Acids Res.* 40, e115 (2012).
 243. Henderson, I. R., Czczulin, J., Eslava, C., Noriega, F. & Nataro, J. P. Characterization of pic, a secreted protease of *Shigella flexneri* and enteroaggregative *Escherichia coli*. *Infect. Immun.* 67, 5587–96 (1999).
 244. Boisen, N., Ruiz-Perez, F., Scheutz, F., Krogfelt, K. A. & Nataro, J. P. High prevalence of serine protease autotransporter cytotoxins among strains of enteroaggregative *Escherichia coli*. *Am. J. Trop. Med. Hyg.* 80, 294–301 (2009).
 245. Johnson, T. J., Wannemuehler, Y. M. & Nolan, L. K. Evolution of the *iss* gene in *Escherichia coli*. *Appl. Environ. Microbiol.* 74, 2360–9 (2008).
 246. Féria, C., Machado, J., Duarte Correia, J., Gonçalves, J. & Gaastra, W. Distribution of *papG* alleles among uropathogenic *Escherichia coli* isolated from different species. *FEMS Microbiol. Lett.* 202, 205–8 (2001).
 247. Hommais, F. et al. The FimH A27V mutation is pathoadaptive for urovirulence in *Escherichia coli* B2 phylogenetic group isolates. *Infect. Immun.* 71, 3619–22 (2003).
 248. Oberc, A. M., Fiebig-Comyn, A. A., Tsai, C. N., Elhenawy, W. & Coombes, B. K. Antibiotics potentiate adherent-invasive *E. coli* infection and expansion. *Inflamm. Bowel Dis.* (2018).
 249. Hancock, R. E., Puente, J. L. & Calva, E. Role of porins in outer membrane permeability. *J. Bacteriol.* 169, 929–33 (1987).
 250. Smith, S. G. J., Mahon, V., Lambert, M. A. & Fagan, R. P. A molecular Swiss army knife: OmpA structure, function and expression. *FEMS Microbiol. Lett.* 273, 1–11 (2007).
 251. Mittal, R., Krishnan, S., Gonzalez-Gomez, I. & Prasadarao, N. V. Deciphering the roles of

- outer membrane protein A extracellular loops in the pathogenesis of *Escherichia coli* K1 meningitis. *J. Biol. Chem.* 286, 2183–93 (2011).
252. Liu, Y. F. et al. Loss of outer membrane protein C in *Escherichia coli* contributes to both antibiotic resistance and escaping antibody-dependent bactericidal activity. *Infect. Immun.* 80, 1815–22 (2012).
 253. Hejair, H. M. A. et al. Functional role of *ompF* and *ompC* porins in pathogenesis of avian pathogenic *Escherichia coli*. *Microb. Pathog.* 107, 29–37 (2017).
 254. Torres, A. G. & Kaper, J. B. Multiple elements controlling adherence of enterohemorrhagic *Escherichia coli* O157:H7 to HeLa cells. *Infect. Immun.* 71, 4985–95 (2003).
 255. Martínez-Martínez, L. Extended-spectrum β -lactamases and the permeability barrier. *Clin. Microbiol. Infect.* 14, 82–89 (2008).
 256. Maruvada, R. & Kim, K. S. Extracellular loops of the *Escherichia coli* outer membrane protein A contribute to the pathogenesis of meningitis. *J. Infect. Dis.* 203, 131–40 (2011).
 257. Wang, H. et al. Biochemical and functional characterization of the periplasmic domain of the outer membrane protein A from enterohemorrhagic *Escherichia coli*. *Microbiol. Res.* 182, 109–115 (2016).
 258. Negm, R. S. & Pistole, T. G. The porin OmpC of *Salmonella typhimurium* mediates adherence to macrophages. *Can. J. Microbiol.* 45, 658–69 (1999).
 259. Benson, S. A., Occi, J. L. L. & Sampson, B. A. Mutations that alter the pore function of the *ompF* porin of *Escherichia coli* K12. *J. Mol. Biol.* 203, 961–970 (1988).
 260. Zhang, E. & Ferenci, T. OmpF changes and the complexity of *Escherichia coli* adaptation to prolonged lactose limitation. *FEMS Microbiol. Lett.* 176, 395–401 (1999).
 261. Ziervogel, B. K. & Roux, B. The binding of antibiotics in OmpF porin. *Structure* 21, 76–87 (2013).
 262. Lucchini, S., Liu, H., Jin, Q., Hinton, J. C. D. & Yu, J. Transcriptional adaptation of *Shigella flexneri* during infection of macrophages and epithelial cells: insights into the strategies of a cytosolic bacterial pathogen. *Infect. Immun.* 73, 88–102 (2005).
 263. Sato, M. et al. Expression of outer membrane proteins in *Escherichia coli* growing at acid pH. *Appl. Environ. Microbiol.* 66, 943–7 (2000).
 264. Sainz, T. et al. Survival to different acid challenges and outer membrane protein profiles of pathogenic *Escherichia coli* strains isolated from pozol, a Mexican typical maize fermented food. *Int. J. Food Microbiol.* 105, 357–367 (2005).
 265. Vannini, A. et al. The crystal structure of the quorum sensing protein TraR bound to its autoinducer and target DNA. *EMBO J.* 21, 4393–401 (2002).
 266. Allsopp, L. P. et al. UpaH is a newly identified autotransporter protein that contributes to biofilm formation and bladder colonization by uropathogenic *Escherichia coli* CFT073. *Infect. Immun.* 78, 1659–69 (2010).
 267. Eaves-Pyles, T. et al. *Escherichia coli* isolated from a Crohn's disease patient adheres, invades, and induces inflammatory responses in polarized intestinal epithelial cells. *Int. J. Med. Microbiol.* 298, 397–409 (2008).
 268. Olson, N. D. et al. Best practices for evaluating single nucleotide variant calling methods for microbial genomics. *Front. Genet.* 6, 235 (2015).
 269. McElroy, K., Thomas, T. & Luciani, F. Deep sequencing of evolving pathogen populations: applications, errors, and bioinformatic solutions. *Microb. Inform. Exp.* 4, 1 (2014).
 270. den Bakker, H. C. et al. Comparative genomics of the bacterial genus *Listeria*: Genome evolution is characterized by limited gene acquisition and limited gene loss. *BMC Genomics* 11, 688 (2010).
 271. Herring, C. D. et al. Comparative genome sequencing of *Escherichia coli* allows observation of bacterial evolution on a laboratory timescale. *Nat. Genet.* 38, 1406–1412 (2006).
 272. Meng, X. et al. Virulence characteristics of extraintestinal pathogenic *Escherichia coli* deletion of gene encoding the outer membrane protein X. *J. Vet. Med. Sci* 78, 1261–1267 (2016).
 273. Pulkkinen, W. S. & Miller, S. I. A *Salmonella typhimurium* virulence protein is similar to a *Yersinia enterocolitica* invasion protein and a bacteriophage lambda outer membrane protein. *J. Bacteriol.* 173, 86–93 (1991).
 274. Westermann, A. J., Gorski, S. a. & Vogel, J. Dual RNA-seq of pathogen and host. *Nat. Rev. Microbiol.* 10, 618–630 (2012).

275. Haas, B. J. et al. How deep is deep enough for RNA-Seq profiling of bacterial transcriptomes? *BMC Genomics* 13, 734 (2012).
276. Schauer, R. Sialic acids as regulators of molecular and cellular interactions. *Curr. Opin. Struct. Biol.* 19, 507–514 (2009).
277. Xiong, L. et al. Arginine metabolism in bacterial pathogenesis and cancer therapy. *Int. J. Mol. Sci.* 17, 363 (2016).
278. Hanna, A., Berg, M., Stout, V. & Razatos, A. Role of capsular colanic acid in adhesion of uropathogenic *Escherichia coli*. *Appl. Environ. Microbiol.* 69, 4474–81 (2003).
279. Lim, S., Kim, M., Choi, J. & Ryu, S. A mutation in *tdcA* attenuates the virulence of *Salmonella enterica* serovar *Typhimurium*. *Mol. Cells* 29, 509–517 (2010).
280. Ladomersky, E. & Petris, M. J. Copper tolerance and virulence in bacteria. *Metallomics* 7, 957–64 (2015).
281. Torres, A. G., Jeter, C., Langley, W. & Matthyse, A. G. Differential binding of *Escherichia coli* O157:H7 to alfalfa, human epithelial cells, and plastic is mediated by a variety of surface structures. *Appl. Environ. Microbiol.* 71, 8008–15 (2005).
282. Lamarche, M. G. et al. Inactivation of the *pst* system reduces the virulence of an avian pathogenic *Escherichia coli* O78 strain. *Infect. Immun.* 73, 4138–45 (2005).
283. Crane, J. K., Byrd, I. W. & Boedeker, E. C. Virulence inhibition by zinc in shiga-toxigenic *Escherichia coli*. *Infect. Immun.* 79, 1696–705 (2011).
284. Crane, J. K., Naeher, T. M., Shulgina, I., Zhu, C. & Boedeker, E. C. Effect of zinc in enteropathogenic *Escherichia coli* infection. *Infect. Immun.* 75, 5974–84 (2007).
285. Hussain, H. I. et al. Virulence and transcriptome profile of multidrug-resistant *Escherichia coli* from chicken. *Sci. Rep.* 7, 8335 (2017).
286. Grabe, G. J. et al. The *Salmonella* effector SpvD is a cysteine hydrolase with a serovar-specific polymorphism influencing catalytic activity, suppression of immune responses, and bacterial virulence. *J. Biol. Chem.* 291, 25853–25863 (2016).
287. Kim, H. S. & Nikaido, H. Different functions of MdtB and MdtC subunits in the heterotrimeric efflux transporter MdtB(2)C complex of *Escherichia coli*. *Biochemistry* 51, 4188–97 (2012).
288. Jandu, N. et al. Enterohemorrhagic *Escherichia coli* O157-H7 gene expression profiling in response to growth in the presence of host epithelia. *PLoS One* 4, e4889 (2009).
289. Rajkumar, A. P. et al. Experimental validation of methods for differential gene expression analysis and sample pooling in RNA-seq. *BMC Genomics* 16, 548 (2015).
290. Elhenawy, W., Tsai, C. N. & Coombes, B. K. Host-specific adaptive diversification of Crohn's disease-associated adherent-invasive *Escherichia coli*. *Cell Host Microbe* (2019).
291. Casadesús, J. & Low, D. Epigenetic gene regulation in the bacterial world. *Microbiol. Mol. Biol. Rev.* 70, 830–56 (2006).
292. Blyn, L. B., Braaten, B. A. & Low, D. A. Regulation of *pap* pilin phase variation by a mechanism involving differential dam methylation states. *EMBO J.* 9, 4045–54 (1990).
293. Proença, J. T., Barral, D. C. & Gordo, I. Commensal-to-pathogen transition: One-single transposon insertion results in two pathoadaptive traits in *Escherichia coli*-macrophage interaction. *Sci. Rep.* 7, 1–12 (2017).
294. Fang, X. et al. Metagenomics-based, strain-level analysis of *Escherichia coli* from a time-series of microbiome samples from a Crohn's disease patient. *Front. Microbiol.* 9, 2559 (2018).
295. Rolhion, N. & Darfeuille-Michaud, A. Adherent-invasive *Escherichia coli* in inflammatory bowel disease. *Inflamm. Bowel Dis.* 13, 1277–1283 (2007).
296. Lapaquette, P. & Darfeuille-michaud, A. Abnormalities in the handling of intracellular bacteria in Crohn's disease. *J. Clin. Gastroenterol.* 00, 1–4 (2010).

● SUPPLEMENTAL MATERIALS ●

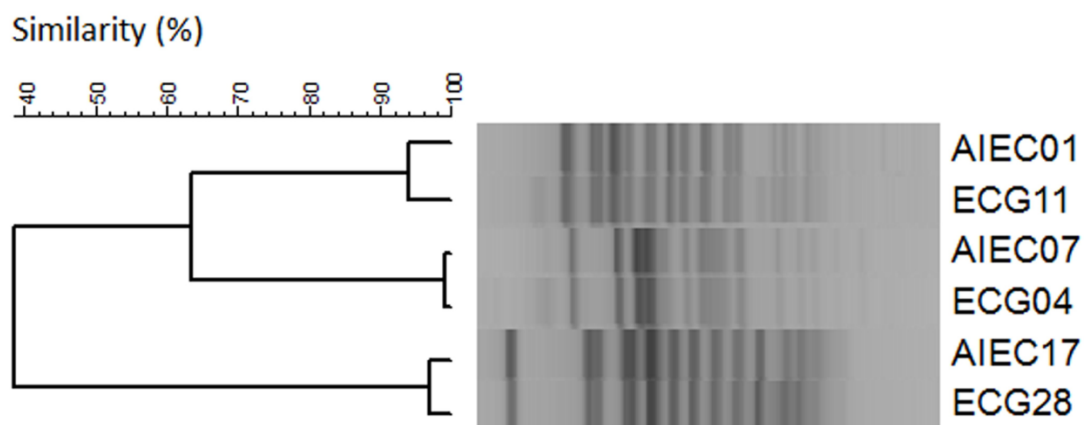


Figure S1. Consensus UPGMA dendrogram generated from the Pearson correlation coefficients of XbaI PFGE profiles of the three pair of strains selected for genome sequencing. Bar indicates profile percentage of similarity.

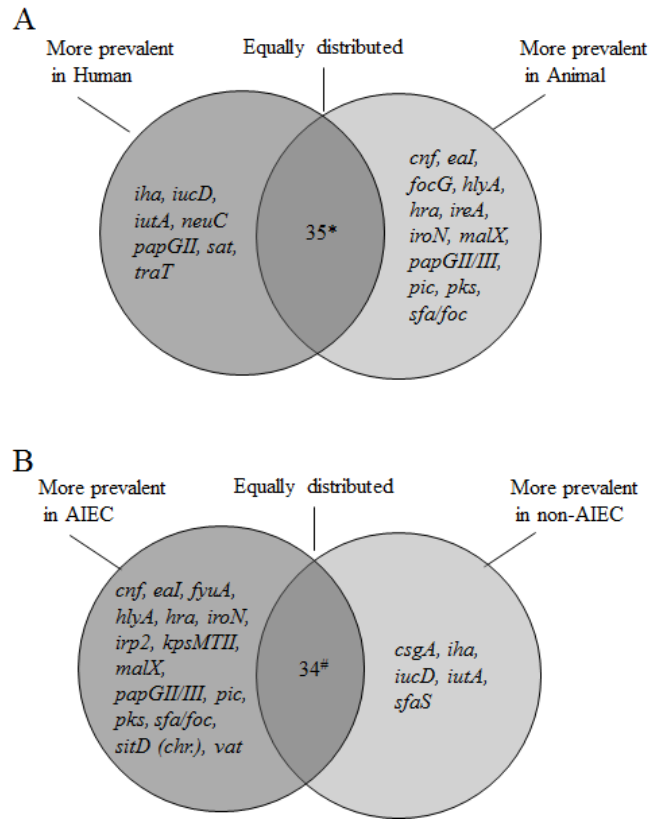


Figure S2. Venn diagram depicting the virulence genes statistically more prevalent in human or animal-isolated strains (A) or in AIEC or non-AIEC strains (B). Genes no differentially distributed are: * *afa/draBC, astA, bmaE, chuA, csgA, cvaB, cvaC, eitA, eitC, etsB, etsC, fimC, fyuA, gafD, gimB, hlyF, ibeA, irp2, iss, kpsMTIII, mat, nfaE, ompA, ompT, papC, papEF, papGI, papGIII, sfaS, sitA, sitD (chr.), sitD (epis.), tia, tsh*, and *vat*. # *afa/draBC, astA, bmaE, chuA, cvaB, cvaC, eitA, eitC, etsB, etsC, fimC, focG, gafD, gimB, hlyF, ibeA, ireA, iss, mat, neuC, nfaE, ompA, ompT, papC, papEF, papGI, papGII, papGIII, sat, sitA, sitD (epis.), tia, traT*, and *tsh*.

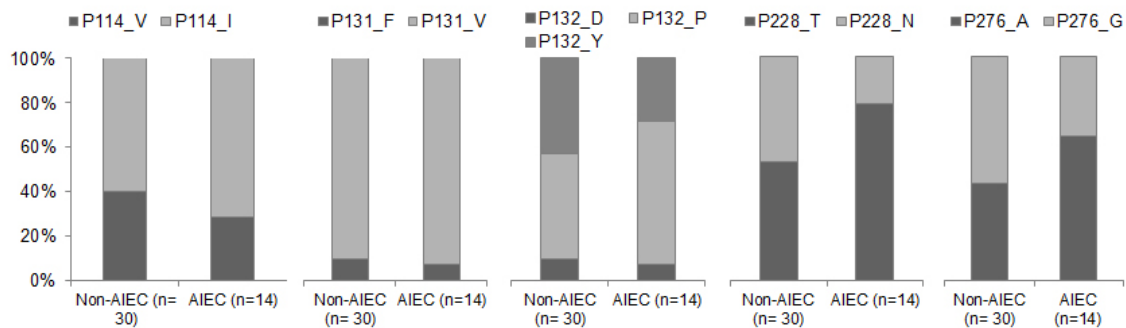


Figure S3. Prevalence of previously detected mutations in OmpA according to pathotype. Each graph represents one amino acid position. Colors indicate the amino acid present.

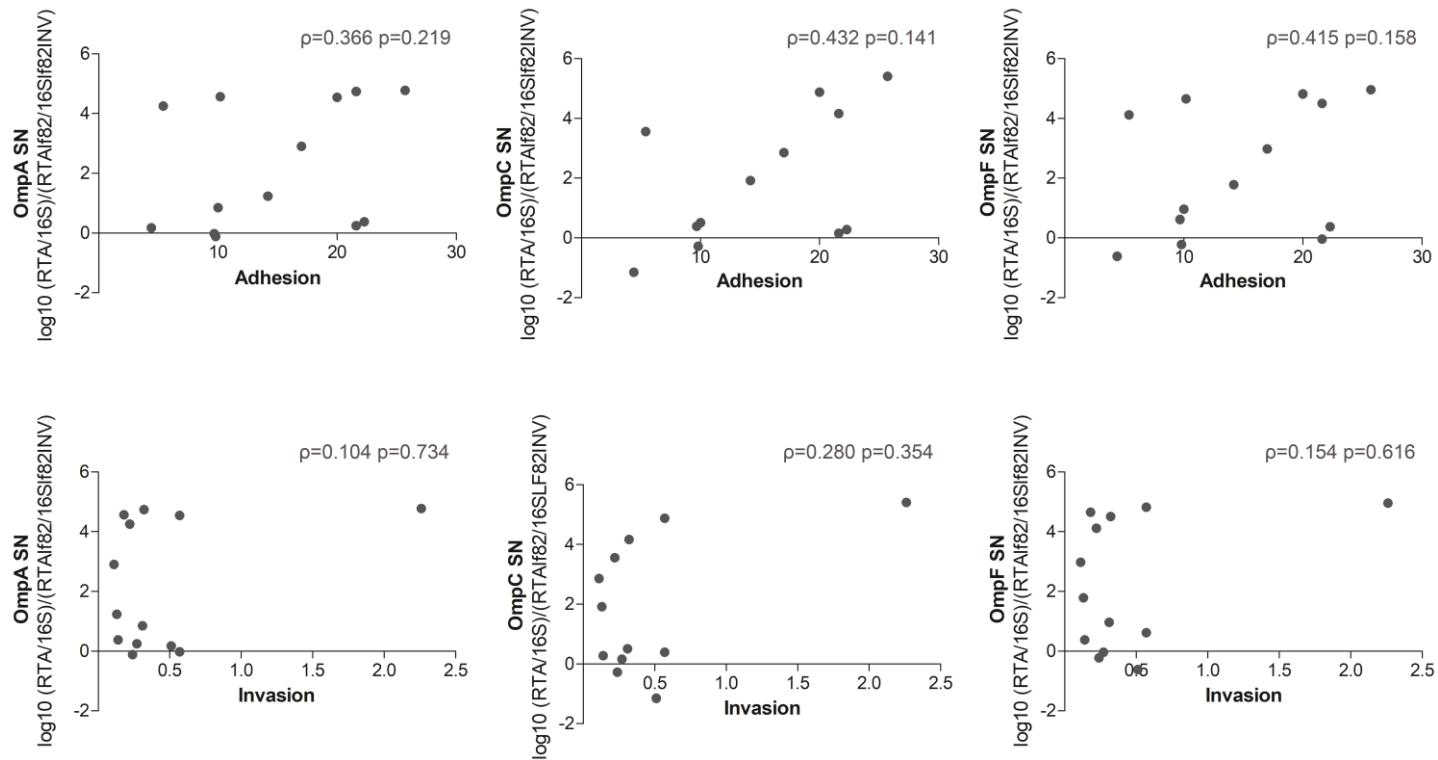


Figure S4. Correlations between OMPs expression and AIEC phenotypic characteristics (adhesiveness and invasiveness) in the SN fraction. Spearman's correlation value (ρ) and significance (p) are indicated. Only AIEC strains are depicted ($n=14$). Adhesion values are depicted as the number of bacteria per I-407 cell and invasion as the percentage of intracellular bacteria relative to the inoculum after 1 h of gentamicin treatment.

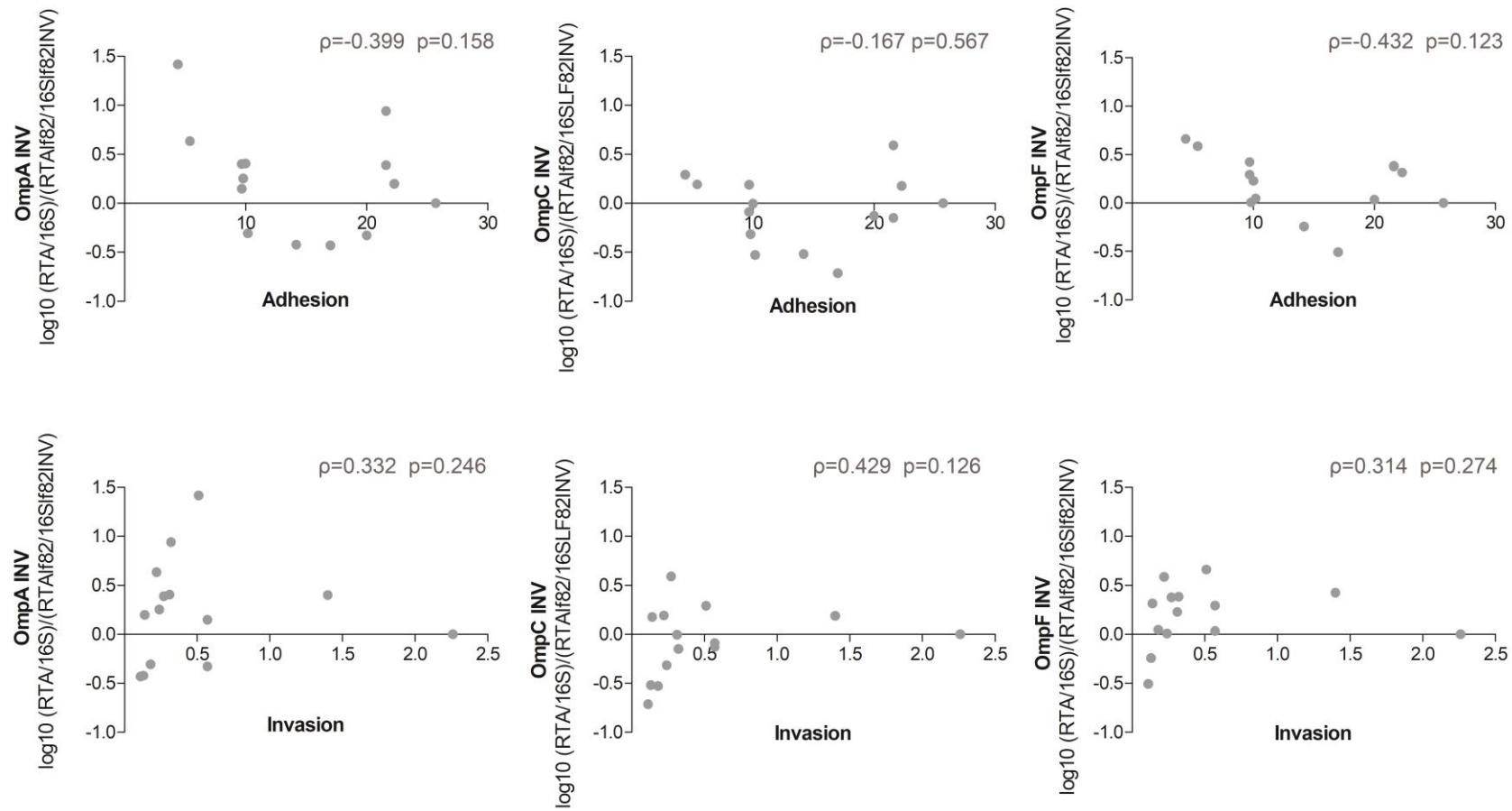


Figure S5. Correlations between OMPs expression and AIEC phenotypic characteristics (adhesiveness and invasiveness) in the INV fraction. Spearman's correlation value (ρ) and significance (p) are indicated. Only AIEC strains are depicted ($n=15$). Adhesion values are depicted as number of bacteria per I-407 cell and invasion as the percentage of intracellular bacteria relative to the inoculum after 1 h of gentamicin treatment.

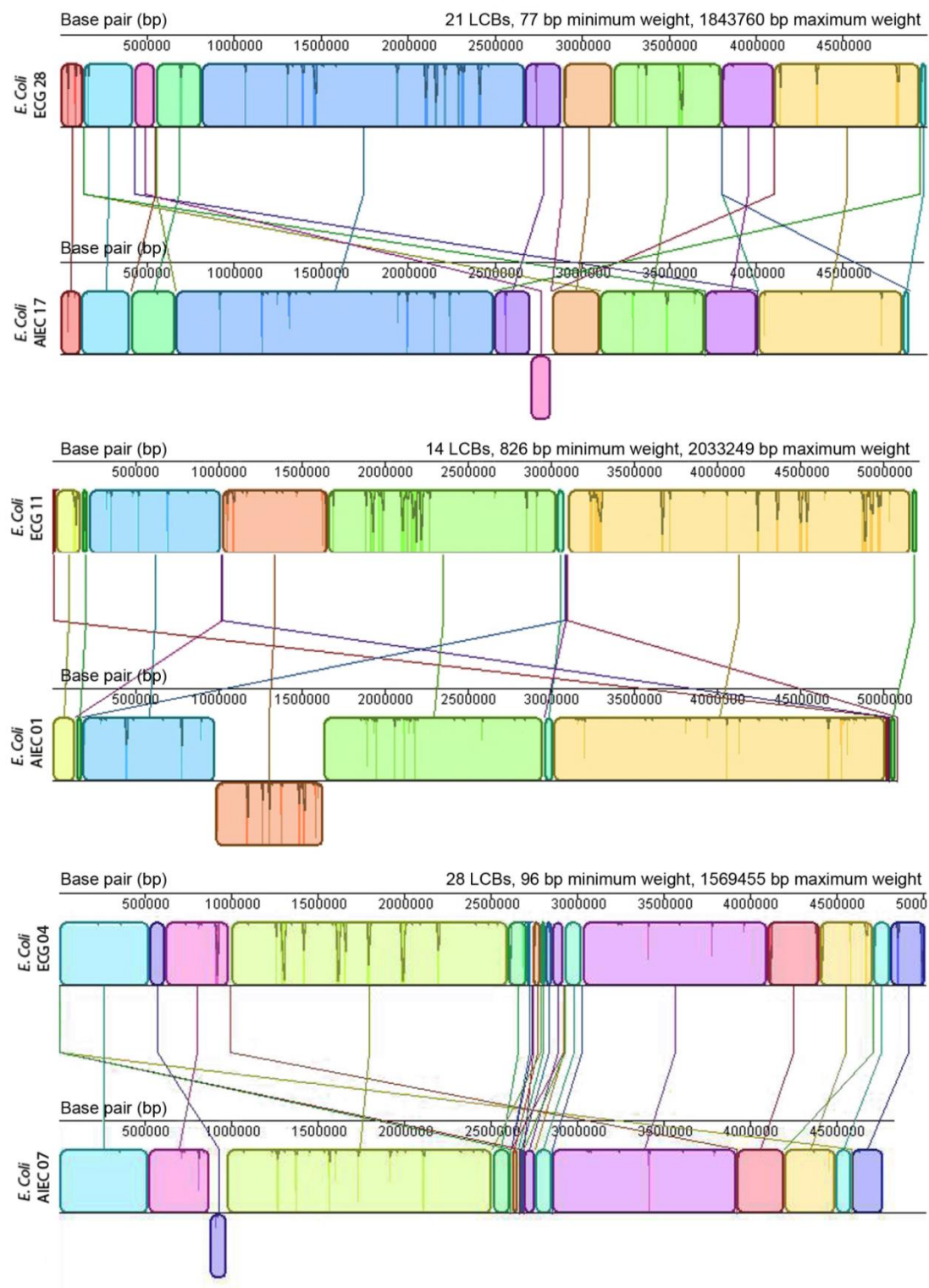


Figure S6. Whole genome map comparison of AIEC/non-AIEC strains with MAUVE 2.3. Boxes of the same colour indicate homologous DNA segments between pairs. Breakpoints in the sequence are represented with the boundaries between the different coloured blocks.

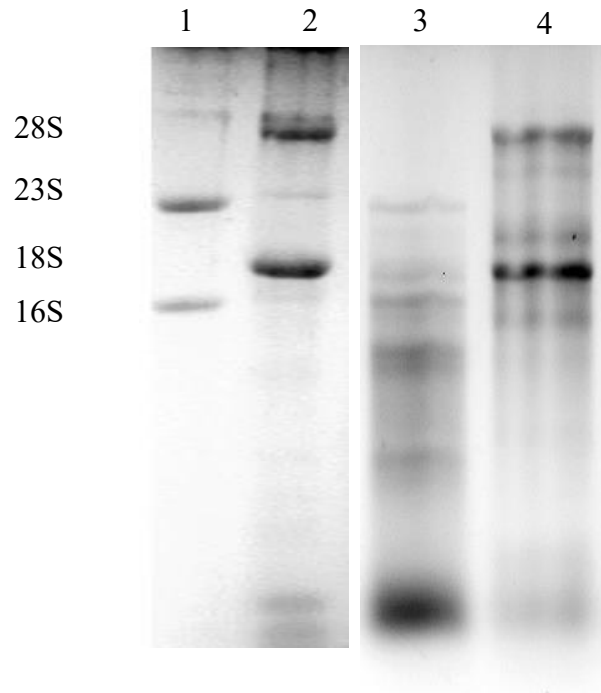


Figure S9. Denaturing agarose gel showing RNA integrity of two SN and INV samples isolated with two different kits. 1: SN total RNA by RiboPure approach A. 2: INV total RNA by RiboPure approach A. 3: SN total RNA by TRIzol approach G. 4: INV total RNA by TRIzol approach G.

Table S1. Information of the patients from whom the UC and CRC strains were isolated.

Strains	Patient ID	Age	Gender	Time since the diagnose (years)	Num. of relapses	Localitization of lesions [¥]	UCDAI	Mayo subscore	Type of lesions (TNM classification)	Surgical resection	Medication	Smoking
UC patients												
PL23F02	107	27	Female	0	0	E2	10	2	na	No	Infliximab	No
PL40G06	121	50	Female	5	nd	E2	nd	2	na	No	Aminosaliclylate	No
GENAIEC13H3	HSC009	52	Male	0	0	E1	6	2	na	No	no	Former smoker
GENAIEC3A9	HSC003	73	Male	10	4	E2	9	2	na	No	Aminosaliclylate	Former smoker
GENAIEC43B3	HT003	36	Male	0	0	E2	6	1	na	No	Aminosaliclylate	No
GENAIEC43B6	HT003	36	Male	0	0	E2	6	1	na	No	Aminosaliclylate	No
GENAIEC43E9	HT003	36	Male	0	0	E2	6	1	na	No	Aminosaliclylate	No
CRC patients												
GENAIEC41B6	HSC021	78	Male	0	na	Rectum	na	na	Neoplasia and metastasis (PT3N0M1)	No	Neoadjuvant therapy	No
GENAIEC42B1	HSC021	78	Male	0	na	Rectum	na	na	Neoplasia and metastasis (PT3N0M1)	No	Neoadjuvant therapy	No

[¥] For UC patients, Montreal classification: E1: proctitis, E2: left-sided colitis, E3: pancolitis. nd: no data; na: not applicable.

Table S3. Distribution of the phylogenetic origin of the strains according to pathotype (A) or origin of isolation (B) in each group of study. Values indicate the percentage of strains present in each condition.

A	VG prevalence all <i>E. coli</i> strains					VG prevalence HUMAN STRAINS					VG prevalence ANIMAL STRAINS			FimH and ChiA variants / AB resistant group of strains *				
	A (N=25)	B1 (N=8)	B2 (N=58)	D (N=12)	P	A (N=9)	B1 (N=8)	B2 (N=29)	D (N=12)	P	A (N=16)	B2 (N=29)	P	A (N=9)	B1 (N=7)	B2 (N=28)	D (N=4)	P
AIEC	36	12.5	62.1	16.7	0.002	44.4	12.5	51.7	16.7	NS	31.3	72.4	0.009	44.4	14.3	53.6	50	NS
Non-AIEC	64	87.5	37.9	83.3		55.6	87.5	48.3	83.3		68.8	27.6		55.6	85.7	46.4	50	
*Excluded: Atypical strain and CRC and UC strains unknown phylogroup																		

B	VG prevalence all <i>E. coli</i> strains					VG prevalence AIEC strains					VG prevalence non-AIEC strains				
	A (N=25)	B1 (N=8)	B2 (N=58)	D (N=12)	P	A (N=9)	B1 (N=1)	B2 (N=36)	D (N=2)	P	A (N=16)	B1 (N=7)	B2 (N=22)	D (N=10)	P
Human	36	100	50	100	<0.001	44.4	100	41.7	100	NS	31.3	100	63.6	100	0.001
Animal	64	0	50	0		55.6	0	58.3	0		68.8	0	36.4	0	

Table S8. Primers used in the study to amplify and analyse differential expression of the selected genes.

Gene code	Comparison	Primer Forward (5'→3')	Rating	T _m (°C)	Primer Reverse (5'→3')	Rating	T _m (°C)	Amplicon length (bp)
XLOC003163	AIEC17vsECG28, INV	TTGAAACCGTAGAAGATGATGC	100	57.32	GCCAGTACAAAGAGAAGATTGCT	90	57.74	151
XLOC000912	AIEC17vsECG28, SN	AATATTTCCGGCAATTCCAC	82	56.93	ATTTGAGCGTTGACACCACA	100	56.84	89
XLOC001058	AIEC17vsECG28, SN	AATACCCGCTTCAGCCATC	100	57.49	GGAGAATTTGCGTCAGTCGT	90	57.5	70
XLOC002831	AIEC17vsECG28, SN	CTCACCGTTCGCAACCAT	92	55.88	TTTTGTGCTGTCTTGAATC	100	55.91	112
XLOC002857	AIEC07vsECG04, INV	AGGGCGACATAATTTTCAGC	90	57.19	GACAATAATGCCACCCAACA	100	56.38	77
XLOC001255	AIEC17vsECG28, SN	CCTCGGTGCTGACGTTATCT	88	57.59	CGGCAGGTAATGGTTTCG	100	57.07	75
XLOC001256	AIEC17vsECG28, SN	CGGGAACGGCAAATAAAAC	100	58.3	CGTCACCGAGAAACAAACCT	100	57.49	78
XLOC000036	AIEC17vsECG28, INV	TCGGGAACACCTCTTTGAAC	100	57.3	CGGTGGTGGAAAGTCTCATTT	100	57.16	90
XLOC001257	AIEC17vsECG28, INV	GTGGAAGCCAATTCGTCAGG	100	58.35	AGCGACAGAATCGGATAGACA	98	57.36	104
XLOC001397	AIEC07vsECG04, INV	GCTGCGGATAGCACGATTAC	99	58.48	CAGGGTGACAGCAAAATACG	100	56.49	170
XLOC1395	AIEC07vsECG04, INV	ACCCGACACCCTATTACCTG	100	56.46	CCATTACGCCCGTCATTT	100	56.47	182
XLOC1396	AIEC07vsECG04, INV	CTGAATTTAAGACTTTACCAGCG	90	56.47	AAACCAGCCAAACGATGC	100	56.45	68
XLOC003046	AIEC17vsECG28, SN	CGTTTGATTATTGAAGAATTAAGG	90	57.89	GCTCTTCTGGATCGGTCCT	91	56.36	184
XLOC000511	AIEC07vsECG04, INV	GAGTCGAACCGGACTAGACG	82	57.08	CGCGTTAACAAAGCGGTTAT	77	58.58	57
XLOC000512	AIEC07vsECG04, INV	GACTTGAACCCGCACAGC	100	56.56	GGATGGTGGAAATCGGTAGAC	100	56.42	67
XLOC000794	AIEC07vsECG04, INV	TTGCCGTATACACACTTTCCA	89	56.75	GGTGAAGTGTCCGAGTGG	100	55.66	54
XLOC1815	AIEC07vsECG04, INV	GTCTCTTAGTTAAATGGATATAACGA	88	57.86	AATCGAACCTGCAATTAGCC	87	57.19	57
gapdh*	Housekeeping gene	CAACTTACGAGCAGATCGAAGC	84	59.64	AGTTTCACGAAGTTGTCGTTCA	82	57.55	170

T_m: primer melting temperature. * Modified from Viveiros et al. 2007.

Table S12. Distribution of FimH amino acid substitutions among the strain collection.

Numbers of strains in each variant are indicated.

	Lectin domain													Pilin domain						
Variant (N)	Amino acid position																			
K12	10	27	33	70	72	74	78	93	106	108	111	117	119	163	166	195	221	237	242	
	A	V	N	N	S	T	S	V	A	Y	P	G	A	V	R	Y	V	T	A	
V1 (N=7)	A																			
V2 (N=1)	A H																			
V3 (N=2)	A F																			
V4 (N=1)	A F																			
V5 (N=1)	A L																			
V6 (N=1)	A K																			
V7 (N=1)	V A																			
V8 (N=4)	A H																			
V9 (N=1)	R																			
V10 (N=1)	P																			
V11 (N=7)																				
V12 (N=2)	A I V																			
V13 (N=3)	A V																			
V14 (N=1)	A I I																			
V15 (N=2)	A N V																			
V16 (N=1)	A I N V																			
V17 (N=2)	S N V																			
V18 (N=8)	A S N																			
V19 (N=2)	A S N V																			
V20 (N=1)	A S N D																			
V21 (N=6)	A S N A																			

Table S13. Distribution of ChiA amino acid substitutions among the strain collection. Numbers of strains in each variant are indicated.

Variant (N)	Amino acid position																					
K12	304	305	314	315-317	326	334	335	336	340	362	370	378	382	388	390	396	414	415	416	427	447	475
	A	V	T	(absent)	S	S	V	N	L	K	K	A	N	E	T	L	V	A	D	D	G	D
V1 (N=13)				PET	N		S			Q	E	V		V		M	I		N	N		
V2 (N=1)				PET	N		S				E	V		V		M	I		N	N		
V3 (N=1)				PET	N		S			Q		V		V		M	I		N	N		
V4 (N=1)				PET	N		S			Q	E	V	D	V		M	I		N	N		
V5 (N=1)				PET	N		S			Q	E	V		V		M	I			N	S	
V6 (N=4)				PET	N		S			Q	E	V		V		M	I	V		N		
V7 (N=5)				PET	N		S			Q	E	V		V		M	I			N		
V8 (N=1)		G	L	PET	N		S		R	Q	E	V		V	R	M	I		N	N		
V9 (N=1)				PET	N		S			Q	E	V				M	I	V				E
V10 (N=1)				PET	N	R						V	D	V		M	I		N	N		
V11 (N=2)					N		G							V								
V12 (N=1)	T						G															
V13 (N=1)							G															
V14 (N=1)	T		H	H			G															
V15 (N=1)								Y														
V16 (N=7)																						

Table S14. Distribution of OmpA amino acid substitutions among the strain collection. Number of strains in each variant and pathotype are indicated.

Variant	Amino acid position	Group of strains			
LF82	26 27 46 87 88 89 114 129 131 132 134-137 139 176 186 200 228 276 N T N S V E V S F D (absent) N H M A T A	AIEC	Non-AIEC	IPEC ¹	ExPEC ¹
Variant 1 (n=5)		2	2	0	1
Variant 2 (n=1)	D	0	1	0	0
Variant 3 (n=3)	D V Y V	1	2	0	0
Variant 4 (n=6)	V Y	1	3	0	2
Variant 5 (n=3)	V Y G	1	1	0	1
Variant 6 (n=15)	I V Y N G	2	9	4	0
Variant 7 (n=1)	D D N I V P GASF D G	0	0	1	0
Variant 8 (n=6)	P D N I A V P GASF D N L N G	1	4	1	0
Variant 9 (n=1)	D D N I I A V P GASF D N L N G	0	1	0	0
Variant 10 (n=1)	Y P D N I I V P GASF D L N G	0	1	0	0
Variant 11 (n=3)	P D N I I A V P GASF D G	1	2	0	0
Variant 12 (n=2)	P D N I I A V P GASF D	0	1	1	0
Variant 13 (n=14)	P D N I I A V P GASF D V	7	6	0	1

¹Gene sequences retrieved from NCBI.

Table S15. Distribution of OmpC amino acid substitutions among the strain collection. Number of strains in each variant and pathotype are indicated.

Variant	Amino acid position																																					
LF82	24	25	28	31	38	47	48	49	50	54	57	85	86	88	89	90	91	92	117	110	138	150	166	174	177	178	179	182	183	184	185	186	187	188	189	191		
	V	Y	D	K	V	N	K	S	E	Q	M	A	P	(absent)	S	E	N	N	I	F	G	F	N	Q	S	V	S	N	D	P	D	F	T	G	H	I		
Variant 1 (n=4)																																						
Variant 2 (n=7)						D			V			S	A		N			V						K				-	-	-	-	-	-	-	-	-	-	M
Variant 3 (n=4)								D	V			S	A		N			V						K				-	-	-	-	-	-	-	-	-	-	M
Variant 4 (n=1)								D	V			S	A		N			V						K				-	-	-	-	-	-	-	-	-	-	M
Variant 5 (n=2)						D			V			S	A		N			V						K				-	-	-	-	-	-	-	-	-	-	M
Variant 6 (n=3)								D	V			S	A		N			V						K				-	-	-	-	-	-	-	-	-	-	M
Variant 7 (n=1)						D			V																													
Variant 8 (n=1)						D			V			S	A		N			V						K				-	-	-	-	-	-	-	-	-	-	M
Variant 9 (n=5)								D	V			S	A		N			V						K	N	P		G	-	-	-	-	-	F	T	G	V	
Variant 10 (n=1)		F	G					D	V			S	A		N			V						K	N	P		G	-	-	-	-	-	F	T	S	V	
Variant 11 (n=3)						D			V			S	A		N			V						K			D	-	-	-	-	-	-	-	-	-	-	M
Variant 12 (n=1)						D			V			S	A		N			V						K			D	-	-	-	-	-	-	-	-	-	-	M
Variant 13 (n=1)						D			V			S	A		N			V						K			D	-	-	-	-	-	-	-	-	-	-	M
Variant 14 (n=1)												S	A		N			V						K			D	-	-	-	-	-	-	-	-	-	-	M
Variant 15 (n=1)												V	T	S	D	N	K	E	V					K				-	-	-	-	-	-	-	-	-	-	M
Variant 16 (n=1)												V	T	S	D	N	K	E	V				Y	K				-	-	-	-	-	-	-	-	-	-	M
Variant 17 (n=1)												V	T	S	A	N	K	E	V				Y	K				-	-	-	-	-	-	-	-	-	-	M
Variant 18 (n=1)									-			V	T	S	D	N	K	E	V				Y	K				-	-	-	-	-	-	-	-	-	-	M
Variant 19 (n=1)												V	T	S	D	N	K	E	V					K				-	-	-	-	-	-	-	-	-	-	M
Variant 20 (n=1)												V	T	S	D	N	K	E	V					K				-	-	-	-	-	-	-	-	-	-	M
Variant 21 (n=1)								D	V			S	A		N					I				T				-	-	-	-	-	-	-	-	-	-	D
Variant 22 (n=2)						D			V															I														
Variant 23 (n=12)						D			V																													
Variant 24 (n=1)						I	D		V																													
Variant 25 (n=1)	I						D		K	K		E				D	S	V		D		D	K		A	H		-	-	-	-	-	-	-	-	-	M	

Table S15. To be continued. ¹Gene sequences retrieved from NCBI.

Variant	Amino acid position																	Group of strains							
	193	197	198	199	200	201	214	220	222	224	231	232	235	236-244	245	246	247	248	250	311	322	AIEC	Non-AIEC	IPEC ¹	ExPEC ¹
LF82	N	K	(absent)	A	L	R	D	V	A	V	W	D	N	(absent)	T	G	L	I	T	V	G				
Variant 1 (n=4)																						2	2	0	0
Variant 2 (n=7)	G							I	G	I	D	S			P	L	Y		N	L		2	3	0	2
Variant 3 (n=4)	G							I	G	I	D	S			P	L	Y		N			1	1	1	1
Variant 4 (n=1)	E							I			D	S			P	L	Y		N	L	-	0	0	1	0
Variant 5 (n=2)	E							I			D	S			P	L	Y		N	L		0	0	2	0
Variant 6 (n=3)	D							I			D	A	T		A	A	Y		N			0	3	0	0
Variant 7 (n=1)											D	A	T		A	A	Y		N			0	0	0	1
Variant 8 (n=1)	D							I	G	I	D	A	T		A	A	Y		N			0	1	0	0
Variant 9 (n=5)	D							I	G	I	D	A	T		A	A	Y		N	L		1	4	0	0
Variant 10 (n=1)	D							I	G	I	D	A	T		A	A	Y		N	L		1	0	0	0
Variant 11 (n=3)	G							I			D	G	SYISNG-VA	R	N	Y						0	2	1	0
Variant 12 (n=1)	G							I			D	A	G	GTYYVDN-VT	H	N	Y					0	1	0	0
Variant 13 (n=1)	G							I			D	A	G	TYVSDNNVV	R	N	Y			L		0	1	0	0
Variant 14 (n=1)	G							I			D	F	GLN--G-YG	E	R	Y	L	N				0	0	1	0
Variant 15 (n=1)	G							I			D	G	SYTSNG-VV	R	N	Y			L			0	1	0	0
Variant 16 (n=1)	G							I			D	F	GL--NG-YG	E	R	Y	L	N				0	1	0	0
Variant 17 (n=1)	G							I			D	F	GL--NG-YG	E	R	Y	L	N				1	0	0	0
Variant 18 (n=1)	G							I			D	F	GL--NG-YG	E	R	Y	L	N				0	1	0	0
Variant 19 (n=1)	G							I			D	A	T		A	A	Y		N			0	1	0	0
Variant 20 (n=1)	D							I			D	A	T		A	A	Y		N			0	1	1	0
Variant 21 (n=1)	Q	G						I	G	I										L		0	1	0	0
Variant 22 (n=2)								I	G	I										L		0	2	0	0
Variant 23 (n=12)																						6	5	0	1
Variant 24 (n=1)																						1	0	0	0
Variant 25 (n=1)	T	D	D	V	F	E	N	I												L		1	0	0	0

Table S16. Distribution of OmpF amino acid substitutions among the strain collection. Number of strains in each variant and pathotype are indicated.

Variant	Amino acid position																												Group of strains					
	48	51	60	99	112	115	118	176	186	187	188	189	190	205	221	222	225	226	227	228	230	264	268	269	270	271	307	308	309	321	AIEC	Non-AIEC	IPEC ¹	ExPEC ¹
LF82	G	E	M	T	Y	V	F	A	D	-	T	A	R	Y	N	L	E	S	S	L	K	T	T	N	T	S	E	G	I	G				
Variant 1 (n=29)																															11	14	0	4
Variant 2 (n=2)																	A	Q	L		N					I					0	0	2	0
Variant 3 (n=13)																	A	Q	P		N										2	9	1	1
Variant 4 (n=3)																	A	Q	P		N	I									0	2	0	1
Variant 5 (n=1)																	A	Q	P		N	I							C		1	0	0	0
Variant 6 (n=1)																	A	Q	P		N							F			1	0	0	0
Variant 7 (n=1)	A	V	K			I											A	Q	P		N						D			0	1	0	0	
Variant 8 (n=2)	D	V	K			I											A	Q	P		N										0	2	0	0
Variant 9 (n=1)	D	V	K			I		T									A	Q	P		N										0	1	0	0
Variant 10 (n=5)	D	V	K	K	F	A	I		A	G	I	P	E	F	D	A	A	E	F	R	Q		E	G	-	-	-	-	-	-	1	3	1	0

¹Gene sequences retrieved from NCBI.

Table S18. OPMs gene expression according to the phylogenetic origin of the strains. The atypical non-AIEC strain was discarded. Gene expression values are given in RTA/16Snorm¹.

	Phylogroup	n	OmpA			OmpC			OmpF		
			mean	desv	p-value	mean	desv	p-value	mean	desv	p-value
SN ²	A	6	315.56	315.17	0.161	546.22	545.10	0.877	903.75	902.70	0.371
	B1	5	7054.71	6923.46		15108.77	14975.26		13245.57	13103.81	
	B2	22	8675.43	3836.12		13215.96	11569.59		8813.16	4595.42	
	D	5	63.33	58.62		73.97	54.79		94.21	79.18	
INV ³	A	7	7.23	2.76	0.417	22.45	15.19	0.226	7.49	3.68	0.171
	B1	5	739.71	666.55		7772.53	7716.07		4158.63	4031.74	
	B2	23	159.33	109.05		65.51	48.57		31.85	24.41	
	D	5	1.77	0.64		2.58	0.73		1.63	0.32	

¹RTA/16S norm= (RTA/16S sample)/(RTA/16S LF82 INV) where RTA=Efficiency [^](Ct target gene reference strain – Ct target gene sample) / Efficiency [^](Ct constitutive gene reference strain – Ct constitutive gene sample). ² SN condition = bacteria growing in suspension . ³INV condition = bacteria adhering and invading intestinal epithelial cells.

Table S19. Distribution of amino acid substitutions in three genes previously associated with AIEC pathogenesis in our AIEC/non-AIEC strain pairs. Amino acid substitutions that were previously associated with AIEC are marked in bold. The first letter corresponds to the amino acid present in the studied strain; the last letter indicates the amino acid found in the commensal strain K-12.

AIEC-associated genes	Strains of study		
	AIEC17-ECG28	AIEC01-ECG11	AIEC07-ECG04
<i>fimH</i> ¹⁶⁵	A27V, S70N , N78S	A27V, H166R	A27V, K32N*
<i>ompA</i> ¹⁵⁹	D46N, V114I , V196A, T224N , A272G	P46N, D47S, N48V, I49E, A125S, P128Y, GASF130-, D135N, V196A, T224N , A272G	P46N, D47S, N48V, I49E, V114I , A125S, P128Y, GASF130-, D135N, N172H, L182M,
<i>chiA</i> ⁸⁵	T100N, G166S, M182T, A200S, T286S, ETPV311, N326S, S335V, Q362K , E370K , V378A , V388E , M396L, I414V, N427D, T517A, E548V , A681D, R696K, S804S, Y810H, G811P	G166S, A200S, T286S, ETPV311-, N326A, S335V, Q362K , E370K , V378A , V388E , M396L, I414V, N416D, N427D, T517A, E548V , R696K	Absent

*Non-synonymous SNP reported in this study by comparative genomics, not associated with AIEC pathotype only present in AIEC07.

Table S25. Level of expression of genes previously associated with AIEC pathogenesis in AIEC and non-AIEC strains analysed in this study.

Gene	UM146 (location in genome)	FPKM AIEC17. SN	FPKM ECG28. SN	fold change	FPKM AIEC17. INV	FPKM ECG28. INV	fold change	FPKM AIEC07. SN	FPKM ECG04. SN	fold change	FPKM AIEC07. INV	FPKM ECG04. INV	fold change
<i>fis</i>	12925-13221	30.4867	31.5483	0.049385	2229.21	2098.74	-0.087010	27.0166	28.8014	0.092291	ND	ND	
<i>lpfA</i>	2542454-2542900	ND	ND		ND	ND		ND	ND		ND	ND	
<i>fucO</i>	554907-556058	43.9419*	42.1843*	-0.058890	13.4742*	10.3821*	-0.376094	213.702*	188.262*	-0.182858	31.3657*	32.2368*	0.039522
<i>fucA</i>	554235-554882	43.9419*	42.1843*	-0.058890	13.4742*	10.3821*	-0.376094	213.702*	188.262*	-0.182858	31.3657*	32.2368*	0.039522
<i>fimH</i>	4654816-4655718	61.3228	59.0376	-0.054790	78.6571	34.4909	-1.18936	ND	ND		ND	ND	
<i>ompA</i>	2624968-2626008	151.326**	152.736**	0.013378	25.3916**	29.1033**	0.196831	61.1365**	63.8808**	0.063349	16.6918**	18.2744**	0.130686
<i>ompC</i>	1193356-1194447	111.514***	109.705***	-0.023595	193.549***	296.601***	0.615821	237.377***	231.636***	-0.035322	122.287***	128.139***	0.067433

ND: Not detected. *covers between 554116-556918. **covers between 2617999-2626537. ***covers between 1182789-1194447.