



**HAL**  
open science

# Metadetect : detection of Shiga toxin-producing Escherichia coli with novel metagenomics approaches and its application on dairy farms in France and Germany.

Sandra Jaudou

## ► To cite this version:

Sandra Jaudou. Metadetect : detection of Shiga toxin-producing Escherichia coli with novel metagenomics approaches and its application on dairy farms in France and Germany.. Microbiologie et Parasitologie. École nationale vétérinaire - Alfort, 2023. Français. NNT : 2023ENVA0004 . tel-04426290

**HAL Id: tel-04426290**

**<https://theses.hal.science/tel-04426290>**

Submitted on 30 Jan 2024

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

NNT : 2023ENVA0004

**METADETECT- DETECTION OF SHIGA TOXIN-PRODUCING  
*ESCHERICHIA COLI* WITH NOVEL METAGENOMICS  
APPROACHES AND ITS APPLICATION ON DAIRY FARMS  
IN FRANCE AND GERMANY**

---

**THESE DE DOCTORAT**

pour obtenir le grade de

**Docteur de l'École nationale vétérinaire d'Alfort**

Spécialité : Microbiologie

École doctorale n°581

Agriculture, alimentation, biologie, environnement et santé (ABIES)

*par*

**Sandra JAUDOU--DUREUIL**

Directrice de thèse : Sabine DELANNOY

Co-encadrants de thèse : Josephine GRÜTZKE et Patrick FACH

**Thèse présentée et soutenue à Maisons-Alfort, le 19 décembre 2023**

**Préparée dans l'Unité « COLiPATH - LSA Laboratoire de Sécurité des Aliments - ANSES »**

**Composition du jury avec voix délibérative :**

Présidente	Catherine SCHOULER, Directrice de recherche, INRAE (Université de Tours)
Rapporteur & Examineur	Mickaël DESVAUX, Directeur de recherche, INRAE (Université Clermont-Auvergne)
Rapporteur & Examineur	Eric OSWALD, Professeur, CHU Toulouse
Examineur	Stefano MORABITO, Chercheur, Istituto Superiore di Sanità (Italie)

*A contribution of new sequencing technologies in risk assessment*

This work was a collaborative project between the French agency for food, environmental and occupational health and safety (Anses) and the German Federal Institute for Risk Assessment (BfR).

I have been working in both institutes in the COLiPATH unit of the Laboratory for Food safety of Maisons-Alfort at ANSES and in both the National Study Center for Sequencing in Risk Assessment (4NSZ) and the National Reference Laboratory for *E. coli* (NRL *E. coli*) of the department of biological safety at the BfR.

---

## Acknowledgment

---

First, I am thankful to Pr Eric Oswald and Dr Mickaël Desvaux for accepting to be reviewers of this PhD thesis as well as Dr Catherine Schouler and Dr Stefano Morabito for accepting to be part of my jury members to evaluate this work.

I could not have undertaken this journey without my PhD committee members, Dr Karine Laroucau, Dr Fabrice Touzain, Pr Stéphane Bonacorsi who generously provided knowledge, expertise and support.

This PhD thesis is the outcome of the last three years of work funded by French agency for food, environmental and occupational health and safety (Anses) and the German Federal Institute for Risk Assessment (BfR).

I would like to express my deepest gratitude to Dr Sabine Delannoy, my daily supervisor for her unlimited support, for always guiding me, for her precious feedback and for her patience. Dr Josephine Grützke my co-supervisor, for her pertinent scientific advices, her attention, her support and her understanding. Dr Patrick Fach, my co-supervisor and head of the COLiPATH unit where I conducted my PhD. I would like to thank them for always encouraging me, trusting me and guiding me through my journey.

Words cannot express my gratitude to Mai-Lan Tran who showed me how to carry out all the experiments in the laboratory, for her precious help, her support, her understanding and her incredible ability to listen. She is not only a colleague but she became a friend. I had the pleasure of sharing my office and working with Fabien Vorimore, who encouraged me, helped me to conduct bio-informatics analyses and gave pertinent advices.

Additionally, this work would not have been possible without the precious collaboration with the National Study Center for Sequencing in Risk Assessment (4NSZ) and National Reference Laboratory for *E. coli* (NRL *E. coli*) groups of the food safety department of the German Federal Institute for Risk Assessment (BfR, Berlin) where I spent 2 times 6 months. I learned a lot not only scientifically but also personally. I am really thankful to each person who has helped in any way. Pr Dr Burkhard Malorny (head of the 4NSZ group) for being there when I needed, his support and for doing anything possible for my stay to be the best. Dr Elisabeth Schuh (head of the NRL *E. coli*) for always taking the time to discuss specific points of the project, her advices and her encouragement. Dr Carlus Deneke (bio-informatician in the 4NSZ group) for the numerous discussions between biology and bioinformatics that were precious to conduct this work, for his help on the bio-informatical part of the project, for his availability and for his support. André Goehler (deputy head of the NRL *E. coli*) for always being available when I needed and for the long hours he had to stay while I was in the lab, for his advices and encouragement. I am also thankful to all of them for their previous advices, time and wonderful support all along the project.

I would like to thank Christine Duvaux-Ponter (deputy of ED ABIÉS) and Alexandre Pery (director of ED ABIÉS) for accepting me in their program but also for their advices and remarkable availability.

The application of the method on natural samples could not have been possible without the help of Pr Stéphane Bonacorsi and his group, Dr Stefano Morabito (head of the European reference laboratory for *E. coli* at the Istituto di Sanità in Rome), colleagues from the LDA39 and a beef company who sent positive samples to test the method.

I am grateful to all members of both BfR teams who welcomed me in their group and made everything for my stay to be comfortable. A special intention to Julia, Wiebke, Hamid, Carlus and Holger. Julia, I would like to thank you for welcoming me in Berlin the first day I arrived and for all the moments we have shared at work but also outside. Hamid, for your humor and presence. Wiebke, Carlus and Holger for spending time with me outside of work.

A special thought for the people I met in Berlin and made me enjoy my stay: Hernan, Isidro, Tamino, Aitor, Giorgia, Ilaria, Chiara and Slavdor. Isidro, Tamino and Aitor for having been there for me, for always supporting, listening, for our dinners and movie times but also for our parties. Giorgia, Ilaria, Chiara and Slavador, with whom I shared wonderful moments and I hope you could learn some bachata steps. Your presence was precious to me and I enjoyed my time with you all.

I also would like to thank the other students, now friends, from the open space with whom I shared precious moments: Christina, Arnaud, Rémi, Nina, Justine, Sandrine, and Baudoin. I am grateful to them for their support, our enjoyable moments and for visiting in Berlin.

Additionally, I am thankful to my friends especially Sobika, Loïc, Yassine and Yoann. Sobika that I know since my bachelor degree, with whom I went for a year abroad, and who has always been there for me. Loïc, a long-time friend who knows me by heart and always supported me and cheered me up. Yassine and Yoann, two wonderful persons that I got to know during my studies and with whom I have experienced many different moments, who still support me today and are there to listen to me and take my mind off things. My family (particularly grandparents and parents and my cousin Alyssa) and my boyfriend for supporting me, for listening whenever I needed it but especially for always believing in me.

Lastly, I would like to mention my dance professors Gilbert and Marceline as well as Pascal for their warm and shining atmosphere. I am thankful to all my dance partners with whom I enjoying dancing which changed my ideas, bachata and kizomba teams with whom I spent such good moments with a special intention to Andrew for always accompanying me to dancing events, your good mood and the moments that we shared which helped me to stay positive.

---

## Résumé substantiel

---

### 1- *Escherichia coli*

Les *Escherichia coli* (*E. coli*) sont des bactéries à coloration de Gram négative. Elles font partie de la flore commensale chez l'Homme et peuvent avoir un rôle bénéfique. Bien que la plupart des *E. coli* ne causent pas de maladies, elles possèdent la capacité d'acquérir des facteurs de virulence qui leur confèrent un caractère pathogène. On distingue deux catégories de *E. coli* pathogènes : les *E. coli* responsables de maladies entériques, que l'on appelle *E. coli* diarrhéiques (ECD) et les *E. coli* extra-intestinales (ExPEC) provoquant des maladies au niveau de sites extra-intestinaux tels que la vessie, le sang ou encore le cerveau.

Les *E. coli* diarrhéiques sont classées en différents pathovars en fonction des facteurs de virulences qu'elles possèdent. Le premier pathovar décrit au sein des ECD sont les *Escherichia coli* enteropathogènes (EPEC). Les EPEC possèdent un système d'adhésion par attachement/effacement (lésion A/E) qui provoque la formation d'un piédestal ainsi que l'effacement des microvillosités des entérocytes. Cette lésion est causée par des facteurs du système de sécrétion de type III (T3SS) dont les gènes codants sont présents au sein d'un îlot de pathogénicité appelé le locus d'effacement des entérocytes (LEE). D'autres pathovars ont été décrits en fonction du système d'adhésion ou de colonisation exprimé. Les EAEC utilisent un mécanisme d'adhésion par agrégation ; les DAEC par adhésion diffuse. Les AIEC colonisent les cellules de manière adhérente et invasive alors que les EIEC utilisent l'invasion cellulaire. Les ETEC colonisent les entérocytes avant de produire des toxines spécifiques. Enfin, les STEC regroupent des souches de *E. coli* capables de produire la Shiga toxine. Un sous-groupe de STEC est capable de causer des symptômes sévères chez l'Homme. Ce sous-groupe appelé *E. coli* enterohémorragique (EHEC) comprend les EHEC typiques qui portent le LEE (notamment le gène *eae* codant pour l'intimine responsable de l'adhésion) et les EHEC atypiques qui n'utilisent pas le système de colonisation des EPEC mais des mécanismes alternatifs. Dans ce travail, les souches portant le LEE sont appelées STEC positives pour le gène *eae*. Les EHEC atypiques et plus généralement les STEC peuvent exprimer d'autres systèmes d'adhésion et notamment ceux caractéristiques d'autres pathovars de *E. coli* diarrhéiques ou extra intestinaux.

## 2- La contamination par les STEC : *Escherichia coli* productrices de Shiga toxine

La contamination par les STEC se fait en plusieurs étapes. Dès que la bactérie est ingérée elle doit posséder la capacité d'adhérer au tube digestif pour s'y fixer et ensuite produire la Shiga toxine. Les symptômes varient, allant de la diarrhée aqueuse bénigne à la diarrhée sanglante pouvant évoluer vers des symptômes plus graves tel que le syndrome hémolytique et urémique (SHU). Le taux de mortalité dû à une contamination par les STEC est de 3 à 5% en Europe. Les STEC sont des bactéries que l'on retrouve majoritairement chez les ruminants où elles y sont commensales et ne causent généralement pas de maladies. Bien que les ruminants contaminés par les STEC soient asymptomatiques, ils peuvent contaminer leur environnement ou encore les aliments. Les fèces de ces ruminants peuvent contaminer les sols, l'eau qui va être utilisée pour irriguer les cultures alimentaires ou encore le lait lors de la traite. Les STEC sont des pathogènes à transmission principalement alimentaire, mais la transmission de souches STEC peut se faire par contact direct avec l'animal. En Europe, et c'est notamment le cas en France et en Allemagne, les principales sources de contamination humaines par les STEC sont les viandes hachées et dérivés, suivies du lait cru et des produits à base de lait cru, les végétaux et les eaux de contact (baignades...).

## 3- Réglementation et méthodes de détection des STEC au sein de produits alimentaires

Au niveau européen, un ensemble de règlements relatifs à l'hygiène alimentaire (paquet hygiène) stipule que les produits alimentaires mis sur le marché ne doivent pas être préjudiciables à la santé humaine et/ou animale. Concernant les STEC, il n'existe pas de critère microbiologique permettant de définir un seuil de contamination. Seul un critère microbiologique est décrit pour le cas particulier des graines germées depuis l'épidémie de SHU la plus sévère en Europe causée par des une souche hybride EAEC/STEC de sérotype O104:H4. Les méthodes de référence pour la détection des STEC au sein de matrices alimentaires (ISO/TS13-136 et MLG5C) comprennent une première étape d'enrichissement afin d'obtenir les souches cibles à un niveau détectable. Une étape de PCR est réalisée afin de détecter les gènes de virulences *stx* et *eae*. Dès que le gène *stx* est détecté, des étapes d'isolement sont réalisées pour essayer d'isoler la souche STEC et la caractériser ensuite. En Europe, la plupart des souches responsables des symptômes les plus sévères tel que le SHU, portent en plus de la Shiga toxine, le LEE (gène *eae*). En France, 90% des SHU pédiatriques sont causés par des souches STEC positives pour le gène *eae*, selon l'avis de l'Anses 2023 sur l'analyses de

données TESSy de l'ECDC. Cependant, en raison de la diversité des *E. coli* qui existent et qui peuvent se retrouver dans un même échantillon il est nécessaire d'isoler la souche STEC afin de pouvoir la caractériser, et plus particulièrement lorsque l'on obtient un signal *stx* et *eae* positif par PCR. En effet, cela permet de distinguer la présence d'une souche STEC qui porte le LEE (*eae*) de la présence d'une souche EPEC qui porte le LEE et une souche STEC qui porte le gène *stx* (culture mixte STEC et EPEC). Faire cette distinction est primordiale car la signification en termes de santé publique n'est pas là même dans un cas comme dans l'autre. Cette étape d'isolement, bien que cruciale pour pouvoir caractériser la souche STEC, représente un challenge. En effet, il n'existe, à ce jour, pas de milieu sélectif spécifique à l'isolement des souches STEC préférentiellement aux autres souches de *E. coli*. Du fait du niveau de contamination des STEC qui peut être très bas (10 cellules), la présence d'autres souches d'*E. coli* rend l'isolement des STEC parfois impossible. En conséquence, l'estimation du risque est biaisée et peut conduire d'une part à la destruction de denrée saine, ou d'autre part à la consommation de denrée contaminée.

#### 4- Objectif de la thèse

Le but de ce projet était donc de développer une méthode qui permette de caractériser les souches STEC directement dans le lait cru sans passer par l'étape d'isolement. Pour cela nous avons testé des approches méta-génomiques qui permettent d'analyser l'ADN de tous les micro-organismes présents au sein de l'échantillon grâce à des approches de séquençage de l'ADN. Deux méthodes de séquençage peuvent être utilisées : le séquençage qui génère de petites séquences d'ADN (séquençage short read) et celui qui génère de longues séquences (séquençage long read). Les méthodes de séquençages short read génèrent de courtes séquences d'ADN (maximum 300bp) et aboutissent souvent à une reconstruction du génome des souches STEC très fragmentée avec au moins 200 contigs. Les deux gènes de virulence *stx* et *eae* étant intégrés à différents endroits du génome, le séquençage short read ne permet pas de caractériser une souche STEC positive pour le gène *eae* au sein d'un méta-génome. En revanche, la méthode de séquençage long read développée par Oxford Nanopore Technology (ONT), génère de longues séquences d'ADN pouvant aller jusqu'à 4 mégabases (Mb) qui pourraient permettre de recouvrir les régions répétées et générer un assemblage avec les gènes *stx* et *eae* co-localisés sur un même contig.



## 5- Développement de la méthode

### 5-1 Optimisation des conditions d'enrichissement

Nous avons déterminé qu'un minimum de 35x de couverture de génome est nécessaire pour obtenir un assemblage correct de souches STEC positives pour le gène *eae* avec l'assembleur Flye. Or, en considérant la flore connexe conséquente présente dans le lait cru ainsi que la quantité de données générées, une phase d'enrichissement de l'échantillon pour multiplier le germe et obtenir, après culture, une concentration suffisante de STEC dans la suspension alimentaire enrichie est nécessaire.

Les conditions d'enrichissement actuelles pour les STEC au sein de matrices alimentaires sont 37°C en présence d'acriflavine pour les produits laitiers dans de l'eau peptonnée tamponnée (EPT). L'acriflavine est un antibiotique permettant de limiter le développement des bactéries à coloration de Gram positive. Néanmoins, ces conditions d'enrichissement et notamment l'utilisation d'acriflavine sont largement discutées au sein de la communauté et il a été envisagé d'enlever l'acriflavine et d'augmenter la température d'enrichissement pour réduire la flore connexe. Différentes conditions ont donc été comparées sur du lait cru de vache artificiellement contaminé avec une souche STEC positive pour le gène *eae* de sérotype O26:H11 préalablement caractérisée. Après comparaison des conditions, l'acriflavine permettait de réduire la présence de bactéries à coloration de Gram positive, très présente au sein du microbiote du lait, permettant ainsi la croissance de la souche STEC. Concernant la température d'incubation, peu de différence a été observée et l'acriflavine s'est montrée plus efficace que l'augmentation de température pour enrichir les souches STEC dans un échantillon complexe comme le lait cru (Publication 3).

### 5-2. Extraction d'ADN à partir de lait cru

La technologie de séquençage ONT requiert une grande quantité d'ADN de haut poids moléculaire (1 µg d'ADN génomique) et est sensible aux impuretés qui peuvent bloquer le passage de l'ADN à travers les pores. Afin d'obtenir de l'ADN qui correspond aux attentes requises par ONT, trois méthodes d'extraction d'ADN différentes (l'extraction d'ADN sur billes, sur colonnes et enfin par précipitation de l'ADN) ont été comparées sur plusieurs souches de *E. coli* dont la majorité étaient des STEC. Les méthodes d'extraction sur billes (kit AMPureXP) et par précipitation (kit MasperPure) sont celles qui ont permis d'obtenir les plus grandes quantités d'ADN avec peu de dégradation, bien que la méthode par précipitation

d'ADN permette en plus l'extraction d'ADN avec le moins d'impuretés. Après séquençage MinION, de plus longues séquences ont été obtenues pour les ADN extraits avec la méthode d'extraction par précipitation permettant un assemblage plus complet de souches STEC (Publication 1). En combinant ces données de séquençage long read aux données de séquençages short read, nous avons complètement assemblé 75 génomes d'*E. coli* (dont 71 STEC) d'origine bovine (Publication 2). La méthode de précipitation de l'ADN a ensuite été testée sur lait cru artificiellement contaminé avec une souche STEC positive pour le gène *eae* (O26:H11) et non contaminé. Les résultats ont clairement montré que la méthode d'extraction par précipitation était la plus adaptée ici car elle permet l'extraction d'ADN de haut poids moléculaire en grande quantité, avec de bons ratios de pureté et une valeur d'intégrité de l'ADN suffisante pour pouvoir être séquencée avec la technologie ONT. Cette méthode a donc été sélectionnée pour la suite du projet. L'utilisation de lait congelé a montré un niveau plus élevé de dégradation et sera donc proscrite dans la suite de ce projet (Additional experiment 1).

### 5-3. Développement d'un pipeline facilitant l'analyse des données

L'analyse des données obtenues après séquençage MinION nécessite certaines compétences en bio-informatique. Afin de faciliter l'analyse de ces données, le pipeline STECmetadetector a été développé dans ce projet pour la caractérisation des STEC à partir de données méta-génomique long read. Les données brutes obtenues après séquençage MinION sont des variations de courant spécifiques à chaque nucléotide. La conversion de ce signal électrique en base se fait lors d'une étape appelé base-calling. L'algorithme Guppy développé par ONT était celui le plus approprié pour l'analyse de nos données. Le pipeline STECmetadetector utilise les données qui ont été converties en bases (après l'étape de base-calling) pour au final caractériser la(es) souche(s) *E. coli* présente(s) dans l'échantillon. Contrairement aux approches d'assemblage de méta-génome classique, le pipeline utilise des données de bonne qualité pour classifier les reads et extraire seulement les reads *E. coli* qui sont utilisés pour détecter la présence des gènes *stx* et *eae* ainsi que les gènes associés au sérotype (permettant d'avoir une information quant à la présence de plusieurs souches de *E. coli*), d'une part ; et qui sont assemblés (Flye ou Canu disponible) pour permettre de caractériser la(es) souche(s), d'autre part.

Bien que le STECmetadetector ait été créé pour caractériser spécifiquement les STEC positives pour le gène *eae*, son utilisation est flexible et peut permettre la détection de gènes de virulence

appartenant à d'autres pathovars de *E. coli*, notamment la caractérisation de souches hybrides. De plus, les bases de données utilisées peuvent être modifiées afin de cribler la présence de plasmides, de gènes de résistance aux antibiotiques, etc. Le STECmetadetector est facilement installable et permet une reproductibilité des résultats générés (Publication 3).

## 6- Limites de la méthode

En utilisant les conditions d'enrichissement optimales pour la croissance des STEC dans du lait cru, la méthode d'extraction sélectionnée et le pipeline STECmetadetector on a pu déterminer la limite de caractérisation de souches STEC positives pour le gène *eae*. Pour cela, du lait cru (*stx* et *eae* négatif) a été artificiellement contaminé avec une souche de sérotype O26:H11 (*stx1a*, *eae* positive, ST21) à des niveaux d'inoculation de 500, 50 et 5 CFU.mL<sup>-1</sup>. Grâce à la méthode développée dans ce projet, la souche inoculée a été caractérisée (assemblage des gènes *stx* et *eae* co-localisés sur le même contig) même à un faible niveau de contamination de 5 CFU.mL<sup>-1</sup> (Publication 3).

Cette étude a montré qu'il est possible de caractériser des souches STEC y compris les souches STEC positives pour le gène *eae* grâce à la méta-génomique long read, mais avec certaines limites. En effet, grâce à la PCR digitale quantitative (qdPCR), nous avons pu quantifier la souche STEC présente après enrichissement et déterminé qu'un minimum de 10<sup>8</sup> copies.mL<sup>-1</sup> est requis pour caractériser la souche. De plus, la présence de plusieurs souches *E. coli* peut interférer avec l'analyse en empêchant la croissance de la souche STEC et donc sa caractérisation. Il n'est pas possible à l'heure actuelle de différencier ces souches sur milieu, ni de contrôler leur croissance lors de l'enrichissement. De plus, elles peuvent être présentes à différents ratios et donc entrer en compétition. Or, aucune étude ne montre la quantité de STEC présente dans un échantillon comparé aux souches *E. coli*. Pour estimer la quantité de *E. coli* commensales qui pourraient empêcher la caractérisation des souches STEC positives pour le gène *eae*, une expérience de co-contamination a été réalisée. Pour cela, la souche STEC positive pour le gène *eae* de sérotype O26:H11 (6423-O26) a été inoculée dans du lait pasteurisé à des ratios croissant et décroissant en présence d'une souche *E. coli* commensale (BfR-Ec-19174, O2:H10, *stx* et *eae* négative, isolée de lait cru).

Bien que la souche STEC ait mieux cultivé que la souche commensale dans les conditions d'enrichissement utilisées ici, la présence de la souche STEC à un ratio STEC : commensale inférieur à 10 : 1 empêche sa caractérisation. Plus particulièrement, les assembleurs disponibles

au moment de la réalisation de ce projet ne permettait pas de distinguer deux souches *E. coli* d'une même espèce. Il n'était pas possible avec les outils actuels, d'assembler le génome d'une souche STEC en deçà de ce ratio (Publication 4).

#### 7- Approches alternatives permettant l'identification de souches STEC positives pour le gène *eae*

Nous avons en parallèle cherché à développer des approches bio-informatiques alternatives permettant l'identification de souche STEC positives pour le gène *eae* au sein de données de séquençage métagénomique. Tout d'abord, l'outil strainberry a récemment été développé et permettrait de séparer les souches d'une même espèce à partir d'un assemblage métagénomique. Strainberry a été inclus dans le STECmetadetector mais ne nous a pas permis de différencier correctement deux souches *E. coli*.

Les gènes *stx* et *eae* peuvent être portés individuellement par des souches d'*E. coli* : *stx* pour les STEC et *eae* pour les EPEC. La présence de souches portant les deux gènes *stx* et *eae* conduit à deux hypothèses. Soit la présence de ces deux gènes au sein d'un génome est aléatoire, soit elle nécessite un fond génétique nécessaire à l'intégration et/ou au maintien des deux facteurs de virulence au sein d'un même chromosome. Bien que le phage Stx peut être transmis à différentes souches *E. coli*, de nombreux travaux ont montré qu'un certain nombre de marqueurs génétiques (essentiellement des effecteurs du système de sécrétion de type 3) semblent préférentiellement associés aux EHEC typiques (STEC positives pour le gène *eae*), constituant une sorte de signature génétique permettant leur identification (Bugarel *et al.*, 2010a, 2010b; Coombes *et al.*, 2008; Delannoy *et al.*, 2013a, 2013b; Imamovic *et al.*, 2010; Karmali *et al.*, 2003; Konczy *et al.*, 2008). Ainsi, l'utilisation de ces marqueurs génétiques combinés à la présence des gènes *stx* et *eae* lors de la première étape de screening des denrées alimentaires, a permis de réduire de façon significative la quantité d'échantillons présumés positifs (Delannoy *et al.*, 2022, 2016). Des approches récentes de machine learning ont montré leur puissance pour exploiter les données de séquençage afin de prédire la pathogénicité des souches STEC ou encore d'attribuer la source de contamination de souches STEC responsables de cas humain en se basant sur leur composition génique. Nous avons donc utilisé le machine learning afin d'exploiter la quantité importante de génomes de STEC séquencés pour identifier

une potentielle signature génétique associée aux EHEC typiques (ici souches STEC positives pour le gène *eae*, sans lien clinique).

Pour ceci, nous avons tout d'abord généré une base de données contenant 1425 génomes de *E. coli* de sérogroupes et d'origines géographiques variés. Les génomes ont été annotés en utilisant le génome de la souche Sakai comme référence. Après une analyse du pan-génome de ces souches, différents modèles de machine learning ont été testés sur les données de présence/absence des gènes qui permettraient de prédire la présence d'une souche portant simultanément les gènes *stx* et *eae*. Grâce à cette approche nous avons pu montrer que la combinaison de présence/absence de 6 marqueurs permettait l'identification de souches STEC positive pour le gène *eae* à partir d'un assemblage méta-génomique (metaFlye). Cette approche montre que non seulement il est possible d'identifier des STEC positives pour le gène *eae* à partir d'assemblages obtenus grâce à des méthodes de méta-génomique long read ; mais aussi qu'il existe probablement un fond génétique propice à la présence des deux gènes au sein d'une même souche (Publication 5). L'avantage apporté par cette méthode est l'identification de souches STEC positives pour le gène *eae* lorsque le ratio STEC : autre *E. coli* est inférieur à 10 : 1, même à de faibles couvertures de génome. Elle nécessite tout de même la génération de données suffisantes pour pouvoir être séquencées et assemblées (au moins 3x de profondeur avec Flye). En revanche, le faible nombre de marqueurs permet de maximiser la probabilité d'identifier la souche en cas de couverture incomplète du génome. Cette approche pourrait être implémentée dans le pipeline STECmetadetector pour permettre l'identification des souches STEC positives pour le gène *eae* même quand les conditions requises ne sont pas atteintes, ne permettant pas la caractérisation. Dans une autre mesure, cette méthode pourrait être développée en laboratoire afin de détecter la présence/absence de ces marqueurs par qPCR lors de la première étape de screening des denrées alimentaires et utilisant un algorithme assez simple qui analyserait les résultats de qPCR (combinaison de présence/absence des marqueurs).

## 8- Application de la méthode développée à des échantillons naturellement contaminés

Une fois la méthode développée et les limites déterminées, nous avons testé cette méthode sur des échantillons présumés positifs, c'est-à-dire positifs pour les gènes *stx* et *eae* mais dont la présence des deux gènes au sein d'une même souche n'a pas été confirmée, ou naturellement contaminés. Nous avons reçu des échantillons présumés positifs de fromage au lait cru et de viande hachée de bœuf, ainsi que des échantillons cliniques (confirmation de souche STEC positive pour le gène *eae*). Il est important de noter que ces échantillons n'ont pas été enrichis dans les mêmes conditions que développé ici (en particulier, en absence d'acriflavine). Les résultats ont confirmé les limites de la méthode. En effet, très peu de données appartenant à l'espèce *E. coli* ont été obtenues et probablement l'utilisation d'enrichissement sélectif aurait permis une meilleure croissance des souches. Lorsque les gènes *stx* et *eae* ont été quantifiés en dessous de la valeur seuil et que le ratio STEC : autres *E. coli* était inférieur à 10 :1, la caractérisation n'a pas été possible. En revanche, pour un échantillon, bien que très peu de données *E. coli* aient été générées, et que la souche STEC a été quantifiée en-dessous de la valeur seuil, elle représentait la seule souche *E. coli* présente et a pu être caractérisée par assemblage. Nous avons ensuite appliqué l'approche de machine learning qui a permis de lever le doute lorsque les deux gènes *stx* et *eae* étaient présents mais assemblés sur différents contigs lorsque la présence d'autres souches *E. coli* ne permettait pas d'avoir un ratio STEC : autre *E. coli* de 10 :1.

## Conclusion :

Ce projet a montré le potentiel de la méta-génomique long read afin de caractériser des souches STEC positives pour le gène *eae* au sein de lait cru. Cette approche permet la caractérisation des souches sans avoir recours à l'étape d'isolement. Cependant, les conditions mentionnées ici (au moins  $10^8$  copies.mL<sup>-1</sup> de STEC après enrichissement et 10 fois plus de STEC que d'autres souches *E. coli*) doivent être respectées. L'utilisation du machine learning permet néanmoins l'identification de souches STEC positives pour le gène *eae* lorsque le ratio obtenu n'est pas respecté mais que la quantité de STEC après enrichissement permet d'obtenir suffisamment de données afin que les gènes soient présents dans l'assemblage. Cette méthode ne vise pas à remplacer l'étape de détection des gènes *stx* et *eae* par qPCR mais à remplacer l'étape intensive d'isolement réalisée au sein de laboratoire de référence afin de caractériser les souches potentiellement pathogènes. A l'heure actuelle, l'application de la méthode développée est limitée par la quantité de données générées ainsi que par les outils d'assemblages ne permettant pas de différencier des souches de *E. coli*. Comme nous l'avons observé sur un échantillon, une souche STEC positive pour le gène *eae* de sérotype O157:H7 a été isolée par le laboratoire national de référence des *E. coli* mais n'a pas été détectée par notre méthode. Il est important de noter que pour d'autres échantillons provenant de la même source, nous avons pu caractériser la présence de deux souches STEC positives pour le gène *eae* qui ont un potentiel pathogène. Avec l'optimisation du séquençage MinION et des outils informatiques, il se peut que la méthode développée ici soit applicable dans le futur. En effet, Oxford Nanopore Technology est en constante évolution afin d'améliorer non seulement la quantité de données générées notamment en rendant le séquençage ciblé possible, mais également en augmentant la précision de ces données. Si la promesse faite par ONT en générant des données aussi précises que celles obtenues en séquençage Illumina, on pourrait imaginer que les différentes méthodes bio-informatiques basées sur l'analyse de variant (SNPs, *k*-mers, etc) permettraient de distinguer des souches et donc d'identifier la présence de souche STEC positive pour le gène *eae* au sein du méta-génome de lait.

---

## Contents

---

Acknowledgment .....	3
Résumé substantiel.....	5
Contents .....	15
Figures .....	18
Table.....	19
Chapter 1 – <i>Escherichia coli</i> .....	20
1. <i>Escherichia coli</i> : From commensal to pathogen.....	20
1.1. Enteropathogenic <i>E. coli</i> (EPEC) .....	22
1.2. Enteroaggregative <i>E. coli</i> (EAEC) and Diffusely-adherent <i>E. coli</i> (DAEC).....	22
1.3. Enterotoxigenic <i>E. coli</i> (ETEC).....	22
1.4. Enteroinvasive <i>E. coli</i> (EIEC) and Adherent invasive <i>E. coli</i> (AIEC) .....	23
1.5. Shiga toxin-producing <i>E. coli</i> and its sub-group EHEC .....	23
1.6. Extra-intestinal <i>Escherichia coli</i> (ExPEC) .....	24
1.7. Hybrid or cross-pathotype <i>E. coli</i> .....	24
1.8. STEC importance.....	24
2. <i>E. coli</i> typing methods.....	26
3. STEC characteristic virulence factor: Shiga toxin .....	28
4. STEC capacity to acquire mobile genetic elements (MGEs) increase their pathogenic potential .....	30
4.1. Bacteriophages are highly present in STEC .....	30
4.2. Pathogenicity islands carried by STEC.....	32
4.2.1. Pathogenicity islands.....	32
4.2.2. Locus of enterocyte effacement (LEE) .....	32
4.2.3. O-islands.....	34
4.2.4. Locus of autoaggregation and adhesion .....	34
4.3. Plasmids .....	35
5. Cross-pathotype/Hybrid <i>E. coli</i> .....	37
Chapter 2: STEC, A food safety perspective .....	38
6. STEC implication in hemolytic uremic syndrome in France and Germany .....	38
6.1. Virulence profile of strains causing HUS in France .....	38
6.2. Virulence profile of strains causing HUS in Germany .....	38



7. STEC contamination sources .....	39
8. HUS-causing STEC outbreaks in France and Germany .....	40
9. The challenge of STEC pathogenicity assessment and their regulation in food .....	42
9.1. Development of classification approaches .....	42
9.2. Classification from serotype to virulence gene profile .....	43
9.3. Recent classifications based on <i>stx</i> subtype and additional adhesion genes .....	44
10. ISO/TS-13136:2012 STEC detection method in food products and its drawback .....	46
<b>Chapter 3- Long-read metagenomics as a new approach for STEC characterization .....</b>	<b>48</b>
11. DNA sequencing .....	48
12. MinION sequencing: from principle to data analysis .....	49
12.1. Principle of nanopore sequencing .....	49
12.2. Converting nanopore sequencing data to base sequences .....	49
12.3. Generation of error-prone data .....	50
13. Assembling STEC genome using long-read sequencing data .....	50
13.1. Long reads advantage for STEC genome assembly .....	50
13.2. Long-read assembly strategies .....	52
14. Long-read metagenomics for food-borne pathogens identification .....	54
14.1. How long-read metagenomics might help STEC characterization from complex matrices .....	54
14.2. Why an assembly-based approach is necessary for STEC characterization? ....	54
<b>Chapter 4: Aims of the work .....</b>	<b>57</b>
<b>Chapter 5: Development of a long-read metagenomics approach for identifying <i>ae</i>-positive STEC .....</b>	<b>59</b>
Chapter5-1: Obtaining complete STEC genomes using long-read sequencing .....	59
Publication 1 .....	61
Evaluation of high molecular weight DNA extraction methods for long-read sequencing of Shiga toxin-producing <i>Escherichia coli</i> .....	61
Publication 2 .....	75
Hybrid Assembly from 75 <i>E. coli</i> Genomes Isolated from French Bovine Food Products between 1995 and 2016 .....	75
Additional experiment 1: Assessment of the selected extraction method performances on raw milk. ....	82
Chapter5-2: Characterizing STEC from raw milk using long-read metagenomics ...	90
Publication 3 .....	91

A step forward for Shiga toxin-producing <i>Escherichia coli</i> identification and characterization in raw milk using long-read metagenomics .....	91
Publication 4 .....	107
Exploring long-read metagenomics for full characterization of Shiga toxin-producing <i>Escherichia coli</i> in presence of commensal <i>E. coli</i> .....	107
Chapter5-3: The benefit of machine learning to identify <i>eae</i> -positive STEC from metagenomics data. ....	120
Publication 5 .....	123
Combination of whole genome sequencing and supervised machine learning provides unambiguous identification of <i>eae</i> -positive Shiga toxin-producing <i>Escherichia coli</i> .....	123
Chapter5-4: Application of the developed method on presumptive positive samples or naturally contaminated samples.....	138
Chapter 6: Discussion .....	147
Conclusion .....	161
Literature .....	163
Abbreviations.....	184
Scientific valorization .....	188

---

## Figures

---

Figure 1: Diarrheic <i>E. coli</i> (DEC) pathotypes of pathogenic <i>Escherichia coli</i> .	21
Figure 2: Bubble chart representing the different cross-pathotypes of <i>E. coli</i> .	25
Figure 3: <i>Escherichia coli</i> ( <i>E. coli</i> ) O-antigen structure and the example of O26 O-Antigen cluster (O-AGC).	27
Figure 4: Mode of action of the Shiga toxin in human enterocytes.	29
Figure 5: Representation of the lytic and lysogenic cycle of bacteriophages that can use the <i>stx</i> -phage.	31
Figure 6: Important pathogenicity islands carried by STEC.	33
Figure 7: Shiga toxin-producing <i>Escherichia coli</i> (STEC) causing hemolytic uremic syndrome (HUS) have acquired virulence factors carried by mobile genetic elements (MGEs).	36
Figure 8: Shiga toxin-producing <i>Escherichia coli</i> (STEC) main contamination sources.	41
Figure 9: Schematic representation of the ISO/TS-13136:2012 procedure for STEC detection in food products.	47
Figure 10: MinION sequencing principle.	51
Figure 11: Difference of short- (A) and long- (B) read sequencing on genome assembly and representation of long-read assembler strategies (C).	53
Figure 12: Traditional workflow for genome assembly from metagenomics data.	55
Figure 13: Comparison of DNA yield (A), A260/A280 (B) and A260/A230 (C) values obtained depending on the extraction method used.	85
Figure 14: Gel electrophoresis of DNA extracted from fresh raw milk artificially contaminated with $10^4$ CFU.mL <sup>-1</sup> of STEC after enrichment at 37°C in BPW.	86
Figure 15: Gel electrophoresis of DNA extracted from frozen enriched raw milk artificially contaminated with $10^5$ CFU.mL <sup>-1</sup> of STEC post-enrichment at 37°C in BPW using Lucigen kit (A) and Zymo kit (B), as well as from frozen enriched raw milk without artificial contamination using both Lucigen and Zymo kits (C).	87
Figure 16: Barplot representations of the most abundant genus (>2%) detected in long-read sequencing from A: feces or rectal swab samples from contaminated patients, B: presumptive-positive raw milk cheese enrichments (37°C for 18-24h without acriflavine) and C: presumptive-positive ground beef enrichment broths (37°C for 18-24h without acriflavine).	143
Figure 17: Representation of the workflow developed for <i>eae</i> -positive STEC characterization from raw milk and its limits.	148

---

## Table

---

Table 1: Seropathotype classification of pathogenic STEC proposed by Karmali and colleagues in 2003. ....	43
Table 2: EFSA classification of highly virulent STEC based on Tor1 assessment of 2020. ....	44
Table 3: to Food and Agriculture Organization of the United Nations (FAO) and World health Organization (WHO) approach for classifying STEC pathogenic potential 2018. FAO/WHO. 2018. ....	45
Table 4: Molecular risk assessment of STEC according to the National Advisory Committee on Microbiological Criteria for Foods (NACMCF) and the United States Department of Agriculture (USDA) NACMCF / USDA, 2019. ....	45
Table 5: Anses new classification of STEC pathogenic potential. Anses (2023). ....	46
Table 6: Quantification and purity ratios of DNA extracted from fresh or frozen raw milk after enrichment with or without artificial contamination using STEC. ....	84
Table 7: Properties of DNA extracted from presumptive-positive milk and ground beef enrichment and naturally contaminated feces or rectal swab enrichment. ....	141
Table 8: Summary of STECmetadetector and machine learning results. ....	142
Table 9: Quantification results obtained from extracted DNA on <i>eae</i> -positive STEC virulence factors using qdPCR. ....	146

---

## Chapter 1 – *Escherichia coli*

---

### 1. *Escherichia coli*: From commensal to pathogen

First characterized in 1885 from infant stools by Theodor Escherich, *Escherichia coli* are Gram-negative bacteria from the *Enterobacteriaceae* family (Croxen *et al.*, 2013). *E. coli* are ubiquitous bacteria found in a variety of niches. They are natural members of mammals' microbial flora (Tenaillon *et al.*, 2010). A variety of *E. coli* strains exists; most of them are usually non-pathogenic (commensal) to humans. Many studies have shown the benefits of *E. coli*, such as assisting in the absorption of vitamin K or preventing the colonization of pathogenic bacterial cell in the human gastrointestinal tract. Nevertheless, some of these strains can cause a broad range of human diseases and are known as pathogenic *E. coli* (Croxen *et al.*, 2013). Pathogenic *E. coli* have acquired the ability to colonize specific tissues. Two categories of pathogenic *E. coli* have been described, the diarrheagenic *E. coli* (DEC) and the extra-intestinal pathogenic *E. coli* (ExPEC) (Nataro and Kaper, 1998; Russo and Johnson, 2000).

DEC is a large group of pathogenic *E. coli* that can cause a variety of intestinal symptoms ranging from watery diarrhea to bloody diarrhea. More severe symptoms can lead to kidney failure or even death (Nataro and Kaper, 1998). Among the DEC, seven pathotypes have been defined based on the virulence factors and adhesion mechanisms they have stably acquired /expressed. These are Enteropathogenic *E. coli* (EPEC); Shiga toxin-producing *E. coli* (STEC) among which a subset are Enterohaemorrhagic *E. coli* (EHEC); Enterotoxigenic *E. coli* (ETEC); Enteroinvasive *E. coli* (EIEC), including *Shigella*; Enteroaggregative *E. coli* (EAEC); Diffusive-adherent *E. coli* (DAEC) and the recently defined Adherent invasive *E. coli* (AIEC) (Darfeuille-Michaud *et al.*, 2004; Kaper *et al.*, 2004). Figure 1 represents the described DEC pathotypes with the exception of EIEC, DAEC and AIEC.

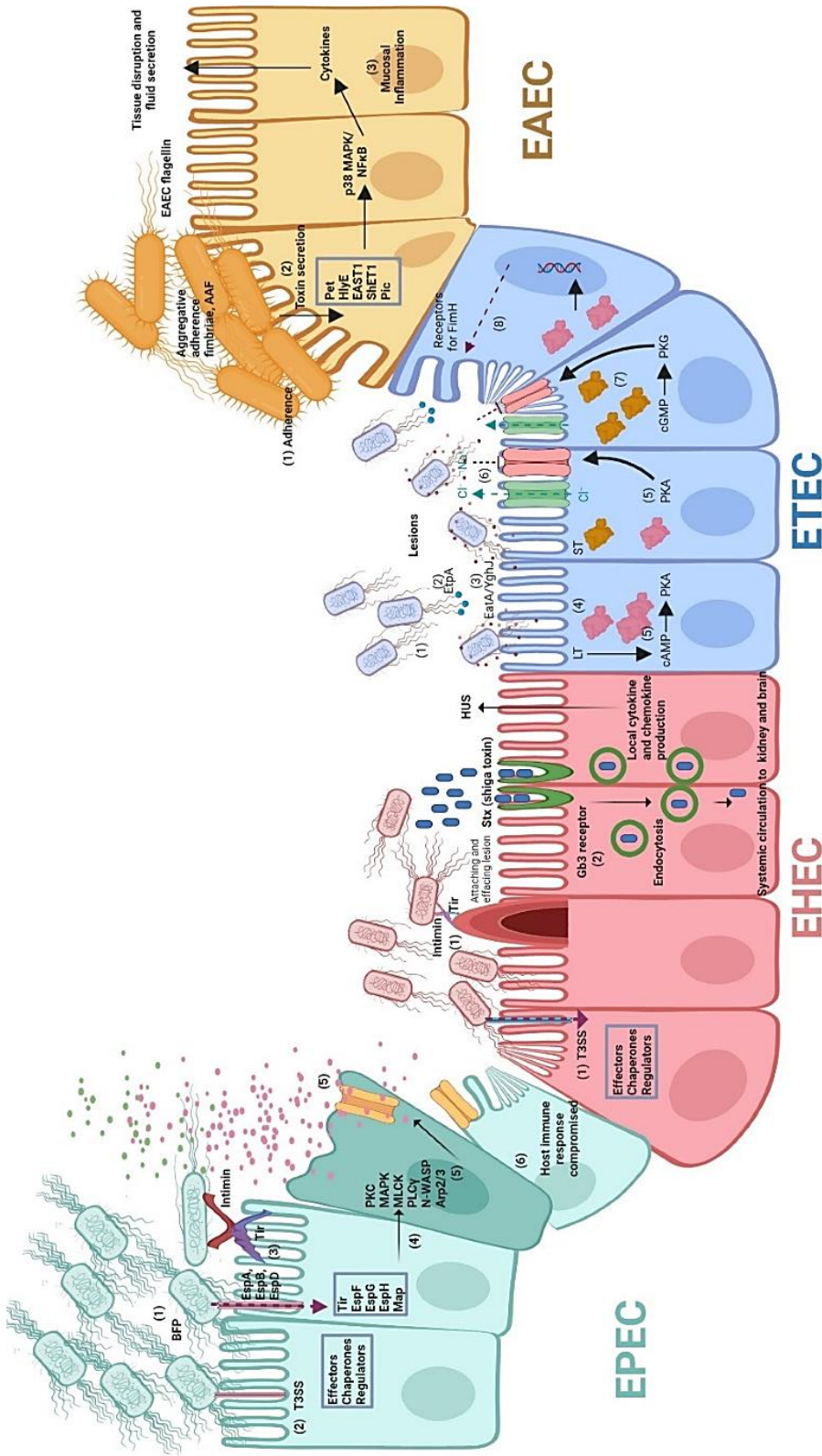


Figure 1 : Diarrhetic *E. coli* (DEC) pathotypes of pathogenic *Escherichia coli*. Taken from Pokharel *et al.*, 2023

### 1.1. Enteropathogenic *E. coli* (EPEC)

The first pathovar identified among diarrheagenic *E. coli* was EPEC, which commonly causes diarrhea in children and animals. EPEC adhere to human intestinal cells by provoking a so-called attaching/effacing (A/E) lesion, which is mediated by a type III secretion system (T3SS). The genes encoding the T3SS are located on a pathogenicity island named Locus of Enterocyte Effacement (LEE). One protein in particular, is the intimin, encoded by the *eae* gene located on the LEE, which is responsible for intimate adhesion of the bacteria to the host cell (Dean-Nystrom *et al.*, 1998). T3SS effectors cause actin cytoskeleton rearrangement, effacement of microvilli and a pedestal formation (Tobe *et al.*, 2006).

### 1.2. Enteroaggregative *E. coli* (EAEC) and Diffusely-adherent *E. coli* (DAEC)

Enteroaggregative *E. coli* (EAEC) is another DEC pathotype able to adhere to enterocytes but using an aggregative adhesion mechanism that resembles a brick pattern. Fimbriar adhesins called aggregative adherence fimbriae (AAFs) are involved in this adhesion mechanism. They are encoded by the *aggR* gene located on a plasmid (pAA). EAEC also carry a gene island *aai* encoding a type VI secretion system (Pakbin *et al.*, 2021). A subclass of EAEC are also responsible for watery diarrhea in children and may be associated with urinary tract infections: Diffusely-adherent *E. coli* (DAEC). Their adhesion mechanism is not fully understood, but they apparently colonize intestinal cells using a diffuse adhesion process by expressing diffuse adherence fimbriae (Meza-Segura *et al.*, 2020).

### 1.3. Enterotoxigenic *E. coli* (ETEC)

One pathotype of DEC not only adheres to intestinal cells, but also produces toxins: Enterotoxigenic *E. coli* (ETEC). ETEC are the main cause of traveler's diarrhea in humans, but also cause severe infant diarrhea. Once they adhere to intestinal cells via colonization factors (CFs), they secrete enterotoxins, either heat-stable and/or heat-labile toxins (STs and LTs, respectively), which mediate watery diarrhea. Colonization mechanism is crucial for ETEC, prior to secreting plasmid-encoded toxin (LT or ST) (Pakbin *et al.*, 2021).

#### 1.4. Enteroinvasive *E. coli* (EIEC) and Adherent invasive *E. coli* (AIEC)

EIEC are a particular type of DEC that use a similar invasion mechanism as *Shigella*. They are intracellular microorganisms that use trans-cytosis to cross cell barriers and navigate from cell to cell leading to their death. They do not express any colonization of flagellar factors (Pakbin *et al.*, 2021). Similarly to EIEC, AIEC are capable of replication inside macrophages. AIEC virulence mechanism is not fully understood and is still being studied. It seems that they are frequently detected in patients suffering from inflammatory bowel diseases, particularly Crohn's disease. It was shown that they adhere and invade the host cells using virulence factors that include type I pili, FimH and invasion protein IbeA (Darfeuille-Michaud *et al.*, 2004, 1998; Mansour *et al.*, 2023).

#### 1.5. Shiga toxin-producing *E. coli* and its sub-group EHEC

Lastly, STEC regroups *E. coli* strains producing a cytotoxin called Shiga toxin (Stx1 and/ or Stx2) (O'Brien *et al.*, 1983). First described in 1977, they were isolated from a patient suffering of diarrhea (Konowalchuk *et al.*, 1977). STEC infections may cause different symptoms including watery diarrhea, hemorrhagic colitis (HC), hemolytic uremic syndrome (HUS) potentially leading to renal failure and thrombocytopenic purpura (Karmali, 1989). The *stx* gene, which encodes the Shiga toxin, is located on a lambdoid bacteriophage named Stx-phage (Krüger and Lucchesi, 2015; Newland *et al.*, 1985). Most STEC causing severe human symptoms are also carrying adhesion or colonization factors. A subset of STEC are Enterohaemorrhagic *E. coli* (EHEC), which are STEC strains causing the life-threatening hemolytic uremic symptom (HUS). Typical EHEC harbor the LEE pathogenicity island from EPEC. On the contrary, atypical EHEC are LEE-negative STEC implicated in HUS cases, but should harbor alternative virulence factors conferring adhesion or colonization of enterocytes (Joseph *et al.*, 2020). In this work, we will use the term STEC to refer to any *E. coli* carrying a Stx-phage, but we will distinguish between LEE-positive STEC (or *eae*-positive) for STEC carrying the LEE and LEE-negative STEC (or *eae*-negative) for STEC that do not harbor the LEE locus.



### 1.6. Extra-intestinal *Escherichia coli* (ExPEC)

Extra-intestinal *E. coli* (ExPEC) are usually asymptomatic in the intestinal gut but carry additional factors that enable their colonization of specific tissues (Joseph *et al.*, 2020; Lo *et al.*, 2015). They have the potential of causing symptoms in extra-intestinal sites such as urinary tract infections (Uropathogenic *E. coli* UPEC), meningitis (Neonatal meningitis-causing *E. coli* NMEC) or even sepsis (Septicemic pathogenic *E. coli* SPEC). The classification of ExPEC is mostly attributed to the presence of extra-intestinal virulence genes, the tissue they colonize and the symptoms that they cause.

### 1.7. Hybrid or cross-pathotype *E. coli*

With the emergence of whole genome sequencing technologies, the genomic plasticity of *E. coli* strains have been highlighted. Since the severe outbreak of 2011 caused by a hybrid STEC, the analysis of virulence genes present in available genomes have demonstrated the existence of *E. coli* strain harboring sets of virulence factors from more than one *E. coli* pathotype. The terms cross-pathotypes or hybrid *E. coli* are used to describe such strains. Nowadays, STEC hybrids or cross pathotypes are increasingly described. In particular, hybrid STEC such as O104:H4 EAEC/STEC or O80:H2 ExpEC/STEC/EPEC which were implicated in severe human symptoms have been intensively studied to understand their evolution process (details in Section 1.5 of Chapter 1). It is noteworthy that some hybrid *E. coli* can carry genes from different pathotypes among diarrheic *E. coli*, but other harbor virulence factors from both diarrheic and extra-intestinal *E. coli*, as represented on Figure 2 (Santos *et al.*, 2020).

### 1.8. STEC importance

Most pathogenic *E. coli* pathotypes are of public health concern but STEC are monitored in Europe since they are zoonotic food-borne pathogens frequently associated with severe symptoms as HC or HUS (EFSA, 2020). Severe forms of STEC infections are mostly manifested in infants, elderly and immunocompromised patients (Ochoa and Cleary, 2003). However, the severe outbreak in 2011 has shown that every individual may be subjected to severe forms of STEC infections (Tozzoli *et al.*, 2014). The mortality rate caused by STEC infection is around 3-5% (Thorpe, 2004).

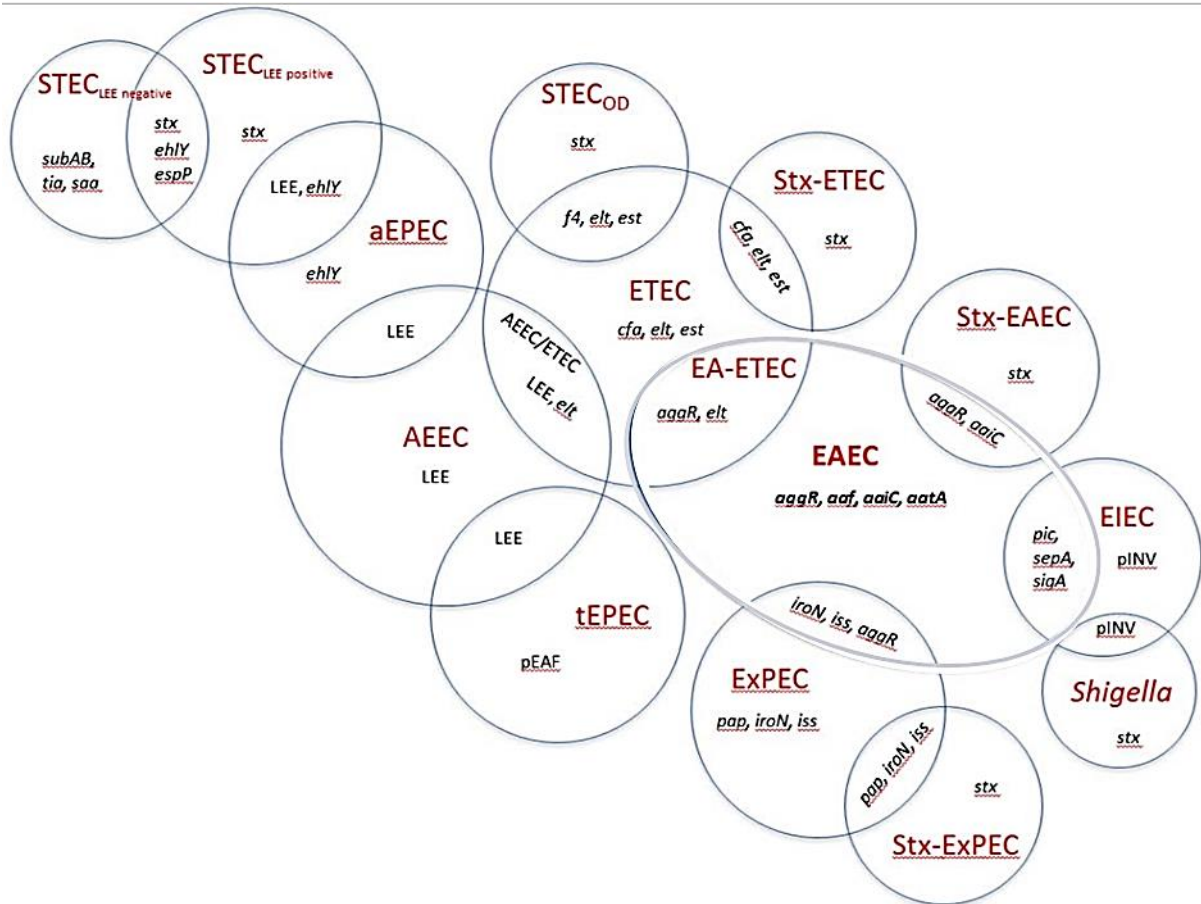


Figure 2 : Bubble chart representing the different cross-pathotypes of *E. coli*. Taken from EFSA, 2015

## 2. *E. coli* typing methods

Due to the vast diversity of strains in the *E. coli* species, typing techniques are particularly important for characterizing strains implicated in food-borne outbreaks (Fratamico *et al.*, 2016b). Phenotypic methods based on their antibiotic resistance, phage susceptibility (bacteriophage typing) or on the carriage of surface antigens (serotyping) have been used to subtype *E. coli* strains (Foxman *et al.*, 2005). Their biochemical properties, especially their capacity to metabolize specific sugars were also targeted (API galleries, (Bouhaddioui *et al.*, 1998)). Advances in molecular techniques and DNA sequencing technologies have facilitated *E. coli* typing. Pulse-field gel electrophoresis (PFGE) uses restriction enzymes to map the strain genome on an agarose gel. Until recently, it was the most used in outbreak investigations of food-borne pathogens and considered the gold standard for *E. coli* typing (Goering, 2010). Although it is still used, alternative PCR-based methods were developed. They usually identify specific regions such as tandem repeats, CRISPR spacers or housekeeping genes (MLST) (Fratamico *et al.*, 2016b). Only serotyping and MLST methods will be developed here since they will be mentioned and they provide epidemiologically important information.

In 1977, Orskov proposed a phenotypic typing method based on the presence and differences of three principal cell surface antigens: O or somatic antigen, H or flagellar antigen and K or capsular antigen (Fig. 3, (Orskov *et al.*, 1977)). Similar to other Gram-negative bacteria, *E. coli* outer membrane is a lipid bilayer into which structures are inserted (Fig. 3A). An important structure anchored is named lipopolysaccharides (LPS) and consists of three parts: lipid A, a core oligosaccharide and a unique polysaccharide named the O-antigen, as presented on Figure 3B (Szalo *et al.*, 2006). The O-antigen is a polysaccharide chain containing repeating units of 2 to 7 sugar residues. Due to the structural diversity found within this O-antigen, it has been a marker targeted for *E. coli* classification (Liu *et al.*, 2020). Not all *E. coli* strains harbor a capsule (K antigen). STEC strains generally do not exhibit any capsule. Thus, methods for serotyping STEC generally solely include O and H antigens. However, the complete antigenic formula O:K:H is used for ExPEC serotyping since they generally harbor a capsule. Serotyping is traditionally performed using a serological method in which antisera from rabbit were raised against the different *E. coli* O-groups and H-types (Ørskov and Ørskov, 1984).

This method is however laborious, expensive, time-consuming and sometimes inaccurate due to cross-reactions (Lacher *et al.*, 2014). Genes encoding the O-antigen are located on the chromosome in a single region named the O-antigen cluster (O-AGC). Biosynthesis of the LPS is usually Wzx and Wzy-pathway dependent, but a few are ABC-transporter dependent (Samuel and Reeves, 2003). An example of O-AGC Wzx and Wzy-pathway dependent is represented on Figure 3C for O-group O26 (Liu *et al.*, 2020). H-typing is based on the flagellar filament structural protein (FliC) encoded by the *fliC* gene (Wang *et al.*, 2003). Molecular serotyping approaches have been developed (including PCR and whole genome sequencing (WGS)) targeting *wzx* and *wzy* (for Wzx/Wzy) or *wzm* and *wzt* (for ABC transporters) and flagellin-associated genes *fliC* (Iguchi *et al.*, 2015; Joensen *et al.*, 2015).

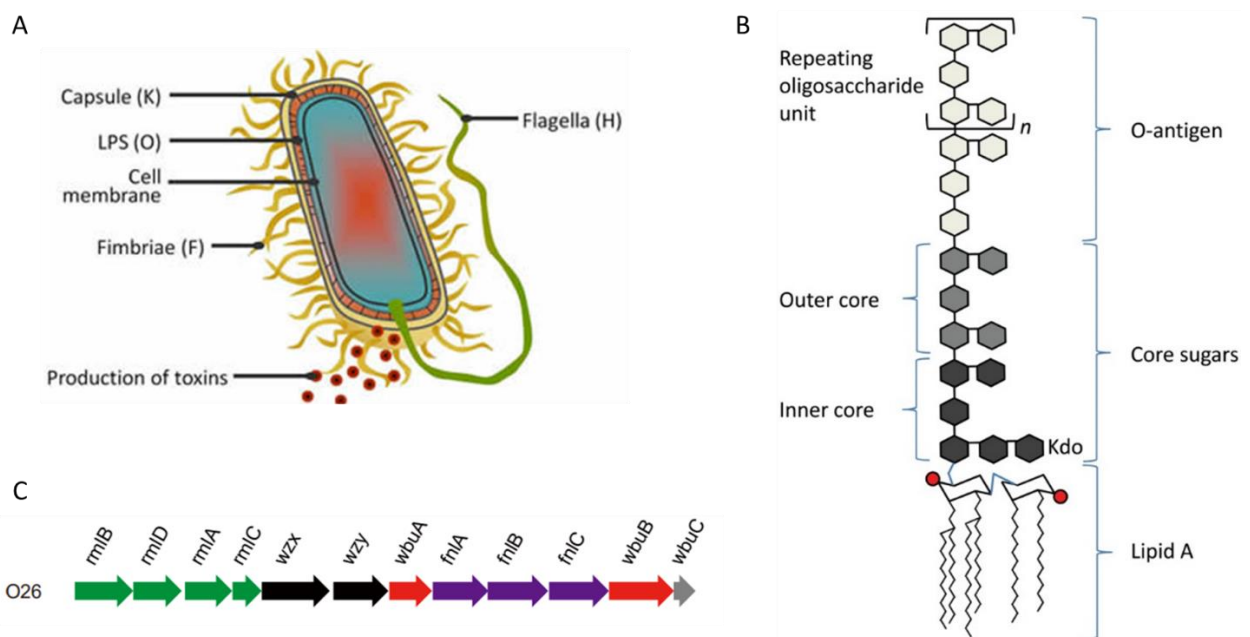


Figure 3 : *Escherichia coli* (*E. coli*) O-antigen structure and the example of O26 O-Antigen cluster (O-AGC).

**A:** The general structure of an *E. coli* membrane. The three surface antigens O, H and K inserted on the cell membrane; taken from [http://www.ecl-lab.ca/contribute\\_images/Ecoli\\_EN.jpg](http://www.ecl-lab.ca/contribute_images/Ecoli_EN.jpg). **B:** General structure of the Lipopolysaccharide (LPS) which includes a lipidic region (lipid A), anchored in the bacterial lipid bilayer membrane, a core polysaccharide and the O-antigen; taken from Maeshima and Fernandez, 2013 **C:** An example of O-AGC Wzx and Wzy-dependent with O-group O26. The targeted genes for O-typing are represented in black (*wzx* and *wzy*); taken from Liu *et al.*, 2020.

A very common molecular approach based on the allelic differences in 5 to 8 housekeeping genes of the bacterial cell was developed and known as multi locus sequence typing (MLST). Two schemes have been described for *E. coli*, the *E. coli* Achtman scheme targeting *adhA*, *fumC*, *gyrB*, *icd*, *mdh*, *purA*, and *recA* genes and the *E. coli* Pasteur scheme analyzing *dinB*, *icdA*, *padB*, *polB*, *putP*, *trpA*, *trpB* and *uidA* genes (Jaureguy *et al.*, 2008; Wirth *et al.*, 2006). Each allelic profile corresponds to a specific sequence type (ST). MLST was originally performed using PCR to amplify the targeted genes and sequenced using Sanger sequencing.

With the emergence of whole genome sequencing, it is possible to fully characterize a strain considering the numerous information available (Franz *et al.*, 2014; Lacher *et al.*, 2014; Larsen Mette V. *et al.*, 2012). This includes (but is not limited to) serotype, MLST, cgMLST and wgMLST, as well as information on the set of virulence genes that the strain can harbor (Parsons *et al.*, 2016).

### 3. STEC characteristic virulence factor: Shiga toxin

The Shiga toxin represents a critical virulence factor for STEC (Nataro and Kaper, 1998). This toxin was named vero cytotoxin (VT) due to its cytotoxic effect on Vero cells (Konowalchuk *et al.*, 1977). Further studies have shown that the VT was similar in structure and action mechanism to the toxin produced by *Shigella dysenteriae* serotype 1: Shiga toxin (O'Brien *et al.*, 1983). Two immunological types of Shiga toxin produced by STEC and various subtypes for both, Stx1 (1a, 1c, 1d and 1e) and Stx2 (2a-2m and 2o) toxins were described so far (Bai *et al.*, 2021, 2018; Gill *et al.*, 2022; Lacher *et al.*, 2016; Meng *et al.*, 2014; Probert *et al.*, 2014; Scheutz *et al.*, 2012). Stx1 and Stx2 possess 99 and 45% of similarity with the *Shigella dysenteriae* Shiga toxin, respectively (Strockbine *et al.*, 1986). STEC can harbor different combination profiles of Shiga toxins. Some harbor only one type of Shiga toxin whereas STEC carrying more than one subtype of one or two types of Shiga toxin have been observed (Shen *et al.*, 2022).

The Shiga toxin consists in two subunits, one A subunit and 5 identical B subunits that are not covalently bound. Once released, the Shiga toxin can cause endothelial cells death. Subunit B attaches to the globotriosyl ceramide 3 (Gb3) specific receptor, enters into the bacterial cell by endocytosis and is transported to the endoplasmic reticulum. There, the A subunit is cleaved, activated and exercise its RNA N-glycosidase activity on 28S rRNA leading to host protein synthesis inhibition and may induce cell apoptosis (Fig. 4; (Kavaliauskiene *et al.*, 2017; Sandvig, 2001)). The expression of Shiga toxin may result in diverse degrees of diarrhea. In severe cases, the released toxin can reach different organs through the bloodstream, for example kidneys, causing HUS or thrombocytopenia and potentially lead to kidney failure (Joseph *et al.*, 2020).

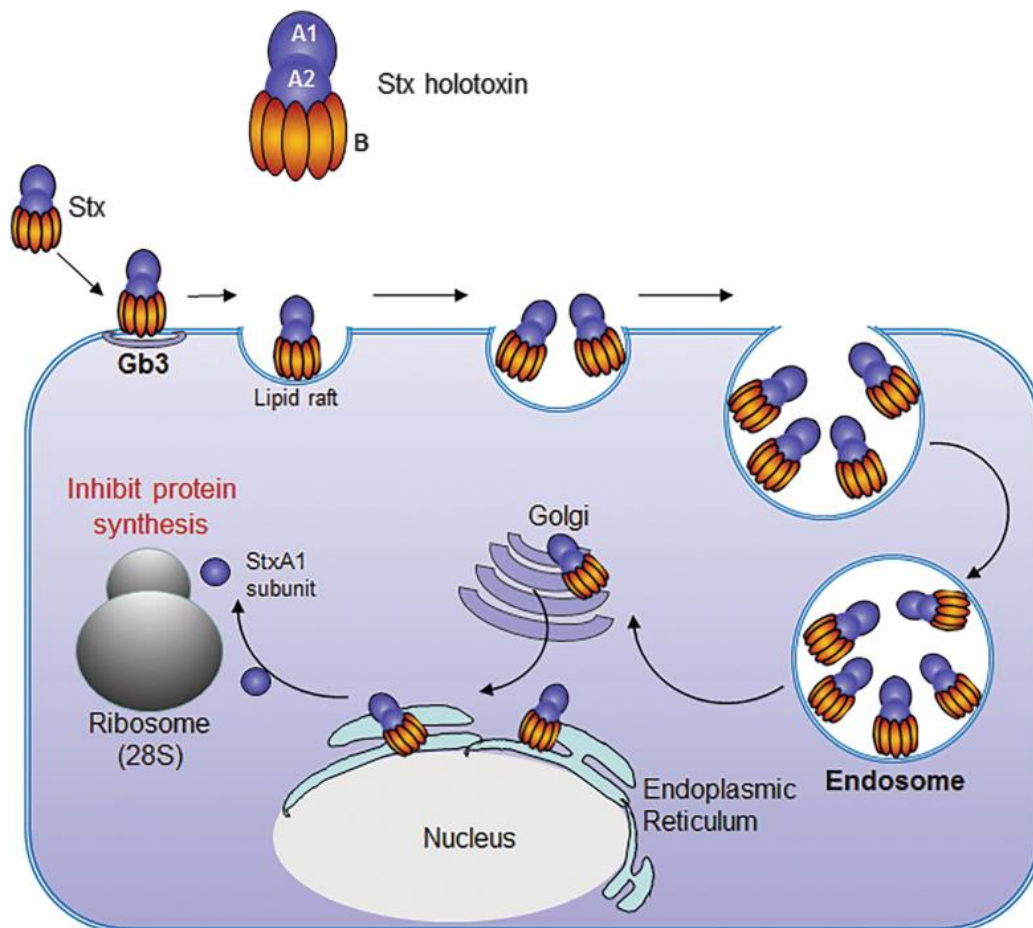


Figure 4 : Mode of action of the Shiga toxin in human enterocytes.

The B subunit of the Shiga toxin binds to its specific Gb3 receptor, the toxin enter the host cell and is transited to the endoplasmic reticulum. There, the A subunit will inhibit protein synthesis. Taken from [https://link.springer.com/chapter/10.1007/978-1-4939-7349-1\\_14#Fig2](https://link.springer.com/chapter/10.1007/978-1-4939-7349-1_14#Fig2) (Bhunia, 2018).

#### 4. STEC capacity to acquire mobile genetic elements (MGEs) increase their pathogenic potential

In addition to the Shiga toxin, other virulence factors may contribute to STEC pathogenicity. Comparative genomics has led to the observation that *E. coli* genome is composed of genes conserved across *E. coli* members (core genome) that are necessary for basic cell function, and a pool of genes that is present in certain individuals (accessory genome) (Rasko *et al.*, 2008). Together, the core and accessory genomes form the pan genome. The accessory genes are usually the result of genetic transfer between bacteria (Kelly *et al.*, 2009). While comparing commensal *E. coli* K12 (MG1655) and LEE-positive STEC EDL933 (*stx*- and *eae*-positive *E. coli*) genomes, 20% of the LEE-positive STEC genome was found to be accessory genes, named mobile genetic elements (MGEs) probably acquired through horizontal gene transfer (HGT) (Reid *et al.*, 2000). HGT is the transfer of genetic material between bacteria and is thought to be the driving force behind *E. coli* ability to acquire or modify its genome, also known as genomic plasticity (Braz *et al.*, 2020). MGEs include phages, plasmids, genomic islands and insertion sequences. There are three main ways for bacteria to transfer DNA: conjugation, transduction and transformation (Burrus and Waldor, 2004; Frost *et al.*, 2005; Hacker and Carniel, 2001). Conjugative transfer is used by some plasmids, transduction is specific for bacteriophage DNA and transformation refers to the stable incorporation of DNA sequences into the bacterial chromosome (Kelly *et al.*, 2009).

##### 4.1. Bacteriophages are highly present in STEC

The main MGE carried by STEC is the lambdoid *stx*-bacteriophage that carries the Shiga toxin-encoding *stxA* and *stxB* genes. Bacteriophages are viruses infecting bacterial cells by injecting their genetic material. Two main types of bacteriophages are lytic and lysogenic phages, as represented on Figure 5. Lytic phages inject their DNA inside the bacteria, use the host cell machinery to replicate their genetic material and finally lyse the bacterial cell to release and spread the produced particles. In contrast, lysogenic phages inject their genetic material that will integrate into the bacterial genome (mostly into the chromosome but plasmid insertions happen) and is named a prophage.

Under certain conditions, lysogenic phages may be induced and enter a lytic cycle (Fig. 5, point 5) (De Paepe *et al.*, 2014). This is the case of the Stx-phage, which is usually integrated into the bacterial chromosome at specific insertion sites (Rodríguez-Rubio *et al.*, 2021; Steyert *et al.*, 2012). When induced, it can be transferred to other *E. coli* strains, but it was also detected in other organisms as observed in *S. dysenteriae* type 1, *Citrobacter freundii*, or *Enterobacter cloacae* (Brabban *et al.*, 2005; Butler, 2012; Khalil *et al.*, 2016; Zhi *et al.*, 2021).

A variety of non-Stx phages are also inserted in STEC genomes. A study showed that STEC can carry up to 20 prophages in their genome (Nakamura *et al.*, 2020). Inserted prophages may carry genes encoding toxins or effector proteins as well as antimicrobial resistance genes (ARG) (De Paepe *et al.*, 2014).

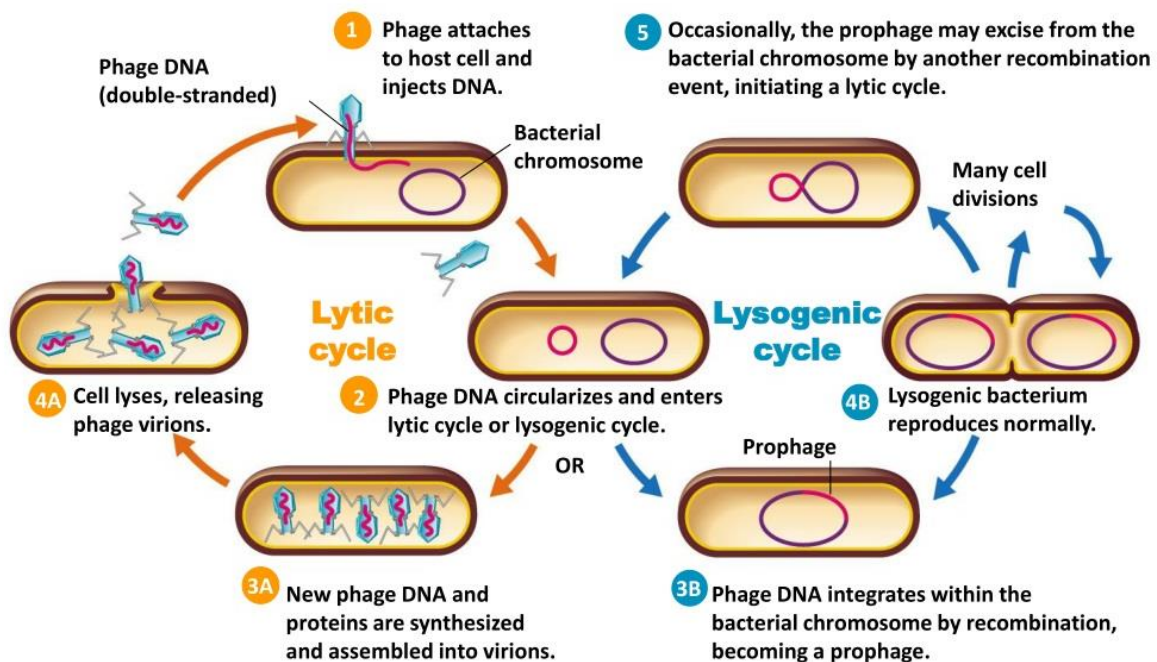


Figure 5 : Representation of the lytic and lysogenic cycle of bacteriophages that can use the *stx*-phage. Taken from <https://image4.slideserve.com/306073/figure-13-12-the-lysogenic-cycle-of-bacteriophage-1.jpg>.



## 4.2. Pathogenicity islands carried by STEC

### 4.2.1. Pathogenicity islands

Genomic islands (GIs) are large regions (10-200 kb) characterized by different GC content compared to the bacterial genome and an alternative codon usage. These two genetic properties suggest their acquisition from a foreign origin. GIs are non-replicative and mostly lack the ability to self-mobilize but may harbor mobility genes that enable them to integrate phage or plasmid sequences. Due to recombination events, GIs may also carry MGEs or parts thereof which may lead to mosaic GI structures (insertion, deletion or genetic rearrangement) (Hacker and Carniel, 2001; Hacker and Kaper, 2000). Genomic islands containing one or more virulence genes are called pathogenicity islands (PAIs) (Blum *et al.*, 1994; Schmidt and Hensel, 2004). PAIs are typical features of pathogenic *E. coli* that played an important role in their evolution.

### 4.2.2. Locus of enterocyte effacement (LEE)

The LEE locus characteristic of EPEC strains (McDaniel *et al.*, 1995) was also identified in STEC strains (Kaper *et al.*, 2004). The LEE PAI, as represented on Figure 6A, consists in 5 poly-cistronic regions (LEE1 to LEE5) of approximately 35 kb that has integrated the *E. coli* genome at specific insertion sites as for example *selC*, *pheU*, *pheV* (Wong *et al.*, 2011). In addition to the intimin protein (Section 1.1, Chapter 1), the LEE encodes all the genes necessary to assemble a functional T3SS, regulators and core effector proteins (T3SS; (Pearson *et al.*, 2016)). The T3SS enables the delivery of effector proteins from the bacterial cell to the cytosol of the targeted cell using a syringe-like structure (Deng *et al.*, 2004, 2001). EspA, a translocator protein forms a long filament, and EspB and EspD form a pore in the host cell membrane, allowing the delivery of effector proteins to the host cell (Garmendia *et al.*, 2005; Gaytán *et al.*, 2016; Ide *et al.*, 2001). Among the transported effector proteins, we find EspF, Tir, EspG, EspH that are LEE-encoded (Serapio-Palacios and Finlay, 2020).

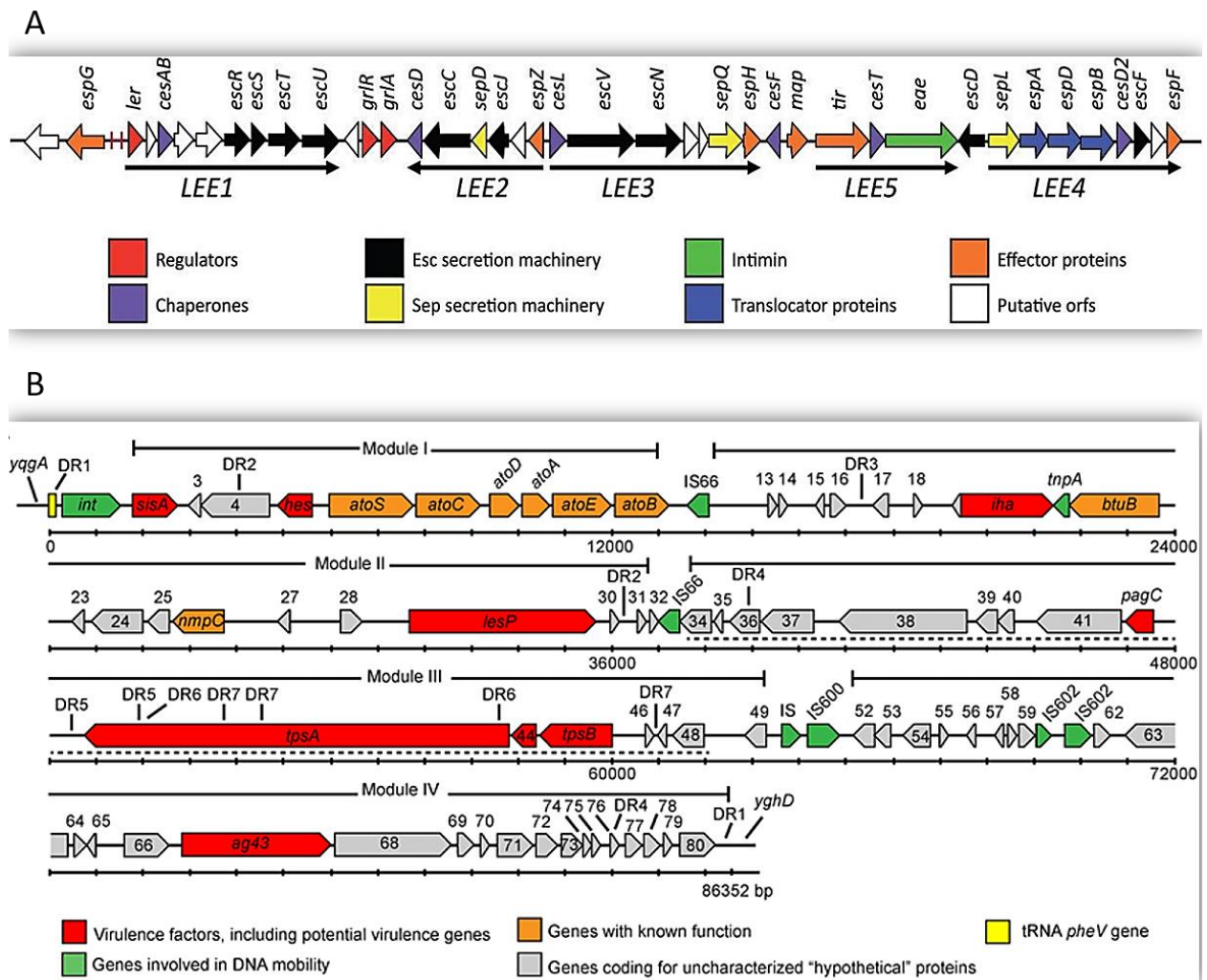


Figure 6 : Important pathogenicity islands carried by STEC.

**A:** Locus of Enterocyte Effacement (LEE) from Tobe *et al.*, 2006; **B:** Locus of Autoaggregation and adhesion (LAA) from (Montero *et al.*, 2017).

#### 4.2.3. O-islands

Phage sequences are present in STEC genomes. First characterized as genomic islands (GIs) in EDL933 (or Sakai), they were named O-islands (OIs or Sp for Sakai strain) (Hayashi *et al.*, 2001; Perna *et al.*, 1998). Several O-islands have been described in EDL933 strain. Although the function of many genes carried by O-islands is still unknown, some of them encode putative virulence factors, others confer advantages for bacterial strain survival (Jiang *et al.*, 2021). Non-LEE encoded effectors are found on cryptic or inducible prophages (NleA/EspI, NleB, NleC, NleD, NleE, EspJ, NleH, EspG, EspM, cif, EspK, EspV) corresponding to OI-57/Sp9, OI-71/Sp9, OI-122/SpLE3, OI-44/Sp4, and OI-50/Sp6 in EDL933/Sakai (Delannoy *et al.*, 2013a; Garmendia *et al.*, 2005; Konczyk *et al.*, 2008; Ogura *et al.*, 2009; Tobe *et al.*, 2006; Wong *et al.*, 2011). The injected virulence factors lead to cytoskeletal rearrangements that induce cell death, resulting in diarrhea.

In addition to the LEE, other pathogenicity islands have been identified in LEE-positive STEC strains that contribute to the A/E lesion (typical EHEC). An example is OI-122/SpLE3, a genomic island of 23 kb that harbors adhesion factors including *efa1-lifA* complex, frequently detected in LEE-positive STEC associated with severe human diseases (Karmali *et al.*, 2003; Wickham *et al.*, 2006). It was shown to be involved in the A/E lesion and inhibits host lymphocyte activation (Klapproth *et al.*, 2000). Two OIs harbor genetic markers Z2098 (OI-57/Sp9) and *ureD* (OI-43/OI-48, SpLE1) that were proposed as molecular markers for typical EHEC diagnostics in combination with *espK* (OI-50/Sp6) and *espV* (OI-44/Sp4) (Delannoy *et al.*, 2016, 2013a; Hayashi *et al.*, 2001).

#### 4.2.4. Locus of autoaggregation and adhesion

The T3SS is not the only adhesion / translocation system used by pathogenic *E. coli*. LEE-negative STEC strain of serotype O91:H14, O91:H21, O113:H21, O104:H4 or even O165:H25 have been implicated in HUS cases (Bielaszewska *et al.*, 2009; Feng *et al.*, 2014; Mellmann *et al.*, 2009; Nakamura *et al.*, 2023). Studies on LEE-negative STEC causing severe human symptoms tried to identify the adhesion mechanism they use (Krause *et al.*, 2018).

In 2017, Montero and colleagues analyzed LEE-negative STEC and described a pathogenicity island of approximately 86 kb composed of 80 genes dispatched on 4 modules, presented on Figure 6B. This PAI was named Locus of Autoaggregation and Adhesion (LAA) since it carries factors responsible for adhesion and autoaggregation. Among them, we find *hes* on module I, *iha* and *lesP* on module II, *pagC*, *tpsA* and *tpsB* on module III and *agn43* located on module IV (Montero *et al.*, 2017). This locus or parts thereof have been found to be carried by emergent LEE-negative STEC of serogroups O113, O91, O128, O146 and O174 implicated in clinical cases (Colello *et al.*, 2018; Montero *et al.*, 2019, 2017; Vélez *et al.*, 2020).

#### 4.3. Plasmids

In addition to bacteriophages and PAIs, plasmids also encode putative virulence factors that contribute to STEC pathogenicity. Plasmids are defined as extrachromosomal double stranded DNA molecules stably inherited in bacterial cells. They are able of autonomous replication. Although they do not carry genes that are essential for the host cell, they may carry putative virulence factors (bacteriocins, siderophores, cytotoxins, adhesion factors) or antimicrobial resistance genes (Johnson and Nolan, 2009; Thomas and Nielsen, 2005). Plasmids carrying adhesion or colonization factors as well as toxins were identified in HUS-causing STEC. A large plasmid of 92 kb named pO157, encodes putative virulence factors *ehxA*, *espP*, *etpD*, *katP* and *toxB* and is frequently present in LEE-positive STEC (Losada *et al.*, 2016; Ogura *et al.*, 2009). An enterohemolysin toxin encoded by *ehxA* may be responsible for enterocyte lysis (Beutin *et al.*, 1990). Another example is the pO113 plasmid, that carries genes encoding a cytotoxin SubA (subtilase cytotoxin, (Paton *et al.*, 2004)) an autotransporter EpeA (Leyton *et al.*, 2003), and adhesins Saa (autoagglutinating, (Paton *et al.*, 2001)), Sab and Iha (Herold *et al.*, 2009; Newton *et al.*, 2009). The pO113 plasmid was detected in LEE-negative STEC strains of serogroups O91, O128, O113 or even O174 (Fig. 7) (Newton *et al.*, 2009). Figure 7 shows an example of LEE-positive STEC carrying the pO157 plasmid and a LEE-negative STEC positive for the LAA and the pO113 plasmid.

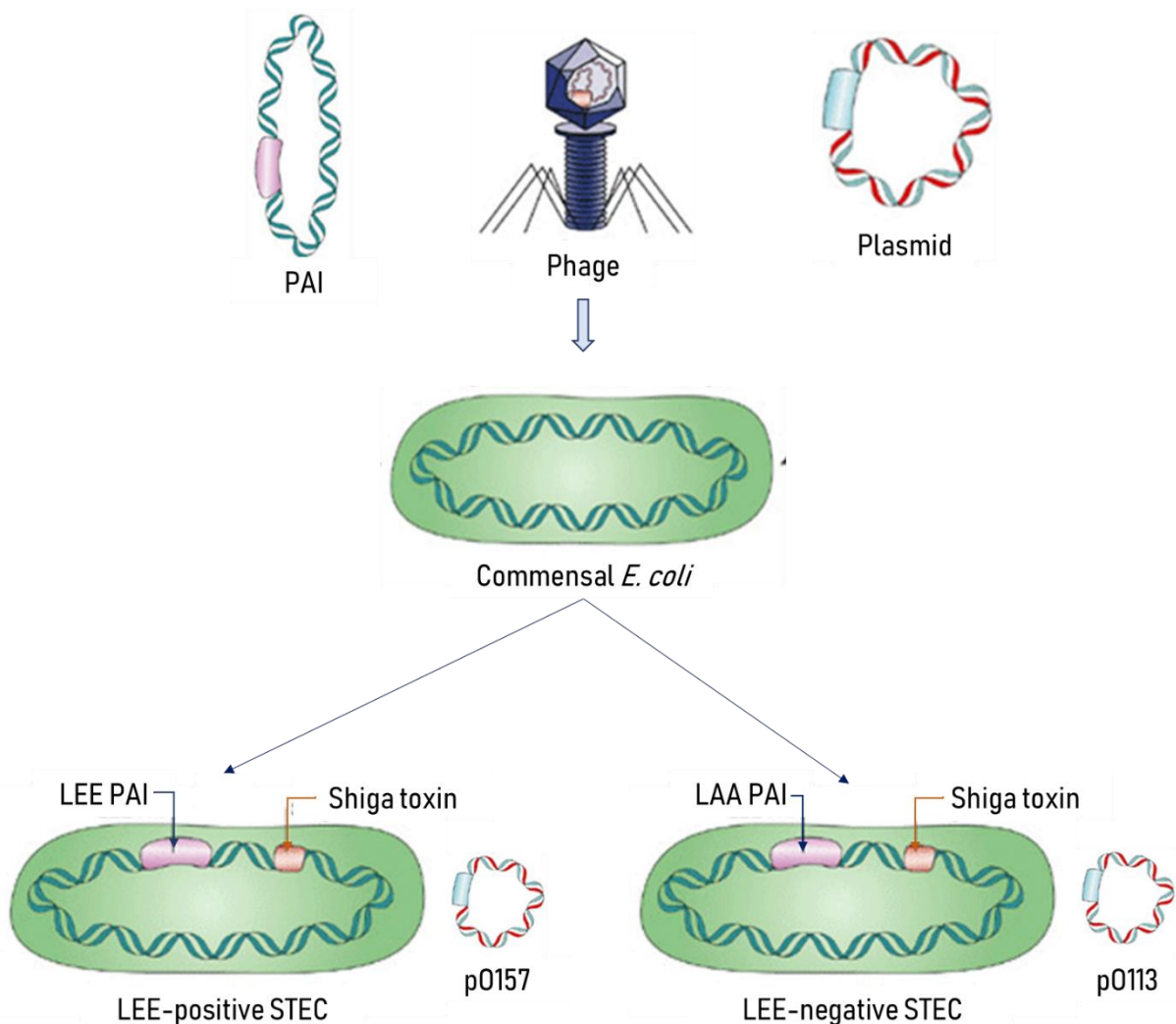


Figure 7 : Shiga toxin-producing *Escherichia coli* (STEC) causing hemolytic uremic syndrome (HUS) have acquired virulence factors carried by mobile genetic elements (MGEs).

Two examples are represented here with strains that carries a pathogenicity island (PAI) either the Locus of Enterocyte Effacement (LEE) or the Locus of Autoaggregation and Adhesion (LAA). The first example represents LEE-positive STEC carrying parts or the entire p0157 plasmid frequently implicated in severe cases. The second example are emerging LEE-negative STEC carrying another pathogenicity island named LAA and may carry the p0113 plasmid (e.g strains O113:H21); adapted from <https://www.nature.com/articles/nrmicro818> (Kaper *et al.*, 2004).

## 5. Cross-pathotype/Hybrid *E. coli*

Despite belonging to the same species, there is a great diversity of *E. coli* strains. Most *E. coli* strains are commensal, sometimes even beneficial to human health, but their potential to acquire MGEs carrying virulence factors has led to the emergence of pathogenic *E. coli* ((Croxen *et al.*, 2013); example on Fig. 7). Subtyping methods as presented earlier (Section 2, Chapter 1) are not able to differentiate commensal from pathogenic *E. coli*. The use of molecular-based approaches has helped in identifying and differentiating the set of virulence markers characteristics of pathogenic *E. coli* from commensal *E. coli*. PCR-based assays are used to characterize the pathogenic *E. coli* involved by analyzing specific genes.

To cause severe symptoms, STEC have to colonize the intestinal tract. The most frequent adhesion factor identified in STEC causing outbreaks is the intimin (*eae* gene) encoded by the LEE. Yet, the increasing number of severe cases caused by LEE-negative STEC have highlighted the evolution of cross-pathotype *E. coli*. STEC harboring virulence genes from another *E. coli* pathotypes are referred to as hybrid or cross-pathotype STEC. Although it is the hybrid STEC/EAEC O104:H4 strain that caused a large epidemic in 2011 in both France and Germany, (Bielaszewska *et al.*, 2011; Muniesa *et al.*, 2012) other STEC/EAEC hybrids have been implicated in human symptoms (Dallman *et al.*, 2012; Morabito *et al.*, 1998; Santos *et al.*, 2020). With the emergence of WGS, the whole database of *E. coli* virulence genes from all pathotypes can be screened. It has led to the description of different STEC cross-pathotypes isolated from HUS- or diarrhea-suffering patients such as ETEC/STEC (Nyholm *et al.*, 2015). More recently, different clones of ExPEC/STEC hybrid of serotype O80:H2 caused HUS cases in Europe. Such hybrid strains were LEE- and *stx2*-positive (though a particular type of intimin (xi) and different combinations of *stx2a*, *stx2c* and *stx2d* genes). Additionally, they harbored factors enabling the colonization of extra-intestinal tissues (*iss*, *iroN*, *cvaV* genes) and an antibiotic resistance gene cassette present on a mosaic plasmid, parts of which are characteristic of the ExPEC pS88 plasmid (Braz *et al.*, 2020; Cointe *et al.*, 2020, 2018)

---

## Chapter 2: STEC, A food safety perspective

---

### 6. STEC implication in hemolytic uremic syndrome in France and Germany

STEC infections caused 6 084 illnesses in humans worldwide including 4 355 cases in the EU from 2017 to 2021 (EFSA and ECDC, 2022). From 2012 to 2017, STEC were implicated in 5 931 cases of bloody diarrhea and 1 653 HUS in Europe (EFSA, 2020). The Top 5 serogroups of STEC implicated in HUS were O157 (38%), O26 (24%), O111 (5%), O80 (4%) and O145 (4%) followed by O55, O121, O103, O91 and O104 (<3%) (EFSA, 2020).

#### 6.1. Virulence profile of strains causing HUS in France

In France, since 1996, only pediatric HUS (in children below 15 years old) are reported by voluntary pediatric nephric hospitals to the National Reference Center (NRC) for *E. coli* (Pasteur Institute). France reported 2 959 cases of pediatric HUS caused by STEC between 1996 and 2021. On all reported HUS cases, 68% were from children below 3 years old (Santé Publique France, 2021). Prior to 2016, the most frequent STEC serogroups were O157 (23%), O26 (11%) and O80 (8%) followed by O111, O145 and O55. Since 2016, an important shift occurred and O26 now represents the major serogroup associated with HUS (35%), followed by O80 (18%), and O157 (10%). In 93% of HUS cases, the isolated STEC strain was positive for *eae*. It has been reported that the most prevalent *stx* subtypes were *stx2a*, *stx2d* and a combination of *stx1a* and *stx2a* (Santé Publique France, 2021).

#### 6.2. Virulence profile of strains causing HUS in Germany

In Germany, all diagnosed clinical cases of STEC infection -and not only HUS- are reported to the Robert Koch Institute (RKI) by public health institutions. The active surveillance regarding pediatric HUS have been initiated in 2008 by the society for pediatric nephrology and the RKI (RKI, 2021). In 2020, 60 HUS cases occurred and were caused by O157 (16.67%) followed by O26 (5%), O111 (5%) and O145 (5%), and in a minor case (1.67%): O8, O80, O114, O172 and Ont (RKI, 2020).

Like in France, most of HUS cases were caused by STEC carrying *stx2* gene subtypes 2a, 2c or 2d and a few cases caused by *stx1*-positive *E. coli* (here 5.8%) (Pörtner *et al.*, 2019). Indeed, the National Reference Center for *Salmonella* and other bacterial enteritis pathogens from RKI and the National Consulting Laboratory for HUS at University hospital of Münster have analyzed the virulence profile of strains isolated from patients with HUS symptoms. They investigated 172 strains causing HUS from 1998 to 2018 excluding cases caused by the O104:H4 outbreak-causing strain. On all isolates, 94% were *stx2*-positive (with 142/172 strains *stx2*-positive only and 19 strains positive for both *stx1* and *stx2* genes) and 6% *stx1*-positive. Similar results were observed on the more recent years within this dataset, from 2015 to 2018, with 95% of strains *stx2*-positive (73/87 only positive for *stx2* and 10 *stx1*- and *stx2*-positive) and 5% exclusively *stx1*-positive. The presence of the *eae* gene was detected in 87% of strains from 1998-2018 (149/172) and 94% from 2015-2018 (82/87) (Pörtner *et al.*, 2019).

## 7. STEC contamination sources

The intestinal tract of warm-blooded animals, and especially from ruminants, is a natural reservoir for STEC. Their primary reservoir is cattle, which are asymptomatic carriers of STEC. Nevertheless, STEC are also found in other ruminants such as sheep, goat, deer and buffaloes (Kim *et al.*, 2020). Cattle can contaminate the environment, their hides and udders through their feces. In the soil or grass, some STEC are able to survive and reach the ground waters used to irrigate cultures (Ferens and Hovde, 2011). Similarly, food products can be contaminated after being in contact with feces material of contaminated animals during processing, storage or distribution. Consequently, water and food ingested by humans can be contaminated, especially ground meat, dairy products including raw milk or vegetables (Gyles, 2007). Although food-borne contaminations constitute the main source of transmission to humans, direct contact with animals or the environment is possible (Fig. 8). A study conducted in the US showed that, although direct and indirect transmission can arise, contaminated food is the major contamination route, representing 66% of cases, compared to person-to-person transmission (20%), water-borne contamination (12%) and animal contact (2%) (Rangel *et al.*, 2005). However, the situation is not a generality and varies from country to country depending on production and consumption habits (Karmali, 2018, 2017).



## 8. HUS-causing STEC outbreaks in France and Germany

Contaminated food products is the primary source of STEC infections in Europe. From 2017 to 2021, 31 STEC outbreaks occurred in Europe with strong evidence of contamination source for five of them. Two were due to contaminated bovine meat products and one from milk. At the European level, ‘meat and meat products’ remains the first category of food responsible for STEC infections followed by ‘milk and milk products’ and ‘fruits and vegetables’ (EFSA and ECDC, 2022).

Prior to 2016, in France HUS outbreaks caused by STEC were due to contaminated ground beef (54%), raw milk cheese (22%), raw milk (5%), water (surface or bathing 19%) and contact with animal (20%) (Bruyand *et al.*, 2019). From 2017, small HUS-STECS outbreaks occurred in France due to contaminated raw milk cheese (2018, 2019, 2020), raw cucumber (2021), and lastly pizza (2022) (Santé Publique France, 2022, 2021, 2020, 2019, 2018). Other sporadic cases occurred but no contamination source was reported (2017-2021). It is important to note that in France, the contamination source is investigated and reported for grouped cases (time and space cases) of HUS and is not always successful (Santé Publique France, 2009).

In Germany, since the big outbreak caused by the O104:H4 EAEC/STEC hybrid, only one national outbreak was reported which was probably associated with contaminated food product consumption. Indeed, in 2017, 14 confirmed HUS cases including one person that died as a result of infection were caused by an *eae*-positive STEC O157, probably associated with mixed minced beef consumption (pork and beef meat) (Vygen-Bonnet *et al.*, 2017). As described for France, other HUS cases were reported but a food origin could not be traced (RKI, 2021, 2019). Although few HUS cases are suspected to be caused by ground beef and dairy product consumption (RKI, 2021, 2018, 2017), direct contact with animals or inter-human transmission are also presumed (RKI, 2019, 2018, 2017). One of 3 HUS outbreaks reported in 2020 was suspected to be caused by raw donkey milk consumed during holidays in France (RKI, 2021).

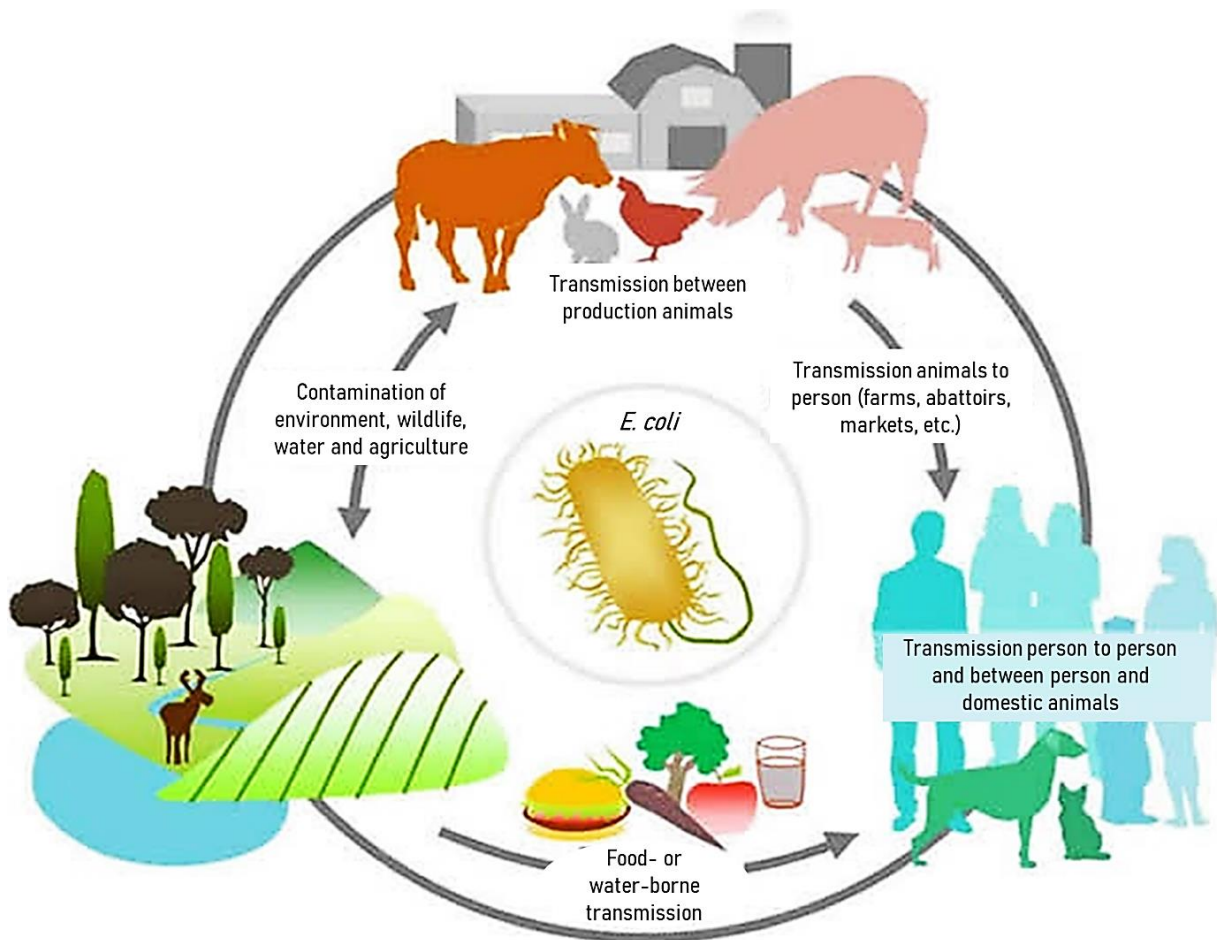


Figure 8 : Shiga toxin-producing *Escherichia coli* (STEC) main contamination sources.

Taken from <https://marlerclark.com/foodborne-illnesses/e-coli/transmission-of-and-infection-with-e-coli>.

Because of its high content in nutrients as fat, lipids, carbohydrates, proteins and vitamins, raw milk is susceptible to microbial contamination. Raw milk contamination by STEC is mainly caused by the contamination of dairy cow teats by feces and is more frequent from end of spring to end of summer. To avoid food-borne contamination, the milk is usually pasteurized or sterilized to destroy the microbial flora. In France, many dairy products are nonetheless made from raw milk, which do not undergo heating processes (Bagel and Sergentet, 2022).

## 9. The challenge of STEC pathogenicity assessment and their regulation in food

### 9.1. Development of classification approaches

The first time *stx*- and *eae*-positive *E. coli* have been associated with pathogenicity was during an outbreak in the USA in 1992. Undercooked ground beef contaminated with LEE-positive STEC of serotype O157:H7 at the “Jack in the box” rapid food chain affected approximately 732 persons. Four affected patients died while 178 had persistent sequels (Bell *et al.*, 1994; Rangel *et al.*, 2005). Rapidly, the Food Safety and Inspection Service (FSIS) implemented the testing for *E. coli* O157:H7 in ground beef, which was classified as an adulterant in this food product (FSIS, 1994). The biggest outbreak reported worldwide occurred two years later in Japan, where a LEE-positive STEC O157:H7 strain (Sakai) infected 12 680 persons, 121 with HUS and three patients died after eating contaminated radish sprouts (Fukushima *et al.*, 1999).

Although additional LEE-positive STEC O157:H7 caused outbreaks, non-O157 strains were implicated in HUS outbreaks in the USA and other parts of the world. In the USA non-O157 STEC implicated in severe symptoms are known as Top-7 and include serogroups O26, O111, O103, O121, O45 and O145, and were also classified as adulterants in raw beef products in 2011 (FSIS and USDA, 2012). Similarly, an increasing number of HUS cases caused by non-O157 STEC were described in Europe known as Top-5 (O26, O103, O111, O145 in addition to O157). Strains from the Top-7 or Top-5 are all *eae* positive. Due to their genomic plasticity and the different STEC strains that caused HUS symptoms, it is difficult to strictly define pathogenic STEC.

In 2002, the European council (CE no 178/2002) introduced the ‘Hygiene package’ to regulate food hygiene’s procedure ‘from farm to fork’. It concerns a variety of food-borne pathogens and aims at preventing food-borne contaminations. Although microbiological criteria have been defined for many food-borne pathogens, it is not the case for STEC (except in sprouts). STEC contamination has to be prevented by each company/entity selling food, considered responsible for the hygienic quality of the product that they propose to consumers (European Parliament and Council of the European Union, 2002).

## 9.2. Classification from serotype to virulence gene profile

In 2003, Mohamed Karmali proposed to assess STEC pathogenicity using a seropathotype classification. As presented on Table 1, this classification was based on serotypes and their incidence and frequency of causing epidemics as well as their association with severe symptoms such as HC or HUS (Karmali *et al.*, 2003). Although the serogroup is epidemiologically important for tracking (incidence, outbreak detection, etc.), we cannot only rely on the serotype to predict pathogenicity. Nowadays, an increasing number of outbreaks or sporadic cases caused by non-Top5 or even hybrid STEC are described. In 2011, the most severe non-O157 STEC outbreak was caused by a hybrid STEC/EAEC O104:H4 strain isolated from sprouts and occurred in Germany (Bielaszewska *et al.*, 2011; Muniesa *et al.*, 2012; Rasko *et al.*, 2011). This strain carried both a Shiga toxin-encoding gene (*stx2a*) and virulence traits of EAEC (*aggR* and *aaiC*, Section 1.7, Chapter 1). This outbreak demonstrated that the classification method proposed by Karmali and colleagues was inadequate.

Table 1: Seropathotype classification of pathogenic STEC proposed by Karmali and colleagues in 2003. Based on Scientific Opinion on VTEC-seropathotype and scientific criteria regarding pathogenicity assessment - - 2013 - EFSA Journal - Wiley Online Library

Seropathotype	Incidence (frequency in human disease)	Frequency of involvement in outbreaks	Association with HC and HUS	Serotypes
A	High	Common	Yes	O157:H7 O157:NM
B	Moderate	Uncommon	Yes	O26:H11, O103:H2, O111:NM, O121:H19 O145:NM
C	Low	Rare	Yes	O91:H21, O104:H21, O113:H21, O5:NM, O121:NM O165:H25
D	Low	Rare	No	Multiple
E	Non-human only	NA	No	Multiple

Consequently, STEC pathogenicity was proposed to be assessed by molecular approaches based on the genetic markers present in the strain (EFSA, 2020). EFSA proposed in 2013 to classify STEC at high risk of causing HUS as *eae*- or *aaiC* and *aggR*- positive STEC of one of the above-mentioned serotypes and O104 (EFSA, 2013). However, no EAEC/STEC hybrid caused important cases since. The current EFSA classification of STEC at high risk of causing HUS is presented on Table 2 (EFSA, 2020). Anses, the French agency for food safety also recommended the surveillance of the emerging O80:H2 STEC/ExPEC (Anses, 2019).

Table 2: EFSA classification of highly virulent STEC based on Tor1 assessment of 2020. Pathogenicity assessment of Shiga toxin-producing *Escherichia coli* (STEC) and the public health risk posed by contamination of food with STEC - - 2020 - EFSA Journal - Wiley Online Library

EFSA Classification	Virulence gene <i>eae</i>	Serogroup	HC	HUS
Group I	<i>eae</i> -positive	Top 5	High risk	High risk
Group II	<i>eae</i> -positive	na	High risk	Unknown
Group III	<i>eae</i> -negative	na	Unknown	Unknown

### 9.3. Recent classifications based on *stx* subtype and additional adhesion genes

Different approaches for assessing STEC pathogenicity are based on their virulence profile rather than O-group. It is recognized worldwide that STEC positive for *stx2a* alone or in association with *stx1*, and carrying the *eae* gene represent a higher risk for humans. As shown in Table 3, STEC at higher risk of causing HUS according to Food and Agriculture Organization of the United Nations (FAO) and World Health Organization (WHO), described in 2018, are *stx2a* and *eae*- or *aggR*-positive. In addition, the presence of *stx2d* in combination with *eae* or *aggR* or alone are also considered at higher risk for causing HUS (Table 3). Similarly, the National Advisory Committee on Microbiological Criteria for Foods (NACMCF) and the United States Department of Agriculture (USDA) classify STEC carrying *stx2a* in combination with *aggR* or *eae* adhesion-conferring genes at high risk of causing severe forms, as represented Table 4 (NACMCF, 2019). Table 5 presents the Anses latest opinion, which categorized STEC at high risk of causing HUS depending on the presence of *eae* or *aggR* genes and specific *stx* sub-types. They also classified *stx2a* and *aggR*- or *eae*-positive profile at higher risk of causing HUS but also included *stx2d* and *aggR* or *eae*-positive virulence profile (Table 5, Anses, 2023).

Table 3: to Food and Agriculture Organization of the United Nations (FAO) and World health Organization (WHO) approach for classifying STEC pathogenic potential 2018. FAO/WHO. 2018. Shiga toxin-producing *Escherichia coli* (STEC) and food: attribution, characterization, and monitoring. (Rome, Italy: FAO), 175 p.

Rank	Virulence genes	Risk of
1	<i>stx2a + eae</i> ou <i>aggR</i>	D/BD/HUS
2	<i>stx2d</i>	D/BD/HUS <sup>2</sup>
3	<i>stx2c + eae</i>	D/BD <sup>3</sup>
4	<i>stx1a + eae</i>	D/BD <sup>3</sup>
5	Other <i>stx</i> sub-types	D

<sup>1</sup>Depending on host susceptibility and other factors e.g. antibiotic treatment

<sup>2</sup>HUS association relies on *stx2d* variant and genetic background of the strain

<sup>3</sup>Certain *stx*-subtypes are associated with BD and occasionally to HUS

Table 4: Molecular risk assessment of STEC according to the National Advisory Committee on Microbiological Criteria for Foods (NACMCF) and the United States Department of Agriculture (USDA) NACMCF / USDA, 2019.

[https://www.fsis.usda.gov/sites/default/files/media\\_file/2020-07/nacmcf-stec-2019.pdf](https://www.fsis.usda.gov/sites/default/files/media_file/2020-07/nacmcf-stec-2019.pdf), p13

Risk (in increasing order)	Virulence genes and O-group
1	<i>stx2a</i> + EAEC
2	<i>stx2 + eae</i> + O157 <i>stx2a</i> > <i>stx2c</i> > <i>stx2a+stx1a</i> > <i>stx1a</i>
3	<i>stx2 + eae</i> + Top-6 <i>stx2a</i> > <i>stx2d</i> > <i>stx2c</i> > <i>stx1a</i>
4	<i>stx + eae</i> + other O-group Toxin order as above
5	<i>stx</i> Toxin order as above

Table 5 : Anses new classification of STEC pathogenic potential. Anses (2023).  
Avis relatif à la définition des souches pathogènes d'*Escherichia coli* productrices de shigatoxines (saisine n°2020-SA-0095). Maisons-Alfort: Anses, 63 p.

Group	Virulence genes	PPV <sup>1</sup> HUS	PPV <sup>1</sup> HC
I <sup>2</sup>	<i>stx2a</i> and/or <i>stx2d</i> <sup>3</sup> - positive <i>eae</i> or <i>aaiC</i> / <i>aggR</i> - positive	69%	13%
II <sup>4</sup>	<i>stx2a</i> and/or <i>stx2d</i> <sup>3</sup> - positive <i>eae</i> or <i>aaiC</i> / <i>aggR</i> - negative	48%	25%
III <sup>2</sup>	Other <i>stx</i> -positive <i>eae</i> ou <i>aaiC/aggR</i> - positive	13%	38%
IV	Other <i>stx</i> -positive <i>eae</i> ou <i>aaiC/aggR</i> - negative	15%	31%

<sup>1</sup>The positive predictive value (PPV) reflects the probability of the strains to induce clinical symptoms.

<sup>2</sup>Top-5 O-group plus O80 represents 80% of HUS cases, 67% of HC and 61% of acute diarrhea

<sup>3</sup>*stx*-subtype alone or in combination

<sup>4</sup>These strains are mainly implicated in adult HUS cases (majorily O-groups O91, O171, O174 and O148)

## 10. ISO/TS-13136:2012 STEC detection method in food products and its drawback

The distinction of STEC strains from other *E. coli* is impossible at the phenotypic level. STEC are nothing more than *E. coli* carrying the Stx-phage and this characteristic may not be used as a distinct factor of STEC or non-STEC *E. coli* in a mixture (Fratamico *et al.*, 2016a). So far, the most reliable method is to use molecular-based detection approaches targeting *stx*-coding genes. Two methods were proposed to harmonize STEC detection in food and animal feed: ISO 16654 (ISO, 2001) for STEC O157:H7 and ISO/TS-13136:2012 (ISO, 2012) for non-O157 STEC. Figure 9 shows the steps of the ISO/TS 13136:2012 detection method. An enrichment step to get bacteria growing to a detectable level is performed and the use of acriflavine -an antibiotic targeting Gram-positive bacteria- was recommended in dairy products. Since there is a diversity of strains, a single enrichment procedure favoring STEC regardless of the other *E. coli* strains is complicated.

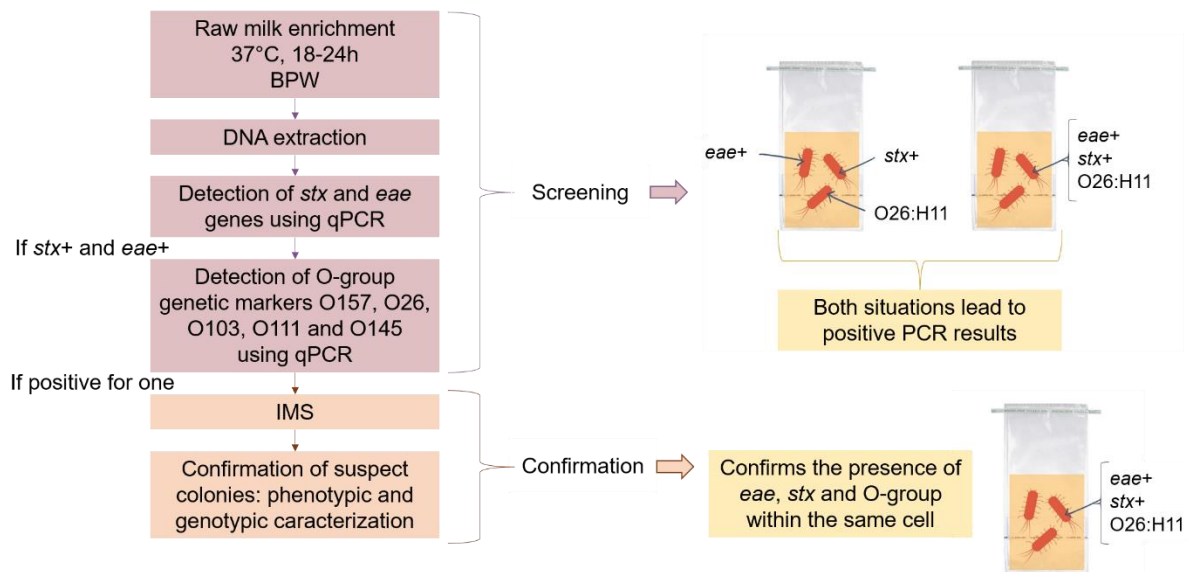


Figure 9: Schematic representation of the ISO/TS-13136:2012 procedure for STEC detection in food products.

Adapted from <https://supermicrobiologistes.fr/detection-stec-dans-aliments/#reglementation>.

The next step consists in a qPCR to detect the *stx* and *eae* genes in the enrichment broth. The abundant *E. coli* flora may disturb STEC detection. A study showed that as much as 50 strains of *E. coli* were identified in a single sample in cattle microbiota (Cookson *et al.*, 2017). Multiple strains (each carrying one of the targeted genes) can be present simultaneously in the broth, resulting in presumptive-positive results (Fig. 9). A few *stx*-carrying species such as *Citrobacter*, *Shigella* might lead to similar results (Brabban *et al.*, 2005; Butler, 2012; Tajeddin *et al.*, 2020; Zhi *et al.*, 2021). Lastly, the presence of free *stx*-phages have been described which could be amplified by PCR and impede STEC isolation attempts (Martínez-Castillo and Muniesa, 2014). Attempted isolation of the strain is necessary to ensure that those genetic markers (*stx*, *eae* and O-group) belong to the same strain. Due to the lack of STEC specific isolation medium and the significant background flora in milk (mainly consisting of Gram-positive bacteria), it is difficult to isolate STEC strains. The problem of unconfirmed presumptive positive results is the uncertainty regarding the decision to commercialize the product. On the one hand, the destruction of food products with unknown presence of HUS-causing STEC has an important economic impact. On the other hand, contaminated food products with a real risk of causing HUS may be consumed.



---

## Chapter 3- Long-read metagenomics as a new approach for STEC characterization

---

### 11. DNA sequencing

Bacteria are classified into families based on their genetic material, which is supported by DNA molecules. DNA sequencing is a method that consists in analyzing the nucleotide (adenine A, guanine G, cytosine C and thymine T) order from a specimen that constitutes its genetic material. First generation sequencing (FGS) methods have been introduced in 1975 by Sanger in parallel to Maxam and Gilbert in 1977 (Maxam and Gilbert, 1977; Sanger, 1975). FGS enabled the decryption of DNA fragments up to 1000 bp (Schadt *et al.*, 2010). The first sequencing of the human genome was performed using Sanger sequencing; it took 13 years for twenty scientific teams around the world (the International Human Genome Sequencing Consortium) and cost several billion US dollars (Lander *et al.*, 2001). DNA sequencing evolved in 2004 with the commercialization of second-generation sequencing (SGS, also called next generation sequencing (NGS)) which reduced the cost of FGS and enabled higher throughput. NGS methods have the advantages of generating highly accurate data of many samples in parallel (multiplexing) with homogenous coverage from low DNA input (1 ng). Different technologies were commercialized but the most widely used method was developed by Illumina. Illumina sequencing has a limited read length (50-300bp) to avoid sequencing errors caused by background signal (Schadt *et al.*, 2010). Third generation sequencing (TGS) determines the sequence of a single DNA molecule in real-time with the objective of increasing sequencing throughput and read length (Schadt *et al.*, 2010). Pacific Biosciences (PacBio) were the first to commercialize the single-molecule real-time sequencing technology, in 2011. Oxford Nanopore Technology (ONT) later released the MinION sequencer, which was launched in 2014. ONT sequencing has the advantage of generating longer reads than PacBio technology with reads up to 4 Mb (Lu *et al.*, 2016).

## 12. MinION sequencing: from principle to data analysis

### 12.1. Principle of nanopore sequencing

The principle of nanopore sequencing, represented on Figure 10, relies on the different structure of the four nucleotides A, T, C, and G (Deamer *et al.*, 2016). The flow cells, on which nanopore sequencing is performed, consist of a synthetic lipid bilayer membrane into which synthetic nanopores are inserted. An ionic current is applied to this membrane. During the library preparation step, adapters are ligated to the DNA fragments, allowing the DNA molecule to bind to a tether protein that guides the DNA through the nanopore. An exonuclease located on the external side of the pore cleaves the DNA strand that will pass through the pore (Fig. 10). Each nucleotide that passes through the nanopore perturbs the applied current in a nucleotide-specific manner (Fig. 10). Unlike other sequencing technologies, the current variation is the raw data of MinION sequencing, which is encoded into fast5 files (Fig. 10) (Lu *et al.*, 2016).

### 12.2. Converting nanopore sequencing data to base sequences

The sequence of bases representing the nucleotide order as determined from a nucleic acid fragment using a sequencing platform is named a read. Most bio-informatics tools accept reads encoded into fastq or fasta files. The fastq and fasta formats carry the sequence order information in bases, but fastq files additionally encode quality data of each base. Fast5 files contain the electric signal of current variation and has to be translated into bases through a process called base-calling (Fig. 10, (Lu *et al.*, 2016)). The base-calling step is time- and resource consuming, but running time can be reduced using parallel computing afforded by graphics processing units (GPUs). This step can be performed “on board” if the system possesses the required capacities (e.g., using the MinION Mk1C platform) or on another machine. The base-caller currently used by the community is called Guppy and so far, three models have been developed with different accuracy levels: fast, high accuracy (hac) and super accuracy (sup). The base-calling models are constantly being improved to reduce read error rates.

### 12.3. Generation of error-prone data

The high error rates observed in ONT-generated reads is mainly due to the way the signal is acquired, particularly the speed of DNA translocation through the pore (450 nucleotides per second, Wang *et al.*, 2021). The data on which base-calling models are trained is a crucial parameter for correctly interpreting the current variation signal. A particular problem faced using ONT sequencing technology, is homopolymers, sequences of a  $k$ -repetitive nucleotide with  $k > 2$ . Because homopolymer sequences give a constant similar signal as they pass through the pore, base-callers struggle to distinguish or precisely assess the number of similar nucleotides. To better address homopolymers, ONT developed a new generation of pores, the R10, with the goal of increasing the accuracy over homopolymers (Reviewed in Wang *et al.*, 2021). This year, ONT has developed a new pore with its specific chemistry: R14, which should generate more reads with an accuracy of about 99.9%.

## 13. Assembling STEC genome using long-read sequencing data

### 13.1. Long reads advantage for STEC genome assembly

Genome assembly is the term used for *in silico* reconstruction of the targeted genome from reads. The accuracy of short-read sequencing has the advantage of providing precise information on specific characteristics of the strain. For example, it is more accurate to determine MLST, serotype, and identify virulence genes and genetic variants. We have previously shown (Section 4, Chapter 1) that STEC strains can harbor non-negligible amount of mobile genetic elements, sometimes repeated (e.g. multiple *stx*-phages). Figure 11 represents the different size of reads obtained from short- and long-read sequencing. Due to the short-reads (50-300bp) generated with SGS, assemblers often fail to resolve structural variants, or repeated elements longer than the read length (Fig. 11A; (Schadt *et al.*, 2010)). Long reads generated with the MinION have the potential to resolve repetitive regions that present an unsolvable challenge for short-read assemblers (Fig. 11B) (Alkan *et al.*, 2011; Pollard *et al.*, 2018).

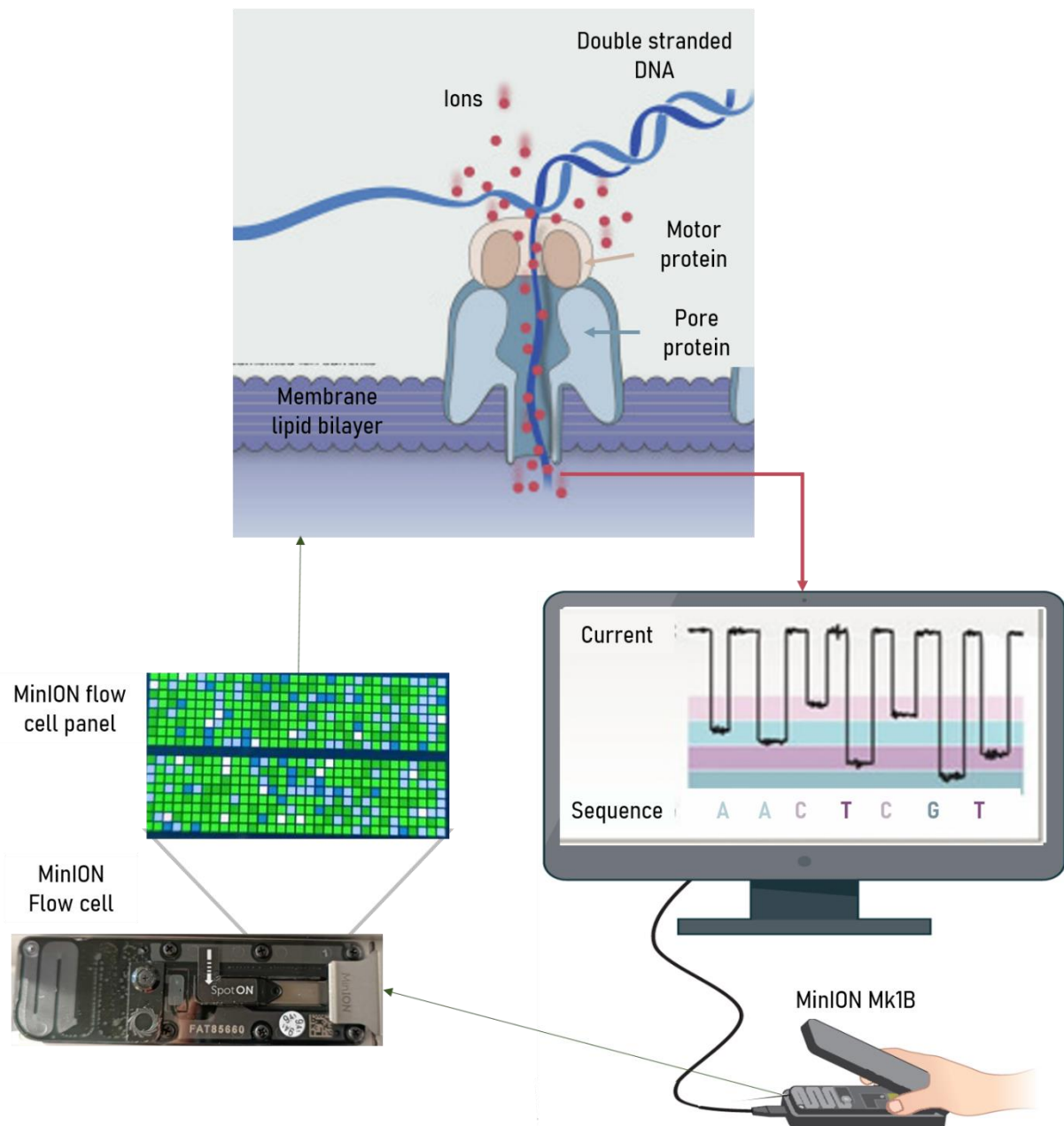


Figure 10 : MinION sequencing principle.

The MinION flow cells inserted onto the Mk1B device, possess a membrane into which nanopores are embedded. The DNA sample is loaded onto the flow cell and the motor protein catches DNA molecules and guides them through the pore. An ionic current is applied through the pore and the passage of nucleotide disturbs the current in a nucleotide-specific way. This raw signal is then converted into sequence (bases) during the base-calling step. Adapted from [https://www.cell.com/the-innovation/fulltext/S2666-6758\(21\)00078-3](https://www.cell.com/the-innovation/fulltext/S2666-6758(21)00078-3) (Applications and potentials of nanopore sequencing in the (epi)genome and (epi)transcriptome era: The Innovation (cell.com)) and Nanopore Sequencing Market Revolutionary Trends (2020-2026) by Industry Statistics | Medgadget.

### 13.2. Long-read assembly strategies

*De novo* assemblers use overlapping sequences present on reads to reassemble the reads into longer sequences called contigs. Reference-based assembly reconstructs the target genome by mapping reads onto a reference genome. The accuracy offered by reference-based assembly is crucial for small nucleotide variations (SNVs) including insertions and deletions (indels) analysis but is not suitable for the analysis of structural variation (SV) analysis ( $\geq 50$ bp) (Fitzgerald *et al.*, 2021). *De novo* assembly, which does not use a reference genome, is challenged by the high error-rate of long reads and by the computational part that is time consuming and resource intensive. Considering the genomic plasticity of *E. coli* strains, *de novo* assembly methods are more suitable than reference-based assembly.

Different *de novo* long-read assemblers have been developed based on two main methods: overlap-layout-consensus (OLC) and de Bruijn graph (DBG) methods. OLC methods are comparing all reads two by two, referred as pairwise comparison. They appear to perform better at assembling long error-prone reads, but are computationally more demanding (Cherukuri and Janga, 2016). Alternative approaches, derived from de Bruijn graph (DBG) methods have the advantages of being time-efficient since they generate *k*-mers that are short sequences of DNA of *k* length, from reads and uses those *k*-mers to find overlaps (Reviewed in Goussarov *et al.*, 2022). Long reads generated with ONT technologies are usually error-prone and errors such as indels can be passed from reads to the assembled genome. One can use a read correction tool prior to assembly, perform hybrid assembly using long and short reads, or correct the generated assembly (contigs) using reads in a process called polishing (Fig. 11C) (Reviewed in Meng *et al.*, 2022). A polishing step reduces the error rate of the reconstructed genome. Most long-read assemblers do not natively include polishing steps. Many tools have been developed specifically for polishing long-read-based assemblies, such as Racon or medaka (NanoporeTech; Vaser *et al.*, 2017). Polishing tools use the quality scores of each base encoded in the MinION fast5 file but require a reasonable read depth and multiple rounds to correct errors (Safar *et al.*, 2023; Zhang *et al.*, 2020).

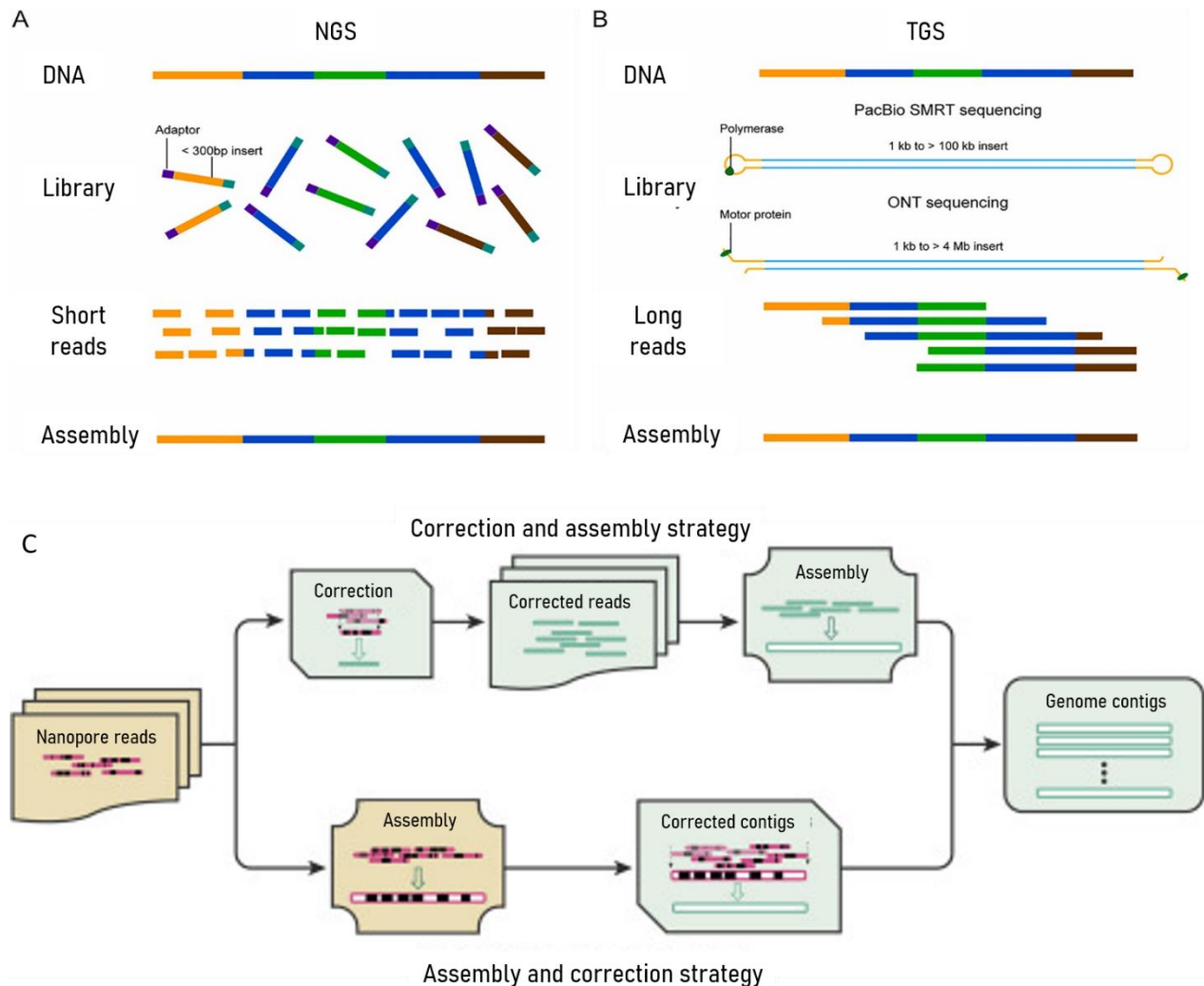


Figure 11 : Difference of short- (A) and long- (B) read sequencing on genome assembly and representation of long-read assembler strategies (C).

Benefit of long-read sequencing for reconstructing STEC genomes; taken from [https://www.researchgate.net/figure/Overview-of-NGS-short-read-and-TGS-long-read-methods-A-In-NGS-by-Illumina-technology\\_fig1\\_355491654](https://www.researchgate.net/figure/Overview-of-NGS-short-read-and-TGS-long-read-methods-A-In-NGS-by-Illumina-technology_fig1_355491654). Long-read assemblers either correct error-prone reads prior to assembly or correct the generated assembly using polishing tools (C); taken from Applications and potentials of nanopore sequencing in the (epi)genome and (epi)transcriptome era: The Innovation (cell.com).

## 14. Long-read metagenomics for food-borne pathogens identification

### 14.1. How long-read metagenomics might help STEC characterization from complex matrices

WGS has been helpful in characterizing and tracking pathogenic bacteria from food samples. Isolation-free approaches would save time in food-borne outbreak investigations (EFSA, 2019). Metagenomics is an isolation-, sometimes culture-independent method that enables the analysis of the totality of the genetic material present in a sample. Short-read metagenomics approaches have been used for identifying STEC contamination but do not permit the identification of STEC carrying additional adhesion mechanisms as LEE-positive STEC (Leonard *et al.*, 2016). The use of long-read sequencing for the detection of food-borne pathogens without isolation has been shown to be suitable for identifying food matrices contaminated with *Listeria* or *Salmonella* (Azinheiro *et al.*, 2022, 2021; Kocurek *et al.*, 2023). While the detection of specific *Salmonella* serovars in food samples is proof of contamination, it is more complex for the identification of *eae*-positive STEC. Particularly, *stx* virulence genes can be carried by different pathotypes of *E. coli*, as it was also shown to be the case for EAEC, ExPEC or even ETEC. The mere detection of the various virulence genes on separate contigs does not bring significant additional information compared to their detection by qPCR. Only the demonstration of their association in a single strain can inform on the pathogenic potential of the strains (Section 9.3, Chapter 2). STEC genome assembly from long-read metagenomics would potentially help identifying *eae*-positive STEC contamination and more generally characterize hybrid STEC strains (Section 1.8, Chapter 1).

### 14.2. Why an assembly-based approach is necessary for STEC characterization?

The advantage of genome assembly is that it provides a great deal of information about the bacterial genome and allows for better characterization of strains that are implicated in food-borne outbreaks. However, assembling complete genomes from metagenomics data is challenging. The presence of bacterial genomes containing genomic repeats but also phylogenetically close genomes can lead to fragmented assemblies or chimeric contigs carrying genomic regions from two or more different strains (Ekaterina Kazantseva *et al.*, 2023;

Vicedomini *et al.*, 2021). In addition, a fraction of reads may not be assembled, resulting in missing genomic regions (Ekaterina Kazantseva *et al.*, 2023). Typically, assemblies performed using metagenomics data generate fragmented and non-overlapping contigs that may belong to the same species. Therefore, an additional step is usually performed to group taxonomically related contigs and is called binning. The aim of the binning step is to reconstruct individual genomes from the metagenomics assembly, referred to as metagenome-assembled genomes (MAGs). Due to possible contamination during the binning step, some parameters have been described such as genome completeness (>90%) and contamination (<5% based on the number of copy of core genes) of the MAGs (Bowers *et al.*, 2017). After validation, the generated MAGs are classified and annotated (Goussarov *et al.*, 2022). Figure 12 represents the traditional workflow that leads to genome characterization from metagenomes.

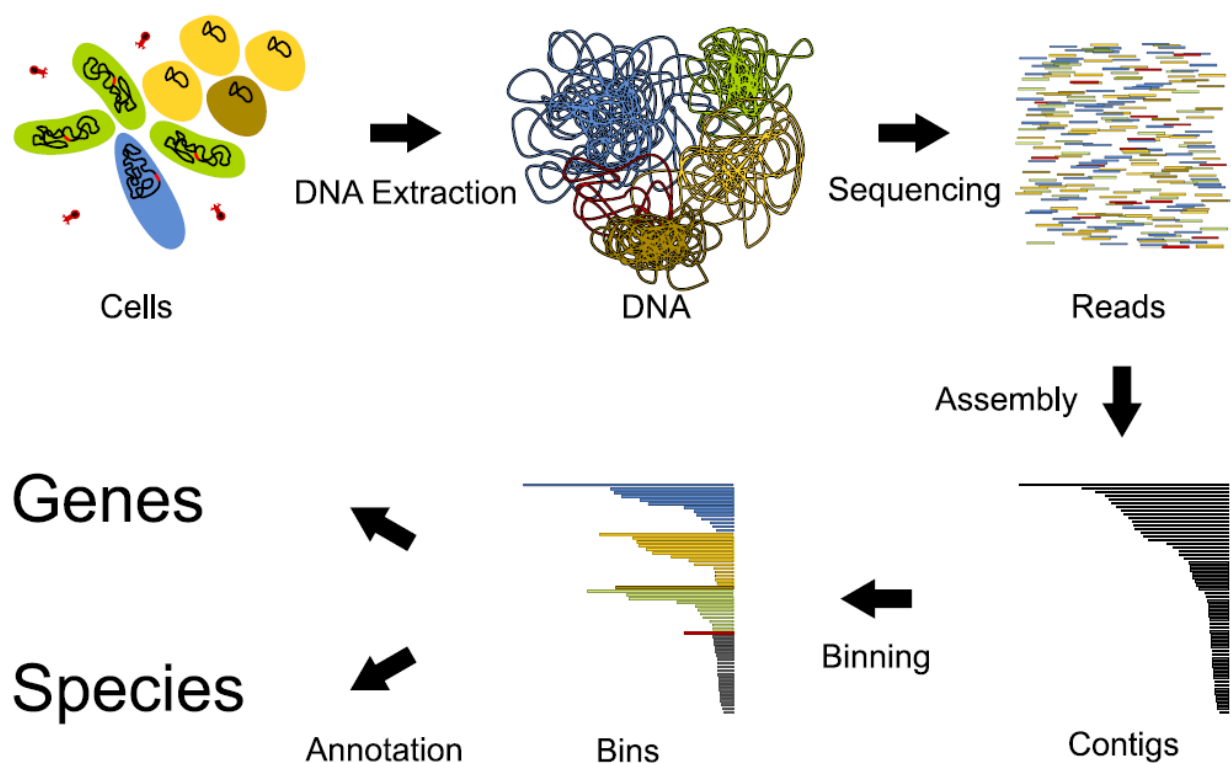


Figure 12 : Traditional workflow for genome assembly from metagenomics data. Taken from Goussarov *et al.*, 2022



The assembler Flye performs well in reconstructing STEC genomes and was the first long-read assembler that has been adapted to assemble metagenomes (MetaFlye), but is limited by the presence of highly similar or related species (Kolmogorov *et al.*, 2020; Vicedomini *et al.*, 2021; Wick and Holt, 2019). Beef samples and wastewater artificially contaminated with STEC have shown the potential of long-read metagenomics to identify STEC contamination (Buytaers *et al.*, 2021; Maguire *et al.*, 2021).

---

## Chapter 4: Aims of the work

---

Shiga toxin-producing *Escherichia coli* (STEC) are zoonotic food-borne pathogens responsible for a variety of intestinal symptoms that can evolve into the life-threatening hemolytic uremic syndrome (HUS). STEC infections require ingestion of the bacteria, adhesion to the host cell and production of the Shiga toxin.

Food products such as ground meat or dairy products might be contaminated with STEC from cattle intestinal tract. To protect the consumer from STEC infections, these food products are tested for the presence of pathogenic STEC before commercialization. Enriched food products are screened using qPCR methods that target virulence genes characteristic of highly virulent STEC. Characterization of the contaminating STEC strains requires an isolation step. However, due to the lack of selective medium, isolation of STEC strains from complex matrices is challenging. We aimed at using long-read sequencing, using Oxford Nanopore Technology, for characterizing STEC directly from raw milk without an isolation step. Although STEC carrying alternate adhesion factors than the LEE have been responsible for severe human infections, LEE-positive STEC were more frequently implicated in HUS. Thus, we decided to use an *eae*-positive STEC of serotype O26:H11, representing the most frequent serotype in Europe causing HUS, as a model organism.

I have first determined that an enrichment step is necessary for identifying STEC from complex matrices such as raw milk using MinION sequencing. Different enrichment conditions to enhance STEC growth in raw milk enrichment broth were compared and 37°C in acriflavine-supplemented BPW was selected. Long-read sequencing requires the extraction in high quantity of HMW gDNA (ONT recommends 1µg of genomic DNA as starting material). The next objective was to find a DNA extraction protocol that permits the recovery of DNA fitting ONT requirements. DNA extraction is a crucial step of the method since it will influence sequencing. Hence, we have tested and compared different extraction methods and found that the salting-out method allowed us to extract high quantity of HMW DNA suitable for ONT sequencing. This allowed us to sequence and completely assemble (closed or almost closed) 75 genomes of *E. coli* strains of bovine origin from our collection (Chapter 5-1).

Meanwhile, due to the complexity of data generated using the ONT sequencing, I have developed a pipeline for *eae*-positive STEC characterization from long-read metagenomics data: STECmetadetector. Using the best enrichment condition for STEC growth in artificially contaminated raw milk, the selected DNA extraction method and the pipeline, I successfully characterized the inoculated strain from an inoculation level of 5 CFU.mL<sup>-1</sup>. The characterization limit was assessed to 10<sup>8</sup> copies.mL<sup>-1</sup> post-enrichment based on qdPCR results (Chapter 5-2). In this work, I observed that the presence of more than one *E. coli* strain may hamper the direct characterization of *eae*-positive STEC strain in enriched raw milk using an assembly-based approach. I performed artificial co-contamination experiments with increasing ratio of commensal/*eae*-positive STEC strains and determined that in addition to the required quantities of STEC post-enrichment (10<sup>8</sup> copies.mL<sup>-1</sup>), the STEC should be 10-times more present than other *E. coli* to fully characterize the *eae*-positive STEC using long-read metagenomics and the STECmetadetector pipeline.

Although genome assembly is crucial to characterize STEC strains, we have tested alternative approaches specifically for *eae*-positive STEC identification. We have generated a large dataset of *E. coli* genomes and harnessed the power of machine learning algorithms to identify some markers that correlate with the presence of *eae*-positive STEC in genome assemblies obtained from long-read *E. coli* sequencing data. The machine learning-based approach correctly predicted the presence of the *eae*-positive STEC even in presence of other *E. coli* and could further be implemented in the STECmetadetector pipeline. Additionally, the identified markers were implemented as qPCR targets for STEC screening (Chapter 5-3).

Lastly, I applied the developed method on presumptive or naturally contaminated samples (Chapter 5-4) to test the applicability of the method, which confirmed the limiting conditions as previously determined.

---

## Chapter 5: Development of a long-read metagenomics approach for identifying *eae*-positive STEC

---

### Chapter5-1: Obtaining complete STEC genomes using long-read sequencing

The main requirement for long-read sequencing is the extraction of genomic DNA in sufficient quantities since Oxford Nanopore Technologies (ONT) recommends using 1 µg of genomic DNA. Besides DNA concentration, purity and integrity of the extracted DNA are two additional important parameters. DNA purity is critical to obtain decent amount of data generated as the presence of contaminants may block the MinION flow cell's pores. Long DNA fragments are needed to generate long reads that are desired to span mobile genetic elements. The objective of this work was to find a DNA extraction method that allows the recovery of HMW DNA in high quantities and matching the requirements of ONT sequencing. We first tested three DNA extraction methods (bead-based, solid phase or salting-out) on STEC pure cultures. The DNA yield, quality and integrity of the extracted DNA were compared. Additionally, the quality of data (particularly read length) generated using MinION sequencing as well as generated assembly metrics were compared for bead-based and salting-out DNA extraction methods. While longer reads were generated from DNA extracted using the salting-out methods, no difference was observed on the STEC assembly generated. Based on all results we have selected the salting-out method (Publication 1).

Due to their genomic complexity, there is a lack of completely assembled STEC genomes from strains originally isolated from food matrices. We aimed to characterize the genome (and in particular the mobilome) of STEC isolates of bovine food origin in order to estimate their genetic diversity. By using both short- and long-read sequencing technologies, we combine the potential of reconstructing the genome structure (long-read sequencing) and a reduced error rate (short-read sequencing) in STEC genome assembly. Samples matching the required conditions for MinION sequencing were sequenced using both Illumina Miseq short-read sequencing and ONT MinION sequencing technologies. Hybrid assemblies were generated by either assembling short reads and scaffolding using long reads (Unicycler) or the other way

around (Canu, Flye, Raven). Using this hybrid approach, we have reconstructed the genomes of 75 *E. coli* strains deposited on NCBI to benefit the community (Publication 2).

Lastly, because raw milk is complex matrix, I tested the performances of the DNA extraction method selected on raw milk. I compared the salting-out method previously selected from STEC pure cultures with a bead-based kit able to recover HMW gDNA (Quick-DNA HMW MagBead extraction kit from Zymo research). Comparison was made from fresh and frozen enriched raw milk samples. Based on the results, I concluded that the salting-out method was the best in this case (Additional experiment 1, unpublished).

## Publication 1

### **Evaluation of high molecular weight DNA extraction methods for long-read sequencing of Shiga toxin-producing *Escherichia coli***

Jaudou Sandra, Tran Mai-Lan, Vorimore Fabien, Fach Patrick, Delannoy Sabine

PLOS ONE 17(7): e0270751. <https://doi.org/10.1371/journal.pone.0270751>

RESEARCH ARTICLE

# Evaluation of high molecular weight DNA extraction methods for long-read sequencing of Shiga toxin-producing *Escherichia coli*

Sandra Jaudou<sup>1</sup>, Mai-Lan Tran<sup>1,2</sup>, Fabien Vorimore<sup>2</sup>, Patrick Fach<sup>1,2</sup>, Sabine Delannoy<sup>1,2\*</sup>

**1** Pathogenic E. coli Unit, Laboratory for Food Safety, Anses, Maisons-Alfort, France, **2** IdentityPath Platform, Laboratory for Food Safety, Anses, Maisons-Alfort, France

\* [sabine.delannoy@anses.fr](mailto:sabine.delannoy@anses.fr)



## Abstract

Next generation sequencing has become essential for pathogen characterization and typing. The most popular second generation sequencing technique produces data of high quality with very low error rates and high depths. One major drawback of this technique is the short reads. Indeed, short-read sequencing data of Shiga toxin-producing *Escherichia coli* (STEC) are difficult to assemble because of the presence of numerous mobile genetic elements (MGEs), which contain repeated elements. The resulting draft assemblies are often highly fragmented, which results in a loss of information, especially concerning MGEs or large structural variations. The use of long-read sequencing can circumvent these problems and produce complete or nearly complete genomes. The ONT MinION, for its small size and minimal investment requirements, is particularly popular. The ultra-long reads generated with the MinION can easily span prophages and repeat regions. In order to take full advantage of this technology it requires High Molecular Weight (HMW) DNA of high quality in high quantity. In this study, we have tested three different extraction methods: bead-based, solid-phase and salting-out, and evaluated their impact on STEC DNA yield, quality and integrity as well as performance in MinION long-read sequencing. Both the bead-based and salting-out methods allowed the recovery of large quantities of HMW STEC DNA suitable for MinION library preparation. The DNA extracted using the salting-out method consistently produced longer reads in the subsequent MinION runs, compared with the bead-based methods. While both methods performed similarly in subsequent STEC genome assembly, DNA extraction based on salting-out appeared to be the overall best method to produce high quantity of pure HMW STEC DNA for MinION sequencing.

## OPEN ACCESS

**Citation:** Jaudou S, Tran M-L, Vorimore F, Fach P, Delannoy S (2022) Evaluation of high molecular weight DNA extraction methods for long-read sequencing of Shiga toxin-producing *Escherichia coli*. PLoS ONE 17(7): e0270751. <https://doi.org/10.1371/journal.pone.0270751>

**Editor:** Mark Eppinger, University of Texas at San Antonio, UNITED STATES

**Received:** August 26, 2021

**Accepted:** June 16, 2022

**Published:** July 13, 2022

**Copyright:** © 2022 Jaudou et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

**Data Availability Statement:** All relevant data are within the manuscript and its [Supporting Information](#) files. All sequencing data have been deposited in the NCBI SRA database (Accession number: PRJNA808207).

**Funding:** The authors received no specific funding for this work.

**Competing interests:** The authors have declared that no competing interests exist.

## Introduction

*Escherichia coli*, a natural inhabitant of the digestive tracts of humans and animals, is a very diverse species, which comprise commensals as well as various pathogens. Among the latter,

Shiga toxin-producing *E. coli* (STEC) are responsible for serious gastro-intestinal diseases ranging from aqueous diarrhea to hemorrhagic colitis and hemolytic uremic syndrome.

Besides the Shiga toxin, many virulence factors appear to be involved in STEC pathogenicity. Most of these genes are localized on mobile genetic elements (MGEs) such as bacteriophages and plasmids [1–4]. These genes, which may account for up to 20% of the genome of *E. coli* confer a high plasticity to the genome of STEC strains.

Whole Genome Sequencing (WGS) currently represents the best method for high-resolution genome typing, exploration of the ‘mobilome’ and characterization. The number of sequence data in databases has exploded over the last 10 years, however, most genomes deposited are drafts genomes composed of a variable number of fragments or contigs. In particular, STEC genome sequences are particularly fragmented, often around 200 contigs. Short-read sequencing data of STEC is difficult to assemble because of the presence of numerous MGEs and repeated elements in the STEC genome, including ribosomal genes, transposons and insertion sequences. This fragmentation results in a loss of information on MGEs (especially plasmids and phages), horizontal gene transfers, copy number variations, as well as structural variations.

The current development of long-read sequencing is thus of particular interest for STEC genomics. The MinION sequencer can produce very long reads (more than 2 Mbp have been obtained [5]) which can span the repeated elements and produce closed or almost closed genomes [6].

One drawback of this technology compared to short read sequencing is the requirement for a high quantity of high molecular weight (HMW) DNA [7]. Many DNA extraction methods have been described and evaluated in specific settings [8–10]. It appears however that there is no universal HMW DNA extraction method. The efficiency of the method appears highly dependent on the nature of the sample: carbohydrates composition, nucleotide composition, topology, and association with proteins [11]. Furthermore, choosing the “right” method is a balance of many additional criteria including costs, handling time and user friendliness, which overall evaluation is specific to each situation.

In order to implement third-generation sequencing routinely in our lab, we needed a method that would consistently produce HMW STEC DNA, in sufficient quantity and quality, but which was also fast, practical and suitable for BSL3 work; hence, the use of toxic chemicals requiring specific handling and disposal (*i.e.* phenol-chloroform) was prohibited. DNA extraction methods usually used to produce DNA for short-read sequencing often comprise numerous centrifugation steps that can be detrimental to DNA integrity. As the spin-column method is routinely performed in many labs, including our own, it was nonetheless included in this comparison.

In this study, we tested three different methods to extract HMW DNA from STEC for the purpose of MinION sequencing: bead-based, column-based (*i.e.* solid phase), and salting-out, and evaluated their impact on STEC DNA quantity, quality and integrity as well as performance in long-read sequencing and subsequent assembly.

## Material and methods

### • STEC strains collection

A total of 83 STEC strains (including 34 strains positive for *eae*, and 49 strains negative for *eae*) from 30 different serogroups and 36 different serotypes were selected from the Anses collection to be sequenced and assembled in this study (S1 Table). Most of the strains selected were of bovine origin isolated in France from dairy products or meat. A few strains were isolated from goat milk and cheese.



### • STEC strains culture

Isolates were stored at  $-80^{\circ}\text{C}$  in 20–30% glycerol, revived on TSYe plates (BioMérieux) overnight at  $37^{\circ}\text{C}$  and cultured in 10 mL of BHI medium (BioMérieux) overnight at  $37^{\circ}\text{C}$  and 320 rpm. For each strain, genomic DNA was extracted from 1 mL overnight culture using various DNA extraction and purification kits.

### • Genomic DNA extraction

Three different DNA extraction principles were used: (i) bead-based, using beads from four different suppliers: AMPureXP (Beckman Coulter™) ( $n = 24$ ), HighPrepPCR (MagBio) ( $n = 7$ ), NucleoMag™ (Macherey-Nagel™, Fisher Scientific) ( $n = 6$ ) and MagAttract HMW DNA Kit (Qiagen) ( $n = 20$ ), (ii) solid-phase using silicium columns (Monarch® Genomic DNA Purification Kit, New England BioLabs®) ( $n = 10$ ), and (iii) salting-out with isopropanol precipitation (MasterPure DNA extraction and purification kit, Lucigen) ( $n = 55$ ) (S2 Table). The Monarch and MasterPure DNA extraction kits were used according to the manufacturer's recommendations, except for the final DNA elution / rehydration, which was performed in 10 mM Tris-HCl (pH 8.5) (EB Buffer, Qiagen), and included an RNaseA treatment (RNaseA  $100\text{ mg}\cdot\text{mL}^{-1}$ , Qiagen) immediately after cell lysis. The procedure for HMW DNA extraction using magnetic beads is described in supplementary material (S1 File). In each procedure, the use of vortex mixer was reduced to a minimum and pipetting steps were performed slowly to limit DNA shearing.

All DNA extracts were kept at  $+4^{\circ}\text{C}$  after extraction (storage up to 10 months).

### • Genomic DNA quantification and quality control

To allow appropriate relaxation and refolding, the gDNA was quantified at least 24 hours after extraction with a Qubit 3.0 Fluorometer (Thermo Fisher Scientific) using the Qubit dsDNA BR Assay Kit (Thermo Fisher Scientific), according to the manufacturer's instruction.

The purity of gDNA was estimated with a NanoDrop ND-2000 spectrophotometer (Thermo Fisher Scientific) by calculating  $A_{260}/A_{280}$  and  $A_{260}/A_{230}$  ratios.

The gDNA integrity was determined with a TapeStation 4200 (Agilent) using Genomic DNA Screentapes (Agilent) following manufacturer's instructions (S3 Table).

### • ONT MinION sequencing

A total of 66 samples were selected for sequencing, 47 samples extracted with the salting-out method and 19 samples extracted with bead-based methods. Libraries were prepared from 1 to 2  $\mu\text{g}$  input DNA using the SQK-LSK109 Ligation Sequencing kit in conjunction with the EXP-NBD104 and EXP-NBD114 Native Barcode Expansion kits (Oxford Nanopore Technologies, Oxford, UK) in accordance with the manufacturer's instructions. Libraries (6 to 14 per flow cell) were loaded onto R9.4.1 flow cells (Oxford Nanopore Technologies) and sequenced using the MK1B MinION device (Oxford Nanopore Technologies) for 48 h without live basecalling.

### • Illumina MiSeq sequencing

Libraries were prepared from 1 ng of gDNA using the Nextera XT DNA Library Preparation Kit (Illumina Inc.) following the manufacturer's instructions. Libraries were sequenced using the MiSeq Reagent Kit v2 (300 or 500-cycles) (Illumina Inc.) on a MiSeq System. The quality of the raw short reads was checked using the CLC Genomic Workbench version 21.0.5.

### • Sequencing data analysis

Basecalling of the raw fast5 data was performed with Guppy basecaller version 4.4.2. Demultiplexing was performed with Guppy barcoder version 4.4.2.

Basic run metrics statistics were calculated using in-house python scripts (S1-S3 Scripts in [S2 File](#)) (total number of reads, total number of bases sequenced, average size of reads, median size of reads, maximum read length, minimum read length, read length N50) and Nanoplot version 1.29.0 (mean read quality, median read quality) [12].

Raw MinION reads were assembled using Raven version 1.2.2 [13] with default parameters and Flye version 2.8.1 (—nano-raw; genome size: 5m; iterations: x5) [14]. Hybrid assemblies using MiSeq and MinION reads were assembled using Unicycler version 0.4.8 [15, 16] with default parameters. The contiguity and number of contigs of each genome assembly were assessed using Quast version 5.0.2 [17] without reference genome.

### • Statistical analysis

The PCA analysis was performed on R Studio version 4.0.3 and the following packages: factoextra v.1.0.7 and FactoMineR v.2.4. The ellipse.level parameter was set to 0.95.

All other statistical analyses were performed on R Studio version 1.2.5019.

Descriptive analyses of percentages (DNA yield and DNA purity, DIN) were performed using the average and standard deviation (s.d.). Non-parametric statistical tests were performed with  $\alpha$  of 5%. The null hypothesis was rejected when p-values were  $< 0.05$ .

The Kruskal-Wallis test was used to analyze yield and purity according to extraction kit or extraction method. Mutual comparison of test groups was performed using a Dunn post-hoc analysis with Holm correction.

The Wilcoxon rank sum test was used to analyze read length and quality metrics as well as assembly contiguity metrics between extraction methods.

The Kendall rank correlation was used to analyze association between read length and STEC assembly contiguity.

### • Data availability

MinION and MiSeq raw sequencing data were deposited in NCBI SRA database under the Bioproject number PRJNA808207.

## Results and discussion

The preparation of high-quality high molecular weight (HMW) genomic DNA (gDNA) is critical to fully exploit the capacity of the MinION sequencing platform. In this study, we evaluated several parameters in order to determine the relative effectiveness of three different STEC genomic DNA extraction methods for MinION sequencing.

A collection of STEC strains (positive or not for *eae*) were selected for DNA extraction ([S1 Table](#)). All the strains originated from cattle-related matrix (meat, raw-milk, raw-milk cheese, animal) and had previously been serotyped and characterized by qPCR for the presence of the *stx* and *eae* genes [18, 19].

### • Quantity and quality of the extracted DNA

The yield, concentration, purity, and integrity of the extracted DNA were evaluated ([Table 1](#) and [S2](#) and [S3 Tables](#)). The DNA concentrations were determined with the Qubit Fluorometer. We used the NanoDrop absorption spectra to assess sample purity and identify potential contamination such as carbohydrates and extraction chemicals. The A260/280 ratio for pure

**Table 1. Performance comparison of DNA extraction methods.**

DNA extraction kit	DNA extraction method	Mean DNA yield (ng / ml culture) (sd)	Mean A260/280 (sd)	Mean A260/230 (sd)
AMPureXP	Bead-based	4835 (3044)	1.76 (0.19)	0.49 (0.18)
HighPrepPCR	Bead-based	3280 (1668)	1.61 (0.08)	0.52 (0.07)
NucleoMag	Bead-based	2756 (948)	1.58 (0.05)	0.64 (0.16)
Monarch	Solid phase	3162 (1078)	1.60 (0.43)	1.11 (0.40)
MasterPure	Salting-out	8687 (3462)	2.05 (0.09)	2.00 (0.26)
MagAttract	Bead-based	1582 (764)	2.08 (0.04)	1.47 (0.44)

<https://doi.org/10.1371/journal.pone.0270751.t001>

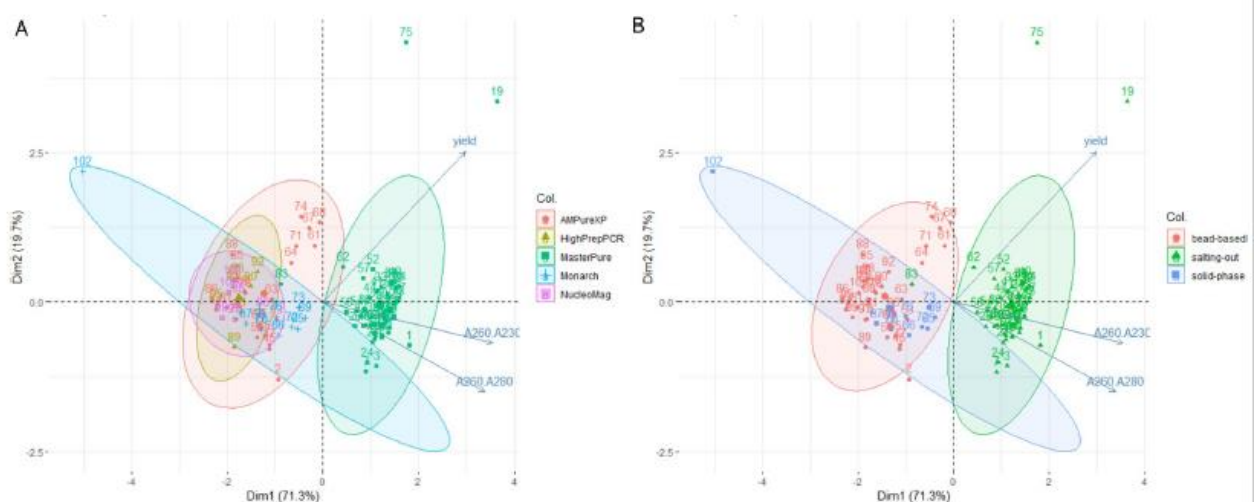
DNA is expected around 1.8. Contaminants absorbing at 280 nm include proteins and phenol. The A260/230 ratio can indicate possible residual chemical contamination such as EDTA, phenol, guanidine salts, or carbohydrates. An A260/230 ratio close to 2.0 is expected for a DNA preparation free of contamination. Integrity of the gDNA was estimated using the DNA integrity number (DIN) obtained from the TapeStation. A DNA sample with little to no short fragments will have a DIN value of >9. The DIN decreases when degradation of the DNA sample increases. For MinION libraries, DNA samples with DIN values >8 are preferred.

Yield and purity of the isolated DNA varied between extraction methods (Table 1).

Although it produced HMW DNA the DNA concentration and total yield obtained with the MagAttract kit without optimization were not deemed sufficient for MinION library preparation and it was thus omitted from the subsequent steps and analyses.

To have a global idea on DNA extraction methods / kits performance we performed a PCA biplot analysis on DNA yields and purity ratios (Fig 1). Two principal components stand out, the first (PC1) representing 71.3% of the variance between data and the second (PC2) which represent 19.7% of the variance. Altogether, these two dimensions represent the three variables DNA yield, and the two purity ratios A260/A280 and A260/A230.

The PCA biplot analysis indicated that the three bead-based extraction kits have similar DNA extraction results (Fig 1A). Consequently, the DNA extractions performed with the beads from the three remaining suppliers were grouped together as “bead-based”. Indeed, the yield, A260/A230 and A260/A280 ratios did not differ significantly between the three suppliers

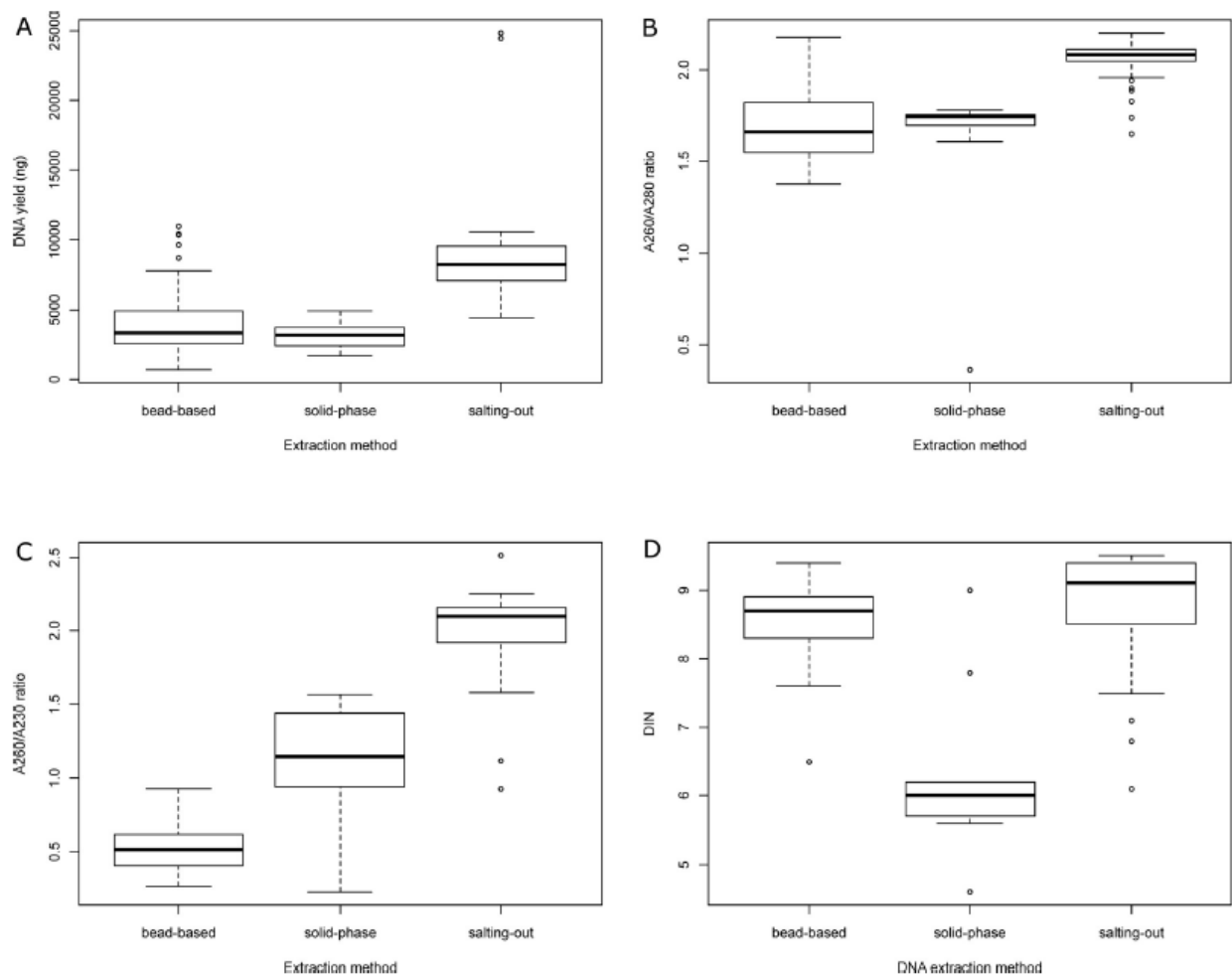


**Fig 1. PCA biplot analysis on DNA yield and quality recovered from STEC strains depending on (A) extraction kits and (B) method.**

<https://doi.org/10.1371/journal.pone.0270751.g001>

(Kruskal-Wallis test,  $p$ -values  $> 0.05$ ). Except for A260/280 ratio obtained with the AMPureXP beads, which was significantly different from that obtained with the NucleoMag beads (Kruskal-Wallis test,  $p$ -value = 0.02768 and Dunn's post hoc test,  $p$ -value = 0.026197) but not from that obtained with the HighPrepPCR beads. It is noteworthy that fewer extractions were performed with the NucleoMag and HighPrepPCR beads compared with AMPureXP beads.

The PCA biplot representation showed that the salting-out was the best method to obtain highly concentrated pure DNA compared to the other extraction methods tested in this study (Fig 1B). Taken together the DNA yields were significantly higher with the salting-out method than with the bead-based and solid-phase methods (Kruskal-Wallis test,  $p$ -value = 4.025e-11 and Dunn's post hoc test,  $p$ -values 1.6e-09 and 1.7e-06, respectively), which did not differ significantly from each other ( $p$ -value = 0.31) (Fig 2A). Similar extraction yields have previously been reported for STEC for these extraction methods [10].



**Fig 2. Quantity and quality of the extracted DNA according to the DNA extraction method.** (A) DNA yield ( $\text{ng}\cdot\text{mL}^{-1}$  culture) as determined with Qubit dsDNA Broad range assay kit, (B) A260/A280 ratio as determined with Nanodrop, (C) A260/A230 ratio as determined with nanodrop, (D) DIN value as determined with Genomic DNA Screenshot.

<https://doi.org/10.1371/journal.pone.0270751.g002>

DNA purity as assessed with the A260/280 and the A260/230 ratio differed significantly between the extraction methods (p-values <0.05). No statistically significant differences were observed between the solid-phase and bead-based extraction methods (p-values >0.05). The salting-out method allowed the extraction of significantly purer DNA compared to the solid-phase and the bead-based methods (p-values <0.05). Only DNA extracted using the salting-out extraction method showed A260/280 purity ratio with acceptable values >1.8 and A260/230 purity ratio close to 2.0 (Fig 2B and 2C). The low ratios obtained with the bead-based and solid-phase methods suggest incomplete removal of proteins and organic compounds. The ratios >1.9 obtained with the salting-out method suggest the presence of RNA in the sample, although, based on our experience, the presence of RNA is not detrimental to MinION library preparation.

The DNA extracted with the solid-phase method showed significant degradation of the DNA compared with the other methods (Fig 2D) with a median DIN value of 6.0 (sd 1.3). These results were expected due to the numerous centrifugation steps in this procedure. These DNA were deemed not suitable for MinION sequencing.

Both the bead-based and salting-out methods produced HMW DNA with fragments >60 kb and median DIN values of 8.7 (sd 0.7) and 9.1 (sd 0.8) respectively, indicative of low amounts of genomic DNA degradation (Fig 2D). Integrity of the DNA was checked up to 10 months after extraction and storage at +4°C and showed no sign of degradation with either bead-based or salting-out method.

The bead-based and salting-out methods were the most appropriate methods to yield sufficient DNA amounts and concentrations with appropriate purity and integrity to perform MinION library preparation. To further demonstrate the suitability for long-read sequencing of the genomic DNA obtained, DNA extracted with the bead-based and salting-out methods were subsequently sequenced by the MinION technology.

### • Quality of the sequencing data

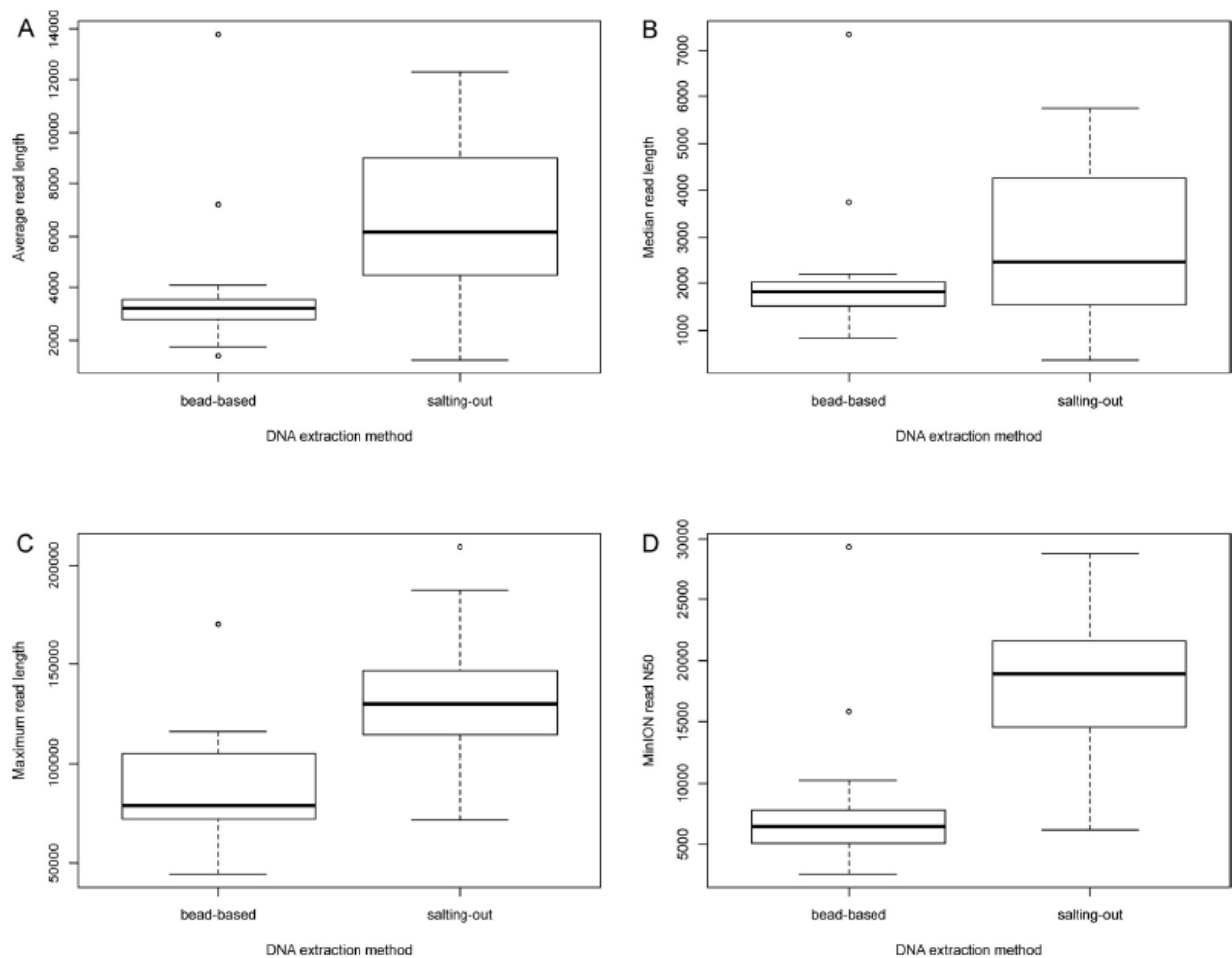
Forty-seven samples extracted with the salting-out method and 19 samples extracted with bead-based methods were sequenced with the MinION (S2 and S4 Tables). Quality control checks were performed on the raw data to evaluate the kit's influence on read quality (mean and median Q score, mean and median read length, read length N50).

Our data indicates that read size distribution is not uniform across extraction methods. The average and median read lengths obtained appeared quite short for some of the isolates, but overall similar read lengths have been reported for STEC and other food strains [20–24]. To avoid short fragments, a more stringent size selection could be applied during library preparation.

The average read length is significantly higher with the salting-out method (Wilcoxon test, p-value = 9.9e-06), but the median read length is not significantly different (p-value = 0.081) (Fig 3A and 3B).

Similarly, the maximum read length and total read length N50 are significantly higher with the salting-out method (Wilcoxon test, p-values 1.4e-06 and 7.4e-09, respectively) (Fig 3C and 3D). The salting-out method produced the longest fragments (up to 209 kbp) while the bead-based method produced fragments up to 169 kbp.

On the contrary, although the difference appeared modest, both the mean and median Q score were significantly higher with the bead-based method than the salting-out method (Wilcoxon test, p-value = 7.1e-05) (Fig 4), presumably due to the presence of much longer DNA fragments in the salting-out DNA extracts.



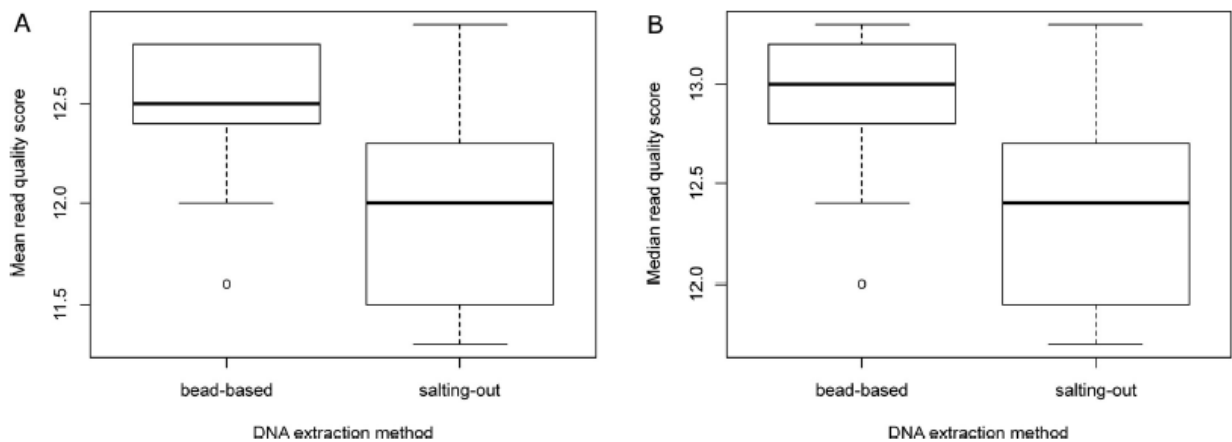
**Fig 3. Read length-related metrics of the sequencing data according to the DNA extraction method.** (A) Average read length (bp), (B) Median read length (bp), (C) Maximum read length (bp), (D) Read length N50 (bp).

<https://doi.org/10.1371/journal.pone.0270751.g003>

### • Contiguity of STEC assemblies

In order to check the influence of the extraction method and more generally of read length on the assembly contiguity we performed crude assemblies of the sequencing data. For each sample ( $n = 66$ ) we generated three assemblies, one hybrid assembly with Unicycler using MiSeq and MinION reads, and two assemblies with MinION data alone using Raven and Flye. The contiguity of the assemblies were determined with Quast (S4 Table).

When considering the data by extraction method, there was no difference in the number of contigs between the two extraction methods whatever the assembler used (Wilcoxon rank sum tests,  $p$ -values  $> 0.05$ ). While there was no difference between the two extraction methods concerning the N50 of the assemblies generated with Raven and Unicycler (Wilcoxon rank sum tests,  $p$ -values  $> 0.05$ ), the Flye assemblies N50 were significantly longer for the samples extracted with the salting-out method than with the bead-based method (Wilcoxon rank sum tests,  $p$ -value = 0.0118).



**Fig 4. Quality metrics of the sequencing data according to the DNA extraction method.** (A) Mean read Q score, (B) Median read Q score.

<https://doi.org/10.1371/journal.pone.0270751.g004>

We studied the correlation between read length, irrespective of the extraction method, and STEC genome assembly contiguity (number of contigs and assembly N50). Our data indicate that within the range of read lengths obtained in this study the influence of read length on assembly contiguity depends on the assembler used.

The quality of the assemblies generated by Raven are not associated with read length within this dataset (Kendall rank correlation tests,  $p$ -values  $> 0.05$ ). On the contrary, the quality of the assemblies generated by Flye and Unicycler appear to be associated with read length to some extent, longer reads generating assemblies with more contiguity (less contigs and a higher N50). Indeed, the average read length is significantly negatively correlated with the number of contigs generated by Flye and Unicycler (Kendall rank correlation test,  $p$ -values = 0.04525 and 0.01787 respectively, Kendall's tau = -0.17178 and -0.20308 respectively), while it is positively correlated with the Flye assemblies N50 (Kendall rank correlation test,  $p$ -value = 0.02, tau = 0.19), but not with the Unicycler assemblies N50 (Kendall rank correlation test,  $p$ -value = 0.09). Similarly the median read length is significantly negatively correlated with the number of contigs generated by Flye and Unicycler (Kendall rank correlation test,  $p$ -values = 0.02263 and 0.00311 respectively, Kendall's tau = -0.19557 and -0.2535 respectively), while it is positively correlated with the Unicycler assemblies N50 (Kendall rank correlation test,  $p$ -value = 0.04946, tau = 0.1655), but not with the Flye assemblies N50 (Kendall rank correlation test,  $p$ -value = 0.3729). The total read length N50 is negatively correlated with the number of contigs generated by Flye (Kendall rank correlation test,  $p$ -value = 0.04769, tau = -0.16987), but not Unicycler (Kendall rank correlation test,  $p$ -value = 0.2203), and it is positively correlated with the Flye assemblies N50 (Kendall rank correlation test,  $p$ -value = 0.00024, tau = 0.30909), but not Unicycler (Kendall rank correlation test,  $p$ -value = 0.2213).

Overall, samples obtained with both extraction methods performed similarly in the assembly step confirming that both extraction methods are suitable to produce HMW gDNA for MinION sequencing.

## Conclusion

In this study HMW DNA extraction methods were compared based on DNA yield, purity and integrity, as well as MinION read length and quality score and subsequent assembly.

Among all methods tested the salting-out method appears to be one of the best compromise in our hands. It produced highly concentrated HMW STEC DNA of required quality for MinION sequencing. In addition to the high yield, the highest integrity of the DNA obtained from this kit was also evidenced by the longest read length from the sequencing result. It should however be noted that studies [9, 10] have suggested that the salting-out method might be less efficient than other methods for the specific extraction of small plasmids (< 5 kb). This should be taken into account depending on the project. Concerning *E. coli*, and STEC in particular, these small plasmids < 5 kb are most likely not involved in virulence or antimicrobial-resistance, but could play a role in bacterial adaptation [25–27].

The bead-based extraction method also produced sequencing-grade HMW STEC DNA and appears a valid alternative to the salting-out method. It is however more tedious to perform with the bead drying time being difficult to standardize.

Finally, the column-based solid-phase method is easy to use and popular but shears DNA, which makes it less suited for long-read sequencing applications. It is also noteworthy that this method generates more waste than the salting-out and bead-based methods.

## Supporting information

**S1 Table. Characteristics of the STEC strains used in this study.**  
(XLSX)

**S2 Table. Extraction and sequencing data.**  
(XLSX)

**S3 Table. DIN values of extracted DNA.**  
(XLSX)

**S4 Table. Assembly metrics using Unicycler, Flye and Raven assemblers.**  
(XLSX)

**S1 File. Genomic DNA extraction with Solid Phase Reverse Immobilization (SPRI) beads.**  
(DOCX)

**S2 File. Python scripts used to calculate MinION reads metrics.**  
(ZIP)

## Author Contributions

**Conceptualization:** Mai-Lan Tran, Patrick Fach, Sabine Delannoy.

**Formal analysis:** Sandra Jaudou, Sabine Delannoy.

**Funding acquisition:** Fabien Vorimore, Patrick Fach, Sabine Delannoy.

**Investigation:** Sandra Jaudou, Mai-Lan Tran, Fabien Vorimore, Sabine Delannoy.

**Methodology:** Patrick Fach, Sabine Delannoy.

**Project administration:** Patrick Fach, Sabine Delannoy.

**Resources:** Sabine Delannoy.

**Supervision:** Patrick Fach, Sabine Delannoy.

**Writing – original draft:** Sandra Jaudou, Mai-Lan Tran, Sabine Delannoy.

**Writing – review & editing:** Sandra Jaudou, Mai-Lan Tran, Patrick Fach, Sabine Delannoy.



## References

1. Tobe T, Beatson SA, Taniguchi H, Abe H, Bailey CM, Fivian A, et al. An extensive repertoire of type III secretion effectors in *Escherichia coli* O157 and the role of lambdoid phages in their dissemination. *Proc Natl Acad Sci U S A*. 2006 Oct 3; 103(40):14941–6. <https://doi.org/10.1073/pnas.0604891103> PMID: 16990433; PMCID: PMC1595455.
2. Ogura Y, Ooka T, Asadulghani, Terajima J, Nougayrède JP, Kurokawa K, et al. Extensive genomic diversity and selective conservation of virulence-determinants in enterohemorrhagic *Escherichia coli* strains of O157 and non-O157 serotypes. *Genome Biol*. 2007; 8(7):R138. <https://doi.org/10.1186/gb-2007-8-7-r138> PMID: 17711596; PMCID: PMC2323221.
3. Noll LW, Worley JN, Yang X, Shridhar PB, Ludwig JB, Shi X, et al. (2018) Comparative genomics reveals differences in mobile virulence genes of *Escherichia coli* O103 pathotypes of bovine fecal origin. *PLoS ONE* 13(2): e0191362. <https://doi.org/10.1371/journal.pone.0191362> PMID: 29389941
4. Nakamura K, Murase K, Sato MP, Toyoda A, Itoh T, Mainil JG, et al. Differential dynamics and impacts of prophages and plasmids on the pangenome and virulence factor repertoires of Shiga toxin-producing *Escherichia coli* O145:H28. *Microb Genom*. 2020 Jan; 6(1):e000323. <https://doi.org/10.1099/mgen.0.000323> PMID: 31935184; PMCID: PMC7067040.
5. Payne A, Holmes N, Rakyan V, Loose M. BulkVis: a graphical viewer for Oxford nanopore bulk FAST5 files. *Bioinformatics*. 2019 Jul 1; 35(13):2193–2198. <https://doi.org/10.1093/bioinformatics/bty841> PMID: 30462145; PMCID: PMC6596899.
6. González-Escalona N, Allard MA, Brown EW, Sharma S, Hoffmann M. Nanopore sequencing for fast determination of plasmids, phages, virulence markers, and antimicrobial resistance genes in Shiga toxin-producing *Escherichia coli*. *PLoS One*. 2019 Jul 30; 14(7):e0220494. <https://doi.org/10.1371/journal.pone.0220494> PMID: 31361781; PMCID: PMC6667211.
7. Schalamun M, Nagar R, Kainer D, Beavan E, Eccles D, Rathjen JP, et al. Harnessing the MinION: An example of how to establish long-read sequencing in a laboratory using challenging plant tissue from *Eucalyptus pauciflora*. *Mol Ecol Resour*. 2019 Jan; 19(1):77–89. <https://doi.org/10.1111/1755-0998.12938> Epub 2018 Oct 5. PMID: 30118581; PMCID: PMC7380007.
8. Chacon-Cortes D, Griffiths L. Methods for extracting genomic DNA from whole blood samples: current perspectives. *Journal of Biorepository Science for Applied Medicine*. 2014; 2:1–9 <https://doi.org/10.2147/BSAM.S46573>
9. Becker L, Steglich M, Fuchs S, Werner G, Nübel U. Comparison of six commercial kits to extract bacterial chromosome and plasmid DNA for MiSeq sequencing. *Sci Rep*. 2016; 6:28063. <https://doi.org/10.1038/srep28063> PMID: 27312200
10. Nouws S, Bogaerts B, Verhaegen B, Denayer S, Piérard D, Marchal K, et al. Impact of DNA extraction on whole genome sequencing analysis for characterization and relatedness of Shiga toxin-producing *Escherichia coli* isolates. *Sci Rep*. 2020 Sep 4; 10(1):14649. <https://doi.org/10.1038/s41598-020-71207-3> PMID: 32887913; PMCID: PMC7474065.
11. Melzak K. A., Sherwood C. S., Turner R. F. B. & Haynes C. A. Driving Forces for DNA Adsorption to Silica in Perchlorate Solutions. *Journal of Colloid and Interface Science*. 1996 181, 635–644, <https://doi.org/10.1006/jcis.1996.0421>
12. De Coster W, D'Hert S, Schultz DT, Cruts M, Van Broeckhoven C. NanoPack: visualizing and processing long-read sequencing data. *Bioinformatics*. 2018 Aug 1; 34(15):2666–2669. <https://doi.org/10.1093/bioinformatics/bty149> PMID: 29547981; PMCID: PMC6061794.
13. Vaser R, Šikić M. Time- and memory-efficient genome assembly with Raven. *Nat Comput Sci*. 2021; 1, 332–336. <https://doi.org/10.1038/s43588-021-00073-4>
14. Kolmogorov M, Yuan J, Lin Y, Pevzner P. Assembly of Long Error-Prone Reads Using Repeat Graphs. *Nature Biotechnology*, 2019 <https://doi.org/10.1038/s41587-019-0072-8> PMID: 30936562
15. Wick RR, Judd LM, Gorrie CL, Holt KE. Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. *PLoS Comput Biol* 2017; 13(6): e1005595. <https://doi.org/10.1371/journal.pcbi.1005595> PMID: 28594827
16. Wick RR, Judd LM, Gorrie CL, Holt KE. Completing bacterial genome assemblies with multiplex MinION sequencing. *Microb Genom*. 2017 Sep 14; 3(10):e000132. <https://doi.org/10.1099/mgen.0.000132> PMID: 29177090; PMCID: PMC5695209.
17. Mikheenko A, Prijbelski A, Saveliev V, Antipov D, Gurevich A. Versatile genome assembly evaluation with QAST-LG. *Bioinformatics*. 2018; 34(13): i142–i150. <https://doi.org/10.1093/bioinformatics/bty266> PMID: 29949969
18. Bugarel M, Beutin L, Martin A, Gill A, Fach P. Micro-array for the identification of Shiga toxin-producing *Escherichia coli* (STEC) seropathotypes associated with Hemorrhagic Colitis and Hemolytic Uremic

- Syndrome in humans. *Int J Food Microbiol*. 2010 Sep 1; 142(3):318–29. <https://doi.org/10.1016/j.ijfoodmicro.2010.07.010> Epub 2010 Jul 14. PMID: 20675003.
19. Delannoy S, Beutin L, Fach P. Discrimination of enterohemorrhagic *Escherichia coli* (EHEC) from non-EHEC strains based on detection of various combinations of type III effector genes. *J Clin Microbiol*. 2013 Oct; 51(10):3257–62. <https://doi.org/10.1128/JCM.01471-13> Epub 2013 Jul 24. PMID: 23884997; PMCID: PMC3811616.
  20. Leidenfrost RM, Pöther DC, Jäckel U, Wünschiers R. Benchmarking the MinION: Evaluating long reads for microbial profiling. *Sci Rep*. 2020 Mar 20; 10(1):5125. <https://doi.org/10.1038/s41598-020-61989-x> PMID: 32198413; PMCID: PMC7083898.
  21. Taylor TL, Volkening JD, DeJesus E, Simmons M, Dimitrov KM, Tillman GE, et al. Rapid, multiplexed, whole genome and plasmid sequencing of foodborne pathogens using long-read nanopore technology. *Sci Rep*. 2019 Nov 8; 9(1):16350. <https://doi.org/10.1038/s41598-019-52424-x> PMID: 31704961; PMCID: PMC6841976.
  22. Goldstein S, Beka L, Graf J, Klassen JL. Evaluation of strategies for the assembly of diverse bacterial genomes using MinION long-read sequencing. *BMC Genomics*. 2019 Jan 9; 20(1):23. <https://doi.org/10.1186/s12864-018-5381-7> PMID: 30626323; PMCID: PMC6325685.
  23. Peritz A, Paoli GC, Chen CY, Gehring AG. Serogroup-level resolution of the "Super-7" Shiga toxin-producing *Escherichia coli* using nanopore single-molecule DNA sequencing. *Anal Bioanal Chem*. 2018 Sep; 410(22):5439–5444. <https://doi.org/10.1007/s00216-018-0877-1> Epub 2018 Jan 27. PMID: 29374775.
  24. Arredondo-Alonso S, Pöntinen AK, Cléon F, Gladstone RA, Schürch AC, Johnsen PJ, et al. A high-throughput multiplexing and selection strategy to complete bacterial genomes. *Gigascience*. 2021 Dec 9; 10(12):giab079. <https://doi.org/10.1093/gigascience/giab079> PMID: 34891160; PMCID: PMC8673558.
  25. Johnson TJ, Nolan LK. Pathogenomics of the virulence plasmids of *Escherichia coli*. *Microbiol Mol Biol Rev*. 2009 Dec; 73(4):750–74. <https://doi.org/10.1128/MMBR.00015-09> Erratum in: *Microbiol Mol Biol Rev*. 2010 Sep; 74(3):477–8. PMID: 19946140; PMCID: PMC2786578.
  26. Branger C, Ledda A, Billard-Pomares T, Doublet B, Barbe V, Roche D, et al. Specialization of small non-conjugative plasmids in *Escherichia coli* according to their family types. *Microb Genom*. 2019 Sep; 5(9):e000281. <https://doi.org/10.1099/mgen.0.000281> Epub 2019 Aug 5. PMID: 31389782; PMCID: PMC6807383.
  27. Stephens C, Arismendi T, Wright M, Hartman A, Gonzalez A, Gill M, et al. F Plasmids Are the Major Carriers of Antibiotic Resistance Genes in Human-Associated Commensal *Escherichia coli*. *mSphere*. 2020 Aug 5; 5(4):e00709–20. <https://doi.org/10.1128/mSphere.00709-20> PMID: 32759337; PMCID: PMC7407071.



## Publication 2

### **Hybrid Assembly from 75 *E. coli* Genomes Isolated from French Bovine Food Products between 1995 and 2016**

Jaudou Sandra, Tran Mai-Lan, Vorimore Fabien, Fach Patrick, Delannoy Sabine

Microbiology Resource Announcement 12, e01095-22.

<https://doi.org/10.1128/mra.01095-22>



# Hybrid Assembly from 75 *E. coli* Genomes Isolated from French Bovine Food Products between 1995 and 2016

✉ Sandra Jaudou,<sup>a</sup> Mai-Lan Tran,<sup>ab</sup> Fabien Vorimore,<sup>b</sup> ✉ Patrick Fach,<sup>a,b</sup> ✉ Sabine Delannoy<sup>a,b</sup>

<sup>a</sup>Pathogenic *E. coli* Unit, Laboratory for Food Safety, Anses, Maisons-Alfort, France

<sup>b</sup>IdentyPath Genomics Platform, Laboratory for Food Safety, Anses, Maisons-Alfort, France

**ABSTRACT** Here, we report the complete (or near-complete) genome sequences of 75 *Escherichia coli* isolates, including 71 *stx*-positive *E. coli* isolates, isolated in France between 1995 and 2016 from food of bovine origin. Genomes were assembled using a combination of long- and short-read sequencing.

Shiga toxin-producing *Escherichia coli* (STEC) food contamination can lead to severe human symptoms (1). Here, we assembled 75 *E. coli* genomes, including 31 *eae*-positive STEC and 39 *eae*-negative STEC genomes, isolated from food of bovine origin within a 21-year period in France.

*E. coli* strains were originally isolated on Trypticase soy broth with yeast extract (TSYE) plates overnight at 37°C. Isolates were revived from 20 to 30% glycerol stock on TSYE plates and incubated overnight at 37°C. One colony was cultured overnight at 37°C in brain heart infusion broth (BHI) with rotation. Genomic DNA was extracted from BHI culture (1 mL) using different extraction methods and sequenced using both MiSeq (Illumina, Inc., San Diego, CA, USA) and MinION (Nanopore, Oxford Science Park, Oxford, UK) technologies (2). Default parameters were used for all software unless otherwise specified. Illumina libraries were constructed using Nextera XT kit and sequenced on v2 micro or standard flow cells (2 × 150 bp or 2 × 250 bp). MinION libraries were constructed from unshredded and non-size-selected DNA using the SQK-LSK109 ligation kit and EXP-NBD104 or EXP-NBD114 barcoding kit and sequenced on FLO-MIN106 flow cells using an Mk1B device.

Illumina reads were trimmed and quality checked using the CLC Genomics Workbench v21 (Qiagen). We generated 299,212 to 3,972,904 reads (mean, 1,368,471 ± 720,963.42) with an average read length of 198.08 (±37.99) bases per sample, representing 12 to 174× (mean, 51 ± 22.65) (Table 1). MinION FAST5 files were basecalled and demultiplexed using guppy\_basecaller and guppy\_barcode v3.4.5+fb1fbfb or v5.0.11+2b6dbff using “-compress-fastq” and “-trim-barcodes” parameters (3). Statistics on raw reads were performed using NanoPlot v1.29.0 (4). Between 14,450 and 1,992,878 reads (mean, 239,919 ± 341,570.565) were generated with  $N_{50}$  values ranging from 2,565 bp to 29,351 bp (mean, 16,330 ± 7,056.38). Assemblies were generated using Unicycler v0.4.8 (-min\_fasta\_length 1,000), Flye v2.6 or v2.8.1-b1676 (-genome-size 5m), and Raven v1.2.2 (5–7). Assembly quality and contiguity were assessed using QUAST v5.0.2 (8) and manual inspection of the assemblies (*stx*-phage and large plasmids integrity). Either the Unicycler-generated assembly was conserved or the Flye and/or Raven assemblies were further polished with short reads using racon v1.4.13, medaka v0.12.1 (optional), and pilon v1.24 polishing tools (3 to 5 rounds) (9, 10; <https://github.com/nanoporetech/medaka>). If none of these strategies produced satisfying assemblies, Canu v1.8 or v2.1.1 (genomeSize 5m) or Unicycler based on Flye or Raven assembly (-existing\_long-read\_assembly) was generated (11). Because assemblers have different performances regarding plasmid or phage reconstruction, several assemblies

**Editor** David Rasko, University of Maryland School of Medicine

**Copyright** © 2023 Jaudou et al. This is an open-access article distributed under the terms of the [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/).

Address correspondence to Sabine Delannoy, [Sabine.Delannoy@anses.fr](mailto:Sabine.Delannoy@anses.fr).

The authors declare no conflict of interest.

**Received** 12 March 2022

**Accepted** 16 January 2023

**Published** 1 February 2023



**TABLE 1** Assembly metrics of *Escherichia coli* draft genomes and NCBI accession numbers for each isolate sequenced

NCBI accession no. (ONT, Illumina)	DNA extraction method(s) <sup>1</sup>	Strain	Serotype <sup>2</sup>	stx subtype(s)	ere	Hybrid assembly	MiniON read length ( $M_{50}$ [bp])	No. of contigs	Genome size (bp)	Assembly $M_{50}$ (bp)	GC content (%)	Genome coverage (X)	GenBank accession no.
SRR18191646	Bead based	1429	O76:H19	1c	-	Unicycler	10,334	3	5,316,065	5,183,674	50.64	107	JAQOMI0000000000
SRR18191606						Canu <sup>3</sup>		5	5,528,843	5,270,995	50.64	107	JAOWPQ0000000000
SRR18191539	Salting out	2206	O79:H48	1a, 2a	-	Unicycler	14,002	16	5,258,581	1,387,201	50.73	184	JAQMIH0000000000
SRR18191528						Flye		6	5,257,865	5,088,236	50.71	184	JAOWPF0000000000
SRR18191517	Salting out	06QMA126.1	O153:O178:H19	1a, 2a, 2d	-	Unicycler	17,384	7	5,368,292	4,589,384	50.59	179	JAQOMG0000000000
SRR18191506						Flye <sup>4</sup>		4	5,348,285	5,086,995	50.67	179	JAOWPE0000000000
SRR18191495	Salting out	06QMA140.2	O149:H1	1d	-	Unicycler	7,397	10	5,300,421	4,864,012	50.66	209	JAQOMF0000000000
SRR18191580													
SRR18191569	Bead based, salting out	06QMA181.1a	O91:H14	1a, 2b, 2b	-	Unicycler	28,527	13	5,618,742	5,174,383	50.74	242	JAQOME0000000000
SRR18191644						Raven <sup>5</sup>		3	5,602,836	5,378,985	50.76	242	JAOWNU0000000000
SRR18191633	Salting out	06QMA227.3	O116:H48	2a	-	Unicycler	13,538	5	4,987,186	4,797,877	50.62	230	JAQOMD0000000000
SRR18191566													
SRR18191555	Salting out	07HMPA386	O46/O134:H38	1a, 2a	-	Unicycler	18,809	5	4,943,051	4,790,646	50.57	150	JAQOMC0000000000
SRR18191544						Flye <sup>6</sup>		10	4,965,900	4,764,076	50.63	150	JAOWPD0000000000
SRR18191613	Salting out	07HMPA903	O113:H21	2a, 2a	-	Unicycler	22,124	6	5,144,288	4,966,076	50.73	238	JAQOMB0000000000
SRR18191610													
SRR18191609	Salting out, bead based	07HMPA966	O91:H10	2a	-	Canu <sup>3</sup>	7,440	10	5,575,342	5,185,943	50.72	706	JAOWPF0000000000
SRR18191608						Flye <sup>6</sup>		3	5,331,927	5,145,461	50.76	706	JAOWPC0000000000
SRR18191607	Salting out	09QMA299.3B	O175:H16	2a	-	Unicycler	25,581	7	5,088,106	4,812,392	50.73	207	JAQOMA0000000000
SRR18191605						Raven <sup>5</sup>		4	5,087,513	4,862,173	50.73	207	JAOWNI0000000000
SRR18191604	Salting out	09QMA303.1	O8H19	2g	-	Unicycler	20,236	3	5,132,730	5,022,134	50.76	168	JAOQLZ0000000000
SRR18191603													
SRR18191602	Salting out	09QMA311.2.3	O175:H16	2d	-	Unicycler	23,479	4	5,179,346	5,125,959	50.77	149	JAOQLY0000000000
SRR18191601													
SRR18191600	Salting out	09QMA33.1	O2H27	2a	-	Unicycler	20,366	11	5,535,100	5,226,378	50.61	223	JAOQLX0000000000
SRR18191599						Flye <sup>6</sup>		2	5,382,235	5,226,869	50.66	223	JAOWPB0000000000
SRR18191542	Salting out	09QMA47.1	O6H10	1c	-	Unicycler	17,841	7	5,283,406	4,794,423	50.53	215	JAOQLW0000000000
SRR18191541						Flye <sup>6</sup>		4	5,268,782	4,794,913	50.54	215	JAOWPA0000000000
SRR18191540	Salting out	1890-4	O26:H11	1a	+	Unicycler	22,209	17	5,801,088	4,709,082	50.64	272	JAOQLV0000000000
SRR18191538						Raven <sup>5</sup>		2	5,834,018	5,743,918	50.68	272	JAOWNH0000000000
SRR18191537	Salting out	2011-O26	O26:H11	1a	+	Unicycler	17,173	33	5,891,761	2,948,002	50.45	262	JAOQLU0000000000
SRR18191536						Flye <sup>6</sup>		9	5,906,758	5,674,537	50.51	262	JAOWOZ0000000000
SRR18191535	Salting out	2026-O145	O145:H28	2c	+	Unicycler	23,580	11	5,509,245	4,447,030	50.60	295	JAOQLT0000000000
SRR18191534						Canu <sup>3</sup>		5	5,826,806	5,487,952	50.58	295	JAOWPO0000000000
SRR18191533	Salting out	2039-O26	O26:H11	1a	+	Raven <sup>5</sup>	9,087	3	5,582,748	5,434,794	50.60	295	JAOWNG0000000000
SRR18191532						Raven		3	5,815,566	5,670,366	50.63	151	JAOWNF0000000000
SRR18191531	Salting out	2048-O145	O145:H28	2c	+	Unicycler	18,920	21	5,735,378	3,340,315	50.52	151	JAOQLS0000000000
SRR18191530						Unicycler		15	5,526,494	5,203,884	50.57	250	JAOQLR0000000000
SRR18191529	Salting out	2142-O103	O103:H25	„m	+	Flye <sup>6</sup>	22,773	10	5,628,166	5,402,071	50.59	250	JAOWOY0000000000
SRR18191527						Raven		10	5,661,007	5,368,150	50.51	174	JAOWNE0000000000
SRR18191526	Salting out	2149-O103	O103:H11	1a	+	Unicycler	14,099	27	5,844,869	5,519,265	50.69	153	JAOQLQ0000000000
SRR18191525						Unicycler		12	5,758,941	5,122,111	50.69	153	JAOWNX0000000000

(Continued on next page)



TABLE 1 (Continued)

NCBI accession no. (ONT, Illumina)	DNA extraction method(s) <sup>1</sup>	Strain	Serotype <sup>2</sup>	stx subtype(s)	ere	Hybrid assembly	MinION read length ( $M_{50}$ [bp])	No. of contigs	Genome size (bp)	Assembly $M_{50}$ (bp)	GC content (%)	Genome coverage (x)	GenBank accession no.
SRR18191524, SRR18191523	Salting out	2236-3b	O26:H11	1a	+	Flye <sup>a</sup>	16,957	23	5,997,576	5,003,748	50.63	154	JAOWOX000000000
SRR18191522, SRR18191521	Bead based	2918-O145	O145:H28	2a	+	Raven <sup>b</sup>	6,716	7	5,553,909	3,535,429	50.63	72	JAOWND000000000
SRR18191520, SRR18191519	Salting out	3273-O103	O103:H2	1a	+	Unicycler	20,998	22	5,492,229	1,347,297	50.54	72	JAOQLP000000000
SRR21721513, SRR18191516	Bead based, salting out	3313-O111	O111:H8	1a	+	Flye <sup>b</sup>	4,229	7	5,570,495	5,272,882	50.63	90	JAOWOV000000000
SRR21721512, SRR18191518	Bead based, salting out	3313-O111	O111:H8	1a	+	Unicycler	10,277	17	5,489,297	3,617,166	50.55	90	JAOQL000000000
SRR18191516, SRR18191515	Salting out	3313-O111	O111:H8	1a	+	Unicycler	21,079	19	5,488,508	1,232,028	50.55	398	JAOWNV000000000
SRR18191514, SRR18191513	Salting out	3382-O103	O103:H2	1a	+	Flye <sup>c</sup>	13,920	18	5,719,857	4,395,063	50.65	128	JAOWOU000000000
SRR18191512, SRR18191511	Bead based	3383-O103	O103:H2	1a	+	Unicycler	5,882	3	5,540,746	5,463,884	50.67	128	JAOWNL000000000
SRR18191509, SRR18191508	Bead based	3383-O26b	O26:H11	1a	+	Flye <sup>b</sup>	6,945	6	5,719,857	4,395,063	50.65	671	JAOWNU000000000
SRR18191507, SRR21721511	Bead based, salting out	429-O26	O26:H11	1a	+	Flye <sup>c</sup>	12,758	37	5,542,730	5,454,423	50.65	671	JAOWOT000000000
SRR18191505, SRR18191504	Salting out	429-O26	O26:H11	1a	+	Flye <sup>c</sup>	8,683	5	5,838,831	5,709,619	50.59	778	JAOWNT000000000
SRR18191503, SRR18191502	Salting out	4747-O26	O26:H11	1a	+	Unicycler	12,758	37	5,822,451	5,697,780	50.62	778	JAOWOS000000000
SRR18191499, SRR18191498	Bead based	97HMPL449	O3:H12	1a	-	Unicycler	5,941	6	5,558,675	4,744,103	50.72	259	JAOWOR000000000
SRR18191497, SRR18191496	Salting out	97HMPL473	O3:H12	1a	-	Unicycler	15,334	6	5,728,320	1,975,710	50.56	270	JAOQLN000000000
SRR18191590, SRR18191589	Bead based, salting out	97HMPL650	O10:H9	1c	-	Unicycler	28,759	4	5,805,978	3,820,369	50.73	148	JAOWOQ000000000
SRR18191588, SRR18191587	Bead based, salting out	97HMPL652	O10:H9	1c	-	Raven <sup>b</sup>	22,595	3	5,659,566	1,082,226	50.60	148	JAOWNR000000000
SRR18191585, SRR21721510	Bead based, salting out	97HMPL657	O136:H12	1a	-	Unicycler	21,241	44	5,972,468	5,737,945	50.75	71	JAOWOP000000000
SRR18191583, SRR18191582	Bead based	97HMPL915	O112ac:H19	2c	-	Unicycler	8,149	2	5,664,587	2,049,284	50.56	148	JAOQLM000000000

(Continued on next page)



TABLE 1 (Continued)

NCBI accession no. (ONT, Illumina)	DNA extraction method(s)	Strain	Serotype <sup>a</sup>	stx subtype(s)	ere	Hybrid assembly	MinION read length [N <sub>50</sub> ] (bp)	No. of contigs	Genome size (bp)	Assembly N <sub>50</sub> (bp)	GC content (%)	Genome coverage (X)	GenBank accession no.
SRR18191581, SRR18191579	Bead based	98HMPL324	O174H21	2c	-	Raven <sup>d</sup>	5,263	4	5,214,333	5,042,330	50.76	908	JAOWN8000000000
SRR18191578, SRR18191577	Salting out	98HMPL325	O17/O44/O77/H18	2d	-	Canu	14,916	6	5,045,734	5,090,743	50.71	199	JAOWPK0000000000
SRR18191576, SRR18191575	Bead based	98HMPL475	O110H9	1c	-	Unicycler	4,854	6	5,010,608	4,847,085	50.59	1148	JAOQLG0000000000
SRR18191574, SRR18191573	Salting out	98HMPL479	O91H14	1a, 2b, 2c	-	Unicycler	8,894	13	5,831,137	5,558,854	50.68	152	JAOQLE0000000000
SRR18191572, SRR18191571	Salting out	98HMPL487	O136H12	1a	-	Canu <sup>d</sup>	20,281	2	6,118,277	5,606,901	50.72	152	JAOWP3000000000
SRR18191570, SRR18191568	Salting out	Cri154393	O157H7	2a, 2c	+	Unicycler	21,282	11	5,430,616	5,174,002	50.58	191	JAOQLD0000000000
SRR18191567, SRR18191598	Salting out	Cri154395	O157H7	2c	+	Unicycler	21,865	17	5,744,891	5,126,924	50.47	134	JAOQLC0000000000
SRR18191597, SRR18191596	Salting out	Cri154397	O157H7	2c, 2c	+	Unicycler	20,389	19	5,807,355	5,666,998	50.47	134	JAOWOM0000000000
SRR18191595, SRR18191594	Salting out	Cri169922	O82H8	-	-	Flye <sup>d</sup>	20,244	6	5,780,274	5,273,188	50.49	134	JAOWNA0000000000
SRR18191593, SRR18191592	Salting out	Cri169937	O51H41	-	-	Raven <sup>d</sup>	13,654	4	5,720,557	4,991,949	50.39	129	JAOQLB0000000000
SRR18191591, SRR18191643	Salting out	CriO103	O103H2	-	-	Flye <sup>d</sup>	16,776	2	4,958,903	5,555,517	50.48	129	JAOWOL0000000000
SRR18191642, SRR18191641	Salting out	CriO157	O157H7	2c	+	Unicycler	22,759	4	5,090,231	4,969,612	50.57	144	JAOQLA0000000000
SRR18191640, SRR18191639	Bead based, salting out	E. coli 12-1	O157H7	2a, 2c	+	Unicycler	25,609	10	4,868,876	2,456,854	50.78	225	JAOQKZ0000000000
SRR18191638, SRR18191637	Salting out	EC14	O174H21	2c	-	Flye <sup>d</sup>	20,105	19	4,976,199	4,761,119	50.80	225	JAOWOX0000000000
SRR18191636, SRR18191635	Salting out	ECA135	O22H16	2b, 2c, 2d	-	Canu <sup>d</sup>	29,351	9	4,958,903	4,892,016	50.63	95	JAOQKY0000000000
SRR18191634, SRR18191632	Bead based, salting out	ECA15	O79H48	1a, 2a	-	Unicycler	19,793	4	5,082,783	4,945,157	50.83	219	JAOQKV0000000000
SRR18191631, SRR18191630	Salting out	ECA193	O8H21	1a, 2c	-	Flye <sup>b</sup>	20,554	12	5,155,830	4,911,711	50.83	219	JAOWOH0000000000
SRR18191629, SRR18191628	Bead based	ECA194	O179H8	1a, 2a	-	Canu <sup>d</sup>	15,826	5	5,459,604	5,386,935	50.63	276	JAOWPH0000000000
SRR18191627, SRR18191626	Bead based	ECA279	O174H2	1a, 2d	-	Flye <sup>b</sup>	2,565	1	5,624,392	4,232,969	50.43	278	JAOQKW0000000000
SRR18191625, SRR18191624	Bead based, salting out	ECA329	O91H21	2a, 2d	-	Unicycler	15,119	3	5,642,849	5,231,425	50.52	278	JAOWOJ0000000000
SRR18191625, SRR18191625	Salting out	ECA329	O91H21	2a, 2d	-	Flye <sup>b</sup>	15,119	4	5,909,043	5,747,499	50.44	284	JAOWPI0000000000
SRR18191565, SRR18191564	Salting out	ECA34	O5H9	1a	+	Unicycler	6,208	5	5,807,203	5,689,286	50.50	284	JAOWOI0000000000

(Continued on next page)



**TABLE 1 (Continued)**

NCBI accession no. (ONT, Illumina)	DNA extraction method(s) <sup>a</sup>	Strain	Serotype <sup>b</sup>	stx subtype(s)	eee	Hybrid assembly	MinION read length ( $N_{50}$ [bp])	No. of contigs	Genome size (bp)	Assembly $N_{50}$ (bp)	GC content (%)	Genome coverage (X)	GenBank accession no.
SRR18191563, SRR18191564	Salting out	ECA34	O5:H9	1a	+	Raven <sup>d</sup>	24,947	4	5,414,456	4,637,316	50.68	115	JAOVMY0000000000
SRR18191563, SRR18191562	Bead based, salting out	ECA36	O5:H9	1a, 1a	+	Unicycler	22,228	14	5,418,653	4,566,021	50.66	100	JAOVNN0000000000
SRR18191561, SRR18191560	Bead based, salting out	ECA37	O5:H9	1a, 1a	+	Flye <sup>e</sup>	24,413	11	5,403,523	5,306,942	50.69	112	JAOVOC0000000000
SRR18191559, SRR18191558	Bead based	ECA399	O174:H21	2c	-	Unicycler	3,390	5	5,141,946	3,448,145	50.78	479	JAOQKR0000000000
SRR18191557, SRR18191556	Salting out	ECA400	O6:H10	2d	-	Canu <sup>f</sup>	24,727	2	5,061,078	5,053,258	50.68	200	JAOVFP0000000000
SRR18191554, SRR18191553	Bead based	ECA447	O22:H16	1a, 2b, 2d	-	Unicycler	6,419	5	5,351,044	5,298,558	50.63	121	JAOQKQ0000000000
SRR18191552, SRR18191551	Salting out	ECA89	OgN-RK2:H16	1c	-	Raven <sup>d</sup>	18,224	2	5,345,303	3,529,088	50.64	121	JAOVMM0000000000
SRR18191550, SRR18191549	Salting out	ECA97	O113:H4	2d	-	Flye	17,597	52	5,336,292	4,922,938	50.72	180	JAOVWY0000000000
SRR18191548, SRR18191547	Salting out	HSVR03	O157:H7	2c	+	Unicycler	22,885	3	5,163,781	4,943,802	50.76	180	JAOVNN0000000000
SRR18191546, SRR18191545	Salting out	HSVR04	O157:H7	2c	+	Unicycler	22,936	4	5,262,159	5,257,847	50.55	149	JAOQRP0000000000
SRR18191543, SRR18191622	Bead based	NC672mucus	Ox13:H2	2c	-	Raven <sup>d</sup>	6,632	40	5,843,094	5,515,994	50.43	113	JAOVOB0000000000
SRR18191621, SRR18191620	Salting out	NC809	O41:H7	2c	-	Flye <sup>e</sup>	19,581	3	5,323,626	5,187,741	50.66	381	JAOVWA0000000000
SRR18191619, SRR18191618	Bead based	NV34	O168:H8	2d	-	Unicycler	6,450	8	5,227,833	4,891,608	50.51	230	JAOQKQ0000000000
SRR18191617, SRR18191616	Bead based	NV36	O75:H8	1c, 2b	-	Raven <sup>d</sup>	6,484	6	5,269,996	4,249,978	50.77	402	JAOVWY0000000000
SRR18191615, SRR18191614	Salting out	Slk8430761	O177:H25	2c, 2d	+	Unicycler <sup>g</sup>	20,113	28	5,575,909	5,306,095	50.65	468	JAOQKM0000000000
SRR18191612, SRR18191611	Salting out	Slk8430767	O177:H25	2c, 2d	+	Unicycler <sup>g</sup>	17,583	30	5,577,269	5,301,141	50.65	113	JAOVNM0000000000
	Long-read assembly polished with pilon.					Flye <sup>e</sup>		28	5,639,340	5,231,455	50.58	180	JAOVNO0000000000
	Long-read assembly polished with racon and medaka.					Flye <sup>e</sup>		30	5,658,335	5,248,496	50.58	180	JAOVNZ0000000000
	Long-read assembly polished with racon (2 rounds) and medaka.					Unicycler		12	5,420,630	5,017,089	50.54	150	JAOQKL0000000000
	Long-read assembly polished with racon, medaka, and pilon.					Raven <sup>d</sup>		3	5,458,939	5,276,849	50.57	150	JAOVMM0000000000

<sup>a</sup> Long-read assembly polished with pilon.

<sup>b</sup> Long-read assembly polished with racon and medaka.

<sup>c</sup> Long-read assembly polished with racon (2 rounds) and medaka.

<sup>d</sup> Long-read assembly polished with racon, medaka, and pilon.

<sup>e</sup> Long-read assembly polished with racon, medaka, and pilon, and contigs <1 kb were removed.

<sup>f</sup> Long-read assembly polished with racon (3 rounds), medaka, and pilon, and contigs <1 kb were removed.

<sup>g</sup> Long-read assembly polished with racon (4 rounds), medaka, and pilon, and contigs <1 kb were removed.

<sup>h</sup> Long-read assembly polished with racon (3 rounds) and pilon.

<sup>i</sup> Long-read assembly polished with racon (3 rounds), pilon, and contigs <1 kb were removed.

<sup>j</sup> If two methods were used for DNA extraction, the first one corresponds to the DNA extraction method used for sequencing using Illumina and the second for MinION sequencing.

<sup>k</sup> Serotypes were determined using *in silico* analysis.

<sup>l</sup> 50× subsampling.

<sup>m</sup> - are used to describe bash arguments/options.

were retained for some of the samples. Whenever possible, Unicycler performed the circularization of the replicons (5).

Assemblies from 1 to 52 contigs (mean,  $10.33 \pm 9.73$ ) with a total assembly length between 4,849,707 and 6,118,277 bp (mean,  $5,468,530.42 \pm 292,599.77$  bp) and mean GC content of 50.64% ( $\pm 0.13$ ) were generated.  $N_{50}$  values ranged from 1,082,226 to 5,821,752 bp (mean,  $4,759,454.32 \pm 1,018,105.98$  bp). The main sample characteristics, including serotype, *stx* and *eae* gene presence (determined using abricate [-min-cov 60 -min-id 80]), genome coverage, and accession numbers are reported in Table 1. All genomes were annotated using the National Center for Biotechnology Information (NCBI) Prokaryotic Genome Annotation Pipeline (PGAP) v6.3 at [http://www.ncbi.nlm.nih.gov/genome/annotation\\_prok](http://www.ncbi.nlm.nih.gov/genome/annotation_prok) (12).

**Data availability.** Short and long raw reads were deposited in the NCBI SRA database under BioProject accession number [PRJNA808207](https://www.ncbi.nlm.nih.gov/bioproject/PRJNA808207). Additionally, assembled genomes were deposited in GenBank under accession numbers as referred to in Table 1 within BioProject accession numbers [PRJNA808207](https://www.ncbi.nlm.nih.gov/bioproject/PRJNA808207), [PRJNA885284](https://www.ncbi.nlm.nih.gov/bioproject/PRJNA885284), [PRJNA884285](https://www.ncbi.nlm.nih.gov/bioproject/PRJNA884285), [PRJNA884276](https://www.ncbi.nlm.nih.gov/bioproject/PRJNA884276), and [PRJNA883638](https://www.ncbi.nlm.nih.gov/bioproject/PRJNA883638).

## REFERENCES

- Karmali MA. 2004. Infection by Shiga toxin-producing *Escherichia coli*: an overview. *Mol Biotechnol* 26:117–122. <https://doi.org/10.1385/MB:26:2:117>.
- Jaudou S, Tran ML, Vorimore F, Fach P, Delannoy S. 2022. Evaluation of high molecular weight DNA extraction methods for long-read sequencing of Shiga toxin-producing *Escherichia coli*. *PLoS One* 17:e0270751. <https://doi.org/10.1371/journal.pone.0270751>.
- Wick RR, Judd LM, Holt KE. 2019. Performance of neural network basecalling tools for Oxford Nanopore sequencing. *Genome Biol* 20:129. <https://doi.org/10.1186/s13059-019-1727-y>.
- De Coster W, D'Hert S, Schultz DT, Cruts M, Van Broeckhoven C. 2018. NanoPack: visualizing and processing long-read sequencing data. *Bioinforma Oxf Engl* 34:2666–2669. <https://doi.org/10.1093/bioinformatics/bty149>.
- Wick RR, Judd LM, Gorrie CL, Holt KE. 2017. Unicycler: resolving bacterial genome assemblies from short and long sequencing reads. *PLoS Comput Biol* 13:e1005595. <https://doi.org/10.1371/journal.pcbi.1005595>.
- Kolmogorov M, Bickhart DM, Behsaz B, Gurevich A, Rayko M, Shin SB, Kuhn K, Yuan J, Polevikov E, Smith TPL, Pevzner PA. 2020. metaFlye: scalable long-read metagenome assembly using repeat graphs. *Nat Methods* 17:1103–1110. <https://doi.org/10.1038/s41592-020-00971-x>.
- Vaser R, Šikić M. 2021. Time- and memory-efficient genome assembly with Raven. *Nat Comput Sci* 1:332–336. <https://doi.org/10.1038/s43588-021-00073-4>.
- Gurevich A, Saveliev V, Vyahhi N, Tesler G. 2013. QUAST: quality assessment tool for genome assemblies. *Bioinforma Oxf Engl* 29:1072–1075. <https://doi.org/10.1093/bioinformatics/btt086>.
- Vaser R, Sović I, Nagarajan N, Šikić M. 2017. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res* 27:737–746. <https://doi.org/10.1101/gr.214270.116>.
- Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA, Zeng Q, Wortman J, Young SK, Earl AM. 2014. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* 9:e112963. <https://doi.org/10.1371/journal.pone.0112963>.
- Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. 2017. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res* 27:722–736. <https://doi.org/10.1101/gr.215087.116>.
- Tatusova T, DiCuccio M, Badretdin A, Chetverin V, Nawrocki EP, Zaslavsky L, Lomsadze A, Pruitt KD, Borodovsky M, Ostell J. 2016. NCBI prokaryotic genome annotation pipeline. *Nucleic Acids Res* 44:6614–6624. <https://doi.org/10.1093/nar/gkw569>.

## **Additional experiment 1: Assessment of the selected extraction method performances on raw milk.**

### Context:

In this project, we aimed at characterizing STEC directly from raw milk using long-read sequencing. In the previous paper (Publication 1), we have compared three different extraction methods on pure STEC cultures. We aimed at obtaining high quantities of HMW DNA with correct purity ratios to be sequenced using the MinION platform. The salting-out method was the best in our hands to obtain DNA that is suitable for obtaining a complete STEC assembly using long-read sequencing. The bead-based extraction method appeared as a valid alternative, also producing sequencing-grade HMW gDNA. However, the performances of the DNA extraction methods from complex matrices such as raw milk might be different.

The first important parameter tested was the DNA concentration obtained since ONT sequencing requires large quantities of DNA. Indeed, it has been shown that the DNA yield recovered from raw milk was different depending on DNA extraction methods (Cremonesi *et al.*, 2021). The second parameter evaluated was the purity of DNA extracts. Raw milk is a complex matrix characterized by a high content in proteins, fats and nutrients (Porcellato *et al.*, 2021). These molecules complicate DNA extraction as they usually remain present in the DNA extract (Quigley *et al.*, 2012). Their presence might also obstruct the pores present on MinION flow cell preventing the DNA from passing through and resulting in the generation of fewer data. Lastly, DNA integrity was assessed as the recovery of long DNA fragments is crucial for STEC genome assembly and necessitates a gentle DNA extraction method to avoid DNA shearing.

Thus, it is important to ensure that the selected method is efficient enough to recover the maximum DNA possible from raw milk (Parente *et al.*, 2020; Porcellato *et al.*, 2021) while removing contaminants to ensure the generation of sufficient data and extracting long DNA fragments. In this study, the performance of two gDNA extraction kits on raw milk were compared. The Lucigen MasterPure kit, as previously determined to be more suitable in our case to extract STEC HMW gDNA from pure cultures, was compared to the Zymo Quick-DNA HMW MagBead kit used by colleagues on complex matrices to obtain HMW gDNA. The comparison was made on DNA yield and concentrations obtained from raw milk, its purity and integrity.

In this analysis, we used DNA extracted from fresh raw milk first enriched at 37°C in BPW and then artificially contaminated with 10<sup>4</sup> CFU of STEC per mL of milk. Additionally, frozen raw milk was enriched in the same conditions, without and with artificial contamination (10<sup>5</sup> CFU of STEC per mL of milk).

## Materials and methods:

### 1. STEC cells for artificial contamination

The 4712-O26 strain (*stx1a*, *eae*-positive, O26:H11), described in publication 2, was used here for artificial contamination of enriched raw milk. STEC cells were revived from glycerol stock, plated onto TSYe plates and incubated at 37°C overnight. STEC pure cultures were prepared by incubating one colony from the plate in 10 mL brain heart infusion (BHI) at 37°C overnight, with agitation.

### 2. Raw milk enrichment

Fresh raw milk was collected from a French farm in the surroundings of Paris and 1 mL was enriched in 1:10 BPW at 37°C with agitation for 18h.

A portion of the enriched raw milk was artificially contaminated with  $10^4$  CFU.mL<sup>-1</sup> of STEC while the remaining milk was stored at -20°C for further experiments. Artificial contamination of enriched raw milk was done using one mL of  $10^4$  CFU.mL<sup>-1</sup> of STEC obtained by serial dilution (1:10) of the STEC pure culture. DNA from frozen raw milk was also extracted after enrichment in the same conditions as previously mentioned, with and without artificial contamination using a  $10^5$  CFU.mL<sup>-1</sup> dilution from STEC pure culture.

### 3. DNA extraction

The presence of fatty acids is problematic and to reduce their presence we centrifuged the enriched milk prior to DNA extraction, removed the milk fat layer, and washed with PBS twice. DNA was extracted in triplicates from 1 mL of enriched raw milk with or without artificial contamination using both Lucigen (Complete DNA and RNA extraction kit, Masterpure) and Zymo (Quick-DNA HMW MagBead extraction kit, Zymo Research) kits. DNA extraction, quantification and qualification was performed as explained in publication 1. DNA extracts were quantified using the Qubit 3.0 Fluorometer and the Qubit dsDNA BR (broad range) Assay-kit (Thermo Fisher Scientific). Their purity was assessed using the A260/A280 and A260/A230 ratios determined using a Nanodrop UV-Vis Spectrophotometer (Thermo Fisher Scientific). The integrity of extracted DNA was assessed using a TapeStation system and Genomic screentape (Agilent) analyzed using the TapeStation Analysis software v4.4.1.

### 4. Visualization and statistical analysis

Boxplots were generated using R v 4.1.2. Statistical tests were also performed using R and the alpha error set to 5%. The Kruskal-Wallis test was used to compare the DNA concentrations between the two kits and the Wilcoxon test to test whether the values of purity ratio obtained differ significantly from the optimal value of 1.8 for A260/A280 ratio and of 2 for the A260/A230 ratio. Wilcoxon test was first performed two sided (alternative= two.sided) and when significantly different results was obtained, the test was performed to know if the values obtained were lower (alternative= less) or higher (alternative= greater) than the reference value for each ratio and DNA extraction kit.

## Results:

1. The salting-out extraction method recovers higher quantities of DNA with lower impurities from milk matrix

Table 6 presents DNA yield and concentration obtained as well as purity and integrity of the extracts. In fresh raw milk artificially contaminated with  $10^4$  CFU.mL<sup>-1</sup> of STEC, the concentration of recovered DNA using the salting-out method was twice as much as extracted using the bead-based kit, and ten times higher in frozen raw milk artificially contaminated with  $10^5$  CFU.mL<sup>-1</sup> of STEC (Table 6). In addition, in milk samples without artificial contamination, more DNA was extracted using the salting-out method (Table 6). Overall, higher DNA amounts were recovered using the salting-out extraction method compared to the bead-based, although the difference is not statistically different when including all samples ( $\chi^2= 3.1875$ , p-value=0.0742, Fig. 13).

Table 6 : Quantification and purity ratios of DNA extracted from fresh or frozen raw milk after enrichment with or without artificial contamination using STEC.

STEC spike-in level (CFU.mL <sup>-1</sup> )	DNA concentration (ng.μL <sup>-1</sup> )	DNA yield (ng)*	A260/280 ratio	A260/230 ratio	DIN	DNA extraction kit* <sup>2</sup>	Raw milk state* <sup>3</sup>
10 <sup>4</sup>	6.02	301	1.7	0.9	9	Zymo	fresh
10 <sup>4</sup>	3.52	176	1.65	0.64	8.8	Zymo	fresh
10 <sup>4</sup>	5.52	276	1.57	0.79	6.1	Zymo	fresh
10 <sup>4</sup>	13.1	458,5	2.08	1.77	9	Lucigen	fresh
10 <sup>4</sup>	12.9	451,5	2.14	2.21	8.7	Lucigen	fresh
10 <sup>4</sup>	10.3	360,5	2.13	2.25	8.1	Lucigen	fresh
10 <sup>5</sup>	193.6	6776	1.89	1.79	6.7	Lucigen	frozen
10 <sup>5</sup>	170.4	5964	1.88	1.8	6.7	Lucigen	frozen
10 <sup>5</sup>	196.8	6888	1.88	1.77	6.8	Lucigen	frozen
10 <sup>5</sup>	17.1	855	1.52	0.83	7.2	Zymo	frozen
10 <sup>5</sup>	13	650	1.52	0.66	7.3	Zymo	frozen
10 <sup>5</sup>	11.3	565	1.52	0.79	7.2	Zymo	frozen
no	31.8	1590	1.47	0.85	7.1	Zymo	frozen
no	27	1350	1.65	1.04	7.1	Zymo	frozen
no	132.4	4634	1.92	1.8	6.8	Lucigen	frozen
no	164.6	5761	1.97	1.99	6.9	Lucigen	frozen

\* DNA yield (ng) was calculated as follow: DNA concentration (ng.μL<sup>-1</sup>) x elution volume (μL), with elution volume of 50 μL for Zymo and 35 μL for Lucigen.

\*<sup>2</sup>Zymo is the bead-based extraction kit and Lucigen represents the salting-out extraction method

\*<sup>3</sup>Fresh raw milk (conserved at +4°C), frozen raw milk upon arrival at -20°C

The optimal purity ratios A260/A280 and A260/A230, based on the light absorbance at 260, 280 and 230nm wavelengths are 1.8 and 2, respectively. The A260/A280 ratio gives information regarding the presence of proteins or phenol compounds and A260/A230 on contaminants such as EDTA, guanidine, etc. The results show that the value obtained for the A260/A280 ratio was significantly lower than the expected value of 1.8 using the bead-based kit (mean=1.575 ± 0.082, v=0, p-value=0.006838) and higher using the salting-out extraction kit (mean=1.986 ± 0.113, v=36, p-value=0.007074). However, the A260/A230 ratio showed optimal values for salting-out method (mean=1.9225 ± 0.203, v=12.5, p-value=0.4822) whereas it was significantly lower for DNA extracted with the bead-based approach (mean=0.8125 ± 0.128, v=0, p-value=0.007074). In conclusion, the salting-out method allowed the extraction of DNA with purity ratios close to the desired values (Table 6 and Fig. 13).

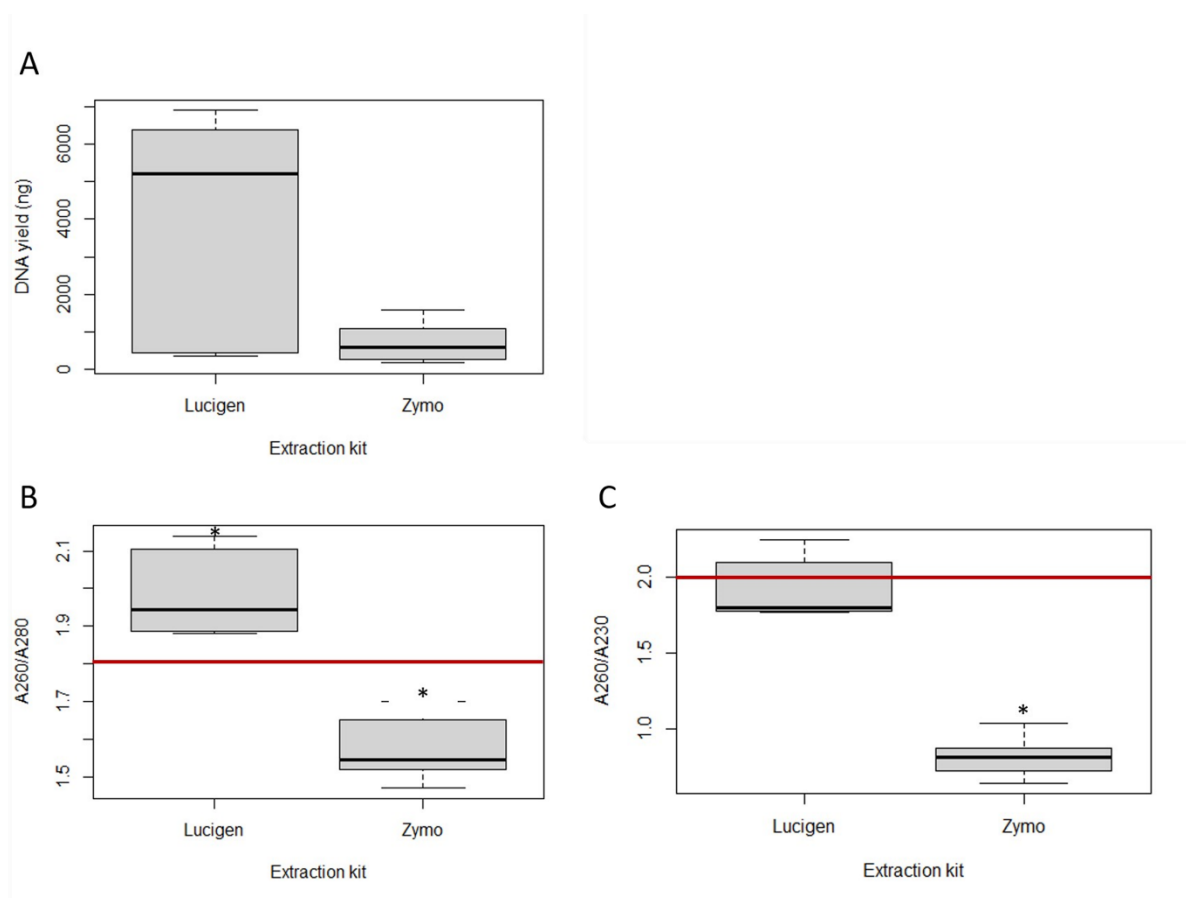


Figure 13 : Comparison of DNA yield (A), A260/A280 (B) and A260/A230 (C) values obtained depending on the extraction method used.

The Lucigen kit was used for salting-out method and Zymo kit for bead-based. The line represents the optimal value for each purity ratio. Significantly different result is labelled with one star (\*) for alpha error fixed to 5%  $* < 0.05$ .

From the results, it is clear that the salting-out is the most promising method to obtain high quantities of DNA with low impurities from raw milk.

## 2. Similar performance of the two methods and matrix-conservation effect on DNA integrity

The DNA integrity is determined based on gel electrophoresis results analyzed with the TapeStation Analysis Software v4.4.1 of the Agilent system. It is represented using a number between 1 and 9. The higher the DNA Integrity Number (DIN), the longer the DNA fragments are. The results showed similar DIN values obtained after extraction using either the salting-out or the bead-based extraction method, though slightly higher using the bead-based extraction kit (Fig. 14 and Fig. 15).

Enriched raw milk has been frozen at  $-20^{\circ}\text{C}$  prior to contamination with  $10^5$  CFU of STEC.mL<sup>-1</sup>. Although, no effect was observed on concentration and purity of the DNA, the DIN value from frozen samples was lower, with a mean value of 6.98 whereas the mean value was approximately 8.28 in fresh raw milk (Fig. 14 and Fig. 15, Table 6).

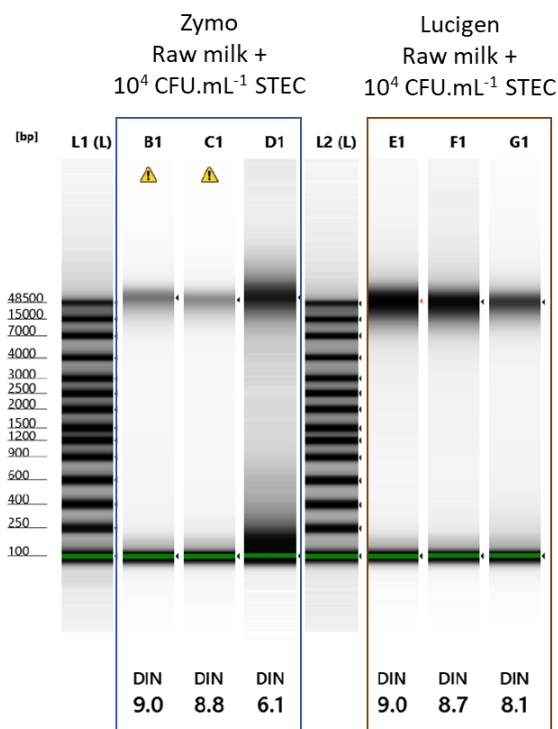


Figure 14 : Gel electrophoresis of DNA extracted from fresh raw milk artificially contaminated with  $10^4$  CFU.mL<sup>-1</sup> of STEC after enrichment at  $37^{\circ}\text{C}$  in BPW.

The electrophoresis was performed on a TapeStation system from Agilent and the Genomic Screentape. The Zymo kit refers to the bead-based method and the Lucigen kit to the salting-out method. The DNA ladder is on lanes L1 and L2. The DNA Integrity Number (DIN) is represented at the bottom and assesses the integrity of the extracted DNA. Caution sign means that the DNA concentration is outside of the recommended range.

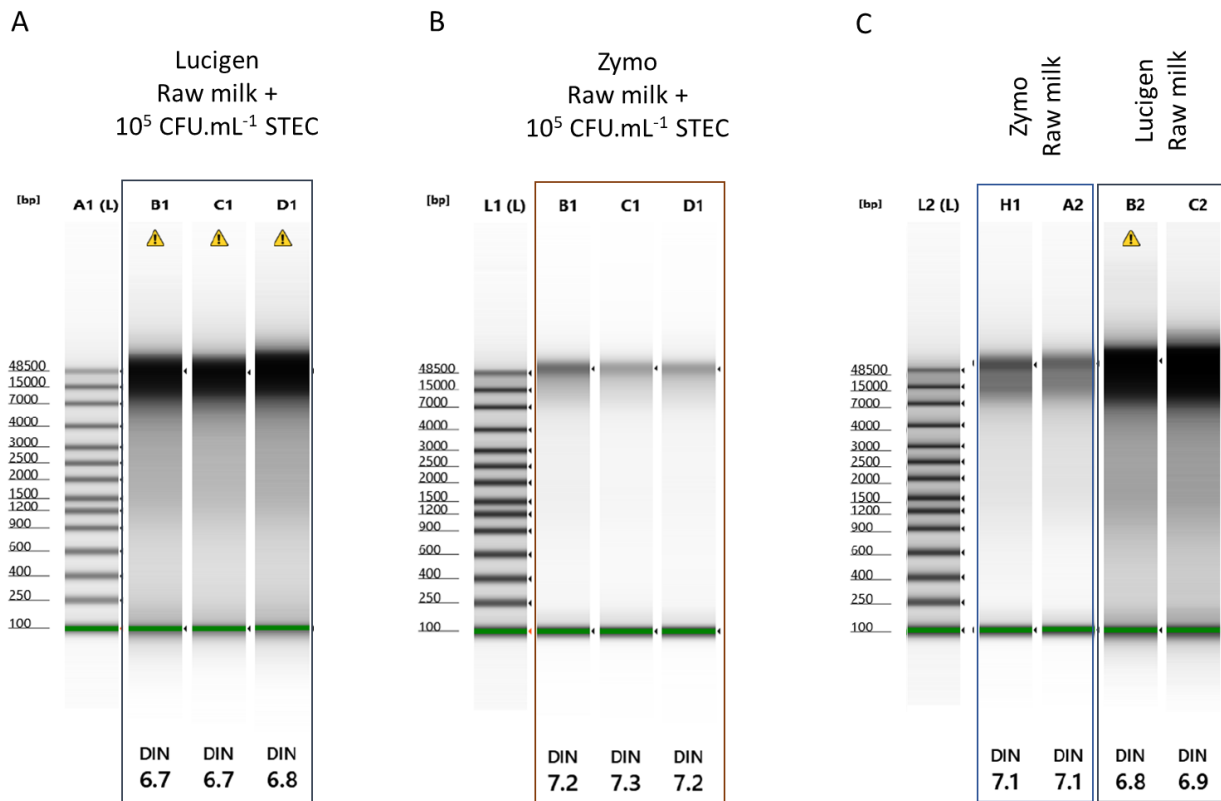


Figure 15 : Gel electrophoresis of DNA extracted from frozen enriched raw milk artificially contaminated with  $10^5$  CFU.mL<sup>-1</sup> of STEC post-enrichment at 37°C in BPW using Lucigen kit (A) and Zymo kit (B), as well as from frozen enriched raw milk without artificial contamination using both Lucigen and Zymo kits (C).

The electrophoresis was performed on a TapeStation system from Agilent and the Genomic Screentape. The Zymo kit refers to the bead-based method and the Lucigen kit to the salting-out method. The DNA ladder is on lanes A1, L1 and L2. The DNA Integrity Number (DIN) is represented at the bottom and assesses the integrity of the extracted DNA. Caution sign means that the DNA concentration is outside of the recommended range.

### Discussion:

Raw milk is a complex matrix in which proteins, fats and nutrients are highly present. The presence of these macromolecules can affect the performances of DNA extraction procedures (Quigley *et al.*, 2012). The extraction of bacterial DNA from raw milk is already a challenge since the DNA of the host can represent more than 90% of the extracted DNA (Ahmadi *et al.*, 2022; Siebert *et al.*, 2021).

The aim of this study was to check whether the MasterPure (Lucigen) DNA extraction that was shown to be the best on STEC pure culture allowed the extraction of HMW DNA from raw milk suitable for long-read sequencing. The first parameters used for comparison were DNA yield and concentration since the DNA input required for long-read sequencing is important.



ONT recommends using 1  $\mu\text{g}$  of DNA in 50  $\mu\text{L}$  for library preparation, but previous experiments prompted us to start with 2  $\mu\text{g}$  of gDNA (a minimum of 40  $\text{ng}\cdot\mu\text{L}^{-1}$ ). From our previous work on STEC pure cultures, higher DNA yield was obtained using salting-out compared to solid phase and bead-based extraction methods. In this analysis, the quantity of DNA recovered was higher using the Masterpure (Lucigen) kit than the Zymo. The yields obtained with the Zymo kit are clearly insufficient to perform ONT sequencing. Because of low DNA quantities obtained after bead-based extraction, the DNA extracts were not sequenced and the performances on STEC assembly not available.

The presence of contaminants impedes the proper sequencing of DNA. Purity ratios A260/A280 and A260/A230 of approximately 1.8 and from 2 are desired to consider the extracted DNA to be “pure”. The A260/A280 ratio gives insights on the presence of proteins or phenol compounds absorbing at 280nm. Proteins can block the nanopores of the MinION flow cells. In our previous work, the salting-out method was shown to be the best to remove proteins compared to the bead-based method. The A260/A230 ratio informs on the presence of contaminants. The presence of residuals such as guanidine used in column-based kits, phenol compound or ethanol usually leads to lower A260/A230 values (Pachchigar *et al.*, 2016). Here again, the DNA extracted with the salting-out approach (Lucigen, Masterpure) was purer than DNA extracted with the bead-based approach (Zymo).

Apart from DNA yield and the purity ratios, for which the salting-out performed well compared to bead-based, HMW gDNA was recovered and the integrity of DNA recovered was similar for both methods. The DNA integrity, as evaluated using the DIN was slightly better using the bead-based method. However, the segregation of DNA was better due to the lower concentration. It is noteworthy that the DIN value was lower when working with frozen (enriched) raw milk than on fresh (enriched) raw milk. It appears that the freezing process breaks DNA into smaller fragments, as a degradation profile could be observed on the gel. The degradation effect of the freezing process has already been described (Dahn *et al.*, 2022).

From this analysis, I determined that the salting-out method (Lucigen, MasterPure) allowed the recovery of relatively pure HMW DNA in large quantities from raw milk samples. Hence, this method was conserved for subsequent steps.

### Main results:

- The solid-phase method did not allow the extraction of sufficient DNA for MinION sequencing and DNA was degraded.
- Bead-based methods allow the extraction of DNA suitable for MinION sequencing. However, contaminants were present, as determined with the measured A260/A230 purity ratio.
- Both bead-based and salting-out methods allowed the extraction of long DNA fragments from both STEC pure culture and raw milk and generated long reads from STEC pure culture, although DNA extracted using salting-out method produced longer reads after MinION sequencing
- The salting-out method showed the best performances with extraction of high quantities of relatively pure DNA extracts from both STEC pure culture and raw milk.
- No difference of assembly contiguity from STEC pure cultures using bead-based or salting-out extracted DNA was observed using Canu and Raven assemblers.
- Flye assemblies were more contiguous using DNA extracted with the salting-out method compared to bead-based.
- DNA degradation was observed from frozen raw milk.
- Combination of short- and long-read sequencing allowed the reconstruction of complete *E. coli* genomes n=75 from food origin.

### Main conclusions:

- Bead-based and salting-out extraction methods allowed the recovery of HMW DNA matching requirements for MinION sequencing from STEC pure cultures.
- The salting-out method was the most appropriate to extract DNA suitable for MinION sequencing and showed similar performance on raw milk and on STEC pure cultures.
- Both bead-based and salting-out methods generated long reads enabling complete and contiguous STEC assemblies.
- Generation of complete STEC genomes contributed to filling the gap of publicly available STEC genomes from food origin.
- The processing of fresh raw milk is preferred since the freezing process degrades DNA.

### Perspectives:

- Characterize the genetic diversity of STEC from food origin

## Chapter5-2: Characterizing STEC from raw milk using long-read metagenomics

After DNA extraction was optimized for MinION sequencing, we aimed to develop a method to identify and characterize an *eae*-positive STEC strain in raw milk samples using long-read metagenomics and an assembly-based approach. Using *in silico* analysis, I have determined that an enrichment step is necessary to obtain enough *E. coli* data from raw milk. I compared different enrichment conditions to recover the highest amount of data from artificially contaminated raw milk using an *eae*-positive STEC. The best condition for characterizing the inoculated *eae*-positive STEC using long-read metagenomics and an assembly-based approach was 37°C in acriflavine-supplemented BPW. Using the selected enrichment condition and salting-out DNA extraction method, I assessed that successful characterization was possible from an inoculation level of 5 CFU.mL<sup>-1</sup> provided that it can reach 10<sup>8</sup> copies.mL<sup>-1</sup> post-enrichment, using the developed method (Publication 3).

The presence of multiple *E. coli* strains that cannot be distinguished using assembly is a challenge for STEC characterization. Thus, we aimed at identifying the limiting ratio of STEC to commensal *E. coli* strains that would impede STEC characterization using the developed method. I artificially contaminated pasteurized milk using two *E. coli* strains (a commensal strain and a STEC strain) extracted from cow raw milk at different inoculation levels (Publication 4). With this study, I could refine the limit of our method and showed that the *eae*-positive STEC should be 10-times in excess compared to other *E. coli* when multiple strains are present in the sample for a successful characterization.

Overall, I have shown that the use of long-read metagenomics from raw milk samples was efficient to characterize *eae*-positive STEC. However, the STEC strain has to grow to 10<sup>8</sup> copies.mL<sup>-1</sup> post-enrichment and if additional strains are present, it has to be in excess of at least 10-times.

## Publication 3

### **A step forward for Shiga toxin-producing *Escherichia coli* identification and characterization in raw milk using long-read metagenomics**

Jaudou Sandra, Deneke Carlus, Tran Mai-Lan, Schuh Elisabeth, Goehler André,  
Vorimore Fabien, Malorny Burkhard, Fach Patrick, Grützke Josephine,  
Delannoy Sabine

Microbial Genomics. <https://doi.org/10.1099/mgen.0.000911>

# A step forward for Shiga toxin-producing *Escherichia coli* identification and characterization in raw milk using long-read metagenomics

Sandra Jaudou<sup>1,2,\*</sup>, Carlus Deneke<sup>2</sup>, Mai-Lan Tran<sup>1,3</sup>, Elisabeth Schuh<sup>4</sup>, André Goehler<sup>4</sup>, Fabien Vorimore<sup>3</sup>, Burkhard Malorny<sup>2</sup>, Patrick Fach<sup>1,3</sup>, Josephine Grützke<sup>†</sup> and Sabine Delannoy<sup>1,3,\*</sup>†

## Abstract

Shiga toxin-producing *Escherichia coli* (STEC) are a cause of severe human illness and are frequently associated with haemolytic uraemic syndrome (HUS) in children. It remains difficult to identify virulence factors for STEC that absolutely predict the potential to cause human disease. In addition to the Shiga-toxin (*stx* genes), many additional factors have been reported, such as intimin (*eae* gene), which is clearly an aggravating factor for developing HUS. Current STEC detection methods classically rely on real-time PCR (qPCR) to detect the presence of the key virulence markers (*stx* and *eae*). Although qPCR gives an insight into the presence of these virulence markers, it is not appropriate for confirming their presence in the same strain. Therefore, isolation steps are necessary to confirm STEC viability and characterize STEC genomes. While STEC isolation is laborious and time-consuming, metagenomics has the potential to accelerate the STEC characterization process in an isolation-free manner. Recently, short-read sequencing metagenomics have been applied for this purpose, but assembly quality and contiguity suffer from the high proportion of mobile genetic elements occurring in STEC strains. To circumvent this problem, we used long-read sequencing metagenomics for identifying *eae*-positive STEC strains using raw cow's milk as a causative matrix for STEC food-borne outbreaks. By comparing enrichment conditions, optimizing library preparation for MinION sequencing and generating an easy-to-use STEC characterization pipeline, the direct identification of an *eae*-positive STEC strain was successful after enrichment of artificially contaminated raw cow's milk samples at a contamination level as low as 5 c.f.u. ml<sup>-1</sup>. Our newly developed method combines optimized enrichment conditions of STEC in raw milk in combination with a complete STEC analysis pipeline from long-read sequencing metagenomics data. This study shows the potential of the innovative methodology for characterizing STEC strains from complex matrices. Further developments will nonetheless be necessary for this method to be applied in STEC surveillance.

Received 18 May 2022; Accepted 12 October 2022; Published 24 November 2022

**Author affiliations:** <sup>1</sup>COLiPATH Unit, Laboratory for Food Safety, ANSES, Maisons-Alfort, France; <sup>2</sup>National Study Center for Sequencing, Department of Biological Safety, German Federal Institute for Risk Assessment, Berlin, Germany; <sup>3</sup>Genomics Platform IdentityPath, Laboratory for Food Safety, ANSES, Maisons-Alfort, France; <sup>4</sup>National Reference Laboratory for *Escherichia coli* including VTEC, Department of Biological Safety, German Federal Institute for Risk Assessment, Berlin, Germany.

**\*Correspondence:** Sandra Jaudou, sandra.jaudou.ext@anses.fr; Sabine Delannoy, sabine.delannoy@anses.fr

**Keywords:** genome assembly; isolation-independent identification; long-read sequencing; metagenomics; raw milk; Shiga toxin-producing *Escherichia coli* (STEC).

**Abbreviations:** BPW, buffered peptone water; HUS, haemolytic uraemic syndrome; ISO, International Organization for Standardization; MGE, mobile genetic element; MLST, multilocus sequence type; qdPCR, quantitative digital PCR; qPCR, real-time PCR; ST, sequence type; STEC, Shiga toxin-producing *Escherichia coli*; TS, technical specification.

Raw sequence data were deposited in the National Center for Biotechnology Information SRA (Sequence Read Archive) under BioSample accession numbers: SRR19090764, SRR19090765, SRR19090766, SRR19090767, SRR19090768, SRR19090769, SRR19090770, SRR19090771, SRR19090772, SRR19090773, SRR19090774, SRR19090775, SRR19090776, SRR19090777, SRR19090778, SRR19090779, SRR19090780, SRR19090781, SRR19090782, SRR19090783, SRR19090784, SRR19090785, SRR19090786, SRR19090787, SRR19090788, SRR19090789, SRR19090790, SRR19090791, SRR19090792, SRR19090793, SRR19090794.

†These authors contributed equally to this work

**Data statement:** All supporting data, code and protocols have been provided within the article or through supplementary data files. Four supplementary tables are available with the online version of this article.

000911 © 2022 The Authors



This is an open-access article distributed under the terms of the Creative Commons Attribution NonCommercial License.

## DATA SUMMARY

All basecalled and demultiplexed fastq files were deposited in the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA) under BioProject PRJNA835223. The code for the STECmetadetector pipeline is freely available from GitLab ([https://gitlab.com/bfr\\_bioinformatics/STECmetadetector](https://gitlab.com/bfr_bioinformatics/STECmetadetector)).

## INTRODUCTION

Shiga toxin-producing *Escherichia coli* (STEC) are a subset of diarrheagenic *E. coli*, which are known to cause various symptoms ranging from watery diarrhoea to haemorrhagic colitis (HC) or haemolytic uraemic syndrome (HUS) [1, 2]. Their main virulence factor is located on bacteriophages, named *stx*-phages, which encode the Shiga toxin, and are integrated into the *E. coli* chromosome [3]. STEC pathogenicity assessment is challenged by their capacity to acquire mobile genetic elements (MGEs) carrying various virulence factors.

Although the main STEC virulence property is the capacity to produce the Shiga toxin (Stx), their ability to acquire MGEs encoding virulence factors may increase the risk of severe human symptoms. Some STEC frequently causing HUS additionally carry a pathogenicity island named locus of enterocyte effacement (LEE). One of the most notable protein encoded by the LEE is intimin (*eae* gene), which is responsible for bacterial adhesion to the intestinal cells [4]. Hybrid STEC strains harbouring virulence factors of other *E. coli* pathotypes and causing HUS are increasingly described. In 2011, an O104:H4 strain expressing aggregative factors from enteroaggregative *E. coli* (EAEC) along with an *stx* gene was responsible for a HUS outbreak in both Germany and France [5]. More recently, an O80:H2 cross-pathotype harbouring extra-intestinal pathogenic *E. coli* (ExPEC) virulence markers and an *stx* gene was described in France [6]. STEC O80 now represents the second main cause of HUS in Europe [7–9].

Though STEC are mostly found in cattle gut microbiota, contaminated food products are a major source of human infections, especially bovine meat and dairy products [9, 10]. In Europe, most severe human cases caused by food consumption are generally associated with *eae*-positive STEC strains of the top five O-groups: O26, O157, O103, O145 and O111 [9]. STEC strains in food and animal feed are screened by real-time PCR on DNA extracted from an enriched food portion [11]. However, PCR performed on food enrichment samples is not sufficient to ensure that *stx*, *eae* and the O-group associated genes belong to the same strain. Consequently, isolation of the STEC from all *stx*-positive samples is attempted for further characterization.

STEC isolation steps are laborious, time-consuming and frequently unsuccessful due, in part, to the lack of selective media. When isolation of STEC strains fails, no strain characterization can be done using conventional approaches. We were interested in exploring metagenomics to facilitate STEC strain characterization from food samples by providing crucial information on virulence factors and typing, in an isolation-independent way. However, the high content of MGEs containing repetitive sequences in STEC complicates their assembly using short-read sequencing, which can result in more than 200 contigs [12]. Recently, long-read sequencing was developed to resolve highly repetitive sequences [13]. Different long-read sequencing technologies exist, among which the most commonly used were commercialized by Pacific Biosciences (PacBio) and Oxford Nanopore Technology (ONT) [13, 14]. The MinION platform, released in 2014 by Oxford Nanopore Technology, known for its portability and its ability to generate very long reads, has been increasingly tested for food-borne pathogen detection [15–20].

Genomic studies using MinION sequencing were conducted on isolated STEC strains and the generated long-reads were proven to be sufficient to identify crucial MGEs responsible for STEC pathogenicity, including plasmids, phages, virulence genes and antimicrobial-resistance genes, which may be lost when assembling with short reads [21]. While whole-genome sequencing still requires a bacterial isolation step for characterizing pure isolates, metagenomics based on sequencing the total DNA from the specimen –including micro-organisms and the matrix DNA – may provide the means to identify and characterize STEC in an isolation-independent way. So far, major studies using long-read metagenomics for identifying STEC were performed on artificial mixtures or artificially contaminated food portions. In 2018, Peritz and colleagues demonstrated the potential of MinION for resolving STEC O-groups from artificial DNA mixtures of five STEC isolates [22]. Metagenomics analysis of artificially contaminated beef, spinach, pasteurized milk or wastewater using STEC strains has demonstrated the utility of MinION sequencing for STEC detection or identification and its limits [17–19, 23].

The objective of this study was to develop a long-read sequencing-based method to characterize STEC without the need to perform a strain isolation step. Here, we used an *eae*-positive O26 STEC strain, since they are frequently associated with the most severe human cases in Europe. We chose an assembly-based approach to detect *stx* and *eae* genes on the same contig, as well as the O-group and H-type associated genes. Because the *stx*-prophage may integrate into the genome at different sites, the *stx* and *eae* genes might be distantly located by up to 2.1 Mb [24]. Due to the high fragmentation of the draft genomes resulting from short-read sequencing methods, these are not appropriate. In contrast, the reads obtained from MinION sequencing that can be up to 4 Mb in length have the potential of resolving repeated sequences and thereby of improving assembly contiguity. Long-read assemblers try to span repetitive regions by using either an overlap-layout-consensus (OLC) or a graph-based approach; thus, closing gaps and generating long contigs that can be as long as the chromosome itself [25, 26].

### Impact Statement

Shiga toxin-producing *Escherichia coli* (STEC) are food-borne pathogens that may cause severe human illnesses such as haemolytic uraemic syndrome. Current detection methods rely on time-consuming and laborious isolation steps for STEC characterization. Because short-read sequencing is not sufficient to span repeated sequences widely present in STEC genomes, we assessed the potential of long-read sequencing to identify and characterize STEC strains in an isolation-independent way. Here, we artificially contaminated raw cow's milk and showed that the inoculated *eae*-positive STEC O26 strain was identifiable in raw milk enriched at 37 °C in acriflavine-supplemented buffered peptone water, from an STEC concentration of 5 c.f.u. (ml raw milk)<sup>-1</sup>. For ease of analysis, we developed an automated STEC characterization pipeline. We present a complete workflow to identify and characterize STEC directly from raw cow's milk using long-read metagenomics and an assembly-based method. This study demonstrates the ability of long-read metagenomics to efficiently characterize *E. coli* strains directly from a complex matrix.

So far, no studies have been conducted to identify and characterize *eae*-positive STEC strains in raw milk using long-read metagenomics. We used raw milk as an example of a typical food product that is known to cause severe STEC outbreaks [27]. Our method aimed to assemble the *eae*-positive STEC genome in bovine raw milk using long-read metagenomics. Here, we artificially contaminated STEC-negative raw cow's milk using an *eae*-positive STEC strain of serotype O26:H11 and optimized its growth conditions. To facilitate long-read metagenomics data analysis, we developed a complete pipeline for *eae*-positive STEC characterization. In this study, we present an entire and accessible workflow for STEC identification and characterization when detected in raw cow's milk using an assembly-based method from long-read metagenomics data. It shows that metagenomics approaches using long-read technologies are becoming an easy and time-efficient alternative to classical STEC characterization methods, and could be suitable to identify and eliminate hazards for the consumer at an early stage in the near future.

## METHODS

### Selected strains

For artificial contamination of raw cow's milk, *eae*-positive STEC (*stx*-positive and *eae*-positive) strains from the ANSES (French Agency for Food, Environmental and Occupational Health and Safety) collection were used. Two different *eae*-positive STEC strains of serotype O26:H11 isolated from bovine raw milk were selected: 4712-O26 (*eae*, *stx1a*) and 6423-O26 (*eae*, *stx1a*) (Table 1). Each strain has been previously characterized using both Illumina and Oxford Nanopore Technology sequencing methods [26]. Sequencing data for 4712-O26 and 6423-O26 were deposited in GenBank under accession numbers SRR18191504, SRR18191503, SRR18191500 and SRR18191499, within BioProject PRJNA808207.

### *In silico* determination of the minimum genome coverage required for *eae*-positive STEC strain identification

Oxford Nanopore Technology sequencing data generated by Jaudou and colleagues using the MinION instrument from pure culture of 10 *eae*-positive STEC strains (Table S1, available with the online version of this article) were selected depending on their N50 value (9087–22 759 bp) [26]. Data were sub-sampled to various genome coverages (3×, 5×, 10×, 15×, 20×, 25×, 30×, 35×, 40×, 50×, 60×, 70×) using the random sub-sampling tool rasusa v0.6.0 [28] (<https://github.com/mbhall88/rasusa>), and further assembled using the default mode of three different long-read assemblers: Flye v2.8.1-b1676 (<https://github.com/fenderglass/Flye>), Raven v1.2.2 and v1.7.0 (<https://github.com/lbcb-sci/raven>), and Canu v2.1.1 (<https://github.com/marbl/canu>) [29–31]. The quality of the resulting assemblies was assessed with QUAST v5.0.2 [32] (<http://quast.sourceforge.net/>). Virulence genes were detected using GENIAL v1.0 (an abricate v0.8.7 [11] wrapper; <https://github.com/p-barbet/GENIAL>) and the VFDB (Virulence Factor Database) version 2020-05-29 (<https://github.com/tseemann/abricate/tree/master/db/vfdb>). Results from GENIAL were used to determine whether both *stx* and *eae* genes were on the same contig. Only when *stx* and *eae* virulence genes were present on a same contig, the coverage for the assembly was considered sufficient for *eae*-positive STEC identification.

**Table 1.** Characteristics of the inoculated STEC O26 strains for artificial contamination of raw milk

Isolate	Serotype	ST	<i>stx</i> subtype	<i>Stx</i> -phage insertion site	<i>eae</i> subtype	<i>stx/eae</i> distance	Isolation year	Isolation source
4712-O26	O26:H11	21	1a	<i>wrbA</i>	Beta1	1.9 Mb	2014	Bovine raw milk
6423-O26	O26:H11	21	1a	<i>wrbA</i>	Beta1	1.9 Mb	2015	Bovine raw milk

### Raw milk screening for *eae*-positive STEC

Five fresh raw milk samples were collected from dairy cows kept on the experimental farm of the German Federal Institute for Risk Assessment (BfR, Berlin, Germany) or bought from a French farm located in the surroundings of Paris, and stored at +4°C for a maximum of 1 day until use. For enrichment, 1 ml of raw milk was diluted 1:10 in buffered peptone water (BPW; bioMérieux or Mast Group) and incubated overnight either at 37 or 41.5°C with agitation. After 18 to 20 h of enrichment, DNA from each enriched milk sample was extracted using InstaGene™ matrix (Bio-Rad), following the manufacturer's instructions. Enriched cultures (1 ml) were centrifuged at a minimum of 10 000 r.p.m. for 3 min and the raw milk fat layer was detached using either a sterile cotton swab or a pipet tip. The pellets were washed twice with 1 ml PBS before DNA extraction. Real-time PCR was performed on each sample for the detection of *stx1*, *stx2*, *eae* and *wecA* or *cdgR*, as described in Methods (in the Real-time PCR analysis of enriched raw milk section). The genetic markers *wecA* or *cdgR* were used as generic *E. coli* gene targets. STEC-negative raw milk samples were further used for artificial contamination.

### Artificial contamination of STEC-negative raw milk and enrichment conditions

Strains used for artificial contamination of raw milk were revived from 20–30% (v/v) glycerol stock on TSYe (Tryptone soy yeast extract agar) or TSA (tryptic soy agar) plates (bioMérieux or Mast Group) and incubated overnight at 37°C. One colony was transferred to brain heart infusion (BHI) broth (bioMérieux or Mast Group) and incubated overnight at 37°C with rotation. Approximate bacterial concentrations of pure overnight cultures and 1:10 serial dilutions were determined by optical density measurement at 600 nm ( $OD_{600}$ ). Actual spike-in levels were calculated by triplicate plate counting on TSYe or TSA plates. Raw milk (1 ml) was initially inoculated with respective 1 ml STEC O26 cultures of  $10^3$  c.f.u. ml<sup>-1</sup> ( $1.25$ – $1.78 \times 10^3$  c.f.u. ml<sup>-1</sup>),  $10^2$  c.f.u. ml<sup>-1</sup> ( $1.6$ – $2.63 \times 10^2$  c.f.u. ml<sup>-1</sup>) or  $10^1$  c.f.u. ml<sup>-1</sup> ( $1.33 \times 10^1$  c.f.u. ml<sup>-1</sup>); aiming for estimated levels of  $0.5 \times 10^3$ ,  $0.5 \times 10^2$  and  $0.5 \times 10^1$  c.f.u. ml<sup>-1</sup> of STEC O26 in the raw milk. Each artificially contaminated sample was inoculated with a single STEC O26 strain. Artificially contaminated raw milk (2 ml) was incubated in BPW (9 ml) with or without acriflavine supplementation (final acriflavine concentration of 12 mg l<sup>-1</sup>) with agitation [International Organization for Standardization/technical specification (ISO/TS) 13136:2012] [11] and incubated for 18 to 20 h either at 37 or 41.5°C. A negative control (raw milk not artificially contaminated with STEC) was similarly processed. All experiments were performed in triplicate.

### DNA extraction from enriched raw milk for MinION sequencing

The enriched raw milk (1 ml) was centrifuged and the raw milk fat layer was removed before the pellet was washed with 1× PBS as described above for raw milk screening. DNA extraction and purification was performed using the MasterPure complete DNA and RNA extraction and purification kit (Lucigen), following the manufacturer's instructions and including an RNase A treatment for 30 min (Qiagen; 100 mg ml<sup>-1</sup>) [26]. Genomic DNA was quantified using a Qubit 3.0 fluorometer and the broad range kit (Agilent), and quality control of the DNA extracts was assessed using the Nanodrop 1.0 or 2.0 spectrophotometer (Agilent).

### Real-time PCR analysis of enriched raw milk

Real-time PCR was performed on extracted DNA using 20 µl of the following mixture for each sample: PerfeCTa 1× qPCR ToughMix low ROX (QuantaBio), probes and primers at a final concentration of 0.3 µM (except for *ntb2* with a final concentration of 0.2 µM), completed with nuclease-free water. Either 5 µl DNA extracted using the InstaGene™ protocol or 2 µl DNA extracted using the Lucigen protocol were added. Strain EDL933 (*stx1a*, *stx2a*, *eae*) was used as positive control. Dsb or *ntb2* plasmid [33–36] was used as an inhibition control. The prepared samples were amplified using the CFX96 real-time detection system (BioRad) with the following program: 10 min at 95°C (5°C s<sup>-1</sup>); followed by 39 cycles of 15 s at 95°C (2°C s<sup>-1</sup>) and 60 s at 60°C (2°C s<sup>-1</sup>); and a final step of 30 s at 40°C (5°C s<sup>-1</sup>). Primers and probes are described in Table S2. All probes were labelled with 6-hexachlorofluorescein (HEX) or 6-carboxyfluorescein (FAM) and BHQ1 (black hole quencher) (Eurofins).

### Quantitative digital PCR (qdPCR) to quantify STEC O26 in enriched raw milk

qdPCR analysis was performed on the DNA extracted with the Lucigen protocol from enriched artificially contaminated raw milk samples. The Fluidigm BioMark system and the qdPCR 37k IFC digital array microfluidic chips were used according to the manufacturer's instructions (Fluidigm) and as previously described [37]. Reactions were performed in 6 µl per sample with 3 µl PerfeCTa 2× qPCR ToughMix low ROX (QuantaBio), 0.6 µl 20× GE sample loading reagent (Fluidigm), 0.3 µl 20× primer stock (containing 18 µM primers forward and reverse, and 4 µM probe), and 1.8 µl DNA, completed by nuclease-free water. A non-template control was included for each target. Real-time digital PCR was run using 6-carboxyfluorescein (FAM)- and black hole quencher (BHQ)-labelled probes. Amplification was performed with the following thermal profile: 2 min at 50°C, 10 min at 95°C, followed by 40 cycles of denaturation at 95°C for 15 s and annealing at 60°C for 60 s (rate of 2°C s<sup>-1</sup>). Each fluorescent signal was acquired after the annealing step. Data were analysed using the Fluidigm digital PCR analysis software v4.1.2. The *wecA* genetic marker was used to quantify total *E. coli* and *wzx*<sub>O26</sub> to quantify the artificially contaminated strain.



## MinION sequencing of raw milk

Libraries for MinION sequencing were prepared from the Lucigen DNA extracts using the LSK-SQK109 ligation kit and the EXP-NBD104 and/or EXP-NBD114 barcoding kits (Oxford Nanopore Technology) following the manufacturer's recommendations, except that 2 µg genomic DNA was used as the starting material. Six to seven metagenomics samples were multiplexed and sequenced with a R.9.4.1 FLO-MIN106 flow cell (Oxford Nanopore Technology) on the Mk1B or Mk1C MinION sequencer (Oxford Nanopore Technology). Different amounts of pooled DNA libraries were loaded on each flow cell ranging from 96 to 530 ng.

## Sequencing data analysis

The sequencing runs were acquired without base-calling, and fast5 files output was selected. Raw fast5 files were base-called using guppy basecaller v4.4.2+ and demultiplexed using guppy barcoder v4.4.2+, with or without the `-require_barcodes_both_ends` option (Table S3).

A snakemake [38] pipeline, STECMetadetector ([https://gitlab.com/bfr\\_bioinformatics/STECmetadetector](https://gitlab.com/bfr_bioinformatics/STECmetadetector)), was developed in this study for user-friendly characterization of STEC strains from long-read metagenomics data. The main module of the STECMetadetector pipeline aims to assemble and characterize *E. coli* reads. First, barcodes and adapters are trimmed using porechop v0.2.4 (<https://github.com/rwick/Porechop>). Short (< 1 kb) and low-quality reads (q-score <7) are filtered using NanoFilt v2.8.0 [39]. Nanoplot v1.39 is used to assess read quality from both raw and filtered reads [39]. Filtered reads are classified using the kraken2 classification tool v2.1.2 [40] and the Minikraken DB 8 GB (October 18 2017) database. The kraken2report file is converted into mpa format using krakentools v1.2 [40] (<https://ccb.jhu.edu/software/krakentools/>). Reads taxonomically assigned to *E. coli* are extracted using krakentools v1.2, mapped to detect *stx*, *eae* and O-group associated genes using minimap2 v2.24 [41] and Center for Genomic Epidemiology (CGE) databases [42], and assembled with Flye assembler v2.9-b1768 [29]. The metagenome mode of Flye may be selected with the extra parameter `--meta`. The generated assembly is characterized using abricate v1.0.1 (<https://github.com/tseemann/abricate>) to detect serotype and the presence of virulence genes. Multilocus sequence type (MLST) is identified using MLST v2.19 (<https://github.com/tseemann/mlst>) and the *E. coli* 1 scheme [43]. The completeness, contamination and strain heterogeneity are screened using Checkm v1.1.3 [44]. Strainberry v1.1 [45] attempts to separate strains if multiple *E. coli* strains are suspected. A post-Strainberry module called Strainberry-parsing may be run to characterize the predicted *E. coli* strains. Several utility R scripts for parsing and summarizing results are integrated in the pipeline and use R packages mentioned in Table S4 [46].

To use the STECMetadetector pipeline, a sample sheet is required and can be provided by the user or created as mentioned in GitLab. STECMetadetector may be called by either using the Python *STECmetadetector.py* wrapper or by using the snakemake command line and editing the *config.yaml* file. The STECMetadetector pipeline can be also run on an HPC (high-performance computing) system with the `--cluster` option. All the pipeline's software and database dependencies can be easily and reproducibly installed using conda [47]. Further information and detailed documentation can be found in the pipeline's repository.

For the particular *eae*-positive STEC identification, *stx* and *eae* genes positions were extracted from the virulence file generated by abricate. Kraken2 output files (*kraken2-classification.tsv* and *kraken2-output.tsv*) from the pipeline were used to generate barplots and dotplots. For ease of presentation, only the most abundant genera have been included in the barplots. Flye assembly statistics were represented using dotplot graphs. All plots were generated using ggplot2 v3.3.5 [48] and reshape2 v1.4.4 [49] R packages on R v4.0.3.

All versions of the software used here are reported in Table S4.

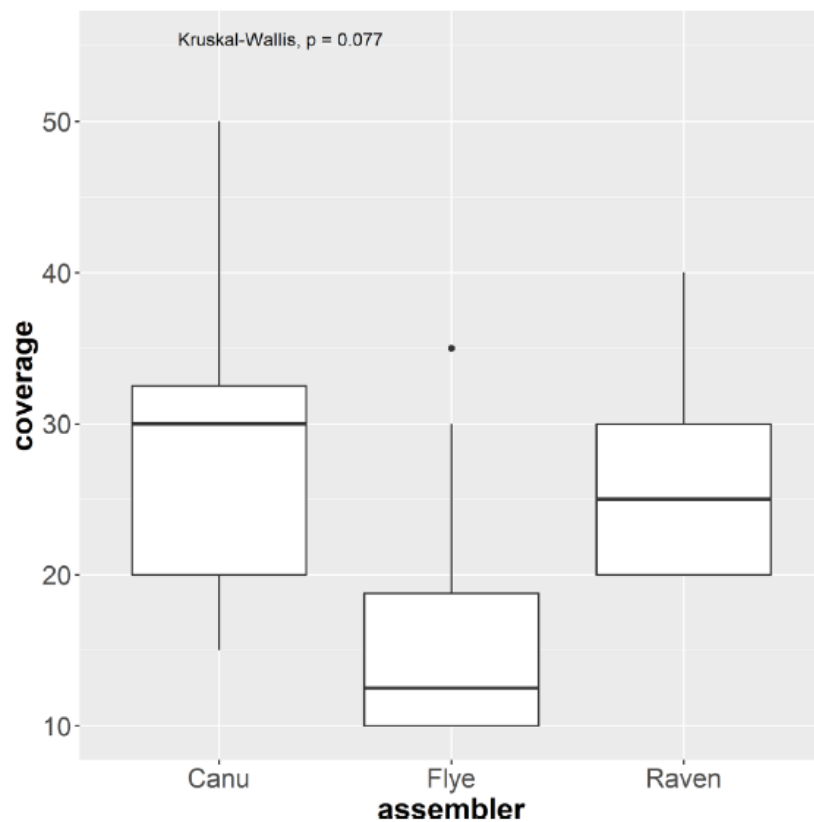
## Statistical analysis

The Kruskal–Wallis non-parametric test and its post hoc Mann–Whitney test were used for statistical analysis, with alpha error being fixed at 5%. The resulting *P* values were represented on an R-generated boxplot using the *ggpubr* R package v0.4.0 (<https://rpkgs.datanovia.com/ggpubr/>) and the `--stat_compare_means` function on R v4.0.3 and R Studio v1.3.1093.

## RESULTS

### Identification of the minimum coverage required to identify *eae*-positive STEC using MinION sequencing

We first aimed to determine the minimum coverage required to assemble the *stx* and *eae* genes on the same contig. For this purpose, we generated *in silico* assemblies using sub-sampled MinION datasets sequenced from *eae*-positive STEC isolates ( $n=10$ ) using three different long-read assemblers. While the difference between the assemblers' performance was not significant (Kruskal–Wallis,  $P=0.077$ ), Flye performed best in identifying *eae*-positive STEC strains from a genome coverage of 10× to 35× in eight strains (Fig. 1). Canu assembled seven strains with genome coverage varying from 15× to 50× and, lastly, Raven required a minimum coverage of 20× for *eae*-positive STEC strain identification, but did not assemble *stx* and *eae* genes on the same contig for five strains (Fig. 1). For one strain, all assemblers failed to assemble both virulence genes on a single contig, even at a high



**Fig. 1.** Boxplot representing the minimal coverage required to assemble *eae*-positive STEC. The minimal coverage at which *stx* and *eae* were co-localized on the same contig is represented depending on the long-read assembler used on subsampled MinION data of 10 *eae*-positive STEC strains. The minimal coverage required for *eae*-positive STEC identification was confirmed when *stx* and *eae* genes were found on the same contig in all of the following assemblies with higher coverages (one exception was accepted). Flye was the most efficient to assemble *eae*-positive STEC with a genome coverage up to 35 $\times$ , although the difference with Canu and Raven assemblers was not significant (Kruskal-Wallis test,  $P$  value  $>0.05$ ).

coverage. Based on the results, the minimum coverage to ensure *eae*-positive STEC identification from genome assembly for most strains was 35 $\times$  with Flye assembler, in the default mode.

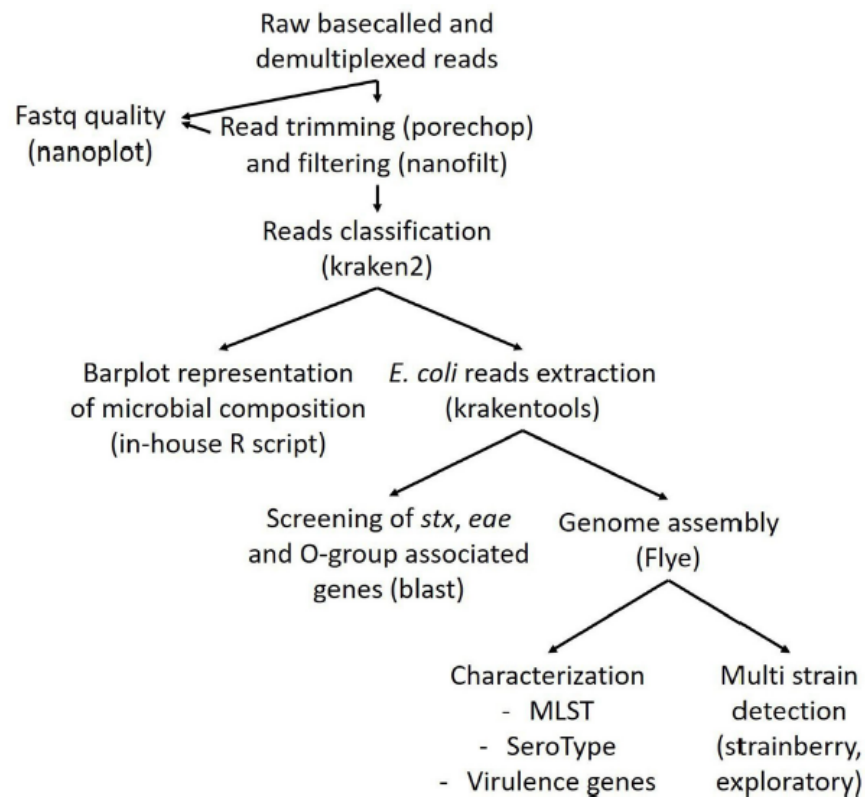
### Generation of a snake make pipeline for STEC characterization from long-read metagenomics data

#### STECmetadetector pipeline

We generated a complete and open-source snakemake pipeline ([https://gitlab.com/bfr\\_bioinformatics/STECmetadetector](https://gitlab.com/bfr_bioinformatics/STECmetadetector)), which orchestrates appropriate state-of-the-art tools to consistently and reproducibly characterize STEC strains from metagenomics data using an assembly-based approach (Fig. 2) [38]. All relevant information from the different modules is aggregated into a complete HTML report file. All steps of this pipeline are presented in Methods (in the Sequencing data analysis section) and represented in Fig. 2.

#### Pipeline results on real metagenomics data

The STECmetadetector pipeline was used to select pertinent data for STEC characterization in artificially contaminated raw milk samples. Short and low-quality reads (length  $<1$  kb and q-score  $<7$ ) were filtered, which resulted into the retention of 8.49–57.5% of all reads but in more than 42.73% of the generated data (in bases) (mean=73.85 $\pm$ 11.88%,  $n=31$ ). Most filtered reads were taxonomically assigned (88.73–99.77%,  $n=30$ ), except for one sample (Milk5\_37+A) with a very low DNA concentration extracted (below limit of quantification; classified reads=51.55%,  $n=1$ ). Between 0.03 and 46.24% and between 13.9 and 80.62% of classified reads were assigned as *E. coli* in uncontaminated and artificially contaminated raw milk samples, respectively (mean=5.73 and 64.46%,  $\pm$ 14.33 and 18.1 %, respectively;  $n=10+21$ ). Flye assembly on *E. coli* extracted reads from artificially contaminated samples showed a mean total assembly length of 5.92 Mb (4.96–6.4 $\pm$ 0.24Mb,  $n=21$ ). The *wzx*<sub>026</sub>, *stx* and *eae* genes were co-localized on the same contig in 19 Flye assemblies ( $n=21$ ; Table S3).



**Fig. 2.** Presentation of the different steps performed by the STECmetadetector pipeline for *eae*-positive STEC identification from long-read metagenomics data.

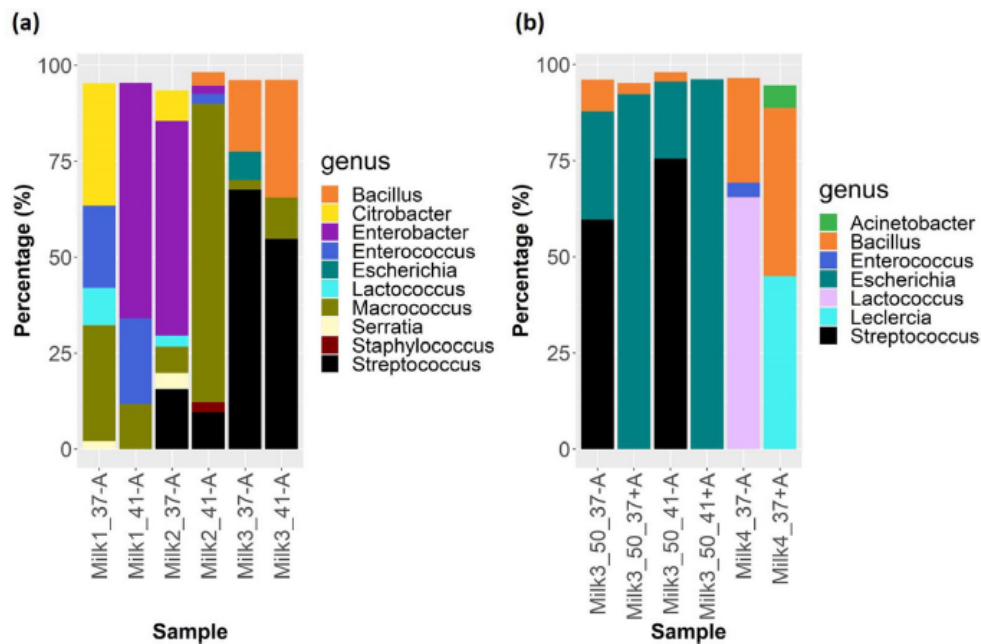
### Determination of the optimal enrichment conditions for multiplying *eae*-positive STEC in raw cow's milk

Based on *in silico* results presented in Results (in the section Identification of the minimum coverage required to identify *eae*-positive STEC using MinION sequencing), we established that an enrichment step was necessary to achieve the minimal coverage required for STEC assembly. We then determined the best enrichment conditions for *eae*-positive STEC multiplication and identification in raw milk using long-read metagenomics. STEC-negative raw milk was artificially contaminated with  $0.5 \times 10^3$  c.f.u. ml<sup>-1</sup> (Milk1) and  $0.5 \times 10^2$  c.f.u. ml<sup>-1</sup> (Milk2 and Milk3) of one STEC strain per ml of raw milk. Four different conditions were compared with samples being incubated at either 37 or 41.5 °C, with or without acriflavine supplementation (final concentration of 12 mg l<sup>-1</sup>). Acriflavine is an antibiotic targeting Gram-positive bacteria recommended for STEC enrichment in milk and dairy products by the ISO/TS 13136:2012.

While the enrichment temperature had no impact on the inoculated strain growth (Table 2), we observed a changing composition of the background flora in non-contaminated raw milk samples, yet without a clear pattern (Fig. 3a). Similarly, acriflavine did not negatively affect the growth of both 4712-O26 and 6423-O26 *eae*-positive STEC strains (Table 2). However, it strictly reduced the proportion of Gram-positive bacteria such as *Lactococcus* spp. (Milk4) and *Streptococcus* spp. (Milk3) that were dominant

**Table 2.** Concentration of STEC O26 in raw milk (c.f.u. ml<sup>-1</sup>) and quantification in raw milk enriched in four different conditions by qdPCR (copies ml<sup>-1</sup>)

Concentration of STEC O26 in raw milk before enrichment (determined by plating on agar plates) [c.f.u. ml <sup>-1</sup> ]	Enrichment conditions	Quantification of STEC O26 in enriched milk by qdPCR [copies ml <sup>-1</sup> ]
$1.06 \times 10^2 - 0.625 \times 10^3$ (n=3)	37 °C with acriflavine	$2.59 \times 10^4 - 1.45 \times 10^6$
$1.00 \times 10^2 - 0.72 \times 10^3$ (n=3)	37 °C without acriflavine	$6.77 \times 10^3 - 1.24 \times 10^6$
$1.06 \times 10^2 - 0.625 \times 10^3$ (n=3)	41.5 °C with acriflavine	$4.47 \times 10^4 - 6.82 \times 10^4$
$1.00 \times 10^2 - 0.72 \times 10^3$ (n=3)	41.5 °C without acriflavine	$6.77 \times 10^4 - 9.95 \times 10^6$



**Fig. 3.** Barplots representing the enrichment temperature and acriflavine supplementation impact on the background flora. MinION-sequencing data were analysed using the STECmetadetector pipeline. (a) Barplot representing the most abundant genera (>2%) in three natural raw milk samples enriched at 37 °C (\_37) or at 41.5 °C (\_41) without acriflavine (-A) supplementation. (b) Barplot representing the most abundant genera (>2%) in two enriched milk samples. One was inoculated at the concentration of  $0.5 \times 10^2$  c.f.u. STEC strain  $\text{ml}^{-1}$  in raw milk (Milk3\_50) and enriched at 37 °C (\_37) or 41.5 °C (\_41) with (+A) or without (-A) acriflavine supplementation. The second milk sample was enriched at 37 °C with (Milk4\_37+A) or without (Milk4\_37-A) acriflavine supplementation.

in those samples, from 65.59 and 59.68% to less than 1 and 2% of the filtered reads, respectively. By preventing Gram-positive bacterial growth, the growth of Gram-negative bacteria was favoured and particularly that of the inoculated STEC strain (Fig. 3b). One notable example was in an artificially contaminated raw milk containing a naturally high proportion of *Streptococcus* spp., which was reduced from 59.68% to less than 2% after acriflavine supplementation, while we observed a drastic increase of *E. coli* reads detected from less than 24.76% to more than 87.57%, at both 37 and 41.5 °C (Fig. 3b). However, the impact of acriflavine on *Bacillus* was different depending on the milk samples (Fig. 3b). Overall, these results clearly showed that acriflavine supplementation may improve the sensitivity of the method.

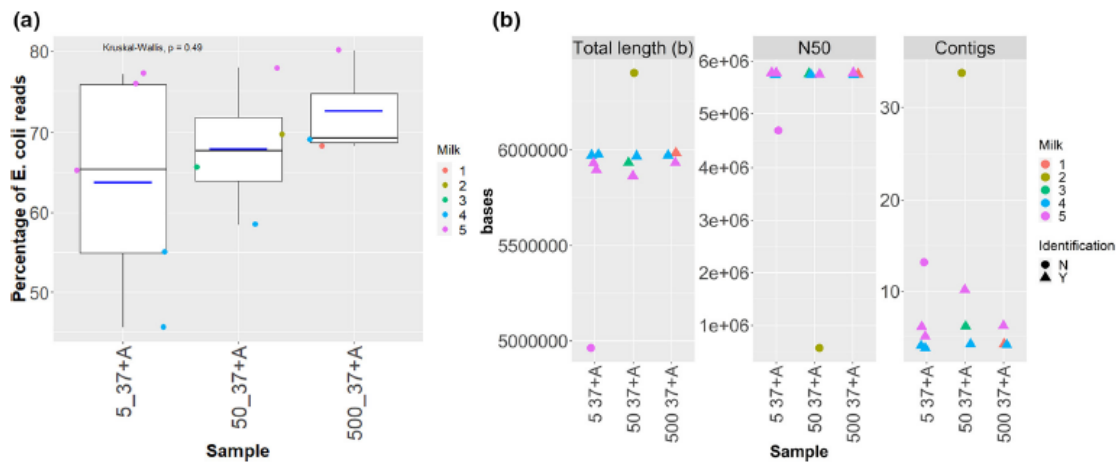
### Determination of the identification limit for an *ee*-positive STEC strain in enriched raw milk using long-read metagenomics

Using the selected enrichment conditions, STEC-negative raw milk was artificially contaminated with one STEC strain at concentrations from  $0.5 \times 10^1$  to  $0.5 \times 10^3$  c.f.u.  $\text{ml}^{-1}$  in raw milk and enriched for 18–20 h. Table 3 shows the contamination levels

**Table 3.** Concentration of STEC O26 in raw milk before and after enrichment

Quantification of *wecA* and *wzx*<sub>O26</sub> genetic markers in enriched raw milk (37 °C with 12 mg acriflavine  $\text{l}^{-1}$ ) was determined by qdPCR (copies  $\text{ml}^{-1}$  in enriched milk). qdPCR results for each non-contaminated raw milk sample enriched at 37 °C without acriflavine are also presented.

Calculated concentration of STEC O26 inoculum (determined by plating on agar plates) (c.f.u. $\text{ml}^{-1}$ )	Estimated concentration of STEC O26 in raw milk before enrichment (determined by $\text{OD}_{600}$ ) (c.f.u. $\text{ml}^{-1}$ )	Quantification of total <i>E. coli</i> ( <i>wecA</i> ) in enriched milk by qdPCR (copies $\text{ml}^{-1}$ )	Quantification of STEC O26 ( <i>wzx</i> <sub>O26</sub> ) in enriched milk by qdPCR (copies $\text{ml}^{-1}$ )
$1.25 \times 10^2 - 1.78 \times 10^2$	500 (n=4)	$8.65 \times 10^8 - 2.37 \times 10^{10}$	$1.16 \times 10^9 - 2.62 \times 10^{10}$
$1.67 \times 10^2 - 2.63 \times 10^2$	50 (n=5)	$4.94 \times 10^8 - 5.13 \times 10^9$	$2.59 \times 10^8 - 2.92 \times 10^9$
$1.33 \times 10^1$	5 (n=8)	$8.59 \times 10^8 - 1.92 \times 10^{10}$	$2.41 \times 10^7 - 1.04 \times 10^{10}$
0	0 (n=5)	Without acriflavine: $0 - 1.37 \times 10^9$	Without acriflavine: 0



**Fig. 4.** *E. coli* reads proportion and *E. coli* assembly metrics obtained from artificially contaminated raw milk. Raw milk samples were artificially contaminated with STEC at the concentrations of  $0.5 \times 10^3$  c.f.u. ml<sup>-1</sup> ( $n=3$ ; 500),  $0.5 \times 10^2$  c.f.u. ml<sup>-1</sup> ( $n=4$ , 50) or  $0.5 \times 10^1$  c.f.u. ml<sup>-1</sup> ( $n=5$ , 5) of raw milk and enriched at 37 °C in the presence of acriflavine (37+A). (a) Boxplot representing the percentage of extracted *E. coli* reads. (b) Dotplot representing assembly metrics obtained from Flye assemblies on extracted *E. coli* reads. A triangle represents the identification of *stx* and *eae* genes on the same contig, whereas their identification on two different contigs is represented using a circle. Milk batches are represented using different colours.

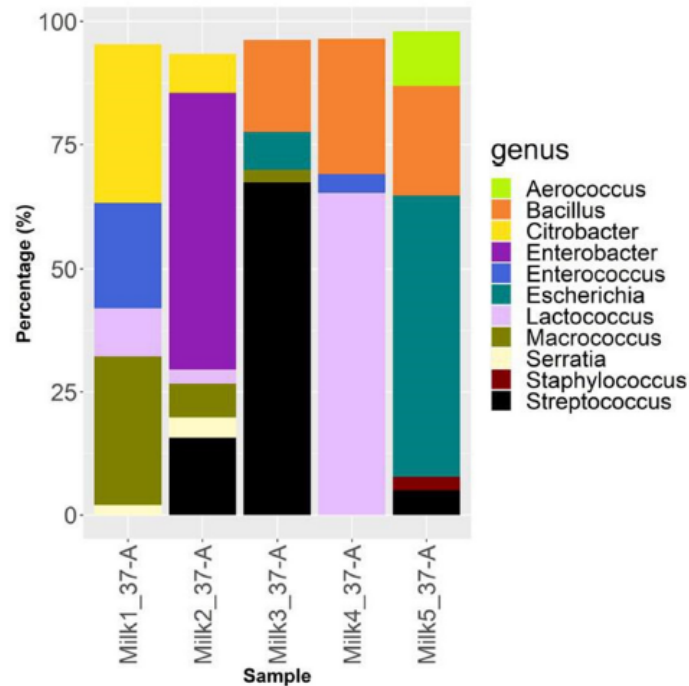
of STEC O26 (c.f.u. ml<sup>-1</sup> in milk) as estimated by optical density and calculated by plating on agar plates. Table 3 also reports quantification of the total *E. coli* (*wecA*) and the spiked STEC O26 (*wzx<sub>O26</sub>*) in enriched milk by qdPCR (copies ml<sup>-1</sup> in enriched milk). DNA extracted from enriched milk samples was sequenced and processed with the STECmetadetector pipeline. The results showed, as expected, that the mean proportion of *E. coli* reads increased with the STEC strain concentration in raw milk, with a percentage of *E. coli* reads of 63.75% ( $\pm 13.56$ ,  $n=5$ ), 67.93% ( $\pm 8.1$ ,  $n=4$ ) and 72.53% ( $\pm 6.54$ ,  $n=3$ ) of filtered reads, for initial STEC concentrations of  $0.5 \times 10^1$ ,  $0.5 \times 10^2$  and  $0.5 \times 10^3$  c.f.u. ml<sup>-1</sup> in raw milk, respectively (Fig. 4a).

Fig. 4b represents assembly metrics from artificially contaminated samples enriched at 37 °C with acriflavine supplementation. In 10 assemblies ( $n=12$  samples), the serotype-associated genes and virulence genes *stx* and *eae* were detected on the same contig, with *stx* and *eae* being at a distance of 1.9 Mb. In two samples, the *eae*-positive O26 STEC was not identified. One sample (Milk2\_50\_37+A) showed a fragmented assembly with 34 contigs and a high genome coverage (111 $\times$ ). When we sub-sampled *E. coli* reads to lower genome coverage, still *stx* and *eae* could not be assembled on the same contig, although the filtered reads N50 value (13 136bp) and the quality (12.6) seemed acceptable. The pathotypes predicted by the STECmetadetector pipeline was 'EHEC?' since *stx* and *eae* genes were detected but could not be assembled on the same contig. In the second sample (Milk5\_5\_37+A\_1), a commensal *E. coli* genome was assembled instead of the STEC that was inoculated at a low contamination level of 5 c.f.u. (ml raw milk)<sup>-1</sup> (Fig. 4b).

Our results showed that the identification of the inoculated STEC strain was possible from an STEC concentration as low as 5 c.f.u. (ml raw milk)<sup>-1</sup> ( $n=4/5$ ) [corresponding to  $2.59 \times 10^8$  –  $2.62 \times 10^{10}$  copies (ml enriched milk)<sup>-1</sup>] (Table 3) under the above-mentioned enrichment conditions. Although commensal *E. coli* were naturally present and quantified to  $1.37 \times 10^9$  copies ml<sup>-1</sup> post-enrichment in non-artificially contaminated milk (Milk5; Table 3), the inoculated STEC O26 strain growth [contamination level of 5 c.f.u. (ml raw milk)<sup>-1</sup>] was sufficient to fully characterized its genome in 2/3 replicates. In the third replicate (Milk5\_5\_37+A\_1), a commensal *E. coli* strain grew better than the STEC and was fully assembled perturbing STEC strain characterization using assembly ( $n=1/3$ ) (Fig. 4b).

### Long-read metagenomics revealed that raw milk microbiota impacts the growth of *eae*-positive STEC

In order to perform all the artificial contamination experiments on fresh raw milk, five different fresh raw milk samples were used in this study. Three milk samples (1, 2 and 3) were bought from a commercial farm in France and two milk samples (4 and 5) were collected from an experimental farm in Germany. Long-read metagenomics revealed different raw milk microbiota for all investigated enriched milk samples. *E. coli* reads were detected only in non-contaminated raw milk samples Milk3 and Milk5 representing 6.8 and 46.23% of filtered reads, respectively (Fig. 5). The use of acriflavine affected the survival of certain Gram-positive bacteria, which consequently increased the proportion of STEC. One pertinent example is Milk3 in which *Streptococcus* spp. dominated and limited the STEC growth in the absence of acriflavine (Fig. 3b). However, acriflavine did not completely prevent the growth of Gram-positive bacteria, as observed with Milk4 that contained a high proportion of



**Fig. 5.** Barplot representing the most abundant bacterial species (>2%) in all studied raw cow's milk samples after enrichment at 37 °C without acriflavine supplementation (37-A, n=5). MinION data obtained from five natural raw milk samples were analysed with the STECmetadetector pipeline and the composition presented.

*Bacillus* spp. (Fig. 3b). While in most samples the STEC strain growth was not negatively affected in acriflavine-supplemented enrichments, in artificially contaminated milk sample 4 the observed percentage of *E. coli* reads post-enrichment was lower than expected (Fig. 4a).

Commensal *E. coli* strains present in raw milk did not appear to compromise the growth of the STEC O26 inoculum. Indeed, two non-inoculated milk samples (Milk3\_37 A and Milk5\_37 A) had detectable levels of *E. coli* ( $1.97 \times 10^7$  and  $1.37 \times 10^9$  copies  $\text{ml}^{-1}$  total *E. coli*, respectively) post-enrichment at 37 °C without acriflavine using qdPCR. Characterization of *E. coli* reads from milk samples Milk3\_37 A and Milk5\_37 A revealed the presence of commensal strains of serotype O149:H20 in Milk3\_37 A, and both O185:H2 [sequence type (ST)2522] and O8:H19 (ST162) strains in Milk5\_37 A enriched at 37 °C without acriflavine (Table S3). The natural presence of O185 and O8 strains in raw milk 5 was confirmed by qdPCR analysis ( $8.75 \times 10^8$  and  $1.84 \times 10^7 - 3 \times 10^7$  copies  $\text{ml}^{-1}$ , respectively).

In artificially STEC O26 contaminated milk batch 3,  $wzx_{O149}$  was not detected when enriched at 37 °C in the presence of acriflavine, but it was detected ( $7.43 \times 10^5$  copies  $\text{ml}^{-1}$ ) when enriched at 37 °C in the absence of acriflavine. Commensal strains markers O185 and O8 were not detected nor assembled in artificially contaminated raw milk samples from milk batch 5, except in one of the three biological replicates of Milk5 inoculated at  $0.5 \times 10^1$  c.f.u.  $\text{ml}^{-1}$ . The O185 marker was quantified to  $4.55 - 8.84 \times 10^9$  copies  $\text{ml}^{-1}$  by qdPCR. Additionally, this commensal strain (O185:H2) was assembled, while the inoculated STEC was not. Although read mapping detected O26 as the major O-group in all other artificially contaminated samples, in this sample (Milk5\_5\_37+A\_1) more reads mapped to O185 (154×) than O26 (13×) (Table S3). Analysis on this replicate showed a difference of approximately one log (copies  $\text{ml}^{-1}$ ) between total *E. coli* and the inoculated strain with both real-time PCR (i.e. a difference of 3  $C_i$  values for both *stx* and *eae* – specific to the inoculated strain – and *wecA* – common to both natural and inoculated strains) and qdPCR [estimated between  $2.41 \times 10^7$  and  $1.09 \times 10^8$  copies  $\text{ml}^{-1}$  for O26 marker specific of the inoculated strain, and between  $8.59 \times 10^8$  and  $1.92 \times 10^9$  copies  $\text{ml}^{-1}$  for *wecA* (Table 3)]. Similarly, in this replicate in which the O185:H2 strain was assembled, qdPCR results confirmed a higher level of the O185 strain [ $6.52 \times 10^8 - 10.09 \times 10^{10}$  copies  $\text{ml}^{-1}$  of O185 ( $wzx_{O185}$ )] compared to the STEC O26 strain. The presence of commensal *E. coli* did not affect the growth of the STEC O26 strain in 2/3 replicates of raw milk artificially contaminated with 5 c.f.u.  $\text{ml}^{-1}$  and enriched in acriflavine-supplemented BPW at 37 °C. However, it hindered its identification using an assembly-based approach in one replicate.

## DISCUSSION

Food-borne pathogens are a worldwide concern. Recently, different approaches for identification of food-borne pathogens were developed. While metagenomics is widely used for the detection and characterization of food-borne pathogens like *Salmonella* or *Listeria*, its efficacy to distinguish STEC that may be *eae*-positive or -negative in complex samples like raw milk is questionable [17]. *E. coli* is a member of the cow gut microbiota, in which commensal and/or pathogenic strains of different pathotypes may coexist. Cookson and colleagues detected as many as 50 different *E. coli* strains in a single sample of cattle faeces [50]. STEC harbour various virulence markers, like the *eae* gene, that are detectable by real-time PCR (qPCR) in complex samples like raw milk [34]. However, it is not resolvable whether these genes belong to the same strain or are present in different strains, and isolation is required to confirm the presence of all virulence factors in a single strain (strain characterization). STEC isolation is laborious, time-consuming and frequently unsuccessful in raw milk, because of the background flora and a lack of a STEC-specific isolation medium. The diversity of STEC emphasizes the need for STEC characterization methods that circumvent the isolation problems posed by current methods. The aim of this study was to explore metagenomics as an innovative STEC isolation-independent method able to identify and characterize STEC in enriched raw milk that could be applied to characterize *stx*-positive samples as an alternative to strain isolation.

So far, only a few studies that describe the detection and characterization of STEC using long-read metagenomics are available [18, 19]. Only one of them explored metagenomics in beef for specifically identifying *eae*-positive STEC of serotype O157:H7 that carry both *stx* and *eae* genes [18]. Given that these genetic markers, *stx* and *eae*, can also be carried by non-pathogenic *E. coli* strains or other bacterial species, this objective represents a real challenge. In this study, we developed a straightforward strategy different from that described by Buytaers *et al.* to identify *eae*-positive STEC using long-read metagenomics of artificially contaminated raw milk [18]. While they used a DNA walking approach linking *stx* and *eae* virulence genes to an *Escherichia* genome, we adopted an assembly-based method. This approach could be applied to characterize any type of *E. coli* contaminating a food matrix, regardless of the existence of reference strains for this pathotype. To mimic real samples, two *eae*-positive STEC O26 isolated from raw cow's milk were selected for this study. Indeed, O26 is one of the STEC serogroups most frequently found in clinical cases in Europe, and is also highly dominant in French and German raw milk [9, 34]. Firstly, we developed a wet-lab method to prepare the samples for long-read sequencing. Secondly, we determined the sensitivity of the method for identification of *eae*-positive STEC after enrichment of artificially contaminated raw milk using long-read metagenomics. Thirdly, we developed a complete and automated analysis pipeline for STEC strain characterization from MinION sequencing data.

Previous studies have determined that the detection limit of STEC strains using MinION sequencing was around  $10^7$  c.f.u. ml<sup>-1</sup> in wastewater [19]. These contamination levels are generally not reached in naturally contaminated raw milk samples without incubating the sample to get *E. coli* growing to a detectable level. Therefore, an enrichment step appears necessary in order to identify *eae*-positive STEC strains from dairy products. As demonstrated in this study, Flye performed best in assembling *eae*-positive STEC ( $n=10$ ), although below a coverage of 35×; *stx* and *eae* virulence genes were not consistently co-localized on a single contig; thus, providing a potential misidentification. A STEC genome at a coverage of 35× represents around 192.5 Mb (genome size=5.5Mb). Shotgun metagenomics studies on raw milk without enrichment revealed tremendous proportions of host DNA varying from 82.1 to 99.5% of sequenced reads [51–53]. Consequently, around 2% of the total DNA sequenced from raw milk in these conditions would correspond to bacterial DNA, which represents 60 Mb of an estimated 3 Gb output sequencing run. By supposing the presence of *E. coli* to be 100% of the bacterial DNA in the milk sample, 60 Mb would represent 10× STEC genome coverage, which is clearly not sufficient for a reliable STEC genome assembly and characterization. In addition, long reads are error-prone, and need to be filtered on quality and length, which, along with the background flora, contributes to reduce the amount of pertinent data available for analysis. Commercial host DNA-removal kits allow a significant increase in the ratio of bacterial DNA reads compared to host DNA reads [51]. However, removing 98% of the extracted DNA can lead to technical difficulties related to the amount of DNA necessary for MinION sequencing (1–2 µg genomic DNA versus only 1 ng genomic DNA for Illumina sequencing). Consequently, an incubation step to get viable *E. coli* growing to a detectable level was preferred, because it not only considerably reduces the proportion of host DNA while keeping viable STEC cells, but also enables sample multiplexing on a flow cell, which dramatically reduces the cost of long-read metagenomics.

Comparison of four different enrichment conditions revealed that even though no difference in terms of *eae*-positive STEC strain identification was observed, the use of acriflavine as recommended by the ISO/TS 13136:2012 [11] was relevant for enrichment efficacy, especially in raw milk containing Gram-positive bacteria. Although the precise benefit of acriflavine during enrichment is still debated [54, 55], in this study, acriflavine was used during the enrichment process because it did not appear to negatively influence the growth of the inoculated strains, while preventing the growth of some of the background flora. Further studies on the impact of acriflavine on the growth of STEC strains from various serotypes would be nonetheless required. We tested two different enrichment temperatures recommended for non-O157 and O157 STEC strains respectively, 37 and 41.5 °C. The inoculated O26:H11 STEC strains were not affected by either 37 or 41.5 °C. In 2008, Baylis demonstrated that the temperature of enrichment alone does not affect the growth of STEC strains, even though it may help reduce the background flora [56]. Here, acriflavine supplementation was more effective at reducing the Gram-positive background flora than performing the enrichment step at 41.5 °C. Hence, all the samples further analysed were enriched at 37 °C for 18–20 h in the presence of acriflavine. Although

it was not tested here, it might be theoretically possible to reduce the enrichment time while maintaining sufficient levels of cells for *eae*-positive STEC characterization using long-read metagenomics.

In this study, we aimed to test the long-read metagenomics potential on STEC assembly from metagenomics samples for strain characterization. Under the optimized enrichment conditions, STEC-negative raw milk samples artificially contaminated with an O26 *eae*-positive STEC strain, at initial concentrations from  $0.5 \times 10^1$  to  $0.5 \times 10^3$  c.f.u. ml<sup>-1</sup> in raw milk, could be clearly identified by MinION long-read metagenomics. The percentage of reads assigned to *E. coli*, as expected, increased with STEC concentration. It is noteworthy that the presence of certain bacterial genera, even in the presence of acriflavine, may adversely affect the growth of STEC strains (i.e. Milk4, which contained 27.25% of *Bacillus*, showed lower percentages of *E. coli* reads after artificial contamination using three concentrations of 4712-O26 culture in raw milk and enrichment with acriflavine).

For easy data analysis of long-read metagenomics, we developed a strain-level characterization pipeline specifically for STEC. The pipeline aims at identifying in particular STEC containing *stx* and *eae* genes in the same strain. This pipeline gathers appropriate tools for STEC genome characterization and enables reproducible analysis on different metagenomics runs. Basecalling and demultiplexing are currently not part of the STECmetadetector pipeline, but should be run on a system enabling Graphics Processing Units (GPUs) to considerably reduce basecalling time. We recommend doing precise basecalling using the high-accuracy or super high-accuracy model, which may help reduce the error-rate and better characterize the strain. Nevertheless, direct basecalling, using the fast basecalling model, and demultiplexing, may be performed on-board using either the MinIT in combination with Mk1B or the Mk1C device, if no GPUs are installed. Once the basecalling and demultiplexing are done, the optimized pipeline designed in this study is useful to fully and precisely characterize *E. coli* strains. Extraction of *E. coli* reads before assembly significantly reduces the overall amount of data and optimizes resource consumption. Although the STECmetadetector pipeline was designed for identifying *eae*-positive STEC, it may be used to characterize other *E. coli* pathogroups and may prove particularly useful for characterization of hybrid or heteropathotype strains.

Using the STECmetadetector pipeline designed in this study, we were able to identify *eae*-positive STEC from long-read metagenomics data and obtained the complete chromosome in one contig from  $2.59 \times 10^8$  copies (ml enriched raw milk)<sup>-1</sup>. Flye assembler, included in the pipeline, not only was shown to be one of the best for *E. coli* genome assembly, but also is currently the only long-read assembler enabling metagenome assembly (metaFlye) and offering the best balance regarding 'time/results quality' [25, 57]. The metaFlye assembler was released to assemble similar reads belonging to different bacterial species present in metagenomics samples but is not designed to distinguish different strains from the same species. To reduce analysis complexity from metagenomes, in part caused by the presence of other bacterial species and potentially multiple *E. coli* strains, we filtered and assembled *E. coli* reads only. In line with Vicedomini and colleagues [45], we showed that even in the presence of several *E. coli* strains in the same sample, Flye usually assembled only one *E. coli* genome completely. This is one major drawback of current assemblers, which may hinder the development of this method. The presence of additional small contigs of different coverages, as well as multiple O-groups and/or multiple alleles of the different MLST genes, may indicate the presence of different *E. coli* strains in the sample.

It is probable that the presence of other *E. coli* strains may compromise *eae*-positive STEC identification since the *eae* and *stx* genes may be distantly located on the *E. coli* chromosome (i.e. 1.9 Mb here), and assemblers are currently not systematically able to distinguish strains. Although we did not aim to analyse the impact of commensal *E. coli* strains (*stx*- and *eae*-) on the STEC strain growth, we had to use different milk samples to perform all the artificial contamination experiments on fresh milk in which commensal *E. coli* were present in 2/5 milk samples. We noticed that it affected the STEC assembly in 1/3 replicates inoculated with 5 c.f.u. (ml raw milk)<sup>-1</sup> (Milk5\_5\_37+A\_1). The inoculation level of 5 c.f.u. ml<sup>-1</sup> thus most likely represents the limit of STEC characterization by long-read metagenomics in the presence of commensal *E. coli*. A recent strain-separation tool named Strainberry was used to distinguish multiple *E. coli* strains [45]. If Strainberry predicted the presence of multiple strains, the post-Strainberry module was used to characterize each resulting strain. We were able to separate two different commensal *E. coli* genomes in uncontaminated raw milk using Strainberry. Identifying a STEC strain in the presence of commensal *E. coli* strains at an initial low inoculation level of 5 c.f.u. ml<sup>-1</sup> in raw milk was successful in two out of three replicates using the optimized assembly-based method described in our study. The outlier was only observed for the replicate where the inoculated strain growth was limited compared to the two other replicates, as confirmed by qdPCR and real-time PCR analysis. It is probable that these conditions constitute the lower limit and that, for this replicate, slightly less than 5 c.f.u. ml<sup>-1</sup> were sampled and inoculated, allowing a better growth of the O185 strain.

The presence of commensal *E. coli* strains together with STEC are not the only challenge for STEC growth in complex matrices. Raw milk microbiota is very variable because of the milk's high content of nutrients and the diverse sources of contamination from the environment within the production chain leading to different bacterial colonization [58, 59]. Gram-positive bacteria, for example *Lactococcus lactis* and *Streptococcus* spp., generally dominate in milk microbiota. A high proportion of *Lactococcus* spp. reads was detected in one sample (Milk4\_37 A) and acriflavine prevented its growth during enrichment. Similarly, in another milk sample (Milk3\_37 A) many reads were assigned to *Streptococcus uberis*, which is known to be a cause of mastitis and may acidify the medium, but they were not detected in an acriflavine-supplemented enrichment sample [60]. Despite certain variability



in the raw milk microbiota, long-read metagenomics seems to be a promising approach for identification of pathogenic STEC in enriched raw milk.

In conclusion, the use of long-read metagenomics for food-borne pathogen identification is emerging and constantly evolving. The work of Buytaers and co-workers was a proof-of principle for STEC identification using long-read metagenomics in beef samples [18]. By using a different approach, we were able to reveal the co-localization of O26:H11 serotype-associated genes, *stx* and *eae*, in the same strain with the objective to identify *eae*-positive STEC strains from artificially contaminated raw milk samples. The complete approach with optimized enrichment conditions for the growth of *eae*-positive STEC strains, and its characterization using long-reads metagenomics using an adequate pipeline (STECmetadetecter), is a significant improvement for STEC identification in metagenomes. The methodology has potential to be applied for characterizing STEC directly in food samples. However, its widespread application in monitoring will be possible only when new, more strain-aware assemblers become available.

#### Funding information

This project was funded by the German Federal Institute for Risk Assessment (BfR) under the grant no. 1322-765 and by the French Agency for Food, Environmental and Occupational Health and Safety (ANSES) under the grant no. ACE16 1120 AAEEH. The study was realized during the PhD of S.J. (shared PhD student between ANSES and BfR).

#### Acknowledgements

We are thankful to Holger Brendebach from the Department of Biological Safety, German Federal Institute for Risk Assessment (BfR), Germany, for valuable help in creating the STECmetadetecter Docker container.

#### Author contribution

S.D., J.G., P.F., B.M. and E.S. conceptualized the project. P.F. and B.M. were in charge of funding acquisition. S.J., M.-L.T., S.D. and J.G. conducted investigations. Formal analysis was conducted by S.J. and M.-L.T. Methodology and resources were from P.F., B.M., J.G., S.D., F.V., C.D., E.S. and A.G. Software was by C.D., J.G., S.J. and F.V. S.J. wrote the original draft, which was reviewed and validated by all authors.

#### Conflicts of interest

The authors declare that there are no conflicts of interest.

#### References

- Levine MM. *Escherichia coli* that cause diarrhea: enterotoxigenic, enteropathogenic, enteroinvasive, enterohemorrhagic, and enteroadherent. *J Infect Dis* 1987;155:377–389.
- Kaper JB, Nataro JP, Mobley HL. Pathogenic *Escherichia coli*. *Nat Rev Microbiol* 2004;2:123–140.
- Nataro JP, Kaper JB. Diarrheagenic *Escherichia coli*. *Clin Microbiol Rev* 1998;11:142–201.
- Karmali MA. Infection by Shiga toxin-producing *Escherichia coli*: an overview. *Mol Biotechnol* 2004;26:117–122.
- Bielaszewska M, Mellmann A, Zhang W, Köck R, Fruth A, et al. Characterisation of the *Escherichia coli* strain associated with an outbreak of haemolytic uraemic syndrome in Germany, 2011: a microbiological study. *Lancet Infect Dis* 2011;11:671–676.
- Mariani-Kurkdjian P, Lemaître C, Bidet P, Perez D, Boggini L, et al. Haemolytic-uraemic syndrome with bacteraemia caused by a new hybrid *Escherichia coli* pathotype. *New Microbes New Infect* 2014;2:127–131.
- Bruyand M, Mariani-Kurkdjian P, Gouali M, de Valk H, King LA, et al. Hemolytic uremic syndrome due to Shiga toxin-producing *Escherichia coli* infection. *Med Mal Infect* 2018;48:167–174.
- Cointe A, Birgy A, Bridier-Nahmias A, Mariani-Kurkdjian P, Walewski V, et al. *Escherichia coli* O80 hybrid pathotype strains producing Shiga toxin and ESBL: molecular characterization and potential therapeutic options. *J Antimicrob Chemother* 2020;75:537–542.
- Koutsoumanis K, Allende A, Alvarez-Ordóñez A, Bover-Cid S, Chemaly M, et al. Pathogenicity assessment of Shiga toxin-producing *Escherichia coli* (STEC) and the public health risk posed by contamination of food with STEC. *EFSA J* 2020;18:e05967.
- Gyles CL. Shiga toxin-producing *Escherichia coli*: an overview. *J Anim Sci* 2007;85:E45–E62.
- International Organization for Standardization. *ISO/TS 13136:2012. Microbiology of Food and Animal Feed – Real-Time Polymerase Chain Reaction (PCR)-Based Method for the Detection of Food-Borne Pathogens – Horizontal Method for the Detection of Shiga Toxin-Producing *Escherichia coli* (STEC) and the Determination of O157, O111, O26, O103, and O145 Serogroups*. Geneva: International Organization for Standardization; 2012.
- Goldstein S, Beka L, Graf J, Klassen JL. Evaluation of strategies for the assembly of diverse bacterial genomes using MinION long-read sequencing. *BMC Genomics* 2019;20:23.
- Schadt EE, Turner S, Kasarskis A. A window into third-generation sequencing. *Hum Mol Genet* 2010;19:R227–R240.
- Lu H, Giordano F, Ning Z. Oxford nanopore MinION sequencing and genome assembly. *Genomics Proteomics Bioinformatics* 2016;14:265–279.
- Neal-McKinney JM, Liu KC, Lock CM, Wu W-H, Hu J. Comparison of MiSeq, MinION, and hybrid genome sequencing for analysis of *Campylobacter jejuni*. *Sci Rep* 2021;11:5676.
- Taylor AJ, Kelly DJ. The function, biogenesis and regulation of the electron transport chains in *Campylobacter jejuni*: new insights into the bioenergetics of a major food-borne pathogen. *Adv Microb Physiol* 2019;74:239–329.
- Azineiro S, Roumani F, Carvalho J, Prado M, Garrido-Maestu A. Suitability of the MinION long read sequencer for semi-targeted detection of foodborne pathogens. *Anal Chim Acta* 2021;1184:339051.
- Buytaers FE, Saltykova A, Denayer S, Verhaegen B, Vanneste K, et al. Towards real-time and affordable strain-level metagenomics-based foodborne outbreak investigations using Oxford nanopore sequencing technologies. *Front Microbiol* 2021;12:738284.
- Maguire M, Kase JA, Roberson D, Muruvanda T, Brown EW, et al. Precision long-read metagenomics sequencing for food safety by detection and assembly of Shiga toxin-producing *Escherichia coli* in irrigation water. *PLoS One* 2021;16:e0245172.
- Siekanić G, Roux E, Lemane T, Guédon E, Nicolas J. Identification of isolated or mixed strains from long reads: a challenge met on *Streptococcus thermophilus* using a MinION sequencer. *Microb Genom* 2021;7:000654.
- González-Escalona N, Allard MA, Brown EW, Sharma S, Hoffmann M. Nanopore sequencing for fast determination of plasmids, phages, virulence markers, and antimicrobial resistance genes in Shiga toxin-producing *Escherichia coli*. *PLoS One* 2019;14:e0220494.

22. Peritz A, Paoli GC, Chen CY, Gehring AG. Serogroup-level resolution of the "Super-7" Shiga toxin-producing *Escherichia coli* using nanopore single-molecule DNA sequencing. *Anal Bioanal Chem* 2018;410:5439–5444.
23. Leonard SR, Mammel MK, Lacher DW, Elkins CA. Application of metagenomic sequencing to food safety: detection of Shiga toxin-producing *Escherichia coli* on fresh bagged spinach. *Appl Environ Microbiol* 2015;81:8183–8191.
24. Steyert SR, Sahl JW, Fraser CM, Teel LD, Scheutz F, et al. Comparative genomics and stx phage characterization of LEE-negative Shiga toxin-producing *Escherichia coli*. *Front Cell Infect Microbiol* 2012;2:133.
25. Wick RR, Holt KE. Benchmarking of long-read assemblers for prokaryote whole genome sequencing. *F1000Res* 2019;8:2138.
26. Jaudou S, Tran ML, Vorimore F, Fach P, Delannoy S. Evaluation of high molecular weight DNA extraction methods for long-read sequencing of Shiga toxin-producing *Escherichia coli*. *PLoS One* 2022;17:e0270751.
27. Mylius M, Dreesman J, Putz M, Pallasch G, Beyrer K, et al. Shiga toxin-producing *Escherichia coli* O103:H2 outbreak in Germany after school trip to Austria due to raw cow milk, 2017 – the important role of international collaboration for outbreak investigations. *Int J Med Microbiol* 2018;308:539–544.
28. Hall MB. Rasusa: randomly subsample sequencing reads to a specified coverage. *JOSS* 2022;7:3941.
29. Kolmogorov M, Bickhart DM, Behsaz B, Gurevich A, Rayko M, et al. metaFlye: scalable long-read metagenome assembly using repeat graphs. *Nat Methods* 2020;17:1103–1110.
30. Vaser R, Šikić M. Time- and memory-efficient genome assembly with Raven. *Nat Comput Sci* 2021;1:332–336.
31. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, et al. Canu: scalable and accurate long-read assembly via adaptive *k*-mer weighting and repeat separation. *Genome Res* 2017;27:722–736.
32. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* 2013;29:1072–1075.
33. Michelet L, Delannoy S, Devillers E, Umhang G, Aspan A, et al. High-throughput screening of tick-borne pathogens in Europe. *Front Cell Infect Microbiol* 2014;4:103.
34. Delannoy S, Tran ML, Fach P. Insights into the assessment of highly pathogenic Shiga toxin-producing *Escherichia coli* in raw milk and raw milk cheeses by high throughput real-time PCR. *Int J Food Microbiol* 2022;366:109564.
35. Anderson A, Pietsch K, Zucker R, Mayr A, Müller-Hohe E, et al. Validation of a duplex real-time PCR for the detection of *Salmonella* spp. in different food products. *Food Anal Methods* 2011;4:259–267.
36. Lamparter MC, Seemann A, Hobe C, Schuh E. Using hydrochloric acid and bile resistance for optimized detection and isolation of Shiga toxin-producing *Escherichia coli* (STEC) from sprouts. *Int J Food Microbiol* 2020;322:108562.
37. Coudray-Meunier C, Fraisse A, Martin-Latil S, Delannoy S, Fach P, et al. A novel high-throughput method for molecular detection of human pathogenic viruses using a nanofluidic real-time PCR system. *PLoS One* 2016;11:e0147832.
38. Mölder F, Jablonski KP, Letcher B, Hall MB, Tomkins-Tinch CH, et al. Sustainable data analysis with Snakemake. *F1000Res* 2021;10:33.
39. De Coster W, D'Hert S, Schultz DT, Cruts M, Van Broeckhoven C. NanoPack: visualizing and processing long-read sequencing data. *Bioinformatics* 2018;34:2666–2669.
40. Wood DE, Lu J, Langmead B. Improved metagenomic analysis with Kraken 2. *Genome Biol* 2019;20:257.
41. Li H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* 2018;34:3094–3100.
42. Joensen KG, Tetzschner AM, Iguchi A, Aarestrup FM, Scheutz F. Rapid and easy in silico serotyping of *Escherichia coli* isolates by use of whole-genome sequencing data. *J Clin Microbiol* 2015;53:2410–2426.
43. Wirth T, Falush D, Lan R, Colles F, Mensa P, et al. Sex and virulence in *Escherichia coli*: an evolutionary perspective. *Mol Microbiol* 2006;60:1136–1151.
44. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res* 2015;25:1043–1055.
45. Vicedomini R, Quince C, Darling AE, Chikhi R. Strainberry: automated strain separation in low-complexity metagenomes using long reads. *Nat Commun* 2021;12:4485.
46. R Core Team. R: a language and environment for statistical computing. Vienna: R Foundation for Statistical Computing, 2018. <https://www.R-project.org/>
47. Grünig B, Dale R, Sjödin A, Chapman BA, Rowe J, et al. Bioconda: sustainable and comprehensive software distribution for the life sciences. *Nat Methods* 2018;15:475–476.
48. Wickham H. *ggplot2: Elegant Graphics for Data Analysis*. New York: Springer; 2009.
49. Wickham H. Reshaping data with the reshape package. *J Stat Soft* 2007;21:12.
50. Cookson AL, Biggs PJ, Marshall JC, Reynolds A, Collis RM, et al. Culture independent analysis using *gnd* as a target gene to assess *Escherichia coli* diversity and community structure. *Sci Rep* 2017;7:841.
51. Grütze K, Gwida M, Deneke C, Brendebach H, Projahn M, et al. Direct identification and molecular characterization of zoonotic hazards in raw milk by metagenomics using *Brucella* as a model pathogen. *Microb Genom* 2021;7:000552.
52. Nikoloudaki O, Lemos Junior WJF, Campanaro S, Di Cagno R, Gobbetti M. Role prediction of Gram-negative species in the resistome of raw cow's milk. *Int J Food Microbiol* 2021;340:109045.
53. Yap M, Gleeson D, O'Toole PW, O'Sullivan O, Cotter PD. Seasonality and geography have a greater influence than the use of chlorine-based cleaning agents on the microbiota of bulk tank raw milk. *Appl Environ Microbiol* 2021;87:e0108121.
54. Amagliani G, Rotundo L, Carloni E, Omiccioli E, Magnani M, et al. Detection of Shiga toxin-producing *Escherichia coli* (STEC) in ground beef and bean sprouts: evaluation of culture enrichment conditions. *Food Res Int* 2018;103:398–405.
55. Mancusi R, Trevisani M. Enumeration of verocytotoxigenic *Escherichia coli* (VTEC) O157 and O26 in milk by quantitative PCR. *Int J Food Microbiol* 2014;184:121–127.
56. Baylis CL. Growth of pure cultures of verocytotoxin-producing *Escherichia coli* in a range of enrichment media. *J Appl Microbiol* 2008;105:1259–1265.
57. Chen Z, Erickson DL, Meng J. Benchmarking long-read assemblers for genomic analyses of bacterial pathogens using Oxford nanopore sequencing. *Int J Mol Sci* 2020;21:E9161.
58. Quigley L, O'Sullivan O, Stanton C, Beresford TP, Ross RP, et al. The complex microbiota of raw milk. *FEMS Microbiol Rev* 2013;37:664–698.
59. Parente E, Ricciardi A, Zotta T. The microbiota of dairy milk: a review. *Int Dairy J* 2020;107:104714.
60. Kabelitz T, Aubry E, van Vorst K, Amon T, Fulde M. The role of *Streptococcus* spp. in bovine mastitis. *Microorganisms* 2021;9:1497.



## Publication 4

### **Exploring long-read metagenomics for full characterization of Shiga toxin-producing *Escherichia coli* in presence of commensal *E. coli***






Jaudou Sandra, Deneke Carlus, Tran Mai-Lan, Salzinger Carina, Vorimore Fabien, Goehler André, Schuh Elisabeth, Malorny Burkhard, Fach Patrick, Grützke Josephine, Delannoy Sabine

Microorganisms, Special Issue VTEC2023, <https://www.mdpi.com/2076-2607/11/8/2043>



Article

# Exploring Long-Read Metagenomics for Full Characterization of Shiga Toxin-Producing *Escherichia coli* in Presence of Commensal *E. coli*

Sandra Jaudou <sup>1,2</sup> , Carlus Deneke <sup>2</sup>, Mai-Lan Tran <sup>1,3</sup>, Carina Salzinger <sup>4</sup>, Fabien Vorimore <sup>3</sup> , André Goehler <sup>4</sup> , Elisabeth Schuh <sup>4</sup>, Burkhard Malorny <sup>2</sup> , Patrick Fach <sup>1,3</sup>, Josephine Grütze <sup>2</sup> and Sabine Delannoy <sup>1,3,\*</sup> 

<sup>1</sup> COLiPATH Unit, Laboratory for Food Safety, ANSES, 94700 Maisons-Alfort, France; sandra.jaudou.ext@anses.fr (S.J.)

<sup>2</sup> National Study Center for Sequencing in Risk Assessment, Department of Biological Safety, German Federal Institute for Risk Assessment, 12277 Berlin, Germany

<sup>3</sup> Genomics Platform IdentityPath, Laboratory for Food Safety, ANSES, 94700 Maisons-Alfort, France

<sup>4</sup> National Reference Laboratory for *Escherichia coli* Including VTEC, Department of Biological Safety, German Federal Institute for Risk Assessment, 12277 Berlin, Germany

\* Correspondence: sabine.delannoy@anses.fr



**Citation:** Jaudou, S.; Deneke, C.; Tran, M.-L.; Salzinger, C.; Vorimore, F.; Goehler, A.; Schuh, E.; Malorny, B.; Fach, P.; Grütze, J.; et al. Exploring Long-Read Metagenomics for Full Characterization of Shiga Toxin-Producing *Escherichia coli* in Presence of Commensal *E. coli*. *Microorganisms* **2023**, *11*, 2043. <https://doi.org/10.3390/microorganisms11082043>

Academic Editor: Peter Neubauer

Received: 26 June 2023

Revised: 26 July 2023

Accepted: 7 August 2023

Published: 9 August 2023



**Copyright** © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

**Abstract:** The characterization of Shiga toxin-producing *Escherichia coli* (STEC) is necessary to assess their pathogenic potential, but isolation of the strain from complex matrices such as milk remains challenging. In previous work, we have shown the potential of long-read metagenomics to characterize *eae*-positive STEC from artificially contaminated raw milk without isolating the strain. The presence of multiple *E. coli* strains in the sample was shown to potentially hinder the correct characterization of the STEC strain. Here, we aimed at determining the STEC:commensal ratio that would prevent the characterization of the STEC. We artificially contaminated pasteurized milk with different ratios of an *eae*-positive STEC and a commensal *E. coli* and applied the method previously developed. Results showed that the STEC strain growth was better than the commensal *E. coli* after enrichment in acriflavine-supplemented BPW. The STEC was successfully characterized in all samples with at least 10 times more STEC post-enrichment compared to the commensal *E. coli*. However, the presence of equivalent proportions of STEC and commensal *E. coli* prevented the full characterization of the STEC strain. This study confirms the potential of long-read metagenomics for STEC characterization in an isolation-free manner while refining its limit regarding the presence of background *E. coli* strains.

**Keywords:** metagenomics; STEC; milk; long-read sequencing; isolation-free characterization

## 1. Introduction

Shiga toxin-producing *Escherichia coli* (STEC) are diarrheic *E. coli* characterized by the presence of a specific virulence factor: the Shiga toxin [1]. A subset of STEC, known as *eae*-positive STEC, additionally possesses the ability to adhere to the intestinal epithelium and cause attaching and effacing (A/E) lesions. Adhesion is conferred by the intimin protein encoded by the *eae* gene located on the locus of enterocyte effacement (LEE) pathogenicity island [2]. EFSA reports showed that *eae*-positive STEC are most frequently associated with severe human illnesses, especially hemolytic uremic syndrome (HUS) [3].

STEC detection methods in food products involve time-consuming isolation steps to characterize the STEC strain [4]. Isolation-free approaches for STEC characterization in food matrices would circumvent this but imply using metagenomics [5,6]. While short-read metagenomics enables STEC detection from complex matrices, it does not permit full characterization of the strain. In addition, different *E. coli* strains may be present in the same sample [7,8]. Thus, strain-level metagenomics for the identification of STEC such as *eae*-positive STEC or carrying other adhesion factors is challenging. Long-read metagenomics have been shown to be applicable for identifying *eae*-

positive STEC in artificially contaminated beef, raw milk, and wastewater using different bio-informatics approaches [9–11].

In a previous study, we tested the feasibility of identifying an *eae*-positive STEC strain from artificially contaminated raw milk samples using long-read metagenomics. Assembly-based methods have the advantage of enabling the characterization of the STEC strain but depend on the contamination level [10,11]. We showed that it was possible to identify *eae*-positive STEC from  $2.59 \times 10^8$  copies.mL<sup>-1</sup> of STEC in the absence of other interfering *E. coli* strains. Nevertheless, this study was a proof-of-principle, and the presence of a diverse background flora, as well as the presence of multiple *E. coli* strains, may hamper the analysis [10]. Since the STEC contamination level can be low (below 100 cells), it is necessary to determine the STEC full characterization limit in the presence of background *E. coli* using an assembly-based metagenomics approach.

In this study, we artificially co-contaminated pasteurized milk using increasing and known inoculation levels of a commensal *E. coli* and an *eae*-positive STEC strain, both originally isolated from raw milk. Because the natural background flora of raw milk and the potential presence of commensal *E. coli* may affect the growth of STEC, pasteurized milk was used to better control the STEC:commensal ratios [10]. Each contaminated milk sample was enriched in acriflavine-supplemented buffered peptone water (BPW) before sequencing. We used the published STECmetadetector pipeline on MinION-generated data [10] and determined the limit of this approach.

## 2. Materials and Methods

### 2.1. *E. coli* Strains Used for Artificial Co-Contamination

Two *E. coli* strains were used in this study for artificial contamination of pasteurized milk. One commensal *E. coli* (*stx* and *eae* negative) strain was isolated from cow raw milk in 2021. This strain (BfR-EC-19174, *stx-eae*-, O2:H10, ST 6527) was isolated following the ISO/TS 13136:2012 method [4]. Its serogroup was determined using seroagglutination [12] and further confirmed with qPCR using primers and probes as described by Delannoy and colleagues [13]. The second strain, 6423-O26, was an *eae*-positive STEC of serotype O26:H11 as already described [10,14]. Both strains BfR-EC-19174 (referred to as the commensal *E. coli*) and 6423-O26 (referred to as the *eae*-positive STEC) were revived from 20–30% glycerol stock on TSA plates. One colony was cultivated in brain heart infusion (BHI) overnight at 37 °C with agitation.

### 2.2. Artificial *E. coli* Mixtures

For each strain, one milliliter of overnight pure culture was used as a starting material for eight serial dilutions (1:10) in BHI or BPW. Optical density (OD) was measured at 600 nm using 1 mL BHI or 1 mL of the following: 1 mL BHI in 9 mL BPW, as a blank. For inoculum determination, triplicate cell counting was performed by plating three dilutions on tryptic soy agar (TSA) plates (100 µL) and incubated overnight at 37 °C. Five mixtures of the two *E. coli* strains were prepared as represented in Table 1.

**Table 1.** Commensal *E. coli* and *eae*-positive STEC artificial mixtures used for co-contamination of pasteurized milk.

Mix	O26:O2 Estimated Ratio	STEC Inoculum (X CFU.mL <sup>-1</sup> )	Commensal Inoculum (X CFU.mL <sup>-1</sup> )	Final STEC:Commensal Estimated Inoculum in Milk (X CFU.mL <sup>-1</sup> )
1	1:1	10 <sup>2</sup>	10 <sup>2</sup>	5:5
2	1:10	10 <sup>2</sup>	10 <sup>3</sup>	5:50
3	1:100	10 <sup>2</sup>	10 <sup>4</sup>	5:500
4	10:1	10 <sup>3</sup>	10 <sup>2</sup>	50:5
5	100:1	10 <sup>4</sup>	10 <sup>2</sup>	500:5

### 2.3. Co-Contamination of Pasteurized Milk

Pasteurized milk with 3.8% fat was bought from a grocery store in Berlin (Germany), and 1 mL was artificially contaminated with 1 mL of each mixture of commensal (BfR-EC-19174) and *eae*-positive STEC (6423-O26) prepared as mentioned in Table 1. For artificial contamination, three biological replicates of artificially contaminated milk samples were enriched for 18–20 h at 37 °C in 8 mL of BPW supplemented with acriflavine at a final concentration of 12 mg.L<sup>-1</sup> [4].

### 2.4. DNA Extraction and Quality Control

One milliliter of milk-enriched mixtures was used for DNA extraction using the MasterPure Lucigen protocol as described by Jaudou et al., 2022 [10]. To ensure the feasibility of MinION sequencing, the quality of the extracted DNA was assessed using Nanodrop 1.0 (A260/A280 and A260/A230 ratios), and DNA was quantified using Qubit 3.0 and the broad range kit following the manufacturer's instructions.

### 2.5. Detection and Quantification of Commensal and *Eae*-Positive STEC Strains Using qPCR and qdPCR, Respectively

Real-time PCR was performed to detect the presence of each strain used for the artificial contamination of pasteurized milk after enrichment and quantitative digital PCR (qdPCR) for quantitative estimation (number of copies.mL<sup>-1</sup>). The *wzx* gene coding for the O-antigen flippase is serogroup-specific and present in a single copy in the *E. coli* genome. Genetic markers *wzx*<sub>O2</sub> and *wzx*<sub>O26</sub> were used to detect and quantify the commensal and the *eae*-positive STEC, respectively, in each milk enrichment. Sequences for primers and probes are given in Table 2. Probes were labeled with 6-carboxyfluorescein (FAM) and black hole quencher (BHQ1) (Eurofins).

Table 2. Primers and probes sequences for *wzx*<sub>O2</sub> and *wzx*<sub>O26</sub>.

Genetic Marker	Sequence (5'-3')	Reference
<i>wzx</i> <sub>O2</sub>	Primer F- GCCAAGTGCAAAGTTTAATCACAAT	[13]
	Primer R- CTTGCCAATTTTCCGAGTATAT	
	Probe [6FAM]- CCTCTGCACCTGTAAGCACTGGCCTT-[BHQ1]	
<i>wzx</i> <sub>O26</sub>	Primer F- CGCGACGGCAGAGAAAATT	[15]
	Primer R- AGCAGGCTTTTATATTCTCCAAC TTT	
	Probe [6FAM]-CCCGGTTAAATCAATACTATTTCACGAGGTTGA-[BHQ1]	

For real-time PCR, a mix (20 µL) containing 1X PerfeCTa qPCR ToughMix low ROX (QuantaBio), primers, and probes at a final concentration of 0.3 µM and completed with nuclease-free water was used for each marker. Extracted DNA (2 µL) was added to the mix. Targeted sequences were amplified using the CFX96 system (BioRad) with a first step at 95 °C for 10 min (5 °C.s<sup>-1</sup>), a second step consisting of 39 cycles at 95 °C for 15 s (2 °C.s<sup>-1</sup>) and at 60 °C for 1 min, and a final step at 40 °C for 30 s (5 °C.s<sup>-1</sup>).

Mixes for qdPCR were prepared as follows: 3 µL PerfeCTa 2X qPCR ToughMix low ROX (QuantaBio), 0.6 µL 20X GE sample loading reagent (Fluidigm), 0.3 µL 20X primer stock prepared using 18 µM for each primer (forward and reverse) and 4 µM of probe, and 1.8 µL of extracted DNA, completed to 6 µL with nuclease-free water. Quantitative digital PCR was run on qdPCR 37k IFC digital array microfluidic chips using a Biomark system (Standard BioTools). For amplification, the thermal profile constituted of a first step at 50 °C for 2 min and a second step at 95 °C for 10 min followed by 40 cycles at 95 °C for 15 sec and 60 sec at 60 °C (2 °C s<sup>-1</sup>). Data analysis was performed with the Fluidigm digital PCR analysis software v4.1.2.

Mixes as well as thermal profiles used for the two PCR methods are also detailed in [10]. Boiled DNA (from 1 mL of overnight culture, boiled at 95 °C for 10 min and centrifuged at 10 000 rpm for 5 min) from CB-16230 and BfR-EC-19174 strains was used

as a positive control for *wzx*<sub>O26</sub> and *wzx*<sub>O2</sub>, respectively. A non-template control was also included.

### 2.6. MinION Sequencing of Artificially Co-Contaminated Milk

Two biological replicates of each enrichment were sequenced using the MinION platform except for Mix3 for which all replicates were sequenced. Libraries were prepared using the SLK-SQK109 ligation sequencing and the EXP-NBD104 or EXP-NBD114 barcoding kits, as previously described [10]. Five to six samples were loaded on FLO-MIN106 R9.4.1 flow cells and sequenced using the Mk1C device without live basecalling or demultiplexing.

### 2.7. Sequencing Data Analysis

Raw data (fast5 files) were basecalled using guppy\_basecaller v6.0.1 + 652ffd1 and the super high accuracy model including a q-score filter of 10. Data were demultiplexed using guppy\_barcode v6.0.1 + 652ffd1 and *-trim-adapters* and *-compress-fastq* parameters. Basecalled and demultiplexed data were processed using the STECmetadetector pipeline v0.1.2 [10] ([https://gitlab.com/bfr\\_bioinformatics/STECmetadetector](https://gitlab.com/bfr_bioinformatics/STECmetadetector), accessed on 26 July 2023). Extracted *E. coli* reads in fastq format were deposited in NCBI BioProject PRJNA982778. We assumed that the inoculated STEC strain was fully characterized when the two virulence genes *stx* and *eae* were co-located on the same contig. The figures were generated using ggplot2 v3.4.0, reshape2 v1.4.4, and patchwork v1.1.2 packages on R v4.1.2 and R studio v2021.9.0.351 [16–18]. The O-antigen clusters for O2 and O50 differ by a maximum of 2 potential SNPs [13,19] and are therefore considered identical and treated as a single molecular serogroup [20]. However, *wzy*<sub>O2</sub> and *wzy*<sub>O50</sub> (1 SNP difference, which cannot be reliably differentiated using MinION sequencing) have separate entries in the database included in the STECmetadetector pipeline. On the contrary, there is a single entry for *wzx*<sub>O2/O50</sub>. In the mapping analyses, reads mapping on *wzy*<sub>O2</sub> and *wzy*<sub>O50</sub> were thus combined.

## 3. Results

### 3.1. Estimation of the Inoculation Level Using Cell Counting and Post-Enrichment Quantification Using qdPCR

In order to determine the STEC:commensal ratio that would prevent the full characterization of the *eae*-positive STEC strain, we used five mixes to contaminate pasteurized milk samples. In Mix 1, we used an equal ratio of STEC and commensal strain before enrichment. In Mixes 2 and 3, we used a commensal strain 10 times and 100 times in excess compared to the STEC strain, respectively. In Mixes 4 and 5, we used a STEC strain 10 times and 100 times in excess compared to the commensal strain, respectively. Actual STEC and commensal strain inoculation levels used for the five mixes were determined using plate counting on TSA. Results are reported in Table 3. Each mix was used to artificially contaminate pasteurized milk before enrichment at 37 °C in acriflavine-supplemented BPW. To estimate the growth of the two strains during enrichment, we used qdPCR on their respective O-group genetic marker (*wzx*<sub>O26</sub> for the STEC and *wzx*<sub>O2</sub> for the commensal). Quantification of the two strains after enrichment is also reported (Table 3).

The results show that the conditions used here for enrichment (37 °C with acriflavine) seemed to favor the growth of the STEC O26 strain compared to the O2 commensal strain ( $8.62 \times 10^7$ – $5.94 \times 10^8$  copies.mL<sup>-1</sup>, mean =  $4.07 \times 10^8$  copies.mL<sup>-1</sup> for the STEC and  $3.74 \times 10^5$ – $2.29 \times 10^8$  copies.mL<sup>-1</sup>, mean =  $5.10 \times 10^7$  copies.mL<sup>-1</sup> for the commensal) regardless of the inoculation level of the commensal O2 strain ( $1.43 \times 10^2$  to  $1.39 \times 10^4$  CFU.mL<sup>-1</sup>). The O26 STEC strain always reached the  $10^8$  copies.mL<sup>-1</sup> threshold previously determined as the minimal quantity required for a single strain assembly [10]. The STEC strain was quantified approximately 1 to 3 logs higher than the commensal strain in all mixtures except Mix 3 (inoculation STEC:commensal ratio of 1:100) (Table 3). In this condition (Mix 3), the commensal strain was inoculated at a concentration 100 times higher than the STEC, but quantification results on O2 and O26 strains were similar post-enrichment



( $1.66 \times 10^8$  and  $3.86 \times 10^8$  copies.mL<sup>-1</sup> for O2 and O26, respectively). Although the O2 inoculation level corresponded to an O26:O2 ratio of 1:221 (Table 3), the STEC strain still appeared to grow to a ratio that reached 2:1 (O26:O2) (Table 3).

Table 3. Estimated inoculum (CFU.mL<sup>-1</sup>) and relative quantification (qdPCR, copies.mL<sup>-1</sup>) of the commensal and *eae*-positive STEC strain after enrichment of artificially co-contaminated milk.

Sample	Desired Inoculum O26:O2 Ratio	O2 Inoculum Estimated Using Cell Counting (CFU.mL <sup>-1</sup> )	O26 Inoculum Estimated Using Cell Counting (CFU.mL <sup>-1</sup> )	Estimated O26:O2 Ratio	Relative Quantification of O2 after Enrichment Using qdPCR ( <i>wzx</i> <sub>O2</sub> , copies.mL <sup>-1</sup> )	Relative Quantification of O26 after Enrichment Using qdPCR ( <i>wzx</i> <sub>O26</sub> , copies.mL <sup>-1</sup> )	O26:O2 Ratio Post-Enrichment
Pmilk_Mix1	1:1	$1.43 \times 10^2$	$6.30 \times 10^1$	1:2	$1.48 \times 10^6$	$2.59 \times 10^6$	175:1
Pmilk_Mix2	1:10	$1.39 \times 10^3$	$6.30 \times 10^1$	1:22	$3.55 \times 10^7$	$4.81 \times 10^6$	13:1
Pmilk_Mix3	1:100	$1.39 \times 10^4$	$6.30 \times 10^1$	1:221	$1.66 \times 10^8$	$3.86 \times 10^8$	2:1
Pmilk_Mix4	10:1	$1.43 \times 10^2$	$6.00 \times 10^2$	4:1	$4.08 \times 10^5$	$4.57 \times 10^8$	1119:1
Pmilk_Mix5	100:1	$1.43 \times 10^2$	$6.00 \times 10^3$	41:1	$2.53 \times 10^5$	$3.46 \times 10^8$	1367:1

Data in italics indicate that the number of cells was extrapolated using cell counting results of previous dilutions.

### 3.2. Characterization of the STEC Strain in Presence of Commensal *E. coli* Using STECmetadector

Two replicates of each condition and three replicates for the condition with equal levels of STEC:commensal post-enrichment (Mix 3) were sequenced using MinION sequencing, and data were analyzed using the STECmetadector pipeline.

The STECmetadector pipeline includes an initial mapping step to detect the presence of multiple O-groups in the set of *E. coli* reads. Figure 1 represents the read mapping depth on O-groups (*wzx* and *wzy* genetic markers) for O26 and O2. Mapping results were in line with quantification results (qdPCR) showing a higher number of reads mapped to O26 compared to O2. When the O2 strain was quantified below  $10^8$  copies.mL<sup>-1</sup> in the enriched samples (Mixes 1, 2, 4, and 5), only a few reads corresponding to O2 could be detected. The low read count attributable to the O2 strain explains the apparent ratio difference between qdPCR and mapping, especially in Mixes 4 and 5.

Assemblies of *E. coli* reads were performed using Flye included in the STECmetadector pipeline. The inoculated STEC strain was considered fully characterized when the two virulence genes *stx* and *eae* were co-localized on the same contig. Figure 2 shows that the *eae*-positive O26 STEC was characterized in 7/11 samples of artificially co-contaminated pasteurized milk. For all of those samples, the chromosome was in one contig with *stx*, *eae*, and serotype genetic markers (*wzx/wzy*<sub>O26</sub> and *fliC*<sub>H11</sub>) co-localized (Figure 2). The STEC strain was successfully characterized using the STECmetadector pipeline in all samples in which the O26 strain attained  $10^8$  copies.mL<sup>-1</sup> and the O26:O2 ratio (determined by qdPCR or mapping) was at least 10:1. The only exception was for one replicate of Mix 1 (Pmilk\_Mix1\_R1) in which O26 was quantified to  $8.62 \times 10^7$  copies.mL<sup>-1</sup> after enrichment (Supplementary File S1).

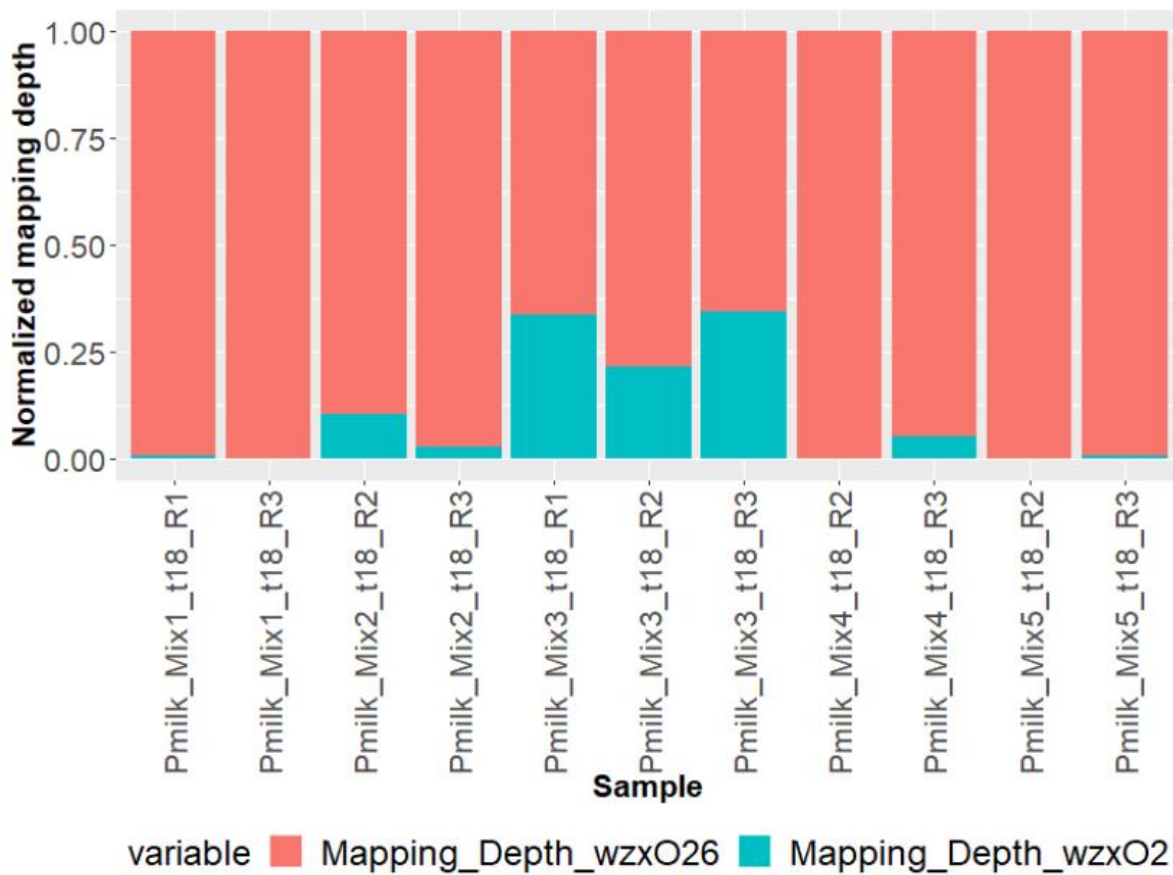
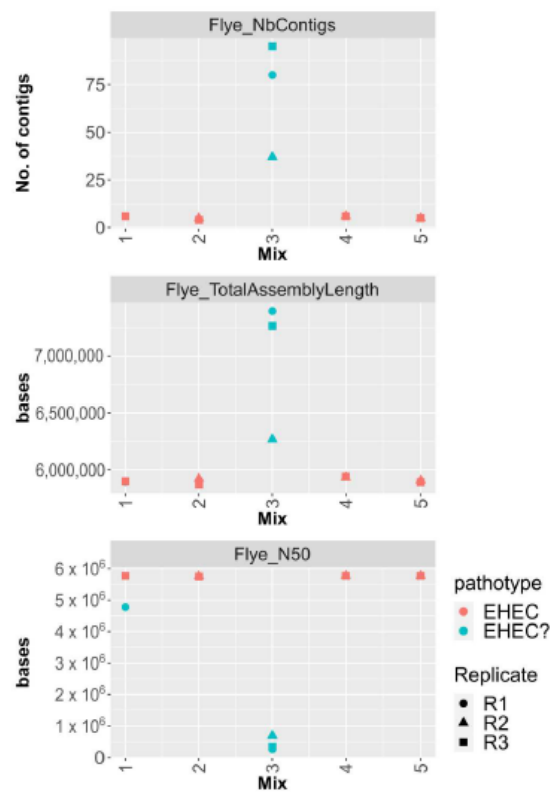


Figure 1. *E. coli* reads mapping depth on *wzx/wzy*<sub>O26</sub> and *wzx/wzy*<sub>O2</sub> genetic markers determined using STECmetadetector pipeline.

When the STEC and commensal strains were in equivalent quantities post-enrichment (i.e., Mix3), the O26 *eae*-positive STEC strain could not be characterized, even though it reached  $10^8$  copies.mL<sup>-1</sup>. All assemblies from these samples (Pmilk\_Mix3) were highly fragmented (Figure 2). In this condition, the STEC:commensal ratio after enrichment was 2:1 as quantified using qdPCR and below 10:1 as determined using mapping (Table 3 and Supplementary File S2).



**Figure 2.** Metrics of assemblies generated using Flye on extracted *E. coli* reads obtained from co-contamination of a commensal *E. coli* and STEC in pasteurized milk. The number of contigs, total assembly length (in bases), and the assembly N50 value are represented for each Mix. Mix 1 corresponds to co-contamination using equal proportions of the two strains and Mixes 2 and 3 to inoculation levels with 10 times and 100 times higher concentration of commensal, while Mixes 4 and 5 corresponds to inoculation levels with 10 times and 100 times higher concentration of STEC, respectively. The biological replicate (R1, R2, and R3) is represented by a specific symbol. A red color represents successful characterization of the STEC strain while a turquoise color represents when *stx* and *eae* were not identified on the same contig.

#### 4. Discussion

LEE-positive STEC constitute the STEC subgroup most frequently associated with severe human symptoms such as hemorrhagic colitis or HUS. Therefore, methods for detecting STEC in food samples often include the detection of both *stx* and *eae* genes by qPCR followed by the characterization of an isolated strain [4]. The characterization of STEC strains detected in contaminated food products is important in order to evaluate their pathogenic potential, but the isolation of a STEC strain from food products is often unsuccessful.

In previous work, we developed long-read metagenomics as an approach to characterize STEC strains from raw milk in an isolation-free manner. We have shown that STEC characterization directly from raw milk is possible but is hindered by the presence of background flora and multiple *E. coli* strains [10]. Cattle are asymptomatic carriers of STEC and can carry from 10 to 10<sup>9</sup> CFU of STEC per gram of feces [21,22]. The main source of milk contamination comes from cattle feces in which not only STEC but also different *E. coli* strains may be present simultaneously [23]. To the best of our knowledge, there is no study that assesses the ratio of STEC to other *E. coli* in milk samples or cattle feces samples. Although we have shown that long-read metagenomics is suitable for characterizing STEC in complex matrices such

as raw milk, it is important to assess the limit of background *E. coli* that would impede the characterization of the STEC using the STECmetadetector pipeline.

Here, we aimed at determining the ratio of commensal *E. coli* to STEC that would prevent the characterization of STEC. Background flora may limit the growth of the *E. coli* strains, which leads to lower output of *E. coli* reads. In our previous work, we used acriflavine, an antibiotic targeting Gram-positive bacteria, to handle the influence of background flora limiting *E. coli* strains growth [4,10]. In this study, even though we do not have background flora in pasteurized milk, we used the same enrichment conditions, including the use of acriflavine as it may potentially affect the growth of STEC and non-pathogenic *E. coli* differently [10,24].

We had previously determined that *eae*-positive STEC could be characterized using long-read metagenomics and an assembly-based approach from  $2.59 \times 10^8$  copies.mL<sup>-1</sup> post-enrichment in the absence of commensal *E. coli* flora in the milk [10]. The results obtained in this study were in line with what was determined before and allowed us to refine the characterization threshold. Overall, we estimated that a minimum of  $10^8$  copies.mL<sup>-1</sup> of STEC is required to achieve STEC characterization. Indeed, here, samples where the *eae*-positive STEC strain was in a 10 times excess over the commensal strain and quantified above  $10^8$  copies.mL<sup>-1</sup> post-enrichment were characterized. In addition, samples in which the commensal strain was below  $10^8$  copies.mL<sup>-1</sup> post-enrichment yielded very few reads corresponding to this strain, which then failed to assemble. The commensal strain reached the characterization threshold only when inoculated 100 times in excess of the STEC strain but resulted in a ratio of 2:1 STEC:commensal post-enrichment. As a consequence, it appears that the STEC:commensal ratio is another important factor to evaluate.

In all pasteurized milk samples artificially co-contaminated, the inoculated *eae*-positive STEC strain was characterized when it was quantified above  $10^8$  copies.mL<sup>-1</sup> post-enrichment and at least one log higher than the commensal strain. The exception was for one replicate inoculated with Mix 1 in which the estimated level of STEC determined using qdPCR was slightly below the  $10^8$  copies.mL<sup>-1</sup> threshold. The limiting condition for characterizing the STEC strain appears to be when the two strains grew to an equivalent ratio post-enrichment.

It is noteworthy that the enrichment conditions seem to affect the growth of the two *E. coli* strains differently. Indeed, the ratio of STEC:commensal has evolved drastically after enrichment from the inoculated levels. The STEC strain was quantified to higher levels than the commensal strain post-enrichment, even when inoculated 10 times less. Owing to their cell wall structure and the presence of the AcrB efflux pump, *E. coli* should not be sensitive to acriflavine [25]. In our previous work, we compared the growth of the STEC strain used in this study (6423-O26) in the presence and absence of acriflavine and observed no negative impact of acriflavine on the growth of this strain. It is, however, possible that the acriflavine used during enrichment affected the growth of the commensal *E. coli* strain (BfR-EC-19174). Some studies have already observed a strain-dependent effect of enrichment conditions [24,26,27]. Another option is a possible competition between the STEC and the commensal *E. coli* strain, with an advantage for the STEC growth. However, analysis of the genome sequences of both strains indicated the presence of multiple bacteriocins, although the level of expression of these bacteriocins was not examined [28].

Despite the possible influence of the enrichment conditions on the growth of the two strains, the limits that were determined in this study are more related to the post-enrichment steps, including DNA extraction, library preparation, data acquisition, and data processing [29]. It has been shown that the DNA extraction procedure can affect the structure of the population, especially in these particular conditions where high quantities of high-molecular-weight DNA are required (MinION sequencing from complex matrices) [30]. Similarly, according to the properties of the DNA (for example, GC content), the library preparation step performs differently [31–33]. Both DNA extraction and the performance of library preparation steps constitute the first bias that may lead to a different population structure compared to the original sample [29,34,35]. The second main bias is introduced by the sequencing technology and the data processing part. MinION sequencing does

not provide a homogenous quantity of data, read length, or quality across flow cells and samples. Some studies have demonstrated that high-GC species are under-represented in sequencing reads and contain a higher error rate [36,37]. In addition, the removal of the shortest reads and low-quality data during filtering steps might increase the bias in the population structure.

In this study, we used the depth of read mapping to the O-group genetic markers to estimate the O26:O2 ratio and check whether it differed from the ratio observed using qdPCR post-enrichment. The ratio determined using qdPCR and mapping was in a similar order of magnitude for samples of Mixes 1, 2, and 3, whereas the ratio observed for Mix 4 and Mix 5 were 100 and 10 times higher, respectively, using qdPCR (O26:O2 of 1000:1) than using the mapping approach (ratios of 10:1 and 100:1, respectively, for each mix). Considering the difference in proportion between the two strains, as estimated using qdPCR, the required amount of sequencing data could not be generated to reach a similar ratio of 1000:1. Our results suggest that the different post-enrichment steps performed here did not influence the O26:O2 ratio.

The main limitation of our approach concerns the assembly process. Indeed, when the two strains reached the required level for full characterization but were present at an equivalent ratio (Mix3 O26:O2 ratio below 10:1), the assembler failed to distinguish the STEC from the commensal strain. Current long-read assemblers (even those allowing metagenomics assemblies) are not able to differentiate different strains from the same species (less than 5% ANI divergence) since they share many genomic regions [38,39]. Although different strains will behave differently during the enrichment step, we showed that to be characterized, the amount of STEC data should be at least 10 times in excess compared to the commensal strain. It is important to note that six samples were multiplexed per MinION flow cell. Different multiplexing conditions would lead to different thresholds.

## 5. Conclusions

Overall, this study refined the limit of the long-read metagenomics approach developed in previous work to characterize *eae*-positive STEC in raw milk. In the conditions of this study, we observed that an *eae*-positive STEC strain can be characterized without the need to isolate the strain if the STEC strain can grow at least 10 times in excess compared to background *E. coli* strains post-enrichment, provided that the background flora does not prevent it from reaching the  $10^8$  copies.mL<sup>-1</sup>. Additionally, this study highlights the need for strain-aware assemblers or alternative approaches that would help identify different strains from a single species.

Under certain conditions, isolation-independent approaches are promising and could be applied in the case of ambiguous results that render decision making difficult, particularly when the STEC strain cannot be isolated. Many studies on STEC prevalence in cattle have been conducted but none on STEC prevalence compared to background *E. coli* [40,41]. Such studies would give insights into the application of long-read metagenomic approaches to characterize STEC among other *E. coli*.

**Supplementary Materials:** The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/microorganisms11082043/s1>. File S1: Relative quantification (qdPCR, copies.mL<sup>-1</sup>) of the commensal (O2) and *eae*-positive STEC (O26) strain after enrichment of artificially co-contaminated milk. File S2: Reads mapping depth on the commensal (O2) and on the *eae*-positive STEC (O26) genetic marker from the *E. coli* set of reads.

**Author Contributions:** S.D., J.G., P.F., B.M., E.S., A.G., F.V., C.D. and S.J. conceptualized the project. Methodology and resources by P.F., B.M., J.G., S.D., F.V., C.D., E.S. and A.G. Formal analysis by S.J. and C.D. S.J., M.-L.T. and C.S. conducted investigations. P.F. and B.M. were in charge of funding acquisition. S.J. wrote the original draft reviewed and validated by all authors. All authors have read and agreed to the published version of the manuscript.

**Funding:** This project was funded by the German Federal Institute for Risk Assessment (BfR) under grant no. 1322–765 and by the French Agency for Food, Environmental and Occupational Health and Safety (ANSES). The study was realized during the Ph.D. of S.J. (shared Ph.D. student between ANSES and BfR).

**Data Availability Statement:** The STECmetadetector pipeline is freely available at [https://gitlab.com/bfr\\_bioinformatics/STECmetadetector](https://gitlab.com/bfr_bioinformatics/STECmetadetector), accessed on 26 July 2023). The extracted *Escherichia coli* reads from each sample were deposited on NCBI under BioProject PRJNA982778 available at <http://www.ncbi.nlm.nih.gov/bioproject/982778>, accessed on 26 July 2023).

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

- Kaper, J.B.; Nataro, J.P.; Mobley, H.L. Pathogenic *Escherichia coli*. *Nat. Rev. Microbiol.* **2004**, *2*, 123–140. [CrossRef]
- Tobe, T.; Beatson, S.A.; Taniguchi, H.; Abe, H.; Bailey, C.M.; Fivian, A.; Younis, R.; Matthews, S.; Marches, O.; Frankel, G.; et al. An Extensive Repertoire of Type III Secretion Effectors in *Escherichia coli* O157 and the Role of Lambdoid Phages in Their Dissemination. *Proc. Natl. Acad. Sci. USA* **2006**, *103*, 14941–14946. [CrossRef]
- EFSA; 2020 EFSA BIOHAZ Panel (EFSA Panel on Biological Hazards); Koutsoumanis, K.; Allende, A.; Alvarez-Ordóñez, A.; Bover-Cid, S.; Chemaly, M.; Davies, R.; De Cesare, A.; Herman, L.; et al. Pathogenicity Assessment of Shiga Toxin-Producing *Escherichia coli* (STEC) and the Public Health Risk Posed by Contamination of Food with STEC. *EFSA J.* **2020**, *18*, 5967. [CrossRef]
- ISO/TS 13136; Microbiology of Food and Animal Feed—Real-Time Polymerase Chain Reaction (PCR)—Based Method for the Detection of Food-Borne Pathogens—Horizontal Method for the Detection of Shiga Toxin-Producing *Escherichia coli* (STEC) and the Determination of O157, O111, O26, O103, and O145 Serogroups. ISO (International Organization for Standardization): Geneva, Switzerland, 2012.
- Leonard, S.R.; Mammel, M.K.; Lacher, D.W.; Elkins, C.A. Application of Metagenomic Sequencing to Food Safety: Detection of Shiga Toxin-Producing *Escherichia coli* on Fresh Bagged Spinach. *Appl. Environ. Microbiol.* **2015**, *81*, 8183–8191. [CrossRef]
- Loman, N.J.; Constantinidou, C.; Christner, M.; Rohde, H.; Chan, J.Z.-M.; Quick, J.; Weir, J.C.; Quince, C.; Smith, G.P.; Betley, J.R.; et al. A Culture-Independent Sequence-Based Metagenomics Approach to the Investigation of an Outbreak of Shiga-Toxigenic *Escherichia coli* O104:H4. *JAMA* **2013**, *309*, 1502–1510. [CrossRef]
- Leonard, S.R.; Mammel, M.K.; Lacher, D.W.; Elkins, C.A. Strain-Level Discrimination of Shiga Toxin-Producing *Escherichia coli* in Spinach Using Metagenomic Sequencing. *PLoS ONE* **2016**, *11*, e0167870. [CrossRef]
- Buytaers, F.E.; Saltykova, A.; Denayer, S.; Verhaegen, B.; Vanneste, K.; Roosens, N.H.C.; Piérard, D.; Marchal, K.; De Keersmaecker, S.C.J. A Practical Method to Implement Strain-Level Metagenomics-Based Foodborne Outbreak Investigation and Source Tracking in Routine. *Microorganisms* **2020**, *8*, 1191. [CrossRef]
- Buytaers, F.E.; Saltykova, A.; Denayer, S.; Verhaegen, B.; Vanneste, K.; Roosens, N.H.C.; Piérard, D.; Marchal, K.; De Keersmaecker, S.C.J. Towards Real-Time and Affordable Strain-Level Metagenomics-Based Foodborne Outbreak Investigations Using Oxford Nanopore Sequencing Technologies. *Front. Microbiol.* **2021**, *12*, 738284. [CrossRef]
- Jaudou, S.; Deneke, C.; Tran, M.-L.; Schuh, E.; Goehler, A.; Vorimore, F.; Malomy, B.; Fach, P.; Grütze, J.; Delannoy, S. A Step Forward for Shiga Toxin-Producing *Escherichia coli* Identification and Characterization in Raw Milk Using Long-Read Metagenomics. *Microb. Genom.* **2022**, *8*, mgen000911. [CrossRef]
- Maguire, M.; Kase, J.A.; Roberson, D.; Muruvanda, T.; Brown, E.W.; Allard, M.; Musser, S.M.; González-Escalona, N. Precision Long-Read Metagenomics Sequencing for Food Safety by Detection and Assembly of Shiga Toxin-Producing *Escherichia coli* in Irrigation Water. *PLoS ONE* **2021**, *16*, e0245172. [CrossRef]
- Ørskov, F.; Ørskov, I. 2 Serotyping of *Escherichia coli* The Terminology Used to Describe the Different Classes of Bacterial Antigens Is Explained in the Preface. However, the Authors Would like to Mention That a Different Convention Is Used in Some Laboratories, for Example O:L, K:L, H:7 Is Equivalent to O1:K1:H7. In *Methods in Microbiology*; Bergan, T., Ed.; Academic Press: London, UK, 1984; Volume 14, pp. 43–112. ISBN 0580-9517.
- Delannoy, S.; Beutin, L.; Mariani-Kurkdjian, P.; Fleiss, A.; Bonacorsi, S.; Fach, P. The *Escherichia coli* Serogroup O1 and O2 Lipopolysaccharides Are Encoded by Multiple O-Antigen Gene Clusters. *Front. Cell. Infect. Microbiol.* **2017**, *7*, 30. [CrossRef]
- Sandra, J.; Mai-Lan, T.; Fabien, V.; Patrick, F.; Sabine, D. Hybrid Assembly from 75 *E. coli* Genomes Isolated from French Bovine Food Products between 1995 and 2016. *Microbiol. Resour. Announc.* **2023**, *12*, e01095-22. [CrossRef]
- Perelle, S.; Dilasser, F.; Grout, J.; Fach, P. Detection by 5'-Nuclease PCR of Shiga-Toxin Producing *Escherichia coli* O26, O55, O91, O103, O111, O113, O145 and O157:H7, Associated with the World's Most Frequent Clinical Cases. *Mol. Cell. Probes* **2004**, *18*, 185–192. [CrossRef]
- R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2018; Available online: <https://www.R-Project.org/> (accessed on 26 July 2023).
- Wickham, H. *Ggplot2: Elegant Graphics for Data Analysis*; Springer: New York, NY, USA, 2016; ISBN 978-3-319-24277-4.
- Wickham, H. Reshaping Data with the Reshape Package. *J. Stat. Soft.* **2007**, *21*, 1–20. [CrossRef]
- Iguchi, A.; Iyoda, S.; Kikuchi, T.; Ogura, Y.; Katsura, K.; Ohnishi, M.; Hayashi, T.; Thomson, N.R. A Complete View of the Genetic Diversity of the *Escherichia coli* O-Antigen Biosynthesis Gene Cluster. *DNA Res.* **2015**, *22*, 101–107. [CrossRef]

20. DebRoy, C.; Fratamico, P.M.; Yan, X.; Baranzoni, G.; Liu, Y.; Needleman, D.S.; Tebbs, R.; O'Connell, C.D.; Allred, A.; Swimley, M.; et al. Comparison of O-Antigen Gene Clusters of All O-Serogroups of *Escherichia coli* and Proposal for Adopting a New Nomenclature for O-Typing. *PLoS ONE* **2016**, *11*, e0147434. [[CrossRef](#)]
21. Chase-Topping, M.; Gally, D.; Low, C.; Matthews, L.; Woolhouse, M. Super-Shedding and the Link between Human Infection and Livestock Carriage of *Escherichia coli* O157. *Nat. Rev. Microbiol.* **2008**, *6*, 904–912. [[CrossRef](#)]
22. Stephens, T.P.; McAllister, T.A.; Stanford, K. Perineal Swabs Reveal Effect of Super Shedders on the Transmission of *Escherichia coli* O157:H7 in Commercial Feedlots. *J. Anim. Sci.* **2009**, *87*, 4151–4160. [[CrossRef](#)]
23. Cookson, A.L.; Biggs, P.J.; Marshall, J.C.; Reynolds, A.; Collis, R.M.; French, N.P.; Brightwell, G. Culture Independent Analysis Using *Gnd* as a Target Gene to Assess *Escherichia coli* Diversity and Community Structure. *Sci. Rep.* **2017**, *7*, 841. [[CrossRef](#)]
24. Baylis, C.L. Growth of Pure Cultures of Verocytotoxin-Producing *Escherichia coli* in a Range of Enrichment Media. *J. Appl. Microbiol.* **2008**, *105*, 1259–1265. [[CrossRef](#)]
25. Murakami, S. Multidrug Efflux Transporter, AcrB—The Pumping Mechanism. *Curr. Opin Struct. Biol.* **2008**, *18*, 459–465. [[CrossRef](#)]
26. Amagliani, G.; Rotundo, L.; Carloni, E.; Omiccioli, E.; Magnani, M.; Brandi, G.; Fratamico, P. Detection of Shiga Toxin-Producing *Escherichia coli* (STEC) in Ground Beef and Bean Sprouts: Evaluation of Culture Enrichment Conditions. *Food Res. Int.* **2018**, *103*, 398–405. [[CrossRef](#)]
27. Mancusi, R.; Trevisani, M. Enumeration of Verocytotoxigenic *Escherichia coli* (VTEC) O157 and O26 in Milk by Quantitative PCR. *Int. J. Food Microbiol.* **2014**, *184*, 121–127. [[CrossRef](#)]
28. Bosák, J.; Hrala, M.; Mícenková, L.; Šmajš, D. Non-Antibiotic Antibacterial Peptides and Proteins of *Escherichia coli*: Efficacy and Potency of Bacteriocins. *Expert Rev. Anti-Infect. Ther.* **2021**, *19*, 309–322. [[CrossRef](#)]
29. Pollock, J.; Glendinning, L.; Wisedchanwet, T.; Watson, M. The Madness of Microbiome: Attempting To Find Consensus “Best Practice” for 16S Microbiome Studies. *Appl. Environ. Microbiol.* **2018**, *84*, e02627-17. [[CrossRef](#)]
30. Goussarov, G.; Mysara, M.; Vandamme, P.; Van Houdt, R. Introduction to the Principles and Methods Underlying the Recovery of Metagenome-Assembled Genomes from Metagenomic Data. *Microbiologyopen* **2022**, *11*, e1298. [[CrossRef](#)]
31. Browne, P.D.; Nielsen, T.K.; Kot, W.; Aggerholm, A.; Gilbert, M.T.P.; Puetz, L.; Rasmussen, M.; Zervas, A.; Hansen, L.H. GC Bias Affects Genomic and Metagenomic Reconstructions, Underrepresenting GC-Poor Organisms. *Gigascience* **2020**, *9*, gaaa008. [[CrossRef](#)]
32. Sato, M.P.; Ogura, Y.; Nakamura, K.; Nishida, R.; Gotoh, Y.; Hayashi, M.; Hisatsune, J.; Sugai, M.; Takehiko, I.; Hayashi, T. Comparison of the Sequencing Bias of Currently Available Library Preparation Kits for Illumina Sequencing of Bacterial Genomes and Metagenomes. *DNA Res.* **2019**, *26*, 391–398. [[CrossRef](#)]
33. Ross, M.G.; Russ, C.; Costello, M.; Hollinger, A.; Lennon, N.J.; Hegarty, R.; Nusbaum, C.; Jaffe, D.B. Characterizing and Measuring Bias in Sequence Data. *Genome Biol.* **2013**, *14*, R51. [[CrossRef](#)]
34. Sevim, V.; Lee, J.; Egan, R.; Clum, A.; Hundley, H.; Lee, J.; Everroad, R.C.; Detweiler, A.M.; Bebout, B.M.; Pett-Ridge, J.; et al. Shotgun Metagenome Data of a Defined Mock Community Using Oxford Nanopore, PacBio and Illumina Technologies. *Sci. Data* **2019**, *6*, 285. [[CrossRef](#)]
35. Stevens, B.M.; Creed, T.B.; Reardon, C.L.; Manter, D.K. Comparison of Oxford Nanopore Technologies and Illumina MiSeq Sequencing with Mock Communities and Agricultural Soil. *Sci. Rep.* **2023**, *13*, 9323. [[CrossRef](#)]
36. Delahaye, C.; Nicolas, J. Sequencing DNA with Nanopores: Troubles and Biases. *PLoS ONE* **2021**, *16*, e0257521. [[CrossRef](#)]
37. Laver, T.; Harrison, J.; O'Neill, P.A.; Moore, K.; Farbos, A.; Paszkiewicz, K.; Studholme, D.J. Assessing the Performance of the Oxford Nanopore Technologies MinION. *Biomol. Detect. Quantif.* **2015**, *3*, 1–8. [[CrossRef](#)]
38. Kazantseva, E.; Donmez, A.; Pop, M.; Kolmogorov, M. StRainy: Assembly-Based Metagenomic Strain Phasing Using Long Reads. *bioRxiv* **2023**. [[CrossRef](#)]
39. Luo, X.; Kang, X.; Schönhuth, A. Enhancing Long-Read-Based Strain-Aware Metagenome Assembly. *Front. Genet.* **2022**, *13*, 868280. [[CrossRef](#)]
40. Ray, R.; Singh, P. Prevalence and Implications of Shiga Toxin-Producing *E. coli* in Farm and Wild Ruminants. *Pathogens* **2022**, *11*, 1332. [[CrossRef](#)]
41. Dewsbury, D.M.A.; Cemicchiaro, N.; Sanderson, M.W.; Dixon, A.L.; Ekong, P.S. A Systematic Review and Meta-Analysis of Published Literature on Prevalence of Non-O157 Shiga Toxin-Producing *Escherichia coli* Serogroups (O26, O45, O103, O111, O121, and O145) and Virulence Genes in Feces, Hides, and Carcasses of Pre- and Peri-Harvest Cattle Worldwide. *Anim. Health Res. Rev.* **2022**, *23*, 1–24. [[CrossRef](#)]

**Disclaimer/Publisher's Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

### Main results:

- An enrichment step is necessary to obtain enough STEC data (35x coverage) for the generation of a complete assembly using Flye.
- Acriflavine was more efficient to reduce the growth of background flora, particularly Gram-positive bacteria, than increasing temperature to 41.5°C.
- Isolation-free characterization of *eae*-positive STEC was successful from an inoculation level of 5 CFU.mL<sup>-1</sup> after enrichment at 37°C in BPW supplemented with acriflavine.
- Quantitative digital PCR showed that around 10<sup>8</sup> copies.mL<sup>-1</sup> of STEC are necessary post-enrichment for successful characterization.
- Generation of the STECmetadetector pipeline to facilitate STEC characterization from long-read metagenomics data.
- Assemblers failed to distinguish two *E. coli* strains (an *eae*-positive STEC and a commensal *E. coli*) when the ratio of STEC to commensal *E. coli* strain was below 10:1.

### Main conclusions:

- An enrichment step is necessary to obtain enough STEC data and the use of acriflavine helped to lower Gram-positive bacterial growth down.
- Characterization of *eae*-positive STEC using long-read metagenomics and an assembly-based approach is possible but necessitates high concentration of STEC post-enrichment and 10 times more STEC compared to additional *E. coli* strains.

### Perspectives:

- Test or develop assembly-free approaches less sensitive to the presence of multiple *E. coli* strains to identify *eae*-positive STEC from metagenomics data.
- Develop strain-aware assemblers to distinguish multiple strains.
- Test the developed method on naturally contaminated samples.
- Envisage DNA enrichment methods: DNA capture or Hi-C applications.



### Chapter5-3: The benefit of machine learning to identify *eae*-positive STEC from metagenomics data.

The long-read metagenomics approach developed in this project has shown the potential to characterize STEC strains including *eae*-positive STEC using an assembly-based method while also suggesting the limits (Chapter 5-2). Some researchers have tested long-read metagenomics to detect or identify STEC from different food matrices (beef, wastewater) (Buytaers *et al.*, 2021; Maguire *et al.*, 2021). The method developed by Maguire and co-workers to characterize STEC in wastewater was based on assembly (Maguire *et al.*, 2021). On the contrary, Buytaers and colleagues have tested an assembly-free approach based on *k*-mers for STEC identification from beef samples artificially and naturally contaminated with STEC strains. The authors first developed a method for short-read data, and expanded it for an application on long-read sequencing data (Buytaers *et al.*, 2021, 2020).

Although the assembly-based approach enables full characterization of STEC strains, I have shown that it is highly limited by the presence of background *E. coli* flora (Publication 3 and 4). The assembly-free approaches developed by other teams have shown the potential of long-read metagenomics to identify STEC present in complex matrices (Buytaers *et al.*, 2021). The challenge here is to ensure that the presence of virulence factors known to be carried by highly pathogenic STEC such as *stx* and *eae* genes are simultaneously present in the same strain. Previous studies based on high-throughput qPCR have demonstrated that the presence of additional genetic markers (such as *espK* and/or *espV*) can predict the presence of *eae*-positive STEC in the enrichment broth, which suggests the existence of a genetic background common to *eae*-positive STEC (Delannoy *et al.*, 2016).

Due to their capacity to acquire virulence factors, there is a large genetic diversity present in the *E. coli* genome. Pan-genomic analyses enabled the identification of the pool of genes common to all *E. coli* strains and the set of genes specific to particular groups (Burgaya *et al.*, 2023; Cummins *et al.*, 2022; Hochhauser *et al.*, 2023; Yang and Gao, 2022). Different pan-genomic analyses combined with machine learning algorithms have been conducted on *E. coli* genomes to predict the clinical outcome of STEC or even determine their isolation source (Im *et al.*, 2021; Lupolova *et al.*, 2021; Njage *et al.*, 2019). Here, we aimed at using pan-genomic analysis and machine learning algorithms to identify *eae*-positive STEC strains using their genetic signature.

In this work, we have compiled a database of 1 425 complete (or almost complete) *E. coli* genomes retrieved from NCBI belonging to different *E. coli* pathotypes. Each genome was annotated according to the annotation of the Sakai strain reference genome (Hayashi *et al.*, 2001). A pan genome program was used to score all genes found in all 1 425 genomes. Each gene was noted either present or absent in each genome. The resulting matrix was used as input for machine learning (ML) analysis using eight algorithms to analyze the presence of genetic regions in *eae*-positive STEC compared to other *E. coli*. Six genetic markers were found to be relevant for identifying *eae*-positive STEC. The combination (presence/absence) of these six

genetic markers allows the identification of *eae*-positive STEC. Interestingly, two of these genetic markers were carried by O-islands from which previous markers were found to be present in *eae*-positive STEC (*espK*, *espV*) (Publication 5).

To confirm that these genetic markers could identify the presence of *eae*-positive STEC among other *E. coli*, we used different datasets. For all tests, predictions were made on Flye assemblies using the metagenome parameter. *In silico* mixtures of various *E. coli* MinION reads at a ratio of 1:1 and coverage varying from 10 to 70x were assembled and used to predict the presence of *eae*-positive STEC. The raw milk samples artificially contaminated with an *eae*-positive STEC alone (Publication 3) or in combination with a commensal *E. coli* using increasing ratios (Publication 3 and Publication 4) were also used. All results were positive for the presence of *eae*-positive STEC, when present. This study has shown that the combination of different genetic markers (maximum of 6) could predict the presence of *eae*-positive STEC among other *E. coli* strains (Publication 5).

This approach is based on assembly and enables the identification of *eae*-positive STEC in presence of other *E. coli* at a ratio of 1:1 even at low coverage (10x) whereas the previous assembly-based approach that we have developed was limited by the presence of multiple *E. coli* at a ratio of 10:1 with excess of *eae*-positive STEC. Perceptively, these results could be used to develop a qPCR for the identification of *eae*-positive STEC in metagenomes together with an algorithm interpreting the PCR results.



## Publication 5

### **Combination of whole genome sequencing and supervised machine learning provides unambiguous identification of *eae*-positive Shiga toxin-producing *Escherichia coli***

Vorimore Fabien<sup>†</sup>, Jaudou Sandra<sup>†</sup>, Tran Mai-Lan, Richard Hugues, Fach Patrick and Delannoy Sabine

<sup>†</sup>These authors have contributed equally to this work

Frontiers in microbiology, Section food microbiology, Volume 14 - 2023 |  
<https://doi.org/10.3389/fmicb.2023.1118158>



#### OPEN ACCESS

##### EDITED BY

Abani Kumar Pradhan,  
University of Maryland, College Park,  
United States

##### REVIEWED BY

Zachary R. Stromberg,  
Pacific Northwest National Laboratory (DOE),  
United States  
Patrick Murigu Kamau Njage,  
Technical University of Denmark, Denmark

##### \*CORRESPONDENCE

Fabien Vorimore  
✉ fabien.vorimore@anses.fr

<sup>†</sup>These authors have contributed equally to this work

RECEIVED 07 December 2022

ACCEPTED 21 April 2023

PUBLISHED 12 May 2023

##### CITATION

Vorimore F, Jaudou S, Tran M-L, Richard H, Fach P and Delannoy S (2023) Combination of whole genome sequencing and supervised machine learning provides unambiguous identification of *eae*-positive Shiga toxin-producing *Escherichia coli*. *Front. Microbiol.* 14:1118158. doi: 10.3389/fmicb.2023.1118158

##### COPYRIGHT

© 2023 Vorimore, Jaudou, Tran, Richard, Fach and Delannoy. This is an open-access article distributed under the terms of the [Creative Commons Attribution License \(CC BY\)](https://creativecommons.org/licenses/by/4.0/). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

# Combination of whole genome sequencing and supervised machine learning provides unambiguous identification of *eae*-positive Shiga toxin-producing *Escherichia coli*

Fabien Vorimore<sup>1,\*†</sup>, Sandra Jaudou<sup>1,2†</sup>, Mai-Lan Tran<sup>1,2</sup>, Hugues Richard<sup>3</sup>, Patrick Fach<sup>1,2</sup> and Sabine Delannoy<sup>1,2</sup>

<sup>1</sup>ANSES, Laboratory for Food Safety, Genomics Platform IdentityPath, Maisons-Alfort, France, <sup>2</sup>ANSES, Laboratory for Food Safety, COLIPATH Unit, Maisons-Alfort, France, <sup>3</sup>Bioinformatics Unit, Genome Competence Center (MF1), Robert Koch Institute, Berlin, Germany

**Introduction:** The objective of this study was to develop, using a genome wide machine learning approach, an unambiguous model to predict the presence of highly pathogenic STEC in *E. coli* reads assemblies derived from complex samples containing potentially multiple *E. coli* strains. Our approach has taken into account the high genomic plasticity of *E. coli* and utilized the stratification of STEC and *E. coli* pathogroups classification based on the serotype and virulence factors to identify specific combinations of biomarkers for improved characterization of *eae*-positive STEC (also named EHEC for enterohemorrhagic *E. coli*) which are associated with bloody diarrhea and hemolytic uremic syndrome (HUS) in human.

**Methods:** The Machine Learning (ML) approach was used in this study on a large curated dataset composed of 1,493 *E. coli* genome sequences and 1,178 Coding Sequences (CDS). Feature selection has been performed using eight classification algorithms, resulting in a reduction of the number of CDS to six. From this reduced dataset, the eight ML models were trained with hyper-parameter tuning and cross-validation steps.

**Results and discussion:** It is remarkable that only using these six genes, EHEC can be clearly identified from *E. coli* read assemblies obtained from in silico mixtures and complex samples such as milk metagenomes. These various combinations of discriminative biomarkers can be implemented as novel marker genes for the unambiguous EHEC characterization from different *E. coli* strains mixtures as well as from raw milk metagenomes.

##### KEYWORDS

machine learning, Shiga toxin-producing *Escherichia coli*, food safety, metagenomics, raw milk

## 1. Introduction

Shiga toxin-producing *Escherichia coli* (STEC) are important zoonotic pathogens comprising more than 400 serotypes (Beutin and Fach, 2015). Pathogenic STEC strains such as enterohemorrhagic *E. coli* (EHEC) may cause hemorrhagic colitis (HC) and hemolytic-uremic syndrome (HUS) in humans. However, it remains difficult to fully define human pathogenic STEC or identify virulence factors for STEC that clearly foresee their capacity to

cause human disease (European Food Safety Authority and European Centre for Disease Prevention and Control, 2021). The production of Shiga toxin (*stx* genes) by highly pathogenic STEC (i.e., EHEC) is the major virulence factor responsible for HUS, but many *E. coli* strains that produce Shiga toxin do not cause HUS. Therefore, the identification of virulent STEC strains based solely on the presence of *stx* genes may be misleading. Shiga toxins comprise a growing family of genes with a vast type diversity (Scheut et al., 2012). The *Stx* family splits into two major branches, *Stx1* and *Stx2*, which are immunologically not cross-reactive and show about 55% difference in their amino acid sequences (Müthing et al., 2009). In addition to producing one or both types of Shiga toxin, typical EHEC strains harbor a genomic pathogenicity, called the “locus of enterocyte effacement” (LEE). This locus was first identified in enteropathogenic *E. coli* (EPEC), a leading cause of infant diarrhea in developing countries. The LEE carries genes encoding proteins involved in the pathogenicity of *E. coli* strains, as they participate in bacterial colonization of the gut and destruction of the intestinal mucosa (Nataro and Kaper, 1998). For example, the intimin-encoding gene (*eae*) is directly involved in the attaching and effacing (A/E) process and serves as an indicator for the A/E function in the bacteria (Zhang et al., 2002). As mentioned above, prediction of STEC pathogenicity using available markers is challenging, but strains positive for Shiga toxin (in particular the *stx2* genes) and *eae* (intimin production) genes have been shown to be associated with a higher risk of causing more severe illness than other virulence factor combinations (European Food Safety Authority, 2007, 2013). STEC are traditionally considered to be zoonotic pathogens that are primarily food- and water-borne, with the main reservoir being the digestive tract of mammals, particularly ruminants (Gill et al., 2022). Consumption of contaminated food, such as undercooked ground meat and unpasteurized dairy products, is the principal source of infection. Current methods for EHEC identification in feed and food samples rely on the molecular detection of *stx*, *eae*, and the top five or top seven EHEC serogroups, followed by strain isolation, as described in the ISO/TS 13136:2012 (EU) and MLG5C.02 (US) reference methods (International Organization for Standardization, 2012; European Food Safety Authority and European Centre for Disease Prevention and Control, 2021). The strain isolation step is necessary to demonstrate that both genes are present in the same strain. Indeed, the major challenge for EHEC identification based on *stx* and *eae* genes detection is that these genes are located on mobile genetic elements, and can be carried by non-pathogenic *E. coli* strains simultaneously present in the food matrix, as well as other *Enterobacteriaceae* (Herold et al., 2004) or even free bacteriophages (Imamovic et al., 2009). The high rates of unconfirmed presumptive positive results observed in food safety tests are a global challenge for the regulatory agencies and industry quality control laboratories performing STEC testing (Delannoy et al., 2016, 2022). It remains a desirable goal for the industry and decision makers to develop cost-effective sensitive detection tests that can guaranty the highest level of food safety. Our objective here was to refine the EHEC diagnostic systems for better identification and characterization of highly pathogenic STEC from any kind of food samples. This work was based on the hypothesis that the co-occurrence of the *stx* and *eae* genes in the same genome would imply the presence of other (variable) genes

and should create complex genetic signatures. We took advantage of a Genome Wide Association Study program (GWAS) to explore a large number of *E. coli* assemblies available from public databases (Franz et al., 2014) and generated a complex matrix summarizing presence and absence for groups of orthologous genes. Machine learning (ML) methods perform admirably in detecting predictive patterns hidden within high dimensional data (Lupolova et al., 2016; Moradigaravand et al., 2018). Supervised learning was used to create ML models that can precisely predict the co-occurrence of *stx* and *eae* genes in a genome or an assembly. After testing on simple *in silico* mixtures of strains, we successfully applied these models on long-read metagenomic sequencing data of artificially *eae*-positive STEC contaminated raw milk samples.

## 2. Materials and methods

### 2.1. Genomic data collection

Available *E. coli* genomes ( $n = 31,230$ ) were retrieved from the GenBank database during the database construction. Based on the genome sequence completeness (full *E. coli* genomes were included in priority), the country of isolation and the *E. coli* pathotype, 1,425 genomes were selected to maximize the diversity. Sixty-eight additional genomes sequenced and assembled in a previous study by Jaudou and colleagues (Jaudou et al., 2023), were added to reach a total of 1,493 genomes. The genome accession numbers and the metadata associated with the selected genomes are reported in Supplementary Table 1. All the genomes were screened against a custom database (Available at [https://github.com/fabgenomics/ML\\_EHEC](https://github.com/fabgenomics/ML_EHEC)) containing all the *stx* subtypes, *eae* and O-group genes using abricate v1.0.1. (<https://github.com/tseemann/abricate>). The phylogroup of each genome was determined using the EzClermont phylotyping tool available at <https://github.com/nickp60/EzClermont>.

### 2.2. Genome annotation and classification

The selected genomes were annotated using the rapid prokaryotic genome annotation software prokka v1.13.3 (Seemann, 2014) using the `proteins` option with the reference genome of *E. coli* O157:H7 str. Sakai (NC 002695.2) (Hayashi et al., 2001). Resulting General Feature Format (.gff) files were processed through a Pangenome analysis pipeline using panaroo v1.2.7 with `-clean-mode strict`, `-remove invalid-genes` and `-merge paralogs` options (Tonkin-Hill et al., 2020). Panaroo collapses genes into putative families with a family sequence identity level of 70% by default and creates groups. The `gene_presence_absence.Rtab` table provided by the panaroo output contains all the groups and genes and the presence/absence information from each genome. These groups and genes were renamed using a custom script (Available at [https://github.com/fabgenomics/ML\\_EHEC](https://github.com/fabgenomics/ML_EHEC)) in which we used the information in the panaroo output file `matrix (gene_presence_absence.csv)` available at <https://doi.org/10.5281/zenodo.7129021>) to retrieve the corresponding locus tag number from the *E. coli* O157:H7 Sakai

strain annotation (ECs number) relative to the group or the gene (CDS). Only groups renamed with the locus tag number by our custom script were retained for further analysis. We created a CDS presence/absence matrix (ECs\_presence\_absence.csv available at <https://doi.org/10.5281/zenodo.7129021>), by adding a new column based on the pathotype of each genome. Genomes that were found to be *stx+/eae+* were assigned to the EHEC pathotype and noted 1 in the table and all the other genomes (STEC for *stx+/eae-*, EPEC for *stx-/eae+* and the abbreviation COM was used for *stx-/eae-*) were considered non-EHEC pathotype and noted 0. A phylogenetic tree was reconstructed using IQtree v2.0.3 (Minh et al., 2020) with the Generalised Time Reversible (GTR) model on the core genome alignment produced by panaroo. The tree was annotated with the molecular serogroups using the CLC genomic workbench v21 (QIAGEN, Aarhus, Denmark).

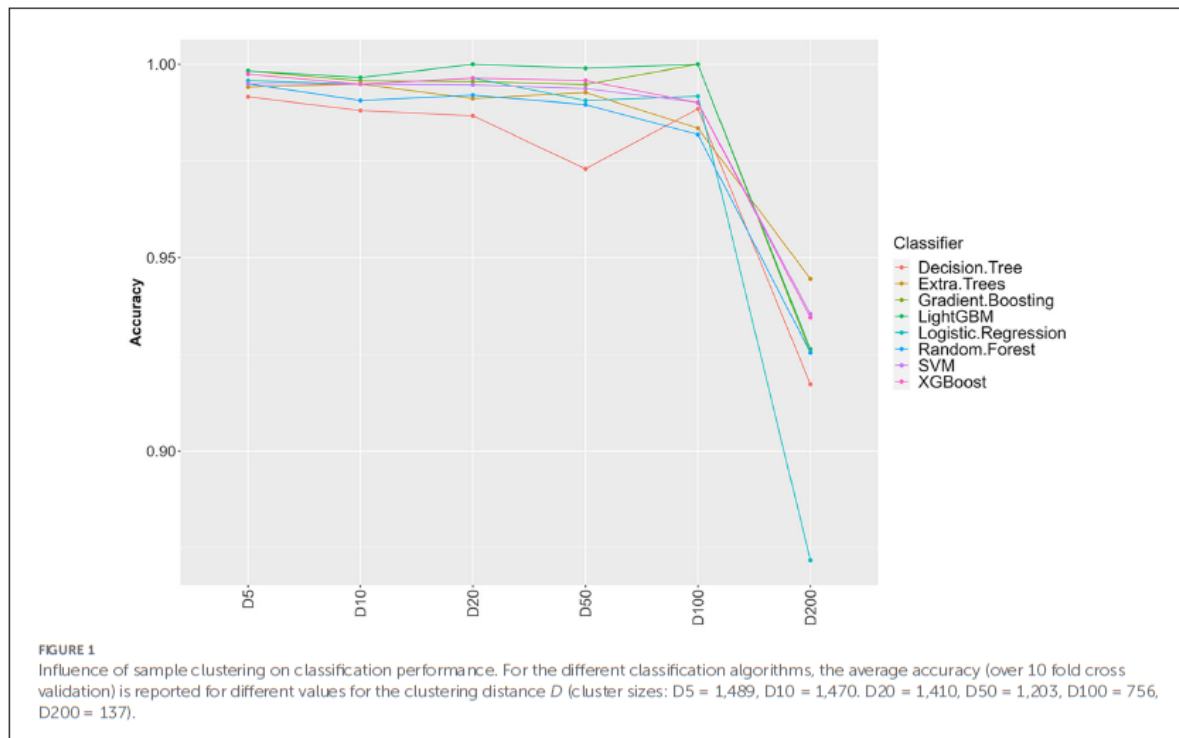
### 2.3. Machine learning model training and evaluation

Before evaluating the performance of the different classifiers, the CDS presence/absence matrix was filtered on non-informative features. Any CDS with less than 10% variance on its presence/absence vector was removed, as these loci do not contain useful information to test machine learning algorithms. This step removed 3,603 CDS, resulting in a dataset with 1,178 CDS. Then, to avoid possible data leakage between training and testing datasets, we grouped the samples based on their similarity. For each pair of samples, their CDS presence/absence vector was used to compute a hamming distance (*i.e.* the number of differences). Any two samples with a hamming distance lower than or equal to  $D$  were allocated to the same cluster. We considered possible values for  $D$  of 5, 10, 50, 100 and 200. Note that the resulting clusters have extremely homogeneous pathotypes (at  $D = 100$ , only one cluster consists of a mixture of EHEC and non-EHEC samples). The cluster table is available at [https://github.com/fabgenomics/ML\\_EHEC](https://github.com/fabgenomics/ML_EHEC). The main dataset ( $n = 1,493$ ) was then subsampled to keep only one genome from each cluster. Then, each subsampled dataset was randomly split using 80% of the samples for training/validation and the remaining 20% for testing. The `train_test_split` module from Sklearn was used, with stratify option to control the proportion of EHEC in both datasets. Eight classification algorithms were trained on each dataset using 10-fold Cross-Validation: Decision Tree, Extra Tree, Gradient Boosting, LGBMClassifier, Logistic Regression, Random Forest, XGBClassifier and Support Vector Machine. The evaluation metric used was the function `cross_val_score` from the sklearn library. For all the cross-validation scores (10 folds *i.e.* 10 scores), the mean accuracy was calculated (Supplementary Table 2 and Figure 1). Further analysis were performed using a distance  $D = 100$  for clustering (dataset Cluster-D100). This implies that two samples in this dataset differ by at least 8.4% (100/1178 CDS) of their gene content. A module from Sklearn library v0.23.1 (`RandomUnderSampler`) was used to select randomly the same amount of non-EHEC genomes to be equal to the number of remaining EHEC genomes from the cluster analysis. The Cluster-D100 dataset was randomly split with

a ratio of 80/20% for training and testing datasets respectively and the stratify option. Eight classification algorithms were used to select the most important features with the `SelectFromModel` library from Sklearn. The most important CDS are listed in Table 1. We arbitrarily chose to select the six most important features to create a new reduced dataset. With this resulting matrix, hyper-parameter tuning was done on each of the eight models using `RandomizedSearchCV` and `GridSearchCV` (scoring on `roc_auc` metric) and cross-validation steps ( $n = 5$ ) when the option was available. Finally, each classifier was retrained with its best hyper-parameter and evaluated on the testing dataset previously set aside (accuracy, precision, recall and F1-score are obtained using the `classification_report` from Sklearn, see Supplementary Table 3). To understand which gene combination led to the prediction of the EHEC pathotype, we generated all  $2^6 = 64$  combinations of the six genes presence/absence and computed, for each ML model, the probability of the EHEC pathotype. We then kept the cases where the probability was  $\geq 0.7$  and transformed the set of gene combinations into simplified boolean expressions (using a boolean Algebra Solver). The results are reported on Figure 2 Charts of the training pipeline and the prediction pipeline are presented in Figures 3A, B, respectively.

### 2.4. Evaluation of the eight models on *in silico* mixtures of *E. coli*

From a previous study conducted by Jaudou et al. (2023), raw ONT MinION reads from two STEC (ECA279 (O174:H2) = SRR18191627 and 97HMPL652 (O110:H9) = SRR18191587), one *stx*-negative *eae*-positive *E. coli* (2142-O103 (O103:H25) = SRR18191529), one *eae*-positive STEC (*E. coli* 12-1 (O157:H7) = SRR18191640) and one commensal *E. coli* [*i.e.*, negative for both *stx* and *eae* (NC809 (O41:H7) = SRR18191621)] were collected. One *stx*-negative *eae*-positive *E. coli* strain (KK072/05 - O156:H8) was newly sequenced during this study following the same protocol described by Jaudou et al. (Jaudou et al., 2023) and the raw ONT MinION reads were deposited to the NCBI database under the number SRX18376762. Raw reads were subsampled using Rasusa v0.6.0 (Hall, 2022) with 5.5Mb for the target number of bases and 10, 20, 30, 40 and 50x for the coverage. From these subsampled reads, twenty-five *in silico* mixtures were generated by concatenating into a 1:1 ratio the same coverage of subsampled reads. Details of the different mixtures are presented in the Table 2. From these mixtures, an assembly was generated using metaFlye v2.9-b1768 (Kolmogorov et al., 2020) with `nano-r` and `meta` parameters. The resulting assemblies were annotated with the same parameters as described in the Genome annotation and classification paragraph (Section 2.2). The produced GFF file of each annotation was integrated individually in the pangenome graph generated with the 1,493 genomes using the `panaroo-integrate` command from the panaroo program. From the new `gene_presence_absence.csv` and the `gene_presence_absence.Rtab` generated by the `panaroo-integrate` command, the row corresponding to the added mixture was extracted using a newly developed python script (Available



**TABLE 1** Top six most important features extracted from the training of the eight classification models.

Rank	Locus_tag	Gene ID	Gene name	Encoded protein	Number of models using the gene for EHEC prediction
1	ECs_1056	62675958	-	Phage excisionase	8
2	ECs_1812	912909	<i>nleA/espI</i>	T3SS secreted effector NleA/EspI	7
3	ECs_1824	912888	<i>nleG</i>	T3SS secreted effector NleG	5
4	ECs_3858	916318	<i>nleE</i>	T3SS secreted effector NleE	4
5	ECs_1815	912903	<i>nleF</i>	T3SS secreted effector NleF	4
6	ECs_1561	913337	<i>espN</i>	T3SS secreted effector EspN	4

The Gene ID of each locus is provided, and when known, the gene name as well.

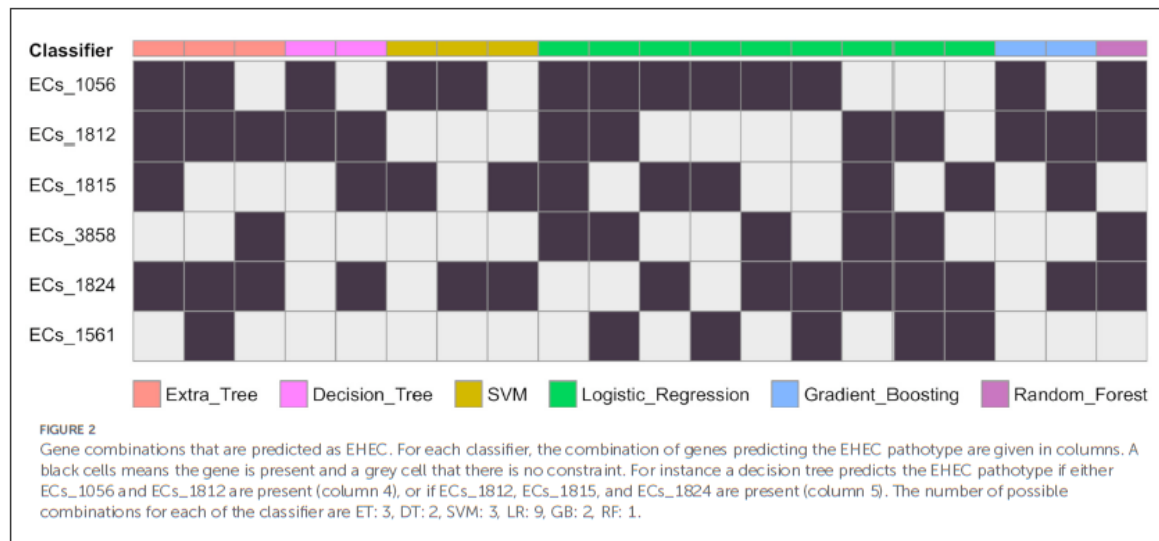
at [https://github.com/fabgenomics/ML\\_EHEC](https://github.com/fabgenomics/ML_EHEC)) to reconstruct the CDS presence/absence table. We extracted only the features required for the tested model and when a feature was absent, we created it and introduced a 0 value. The data extracted were used to perform the predictions (Figure 3B). The `predict_proba` method from all the algorithms was used to estimate the probability that the sample is an EHEC.

## 2.5. Evaluation of the eight models on experimentally-contaminated raw milk

Eight metagenomes from artificially contaminated raw milks described in a previous study (Jaudou et al., 2022) were

downloaded from the Genbank public database (Table 3). The estimated level of contamination was  $0.5 \times 10^3$ ,  $0.5 \times 10^2$  and  $0.5 \times 10^1$  CFU.mL<sup>-1</sup> of EHEC O26 plus one EHEC-free milk. Raw reads were processed using the STECmetadetector pipeline developed by Jaudou et al. (2022) available at [https://gitlab.com/Bfr\\_bioinformatics/STECmetadetector](https://gitlab.com/Bfr_bioinformatics/STECmetadetector) and the extracted *E. coli* reads were assembled using metaFlye v2.9-b1768 with the same parameters as described in Section 2.4. The resulting assemblies were annotated with the same parameters as described in the Genome annotation and classification paragraph (Section 2.2). The resulting GFF file was treated with the same process than the *in-silico* mixture GFF file and the EHEC predictions were performed, as described in Section 2.3 (Figure 3B).





### 3. Results

#### 3.1. *Escherichia coli* pathotype assignment based on genomic information

To take advantages of ML to find patterns and preserve its generalization potential, we constituted an *E. coli* database for which we selected in priority complete *E. coli* genomes with a minimum of required metadata (origin, isolation date and location) and verified their pathotypes (*stx* and *eae* genes presence). A total of 1,493 genomes were downloaded from the GenBank database. The geographic origin of the strains was 33, 26, 23, 7, and 3% from Europe, Asia, America, Africa and Oceania, respectively and the 8% remaining were missing the country of origin. During the genome selection, we tried to respect an equal proportions of *stx/eae*-positive strains (i.e., EHEC) and non-EHEC strains based on the metadata provided. Genomes simultaneously positive for at least one *stx* gene and the *eae* gene ( $n = 632$ ) were assigned to the EHEC group. The other genomes ( $n = 861$ ) were assigned to the non-EHEC group (Available at <https://doi.org/10.5281/zenodo.7129021>). In addition, the custom database was reporting all the O-group sequences so that the serogroup information, in particular the most frequent EHEC serogroup, was available (Supplementary Table 1). The top seven most represented serogroups were O26 ( $n = 160$ ), O157 ( $n = 126$ ), O103 ( $n = 66$ ), O121 ( $n = 61$ ), O145 ( $n = 60$ ), O111 ( $n = 36$ ), and O45 ( $n = 9$ ). These serogroups comprised mostly EHEC strains with 126 EHEC O26 strains, 124 EHEC O157 strains, 57 EHEC O103 strains, 55 EHEC O121 strains, 59 EHEC O145 strains, and 31 EHEC O111 strains. The phylogroup analysis showed that the diversity of the species is well represented. The phylogeny of the final dataset is illustrated in Figure 4.

#### 3.2. Generation of the input dataset

From 37,380 groups generated by panaroo, 13,952 remained with an ECs annotation. Some ECs were duplicated in the table due

to the genetic diversity of some genes and the panaroo identity level threshold. We aggregated the results for these genes, which resulted in 4,780 unique CDS and kept the presence/absence information. Before splitting the ECs presence/absence table (Available at <https://doi.org/10.5281/zenodo.7129021>), we first selected CDS with enough variation between the samples (see Section 2). This filtering step removed 3,602 CDS resulting in a dataset containing 1,178 CDS. To avoid optimistic performance estimates that would result from near duplicate samples both in training and testing sets, we performed a clustering based on gene content similarity (see Section 2). Clustering at distance  $D$  ensures that two clusters in the dataset have at least  $D$  genes that are different. We evaluated the change in accuracy for increasing values of  $D$ , starting from 5 and up to 200 (Figure 1). Up to  $D = 100$  (8.4% of the CDS), the accuracy of all classifiers remains very high (above 97%). This indicates that robust information can be extracted from gene profiles to predict EHEC pathotype with high accuracy. The performances drop by around 5–10% for  $D = 200$  for an accuracy of around 93%. While still high, this decrease could be attributed to the low number of clusters that remain for prediction ( $n = 137$ , 15 EHEC and 122 non-EHEC). From these results, a value of  $D = 100$  was chosen to build the dataset for further analysis, as it combines both good performance with a sufficient sample size ( $n = 756$ ). However, this dataset was imbalanced. To avoid overfitting problems, we subsampled the non-EHEC class to be equal to the EHEC class and ended with a balanced dataset (see Section 2). The matrix was then randomly split into a training dataset containing 80% of the genomes used in the study ( $n = 139$ ) with 69 non-EHEC and 70 EHEC. The testing dataset contained the remaining 20% ( $n = 35$ ) with 18 non-EHEC and 17 EHEC.

#### 3.3. CDS selection using eight ML classifiers

We conducted a comprehensive analysis of the top features extracted from the training of eight classification models, and

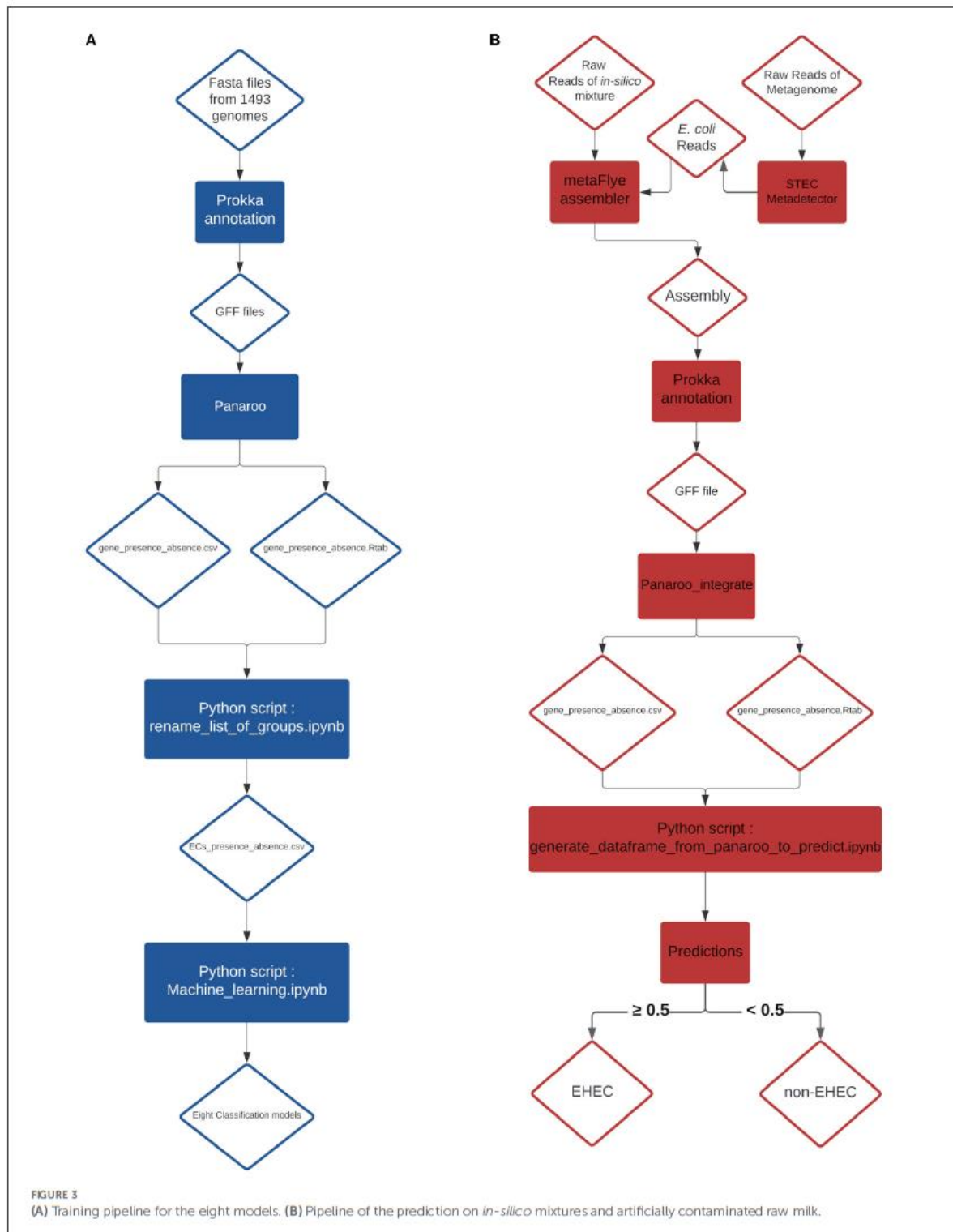


TABLE 2 Prediction of the class probabilities on the 25 generated mixtures of pure *E. coli* cultures.

Strains and genome coverage used for <i>in-silico</i> mixture	<i>E. coli</i> Pathotype mixture*	Class**	LGBM	LR	DT	XGB	RF	SVM	GB	ET
ECA279 + NC809 10x	STEC-COM	0	0.00	0.02	0.00	0.00	0.04	0.01	0.02	0.03
ECA279 + NC809 20x	STEC-COM	0	0.00	0.02	0.00	0.00	0.04	0.01	0.02	0.03
ECA279 + NC809 30x	STEC-COM	0	0.00	0.02	0.00	0.00	0.04	0.01	0.02	0.03
ECA279 + NC809 40x	STEC-COM	0	0.00	0.02	0.00	0.00	0.04	0.01	0.02	0.03
ECA279 + NC809 50x	STEC-COM	0	0.00	0.02	0.00	0.00	0.04	0.01	0.02	0.03
97HMPL652 + 2142-O103 10x	STEC-EPEC	0	0.22	0.27	0.00	0.27	0.46	0.21	0.02	0.47
97HMPL652 + 2142-O103 20x	STEC-EPEC	0	0.22	0.27	0.00	0.27	0.46	0.21	0.02	0.47
97HMPL652 + 2142-O103 30x	STEC-EPEC	0	0.98	0.96	1.00	1.00	0.87	1.00	0.98	0.88
97HMPL652 + 2142-O103 40x	STEC-EPEC	0	0.73	0.72	0.75	0.75	0.73	0.88	0.98	0.77
97HMPL652 + 2142-O103 50x	STEC-EPEC	0	0.73	0.72	0.75	0.75	0.73	0.88	0.98	0.77
97HMPL652 + KK072/05 10x	STEC-EPEC	0	0.05	0.20	0.00	0.01	0.26	0.01	0.02	0.25
97HMPL652 + KK072/05 20x	STEC-EPEC	0	0.05	0.20	0.00	0.01	0.26	0.01	0.02	0.25
97HMPL652 + KK072/05 30x	STEC-EPEC	0	0.05	0.20	0.00	0.01	0.26	0.01	0.02	0.25
97HMPL652 + KK072/05 40x	STEC-EPEC	0	0.05	0.20	0.00	0.01	0.26	0.01	0.02	0.25
97HMPL652 + KK072/05 50x	STEC-EPEC	0	0.05	0.20	0.00	0.01	0.26	0.01	0.02	0.25
Ecolli2-1 + NC809 10x	EHEC-COM	1	1.00	0.99	1.00	1.00	0.95	1.00	0.98	0.98
Ecolli2-1 + NC809 20x	EHEC-COM	1	1.00	0.99	1.00	1.00	0.95	1.00	0.98	0.98
Ecolli2-1 + NC809 30x	EHEC-COM	1	1.00	0.99	1.00	1.00	0.95	1.00	0.98	0.98
Ecolli2-1 + NC809 40x	EHEC-COM	1	0.97	0.92	0.75	0.75	0.82	0.88	0.98	0.86
Ecolli2-1 + NC809 50x	EHEC-COM	1	1.00	0.99	1.00	1.00	0.95	1.00	0.98	0.98
ECA279 + Ecolli2-1 10x	STEC-EHEC	1	0.97	0.92	0.75	0.75	0.82	0.88	0.98	0.86
ECA279 + Ecolli2-1 20x	STEC-EHEC	1	0.97	0.92	0.75	0.75	0.82	0.88	0.98	0.86
ECA279 + Ecolli2-1 30x	STEC-EHEC	1	0.97	0.92	0.75	0.75	0.82	0.88	0.98	0.86
ECA279 + Ecolli2-1 40x	STEC-EHEC	1	0.97	0.92	0.75	0.75	0.82	0.88	0.98	0.86
ECA279 + Ecolli2-1 50x	STEC-EHEC	1	0.97	0.92	0.75	0.75	0.82	0.88	0.98	0.86

\*EHEC, enterohemorrhagic *E. coli*; STEC, Shiga toxin-producing *Escherichia coli*; EPEC, enteropathogenic *E. coli*; COM, commensal *Escherichia coli*.

\*\* Non-EHEC = 0; EHEC = 1.

LGBM, LGBMClassifier; LR, Logistic Regression; DT, Decision Tree; XGB, XGBClassifier; RF, Random Forest; SVM, Support Vector Machine; GB, Gradient Boosting; ET, Extra Tree.

the results are summarized in Table 1. The table shows the top six most important features, ranked based on the number of times they were used by the models. The most important feature was found to be ECs\_1056, which corresponds to a phage excisionase gene. This feature was used by eight of the models. The second most important feature was ECs\_1812, which corresponds to the *nleA/espI* gene, coding for a type III secretion system (T3SS) secreted effector protein, and was used by seven of the models. The third most important feature was ECs\_1824, which corresponds to the *nleG* gene, which encodes another T3SS secreted effector protein, and was used by five of the models. The remaining features, namely ECs\_3858, ECs\_1815, and ECs\_1561, were used by four models each, and correspond to the *nleE*, *nleF*, and *espN* genes, respectively, all of which encodes T3SS secreted effector proteins. Multiple classifiers achieved near perfect performance when evaluated on

the D100 dataset. To better understand which combinations of genes contribute to the prediction of the EHEC pathotype, we generated all 64 ( $2^6$ ) genes presence/absence profiles and recorded in which case one of the models predicted an EHEC pathotype with confidence (Figure 2 and Section 2). This confirms that some genes, such as ECs\_1056 (phage excisionase) and ECs\_1824 (*nleG*) have an importance (they are needed in the majority of the predictions). Decision Tree and Gradient boosting learned the same decision rule: "ECs\_1812 (*nleA*) and (ECs\_1056 or (ECs\_1815 (*nleF*) and ECs\_1824))". The SVM classifier predicts EHEC for any two combination of ECs\_1056, ECs\_1815, and ECs\_1824. All those classifiers can predict an EHEC pathotype with as little as two genes. On the other hand, Extra tree and logistic regression make predictions involving three to four genes, showing that they can have different sensitivity.

TABLE 3 Prediction of the class probabilities on the milk metagenomes.

Strain	Accession number*	EHEC spiking level**	Class***	LGBM	LR	DT	XGB	RF	SVM	GB	ET
4712-O26	SRR19090780	0.5x10 <sup>5</sup>	1	1.00	0.99	1.00	1.00	0.95	1.00	0.98	0.98
6423-O26	SRR19090775	0.5x10 <sup>3</sup>	1	1.00	0.99	1.00	1.00	0.95	1.00	0.98	0.98
4712-O26	SRR19090792	0.5x10 <sup>2</sup>	1	1.00	0.99	1.00	1.00	0.95	1.00	0.98	0.98
6423-O26	SRR19090774	0.5x10 <sup>2</sup>	1	1.00	0.99	1.00	1.00	0.95	1.00	0.98	0.98
4712-O26	SRR19090778	0.5x10 <sup>1</sup>	1	1.00	0.99	1.00	1.00	0.95	1.00	0.98	0.98
6423-O26	SRR19090772	0.5x10 <sup>1</sup>	1	1.00	0.99	1.00	1.00	0.95	1.00	0.98	0.98
6423-O26	SRR19090769	0.5x10 <sup>1</sup>	1	1.00	0.99	1.00	1.00	0.95	1.00	0.98	0.98
EHEC-neg	SRR19090777	0	0	0.02	0.15	0.00	0.03	0.18	0.21	0.02	0.13

\*From Jaudou et al. (2022).

\*\* CFU.mL<sup>-1</sup>.

\*\*\*Non-EHEC = 0; EHEC = 1.

LGBM, LGBMClassifier; LR, Logistic Regression; DT, Decision Tree; XGB, XGBClassifier; RF, Random Forest; SVM, Support Vector Machine; GB, Gradient Boosting; ET, Extra Tree.

### 3.4. Performance of the eight models on the selected features

Supplementary Table 3 shows the evaluation metrics obtained by training the eight classifiers using the selected six features: ECs\_1056, ECs\_1812, ECs\_1824, ECs\_3858, ECs\_1815, ECs\_1561. All classifiers achieved high accuracy scores, ranging from 0.97 to 1.00, indicating a good performance in predicting the target variable. Logistic Regression, Extra Trees XGBClassifier, LGBMClassifier, Decision Tree, and SVM achieved perfect accuracy scores of 1.00, while Random Forest and Gradient Boosting achieved a slightly lower accuracy score of 0.97. Extra Tree and Random Forest achieved a precision of 0.94 on the EHEC class and a recall of 0.94 on the non-EHEC class. All the other classifiers achieved perfect precision, recall, and F1-scores, indicating that the selected features were informative and sufficient to discriminate perfectly between the two classes.

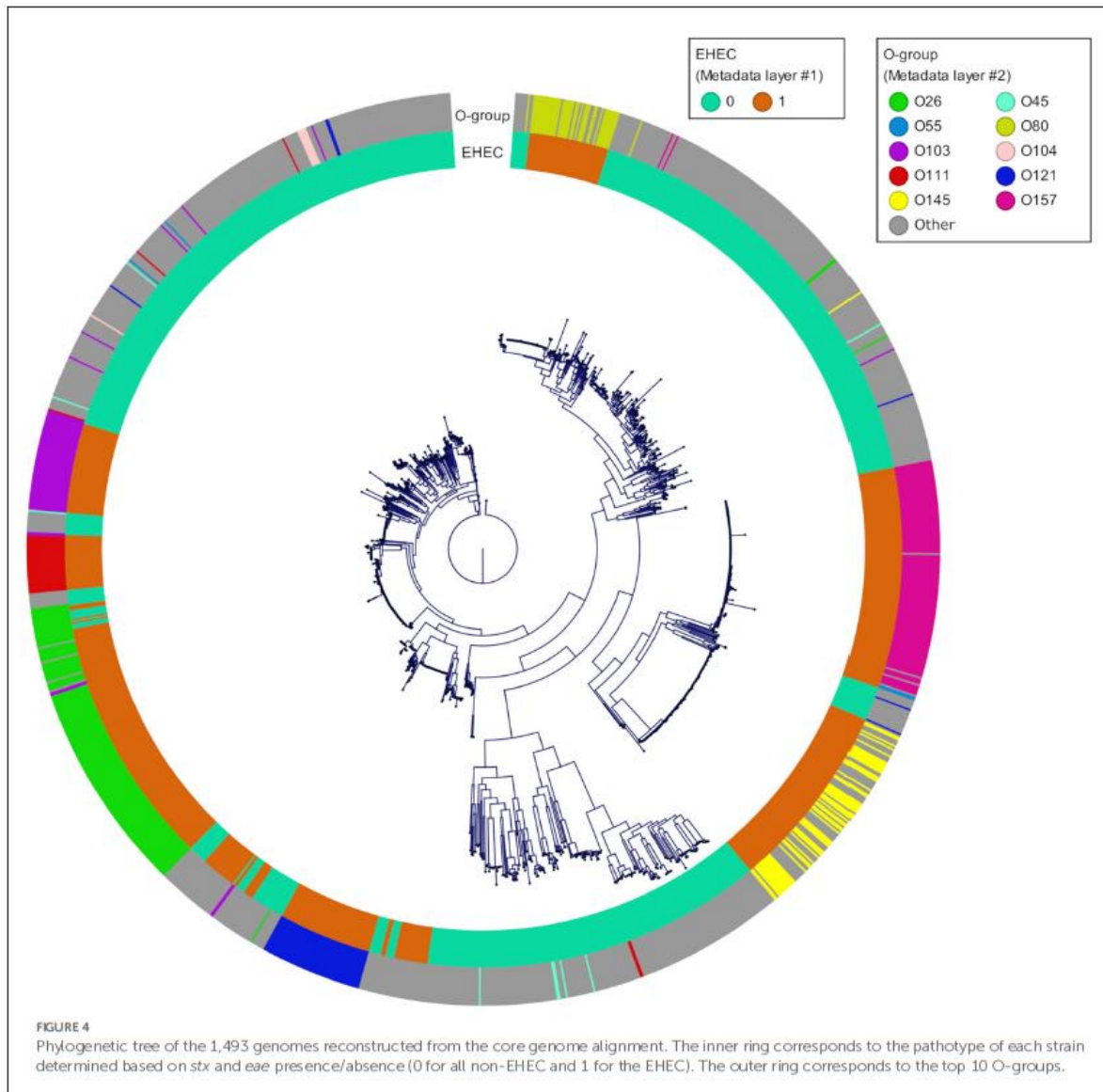
### 3.5. EHEC prediction on *in silico* *E. coli* mixtures

We first tested the ability of the different models to predict the presence of an EHEC strain in a mixture of *E. coli* strains. For this purpose, *in silico* mixtures of raw MinION reads were assembled. Assemblies produced using meta-Flye ranged from 5.63 to 7.94 Mb (mean = 6.57 Mb) and the number of contigs from 80 to 155 (mean = 106) (Supplementary Table 4). In all mixtures of *E. coli* strains, the assembly size was longer than a normal *E. coli* assembly (4.8–5.5 Mb). Shorter assemblies were produced by meta-Flye with the STEC-COM mixture (5.63–6.16 Mb). On the contrary, the EHEC-COM mixture produced longer assemblies (7.35–7.94 Mb). The eight models were then used to perform predictions on the *in silico* mixtures (Table 2). Predictions of the pathotype ranged from 0 to 1 and the average prediction from 0.01 to 0.99. The cut-off for binary classification is usually set to 0.5. Above or equal to this cut-off

value the presence of an EHEC is predicted, and below a non-EHEC is predicted. This cut-off of 0.5 was used to report the results presented in this study. The eight classifiers were able to predict the correct class 22 times over 25 predictions (88%). However, all models incorrectly predicted the presence of an EHEC three times for the STEC-EPEC mixture with the strain 97HMLP652 and 2142-O103 for a coverage of 30x, 40x and 50x, respectively. For all non-EHEC containing mixtures, the higher value was 0.47 for the same STEC-EPEC mixture with the Extra Tree classifier (Table 2). For the EHEC class the lower value was 0.75 for the EHEC-COM mixture and the STEC-EHEC mixture with the Decision Tree classifier and the XGBClassifier. Taken together, these data indicate that all the classifiers were able to predict with high confidence the presence of an EHEC in *E. coli* mixtures that combine different *E. coli* pathotypes. Only three false positives were predicted for the most difficult mixture combining a STEC and an EPEC.

### 3.6. EHEC prediction on artificially-contaminated raw milk

We then tested the performance of the eight models on complex mixtures using artificially contaminated raw milk. A bioinformatic pipeline called STECmetadector developed by Jaudou et al. (2022), was used to specifically extract *E. coli* reads from raw milk samples artificially contaminated with an O26 EHEC strain (from 0 to 500 CFU.mL<sup>-1</sup>). The *E. coli* reads were assembled using the same method as described in Section 2 (paragraph 2.5). Assemblies ranged from 5.2 Mb to 6.83 Mb (mean = 5.95 Mb) and the numbers of contigs from 5 to 62 (mean = 15). The eight classifiers were used to predict the pathotype of artificially contaminated raw milks (Table 3). All models were able to accurately predict the EHEC pathotype in the sample with high confidence, at the three contamination levels tested, regardless of the strain used for the artificial contamination. Notably, the raw milk used for spiking with the 6423-O26 strain was naturally containing commensal *E. coli* of serotype O185:H2 and O8:H19 (Jaudou et al., 2022).



Predictions of the pathotype ranged from 0.95 to 1 for the class EHEC for all contamination levels. The negative control (a non-contaminated raw milk), was classified accurately as non-EHEC by all eight classifiers with the higher value of 0.21 for the SVM (Table 3).

#### 4. Discussion

The correct detection and identification of highly pathogenic STEC from food remains challenging. Conventional detection methods based on the detection of the *stx* and *eae* genes (as well as genes from the most frequent serogroups) require an isolation step to ensure the correct characterization of the strain. Detection

of EHEC in food samples based on the presence of a small number of additional genes that are more specifically associated with strains possessing simultaneously the *stx* and *eae* genes would represent a significant improvement for screening food samples (Delannoy et al., 2016, 2022). With such an approach, the number of presumptive positive samples that require further investigation by isolation and genotypic characterization can be reduced by around 50% (Delannoy et al., 2016, 2022), allowing to save money and time. Still, the amount of unconfirmed presumptive positive samples may be a problem for both the food industry and the decision maker. In a previous study, we showed that long-read metagenomics was efficient in identifying *eae*-positive STEC strains from complex matrices such as raw milk in an isolation-independent way (Jaudou et al., 2022). However, we

have highlighted that the presence of multiple *E. coli* strains may hinder the identification of the *eae*-positive STEC due to the assembly-based approach used. We wanted to continue exploring the potential of long-read metagenomics and take full advantage of ML algorithms by applying them to predict the presence of an EHEC strain directly from *E. coli* reads assembly, even in the presence of multiple *E. coli* strains.

As of February 2021, 31,230 *E. coli* genomes were available in the NCBI Genbank database, around 10% of which are genomes of O157:H7 strains. To build our database, we downloaded complete *E. coli* genomes as well as some scaffolded genomes that contained accompanying metadata, while taking care of having the top 10 EHEC serotypes represented, as well as less frequent ones (Figure 4). Because the geographical distribution of certain clones may be skewed, we also included strains originating from all continents. During the genome selection, our objective was to obtain a database constituted at 50% of EHEC genomes (targets) and 50% of non-EHEC genomes (non-targets). When selecting the non-EHEC genomes we were careful to include various pathotypes such as EPEC (*eae+* only), STEC (*stx+* only), commensals (*stx* and *eae* negative) and some Extra-intestinal pathogenic *E. coli* (ExPEC) strains. Despite the size of our database and the precautions we took to build it, our final dataset after dereplication, was composed of 87 EHEC and 87 non-EHEC. We originally included large numbers of genomes for each of the top 10 serotypes observed in clinical cases worldwide in order to be representative of the frequency of isolation of the various serotypes. However, the diversity within each serotype appears limited. Indeed, several studies on various EHEC serotypes have shown that even the most diverse ones (in terms of SNPs) show a high degree of synteny and collinearity between isolates of different clades or lineages (Dallman et al., 2015; Ogura et al., 2017; Nishida et al., 2021). Also, the pool of genes included in the dataset is the pool present in the Sakai annotation. Therefore, by reducing the available pangenome and increasing the probability for each strain to possess one version of each CDS, we increased the similarity between the genomes in the dataset. To avoid possible data leakage, we chose to group all the genomes that had less than 100 genes difference in their repertoire (8.4% of the genes considered). This is a drastic filtering step, but it is, to our knowledge, the most reliable to avoid reporting biased performance estimates.

One of the first choices when designing the pipeline is whether to use raw reads or assembled data. Initial tests showed that performing the annotation directly on long reads generated a very large amount of data that was computationally too intensive for the downstream processing (not shown). Based on these results we chose to work with assemblies and used the Flye long-read assembler with the metagenome option in order to deal with highly non-uniform coverage, in particular with low level artificially-EHEC contaminated milks. An early step of the pipeline consists in the annotation of the assembled genomes. The advantage of the annotation software used, prokka, is that a reference genome can be used to standardize the annotation. In our case, we used the O157:H7 Sakai genome as reference over the K12 *E. coli* reference genome because it is an EHEC carrying around 20% more integrated genomic elements than K12 *E. coli*, like pathogenicity islands and phages (Hayashi et al., 2001). To generate the first

matrix of gene presence/absence we chose panaroo among GWAS programs such as Roary (Page et al., 2015), PIRATE (Bayliss et al., 2019), or PPanGoLiN (Gautreau et al., 2020) because it offers the possibility to add one new genome to an existing pangenome graph. This feature is the keystone of our pipeline because it is very important to add the new genome into the existing pangenome graph so as not to modify the original matrix used for training the models. Panaroo collapses genes into putative families with a family sequence identity level of 70% in the default mode. During the analysis of the generated matrix, we identified locus tags that were split in different groups and regrouped them. Indeed, the allelic variability of STEC virulence genes can be important (Michelacci et al., 2016).

Other studies have used different algorithms like Support Vector Machine, Gradient Boosting or Random Forest (Lupolova et al., 2016; Njage et al., 2019; Im et al., 2021; Shaik et al., 2022) but the nature of the data and the predictive outcome were different. In this study we used the power of ML to evaluate a high number of genes (1,178 CDS). We successfully decreased the number of genes needed for EHEC presence prediction down to six genes while keeping a high accuracy. It is remarkable that none of these six genes are related to the Shiga toxins. Surprisingly, neither *stx1* subunit A and B nor *stx2* subunit A and B are needed to predict an EHEC. Because it is present in all EPEC strains (*eae*-positive *E. coli*, non-target) the absence of *eae* in the six genes scheme is expected. In the reduced set of selected genes, we found five Type 3 Secretion System (T3SS) effectors and a phage excisionase. The T3SS represents an important component of the *E. coli* mobile gene pool. Although the LEE carries constitutive elements of the T3SS, additional effectors are encoded by prophages inserted into the genome (Tobe et al., 2006). A large number of studies have described T3SS effectors as associated virulence markers (Coombes et al., 2008; Konczyk et al., 2008; Bugarel et al., 2010a,b, 2011; Imamovic et al., 2010; Creuzburg et al., 2011). Here, the most important features identified for EHEC prediction are located on four genomic islands that harbors putative virulence factors already demonstrated to be present in EHEC strains: Sp4 (ECs 1056 / phage excisionase), Sp6 (ECs 1561 / *espN*), Sp9 (ECs 1812 / *nleA*, ECs 1815 / *nleF*, ECs 1824 / *nleG*) and SpLE3 (ECs 3858 / *nleE*) (Tobe et al., 2006; Rasko et al., 2008; Bugarel et al., 2010a,b, 2011; Delannoy et al., 2013). The *nleA* gene (ECs 1812 - Sp9), which was found to be the second most important feature in our study, has been shown to play a key role in the virulence of various pathogenic bacteria, including *E. coli* (Rasko et al., 2008). Similarly, the *nleG* gene (ECs 1824 - Sp9), which was the third most important feature, has been shown to be important for the virulence of enterohemorrhagic *E. coli* (Tobe et al., 2006). Other T3SS effectors located in these four genomic islands have previously been shown to be associated with EHEC. For example, the Sp4 genomic island also harbors *espV* (ECs 1127), which, in combination with *espK* (ECs 1568 - Sp6) have been demonstrated to be present in EHEC strains and proposed as genetic markers to reduce false-positive results in food testing (Delannoy et al., 2013, 2016). Similarly, combinations of genes from Sp9 and SpLE3 were demonstrated to be strong signatures of typical EHECs (Bugarel et al., 2010a,b, 2011). These four genomic islands are recurrently found as harboring important features with all models and were previously experimentally found associated

with EHEC. This strongly suggests that these genomic islands are stably associated with both the LEE and the presence of an *stx*-phage and may have co-evolved (Guo et al., 2012). Although, the precise order of acquisition of these mobile genetic elements remains to be determined. The only incorrect EHEC predictions of the models using *in-silico* mixtures were obtained with the STEC/EPEC mixtures containing the aEPEC strain 2142-O103. Although negative for the *stx* gene this strain harbors the different genomic islands: Sp4, Sp6, Sp9, and SpLE3. It also belongs to a known EHEC serotype (O103:H25) that has been associated with an HUS outbreak (Schimmer et al., 2008). It is thus likely that this strain represents what had previously been named an EHEC-like or EHEC-LST (Bielaszewska et al., 2007; Mellmann et al., 2008; Bugarel et al., 2011), meaning that it could constitute an EHEC progeny that has lost the *stx* phage at one point. The existence of such EHEC-like strains constitute a caveat of our approach, as it is impossible for our model to distinguish a STEC/EHEC-like mixture accurately. However, from a risk management perspective, it could be beneficial to detect this kind of strains when in the presence of other STEC strains due to the potential of these EHEC-like to acquire the *stx* phage and become typical EHEC (Bielaszewska et al., 2007; Mellmann et al., 2008). Correct and timely identification of EHEC is crucial in food microbiology as well as for surveillance of STEC-mediated disease. The growing genomic sequence data offers additive information that may support the identification of discriminative EHEC markers (Kiel et al., 2018). To extend EHEC diagnostics in the post-genomic era beyond the detection of the O157:H7 and the non-O157 serogroups from the Top 7, we developed suitable pipelines that integrate high throughput sequence data, to predict with high specificity and sensitivity EHEC strains. Different combinations of discriminative genetic markers were identified and validated to target the main STEC subgroup (*eae*-positive STEC) associated with severe human infections and outbreaks worldwide. Our study is in line with recent papers showing the potential and power of GWAS and Machine Learning approaches for designing biomarkers that target foodborne pathogens (Feucherolles et al., 2021; Sévellec et al., 2022). Here, the description of these new EHEC biomarkers is the confirmation that *stx* and *eae* are not the only genetic markers that are the hallmark of EHEC, but that EHEC characterization is much more complex than the simultaneous identification of *stx* and *eae* genes. There are in fact associated factors (type III effectors are some of them as shown in this study) which, by their presence or absence, provide a fairly precise predictive model on the co-localization of *stx* and *eae* in a single strain. The new EHEC markers found using ML in our study could predict EHEC with very high accuracy in a large genome dataset and artificially contaminated raw milk metagenomes. The correct prediction of the EHEC strain while co-occurring with another *E. coli* strain at a ratio of 1:1 is remarkable. Most programs that aim at distinguishing strains from the same species relies on coverage differences (*i.e.* for assemblers and binning tools). These findings open the door

for the development of new diagnostics tests for a better screening of EHEC in foods products. As long as DNA sequence-based diagnostics of mixed populations cannot resolve whether relevant markers like *stx* and *eae* genes are present in the same genome, some risk of generating false-positive results exist. Including the combination of additional EHEC-related markers like those we described here, in the detection scheme, would supports a better hazard characterization of typical EHEC.

## Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/Supplementary material.

## Author contributions

FV, SJ, SD, and PF conceptualized the project. PF and SD were in charge of funding acquisition. SJ downloaded the fasta from public database. SJ and M-LT did the milk artificially contaminated sequencing and assembly. FV did the *in silico* mixtures analysis, machine learning model development, database creation, and wrote the original draft. Methodology and resources by FV, SJ, SD, and HR. HR did the cluster analysis. All authors contributed to manuscript revision, read, and approved the submitted version.

## Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

## Supplementary material

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fmicb.2023.1118158/full#supplementary-material>

## References

- Bayliss, S. C., Thorpe, H. A., Coyle, N. M., Sheppard, S. K., and Feil, E. J. (2019). PIRATE: A fast and scalable pan-genomics toolbox for clustering diverged orthologues in bacteria. *Gigascience* 8, g1z119. doi: 10.1093/gigascience/g1z119
- Beutin, L., and Fach, P. (2015). Detection of Shiga toxin-producing *Escherichia coli* from nonhuman sources and strain typing. *295 Microbiol. Spectrum* 2, EHEC-0001-2013. doi: 10.1128/9781555818791.ch14
- Bielaszewska, M., Prager, R., Kock, R., Mellmann, A., Zhang, W., Tschape, H., et al. (2007). Shiga toxin gene loss and transfer *in vitro* and *in vivo* during enterohemorrhagic *Escherichia coli* O26 infection in humans. *Appl. Environ. Microbiol.* 73, 3144–3150. doi: 10.1128/AEM.02937-06
- Bugarel, M., Beutin, L., and Fach, P. (2010a). Low-density microarray targeting non-locus of enterocyte effacement effectors (*nle* genes) and major virulence factors of Shiga toxin-producing *Escherichia coli* (STEC): a new approach for molecular risk assessment of STEC isolates. *Appl. Environ. Microbiol.* 76, 203–211. doi: 10.1128/AEM.01921-09
- Bugarel, M., Beutin, L., Martin, A., Gill, A., and Fach, P. (2010b). Micro-array for the identification of Shiga toxin-producing *Escherichia coli* (STEC) seropathotypes associated with hemorrhagic colitis and hemolytic uremic syndrome in humans. *Int. J. Food Microbiol.* 142, 318–329. doi: 10.1016/j.ijfoodmicro.2010.07.010
- Bugarel, M., Beutin, L., Scheutz, F., Loukiadis, E., and Fach, P. (2011). Identification of genetic markers for differentiation of Shiga toxin-producing, enteropathogenic, and avirulent strains of *Escherichia coli* O26. *Appl. Environ. Microbiol.* 77, 2275–2281. doi: 10.1128/AEM.02832-10
- Coombs, B. K., Wickham, M. E., Mascarenhas, M., Gruenheid, S., Finlay, B. B., and Karmali, M. A. (2008). Molecular analysis as an aid to assess the public health risk of non-O157 Shiga toxin-producing *Escherichia coli* strains. *Appl. Environ. Microbiol.* 74, 2153–2160. doi: 10.1128/AEM.02566-07
- Creuzburg, K., Middendorf, B., Mellmann, A., Martaler, T., Holz, C., Fruth, A., et al. (2011). Evolutionary analysis and distribution of type III effector genes in pathogenic *Escherichia coli* from human, animal and food sources. *Environ. Microbiol.* 13, 439–452. doi: 10.1111/j.1462-2920.2010.02349.x
- Dallman, T. J., Ashton, P. M., Byrne, L., Perry, N. T., Petrovska, L., Ellis, R., et al. (2015). Applying phylogenomics to understand the emergence of Shiga-toxin-producing *Escherichia coli* O157:H7 strains causing severe human disease in the UK. *Microbial Genomics* 1, e000029. doi: 10.1099/mgen.0.000029
- Delannoy, S., Beutin, L., and Fach, P. (2013). Discrimination of enterohemorrhagic *Escherichia coli* (EHEC) from non-EHEC strains based on detection of various combinations of type III effector genes. *J. Clin. Microbiol.* 51, 3257–3262. doi: 10.1128/JCM.01471-13
- Delannoy, S., Chaves, B. D., Ison, S. A., Webb, H. E., Beutin, L., Delaval, J., et al. (2016). Revisiting the STEC testing approach: using *espK* and *espV* to make enterohemorrhagic *Escherichia coli* (EHEC) detection more reliable in beef. *Front. Microbiol.* 7, 1. doi: 10.3389/fmicb.2016.00001
- Delannoy, S., Tran, M.-L., and Fach, P. (2022). Insights into the assessment of highly pathogenic Shiga toxin-producing *Escherichia coli* in raw milk and raw milk cheeses by high throughput real-time PCR. *Int. J. Food Microbiol.* 366, 109564. doi: 10.1016/j.ijfoodmicro.2022.109564
- European Food Safety Authority and European Centre for Disease Prevention and Control. (2021). The European union one health 2019 zoonoses report. *EFSA J.* 19, e06406. doi: 10.2903/j.efsa.2021.6406
- European Food Safety Authority. (2007). Scientific opinion of the panel on biological hazards (biohaz)-monitoring of verotoxigenic *Escherichia coli* (VTEC) and identification of human pathogenic VTEC types. *EFSA J.* 5, 579. doi: 10.2903/j.efsa.2007.579
- European Food Safety Authority. (2013). Scientific opinion on VTEC-seropathotype and scientific criteria regarding pathogenicity assessment. *EFSA J.* 11, 3138–3244. doi: 10.2903/j.efsa.2013.3138
- Feucherolles, M., Nennig, M., Becker, S. L., Martiny, D., Losch, S., Penny, C., et al. (2021). Combination of MALDI-TOF mass spectrometry and machine learning for rapid antimicrobial resistance screening: the case of *Campylobacter* spp. *Front. Microbiol.* 12, 804484. doi: 10.3389/fmicb.2021.804484
- Franz, E., Delaquis, P., Morabito, S., Beutin, L., Gobius, K., Rasko, D. A., et al. (2014). Exploiting the explosion of information associated with whole genome sequencing to tackle Shiga toxin-producing *Escherichia coli* (STEC) in global food production systems. *Int. J. Food Microbiol.* 187, 57–72. doi: 10.1016/j.ijfoodmicro.2014.07.002
- Gautreau, G., Bazin, A., Gachet, M., Planel, R., Burlot, L., Dubois, M., et al. (2020). PpanGGOLiN: depicting microbial diversity via a partitioned pangene graph. *PLoS Comput. Biol.* 16, e1007732. doi: 10.1371/journal.pcbi.1007732
- Gill, A., Dussault, F., McMahon, T., Petronella, N., Wang, X., Cebelinski, E., et al. (2022). Characterization of atypical Shiga toxin gene sequences and description of stx2j, a new subtype. *J. Clin. Microbiol.* 60, e02229–e02221. doi: 10.1128/jcm.02229-21
- Guo, F., Wei, W., Wang, X., Lin, H., Ding, H., Huang, J., et al. (2012). Co-evolution of genomic islands and their bacterial hosts revealed through phylogenetic analyses of 17 groups of homologous genomic islands. *Genet. Mol. Res.* 11, 3735–3743. doi: 10.4238/2012.October.15.5
- Hall, M. B. (2022). Rasusa: randomly subsample sequencing reads to a specified coverage. *J. Open Source Softw.* 7, 3941. doi: 10.21105/joss.03941
- Hayashi, T., Makino, K., Ohnishi, M., Kurokawa, K., Ishii, K., Yokoyama, K., et al. (2001). Complete genome sequence of enterohemorrhagic *Escherichia coli* O157:H7 and genomic comparison with a laboratory strain K-12. *DNA Res.* 8, 11–22. doi: 10.1093/dnares/8.1.11
- Herold, S., Karch, H., and Schmidt, H. (2004). Shiga toxin-encoding bacteriophages-genomes in motion. *Int. J. Med. Microbiol.* 294, 115–121. doi: 10.1016/j.ijmm.2004.06.023
- Im, H., Hwang, S.-H., Kim, B. S., and Choi, S. H. (2021). Pathogenic potential assessment of the Shiga toxin-producing *Escherichia coli* by a source attribution considered machine learning model. *Proc. Natl. Acad. Sci.* 118, e2018877118. doi: 10.1073/pnas.2018877118
- Imamovic, L., Jofre, J., Schmidt, H., Serra-Moreno, R., and Muniesa, M. (2009). Phage-mediated Shiga toxin 2 gene transfer in food and water. *Appl. Environ. Microbiol.* 75, 1764–1768. doi: 10.1128/AEM.02273-08
- Imamovic, L., Tozzoli, R., Michelacci, V., Minelli, F., Marziano, M. L., Caprioli, A., et al. (2010). O1-57, a genomic island of *Escherichia coli* O157, is present in other seropathotypes of Shiga toxin-producing *E. coli* associated with severe human disease. *Infect. Immunity* 78, 4697–4704. doi: 10.1128/IAI.00512-10
- International Organization for Standardization. (2012). *Microbiology of food and animal feed. Real-time polymerase chain reaction (PCR)-based method for the detection of food-borne pathogens. Horizontal method for the detection of Shiga toxin-producing Escherichia coli (STEC) and the determination of O157, O111, O26, O103 and O145 serogroups. ISO/TS 13136:2012*. 22 p.
- Jaudou, S., Deneke, C., Tran, M.-L., Schuh, E., Goehler, A., Vorimore, F., et al. (2022). A step forward for Shiga toxin-producing *Escherichia coli* identification and characterization in raw milk using long-read metagenomics. *Microbial Genomics* 8, mgen00911. doi: 10.1099/mgen.0.000911
- Jaudou, S., Tran, M.-L., Vorimore, F., Fach, P., and Delannoy, S. (2023). Hybrid assembly from 75 *E. coli* genomes isolated from French bovine food products between 1995 and 2016. *Microbiol. Resour. Annot.* 12, e01095-22. doi: 10.1128/mra.01095-22
- Kiel, M., Sagory-Zalkind, P., Miganeh, C., Stork, C., Leimbach, A., Sekse, C., et al. (2018). Identification of novel biomarkers for multiplex serotypes of Shiga toxin-producing *Escherichia coli* and the development of priority PCR for their detection. *Front. Microbiol.* 9, 1321. doi: 10.3389/fmicb.2018.011321
- Kolmogorov, M., Bickhart, D. M., Behsaz, B., Gurevich, A., Rayko, M., Shin, S. B., et al. (2020). metalyze: scalable long-read metagenome assembly using repeat graphs. *Nat. Methods* 17, 1103–1110. doi: 10.1038/s41592-020-00971-x
- Konczyk, P., Ziebell, K., Mascarenhas, M., Choi, A., Michaud, C., Kropinski, A. M., et al. (2008). Genomic O island 122, locus for enterocyte effacement, and the evolution of virulent verocytotoxin-producing *Escherichia coli*. *J. Bacteriol.* 190, 5832–5840. doi: 10.1128/JB.00480-08
- Lupolova, N., Dallman, T. J., Matthews, L., Bono, J. L., and Gally, D. L. (2016). Support vector machine applied to predict the zoonotic potential of *E. coli* O157 cattle isolates. *Proc. Natl. Acad. Sci. U. S. A.* 113, 11312–11317. doi: 10.1073/pnas.1606567113
- Mellmann, A., Lu, S., Karch, H., Xu, J., Harmsen, D., Schmidt, M. A., et al. (2008). Recycling of Shiga toxin 2 genes in sorbitol-fermenting enterohemorrhagic *Escherichia coli* O157:NM. *Appl. Environ. Microbiol.* 74, 67–72. doi: 10.1128/AEM.01906-07
- Michelacci, V., Orsini, M., Knijn, A., Delannoy, S., Fach, P., Caprioli, A., et al. (2016). Development of a high resolution virulence allelic profiling (HReVAP) approach based on the accessory genome of *Escherichia coli* to characterize Shiga-toxin producing *E. coli* (STEC). *Front. Microbiol.* 7, 202. doi: 10.3389/fmicb.2016.00202
- Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., Von Haeseler, A., et al. (2020). IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.* 37, 1530–1534. doi: 10.1093/molbev/msaa015
- Moradigaravand, D., Palm, M., Farewell, A., Mustonen, V., Warringer, J., and Parts, L. (2018). Prediction of antibiotic resistance in *Escherichia coli* from large-scale pan-genome data. *PLoS Comput. Biol.* 14, e1006258. doi: 10.1371/journal.pcbi.1006258
- Müthing, J., Schweppe, C. H., Karch, H., and Friedrich, A. W. (2009). Shiga toxins, glycosphingolipid diversity, and endothelial cell injury. *Thromb. Haemost.* 101, 252–264. doi: 10.1160/TH08-05-0317
- Nataro, J. P., and Kaper, J. B. (1998). Diarrheagenic *Escherichia coli*. *Clin. Microbiol. Rev.* 11, 142–201. doi: 10.1128/CMR.11.1.142



Nishida, R., Nakamura, K., Taniguchi, I., Murase, K., Ooka, T., Ogura, Y., et al. (2021). The global population structure and evolutionary history of the acquisition of major virulence factor-encoding genetic elements in Shiga toxin-producing *Escherichia coli* O121:H19. *Microbial Genomics* 7, 000716. doi: 10.1099/mgen.0.000716

Njage, P. M. K., Leekitchareonphon, P., and Hald, T. (2019). Improving hazard characterization in microbial risk assessment using next generation sequencing data and machine learning: predicting clinical outcomes in shiga toxin-producing *Escherichia coli*. *Int. J. Food Microbiol.* 292, 72–82. doi: 10.1016/j.ijfoodmicro.2018.11.016

Ogura, Y., Gotoh, Y., Itoh, T., Sato, M. P., Seto, K., Yoshino, S., et al. (2017). Population structure of *Escherichia coli* O26:H11 with recent and repeated *stx2* acquisition in multiple lineages. *Microbial Genomics* 3, e000141. doi: 10.1099/mgen.0.000141

Page, A. J., Cummins, C. A., Hunt, M., Wong, V. K., Reuter, S., Holden, M. T., et al. (2015). Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics* 31, 3691–3693. doi: 10.1093/bioinformatics/btv421

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

Scheutz, F., Teel, L. D., Beutin, L., Piérard, D., Buvens, G., Karch, H., et al. (2012). Multicenter evaluation of a sequence-based protocol for subtyping Shiga toxins and standardizing *stx* nomenclature. *J. Clin. Microbiol.* 50, 2951–2963. doi: 10.1128/JCM.00860-12

Schimmer, B., Nygard, K., Eriksen, H.-M., Lassen, J., Lindstedt, B.-A., Brandal, L. T., et al. (2008). Outbreak of haemolytic uraemic syndrome in norway caused by *stx2*-positive *Escherichia coli* O103:H25 traced to cured mutton sausages. *BMC Infect. Dis.* 8, 41. doi: 10.1186/1471-2334-8-41

Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. *Bioinformatics* 30, 2068–2069. doi: 10.1093/bioinformatics/btu153

Sévellec, Y., Ascencio, E., Douarre, P.-E., Félix, B., Gal, L., Garmyn, D., et al. (2022). *Listeria monocytogenes*: investigation of fitness in soil does not support the relevance of ecotypes. *Front. Microbiol.* 13, 917588. doi: 10.3389/fmicb.2022.917588

Shaik, S., Singh, A., Suresh, A., and Ahmed, N. (2022). Genome informatics and machine learning-based identification of antimicrobial resistance-encoding features and virulence attributes in *Escherichia coli* genomes representing globally prevalent lineages, including high-risk clonal complexes. *Mbio* 13, e03796–e03721. doi: 10.1128/mbio.03796-21

Tobe, T., Beatson, S. A., Taniguchi, H., Abe, H., Bailey, C. M., Fivian, A., et al. (2006). An extensive repertoire of type iii secretion effectors in *Escherichia coli* O157 and the role of lambdaoid phages in their dissemination. *Proc. Natl. Acad. Sci. U. S. A.* 103, 14941–14946. doi: 10.1073/pnas.0604891103

Tonkin-Hill, G., MacAlasdair, N., Ruis, C., Weimann, A., Horesh, G., Lees, J. A., et al. (2020). Producing polished prokaryotic pangenomes with the panaroo pipeline. *Genome Biol.* 21, 1–21. doi: 10.1186/s13059-020-02090-4

Zhang, W., Kohler, B., Oswald, E., Beutin, L., Karch, H., Morabito, S., et al. (2002). Genetic diversity of intimin genes of attaching and effacing *Escherichia coli* strains. *J. Clin. Microbiol.* 40, 4486–4492. doi: 10.1128/JCM.40.12.4486-4492.2002

### Main results:

- Identification of six genetic markers which combination of presence/absence helps identifying *eae*-positive STEC from a metagenome assembly.
- Culture-free identification of *eae*-positive STEC in *in silico* *E. coli* mixtures assemblies even at ratio of 1:1 and low coverage with all tested algorithms.

### Main conclusions:

- Circumvent the problem posed by the assembly-based approach regarding the presence of multiple *E. coli* strains.

### Perspectives:

- Development of PCR techniques targeting the six markers and of an algorithm to analyze the results.
- Find alternative assembly-free approaches.
- Include this approach in the STECmetadetector pipeline.

## Chapter5-4: Application of the developed method on presumptive positive samples or naturally contaminated samples.

### Context:

With the previous studies, I have shown the potential of our method to characterize STEC from artificially contaminated raw milk samples using a low inoculation level of 5 CFU.mL<sup>-1</sup>. With the developed workflow, I successfully characterized the inoculated STEC strains. However, meeting the conditions cannot always be feasible on naturally contaminated samples. Indeed, the STEC strain, which concentration in naturally contaminated samples might be as low as 10 cells (EFSA, 2020), has to grow to a level of 10<sup>8</sup> copies.mL<sup>-1</sup> and be present 10-times more than additional *E. coli* strains post-enrichment. In particular, the presence of multiple *E. coli* strains is highly probable. Consequently, the next step is to apply the complete workflow developed here on naturally contaminated samples and test if the limits are also valid.

Hence, I have taken the opportunity to test the workflow on samples positive for *stx* and *eae* genes (with qPCR) from which an *eae*-positive STEC was further isolated (positive) or not (presumptive-positive). I received enrichment cultures of presumptive-positive (positive for both *stx* and *eae* genes using qPCR) raw-milk cheese and ground beef samples kindly provided by a departmental laboratory (LDA39 laboratory) and a meat-producing company, respectively. Additionally, our method was applied on enriched feces or swab samples from the national reference center for *E. coli* (Robert Debré hospital), in which *eae*-positive STEC presence was confirmed after isolation of a STEC strain. Despite the different enrichment conditions used, it was the opportunity to assess the benefits or limits of the method developed in this project and test its application to other matrices as raw-milk cheese, beef and feces. The long-read metagenomics method developed in this project was tested on those samples as well as the machine learning (ML) approach.

The application of the developed method on presumptive-positive or naturally contaminated samples confirmed the previously identified limitations. While results of the machine learning approach indicated the presence of an *eae*-positive STEC, I was not able to characterize the STEC strain with the STECmetadector pipeline.

## Material and method:

### 1. Samples enrichment

A total of 11 enriched raw-milk cheese samples were received from the LDA39 laboratory. A meat-producing company sent four enrichment broths of ground beef. Lastly, five clinical samples (rectal swab (n=2) or feces (n=3) enrichment) were received from the National reference center for *E. coli* (Robert Debré Hospital) (Table 7). Enrichment was done by each laboratory at 37°C in BPW (without acriflavine). Except for the clinical samples that were frozen, raw milk and ground beef enrichment broth were fresh. The samples were conserved and transported at +4°C.

### 2. DNA extraction, quantification and qualification

Analyses were performed as described in Publication 3. Prior to DNA extraction, one washing step was performed on raw-milk cheese, ground beef and clinical samples. DNA was extracted from one milliliter of each sample in triplicates using the Lucigen MasterPure extraction kit, except for the clinical samples from which 400 µL were used in duplicate. DNA extracts were quantified using the Qubit 3.0 Fluorometer and the Qubit dsDNA BR (Broad Range) Assay-kit (Thermo Fisher Scientific); its quality assessed using a Nanodrop UV-Vis Spectrophotometer (Thermo Fisher Scientific). Their purity was assessed using a Nanodrop UV-Vis Spectrophotometer (Thermo Fisher Scientific). The integrity of extracted DNA was assessed using a TapeStation system and Genomic screentape (Agilent) analyzed using the TapeStation Analysis software v4.4.1 (except for ground beef samples).

### 3. Detection (using real-time PCR) and quantification (using quantitative digital PCR) of *stx*, *eae* and *E. coli* generic (*wecA* or *cdgR*) genes

Real-time PCR was performed on each sample to check whether *stx* and *eae* genes were present as well as *cdgR* or *wecA*. Quantitative digital PCR was applied using the Biomark HD system as previously described to quantify the level of STEC present in the enriched samples and compared to the previously assessed limit. Additional genetic markers *espK* and *espV* were quantified (Delannoy *et al.*, 2016). From positive samples, MinION libraries were prepared using the LSK-SQK109 library preparation kit and EXP-NBD104 or EXP-NBD114 barcoding kit and sequenced on FLO-MIN106 flow cells, as described previously. For the raw-milk cheese samples, *stx* and *eae* genes were quantified below the level required for successful STEC characterization. Consequently, we sequenced only four samples that are presented here. All five clinical samples from Robert Debré and the four samples from ground beef were sequenced on two different flow cells.

MinION raw data were basecalled and demultiplexed using guppy v6.0.0+ab79250 for raw-milk cheese and clinical samples and v6.4.2+97a7f06 for beef samples, using the hac model, a minimum q-score of 9 and --trim-adapters, --trim-barcodes and --compress-fastq parameter.

#### 4. Data analysis and assessment of *eae*-positive STEC contamination

The data were analyzed using the STECmetadetector pipeline and further processed using the machine learning approach, as described in publication 3 and 4, respectively. Basecalled and demultiplexed data were processed with the STECmetadetector to try to characterize the STEC strain, and with the machine learning approach to identify the presence of *eae*-positive STEC. Characterization was confirmed when *stx* and *eae* genes were co-localized on the same contig.

With the STECmetadetector pipeline, almost 100% of the reads were classified. To simplify the representation, only the genus representing >2% of abundance were shown here. Barplots were generated using the summary file from the STECmetadetector pipeline and the *ggplot2* R package v3.4.1 (Wickham, 2016) using R v4.1.2.

For each sample, the extracted *E. coli* reads were assembled using Flye v2.9-b1768 and the *--meta* parameter (Kolmogorov *et al.*, 2020). The assembly was annotated with prokka using the Sakai genome as reference (Hayashi *et al.*, 2001; Seemann, 2014). Panaroo, a pan-genomic analysis program, was used to generate a presence/absence table of the genetic features of Sakai genome (Tonkin-Hill *et al.*, 2020). This table was added to the global matrix generated from the complete database of 1 425 *E. coli* genomes constructed for the ML approach (Publication 4). Predictions regarding the presence of an *eae*-positive STEC were made using the eight algorithms tested in the ML approach.

#### Results and discussion:

The aim of this study was to apply the developed workflow on presumptive-positive samples or naturally contaminated samples. The results presented here included four samples from the raw-milk cheese enrichments, four samples from ground beef enrichment broths and five clinical samples (either rectal swab or feces enrichment broth). Table 7 summarizes the DNA concentration, yield, purity ratios and integrity measured for each sample. On all samples, it was possible to extract high quantities of HMW DNA but the purity ratios showed the presence of contaminants (Table 7). Additionally, the DNA integrity number (DIN) reveals the presence of short DNA fragments, which may be matrix dependent, but also led to highly fragmented assemblies. In particular, enrichments from clinical samples were part of a collection and were thus, frozen for a certain period of time. As we have shown previously (additional experiment 1), the freezing process might lead to DNA degradation. As it is shown in Table 7, the DIN value was higher in rectal swab enrichments (DIN>8), which was sufficient to perform MinION sequencing. Although the DIN value was not measured for ground beef samples, the N50-value measured on filtered and extracted *E. coli* reads used for assembly was sufficient for most samples, as shown in Table 8.

Prior to sequencing, we used real-time PCR to check for the presence of *stx* and *eae* genes. Results have confirmed the presence of both *stx* and *eae* genes in all samples presented here. *Stx*- and *eae*-positive samples were further sequenced and processed using the STECmetadetector pipeline. As these samples were obtained from various laboratories, they

were not enriched using our tested enrichment conditions. In particular, acriflavine was not used in any of the enrichments. The taxonomic assignment of reads within the STECmetadetector pipeline allowed us to estimate the proportion of *Escherichia* among the microbial flora. The microbial composition (>2% at genus level) is represented using barplots for each matrix in Figure 16. Results showed highly variable microbial flora between matrices and different *Escherichia* reads proportions within samples. In clinical samples, the proportion varied from 22.01% to 67.09%, from 3.6% to 46.48% in raw-milk cheese samples and in meat samples from 4.8% to 49.59% (Fig. 16).

Matrix	Sample	DNA concentration (ng.µL <sup>-1</sup> )	DNA yield (ng)	A260/A280	A260/A230	DIN
Cheese	608961	77.4	2709	1.80	1.18	7
Cheese	608839	101.8	3563	1.81	1.26	6.9
Cheese	608980	85.2	2982	1.82	1.29	7.8
Cheese	609077	85.8	3003	1.80	1.15	7.8
Ground beef	V1001	67.6	2366	1.81	1.49	na
Ground beef	V107	442	15470	1.84	1.47	na
Ground beef	V394	90.6	3171	1.89	1.72	na
Ground beef	V512	179.8	6293	1.84	1.47	na
Feces	Bouillon1	119+116.8*	4165+4088*	1.78/1.77*	1.13/1.11*	6.4
Feces	Bouillon2	60.2	2107	1.78	0.97	6.7
Rectal swab	Bouillon3	35.2	1232	1.79	0.95	8.5
Feces	Bouillon4	52+35*	1820+1225*	na/1.78*	na/0.85*	6.4
Rectal swab	Bouillon5	53.8+48.2*	1883+1687*	na/1.73*	na/0.74*	8.9

Table 7 : Properties of DNA extracted from presumptive-positive milk and ground beef enrichment and naturally contaminated feces or rectal swab enrichment.

Two technical replicates were used for some samples and are represented with \*.

After taxonomic assignment, an extraction of *E. coli* reads is performed, which are then mapped onto O-group genetic markers to detect the presence of multiple strains as well as *eae* and *stx* genes. Results from read mapping on O-group, *stx* and *eae* genes, showed that for most samples (11/13) a variety of *E. coli* strains were present but the *stx* and *eae* genes were barely detected (below 10 reads), as presented in Table 8. In these samples, no STEC strain could be completely assembled as the generated assemblies were highly fragmented (Table 8). The presence of *stx* and *eae* genes was not detected (<3x) in 7/13 samples.

Table 8 : Summary of STECmetadetector and machine learning results.

The mapping of reads onto O-groups and virulence factors (*stx* and *eae* genes) are represented as well as metric obtained from the generated assembly. Characterization of the assemblies is also reported with serotype, *stx* sub-type, *eae* presence and predicted pathotype using both the STECmetadetector pipeline and the machine learning approach.

Matrix	Sample	Reads N50-value (b)*	# O groups	Read depth on <i>eae</i> (x)	Read depth on <i>stx</i> (x)	# of contigs	Total assembly length (Mb)	Largest contig (Mb)	ST	Serotype	<i>stx eae</i>	STECmetadetector	Machine learning
Cheese	608961	8 141	16	1	<i>stx1</i> (2)	48	5.89	1.34	na	O113:O8:H4	na na	NON-STEC	No
Cheese	608839	6 425	11	na	<i>stx2</i> (1)	44	6.04	1.35	399	O6/O25/Onovel31:H12	na na	NON-STEC	No
Cheese	608980	10 576	7	na	<i>stx1</i> (8)	75	6.27	1.17	na	Onovel3/O123:H16	<i>stx1</i> na	STEC	No
Cheese	609077	9 393	14	na	na	56	5.30	0.94	na	O10/Onovel3/O25/Onovel31:	na na	NON-STEC	No
Ground beef	V1001	15 855	1	10	<i>stx2</i> (9)	8	5.32	1.4	na	O177:H25	<i>stx2d</i> +	EHEC	Yes
Ground beef	V107	4 633	18	2	<i>stx2</i> (3)	93	5.86	0.36	na	Onovel130:H21	na na	NON-STEC	No
Ground beef	V394	9 385	7	16	<i>stx2</i> (2)	48	5.74	0.52	na	O182:H25	<i>stx2a</i> +	EHEC?	Yes
Ground beef	V512	7 029	13	0	na	81	5.65	0.64	na	O8; O5; H11	na na	NON-STEC	No
Feces	Bouillon1	3 169	4	44	54	71	6.24	1.57	21	O26:H11; H6	<i>stx1a</i> +	EHEC?	Yes
Feces	Bouillon2	4 937	1	Na	Na	4	5	4.86	14 1	O2 : H6	na na	NON-STEC	No
Rectal swab	Bouillon3	16 689	4	2	Na	49	5.86	1.58	na	O25 :H6 ; O16	na na	NON-STEC	No
Feces	Bouillon4	6 171	7	1	<i>stx2</i> (11), <i>stx1</i> (1)	12	5.35	4.44	na	O15:H1	na na	NON-STEC	No
Rectal swab	Bouillon5	17 089	7	341	<i>stx2</i> (240)	115	7.17	0.45	na	O26; O1; H7	<i>stx1a</i> +	EHEC?	Yes

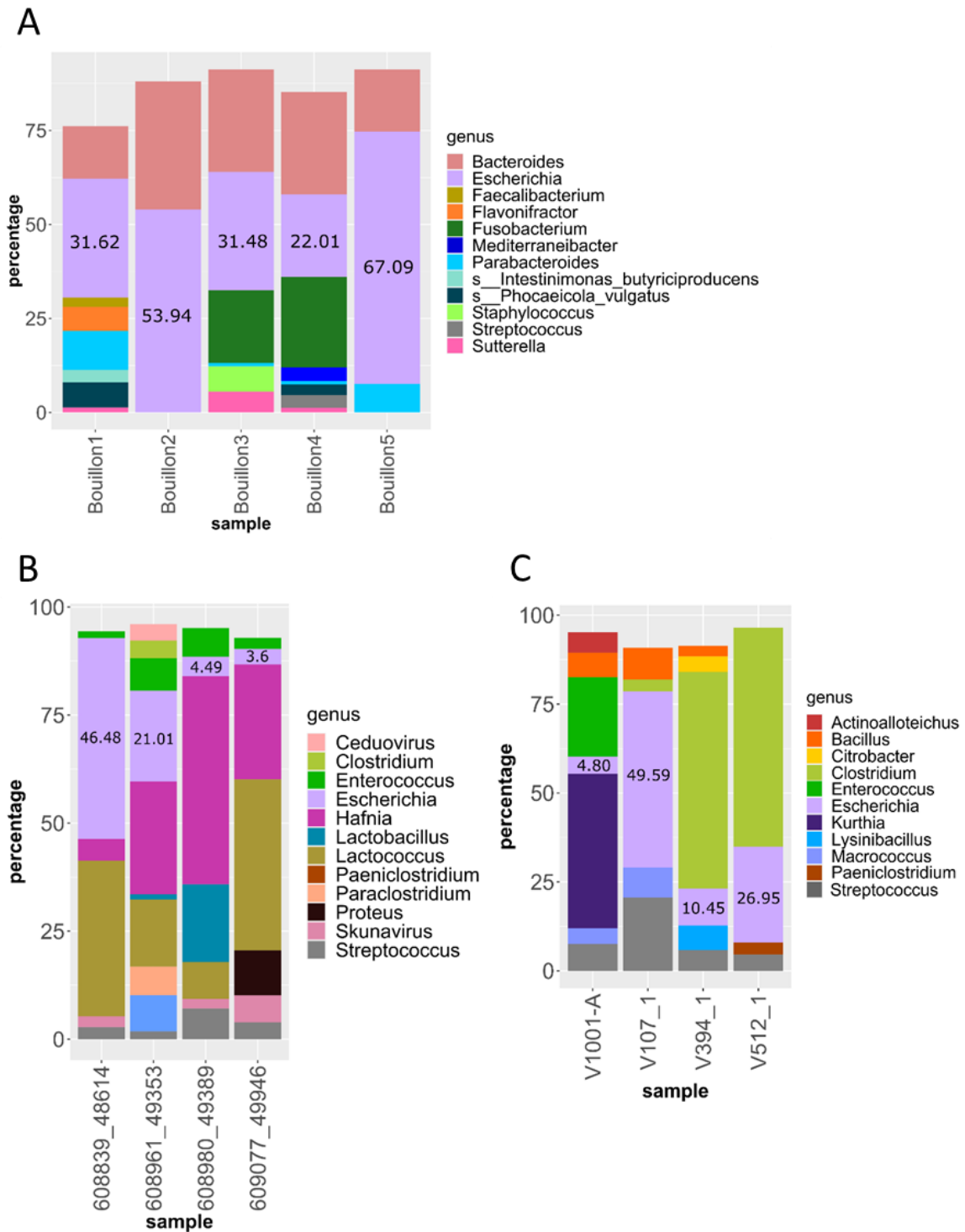


Figure 16 : Barplot representations of the most abundant genus (>2%) detected in long-read sequencing from A: feces or rectal swab samples from contaminated patients, B: presumptive-positive raw milk cheese enrichments (37°C for 18-24h without acriflavine) and C: presumptive-positive ground beef enrichment broths (37°C for 18-24h without acriflavine). Values represent *Escherichia* reads abundance in percent.



Table 9 presents quantification results on *stx* and *eae* genes, but also *E. coli* generic marker (either *wecA* or *cdgR*) as well as *espK* and *espV* in number of copies per mL of DNA extract (copies.mL<sup>-1</sup>). Quantification results show that the STEC did not reach the required level post-enrichment to be fully characterizable in all raw-milk cheese samples, three clinical samples and three ground beef enrichments. In addition, the overall amount of *E. coli* data sequenced was very low (Fig. 16) and several *E. coli* strains appeared to be present in the samples (Table 8), probably at higher concentrations than the STEC (*wecA* or *cdgR*) which presumably prevented its proper characterization.

Indeed, from all raw-milk cheese samples, the *stx* gene was detected in only one sample (608839) with a read depth of 8. In this sample, an *E. coli* MAG was obtained but the assembly was fragmented (44 contigs) making it impossible to characterize the strain (Table 8). The presence of *espK* and *espV* markers was quantified to 10<sup>5</sup> copies.mL<sup>-1</sup> in only one raw-milk cheese sample (608961) using qdPCR which was below the threshold (10<sup>8</sup> copies.mL<sup>-1</sup>). Following our workflow, it was not possible to detect the STEC as it did not grow to the required level, many *E. coli* strains were present (1:100 ratio of STEC:other *E. coli*) and represented only 21.01% of the genus sequenced in the sample.

In all clinical samples, the presence of an *eae*-positive STEC was confirmed by the NRC. Using qdPCR, *eae* and *stx* genes were quantified below the threshold as represented Table 9 and were not detected (<3x) in 3/5 samples (except Bouillon 4, *stx2* read depth of 11x). Indeed, these three samples (Bouillon2-4) had low contamination levels as they were detected with late Ct-values by the RD hospital. On the contrary, in samples Bouillon 1 and Bouillon 5, both *stx* and *eae* genes were quantified above the threshold value using qdPCR (Table 9) and were detected in each sequenced data (Bouillon1 and Bouillon5 in Table 8). However, due to the low proportion of *Escherichia* data (Fig. 16) and the presence of different *E. coli* strains (Table 8), the resulting assembly showed *stx* and *eae* genes on two different contigs, which did not allow characterization of an *eae*-positive STEC. Although characterization was not possible using the STECmetadetector, the machine learning approach allowed the identification of *eae*-positive STEC in these two samples (Table 8).

Similarly, in ground beef, both *stx* and *eae* genes were located on different contigs for one sample (V394) using the STECmetadetector pipeline but the presence of an *eae*-positive STEC was identified with the ML approach. In this sample, the genetic markers were quantified above the determined threshold. However, the low amount of data generated led to a genome coverage of 9x which was not sufficient to correctly assemble the *eae*-positive STEC genome (required 35x coverage was determined previously as described in Publication 3). Two samples did not contain the *stx* gene (V512 and V107). One of these samples was found negative for *espK* and *espV*, suggesting the presence of an EPEC and an STEC simultaneously. However, sample V512 had low concentration of STEC post-enrichment (around 10<sup>4</sup> copies.mL<sup>-1</sup>), which, combined with the presence of multiple strains and low proportion of *Escherichia* reads (26.95%) did not allow its detection. Nevertheless, one sample (sample V1001) was characterized as an *eae*-positive STEC with both genes on the same contig using the STECmetadetector pipeline. Although this sample showed a low proportion of *Escherichia*

reads (4.8%) and the quantification did not reach the threshold ( $10^7$  copies.mL<sup>-1</sup>), it was the only *E. coli* strain present in the sample (Table 8, 9; Fig. 16).

Interestingly, in all raw-milk cheese samples, beef samples and in clinical samples where no STEC was characterized, quantification of *eae*, *stx* showed quantification levels lower than the threshold of  $10^8$  copies.mL<sup>-1</sup> that was determined to be necessary to characterize the STEC. Yet, for samples in which the threshold was reached (Bouillon1, Bouillon5), it is the presence of additional *E. coli* or the low amount of data (V394) that impeded characterization. In fact, a STEC:*E. coli* ratio below 10:1 did not permit the characterization of the strain, though the two genes were detected in the generated assembly. The identification of an *eae*-positive STEC of serotype O177:H25 using both the STECmetadetector and ML is of importance since similar strains have previously been described from cattle (*stx2a* or *stx2d*, *eae*- positive) (Montso *et al.*, 2022, 2019; Sheng *et al.*, 2018). In sample V394, the presence of an *eae*-positive STEC was determined using the ML approach. As few data were sequenced belonging to that strain, the STECmetadetector could not show the presence of both genes in the same genome. Yet, the detection of O182:H25 is also of significant importance as *eae*-positive STEC of this serotype were also reported in cattle (Mussio *et al.*, 2023; van Hoek *et al.*, 2023) and belongs to the main non-O157 O-groups reported to cause human illness in Europe (EFSA and ECDC, 2017). The machine learning approach allowed us to confirm the presence of an *eae*-positive STEC. However, it does not enable any characterization of the strain. By combining the two approaches, it could be possible to find characteristics of the *eae*-positive STEC identified since it is based on the same assembly.

Nevertheless, the ML also depends on the quantity of *E. coli* data sequenced that proved difficult to reach on clinical samples (Table 9). In addition, both the low percentage of *E. coli* reads and the presence of multiple strains prevented the sequencing of enough STEC data. Indeed, as it is based on assembly it needs at least three reads that carry each gene, otherwise it will not be included in the Flye assembly. Additionally, as observed on clinical samples, the limits may lead to false-negative samples as the *eae*-positive STEC present in Bouillon 3-4-5 was not even detected but were confirmed by the NRC, Table 8. Similarly, in ground beef, the presence of an *eae*-positive STEC of serotype O157:H7 was not detected using our approach but was isolated by the NRL *E. coli*. Here, an optimized enrichment broth might have reduced the proportion of background bacteria resulting in a higher sequence data output for the STEC strain and thus, preventing false-negative results with our method.

Overall, the results obtained on presumptive-positive or naturally contaminated samples confirmed the previously identified limitations of the developed workflow.

Table 9 : Quantification results obtained from extracted DNA on *eae*-positive STEC virulence factors using qdPCR.

Quantified virulence markers were *stx1*, *stx2* (or *stx2a\**), *eae*, *espK* and *espV*. *E. coli* generic marker was also quantified using either *cdgR* or *wecA*.

Matrix	Sample	<i>stx1</i>	<i>stx2 (stx2a*)</i>	<i>eae</i>	<i>wecA/cdgR</i>	<i>espK</i>	<i>espV</i>	Pathotype	Machine learning	Presence
Milk	608961	2.32.10 <sup>6</sup>	negative	2.50.10 <sup>6</sup>	2.78.10 <sup>8</sup>	4.18.10 <sup>6</sup>	1.85.10 <sup>6</sup>	NON-STE <sup>C</sup>	No	?
Milk	608839	negative	negative	7.28.10 <sup>4</sup>	1.98.10 <sup>8</sup>	negative	negative	NON-STE <sup>C</sup>	No	?
Milk	608980	2.43.10 <sup>6</sup>	negative	2.09.10 <sup>6</sup>	8.25.10 <sup>6</sup>	negative	negative	STE <sup>C</sup>	No	?
Milk	609077	negative	1.86.10 <sup>5</sup>	na	9.31.10 <sup>6</sup>	negative	negative	NON-STE <sup>C</sup>	No	?
Ground beef	V1001	negative	3.78.10 <sup>7</sup>	3.37.10 <sup>7</sup>	10 <sup>7</sup>	negative	2.89.10 <sup>7</sup>	EHEC?	Yes	Not sent to NRL
Ground beef	V107	<b>negative</b>	<b>negative</b>	<b>negative</b>	na	negative	negative	NON-STE <sup>C</sup>	No	Not sent to NRL
Ground beef	V394	negative	2.11.10 <sup>7</sup>	1.36.10 <sup>8</sup>	1.66.10 <sup>8</sup>	negative	1.15.10 <sup>8</sup>	EHEC?	Yes	Not sent to NRL
Ground beef	V512	negative	1.14.10 <sup>5</sup>	1.25.10 <sup>5</sup>	8.68.10 <sup>8</sup>	8.87.10 <sup>4</sup>	9.82.10 <sup>4</sup>	NON-STE <sup>C</sup>	No	Yes O157:H7, <i>stx2+</i> , <i>eae+</i>
Feces	Bouillon1	negative	8.12.10 <sup>7*</sup>	2.19.10 <sup>8</sup>	<b>1,58.10<sup>8</sup></b>	1.78.10 <sup>8</sup>	1.80.10 <sup>8</sup>	EHEC?	Yes	Yes O26:H11, ST21, <i>stx2a</i> , <i>eae</i> beta
Feces	Bouillon2	negative	5.19.10 <sup>4</sup>	7.65.10 <sup>4</sup>	>10 <sup>5</sup>	3.32.10 <sup>4</sup>	3.44.10 <sup>4</sup>	NON-STE <sup>C</sup>	No	Yes O26:H11, ST21, <i>stx2a</i> , <i>eae</i> beta
Rectal swab	Bouillon3	negative	2.87.10 <sup>4*</sup>	5.43.10 <sup>5</sup>	<b>5,09.10<sup>7</sup></b>	4.97.10 <sup>5</sup>	3.85.10 <sup>5</sup>	NON-STE <sup>C</sup>	No	Yes O26:H11, ST21, <i>stx2a</i> , <i>eae</i> beta
Feces	Bouillon4	5.71.10 <sup>7</sup>	3.33.10 <sup>6*</sup>	6.5.10 <sup>7</sup>	<b>4,23.10<sup>7</sup></b>	5.21.10 <sup>7</sup>	4.97.10 <sup>7</sup>	NON-STE <sup>C</sup>	No	Yes O157:H7, ST11, <i>stx1a</i> , <i>stx2c</i> , <i>eae</i> gamma
Rectal swab	Bouillon5	negative	1.37.10 <sup>8*</sup>	1.19.10 <sup>8</sup>	<b>1,96.10<sup>8</sup></b>	9.04.10 <sup>7</sup>	8.01.10 <sup>7</sup>	EHEC?	Yes	Yes O26:H11, ST21, <i>stx2a</i> , <i>eae</i> beta

---

## Chapter 6: Discussion

---

The current methods for STEC detection and characterization from food matrices can leave a significant proportion of samples as presumptive positives, which cannot be confirmed with a successful isolation of strains. This situation is a conundrum for all involved parties: food producers, reference laboratories as well as decision makers. We endeavored to develop new approaches to characterize STEC strains directly from food samples, without the time-consuming and laborious isolation steps. The Metadetect project aimed at assessing the potential of new long-read metagenomics approaches to better characterize STEC from contaminated raw milk. The work accomplished during this project allowed us to propose a complete workflow from raw milk enrichment to data analysis for full STEC characterization, as represented on Figure 17. Altogether, these approaches show the potential of long-read sequencing for identifying and when possible characterizing STEC from various food samples but also show the limits faced at that time (Fig. 17, represented with caution signs).

Following the complete workflow, we demonstrated that a major limitation of our method is the requirement of a minimum of  $10^8$  copies.mL<sup>-1</sup> of STEC post-enrichment/prior to sequencing to characterize the STEC when no additional background *E. coli* is present, in raw milk. We have shown that the presence of multiple *E. coli* strains hampered STEC characterization and determined that the proportion of STEC to commensal strains should be above 10:1 to be characterized. Although the quantity of STEC data was also shown to be important for the machine learning-based approach, the latter was not sensitive regarding the presence of other *E. coli* strains. When applying the developed method on enrichment broths of different matrices: beef, raw-milk cheese and feces kindly provided by collaborating laboratories, we confirmed these were the limiting conditions. Here, we will discuss the complete workflow with its limits, possible improvements that may be envisaged and its application in real life.

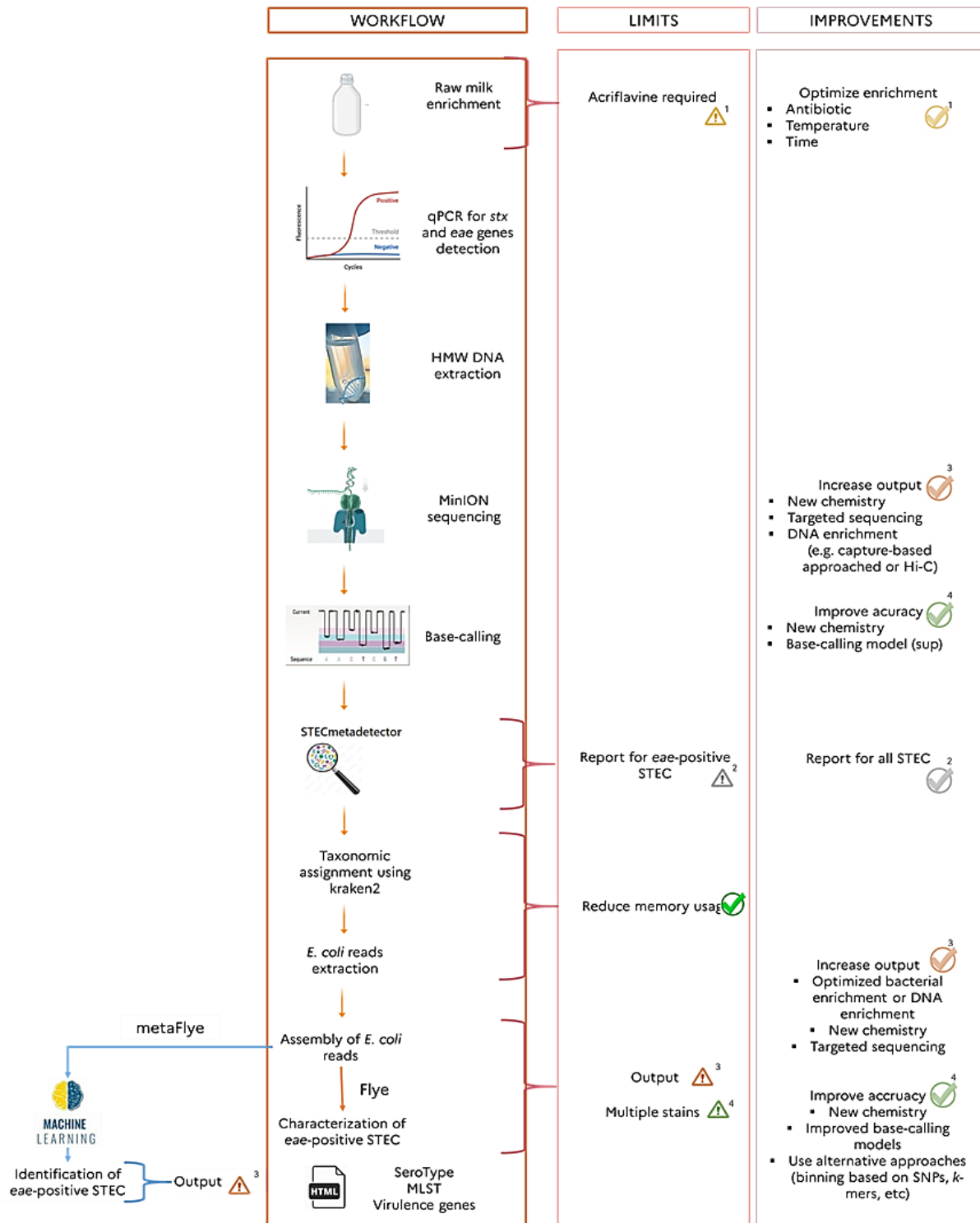


Figure 17: Representation of the workflow developed for *eae*-positive STEC characterization from raw milk and its limits.

The first two triangles do not represent limits by themselves but points that can be improved. The limits of the methods are represented with caution signs 3 and 4 for the assembly limits regarding the quantity of STEC data and the characterization limit in a multiple strain sample, respectively. Possible improvements are represented with validation signs at different steps of the workflow and will be discussed here.

## 1. Determine the enrichment conditions

As represented in Figure 17 (caution sign 1), we first included an enrichment step. For this, I used *eae*-positive STEC MinION sequencing data of isolated strains to compare the amount of data required to generate a contiguous assembly with the amount of data generated from raw milk with our method. I found that most STEC assemblies were complete (chromosome in 1 or 2 contigs) at 35x genome coverage using Flye assembler (Kolmogorov *et al.*, 2020; Maguire *et al.*, 2021). The study of Maguire and colleagues determined that STEC could be completely assembled from  $10^{7-8}$  cfu.mL<sup>-1</sup>, in wastewater (Maguire *et al.*, 2021). Such high concentrations are rarely reached in naturally contaminated samples, as STEC contamination may be low (Caprioli *et al.*, 2005; EFSA, 2020). Hence, I included an enrichment step to reach the required STEC concentration ( $10^8$  CFU.mL<sup>-1</sup>, (Maguire *et al.*, 2021)) and genome coverage of 35x from raw milk samples. This step is necessary as STEC concentration below the threshold of  $10^8$  copies.mL<sup>-1</sup> resulted in the generation of insufficient data; resulting into fragmented assemblies that did not allow STEC characterization (Fig. 17, caution sign 3).

The ISO/TS-13136:2012 technical specification recommends an enrichment step for 18-20h at 37°C in BPW supplemented with acriflavine at a final concentration of 12 g.L<sup>-1</sup> for dairy products (ISO, 2012). However, the use of acriflavine is widely discussed as it was shown to affect the growth of some STEC strains (Amagliani *et al.*, 2018; Mancusi and Trevisani, 2014). A revision of the ISO/TS-13136 discussed at that time suggested to enrich at 41.5°C without using acriflavine. I compared the growth of *eae*-positive STEC of serotype O26:H11 in raw milk using both temperatures with or without acriflavine based on the proportion of reads obtained after sequencing and the contiguity of the assembly. The use of acriflavine was beneficial for the two *eae*-positive STEC strains used in this study, as it reduced the growth of Gram-positive bacteria, which are usually predominant in raw milk and can impede STEC growth thereby reducing the proportion of *E. coli* reads available for analysis. Therefore, I included acriflavine in the enrichment broth (Fig. 17, caution sign1). When we applied the method on presumptive-positive or naturally contaminated samples for which all enrichments were performed at 37°C without acriflavine, low proportions of *E. coli* were detected and growth to the required quantities was not achieved for most of the samples. More studies on the effect of acriflavine on different STEC strains together with the information about their

pathogenicity potential could provide information on whether the method is suitable for application on naturally contaminated samples for all STEC strains.

An ideal case would be to remove the enrichment step, though it is not an additional step compared to the ISO/TS-13136:2012 as is it performed before the detection of *stx* and *eae* genes. For screening of specific genes using qPCR, enrichment conditions do not necessitate to be as stringent since it amplifies specific targets and is sensitive regardless of the presence of additional strains. It has been shown that shotgun sequencing approaches (without enrichment) could rapidly detect or characterize low-abundant pathogens directly from food matrices (without enrichment) (Grützke *et al.*, 2021; McArdle and Kaforou, 2020). However, characterizing STEC carrying an adhesion system acquired from other *E. coli* or organisms is more complex. Without enrichment, Azhineiro and colleagues could detect the presence of the *stx* gene within 2 hours of MinION sequencing, but not characterize the STEC strain (Azhineiro *et al.*, 2021). Buytaers and colleagues achieved strain-level characterization of STEC from artificially contaminated beef samples after 12 hours of MinION sequencing without previous enrichment by using host DNA depletion kit, considerably reducing host DNA (Buytaers *et al.*, 2021). However, they sequenced a single sample per flow cell, thereby considerably increasing the amount of data generated by sample; but also considerably increasing the cost. To avoid enrichment, host depletion kits can remove some host DNA, but also results in low DNA yields that are typically below ONT requirements. In this study, I have shown that, to date, the enrichment step is necessary as we aimed at characterizing *eae*-positive STEC and not detect or identify STEC presence solely based on *stx* gene. In addition, including an enrichment step provided a beneficial side-effect as it reduces the host/bacterial DNA ratio. Further enrichment conditions optimization can include the use of antibiotics, adapting temperature or reducing incubation time (Fig. 17, improvement sign 1). It is noteworthy that here I used the classical enrichment time of 18-24h but this time could probably be reduced to 8 hours that corresponds to *E. coli* stationary phase in BPW.

## 2. Extracting HMW DNA from raw milk

The next step is the extraction of HMW DNA from raw milk. Indeed, the extraction of a high quantity of pure intact gDNA is a prerequisite for MinION sequencing. At the start of this project, several studies comparing extraction methods to obtain DNA suitable for MinION sequencing from different matrices were available, but none specific to STEC, and none from raw milk (Bouso and Planet, 2019; Ghaheri *et al.*, 2016; Mayjonade *et al.*, 2016; Penouilh-Suzette *et al.*, 2020; Schalamun *et al.*, 2019). DNA extraction methods from raw milk focused on subsequent qPCR or short-read sequencing analysis, which do not take into account DNA integrity/degradation, a critical parameter for MinION sequencing (Cremonesi *et al.*, 2021; Parente *et al.*, 2020; Quigley *et al.*, 2012; Siebert *et al.*, 2021; Usman *et al.*, 2014). Our objective was to find (i) a DNA extraction method with an efficient cell lysis step (to recover the maximum amount of DNA), while being gentle enough (to protect the DNA from degradation and obtain contiguous STEC assemblies after sequencing using ONT technology) and (ii) a purification step efficient enough to remove raw milk contaminants (Parente *et al.*, 2020; Porcellato *et al.*, 2021).

Although many studies favored the phenol-chloroform DNA extraction method (Bouso and Planet, 2019; Ghaheri *et al.*, 2016; Maghini *et al.*, 2021; Mayjonade *et al.*, 2016; Schalamun *et al.*, 2019; Trigodet *et al.*, 2022), we excluded this approach for practical reasons in favor of commercially available DNA extraction kits. These kits usually include a cell lysis step (mechanical, enzymatic or chemical) followed by a DNA purification step performed on magnetic beads or on a column on which the DNA is captured/retained and then eluted, or by precipitating (salting-out) the DNA. Both the cell lysis procedure and the purification method have an impact on quantity, quality and integrity of DNA extracted. In our comparison, we used only enzymatic cell lysis procedure and tested the three different DNA purification methods on STEC pure cultures. Enzymatic cell lysis combined with bead-based or salting-out DNA purification methods proved to be more gentle and avoided DNA shearing (Gand *et al.*, 2023; Salonen *et al.*, 2010; Zhang *et al.*, 2022), although salting-out showed the lowest degradation profile (Eagle *et al.*, 2023; Jones *et al.*, 2021; Nouws *et al.*, 2020; Schalamun *et al.*, 2019; Trigodet *et al.*, 2022). The DNA extracted using the bead-based (AMPureXP) and the salting-out (MasterPure) kits, which were the only two allowing the extraction of DNA matching ONT requirements, was sequenced and we compared their performance on STEC genome assembly.



Both methods were shown to be suitable to obtain complete STEC assemblies. Similar results were obtained with *Salmonella* (Eagle *et al.*, 2023). However, the salting-out methods allowed the generation of longer reads and higher coverage (Eagle *et al.*, 2023). Longer reads are important as they can span the numerous repeated regions present in STEC (mostly phage-related sequences over 7 kb) to produce contiguous STEC genome assemblies (Koren *et al.*, 2013).

As this approach aimed at characterizing STEC from raw milk, I have further verified that the salting-out method (MasterPure kit) was efficient on raw milk samples. From raw milk, this method allowed the extraction of sufficient quantities of long DNA fragments to perform MinION sequencing, while removing contaminants. With this study, I could select an appropriate method to extract HMW DNA suitable for MinION sequencing from raw milk with the capacity to generate contiguous STEC assemblies. Nevertheless, the freezing process of raw milk was responsible for DNA degradation, and I recommend working on fresh raw milk. Additionally, I tested the salting-out (MasterPure kit) DNA extraction method on ground beef and raw-milk cheese enrichment broth (presumptive-positive samples) and feces (or rectal swab) samples (STEC contamination confirmed) from which high quantities of DNA could be recovered but the quality (purity and integrity) was not as good as expected. This demonstrates that each matrix has its own specificities and that the DNA extraction method should be optimized accordingly for each specific experimental study.

### 3. Screening of virulence genes *stx* and *eae* using qPCR

DNA extracts are further screened for the presence of *stx* and *eae* genes. Positive samples as detected by qPCR performed according to the ISO/TS13136:2012 are then sequenced and processed as developed here. So far, the qPCR methods remain sensitive and cost-effective.

In the future, it could be envisaged to omit the qPCR step if sufficient data are reliably generated after raw milk enrichment. Indeed, direct sequencing of food samples using the MinION was shown to be sensitive to detect STEC (based on *stx* gene presence) (Maguire *et al.*, 2023). Furthermore, while the sequencing cost currently constitutes an obstacle, it has decreased steadily over the last decade and may continue to do so to the point where it may be a viable alternative; especially considering that, contrary to qPCR equipment, MinION sequencing does not require any investment cost or maintenance cost.

In our case, we did not aim at detecting but characterizing *eae*-positive STEC that requires an enrichment and the screening step to reduce the number of samples to sequence. Nevertheless, if we could combine both the potential to generate sufficient data from raw milk without the need to enrich and perform direct MinION sequencing, it would save time to testing laboratories considerably. Some improvements still have to be considered regarding the amount of data generated and in terms of data accuracy to allow the characterization in presence of multiple strains. These improvements are discussed in Section 4.3, Discussion.

#### 4. Developing a tool for STEC assembly from long-read metagenomics data

##### 4.1. From base-calling to STEC characterization

The developed method, except from being optimized for an application in long-read sequencing of raw milk, does not deviate from the ISO recommendations, so far. Here I tested the use of MinION sequencing to characterize STEC instead of the isolation step followed by isolate characterization performed using qPCR and/or short-read sequencing. I sequenced 5-6 samples per MinION flow cell and performed base-calling locally on a GPU-enhanced computer to use models with increased accuracy (hac or sup model). As bioinformatics skills are needed to analyze MinION sequencing data (post-base-calling), I aimed at developing a pipeline to facilitate data analysis. We developed a snakemake pipeline that allows reproducible analysis of MinION sequencing data and automatic analysis without the need to run the different commands manually (Mölder *et al.*, 2021). In addition, it runs different jobs in parallel, which saves time and computing resources.

After filtering and trimming the sequencing data, the STECmetadetector pipeline performs a taxonomic assignment of the reads using kraken2 (Wood *et al.*, 2019). In classical metagenomics analysis, reads are first assembled using a metagenome assembler and resulting contigs are sorted into bins for further classification and reconstruction of genomes into so-called MAGs (Metagenome-assembled genomes) (Goussarov *et al.*, 2022). However, metagenomics assemblies require many resources, in particular memory usage that can be a limiting factor. To circumvent this problem, I reduced memory usage and the amount of required resources by extracting *E. coli* reads (Fig. 17) (Siekanić Grégoire, 2021). *E. coli* reads are screened for the presence of multiple strains based on the O-groups as well as on *stx* and

*eae* genes, which serves as a quality control and informs on the subtypes of the genes that are epidemiologically important (e.g., *stx2d*, *stx2a*, as discussed in Section 9.3, Chapter 2). *E. coli* reads are then assembled using Flye or Canu (Kolmogorov *et al.*, 2020; Koren *et al.*, 2017). Each assembler has its own properties but both Flye and Canu allow plasmid reconstruction, an important feature for STEC assemblies (Johnson and Nolan, 2009; Safar *et al.*, 2023; Wick and Holt, 2019). Although Flye performed better for assembling STEC genomes using MinION data, as Canu produced longer, repetitive assemblies and was time-consuming (Sanderson *et al.*, 2023; Wick and Holt, 2019), we included an option to the pipeline for the usage of Canu. The STECmetadetector pipeline produces a html file summarizing the main results including the microbial flora, the detected O-groups, *stx* and *eae* genes and the characterization of the generated assembly (*E. coli* virulence genes, serotype and MLST, Fig. 17).

#### 4.2. Further use of the STECmetadetector pipeline

Here we generated a pipeline to characterize STEC from MinION data, with a specific attention regarding *eae*-positive STEC. However, this pipeline can be applied to characterize any pathogenic *E. coli*, as for example hybrid *E. coli*. However, for reasons of convenience, only STEC are mentioned in the final report file (Fig. 17, caution sign 2). In fact, not only the whole *E. coli* virulence genes database from the CGE and specific EHEC markers are included, but also, custom databases may be used. I had the opportunity to participate in a proficiency test organized by the VTEC EURL (Rome) and applied the developed method on artificially contaminated cheese samples. I could correctly identify the presence of an EAEC (*aggR*-positive) strain of serotype O104:H4 ST678 and an STEC (*stx1c* - positive) of serotype O178:H4 in the samples. The absence of microbial flora allowed us to characterize the strains. However, the generation of contiguous assemblies was not possible. This is most probably related to the matrix itself as the A260/A230 purity ratio was low and impurities might have blocked pores resulting in lower flow cell output compared to what we obtained from raw milk. Indeed, slightly less than 300 Mb of *Escherichia* reads were obtained per sample, which corresponds to a genome coverage below 50x. Although 35x coverage was determined to be sufficient, the generated reads were fragmented with an N50-value below 4 kb (N-50 value of 2 250b for sample 1 and of 3 500b for sample 2).

(Presentation is available at [https://www.iss.it/documents/20126/8707308/221011\\_EURL-VTEC\\_Rome\\_JAUDOU.pdf/bbdc373-2384-18ad-7c57-d2e63f67c4c9?t=1684859293955](https://www.iss.it/documents/20126/8707308/221011_EURL-VTEC_Rome_JAUDOU.pdf/bbdc373-2384-18ad-7c57-d2e63f67c4c9?t=1684859293955)).

### 4.3. Assembly limits

The main limitation faced using this approach is at the assembly level. Not only should the quantity of STEC be sufficient for assembling the complete genome, but also the presence of multiple *E. coli* strains may lead to STEC characterization failure. Increasing the output of STEC data might help characterizing the strain using an assembly-based approach (Fig. 17, improvement sign 3). It is important to enrich with the proposed conditions to maximize the chances of attaining the required threshold of  $10^8$  copies.mL<sup>-1</sup> of STEC post-enrichment and 35x genome coverage (Fig. 17, caution sign 1). Despite the enrichment conditions of raw milk, further improvements might be envisaged as enriching the DNA of the bacteria rather than the bacteria itself. PCR-based enrichment that targets specific genomic regions has the advantage of requiring low DNA inputs and low hands-on time. Other approaches such as hybridization with for example, capture methods were tested to enrich in the plastid genomes of various plant species (Bethune *et al.*, 2019). We could imagine applying those methods by targeting *E. coli* core genome and *eae*-positive STEC specific virulence factors (*stx*, *eae*, Sp4, Sp6, SpLE1 and SpLE3). Despite targeting specific regions of the genome, the capture of long fragments could allow enrichment of the whole genome. Alternatively, Hi-C methods, which captures DNA-DNA interaction *in vivo*, were shown to help *de novo* assembly as well as assigning phages to their host from metagenomics data and resolve MAGs (DeMaere and Darling, 2019; Hill *et al.*, 2022; Stalder *et al.*, 2019). Exploring the application of Hi-C methods to long-reads to allow cross-linking of DNA fragments for resolving the carriage of *eae* and *stx* genes by the same cell would be of interest. At sequencing level, ONT is developing targeted sequencing, in which only the reads that match a reference genome are sequenced, which could help in sequencing only reads mapping to *E. coli* genome and shorten data processing. Nevertheless, due to the diversity found within *E. coli* strains, some genomic features might be excluded. Overall, these improvements might lead to higher STEC data output (Fig. 17, improvements 1 or 3).

Despite improving the outcome of STEC data from MinION sequencing, further improvements regarding accuracy could increase differentiation level of *E. coli* strains present simultaneously in the sample, as represented on Figure 17 with caution sign 4. In fact, assemblers require at

least 5% divergence to distinguish individual genomes (Luo *et al.*, 2022) which corresponds to the species delineation limit (Ciuffo *et al.*, 2018; Jain *et al.*, 2018; Richter and Rosselló-Móra, 2009). Despite the variety present within *E. coli* strains, it appears difficult to distinguish different *E. coli* strains with current long-read assemblers. Therefore, Flye generated fragmented assemblies when several *E. coli* strains were present (Vicedomini *et al.*, 2021; Wick and Holt, 2019). Most STEC carry virulence factors acquired from MGEs via HGT, which vary in codon usage and GC content and may not be recognized by the assembler as belonging to the STEC. The LEE (which carries *eae* gene) is a good example, as its GC content (38%) is different from the *E. coli* core genome (50%). In this study, when more than one *E. coli* strain was present in the sample, the LEE was consistently isolated on a separate contig. In a recent study, a mixture of *E. coli* strains including EPEC, EAEC, ExpEC and STEC, with STEC to other *E. coli* ratio of 17:100 and 2:100, was processed using the MinION and its data analysis program (WIMPS). It is noteworthy that they did not aim at characterizing STEC but detecting STEC in case of phage loss. Yet, when the STEC proportion was 2% compared to other *E. coli*, it could not even be detected as the *stx* gene was not assembled (Tunsjø *et al.*, 2023). These studies also support the fact that full STEC characterization is hampered by the endogenous *E. coli* strains present.

As the presence of multiple *E. coli* strains is an obstacle to the assembly of the LEE in the STEC chromosome, it is crucial to find post-assembly approaches that can differentiate the presence of *eae*-positive STEC from the presence of an EPEC and an STEC strain (Fig. 17, improvement sign 4). Recently, strainberry was developed to separate multiple strains from the same species from long-read sequencing data post-assembly (Vicedomini *et al.*, 2021), which was included (the only tool available at that time) in the STECmetadetector pipeline. However, it did not perform well on our samples and generated chimeric assemblies. More recently, a pipeline using a similar approach was developed for long-read metagenomics named MetaBooster pipeline that uses VeChat and Canu to correct the reads prior to assembly, and uses strainberry to separate the strains, which should generate higher strain-resolutive assemblies (Luo *et al.*, 2022). The results observed in this project highlight the need for strain-aware assemblers or alternative approaches that circumvent problems encountered when using current long-read assemblers.

## 5. - Alternative approaches: machine learning

Although the assembly is an ideal tool for characterizing STEC strains, the use of alternative approaches might help identifying potentially pathogenic STEC. Assembling genomes from metagenomics data typically results in highly fragmented assemblies with scaffolds, which may belong to the same genome. In this case, it is unlikely that *stx* and *eae* genes, which may be more than 2 Mb apart on the chromosome, are co-localized on the same contig. As we have shown, the quantity of STEC should be 10-times higher than other *E. coli* strains. Finding an alternative approach that allows the identification of *eae*-positive STEC even at low coverage and when the STEC to other *E. coli* proportion is comparable, would significantly improve the method.

Here, we aimed at finding another approach assessing the presence and co-localization of *stx* and *eae* genes. Within the diversity of *E. coli* strains some only carry the *stx* gene (STEC), others only the LEE (EPEC), while some have both *stx* and *eae* genes (typical EHEC, or *eae*-positive STEC causing HUS). The simultaneous presence of these genes in a strain leads to two hypotheses. Either the presence of these two genes in a genome is random or their co-integration and/or maintenance in a chromosome necessitates a specific genetic background. Although the *stx*-phage integration may be mediated to different *E. coli* strains, different studies have shown that certain genetic markers (mainly T3SS effectors) appeared to be preferentially associated to typical EHEC (*eae*-positive STEC), constituting a genetic signature for their identification (Bugarel *et al.*, 2010a, 2010b; Coombes *et al.*, 2008; Delannoy *et al.*, 2016, 2013b, 2013a; Imamovic *et al.*, 2010; Karmali *et al.*, 2003; Konczy *et al.*, 2008). Moreover, these genetic markers in combination with *stx* and *eae* were proved to be efficient to lower the number of presumptive-positive raw milk samples (Delannoy *et al.*, 2022). Machine learning approaches have shown their potential in predicting the pathogenic potential of STEC based on their genetic profile (Im *et al.*, 2021; Njage *et al.*, 2019) or in source attribution studies (Lupolova *et al.*, 2021, 2016). Altogether, these studies show that the diversity within *E. coli* strains, including *eae*-positive STEC, might depend on their genetic background.

To test this hypothesis, we constructed a database of approximately 1 425 complete (or almost complete) *E. coli* genomes from NCBI GenBank further annotated using the Sakai reference genome. Using a pan-genomic analysis and the power of machine learning, we could identify 6 markers which combination of presence/absence could accurately predict the presence of *eae*-

positive STEC from assemblies even when additional *E. coli* strains were present in a ratio of STEC to commensal below 10:1. Interestingly, those markers were located on genomic island carrying the genetic markers identified previously as associated with *eae*-positive STEC, particularly Sp4 (OI-44) and Sp6 (OI-50) (Delannoy *et al.*, 2016). This method is still based on assembly and depends on the enrichment step performance as it requires enough STEC to be sequenced, but appears to be a great alternative to the STECmetadetector in case where other *E. coli* strains are present (Fig. 17, caution sign 3). We could precisely identify the presence of *eae*-positive STEC by analyzing the presence/absence of six genetic markers assessing their carriage by the same cell.

The small number of markers used by the machine learning algorithms provides an advantage in the case of low abundance of *eae*-positive STEC, as it only requires the data for six genes. Indeed, characterization of *eae*-positive STEC using assembly only depends on the presence of highly related bacteria in the samples and the proportion of data generated for the STEC compared to the other *E. coli*. However, the machine learning approach is less sensitive to the presence of multiple strains and requires the six markers to be sequenced to at least 3x (depth) for being included in the generated assembly. Moreover, the reduced number of markers allows further development for an application in the laboratory using PCR methods (including a short algorithm to interpret the results and conclude on the presence of *eae*-positive STEC in the sample). The machine learning approach can be implemented in the STECmetadetector pipeline for identification of *eae*-positive STEC. A further development could include the detection of these genetic markers during the PCR analysis and positive samples processed using the STECmetadetector pipeline on MinION sequencing data for characterization of the strain. Altogether, these results also tend to show that there should be a particular genetic structure favoring the gathering of *stx*-phage and the LEE in a same chromosome, though it does not prove it.

## 6. Other alternative approaches

Overall, if the conditions determined to be necessary (STEC quantified to  $10^8$  copies.mL<sup>-1</sup> post-enrichment and minimum STEC:commensal ratio of 10:1) are not met, it is currently not possible to fully characterize *eae*-positive STEC from naturally contaminated raw milk samples using the assembly-based approach and the tools available (Tunnsjø *et al.*, 2023). Indeed, a species-level consensus assembly is generated suppressing the strain-level variability. However, with the increasing number of tools and approaches being developed to achieve strain-level characterization, this may soon become a reality.

Different post-assembly approaches have been developed and some of them were tested in this project. First, binning approaches were tested, although very few programs were developed and only designed for short reads (e.g. MaxBin2, STRAINEST: <https://github.com/compmetagen/strainest>) (Goussarov *et al.*, 2022; Wu *et al.*, 2016). Binning can be performed on reads or contigs based on different parameters as coverage/depth, *k*-mer frequencies, or tetranucleotide frequencies. A promising read binning tool was developed specifically for long reads, LRBinner (Wickramarachchi and Lin, 2022) but was not tested during this project. On the contrary, contig binning approaches, which group scaffolds belonging to the same genome into bins, were tested but generated chimeric contigs. Similarly, another approach that is based on variant calling e.g., used by Strainberry, did not allow the separation of the different *E. coli* strains, instead generating chimeric phases (contigs) (e.g. O26:H10 and O2:H11, in which the H-type were exchanged). Here, long reads are mapped to assembled contigs (long-read sequencing data) leading to the separation of the variants (named haplotypes) that are further assembled and scaffolded (Vicedomini *et al.*, 2021). Recently, stRainy using a similar approach (assembly graph, SNPs, resolve haplotypes) was developed and was shown to perform better than strainberry (Ekaterina Kazantseva *et al.*, 2023). Very recently, a new assembly approach for generating MAGs was developed but was unfortunately not tested in this project. The metaMDBG assembler was tested on a mock community containing 21 strains including 5 *E. coli* strains and generated a single circular contig for the two most abundant *E. coli* strains (Benoit *et al.*, 2023).

All these approaches are promising, yet, the high error rate of MinION sequencing data (using the R.9.4.1 chemistry) limits their efficiency. However, the generation of data with accuracy similar to Illumina might allow the successful application of these post-assembly strain



separation tools. ONT is in the process of strengthening the weak point of their technology: accuracy. First, they constantly improve their base-calling process (to convert the raw signal obtained from the MinION to sequences). So far, they provide three base-calling models with increasing accuracy and time of analysis. The recently released super accuracy model (sup), allows more precise characterization of the strain, especially for analysis such as MLST, frequently undetermined when the data were base-called using the hac model. However, both hac and sup models currently require significant computing resources. Additionally, ONT released different generations of pores and chemistry, and the new generation R.14 should provide data reaching 99.9% of accuracy. Combined with targeted sequencing, it shows that ONT sequencing will potentially be able to generate pertinent accurate data rapidly (Fig. 17, improvement sign 4). Some studies conducted to test ONT sequencing in case of outbreaks have proved its efficacy to detect STEC in 2 hours but not for characterization yet (Azinheiro *et al.*, 2021).

With the improvements of ONT especially regarding the accuracy, other bioinformatics approaches of distinguishing different strains would become available. Indeed, for now the use of *k*-mers, SNPs or variant calling approaches are not ideal considering the error-prone reads. However, with increased accuracy we could imagine the generation of a pipeline which combines a binning step to bin reads according to their abundance, *k*-mers frequency or diversity or even using variant calling approaches, with assemblies to resolve strain-level variations. In particular, *k*-mers approaches seem promising to rapidly identify the presence of *eae*-positive STEC since they are based on reads, are less resource-demanding and quicker than assembly-based and read-binning approaches (Buytaers *et al.*, 2021). In this project, we generated a mash reference using our database composed on 1 425 genomes initially constructed for the machine learning approach. We then developed a python script using mash screen (Ondov *et al.*, 2019) to find the closest genome in this reference to each read in a sample and predict the composition of the sample (in particular, the presence of *eae*-positive STEC). Although it did not work for one sample (V394), it seems promising as it allowed the identification of *eae*-positive STEC in all other samples and *in silico* mixtures tested in this project and correctly indicated the presence of multiple strains in most cases.

---

## Conclusion

---

Overall, in this project we have demonstrated that it is possible to characterize STEC strains (particularly *eae*-positive STEC) from raw milk and that this method may be applicable to other samples such as raw-milk cheese, beef or clinical samples, with further optimization of DNA extraction. The developed method as it is now does not aim at replacing the qPCR screening step but rather the characterization of positive samples performed by reference laboratories on isolated strains. It was shown to be efficient and could avoid the laborious work done to isolate the strains. Nevertheless, the limits observed here on the quantity of STEC data generated and the assembly tools that does not allow strain separation renders the application of the method limited. At present, MinION sequencing is not optimized for application in cases where *eae*-positive STEC urgently needs to be characterized. As we have observed in sample like V512 contaminated with  $10^4$  copies.mL<sup>-1</sup> of O157:H7 *eae*-positive STEC, the strain was isolated by the NRL *E. coli* but not detected by our approach. It shows that the method is currently inadequate when the contamination level is low. For such samples, the traditional workflow currently remains the best. Nevertheless, it is noteworthy that the O157 strain was isolated because its serogroup belonged to the Top-5 and specific antibodies were used during immunoseparation to improve the recovery of the strain and allow its isolation. Isolation procedure without the use of immunoseparation might not be as successful. Furthermore, due to the new pathogenic STEC classifications (Section 9.3, Chapter 2), serogroups are not used. Therefore, the immunoseparation could soon become obsolete. Here, this approach allowed the identification and characterization of samples from the same source highly contaminated with potentially pathogenic *eae*-positive STEC of non-Top5 O-groups that would not have been detected using classical approaches (V394 and V1001). If the use of serogroup-based methods is abandoned in favor to *stx*-subtyping approaches, as recommended by new classifications, our method might be of interest. Additionally, ONT keeps optimizing their technology not only to improve data output and particularly targeted sequencing, but also to increase data accuracy to reach the accuracy of Illumina sequencing. We can imagine that if ONT continues the development of its technology, the enrichment step would one day not be required anymore and that background *E. coli* will not constitute a challenge to characterize STEC from complex matrices. Indeed, bio-informatic approaches based on variant analysis to distinguish strains

would be applicable (such as SNPs, *k*-mers, etc) to characterize different *E. coli* strains including *eae*-positive STEC from raw milk samples.

In the meantime, this work provides a basis for STEC characterization in raw milk and could further be improved. It allows isolation-free characterization of *eae*-positive STEC from raw milk and would be an alternative approach in cases where the strain cannot be isolated.

---

## Literature

---

- Ahmadi, A., Khezri, A., Nørstebø, H., Ahmad, R., 2022. A culture-, amplification-independent, and rapid method for identification of pathogens and antibiotic resistance profile in bovine mastitis milk. *Front. Microbiol.* 13, 1104701. <https://doi.org/10.3389/fmicb.2022.1104701>
- Alkan, C., Sajjadian, S., Eichler, E.E., 2011. Limitations of next-generation genome sequence assembly. *Nat. Methods* 8, 61–65. <https://doi.org/10.1038/nmeth.1527>
- Amagliani, G., Rotundo, L., Carloni, E., Omiccioli, E., Magnani, M., Brandi, G., Fratamico, P., 2018. Detection of Shiga toxin-producing *Escherichia coli* (STEC) in ground beef and bean sprouts: Evaluation of culture enrichment conditions. *Food Res. Int. Ott. Ont* 103, 398–405. <https://doi.org/10.1016/j.foodres.2017.10.059>
- Anses, 2023. Avis relatif à la définition des souches pathogènes d'*Escherichia coli* productrices de Shigatoxines (saisine n°2020-SA-0095). Maisons-Alfort : Anses, 63 p. Available at: <https://www.anses.fr/fr/system/files/BIORISK2020SA0095.pdf>.
- Anses, 2019. Fiche de description de danger biologique transmissible par les aliments - : “*Escherichia coli* Entérohémorragiques (EHEC)” - Mai 2019, available at: <https://www.anses.fr/fr/system/files/BIORISK2017SA0224Fi.pdf>.
- Azineiro, S., Roumani, F., Carvalho, J., Prado, M., Garrido-Maestu, A., 2021. Suitability of the MinION long read sequencer for semi-targeted detection of foodborne pathogens. *Anal. Chim. Acta* 1184, 339051. <https://doi.org/10.1016/j.aca.2021.339051>
- Azineiro, S., Roumani, F., Costa-Ribeiro, A., Prado, M., Garrido-Maestu, A., 2022. Application of MinION sequencing as a tool for the rapid detection and characterization of *Listeria monocytogenes* in smoked salmon. *Front. Microbiol.* 13, 931810. <https://doi.org/10.3389/fmicb.2022.931810>
- Bagel, A., Sergentet, D., 2022. Shiga Toxin-Producing *Escherichia coli* and Milk Fat Globules. *Microorganisms* 10, 496. <https://doi.org/10.3390/microorganisms10030496>
- Bai, X., Fu, S., Zhang, J., Fan, R., Xu, Y., Sun, H., He, X., Xu, J., Xiong, Y., 2018. Identification and pathogenomic analysis of an *Escherichia coli* strain producing a novel Shiga toxin 2 subtype. *Sci. Rep.* 8, 6756. <https://doi.org/10.1038/s41598-018-25233-x>
- Bai, X., Scheutz, F., Dahlgren, H.M., Hedenström, I., Jernberg, C., 2021. Characterization of Clinical *Escherichia coli* Strains Producing a Novel Shiga Toxin 2 Subtype in Sweden and Denmark. *Microorganisms* 9, 2374. <https://doi.org/10.3390/microorganisms9112374>
- Bell, B.P., Goldoft, M., Griffin, P.M., Davis, M.A., Gordon, D.C., Tarr, P.I., Bartleson, C.A., Lewis, J.H., Barrett, T.J., Wells, J.G., 1994. A multistate outbreak of *Escherichia coli* O157:H7-associated bloody diarrhea and hemolytic uremic syndrome from hamburgers. The Washington experience. *JAMA* 272, 1349–1353.
- Benoit, G., Raguideau, S., James, R., Phillippy, A.M., Chikhi, R., Quince, C., 2023. Efficient High-Quality Metagenome Assembly from Long Accurate Reads using Minimizer-space de Bruijn Graphs (preprint). *Bioinformatics*. <https://doi.org/10.1101/2023.07.07.548136>
- Bethune, K., Mariac, C., Couderc, M., Scarcelli, N., Santoni, S., Ardisson, M., Martin, J.-F., Montúfar, R., Klein, V., Sabot, F., Vigouroux, Y., Couvreur, T.L.P., 2019. Long-

- fragment targeted capture for long-read sequencing of plastomes. *Appl. Plant Sci.* 7, e1243. <https://doi.org/10.1002/aps3.1243>
- Beutin, L., Bode, L., Ozel, M., Stephan, R., 1990. Enterohemolysin production is associated with a temperate bacteriophage in *Escherichia coli* serogroup O26 strains. *J. Bacteriol.* 172, 6469–6475. <https://doi.org/10.1128/jb.172.11.6469-6475.1990>
- Bhunja, A.K., 2018. *Escherichia coli*, in: Bhunja, A.K. (Ed.), *Foodborne Microbial Pathogens: Mechanisms and Pathogenesis*. Springer New York, New York, NY, pp. 249–269. [https://doi.org/10.1007/978-1-4939-7349-1\\_14](https://doi.org/10.1007/978-1-4939-7349-1_14)
- Bielaszewska, M., Mellmann, A., Zhang, W., Köck, R., Fruth, A., Bauwens, A., Peters, G., Karch, H., 2011. Characterisation of the *Escherichia coli* strain associated with an outbreak of haemolytic uraemic syndrome in Germany, 2011: a microbiological study. *Lancet Infect. Dis.* 11, 671–676. [https://doi.org/10.1016/S1473-3099\(11\)70165-7](https://doi.org/10.1016/S1473-3099(11)70165-7)
- Bielaszewska, M., Stoewe, F., Fruth, A., Zhang, W., Prager, R., Brockmeyer, J., Mellmann, A., Karch, H., Friedrich, A.W., 2009. Shiga toxin, cytolethal distending toxin, and hemolysin repertoires in clinical *Escherichia coli* O91 isolates. *J. Clin. Microbiol.* 47, 2061–2066. <https://doi.org/10.1128/JCM.00201-09>
- Blum, G., Ott, M., Lischewski, A., Ritter, A., Imrich, H., Tschäpe, H., Hacker, J., 1994. Excision of large DNA regions termed pathogenicity islands from tRNA-specific loci in the chromosome of an *Escherichia coli* wild-type pathogen. *Infect. Immun.* 62, 606–614. <https://doi.org/10.1128/iai.62.2.606-614.1994>
- Bouhaddioui, B., Ben Aissa, R., Boudabous, A., 1998. [Characterization of *Escherichia coli* strains isolated from man and seafood]. *Bull. Soc. Pathol. Exot.* 1990 91, 283–286.
- Bouso, J.M., Planet, P.J., 2019. Complete nontuberculous mycobacteria whole genomes using an optimized DNA extraction protocol for long-read sequencing. *BMC Genomics* 20, 793. <https://doi.org/10.1186/s12864-019-6134-y>
- Brabban, A.D., Hite, E., Callaway, T.R., 2005. Evolution of foodborne pathogens via temperate bacteriophage-mediated gene transfer. *Foodborne Pathog. Dis.* 2, 287–303. <https://doi.org/10.1089/fpd.2005.2.287>
- Braz, V.S., Melchior, K., Moreira, C.G., 2020. *Escherichia coli* as a Multifaceted Pathogenic and Versatile Bacterium. *Front. Cell. Infect. Microbiol.* 10, 548492. <https://doi.org/10.3389/fcimb.2020.548492>
- Bruyand, M., Mariani-Kurkdjian, P., Le Hello, S., King, L.-A., Van Cauteren, D., Lefevre, S., Gouali, M., Jourdan-da Silva, N., Mailles, A., Donguy, M.-P., Loukiadis, E., Sergentet-Thevenot, D., Loirat, C., Bonacorsi, S., Weill, F.-X., De Valk, H., Réseau français hospitalier de surveillance du SHU pédiatrique, 2019. Paediatric haemolytic uraemic syndrome related to Shiga toxin-producing *Escherichia coli*, an overview of 10 years of surveillance in France, 2007 to 2016. *Euro Surveill. Bull. Eur. Sur Mal. Transm. Eur. Commun. Dis. Bull.* 24, 1800068. <https://doi.org/10.2807/1560-7917.ES.2019.24.8.1800068>
- Bugarel, M., Beutin, L., Fach, P., 2010a. Low-density macroarray targeting non-locus of enterocyte effacement effectors (nle genes) and major virulence factors of Shiga toxin-producing *Escherichia coli* (STEC): a new approach for molecular risk assessment of STEC isolates. *Appl. Environ. Microbiol.* 76, 203–211. <https://doi.org/10.1128/AEM.01921-09>
- Bugarel, M., Beutin, L., Martin, A., Gill, A., Fach, P., 2010b. Micro-array for the identification of Shiga toxin-producing *Escherichia coli* (STEC) seropathotypes associated with Hemorrhagic Colitis and Hemolytic Uremic Syndrome in humans. *Int. J. Food Microbiol.* 142, 318–329. <https://doi.org/10.1016/j.ijfoodmicro.2010.07.010>

- Burgaya, J., Marin, J., Royer, G., Condamine, B., Gachet, B., Clermont, O., Jaureguy, F., Burdet, C., Lefort, A., de Lastours, V., Denamur, E., Galardini, M., Blanquart, F., Colibafı/Septicoli & Coliville groups, 2023. The bacterial genetic determinants of *Escherichia coli* capacity to cause bloodstream infections in humans. *PLoS Genet.* 19, e1010842. <https://doi.org/10.1371/journal.pgen.1010842>
- Burrus, V., Waldor, M.K., 2004. Shaping bacterial genomes with integrative and conjugative elements. *Res. Microbiol.* 155, 376–386. <https://doi.org/10.1016/j.resmic.2004.01.012>
- Butler, T., 2012. Haemolytic uraemic syndrome during shigellosis. *Trans. R. Soc. Trop. Med. Hyg.* 106, 395–399. <https://doi.org/10.1016/j.trstmh.2012.04.001>
- Buytaers, F.E., Saltykova, A., Denayer, S., Verhaegen, B., Vanneste, K., Roosens, N.H.C., Piérard, D., Marchal, K., De Keersmaecker, S.C.J., 2021. Towards Real-Time and Affordable Strain-Level Metagenomics-Based Foodborne Outbreak Investigations Using Oxford Nanopore Sequencing Technologies. *Front. Microbiol.* 12, 738284. <https://doi.org/10.3389/fmicb.2021.738284>
- Buytaers, F.E., Saltykova, A., Denayer, S., Verhaegen, B., Vanneste, K., Roosens, N.H.C., Piérard, D., Marchal, K., De Keersmaecker, S.C.J., 2020. A Practical Method to Implement Strain-Level Metagenomics-Based Foodborne Outbreak Investigation and Source Tracking in Routine. *Microorganisms* 8, 1191. <https://doi.org/10.3390/microorganisms8081191>
- Caprioli, A., Morabito, S., Brugère, H., Oswald, E., 2005. Enterohaemorrhagic *Escherichia coli*: emerging issues on virulence and modes of transmission. *Vet. Res.* 36, 289–311. <https://doi.org/10.1051/vetres:2005002>
- Cherukuri, Y., Janga, S.C., 2016. Benchmarking of *de novo* assembly algorithms for Nanopore data reveals optimal performance of OLC approaches. *BMC Genomics* 17 Suppl 7, 507. <https://doi.org/10.1186/s12864-016-2895-8>
- Ciufo, S., Kannan, S., Sharma, S., Badretdin, A., Clark, K., Turner, S., Brover, S., Schoch, C.L., Kimchi, A., DiCuccio, M., 2018. Using average nucleotide identity to improve taxonomic assignments in prokaryotic genomes at the NCBI. *Int. J. Syst. Evol. Microbiol.* 68, 2386–2392. <https://doi.org/10.1099/ijsem.0.002809>
- Cointe, A., Birgy, A., Bridier-Nahmias, A., Mariani-Kurkdjian, P., Walewski, V., Lévy, C., Cohen, R., Fach, P., Delannoy, S., Bidet, P., Bonacorsi, S., 2020. *Escherichia coli* O80 hybrid pathotype strains producing Shiga toxin and ESBL: molecular characterization and potential therapeutic options. *J. Antimicrob. Chemother.* 75, 537–542. <https://doi.org/10.1093/jac/dkz484>
- Cointe, A., Birgy, A., Mariani-Kurkdjian, P., Liguori, S., Courroux, C., Blanco, J., Delannoy, S., Fach, P., Loukiadis, E., Bidet, P., Bonacorsi, S., 2018. Emerging Multidrug-Resistant Hybrid Pathotype Shiga Toxin-Producing *Escherichia coli* O80 and Related Strains of Clonal Complex 165, Europe. *Emerg. Infect. Dis.* 24, 2262–2269. <https://doi.org/10.3201/eid2412.180272>
- Colello, R., Vélez, M.V., González, J., Montero, D.A., Bustamante, A.V., Del Canto, F., Etcheverría, A.I., Vidal, R., Padola, N.L., 2018. First report of the distribution of Locus of Adhesion and Autoaggregation (LAA) pathogenicity island in LEE-negative Shiga toxin-producing *Escherichia coli* isolates from Argentina. *Microb. Pathog.* 123, 259–263. <https://doi.org/10.1016/j.micpath.2018.07.011>
- Cookson, A.L., Biggs, P.J., Marshall, J.C., Reynolds, A., Collis, R.M., French, N.P., Brightwell, G., 2017. Culture independent analysis using *gnd* as a target gene to assess *Escherichia coli* diversity and community structure. *Sci. Rep.* 7, 841. <https://doi.org/10.1038/s41598-017-00890-6>

- Coombes, B.K., Wickham, M.E., Mascarenhas, M., Gruenheid, S., Finlay, B.B., Karmali, M.A., 2008. Molecular analysis as an aid to assess the public health risk of non-O157 Shiga toxin-producing *Escherichia coli* strains. *Appl. Environ. Microbiol.* 74, 2153–2160. <https://doi.org/10.1128/AEM.02566-07>
- Cremonesi, P., Severgnini, M., Romanò, A., Sala, L., Luini, M., Castiglioni, B., 2021. Bovine Milk Microbiota: Comparison among Three Different DNA Extraction Protocols To Identify a Better Approach for Bacterial Analysis. *Microbiol. Spectr.* 9, e0037421. <https://doi.org/10.1128/Spectrum.00374-21>
- Croxen, M.A., Law, R.J., Scholz, R., Keeney, K.M., Wlodarska, M., Finlay, B.B., 2013. Recent advances in understanding enteric pathogenic *Escherichia coli*. *Clin. Microbiol. Rev.* 26, 822–880. <https://doi.org/10.1128/CMR.00022-13>
- Cummins, E.A., Hall, R.J., Connor, C., McInerney, J.O., McNally, A., 2022. Distinct evolutionary trajectories in the *Escherichia coli* pangenome occur within sequence types. *Microb. Genomics* 8, mgen000903. <https://doi.org/10.1099/mgen.0.000903>
- Dahn, H.A., Mountcastle, J., Balacco, J., Winkler, S., Bista, I., Schmitt, A.D., Pettersson, O.V., Formenti, G., Oliver, K., Smith, M., Tan, W., Kraus, A., Mac, S., Komoroske, L.M., Lama, T., Crawford, A.J., Murphy, R.W., Brown, S., Scott, A.F., Morin, P.A., Jarvis, E.D., Fedrigo, O., 2022. Benchmarking ultra-high molecular weight DNA preservation methods for long-read and long-range sequencing. *GigaScience* 11, giac068. <https://doi.org/10.1093/gigascience/giac068>
- Dallman, T., Smith, G.P., O'Brien, B., Chattaway, M.A., Finlay, D., Grant, K.A., Jenkins, C., 2012. Characterization of a verocytotoxin-producing Enteroaggregative *Escherichia coli* serogroup O111:H21 strain associated with a household outbreak in Northern Ireland. *J. Clin. Microbiol.* 50, 4116–4119. <https://doi.org/10.1128/JCM.02047-12>
- Darfeuille-Michaud, A., Boudeau, J., Bulois, P., Neut, C., Glasser, A.-L., Barnich, N., Bringer, M.-A., Swidsinski, A., Beaugerie, L., Colombel, J.-F., 2004. High prevalence of adherent-invasive *Escherichia coli* associated with ileal mucosa in Crohn's disease. *Gastroenterology* 127, 412–421. <https://doi.org/10.1053/j.gastro.2004.04.061>
- Darfeuille-Michaud, A., Neut, C., Barnich, N., Lederman, E., Di Martino, P., Desreumaux, P., Gambiez, L., Joly, B., Cortot, A., Colombel, J.F., 1998. Presence of adherent *Escherichia coli* strains in ileal mucosa of patients with Crohn's disease. *Gastroenterology* 115, 1405–1413. [https://doi.org/10.1016/s0016-5085\(98\)70019-8](https://doi.org/10.1016/s0016-5085(98)70019-8)
- De Paepe, M., Leclerc, M., Tinsley, C.R., Petit, M.-A., 2014. Bacteriophages: an underestimated role in human and animal health? *Front. Cell. Infect. Microbiol.* 4, 39. <https://doi.org/10.3389/fcimb.2014.00039>
- Deamer, D., Akeson, M., Branton, D., 2016. Three decades of nanopore sequencing. *Nat. Biotechnol.* 34, 518–524. <https://doi.org/10.1038/nbt.3423>
- Dean-Nystrom, E.A., Bosworth, B.T., Moon, H.W., O'Brien, A.D., 1998. *Escherichia coli* O157:H7 requires intimin for enteropathogenicity in calves. *Infect. Immun.* 66, 4560–4563. <https://doi.org/10.1128/IAI.66.9.4560-4563.1998>
- Delannoy, S., Beutin, L., Fach, P., 2013a. Towards a molecular definition of Enterohemorrhagic *Escherichia coli* (EHEC): detection of genes located on O island 57 as markers to distinguish EHEC from closely related Enteropathogenic *E. coli* strains. *J. Clin. Microbiol.* 51, 1083–1088. <https://doi.org/10.1128/JCM.02864-12>
- Delannoy, S., Beutin, L., Fach, P., 2013b. Discrimination of enterohemorrhagic *Escherichia coli* (EHEC) from non-EHEC strains based on detection of various combinations of type III effector genes. *J. Clin. Microbiol.* 51, 3257–3262. <https://doi.org/10.1128/JCM.01471-13>

- Delannoy, S., Chaves, B.D., Ison, S.A., Webb, H.E., Beutin, L., Delaval, J., Billet, I., Fach, P., 2016. Revisiting the STEC Testing Approach: Using *espK* and *espV* to Make Enterohemorrhagic *Escherichia coli* (EHEC) Detection More Reliable in Beef. *Front. Microbiol.* 7, 1. <https://doi.org/10.3389/fmicb.2016.00001>
- Delannoy, S., Tran, M.-L., Fach, P., 2022. Insights into the assessment of highly pathogenic Shiga toxin-producing *Escherichia coli* in raw milk and raw milk cheeses by High Throughput Real-time PCR. *Int. J. Food Microbiol.* 366, 109564. <https://doi.org/10.1016/j.ijfoodmicro.2022.109564>
- DeMaere, M.Z., Darling, A.E., 2019. bin3C: exploiting Hi-C sequencing data to accurately resolve metagenome-assembled genomes. *Genome Biol.* 20, 46. <https://doi.org/10.1186/s13059-019-1643-1>
- Deng, W., Li, Y., Vallance, B.A., Finlay, B.B., 2001. Locus of enterocyte effacement from *Citrobacter rodentium*: sequence analysis and evidence for horizontal transfer among attaching and effacing pathogens. *Infect. Immun.* 69, 6323–6335. <https://doi.org/10.1128/IAI.69.10.6323-6335.2001>
- Deng, W., Puente, J.L., Gruenheid, S., Li, Y., Vallance, B.A., Vázquez, A., Barba, J., Ibarra, J.A., O'Donnell, P., Metalnikov, P., Ashman, K., Lee, S., Goode, D., Pawson, T., Finlay, B.B., 2004. Dissecting virulence: systematic and functional analyses of a pathogenicity island. *Proc. Natl. Acad. Sci. U. S. A.* 101, 3597–3602. <https://doi.org/10.1073/pnas.0400326101>
- Eagle, S.H.C., Robertson, J., Bastedo, D.P., Liu, K., Nash, J.H.E., 2023. Evaluation of five commercial DNA extraction kits using *Salmonella* as a model for implementation of rapid Nanopore sequencing in routine diagnostic laboratories. *Access Microbiol.* 5, 000468.v3. <https://doi.org/10.1099/acmi.0.000468.v3>
- EFSA, 2020. EFSA BIOHAZ Panel (EFSA Panel on Biological Hazards), Koutsoumanis K, Allende A, Alvarez-Ordóñez A, Bover-Cid S, Chemaly M, Davies R, De Cesare A, Herman L, Hilbert F, Lindqvist R, Nauta M, Peixe L, Ru G, Simmons M, Skandamis P, Suffredini E, Jenkins C, Monteiro Pires S, Morabito S, Niskanen T, Scheutz F, da Silva Felício MT, Messens W and Bolton D, 2020b. Pathogenicity assessment of Shiga toxin-producing *Escherichia coli* (STEC) and the public health risk posed by contamination of food with STEC. *EFSA Journal* 2020;18(1):5967, 105 pp. <https://doi.org/10.2903/j.efsa.2020.5967>.
- EFSA, 2019. EFSA BIOHAZ Panel (EFSA Panel on Biological Hazards), Koutsoumanis K, Allende A, Alvarez-Ordóñez A, Bolton D, Bover-Cid S, Chemaly M, Davies R, De Cesare A, Hilbert F, Lindqvist R, Nauta M, Peixe L, Ru G, Simmons M, Skandamis P, Suffredini E, Jenkins C, Malorny B, Ribeiro Duarte AS, Torpdahl M, da Silva Felício MT, Guerra B, Rossi M and Herman L, 2019. Whole genome sequencing and metagenomics for outbreak investigation, source attribution and risk assessment of foodborne microorganisms. *EFSA Journal* 2019;17(12):5898, 78 pp. <https://doi.org/10.2903/j.efsa.2019.5898>.
- EFSA, 2015. Public health risks associated with Enterohemorrhagic *Escherichia coli* (EHEC) as a food-borne pathogen. *EFSA J.* 13, 4330. <https://doi.org/10.2903/j.efsa.2015.4330>
- EFSA, 2013. BIOHAZ Panel (EFSA Panel on Biological Hazards), 2013c. Scientific Opinion on VTEC-seropathotype and scientific criteria regarding pathogenicity assessment. *EFSA Journal* 2013;11(4):3138, 106 pp.
- EFSA and ECDC, 2022. The European Union One Health 2021 Zoonoses Report. *EFSA J.* 20, e07666. <https://doi.org/10.2903/j.efsa.2022.7666>



- EFSA and ECDC, 2017. The European Union summary report on trends and sources of zoonoses, zoonotic agents and food-borne outbreaks in 2016. *EFSA J. Eur. Food Saf. Auth.* 15, e05077. <https://doi.org/10.2903/j.efsa.2017.5077>
- Ekaterina Kazantseva, Ataberk Donmez, Mihai Pop, Mikhail Kolmogorov, 2023. stRainy: assembly-based metagenomic strain phasing using long reads. *bioRxiv* 2023.01.31.526521. <https://doi.org/10.1101/2023.01.31.526521>
- European Parliament and Council of the European Union, 2002. Règlement (CE) n° 178/2002 du Parlement européen et du Conseil du 28 janvier 2002 établissant les principes généraux et les prescriptions générales de la législation alimentaire, instituant l'Autorité européenne de sécurité des aliments et fixant des procédures relatives à la sécurité des denrées alimentaires, *Journal officiel* n° L 031 du 01/02/2002 p. 0001 - 0024, available at: <https://eur-lex.europa.eu/legal-content/FR/TXT/?uri=celex%3A32002R0178>.
- Feng, P.C.H., Delannoy, S., Lacher, D.W., Dos Santos, L.F., Beutin, L., Fach, P., Rivas, M., Hartland, E.L., Paton, A.W., Guth, B.E.C., 2014. Genetic diversity and virulence potential of Shiga toxin-producing *Escherichia coli* O113:H21 strains isolated from clinical, environmental, and food sources. *Appl. Environ. Microbiol.* 80, 4757–4763. <https://doi.org/10.1128/AEM.01182-14>
- Ferens, W.A., Hovde, C.J., 2011. *Escherichia coli* O157:H7: animal reservoir and sources of human infection. *Foodborne Pathog. Dis.* 8, 465–487. <https://doi.org/10.1089/fpd.2010.0673>
- Fitzgerald, S.F., Lupolova, N., Shaaban, S., Dallman, T.J., Greig, D., Allison, L., Tongue, S.C., Evans, J., Henry, M.K., McNeilly, T.N., Bono, J.L., Gally, D.L., 2021. Genome structural variation in *Escherichia coli* O157:H7. *Microb. Genomics* 7, 000682. <https://doi.org/10.1099/mgen.0.000682>
- Foxman, B., Zhang, L., Koopman, J.S., Manning, S.D., Marrs, C.F., 2005. Choosing an appropriate bacterial typing technique for epidemiologic studies. *Epidemiol. Perspect. Innov. EPI* 2, 10. <https://doi.org/10.1186/1742-5573-2-10>
- Franz, E., Delaquis, P., Morabito, S., Beutin, L., Gobius, K., Rasko, D.A., Bono, J., French, N., Osek, J., Lindstedt, B.-A., Muniesa, M., Manning, S., LeJeune, J., Callaway, T., Beatson, S., Eppinger, M., Dallman, T., Forbes, K.J., Aarts, H., Pearl, D.L., Gannon, V.P.J., Laing, C.R., Strachan, N.J.C., 2014. Exploiting the explosion of information associated with whole genome sequencing to tackle Shiga toxin-producing *Escherichia coli* (STEC) in global food production systems. *Int. J. Food Microbiol.* 187, 57–72. <https://doi.org/10.1016/j.ijfoodmicro.2014.07.002>
- Fratamico, P.M., DebRoy, C., Liu, Y., Needleman, D.S., Baranzoni, G.M., Feng, P., 2016a. Advances in Molecular Serotyping and Subtyping of *Escherichia coli*. *Front. Microbiol.* 7, 644. <https://doi.org/10.3389/fmicb.2016.00644>
- Fratamico, P.M., DebRoy, C., Needleman, D.S., 2016b. Editorial: Emerging Approaches for Typing, Detection, Characterization, and Traceback of *Escherichia coli*. *Front. Microbiol.* 7, 2089. <https://doi.org/10.3389/fmicb.2016.02089>
- Frost, L.S., Leplae, R., Summers, A.O., Toussaint, A., 2005. Mobile genetic elements: the agents of open source evolution. *Nat. Rev. Microbiol.* 3, 722–732. <https://doi.org/10.1038/nrmicro1235>
- FSIS, 1994. (Food Safety and Inspection Service) Directive 10,010.1, Verification Activities for *Escherichia coli* O157:H7 in Raw Beef Products.

- FSIS and USDA, 2012. (Food Safety and Inspection Service and The United States Department of agriculture), Shiga Toxin-Producing *Escherichia coli* in Certain Raw Beef Products. Fed. Regist.
- Fukushima, H., Hashizume, T., Morita, Y., Tanaka, J., Azuma, K., Mizumoto, Y., Kaneno, M., Matsuura, M., Konma, K., Kitani, T., 1999. Clinical experiences in Sakai City Hospital during the massive outbreak of enterohemorrhagic *Escherichia coli* O157 infections in Sakai City, 1996. *Pediatr. Int. Off. J. Jpn. Pediatr. Soc.* 41, 213–217. <https://doi.org/10.1046/j.1442-200x.1999.4121041.x>
- Gand, M., Bloemen, B., Vanneste, K., Roosens, N.H.C., De Keersmaecker, S.C.J., 2023. Comparison of 6 DNA extraction methods for isolation of high yield of high molecular weight DNA suitable for shotgun metagenomics Nanopore sequencing to detect bacteria. *BMC Genomics* 24, 438. <https://doi.org/10.1186/s12864-023-09537-5>
- Garmendia, J., Frankel, G., Crepin, V.F., 2005. Enteropathogenic and enterohemorrhagic *Escherichia coli* infections: translocation, translocation, translocation. *Infect. Immun.* 73, 2573–2585. <https://doi.org/10.1128/IAI.73.5.2573-2585.2005>
- Gaytán, M.O., Martínez-Santos, V.I., Soto, E., González-Pedrajo, B., 2016. Type Three Secretion System in Attaching and Effacing Pathogens. *Front. Cell. Infect. Microbiol.* 6, 129. <https://doi.org/10.3389/fcimb.2016.00129>
- Ghaheri, M., Kahrizi, D., Yari, K., Babaie, A., Suthar, R.S., Kazemi, E., 2016. A comparative evaluation of four DNA extraction protocols from whole blood sample. *Cell. Mol. Biol. Noisy--Gd. Fr.* 62, 120–124.
- Gill, A., Dussault, F., McMahon, T., Petronella, N., Wang, X., Cebelinski, E., Scheutz, F., Weedmark, K., Blais, B., Carrillo, C., 2022. Characterization of Atypical Shiga Toxin Gene Sequences and Description of Stx2j, a New Subtype. *J. Clin. Microbiol.* 60, e0222921. <https://doi.org/10.1128/jcm.02229-21>
- Goering, R.V., 2010. Pulsed field gel electrophoresis: a review of application and interpretation in the molecular epidemiology of infectious disease. *Infect. Genet. Evol. J. Mol. Epidemiol. Evol. Genet. Infect. Dis.* 10, 866–875. <https://doi.org/10.1016/j.meegid.2010.07.023>
- Goussarov, G., Mysara, M., Vandamme, P., Van Houdt, R., 2022. Introduction to the principles and methods underlying the recovery of metagenome-assembled genomes from metagenomic data. *MicrobiologyOpen* 11, e1298. <https://doi.org/10.1002/mbo3.1298>
- Grützke, J., Gwida, M., Deneke, C., Brendebach, H., Projahn, M., Schattschneider, A., Hofreuter, D., El-Ashker, M., Malorny, B., Al Dahouk, S., 2021. Direct identification and molecular characterization of zoonotic hazards in raw milk by metagenomics using *Brucella* as a model pathogen. *Microb. Genomics* 7, 000552. <https://doi.org/10.1099/mgen.0.000552>
- Gyles, C.L., 2007. Shiga toxin-producing *Escherichia coli*: an overview. *J. Anim. Sci.* 85, E45-62. <https://doi.org/10.2527/jas.2006-508>
- Hacker, J., Carniel, E., 2001. Ecological fitness, genomic islands and bacterial pathogenicity. A Darwinian view of the evolution of microbes. *EMBO Rep.* 2, 376–381. <https://doi.org/10.1093/embo-reports/kve097>
- Hacker, J., Kaper, J.B., 2000. Pathogenicity islands and the evolution of microbes. *Annu. Rev. Microbiol.* 54, 641–679. <https://doi.org/10.1146/annurev.micro.54.1.641>
- Hayashi, T., Makino, K., Ohnishi, M., Kurokawa, K., Ishii, K., Yokoyama, K., Han, C.G., Ohtsubo, E., Nakayama, K., Murata, T., Tanaka, M., Tobe, T., Iida, T., Takami, H., Honda, T., Sasakawa, C., Ogasawara, N., Yasunaga, T., Kuhara, S., Shiba, T., Hattori,

- M., Shinagawa, H., 2001. Complete genome sequence of Enterohemorrhagic *Escherichia coli* O157:H7 and genomic comparison with a laboratory strain K-12. *DNA Res. Int. J. Rapid Publ. Rep. Genes Genomes* 8, 11–22. <https://doi.org/10.1093/dnares/8.1.11>
- Herold, S., Paton, J.C., Paton, A.W., 2009. Sab, a novel autotransporter of locus of enterocyte effacement-negative Shiga-toxigenic *Escherichia coli* O113:H21, contributes to adherence and biofilm formation. *Infect. Immun.* 77, 3234–3243. <https://doi.org/10.1128/IAI.00031-09>
- Hill, B.M., Bisht, K., Atkins, G.R., Gomez, A.A., Rumbaugh, K.P., Wakeman, C.A., Brown, A.M.V., 2022. Lysis-Hi-C as a method to study polymicrobial communities and eDNA. *Mol. Ecol. Resour.* 22, 1029–1042. <https://doi.org/10.1111/1755-0998.13535>
- Hochhauser, D., Millman, A., Sorek, R., 2023. The defense island repertoire of the *Escherichia coli* pan-genome. *PLoS Genet.* 19, e1010694. <https://doi.org/10.1371/journal.pgen.1010694>
- Ide, T., Laarmann, S., Greune, L., Schillers, H., Oberleithner, H., Schmidt, M.A., 2001. Characterization of translocation pores inserted into plasma membranes by type III-secreted Esp proteins of Enteropathogenic *Escherichia coli*. *Cell. Microbiol.* 3, 669–679. <https://doi.org/10.1046/j.1462-5822.2001.00146.x>
- Iguchi, A., Iyoda, S., Seto, K., Morita-Ishihara, T., Scheutz, F., Ohnishi, M., Pathogenic *E. coli* Working Group in Japan, 2015. *Escherichia coli* O-Genotyping PCR: a Comprehensive and Practical Platform for Molecular O Serogrouping. *J. Clin. Microbiol.* 53, 2427–2432. <https://doi.org/10.1128/JCM.00321-15>
- Im, H., Hwang, S.-H., Kim, B.S., Choi, S.H., 2021. Pathogenic potential assessment of the Shiga toxin-producing *Escherichia coli* by a source attribution-considered machine learning model. *Proc. Natl. Acad. Sci. U. S. A.* 118, e2018877118. <https://doi.org/10.1073/pnas.2018877118>
- Imamovic, L., Tozzoli, R., Michelacci, V., Minelli, F., Marziano, M.L., Caprioli, A., Morabito, S., 2010. OI-57, a genomic island of *Escherichia coli* O157, is present in other seropathotypes of Shiga toxin-producing *E. coli* associated with severe human disease. *Infect. Immun.* 78, 4697–4704. <https://doi.org/10.1128/IAI.00512-10>
- ISO, 2012. ISO (International Organization for Standardization) (2012). ISO/TS 13136:2012, Microbiology of Food and Animal Feed—Real-Time Polymerase Chain Reaction (PCR)–Based Method for the Detection of Food-Borne Pathogens—Horizontal Method for the Detection of Shiga Toxin-Producing *Escherichia coli* (STEC) and the Determination of O157, O111, O26, O103, and O145 Serogroups.
- ISO, 2001. ISO (International Organization for Standardization) (2001). ISO-16654, 2001, ISO16654 Microbiology of food and animal feeding stuffs — Horizontal method for the detection of *Escherichia coli* O157.
- Jain, C., Rodriguez-R, L.M., Phillippy, A.M., Konstantinidis, K.T., Aluru, S., 2018. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat. Commun.* 9, 5114. <https://doi.org/10.1038/s41467-018-07641-9>
- Jaureguy, F., Landraud, L., Passet, V., Diancourt, L., Frapy, E., Guigon, G., Carbonnelle, E., Lortholary, O., Clermont, O., Denamur, E., Picard, B., Nassif, X., Brisse, S., 2008. Phylogenetic and genomic diversity of human bacteremic *Escherichia coli* strains. *BMC Genomics* 9, 560. <https://doi.org/10.1186/1471-2164-9-560>
- Jiang, L., Yang, W., Jiang, X., Yao, T., Wang, L., Yang, B., 2021. Virulence-related O islands in Enterohemorrhagic *Escherichia coli* O157:H7. *Gut Microbes* 13, 1992237. <https://doi.org/10.1080/19490976.2021.1992237>

- Joensen, K.G., Tetzschner, A.M.M., Iguchi, A., Aarestrup, F.M., Scheutz, F., 2015. Rapid and Easy *In Silico* Serotyping of *Escherichia coli* Isolates by Use of Whole-Genome Sequencing Data. *J. Clin. Microbiol.* 53, 2410–2426. <https://doi.org/10.1128/JCM.00008-15>
- Johnson, T.J., Nolan, L.K., 2009. Pathogenomics of the virulence plasmids of *Escherichia coli*. *Microbiol. Mol. Biol. Rev.* MMBR 73, 750–774. <https://doi.org/10.1128/MMBR.00015-09>
- Jones, A., Torkel, C., Stanley, D., Nasim, J., Borevitz, J., Schwessinger, B., 2021. High-molecular weight DNA extraction, clean-up and size selection for long-read sequencing. *PloS One* 16, e0253830. <https://doi.org/10.1371/journal.pone.0253830>
- Joseph, A., Cointe, A., Mariani Kurkdjian, P., Rafat, C., Hertig, A., 2020. Shiga Toxin-Associated Hemolytic Uremic Syndrome: A Narrative Review. *Toxins* 12, 67. <https://doi.org/10.3390/toxins12020067>
- Kaper, J.B., Nataro, J.P., Mobley, H.L., 2004. Pathogenic *Escherichia coli*. *Nat. Rev. Microbiol.* 2, 123–140. <https://doi.org/10.1038/nrmicro818>
- Karmali, M.A., 2018. Factors in the emergence of serious human infections associated with highly pathogenic strains of Shiga toxin-producing *Escherichia coli*. *Int. J. Med. Microbiol. IJMM* 308, 1067–1072. <https://doi.org/10.1016/j.ijmm.2018.08.005>
- Karmali, M.A., 2017. Emerging Public Health Challenges of Shiga Toxin-Producing *Escherichia coli* Related to Changes in the Pathogen, the Population, and the Environment. *Clin. Infect. Dis. Off. Publ. Infect. Dis. Soc. Am.* 64, 371–376. <https://doi.org/10.1093/cid/ciw708>
- Karmali, M.A., 1989. Infection by verocytotoxin-producing *Escherichia coli*. *Clin. Microbiol. Rev.* 2, 15–38. <https://doi.org/10.1128/CMR.2.1.15>
- Karmali, M.A., Mascarenhas, M., Shen, S., Ziebell, K., Johnson, S., Reid-Smith, R., Isaac-Renton, J., Clark, C., Rahn, K., Kaper, J.B., 2003. Association of genomic O island 122 of *Escherichia coli* EDL 933 with verocytotoxin-producing *Escherichia coli* seropathotypes that are linked to epidemic and/or serious disease. *J. Clin. Microbiol.* 41, 4930–4940. <https://doi.org/10.1128/JCM.41.11.4930-4940.2003>
- Kavaliauskiene, S., Dyve Lingelem, A.B., Skotland, T., Sandvig, K., 2017. Protection against Shiga Toxins. *Toxins* 9, 44. <https://doi.org/10.3390/toxins9020044>
- Kelly, B.G., Vespermann, A., Bolton, D.J., 2009. The role of horizontal gene transfer in the evolution of selected foodborne bacterial pathogens. *Food Chem. Toxicol. Int. J. Publ. Br. Ind. Biol. Res. Assoc.* 47, 951–968. <https://doi.org/10.1016/j.fct.2008.02.006>
- Khalil, R.K.S., Skinner, C., Patfield, S., He, X., 2016. Phage-mediated Shiga toxin (Stx) horizontal gene transfer and expression in non-Shiga toxigenic *Enterobacter* and *Escherichia coli* strains. *Pathog. Dis.* 74, ftw037. <https://doi.org/10.1093/femspd/ftw037>
- Kim, J.-S., Lee, M.-S., Kim, J.H., 2020. Recent Updates on Outbreaks of Shiga Toxin-Producing *Escherichia coli* and Its Potential Reservoirs. *Front. Cell. Infect. Microbiol.* 10, 273. <https://doi.org/10.3389/fcimb.2020.00273>
- Klapproth, J.M., Scaletsky, I.C., McNamara, B.P., Lai, L.C., Malstrom, C., James, S.P., Sonnenberg, M.S., 2000. A large toxin from pathogenic *Escherichia coli* strains that inhibits lymphocyte activation. *Infect. Immun.* 68, 2148–2155. <https://doi.org/10.1128/IAI.68.4.2148-2155.2000>
- Kocurek, B., Ramachandran, P., Grim, C.J., Morin, P., Howard, L., Ottesen, A., Timme, R., Leonard, S.R., Rand, H., Strain, E., Tadesse, D., Pettengill, J.B., Lacher, D.W., Mammel, M., Jarvis, K.G., 2023. Application of quasimetagenomics methods to

- define microbial diversity and subtype *Listeria monocytogenes* in dairy and seafood production facilities. *Microbiol. Spectr.* e0148223.  
<https://doi.org/10.1128/spectrum.01482-23>
- Kolmogorov, M., Bickhart, D.M., Behsaz, B., Gurevich, A., Rayko, M., Shin, S.B., Kuhn, K., Yuan, J., Polevikov, E., Smith, T.P.L., Pevzner, P.A., 2020. metaFlye: scalable long-read metagenome assembly using repeat graphs. *Nat. Methods* 17, 1103–1110.  
<https://doi.org/10.1038/s41592-020-00971-x>
- Konczy, P., Ziebell, K., Mascarenhas, M., Choi, A., Michaud, C., Kropinski, A.M., Whittam, T.S., Wickham, M., Finlay, B., Karmali, M.A., 2008. Genomic O island 122, locus for enterocyte effacement, and the evolution of virulent verocytotoxin-producing *Escherichia coli*. *J. Bacteriol.* 190, 5832–5840. <https://doi.org/10.1128/JB.00480-08>
- Konowalchuk, J., Speirs, J.I., Stavric, S., 1977. Vero response to a cytotoxin of *Escherichia coli*. *Infect. Immun.* 18, 775–779. <https://doi.org/10.1128/iai.18.3.775-779.1977>
- Koren, S., Harhay, G.P., Smith, T.P.L., Bono, J.L., Harhay, D.M., Mcvey, S.D., Radune, D., Bergman, N.H., Phillippy, A.M., 2013. Reducing assembly complexity of microbial genomes with single-molecule sequencing. *Genome Biol.* 14, R101.  
<https://doi.org/10.1186/gb-2013-14-9-r101>
- Koren, S., Walenz, B.P., Berlin, K., Miller, J.R., Bergman, N.H., Phillippy, A.M., 2017. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res.* 27, 722–736. <https://doi.org/10.1101/gr.215087.116>
- Krause, M., Barth, H., Schmidt, H., 2018. Toxins of Locus of Enterocyte Effacement-Negative Shiga Toxin-Producing *Escherichia coli*. *Toxins* 10, 241.  
<https://doi.org/10.3390/toxins10060241>
- Krüger, A., Lucchesi, P.M.A., 2015. Shiga toxins and *stx* phages: highly diverse entities. *Microbiology.* <https://doi.org/10.1099/mic.0.000003>
- Lacher, D.W., Gangiredla, J., Jackson, S.A., Elkins, C.A., Feng, P.C.H., 2014. Novel microarray design for molecular serotyping of Shiga toxin-producing *Escherichia coli* strains isolated from fresh produce. *Appl. Environ. Microbiol.* 80, 4677–4682.  
<https://doi.org/10.1128/AEM.01049-14>
- Lacher, D.W., Gangiredla, J., Patel, I., Elkins, C.A., Feng, P.C.H., 2016. Use of the *Escherichia coli* Identification Microarray for Characterizing the Health Risks of Shiga Toxin-Producing *Escherichia coli* Isolated from Foods. *J. Food Prot.* 79, 1656–1662. <https://doi.org/10.4315/0362-028X.JFP-16-176>
- Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J.P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, Y., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J.C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R.H., Wilson, R.K., Hillier, L.W., McPherson, J.D., Marra, M.A., Mardis, E.R., Fulton, L.A., Chinwalla, A.T., Pepin, K.H., Gish, W.R., Chissoe, S.L., Wendl, M.C., Delehaunty, K.D., Miner, T.L., Delehaunty, A., Kramer, J.B., Cook, L.L., Fulton, R.S., Johnson, D.L., Minx, P.J., Clifton, S.W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N.,

- Cheng, J.F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., Gibbs, R.A., Muzny, D.M., Scherer, S.E., Bouck, J.B., Sodergren, E.J., Worley, K.C., Rives, C.M., Gorrell, J.H., Metzker, M.L., Naylor, S.L., Kucherlapati, R.S., Nelson, D.L., Weinstock, G.M., Sakaki, Y., Fujiyama, A., Hattori, M., Yada, T., Toyoda, A., Itoh, T., Kawagoe, C., Watanabe, H., Totoki, Y., Taylor, T., Weissenbach, J., Heilig, R., Saurin, W., Artiguenave, F., Brottier, P., Bruls, T., Pelletier, E., Robert, C., Wincker, P., Smith, D.R., Doucette-Stamm, L., Rubenfield, M., Weinstock, K., Lee, H.M., Dubois, J., Rosenthal, A., Platzer, M., Nyakatura, G., Taudien, S., Rump, A., Yang, H., Yu, J., Wang, J., Huang, G., Gu, J., Hood, L., Rowen, L., Madan, A., Qin, S., Davis, R.W., Federspiel, N.A., Abola, A.P., Proctor, M.J., Myers, R.M., Schmutz, J., Dickson, M., Grimwood, J., Cox, D.R., Olson, M.V., Kaul, R., Raymond, C., Shimizu, N., Kawasaki, K., Minoshima, S., Evans, G.A., Athanasiou, M., Schultz, R., Roe, B.A., Chen, F., Pan, H., Ramser, J., Lehrach, H., Reinhardt, R., McCombie, W.R., de la Bastide, M., Dedhia, N., Blöcker, H., Hornischer, K., Nordsiek, G., Agarwala, R., Aravind, L., Bailey, J.A., Bateman, A., Batzoglou, S., Birney, E., Bork, P., Brown, D.G., Burge, C.B., Cerutti, L., Chen, H.C., Church, D., Clamp, M., Copley, R.R., Doerks, T., Eddy, S.R., Eichler, E.E., Furey, T.S., Galagan, J., Gilbert, J.G., Harmon, C., Hayashizaki, Y., Haussler, D., Hermjakob, H., Hokamp, K., Jang, W., Johnson, L.S., Jones, T.A., Kasif, S., Kasprzyk, A., Kennedy, S., Kent, W.J., Kitts, P., Koonin, E.V., Korf, I., Kulp, D., Lancet, D., Lowe, T.M., McLysaght, A., Mikkelsen, T., Moran, J.V., Mulder, N., Pollara, V.J., Ponting, C.P., Schuler, G., Schultz, J., Slater, G., Smit, A.F., Stupka, E., Szustakowki, J., Thierry-Mieg, D., Thierry-Mieg, J., Wagner, L., Wallis, J., Wheeler, R., Williams, A., Wolf, Y.I., Wolfe, K.H., Yang, S.P., Yeh, R.F., Collins, F., Guyer, M.S., Peterson, J., Felsenfeld, A., Wetterstrand, K.A., Patrinos, A., Morgan, M.J., de Jong, P., Catanese, J.J., Osoegawa, K., Shizuya, H., Choi, S., Chen, Y.J., Szustakowki, J., International Human Genome Sequencing Consortium, 2001. Initial sequencing and analysis of the human genome. *Nature* 409, 860–921. <https://doi.org/10.1038/35057062>
- Larsen Mette V., Cosentino Salvatore, Rasmussen Simon, Friis Carsten, Hasman Henrik, Marvig Rasmus Lykke, Jelsbak Lars, Sicheritz-Pontén Thomas, Ussery David W., Aarestrup Frank M., Lund Ole, 2012. Multilocus Sequence Typing of Total-Genome-Sequenced Bacteria. *J. Clin. Microbiol.* 50, 1355–1361. <https://doi.org/10.1128/JCM.06094-11>
- Leonard, S.R., Mammel, M.K., Lacher, D.W., Elkins, C.A., 2016. Strain-Level Discrimination of Shiga Toxin-Producing *Escherichia coli* in Spinach Using Metagenomic Sequencing. *PloS One* 11, e0167870. <https://doi.org/10.1371/journal.pone.0167870>
- Leyton, D.L., Sloan, J., Hill, R.E., Doughty, S., Hartland, E.L., 2003. Transfer region of pO113 from Enterohemorrhagic *Escherichia coli*: similarity with R64 and identification of a novel plasmid-encoded autotransporter, EpeA. *Infect. Immun.* 71, 6307–6319. <https://doi.org/10.1128/IAI.71.11.6307-6319.2003>
- Liu, B., Furevi, A., Perepelov, A.V., Guo, X., Cao, H., Wang, Q., Reeves, P.R., Knirel, Y.A., Wang, L., Widmalm, G., 2020. Structure and genetics of *Escherichia coli* O antigens. *FEMS Microbiol. Rev.* 44, 655–683. <https://doi.org/10.1093/femsre/fuz028>
- Lo, Y., Zhang, L., Foxman, B., Zöllner, S., 2015. Whole-genome sequencing of Uropathogenic *Escherichia coli* reveals long evolutionary history of diversity and virulence. *Infect. Genet. Evol. J. Mol. Epidemiol. Evol. Genet. Infect. Dis.* 34, 244–250. <https://doi.org/10.1016/j.meegid.2015.06.023>

- Losada, L., DebRoy, C., Radune, D., Kim, M., Sanka, R., Brinkac, L., Kariyawasam, S., Shelton, D., Fratamico, P.M., Kapur, V., Feng, P.C.H., 2016. Whole genome sequencing of diverse Shiga toxin-producing and non-producing *Escherichia coli* strains reveals a variety of virulence and novel antibiotic resistance plasmids. *Plasmid* 83, 8–11. <https://doi.org/10.1016/j.plasmid.2015.12.001>
- Lu, H., Giordano, F., Ning, Z., 2016. Oxford Nanopore MinION Sequencing and Genome Assembly. *Genomics Proteomics Bioinformatics* 14, 265–279. <https://doi.org/10.1016/j.gpb.2016.05.004>
- Luo, X., Kang, X., Schönhuth, A., 2022. Enhancing Long-Read-Based Strain-Aware Metagenome Assembly. *Front. Genet.* 13, 868280. <https://doi.org/10.3389/fgene.2022.868280>
- Lupolova, N., Chalka, A., Gally, D.L., 2021. Predicting Host Association for Shiga Toxin-Producing *E. coli* Serogroups by Machine Learning. *Methods Mol. Biol.* Clifton NJ 2291, 99–117. [https://doi.org/10.1007/978-1-0716-1339-9\\_4](https://doi.org/10.1007/978-1-0716-1339-9_4)
- Lupolova, N., Dallman, T.J., Matthews, L., Bono, J.L., Gally, D.L., 2016. Support vector machine applied to predict the zoonotic potential of *E. coli* O157 cattle isolates. *Proc. Natl. Acad. Sci. U. S. A.* 113, 11312–11317. <https://doi.org/10.1073/pnas.1606567113>
- Maeshima, N., Fernandez, R.C., 2013. Recognition of lipid A variants by the TLR4-MD-2 receptor complex. *Front. Cell. Infect. Microbiol.* 3, 3. <https://doi.org/10.3389/fcimb.2013.00003>
- Maghini, D.G., Moss, E.L., Vance, S.E., Bhatt, A.S., 2021. Improved high-molecular-weight DNA extraction, nanopore sequencing and metagenomic assembly from the human gut microbiome. *Nat. Protoc.* 16, 458–471. <https://doi.org/10.1038/s41596-020-00424-x>
- Maguire, M., Kase, J.A., Roberson, D., Muruvanda, T., Brown, E.W., Allard, M., Musser, S.M., González-Escalona, N., 2021. Precision long-read metagenomics sequencing for food safety by detection and assembly of Shiga toxin-producing *Escherichia coli* in irrigation water. *PloS One* 16, e0245172. <https://doi.org/10.1371/journal.pone.0245172>
- Maguire, M., Ramachandran, P., Tallent, S., Mammel, M.K., Brown, E.W., Allard, M.W., Musser, S.M., González-Escalona, N., 2023. Precision metagenomics sequencing for food safety: hybrid assembly of Shiga toxin-producing *Escherichia coli* in enriched agricultural water. *Front. Microbiol.* 14, 1221668. <https://doi.org/10.3389/fmicb.2023.1221668>
- Mancusi, R., Trevisani, M., 2014. Enumeration of verocytotoxigenic *Escherichia coli* (VTEC) O157 and O26 in milk by quantitative PCR. *Int. J. Food Microbiol.* 184, 121–127. <https://doi.org/10.1016/j.ijfoodmicro.2014.03.020>
- Mansour, S., Asrar, T., Elhenawy, W., 2023. The multifaceted virulence of adherent-invasive *Escherichia coli*. *Gut Microbes* 15, 2172669. <https://doi.org/10.1080/19490976.2023.2172669>
- Martínez-Castillo, A., Muniesa, M., 2014. Implications of free Shiga toxin-converting bacteriophages occurring outside bacteria for the evolution and the detection of Shiga toxin-producing *Escherichia coli*. *Front. Cell. Infect. Microbiol.* 4, 46. <https://doi.org/10.3389/fcimb.2014.00046>
- Maxam, A.M., Gilbert, W., 1977. A new method for sequencing DNA. *Proc. Natl. Acad. Sci. U. S. A.* 74, 560–564. <https://doi.org/10.1073/pnas.74.2.560>
- Mayjonade, B., Gouzy, J., Donnadieu, C., Pouilly, N., Marande, W., Callot, C., Langlade, N., Muñoz, S., 2016. Extraction of high-molecular-weight genomic DNA for long-read

- sequencing of single molecules. *BioTechniques* 61, 203–205.  
<https://doi.org/10.2144/000114460>
- McArdle, A.J., Kaforou, M., 2020. Sensitivity of shotgun metagenomics to host DNA: abundance estimates depend on bioinformatic tools and contamination is the main issue. *Access Microbiol.* 2, acmi000104. <https://doi.org/10.1099/acmi.0.000104>
- McDaniel, T.K., Jarvis, K., Donnenberg, M.S., Kaper, J.B., 1995. A genetic locus of enterocyte effacement conserved among diverse enterobacterial pathogens. *Proc. Natl. Acad. Sci. U. S. A.* 92, 1664–1668.
- Mellmann, A., Fruth, A., Friedrich, A.W., Wieler, L.H., Harmsen, D., Werber, D., Middendorf, B., Bielaszewska, M., Karch, H., 2009. Phylogeny and disease association of Shiga toxin-producing *Escherichia coli* O91. *Emerg. Infect. Dis.* 15, 1474–1477. <https://doi.org/10.3201/eid1509.090161>
- Meng, Q., Bai, Xiangning, Zhao, A., Lan, R., Du, H., Wang, T., Shi, C., Yuan, X., Bai, Xuemei, Ji, S., Jin, D., Yu, B., Wang, Y., Sun, H., Liu, K., Xu, J., Xiong, Y., 2014. Characterization of Shiga toxin-producing *Escherichia coli* isolated from healthy pigs in China. *BMC Microbiol.* 14, 5. <https://doi.org/10.1186/1471-2180-14-5>
- Meng, Y., Lei, Y., Gao, J., Liu, Y., Ma, E., Ding, Y., Bian, Y., Zu, H., Dong, Y., Zhu, X., 2022. Genome sequence assembly algorithms and misassembly identification methods. *Mol. Biol. Rep.* 49, 11133–11148. <https://doi.org/10.1007/s11033-022-07919-8>
- Meza-Segura, M., Zaidi, M.B., Vera-Ponce de León, A., Moran-Garcia, N., Martinez-Romero, E., Nataro, J.P., Estrada-Garcia, T., 2020. New Insights Into DAEC and EAEC Pathogenesis and Phylogeny. *Front. Cell. Infect. Microbiol.* 10, 572951. <https://doi.org/10.3389/fcimb.2020.572951>
- Mölder, F., Jablonski, K.P., Letcher, B., Hall, M.B., Tomkins-Tinch, C.H., Sochat, V., Forster, J., Lee, S., Twardziok, S.O., Kanitz, A., Wilm, A., Holtgrewe, M., Rahmann, S., Nahnsen, S., Köster, J., 2021. Sustainable data analysis with Snakemake. *F1000Research* 10, 33. <https://doi.org/10.12688/f1000research.29032.2>
- Montero, D.A., Canto, F.D., Velasco, J., Colello, R., Padola, N.L., Salazar, J.C., Martin, C.S., Oñate, A., Blanco, J., Rasko, D.A., Contreras, C., Puente, J.L., Scheutz, F., Franz, E., Vidal, R.M., 2019. Cumulative acquisition of pathogenicity islands has shaped virulence potential and contributed to the emergence of LEE-negative Shiga toxin-producing *Escherichia coli* strains. *Emerg. Microbes Infect.* 8, 486–502. <https://doi.org/10.1080/22221751.2019.1595985>
- Montero, D.A., Velasco, J., Del Canto, F., Puente, J.L., Padola, N.L., Rasko, D.A., Farfán, M., Salazar, J.C., Vidal, R., 2017. Locus of Adhesion and Autoaggregation (LAA), a pathogenicity island present in emerging Shiga Toxin-producing *Escherichia coli* strains. *Sci. Rep.* 7, 7011. <https://doi.org/10.1038/s41598-017-06999-y>
- Montso, P.K., Mlambo, V., Ateba, C.N., 2022. Data on complete genome sequence and annotation of two multidrug resistant atypical Enteropathogenic *Escherichia coli* O177 serotype isolated from cattle faeces. *Data Brief* 42, 108167. <https://doi.org/10.1016/j.dib.2022.108167>
- Montso, P.K., Mlambo, V., Ateba, C.N., 2019. The First Isolation and Molecular Characterization of Shiga Toxin-Producing Virulent Multi-Drug Resistant Atypical Enteropathogenic *Escherichia coli* O177 Serogroup From South African Cattle. *Front. Cell. Infect. Microbiol.* 9, 333. <https://doi.org/10.3389/fcimb.2019.00333>
- Morabito, S., Karch, H., Mariani-Kurkdjian, P., Schmidt, H., Minelli, F., Bingen, E., Caprioli, A., 1998. Enteroaggregative, Shiga toxin-producing *Escherichia coli* O111:H2



- associated with an outbreak of hemolytic-uremic syndrome. *J. Clin. Microbiol.* 36, 840–842. <https://doi.org/10.1128/JCM.36.3.840-842.1998>
- Muniesa, M., Hammerl, J.A., Hertwig, S., Appel, B., Brüßow, H., 2012. Shiga toxin-producing *Escherichia coli* O104:H4: a new challenge for microbiology. *Appl. Environ. Microbiol.* 78, 4065–4073. <https://doi.org/10.1128/AEM.00217-12>
- Mussio, P., Martínez, I., Luzardo, S., Navarro, A., Leotta, G., Varela, G., 2023. Phenotypic and genotypic characterization of Shiga toxin-producing *Escherichia coli* strains recovered from bovine carcasses in Uruguay. *Front. Microbiol.* 14, 1130170. <https://doi.org/10.3389/fmicb.2023.1130170>
- NACMCF, 2019. National Advisory Committee On Microbiological Criteria For, Foods. 2019. “Response to Questions Posed by the Food and Drug Administration Regarding Virulence Factors and Attributes that Define Foodborne Shiga Toxin-Producing *Escherichia coli* (STEC) as Severe Human Pathogens (dagger).” *J Food Prot* 82 (5): 724-767. <https://doi.org/10.4315/0362-028X.JFP-18-479>.
- Nakamura, K., Murase, K., Sato, M.P., Toyoda, A., Itoh, T., Mainil, J.G., Piérard, D., Yoshino, S., Kimata, K., Isobe, J., Seto, K., Etoh, Y., Narimatsu, H., Saito, S., Yatsuyanagi, J., Lee, K., Iyoda, S., Ohnishi, M., Ooka, T., Gotoh, Y., Ogura, Y., Hayashi, T., 2020. Differential dynamics and impacts of prophages and plasmids on the pangenome and virulence factor repertoires of Shiga toxin-producing *Escherichia coli* O145:H28. *Microb. Genomics* 6, e000323. <https://doi.org/10.1099/mgen.0.000323>
- Nakamura, K., Seto, K., Lee, K., Ooka, T., Gotoh, Y., Taniguchi, I., Ogura, Y., Mainil, J.G., Piérard, D., Harada, T., Etoh, Y., Ueda, S., Hamasaki, M., Isobe, J., Kimata, K., Narimatsu, H., Yatsuyanagi, J., Ohnishi, M., Iyoda, S., Hayashi, T., 2023. Global population structure, genomic diversity and carbohydrate fermentation characteristics of clonal complex 119 (CC119), an understudied Shiga toxin-producing *E. coli* (STEC) lineage including O165:H25 and O172:H25. *Microb. Genomics* 9, mgen000959. <https://doi.org/10.1099/mgen.0.000959>
- NanoporeTech, n.d. Nanoporetech/medaka: Sequence correction provided by ONT Research <https://github.com/nanoporetech/medaka>.
- Nataro, J.P., Kaper, J.B., 1998. Diarrheagenic *Escherichia coli*. *Clin. Microbiol. Rev.* 11, 142–201. <https://doi.org/10.1128/CMR.11.1.142>
- Newland, J.W., Strockbine, N.A., Miller, S.F., O’Brien, A.D., Holmes, R.K., 1985. Cloning of Shiga-Like Toxin Structural Genes from a Toxin Converting Phage of *Escherichia coli*. *Science* 230, 179–181. <https://doi.org/10.1126/science.2994228>
- Newton, H.J., Sloan, J., Bulach, D.M., Seemann, T., Allison, C.C., Tauschek, M., Robins-Browne, R.M., Paton, J.C., Whittam, T.S., Paton, A.W., Hartland, E.L., 2009. Shiga toxin-producing *Escherichia coli* strains negative for locus of enterocyte effacement. *Emerg. Infect. Dis.* 15, 372–380. <https://doi.org/10.3201/eid1503.080631>
- Njage, P.M.K., Leekitcharoenphon, P., Hald, T., 2019. Improving hazard characterization in microbial risk assessment using next generation sequencing data and machine learning: Predicting clinical outcomes in Shigatoxigenic *Escherichia coli*. *Int. J. Food Microbiol.* 292, 72–82. <https://doi.org/10.1016/j.ijfoodmicro.2018.11.016>
- Nouws, S., Bogaerts, B., Verhaegen, B., Denayer, S., Piérard, D., Marchal, K., Roosens, N.H.C., Vanneste, K., De Keersmaecker, S.C.J., 2020. Impact of DNA extraction on whole genome sequencing analysis for characterization and relatedness of Shiga toxin-producing *Escherichia coli* isolates. *Sci. Rep.* 10, 14649. <https://doi.org/10.1038/s41598-020-71207-3>

- Nyholm, O., Heinikainen, S., Pelkonen, S., Hallanvuo, S., Haukka, K., Siitonen, A., 2015. Hybrids of Shigatoxigenic and Enterotoxigenic *Escherichia coli* (STEC/ETEC) Among Human and Animal Isolates in Finland. *Zoonoses Public Health* 62, 518–524. <https://doi.org/10.1111/zph.12177>
- O'Brien, A., Lively, T., Chen, M., Rothman, S., Formal, S., 1983. *Escherichia coli* O157:H7 strains associated with haemorrhagic colitis in the United States produce a *Shigella dysenteriae* 1 (Shiga)-like cytotoxin. *Orig. Publ. Vol. 1 Issue 8326* 321, 702. [https://doi.org/10.1016/S0140-6736\(83\)91987-6](https://doi.org/10.1016/S0140-6736(83)91987-6)
- Ochoa, T.J., Cleary, T.G., 2003. Epidemiology and spectrum of disease of *Escherichia coli* O157. *Curr. Opin. Infect. Dis.* 16, 259–263. <https://doi.org/10.1097/00001432-200306000-00013>
- Ogura, Y., Ooka, T., Iguchi, A., Toh, H., Asadulghani, M., Oshima, K., Kodama, T., Abe, H., Nakayama, K., Kurokawa, K., Tobe, T., Hattori, M., Hayashi, T., 2009. Comparative genomics reveal the mechanism of the parallel evolution of O157 and non-O157 Enterohemorrhagic *Escherichia coli*. *Proc. Natl. Acad. Sci. U. S. A.* 106, 17939–17944. <https://doi.org/10.1073/pnas.0903585106>
- Ondov, B. D., Starrett, G. J., Sappington, A., Kostic, A., Koren, S., Buck, C. B., & Phillippy, A. M., 2019. Mash Screen: high-throughput sequence containment estimation for genome discovery. *Genome biology*, 20(1), 232. <https://doi.org/10.1186/s13059-019-1841-x>
- Ørskov, F., Ørskov, I., 1984. 2 Serotyping of *Escherichia coli*\*\*The terminology used to describe the different classes of bacterial antigens is explained in the Preface. However, the authors would like to mention that a different convention is used in some laboratories, for example O:l, K:l, H:7 is equivalent to O1:K1:H7., in: Bergan, T. (Ed.), *Methods in Microbiology*. Academic Press, pp. 43–112. [https://doi.org/10.1016/S0580-9517\(08\)70447-1](https://doi.org/10.1016/S0580-9517(08)70447-1)
- Orskov, I., Orskov, F., Jann, B., Jann, K., 1977. Serology, chemistry, and genetics of O and K antigens of *Escherichia coli*. *Bacteriol. Rev.* 41, 667–710. <https://doi.org/10.1128/br.41.3.667-710.1977>
- Pachchigar, K., Khunt, A., Hetal, B., 2016. DNA QUANTIFICATION. pp. 4–7.
- Pakbin, B., Brück, W.M., Rossen, J.W.A., 2021. Virulence Factors of Enteric Pathogenic *Escherichia coli*: A Review. *Int. J. Mol. Sci.* 22, 9922. <https://doi.org/10.3390/ijms22189922>
- Parente, E., Ricciardi, A., Zotta, T., 2020. The microbiota of dairy milk: A review. *Int. Dairy J.* 107, 104714. <https://doi.org/10.1016/j.idairyj.2020.104714>
- Parsons, B.D., Zelyas, N., Berenger, B.M., Chui, L., 2016. Detection, Characterization, and Typing of Shiga Toxin-Producing *Escherichia coli*. *Front. Microbiol.* 7, 478. <https://doi.org/10.3389/fmicb.2016.00478>
- Paton, A.W., Srimanote, P., Talbot, U.M., Wang, H., Paton, J.C., 2004. A new family of potent AB(5) cytotoxins produced by Shiga toxigenic *Escherichia coli*. *J. Exp. Med.* 200, 35–46. <https://doi.org/10.1084/jem.20040392>
- Paton, A.W., Srimanote, P., Woodrow, M.C., Paton, J.C., 2001. Characterization of Saa, a novel autoagglutinating adhesin produced by locus of enterocyte effacement-negative Shiga-toxigenic *Escherichia coli* strains that are virulent for humans. *Infect. Immun.* 69, 6999–7009. <https://doi.org/10.1128/IAI.69.11.6999-7009.2001>
- Pearson, J.S., Giogha, C., Wong Fok Lung, T., Hartland, E.L., 2016. The Genetics of Enteropathogenic *Escherichia coli* Virulence. *Annu. Rev. Genet.* 50, 493–513. <https://doi.org/10.1146/annurev-genet-120215-035138>

- Penouilh-Suzette, C., Fourré, S., Besnard, G., Godiard, L., Pecrix, Y., 2020. A simple method for high molecular-weight genomic DNA extraction suitable for long-read sequencing from spores of an obligate biotroph oomycete. *J. Microbiol. Methods* 178, 106054. <https://doi.org/10.1016/j.mimet.2020.106054>
- Perna, N.T., Mayhew, G.F., Pósfai, G., Elliott, S., Donnenberg, M.S., Kaper, J.B., Blattner, F.R., 1998. Molecular evolution of a pathogenicity island from Enterohemorrhagic *Escherichia coli* O157:H7. *Infect. Immun.* 66, 3810–3817. <https://doi.org/10.1128/IAI.66.8.3810-3817.1998>
- Pokharel, P., Dhakal, S., Dozois, C.M., 2023. The Diversity of *Escherichia coli* Pathotypes and Vaccination Strategies against This Versatile Bacterial Pathogen. *Microorganisms* 11, 344. <https://doi.org/10.3390/microorganisms11020344>
- Pollard, M.O., Gurdasani, D., Mentzer, A.J., Porter, T., Sandhu, M.S., 2018. Long reads: their purpose and place. *Hum. Mol. Genet.* 27, R234–R241. <https://doi.org/10.1093/hmg/ddy177>
- Porcellato, D., Smistad, M., Bombelli, A., Abdelghani, A., Jørgensen, H.J., Skeie, S.B., 2021. Longitudinal Study of the Bulk Tank Milk Microbiota Reveals Major Temporal Shifts in Composition. *Front. Microbiol.* 12, 616429. <https://doi.org/10.3389/fmicb.2021.616429>
- Pörtner, K., Fruth, A., Flieger, A., Middendorf-Bauchart, B., Mellmann, A., Falkenhorst, G., 2019. Überarbeitung der RKI Empfehlungen für die Wiedenzulassung zu Gemeinschaftseinrichtungen gemäß § 34 IfSG nach EHEC-Infektion. *Epidemiol. Bull.* 506–509. <http://dx.doi.org/10.25646/6414>
- Probert, W.S., McQuaid, C., Schrader, K., 2014. Isolation and identification of an *Enterobacter cloacae* strain producing a novel subtype of Shiga toxin type 1. *J. Clin. Microbiol.* 52, 2346–2351. <https://doi.org/10.1128/JCM.00338-14>
- Quigley, L., O'Sullivan, O., Beresford, T.P., Paul Ross, R., Fitzgerald, G.F., Cotter, P.D., 2012. A comparison of methods used to extract bacterial DNA from raw milk and raw milk cheese. *J. Appl. Microbiol.* 113, 96–105. <https://doi.org/10.1111/j.1365-2672.2012.05294.x>
- Rangel, J.M., Sparling, P.H., Crowe, C., Griffin, P.M., Swerdlow, D.L., 2005. Epidemiology of *Escherichia coli* O157:H7 outbreaks, United States, 1982-2002. *Emerg. Infect. Dis.* 11, 603–609. <https://doi.org/10.3201/eid1104.040739>
- Rasko, D.A., Rosovitz, M.J., Myers, G.S.A., Mongodin, E.F., Fricke, W.F., Gajer, P., Crabtree, J., Sebahia, M., Thomson, N.R., Chaudhuri, R., Henderson, I.R., Sperandio, V., Ravel, J., 2008. The pangenome structure of *Escherichia coli*: comparative genomic analysis of *E. coli* commensal and pathogenic isolates. *J. Bacteriol.* 190, 6881–6893. <https://doi.org/10.1128/JB.00619-08>
- Rasko, D.A., Webster, D.R., Sahl, J.W., Bashir, A., Boisen, N., Scheutz, F., Paxinos, E.E., Sebra, R., Chin, C.-S., Iliopoulos, D., Klammer, A., Peluso, P., Lee, L., Kislyuk, A.O., Bullard, J., Kasarskis, A., Wang, S., Eid, J., Rank, D., Redman, J.C., Steyert, S.R., Frimodt-Møller, J., Struve, C., Petersen, A.M., Krogfelt, K.A., Nataro, J.P., Schadt, E.E., Waldor, M.K., 2011. Origins of the *E. coli* strain causing an outbreak of hemolytic-uremic syndrome in Germany. *N. Engl. J. Med.* 365, 709–717. <https://doi.org/10.1056/NEJMoa1106920>
- Reid, S.D., Herbelin, C.J., Bumbaugh, A.C., Selander, R.K., Whittam, T.S., 2000. Parallel evolution of virulence in pathogenic *Escherichia coli*. *Nature* 406, 64–67. <https://doi.org/10.1038/35017546>

- Richter, M., Rosselló-Móra, R., 2009. Shifting the genomic gold standard for the prokaryotic species definition. *Proc. Natl. Acad. Sci. U. S. A.* 106, 19126–19131.  
<https://doi.org/10.1073/pnas.0906412106>
- RKI, 2021. (Robret Koch-Instituts), Infektionsepidemiologisches Jahrbuch meldepflichtiger Krankheiten für 2020, available at:  
[https://www.rki.de/DE/Content/Infekt/Jahrbuch/Jahrbuch\\_2020.html](https://www.rki.de/DE/Content/Infekt/Jahrbuch/Jahrbuch_2020.html) 116.
- RKI, 2020. (Robret Koch-Instituts), Infektionsepidemiologisches Jahrbuch für 2020, available at [https://www.rki.de/DE/Content/Infekt/Jahrbuch/Jahrbuch\\_2020.html](https://www.rki.de/DE/Content/Infekt/Jahrbuch/Jahrbuch_2020.html), p116 166.
- RKI, 2019. (Robret Koch-Instituts), Infektionsepidemiologisches Jahrbuch meldepflichtiger Krankheiten für 2019, available at:  
[https://www.rki.de/DE/Content/Infekt/Jahrbuch/Jahrbuch\\_2019.pdf](https://www.rki.de/DE/Content/Infekt/Jahrbuch/Jahrbuch_2019.pdf) 136.
- RKI, 2018. (Robret Koch-Instituts), Infektionsepidemiologisches Jahrbuch für 2018, available at: <https://www.rki.de/DE/Content/Infekt/Jahrbuch/Jahrbuecher/2018.html>.
- RKI, 2017. (Robret Koch-Instituts), Infektionsepidemiologisches Jahrbuch für 2017, available at: <https://www.rki.de/DE/Content/Infekt/Jahrbuch/Jahrbuecher/2017.html>.
- Rodríguez-Rubio, L., Haarmann, N., Schwidder, M., Muniesa, M., Schmidt, H., 2021. Bacteriophages of Shiga Toxin-Producing *Escherichia coli* and Their Contribution to Pathogenicity. *Pathog. Basel Switz.* 10, 404.  
<https://doi.org/10.3390/pathogens10040404>
- Russo, T.A., Johnson, J.R., 2000. Proposal for a new inclusive designation for extraintestinal pathogenic isolates of *Escherichia coli*: ExPEC. *J. Infect. Dis.* 181, 1753–1754.  
<https://doi.org/10.1086/315418>
- Safar, H.A., Alatar, F., Nasser, K., Al-Ajmi, R., Alfouzan, W., Mustafa, A.S., 2023. The impact of applying various de novo assembly and correction tools on the identification of genome characterization, drug resistance, and virulence factors of clinical isolates using ONT sequencing. *BMC Biotechnol.* 23, 26. <https://doi.org/10.1186/s12896-023-00797-3>
- Salonen, A., Nikkilä, J., Jalanka-Tuovinen, J., Immonen, O., Rajilić-Stojanović, M., Kekkonen, R.A., Palva, A., de Vos, W.M., 2010. Comparative analysis of fecal DNA extraction methods with phylogenetic microarray: effective recovery of bacterial and archaeal DNA using mechanical cell lysis. *J. Microbiol. Methods* 81, 127–134.  
<https://doi.org/10.1016/j.mimet.2010.02.007>
- Sanderson, N.D., Kapel, N., Rodger, G., Webster, H., Lipworth, S., Street, T.L., Peto, T., Crook, D., Stoesser, N., 2023. Comparison of R9.4.1/Kit10 and R10/Kit12 Oxford Nanopore flowcells and chemistries in bacterial genome reconstruction. *Microb. Genomics* 9, mgen000910. <https://doi.org/10.1099/mgen.0.000910>
- Sandvig, K., 2001. Shiga toxins. *Toxicol. Off. J. Int. Soc. Toxicology* 39, 1629–1635.  
[https://doi.org/10.1016/s0041-0101\(01\)00150-7](https://doi.org/10.1016/s0041-0101(01)00150-7)
- Sanger, F., 1975. The Croonian Lecture, 1975. Nucleotide sequences in DNA. *Proc. R. Soc. Lond. B Biol. Sci.* 191, 317–333. <https://doi.org/10.1098/rspb.1975.0131>
- Santé Publique France, 2022. Investigation de cas groupés de syndrome hémolytique et urémique (SHU) et d'infections à *E. coli* producteurs de Shiga-toxine (STEC) en lien avec la consommation de pizzas Fraîch'Up de marque Buitoni®. Point de situation au 4 mai 2022., available at: <https://www.santepubliquefrance.fr/les-actualites/2022/investigation-de-cas-groupes-de-syndrome-hemolytique-et-uremique-shu-et-d-infections-a-e.-coli-producteurs-de-shiga-toxine-stec-en-lien-avec-la3>.
- Santé Publique France, 2021. Surveillance du syndrome hémolytique et urémique post-diarrhéique chez l'enfant de moins de 15 ans en France en 2021, available at

- <https://www.santepubliquefrance.fr/maladies-et-traumatismes/maladies-infectieuses-d-origine-alimentaire/syndrome-hemolytique-et-uremique-pediatrique/documents/bulletin-national/donnees-de-surveillance-du-syndrome-hemolytique-et-uremique-en-2021>.
- Santé Publique France, 2020. Surveillance du syndrome hémolytique et urémique post-diarrhéique chez l'enfant de moins de 15 ans en France en 2020, available at: <https://www.santepubliquefrance.fr/maladies-et-traumatismes/maladies-infectieuses-d-origine-alimentaire/syndrome-hemolytique-et-uremique-pediatrique/documents/bulletin-national/donnees-de-surveillance-du-syndrome-hemolytique-et-uremique-en-2020>.
- Santé Publique France, 2019. Surveillance du syndrome hémolytique et urémique post-diarrhéique chez l'enfant de moins de 15 ans en France en 2019, available at: <https://www.santepubliquefrance.fr/maladies-et-traumatismes/maladies-infectieuses-d-origine-alimentaire/syndrome-hemolytique-et-uremique-pediatrique/documents/bulletin-national/donnees-de-surveillance-du-syndrome-hemolytique-et-uremique-en-2019>.
- Santé Publique France, 2018. Surveillance du syndrome hémolytique et urémique post-diarrhéique chez l'enfant de moins de 15 ans en France en 2018, available at: <https://www.santepubliquefrance.fr/maladies-et-traumatismes/maladies-infectieuses-d-origine-alimentaire/syndrome-hemolytique-et-uremique-pediatrique/documents/bulletin-national/donnees-de-surveillance-du-syndrome-hemolytique-et-uremique-en-2018>.
- Santé Publique France, 2009. Données de surveillance du syndrome hémolytique et urémique en 2009, Santé Publique France, available at: <https://www.santepubliquefrance.fr/maladies-et-traumatismes/maladies-infectieuses-d-origine-alimentaire/syndrome-hemolytique-et-uremique-pediatrique/documents/bulletin-national/donnees-de-surveillance-du-syndrome-hemolytique-et-uremique-en-2009>.
- Santos, A.C. de M., Santos, F.F., Silva, R.M., Gomes, T.A.T., 2020. Diversity of Hybrid- and Hetero-Pathogenic *Escherichia coli* and Their Potential Implication in More Severe Diseases. *Front. Cell. Infect. Microbiol.* 10, 339. <https://doi.org/10.3389/fcimb.2020.00339>
- Schadt, E.E., Turner, S., Kasarskis, A., 2010. A window into third-generation sequencing. *Hum. Mol. Genet.* 19, R227-240. <https://doi.org/10.1093/hmg/ddq416>
- Schalamun, M., Nagar, R., Kainer, D., Beavan, E., Eccles, D., Rathjen, J.P., Lanfear, R., Schwessinger, B., 2019. Harnessing the MinION: An example of how to establish long-read sequencing in a laboratory using challenging plant tissue from *Eucalyptus pauciflora*. *Mol. Ecol. Resour.* 19, 77–89. <https://doi.org/10.1111/1755-0998.12938>
- Scheutz, F., Teel, L.D., Beutin, L., Piérard, D., Buvens, G., Karch, H., Mellmann, A., Caprioli, A., Tozzoli, R., Morabito, S., Strockbine, N.A., Melton-Celsa, A.R., Sanchez, M., Persson, S., O'Brien, A.D., 2012. Multicenter evaluation of a sequence-based protocol for subtyping Shiga toxins and standardizing Stx nomenclature. *J. Clin. Microbiol.* 50, 2951–2963. <https://doi.org/10.1128/JCM.00860-12>
- Schmidt, H., Hensel, M., 2004. Pathogenicity islands in bacterial pathogenesis. *Clin. Microbiol. Rev.* 17, 14–56. <https://doi.org/10.1128/CMR.17.1.14-56.2004>
- Seemann, T., 2014. Prokka: rapid prokaryotic genome annotation. *Bioinforma. Oxf. Engl.* 30, 2068–2069. <https://doi.org/10.1093/bioinformatics/btu153>

- Serapio-Palacios, A., Finlay, B.B., 2020. Dynamics of expression, secretion and translocation of type III effectors during Enteropathogenic *Escherichia coli* infection. *Curr. Opin. Microbiol.* 54, 67–76. <https://doi.org/10.1016/j.mib.2019.12.001>
- Shen, J., Zhi, S., Guo, D., Jiang, Y., Xu, X., Zhao, L., Lv, J., 2022. Prevalence, Antimicrobial Resistance, and Whole Genome Sequencing Analysis of Shiga Toxin-Producing *Escherichia coli* (STEC) and Enteropathogenic *Escherichia coli* (EPEC) from Imported Foods in China during 2015-2021. *Toxins* 14, 68. <https://doi.org/10.3390/toxins14020068>
- Sheng, H., Duan, M., Hunter, S.S., Minnich, S.A., Settles, M.L., New, D.D., Chase, J.R., Fagnan, M.W., Hovde, C.J., 2018. High-Quality Complete Genome Sequences of Three Bovine Shiga Toxin-Producing *Escherichia coli* O177:H- (*fliC<sub>H25</sub>*) Isolates Harboring Virulent *stx2* and Multiple Plasmids. *Genome Announc.* 6, e01592-17. <https://doi.org/10.1128/genomeA.01592-17>
- Siebert, A., Hofmann, K., Staib, L., Doll, E.V., Scherer, S., Wenning, M., 2021. Amplicon-sequencing of raw milk microbiota: impact of DNA extraction and library-PCR. *Appl. Microbiol. Biotechnol.* 105, 4761–4773. <https://doi.org/10.1007/s00253-021-11353-4>
- Siekanić Grégoire, 2021. Identification of strains of a bacterial species from long reads [Doctoral thesis, Université Rennes 1], available at: <https://ged.univ-rennes1.fr/nuxeo/site/esupversions/81fa81e0-01f4-4d34-8bb3-d71c270efa52?inline>.
- Stalder, T., Press, M.O., Sullivan, S., Liachko, I., Top, E.M., 2019. Linking the resistome and plasmidome to the microbiome. *ISME J.* 13, 2437–2446. <https://doi.org/10.1038/s41396-019-0446-4>
- Steyert, S.R., Sahl, J.W., Fraser, C.M., Teel, L.D., Scheutz, F., Rasko, D.A., 2012. Comparative genomics and *stx* phage characterization of LEE-negative Shiga toxin-producing *Escherichia coli*. *Front. Cell. Infect. Microbiol.* 2, 133. <https://doi.org/10.3389/fcimb.2012.00133>
- Strockbine, N.A., Marques, L.R., Newland, J.W., Smith, H.W., Holmes, R.K., O'Brien, A.D., 1986. Two toxin-converting phages from *Escherichia coli* O157:H7 strain 933 encode antigenically distinct toxins with similar biologic activities. *Infect. Immun.* 53, 135–140. <https://doi.org/10.1128/iai.53.1.135-140.1986>
- Szalo, I.M., Taminiau, B., Mainil, J., 2006. *Escherichia coli* lipopolysaccharide: Structure, biosynthesis and roles. *Ann. Med. Veterinaire* 150, 108–124.
- Tajeddin, E., Ganji, L., Hasani, Z., Ghoalm Mostafaei, F.S., Azimirad, M., Torabi, P., Mohebbi, S.R., Aghili, N., Gouya, M.M., Eshrati, B., Rahbar, M., Mirab Samiee, S., Farzami, M.R., Zali, M.R., Alebouyeh, M., 2020. Shiga toxin-producing bacteria as emerging enteric pathogens associated with outbreaks of foodborne illness in the Islamic Republic of Iran. *East. Mediterr. Health J. Rev. Sante Mediterr. Orient. Al-Majallah Al-Sihhiyah Li-Sharq Al-Mutawassit* 26, 976–981. <https://doi.org/10.26719/emhj.19.102>
- Tenaillon, O., Skurnik, D., Picard, B., Denamur, E., 2010. The population genetics of commensal *Escherichia coli*. *Nat. Rev. Microbiol.* 8, 207–217. <https://doi.org/10.1038/nrmicro2298>
- Thomas, C.M., Nielsen, K.M., 2005. Mechanisms of, and barriers to, horizontal gene transfer between bacteria. *Nat. Rev. Microbiol.* 3, 711–721. <https://doi.org/10.1038/nrmicro1234>
- Thorpe, C.M., 2004. Shiga Toxin—Producing *Escherichia coli* Infection. *Clin. Infect. Dis.* 38, 1298–1303. <https://doi.org/10.1086/383473>

- Tobe, T., Beatson, S.A., Taniguchi, H., Abe, H., Bailey, C.M., Fivian, A., Younis, R., Matthews, S., Marches, O., Frankel, G., Hayashi, T., Pallen, M.J., 2006. An extensive repertoire of type III secretion effectors in *Escherichia coli* O157 and the role of lambdoid phages in their dissemination. *Proc. Natl. Acad. Sci. U. S. A.* 103, 14941–14946. <https://doi.org/10.1073/pnas.0604891103>
- Tonkin-Hill, G., MacAlasdair, N., Ruis, C., Weimann, A., Horesh, G., Lees, J.A., Gladstone, R.A., Lo, S., Beaudoin, C., Floto, R.A., Frost, S.D.W., Corander, J., Bentley, S.D., Parkhill, J., 2020. Producing polished prokaryotic pangenomes with the Panaroo pipeline. *Genome Biol.* 21, 180. <https://doi.org/10.1186/s13059-020-02090-4>
- Tozzoli, R., Grande, L., Michelacci, V., Ranieri, P., Maugliani, A., Caprioli, A., Morabito, S., 2014. Shiga toxin-converting phages and the emergence of new pathogenic *Escherichia coli*: a world in motion. *Front. Cell. Infect. Microbiol.* 4, 80. <https://doi.org/10.3389/fcimb.2014.00080>
- Trigodet, F., Lolans, K., Fogarty, E., Shaiber, A., Morrison, H.G., Barreiro, L., Jabri, B., Eren, A.M., 2022. High molecular weight DNA extraction strategies for long-read sequencing of complex metagenomes. *Mol. Ecol. Resour.* 22, 1786–1802. <https://doi.org/10.1111/1755-0998.13588>
- Tunnsjø, H.S., Ullmann, I.F., Charnock, C., 2023. A preliminary study of the use of MinION sequencing to specifically detect Shiga toxin-producing *Escherichia coli* in culture swipes containing multiple serovars of this species. *Sci. Rep.* 13, 8239. <https://doi.org/10.1038/s41598-023-35279-1>
- Usman, T., Yu, Y., Liu, C., Fan, Z., Wang, Y., 2014. Comparison of methods for high quantity and quality genomic DNA extraction from raw cow milk. *Genet. Mol. Res. GMR* 13, 3319–3328. <https://doi.org/10.4238/2014.April.29.10>
- van Hoek, A.H.A.M., Lee, S., van den Berg, R.R., Rapallini, M., van Overbeeke, L., Opsteegh, M., Bergval, I., Wit, B., van der Weijden, C., van der Giessen, J., van der Voort, M., 2023. Virulence and antimicrobial resistance of Shiga toxin-producing *Escherichia coli* from dairy goat and sheep farms in The Netherlands. *J. Appl. Microbiol.* 134, lxad119. <https://doi.org/10.1093/jambio/lxad119>
- Vaser, R., Sović, I., Nagarajan, N., Šikić, M., 2017. Fast and accurate *de novo* genome assembly from long uncorrected reads. *Genome Res.* 27, 737–746. <https://doi.org/10.1101/gr.214270.116>
- Vélez, M.V., Colello, R., Etcheverría, A.I., Vidal, R.M., Montero, D.A., Acuña, P., Guillén Fretes, R.M., Toro, M., Padola, N.L., 2020. Distribution of Locus of Adhesion and Autoaggregation and *hes* Gene in STEC Strains from Countries of Latin America. *Curr. Microbiol.* 77, 2111–2117. <https://doi.org/10.1007/s00284-020-02062-8>
- Vicedomini, R., Quince, C., Darling, A.E., Chikhi, R., 2021. Strainberry: automated strain separation in low-complexity metagenomes using long reads. *Nat. Commun.* 12, 4485. <https://doi.org/10.1038/s41467-021-24515-9>
- Vygen-Bonnet, S., Rosner, B., Wilking, H., Fruth, A., Prager, R., Kossow, A., Lang, C., Simon, S., Seidel, J., Faber, M., Schielke, A., Michaelis, K., Holzer, A., Kamphausen, R., Kalhöfer, D., Thole, S., Mellmann, A., Flieger, A., Stark, K., 2017. Ongoing haemolytic uraemic syndrome (HUS) outbreak caused by sorbitol-fermenting (SF) Shiga toxin-producing *Escherichia coli* (STEC) O157, Germany, December 2016 to May 2017. *Eurosurveillance*.
- Wang, L., Rothemund, D., Curd, H., Reeves, P.R., 2003. Species-wide variation in the *Escherichia coli* flagellin (H-antigen) gene. *J. Bacteriol.* 185, 2936–2943. <https://doi.org/10.1128/JB.185.9.2936-2943.2003>

- Wang, Yunhao, Zhao, Y., Bollas, A., Wang, Yuru, Au, K.F., 2021. Nanopore sequencing technology, bioinformatics and applications. *Nat. Biotechnol.* 39, 1348–1365. <https://doi.org/10.1038/s41587-021-01108-x>
- Wick, R.R., Holt, K.E., 2019. Benchmarking of long-read assemblers for prokaryote whole genome sequencing. *F1000Research* 8, 2138. <https://doi.org/10.12688/f1000research.21782.4>
- Wickham, H., 2016. *ggplot2: Elegant Graphics for Data Analysis, Use R!* Springer International Publishing.
- Wickham, M.E., Lupp, C., Mascarenhas, M., Vazquez, A., Coombes, B.K., Brown, N.F., Coburn, B.A., Deng, W., Puente, J.L., Karmali, M.A., Finlay, B.B., 2006. Bacterial genetic determinants of non-O157 STEC outbreaks and hemolytic-uremic syndrome after infection. *J. Infect. Dis.* 194, 819–827. <https://doi.org/10.1086/506620>
- Wickramarachchi, A., Lin, Y., 2022. Binning long reads in metagenomics datasets using composition and coverage information. *Algorithms Mol. Biol. AMB* 17, 14. <https://doi.org/10.1186/s13015-022-00221-z>
- Wirth, T., Falush, D., Lan, R., Colles, F., Mensa, P., Wieler, L.H., Karch, H., Reeves, P.R., Maiden, M.C.J., Ochman, H., Achtman, M., 2006. Sex and virulence in *Escherichia coli*: an evolutionary perspective. *Mol. Microbiol.* 60, 1136–1151. <https://doi.org/10.1111/j.1365-2958.2006.05172.x>
- Wong, A.R.C., Pearson, J.S., Bright, M.D., Munera, D., Robinson, K.S., Lee, S.F., Frankel, G., Hartland, E.L., 2011. Enteropathogenic and Enterohaemorrhagic *Escherichia coli*: even more subversive elements. *Mol. Microbiol.* 80, 1420–1438. <https://doi.org/10.1111/j.1365-2958.2011.07661.x>
- Wood, D.E., Lu, J., Langmead, B., 2019. Improved metagenomic analysis with Kraken 2. *Genome Biol.* 20, 257. <https://doi.org/10.1186/s13059-019-1891-0>
- Wu, Y.-W., Simmons, B.A., Singer, S.W., 2016. MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinform. Oxf. Engl.* 32, 605–607. <https://doi.org/10.1093/bioinformatics/btv638>
- Yang, T., Gao, F., 2022. High-quality pan-genome of *Escherichia coli* generated by excluding confounding and highly similar strains reveals an association between unique gene clusters and genomic islands. *Brief. Bioinform.* 23, bbac283. <https://doi.org/10.1093/bib/bbac283>
- Zhang, H., Jain, C., Aluru, S., 2020. A comprehensive evaluation of long read error correction methods. *BMC Genomics* 21, 889. <https://doi.org/10.1186/s12864-020-07227-0>
- Zhang, L., Chen, T., Wang, Y., Zhang, S., Lv, Q., Kong, D., Jiang, H., Zheng, Y., Ren, Y., Huang, W., Liu, P., Jiang, Y., 2022. Comparison Analysis of Different DNA Extraction Methods on Suitability for Long-Read Metagenomic Nanopore Sequencing. *Front. Cell. Infect. Microbiol.* 12, 919903. <https://doi.org/10.3389/fcimb.2022.919903>
- Zhi, S., Parsons, B.D., Szelewicki, J., Yuen, Y.T.K., Fach, P., Delannoy, S., Li, V., Ferrato, C., Freedman, S.B., Lee, B.E., Pang, X.-L., Chui, L., 2021. Identification of Shiga-Toxin-Producing *Shigella* Infections in Travel and Non-Travel Related Cases in Alberta, Canada. *Toxins* 13, 755. <https://doi.org/10.3390/toxins13110755>



---

## Abbreviations

---

### A

A/E: attaching/effacing

AAFs: Aggregative adherence fimbriae

AIEC: Adherent invasive *Escherichia coli*

ANSES: Agence nationale de sécurité sanitaire, alimentation, environnement, travail

ARG(s): Antimicrobial resistance gene(s)

### B

BPW: Buffered peptone water

BfR: Bundesinstitute für Risikobewertung

BHI: Brain heart infusion

### C

CF(s): colonization factor(s)

CFU: colony forming unit(s)

cgMLST: core genome multi locus sequence type

CRISPR: clustered regularly interspaced short palindromic region

### D

DAEC: Diffusely-adherent *Escherichia coli*

DBG: De Bruijn graph

DEC: Diarrheagenic *Escherichia coli*

DIN: DNA integrity number

DNA: deoxyribo-nucleic acid

### E

EAEC: Enteroaggregative *Escherichia coli*

ECDC: European Center for Disease Prevention and Control

EFSA: European Food Safety Administration

EHEC: Enterohaemorrhagic *Escherichia coli*

EIEC: Enteroinvasive *Escherichia coli*

EPEC: Enteropathogenic *Escherichia coli*

ETEC: Enterotoxigenic *Escherichia coli*

EU: European Union

EURL: European Union Reference Laboratory

ExPEC: Extra-intestinal pathogenic *Escherichia coli*

## F

FAO: Food and Agriculture Organization of the United Nations

FGS: First-generation sequencing

FSIS: Food safety and Inspection service

## G

Gb3: Globotriosylceramide 3

gDNA: genomic DNA

GI(s): Genomic island(s)

GPU(s): Graphics processing unit(s)

## H

HC: Hemorrhagic colitis

HGT: Horizontal gene transfer

HMW: High molecular weight

HUS: Hemolytic uremic syndrome

## I

ISO: International Organization for Standardization

## L

LAA: Locus of autoaggregation and adhesion

LEE: Locus of enterocyte effacement

LPS: Lipopolysaccharide

LT(s): Heat-labile toxin(s)

## M

MAG(s): Metagenome-assembled genome(s)

MGE(s): Mobile genetic element(s)

ML: Machine learning

MLST: Multi locus sequence type

## N

NACMCF: National Advisory Committee on Microbiological Criteria for Foods

NGS: Next-generation sequencing

NMEC: Neonatal meningitis-causing *Escherichia coli*

NRC: National reference center

NRL: National reference laboratory

## O

O-AGC: O-antigen cluster

OI(s): O-island(s)

OLC: Overlap-layout-consensus

ONT: Oxford Nanopore Technology

## P

PAI(s): Pathogenicity island(s)

PCR: Polymerase chain reaction

PFGE: Pulse-field gel electrophoresis

PT: Proficiency test

## Q

qPCR: Quantitative PCR

qdPCR: Quantitative digital PCR

## R

RKI: Robert Koch Institute

RNA: Ribonucleic acid

rRNA: Ribosomal RNA

## S

SGS: Second-generation sequencing

SNV(s): Small nucleotide variation(s)

SV: Structural variation

SPEC: Septicemic pathogenic *Escherichia coli*

ST: Sequence type

ST(s) Heat-stable toxin(s)

STEC: Shiga toxin-producing *Escherichia coli*

## T

T3SS: Type three-secretion system

TGS: Third-generation sequencing

TS: Technical specification

## U

UPEC: Uropathogenic *Escherichia coli*

US: United States

USA: United States of America

USDA: The United States Department of Agriculture

## V

VT: Verotoxigenic toxin

## W

wgMLST: whole genome MLST

WGS: Whole-genome sequencing

WHO: World health organization

---

## Scientific valorization

---

### Scientific papers:

**Jaudou S**, Tran ML, Vorimore F, Fach P, Delannoy S. Evaluation of high molecular weight DNA extraction methods for long-read sequencing of Shiga toxin-producing *Escherichia coli*. **PLoS One**. **2022** Jul 13;17(7):e0270751. doi: 10.1371/journal.pone.0270751.

**Jaudou S**, Deneke C, Tran ML, Schuh E, Goehler A, Vorimore F, Malorny B, Fach P, Grützke J, Delannoy S. A step forward for Shiga toxin-producing *Escherichia coli* identification and characterization in raw milk using long-read metagenomics. **Microb Genom**. **2022** Nov;8(11):mgen000911. doi: 10.1099/mgen.0.000911.

**Jaudou S**, Tran ML, Vorimore F, Fach P, Delannoy S. Hybrid Assembly from 75 *E. coli* Genomes Isolated from French Bovine Food Products between 1995 and 2016. **Microbiol Resour Announc**. **2023** Mar 16;12(3):e0109522. doi: 10.1128/mra.01095-22.

Vorimore F#, **Jaudou S**#, Tran ML, Richard H, Fach P, Delannoy S. Combination of whole genome sequencing and supervised machine learning provides unambiguous identification of *eae*-positive Shiga toxin-producing *Escherichia coli*. **Front Microbiol**. **2023** May 12;14:1118158. doi: 10.3389/fmicb.2023.1118158.

**Jaudou S**, Deneke C, Tran ML, Salzinger C, Vorimore F, Goehler A, Schuh E, Malorny B, Fach P, Grützke J, Delannoy S. Exploring Long-Read Metagenomics for Full Characterization of Shiga Toxin-Producing *Escherichia coli* in Presence of Commensal *E. coli*. **Microorganisms**. **2023** Aug 9;11(8):2043. doi: 10.3390/microorganisms11082043.

### Posters:

- Poster, **VTEC2023 (Banff, Canada)**, May 2023, « Evaluation of the usefulness of acriflavine in the enrichment of STEC from milk samples »
- Poster, **ANSES scientific and doctoral days**, October 2022, « A step forward for Shiga toxin-producing *Escherichia coli* identification and characterization in raw milk using long-read metagenomics »
- Poster, **ANSES scientific and doctoral days**, October 2021, « Evaluation of high molecular weight DNA extraction methods for long-read sequencing of Shiga toxin-producing *Escherichia coli* »

### Communications:

- Oral talk, 3-minute thesis, **ANSES scientific and doctoral days**, October 2023 « Long-read sequencing to characterize *eae*-positive STEC in raw milk »
- Oral talk, **Outstanding Oral Talk – Second Place, VTEC2023 (Banff, Canada)**, May 2023 « Application of MinION sequencing as a tool for the characterization of STEC in raw milk »,
- Oral talk, **17th Annual Workshop of the NRLs for *E. coli* in the EU, Istituto Superiore di Sanità - Rome (online)**, October 2022 « Characterization of *E. coli* from artificially contaminated cheese samples (PT33) using long-read metagenomics »

# **METADETECT: DETECTION OF SHIGA TOXIN-PRODUCING *ESCHERICHIA COLI* WITH NOVEL METAGENOMICS APPROACHES AND ITS APPLICATION ON DAIRY FARMS IN FRANCE AND GERMANY**

## **SUMMARY:**

Current methods for characterizing Shiga toxin producing *Escherichia coli* (STEC) require strain isolation, which is complicated by the fact that there is no specific isolation medium that clearly distinguishes STEC from non-pathogenic commensal *E. coli*. Therefore, obtaining strain information using a metagenomics approach would avoid the need to isolate a strain for its full characterization. In this project, in collaboration with the BfR in Germany, we will evaluate whether new long-read metagenomics approaches can unambiguously determine whether specific markers (*stx* and *eae* genes) of typical EHEC (Enterohaemorrhagic *E. coli*) are co-located in the same strain. Third generation hybrid sequencing approaches will be evaluated. Appropriate bioinformatics pipelines, developed in collaboration with BfR, will be evaluated for the analysis of the metagenomic results. These methods will be applied to presumptive positive or naturally contaminated samples.

**KEYWORDS:** metagenomics, Enterohemorrhagic *Escherichia coli*, long-read sequencing, Shiga toxin-producing *Escherichia coli*, raw milk, food microbiology.

# METADETECT : DETECTION DES *ESCHERICHIA COLI* PRODUCTRICES DE SHIGA TOXINE A L'AIDE DE NOUVELLES APPROCHES DE METAGENOMIQUES ET APPLICATION DANS DES EXPLOITATIONS LAITIERES EN FRANCE ET EN ALLEMAGNE

## RÉSUMÉ :

Les méthodologies actuelles de caractérisation d'*Escherichia coli* productrice de toxine Shiga (STEC) nécessitent l'isolement de la souche, ce qui est compliqué par le fait qu'il n'existe pas de milieu d'isolement spécifique qui distingue clairement les STEC des *E. coli* commensaux non pathogènes. Par conséquent, obtenir des informations sur les souches en utilisant une approche métagénomique éviterait d'isoler les souches pour les caractériser complètement. Dans le cadre du projet, en collaboration avec le BfR en Allemagne, nous évaluerons si de nouvelles approches de métagénomique à lecture longue pourraient déterminer sans ambiguïté si des marqueurs spécifiques d'EHEC typiques (*E. coli* entérohémorragique) sont co-localisés dans une même souche. Les approches de séquençage hybrides de deuxième et troisième génération seront évaluées. Des pipelines bio-informatiques seront évalués pour analyser les résultats de l'analyse métagénomique. Ces méthodes seront appliquées sur des échantillons naturellement contaminés (présomptifs positifs ou confirmés).

**MOTS CLÉS :** Métagénomique, *Escherichia coli* Entérohémorragique, séquençage long-read, *Escherichia coli* productrices de Shiga toxine, lait cru, microbiologie alimentaire.