# Come utilizzare i risultati della caratterizzazione dei ceppi STEC ai fini della sorveglianza in chiave One-Health

Stefano Andreoni[1], Valeria Michelacci[2], Rosangela Tozzoli[2], Eleonora Ventola[2]

[1]*Azienda Ospedaliero Universitaria Maggiore della Carità, Novara*

[2]*Dipartimento di Sicurezza Alimentare, Nutrizione e Sanità Pubblica Veterinaria Istituto Superiore di Sanità, Roma*

## Indice

## Caratterizzazione molecolare di ceppi di *Escherichia coli* produttori di Shiga tossina

Gli *Escherichia coli* produttori di Shiga-tossina sono per definizione ceppi che producono la Shiga-tossina (Stx) e sono quindi accomunati dalla presenza nel proprio genoma di geni *stx* che le codificano. Tali geni sono veicolati da batteriofagi, che lisogenizzano ceppi di *E. coli*. Linee clonali distinte sono andate incontro evolutivamente a tale evento di lisogenizzazione, dando origine a popolazioni di ceppi STEC molto diverse tra loro, che in alcuni casi possiedono numerosi altri fattori di virulenza accessori. La caratterizzazione fine dei ceppi STEC è quindi di particolare importanza, permettendo di poter distinguere gli stipiti circolanti e indirizzando correttamente le indagini epidemiologiche dei casi di infezione. Ad esempio, il rilevamento di ceppi STEC con caratteristiche diverse, isolati da più pazienti con infezione in atto può consentire di escludere con elevata probabilità che esista una correlazione epidemiologica tra i casi, cioè che i pazienti siano stati esposti alla stessa fonte di infezione. Analogamente caratteristiche differenti di ceppi STEC isolati da un caso clinico e da un alimento sospettato come possibile veicolo di infezione, permette di escluderne la correlazione. Al contrario, l'identificazione di uno stesso ceppo STEC da due casi di infezione con un possibile collegamento epidemiologico (es. provenienza dalla stessa area geografica) o in un paziente e in una possibile sorgente di infezione (es. un alimento) rappresenta un'evidenza 'robusta' dell'esistenza di un nesso epidemiologico che può consentire di intervenire tempestivamente con azioni volte ad interrompere la catena di trasmissione dell'infezione.

L'utilizzo di metodi molecolari permette questa fine tipizzazione, in quanto in grado non solo di evidenziare la presenza dei geni codificanti le Stx, ma anche

sottotipizzarli, determinare il sierogruppo di appartenenza, identificare geni di virulenza accessori e infine stabilire correlazioni a livello genomico.
La caratterizzazione fine degli isolati di *E. coli* e la genomica in particolare, ci permette cioè di discriminare con grande accuratezza il livello di correlazione tra due o più ceppi, confrontando le loro caratteristiche. Quanto maggiore è il numero di caratteri che consideriamo nel confronto, tanto migliore sarà il livello di discriminazione dell'analisi. Un po' come quando cerchiamo di stabilire se due persone che viaggiano in treno accanto a noi sono tra loro parenti. Non ci basta guardare il colore degli occhi per esprimere un giudizio, ma proviamo a farlo considerando anche il colore dei capelli, il colore della pelle, la statura, la forma del naso e delle orecchie, la forma delle mani, l'ovale del volto ecc. aggiungiamo cioè dettagli che ci permettano una migliore discriminazione. La stessa cosa facciamo con gli isolati. Non ci accontentiamo di determinare gli antigeni somatici O e flagellare H, ma abbiamo bisogno di aggiungere la caratterizzazione dei geni *stx1*, *stx2*, *eae* e altre caratteristiche genetiche. I geni di virulenza accessori sono importanti nella caratterizzazione dei ceppi STEC, in quanto la sola presenza dei geni *stx* non sembra sufficiente per causare le forme più severe di malattia. I ceppi STEC più frequentemente isolati da casi di diarrea emorragica o di Sindrome Emolitico-Uremica (SEU) generalmente possiedono geni di virulenza accessori coinvolti nella colonizzazione dell'ospite e nel meccanismo patogenetico in generale La maggior parte di questi è in grado di produrre la tipica lesione istopatologica denominata *"Attaching and Effacing"* (A/E) grazie alla presenza nel genoma dell'isola di patogenicità (una regione genomica acquisita evolutivamente per via orizzontale da altri microrganismi che veicola geni di virulenza) denominata Locus for Enterocyte Effacement (LEE) (1), per la quale il gene *eae* rappresenta un marcatore.
Sono stati comunque descritti anche ceppi STEC "*ibridi*", ovvero stipiti capaci di produrre le Stx, ma che possiedono caratteristiche di virulenza in comune con altri patotipi di *E. coli*: tali ceppi possono essere evidenziati mediante la ricerca dei geni di virulenza caratteristici di altri patotipi, come il gene *aggR*, coinvolto nel meccanismo di colonizzazione dei ceppi *E. coli* Enteroaggregativi (EAEC), e

i geni *st* ed *lt* codificanti le principali tossine associate agli *E. coli* enterotossigenici (ETEC).

I sierogruppi STEC più frequentemente associati a malattia grave nell'uomo sono i cosiddetti "top-5" e cioè O157, O26, O103, O145 e O111, accomunati dalla presenza dell'isola di patogenicità LEE nel proprio genoma. Nove ulteriori sierogruppi (O45, O55, O80, O91, O104, O113, O121, O128 e O146) causano, insieme a questi, la maggior parte delle infezioni gravi nell'uomo a livello mondiale. L'insieme di questi 14 sierogruppi viene infatti definito "top-14". L'identificazione del sierogruppo dei ceppi STEC può dare un rapido riscontro in termini epidemiologici, permettendo ad esempio di escludere in tempi rapidi la correlazione tra due casi di infezione, qualora gli STEC isolati da essi fossero di sierogruppo diverso.

Anche la sottotipizzazione dei geni *stx* non è soltanto un esercizio tassonomico. La famiglia dei geni *stx* si divide in due tipi principali, *stx1* e *stx2*, sulla base delle differenze antigeniche delle tossine da questi prodotte. Ciascuno di questi due tipi principali si compone di diverse varianti alleliche, definite sottotipi, in particolare tre sottotipi per i geni *stx1* (*stx1a*, *stx1c* e *stx1d*) e sette sottotipi per i geni *stx2* (da *stx2a* a *stx2g*). Alcuni dei sottotipi sono associati ad infezione caratterizzata da sintomatologia più grave, come la SEU, mentre altri sono associati principalmente a casi di malattia con esiti più lievi o sono prodotti da ceppi di *E. coli* che generalmente non causano malattia nell'uomo (2). Inoltre molti studi hanno indicato una associazione più alta con il rischio di sviluppare la SEU, la colite emorragica o entrambe per i sottitipi *stx2a* e *stx2d* (2, 3).

Oggi il paradigma di patogenicità basato sulla determinazione del sierogruppo O è stato superato e sostituito proprio da un sistema basato sulla tipizzazione e sottotipizzazione dei geni *stx* e di altri geni di virulenza, poiché l'estrema plasticità genomica degli STEC rende possibile l'emergere di nuovi sierogruppi O ad elevata patogenicità per l'uomo attraverso l'acquisizione di geni di virulenza da altri microrganismi per via orizzontale. La classificazione più recente dei ceppi STEC in base al contenuto in geni di virulenza e al potenziale patogeno stimato è riportata in Tabella 1 (3).

**Tabella 1.** Combinazioni dei geni di virulenza di STEC e potenziale stimato di causare diarrea (D), diarrea ematica (DE) e Sindrome Emolitico-Uremica (SEU) [1] (3)

| Livello | Gene/Geni | Potenziale per: |
|---|---|---|
| 1 | *stx2a* + *eae* o *aggR* | D/DE/SEU |
| 2 | *stx2d* | D/DE/SEU[2] |
| 3 | *stx2c* + *eae* | D/DE[3] |
| 4 | *stx1a* + *eae* | D/DE[3] |
| 5 | Altri sottotipi di *stx* | D |

NOTE: 1. a seconda della suscettibilità dell'ospite o di altri fattori; per esempio: trattamento antibiotico
2. associazione con SEU dipendente dalla variante *stx2d* e dal background del ceppo
3. è stato riportato che alcuni sottotipi causano DE e, in rare occasioni, SEU

Ad oggi sono disponibili metodi di facile applicazione, che permettono l'identificazione dei sierogruppi top-14, dei principali geni di virulenza tipici degli STEC e la sottotipizzazione dei geni *stx* mediante semplici reazioni di PCR convenzionale o PCR Real Time. Di seguito illustreremo brevemente le principali metodiche a disposizione, rimandando per gli approfondimenti alla lettura dei testi completi dei metodi, disponibili attraverso la pagina del sito web del Laboratorio Nazionale ed Europeo di Riferimento per *Escherichia coli* dedicata ai metodi di laboratorio (https://www.iss.it/en/vtec-laboratory-methods).

## Metodi per la caratterizzazione molecolare mediante PCR convenzionale

### Identificazione e caratterizzazione di ceppi STEC attraverso l'amplificazione dei principali geni di virulenza tramite PCR convenzionale

Al fine di determinare la presenza dei principali fattori di virulenza dei ceppi STEC, è disponibile un metodo che sfrutta una PCR convenzionale multiplex in grado di identificare simultaneamente la presenza dei geni *stx1, stx2* e del gene *eae*. La coppia di oligonucleotidi utilizzata per identificare i geni *stx1* è in

grado di rilevare la presenza di tutti e tre i principali sottotipi *stx1* finora descritti. Al contrario, a causa dell'elevata differenza nucleotidica del sottotipo *stx2f*, non è stato possibile selezionare un'unica coppia di oligonucleotidi in grado di identificare tutti e sette i sottotipi *stx2*. In particolare, solo per identificare il sottotipo *stx2f* è necessario aggiungere alla miscela di reazione una ulteriore coppia di oligonucleotidi specifica per il sottotipo *stx2f*. Nella stessa miscela di reazione, oltre alle tre coppie descritte finora, ne viene aggiunta una quarta specifica per la rilevazione del gene *eae*. Il metodo prevede la preparazione del DNA stampo a partire da singole colonie prelevate da piastre di crescita risospese in acqua sterile, che viene poi sottoposta a semplice bollitura. La reazione prevede un profilo termico particolare, ottimizzato per questa PCR multiplex, che impiega circa tre ore e mezza per essere ultimato in un termociclatore convenzionale. Il risultato può essere poi visualizzato mediante elettroforesi su gel di agarosio alla concentrazione di almeno 1,5%.
Il testo completo del metodo è disponibile a questo link diretto:
https://www.iss.it/documents/20126/0/EURL_VTEC_Method_01_Rev+1.pdf/3941c7fc-98ba-a5c5-4f2b-f537218efa46?t=1644308327642

## Identificazione dei sierogruppi STEC più frequentemente associati alle infezioni umane (top-14) mediante amplificazione in PCR convenzionale dei geni associati agli antigeni O

Il metodo prevede l'identificazione dei geni associati agli antigeni O più frequentemente associati a malattia grave nell'uomo attraverso reazioni di PCR convenzionale. In particolare i sierogruppi compresi nel campo di applicazione del metodo sono i seguenti: O26, O45, O55, O80, O91, O103, O104, O111, O113, O121, O128, O145, O146 e O157. I geni target del metodo sono diverse varianti alleliche dei geni *wzx* e *wzy*, associate ai diversi sierogruppi, il gene *wbgN* nel caso del sierogruppo O55 e il gene *rfb* nel caso del sierogruppo O104. Le reazioni specifiche per i sierogruppi O26, O103, O111, O121, O145 e O157 (che rappresentano i sierogruppi top-5, con l'aggiunta del sierogruppo O121) possono essere convenientemente eseguite come una unica PCR multiplex, avendo l'accortezza di utilizzare una polimerasi HotStart per

garantire la specificità di reazione. È possibile preparare miscele di oligonucleotidi pronte all'uso contenenti per ciascun primer una concentrazione 10 volte superiore a quella prevista per la soluzione finale, permettendo quindi di allestire reazioni in grado di identificare questi sei sierogruppi molto rilevanti in chiave epidemiologica con un'unica reazione di PCR, facile e veloce da preparare. La purificazione del DNA stampo può avvenire attraverso semplice bollitura di una colonia dispersa in acqua sterile e la rilevazione dei risultati viene eseguita mediante elettroforesi su gel di agarosio, ad una concentrazione del 2% nel caso di PCR multiplex.
Il testo completo del metodo è disponibile a questo link diretto:
https://www.iss.it/documents/20126/0/EURL_VTEC_Method_03_Rev+2.pdf/5f7cf968-b58e-2524-501c-ef3cc5dd5fde?t=1644309161824

## Identificazione dei sottotipi dei geni codificanti le Shiga-tossine mediante PCR convenzionale

Il metodo prevede la sottotipizzazione attraverso reazioni di PCR convenzionale dei geni *stx1* e *stx2* codificanti le Shiga-tossine in un totale di dieci sottotipi: tre sottotipi per la *stx1* (*stx1a*, *stx1c* e *stx1d*) e sette sottotipi per la *stx2* (da *stx2a* a *stx2g*). Per distinguere i tre sottotipi dei geni *stx1* viene eseguita una unica PCR triplex, mentre per distinguere i sottotipi *stx2* vengono eseguite sette diverse reazioni di PCR, che possono essere suddivise in due gruppi (*stx2a-stx2b-stx2c* e *stx2d-stx2e-stx2f-stx2g*) per profilo termico di amplificazione. Il metodo è stato validato utilizzando una *Taq* polimerasi "HotStart" per ridurre la possibilità di amplificazione di prodotti aspecifici, ed il metodo necessita l'ottimizzazione delle condizioni di reazione nei diversi laboratori con l'utilizzo di diversi reagenti. La preparazione del DNA stampo prevede la dispersione di 25 µl di coltura pura *overnight* del ceppo da sottotipizzare in 975 µl di acqua sterile e la bollitura della miscela così preparata. La rilevazione dei risultati viene eseguita mediante elettroforesi su gel di agarosio, ad una concentrazione del 2% nel caso della PCR triplex per i geni *stx1*.

Il testo completo del metodo è disponibile a questo link diretto:
https://www.iss.it/documents/20126/0/EURL_VTEC_Method_06_Rev+2.pdf/abf8f392-e0ec-9bc8-9873-947d6b40d27b?t=1644309235269

## Metodi per caratterizzazione molecolare con PCR Real Time

### Identificazione e caratterizzazione di ceppi STEC attraverso l'amplificazione dei principali geni di virulenza mediante PCR Real Time

Il metodo prevede l'identificazione tramite reazioni di PCR Real Time dei geni *stx1*, *stx2* ed *eae*. I geni *stx1* e *stx2* possono essere amplificati utilizzando un'unica coppia di oligonucleotidi in grado di appaiarsi alle sequenze di entrambi i tipi di geni, che vengono però discriminati attraverso l'utilizzo di due sonde molecolari diverse marcate con fluorofori che emettano fluorescenza a diverse lunghezze d'onda. È importante notare che, per via della elevata differenza della sequenza nucleotidica dei geni *stx2f*, tale sottotipo non è rilevato dal metodo indicato. Una reazione a parte deve essere allestita invece con due oligonucleotidi e una sonda specifici per il gene *eae*. Le reazioni per i geni *stx* ed *eae* possono essere eseguite contemporaneamente su un unico strumento per PCR Real Time, in quanto utilizzano lo stesso profilo termico di amplificazione, permettendo la rilevazione diretta dei tre principali geni di virulenza degli STEC in tempi molto brevi, pari a circa un'ora e trenta minuti. Il DNA stampo deve essere preparato partendo da un millilitro di una coltura pura del ceppo batterico da analizzare, attraverso un metodo di purificazione idoneo all'utilizzo in PCR Real Time, che può essere un kit basato su biglie ricoperte con tecnologia chelex, come quelli spesso usati per estrarre il DNA direttamente da campioni clinici o di alimenti, oppure un kit basato sull'utilizzo di colonnine cromatografiche. Quando si vogliono testare colture disponibili su terreno solido, una colonia può essere inoculata e posta in terreno liquido ricco (ad es. TSB) nell'incubatore a 37 °C per una notte, oppure dispersa in 1 ml di acqua sterile prima di procedere con l'estrazione.
Il testo completo del metodo è disponibile a questo link diretto:
https://www.iss.it/documents/20126/0/EURL_VTEC_Method_02_Rev_1.pdf/b5173cbd-5789-c729-b0c3-81da039e88c7?t=1644309049719

## Identificazione di ceppi di *E. coli* che producono il sottotipo Stx2f di Shiga-tossina mediante PCR Real Time

Il sottotipo Stx2f di Shiga-tossina è associato raramente a casi di malattia grave nell'uomo. Tuttavia, negli ultimi dieci anni sono stati riportati episodi sporadici di SEU causati da ceppi STEC che producevano tale sottotipo Stx. Per questo motivo è importante avere a disposizione un metodo in grado di rilevare la presenza dei geni *stx2f* in ceppi di *E. coli*. Il sottotipo *stx2f* è caratterizzato da una sequenza nucleotidica particolarmente diversa rispetto ai geni degli altri sottotipi *stx2*. Per questo motivo sono stati disegnati due oligonucleotidi e una sonda molecolare specifici per *stx2f* da utilizzare in reazioni distinte di PCR Real Time. Per il resto il metodo consiste esattamente degli stessi passaggi del metodo appena illustrato, dedicato alla identificazione e caratterizzazione di ceppi STEC mediante PCR Real Time, e utilizza lo stesso profilo termico di amplificazione. Ciò consente, qualora si volesse, di eseguire le reazioni per *stx1*, *stx2* e il sottotipo *stx2f* ed il gene *eae* in un unico ciclo di amplificazione.
Il testo completo del metodo è disponibile a questo link diretto:
https://www.iss.it/documents/20126/0/EURL_VTEC_Method_10_Rev+0.pdf/a4eb6e2b-fd13-c52a-112a-255c005b4872?t=1644309297880

## Identificazione dei sierogruppi STEC principalmente associati alle infezioni umane attraverso amplificazione dei geni associati agli antigeni O in PCR Real Time

Il metodo prevede l'identificazione dei geni associati agli antigeni O più frequentemente associati a malattia grave nell'uomo attraverso reazioni di PCR Real Time. In particolare i sierogruppi compresi nel campo di applicazione del metodo sono i seguenti: O26, O45, O55, O80, O91, O103, O104, O111, O113, O121, O128, O145, O146 e O157. I geni target del metodo sono diverse varianti alleliche dei geni *wzx* e *wzy*, associate ai diversi sierogruppi, il gene *wbdl* nel caso del sierogruppo O111 e il gene *rfbE* nel caso del sierogruppo O157. Il metodo consiste degli stessi passaggi dei metodi precedenti e tutte le reazioni devono essere eseguite con lo stesso profilo termico in uso per i due metodi appena illustrati (dedicati ai geni principali di virulenza e al sottotipo

*stx2f*), con l'eccezione delle reazioni dedicate ai sierogruppi O80 e O103, che vengono eseguiti con una temperatura più bassa per la fase di appaiamento degli oligonucleotidi.
Il testo completo del metodo è disponibile a questo link diretto:
https://www.iss.it/documents/20126/0/EURL_VTEC_Method_11_Rev_1.pdf/36a945da-ce05-1cab-df19-637f8169be3d?t=1644309311442

## La sorveglianza genomica di ceppi STEC

Lo sviluppo delle tecnologie di sequenziamento ad alta intensità, denominate nel complesso Next Generation Sequencing (NGS), ha reso possibile il sequenziamento del genoma completo di ceppi batterici in tempi molto rapidi (circa due giorni da quando si ha a disposizione il DNA purificato da analizzare) e a costi contenuti, in alcuni casi sono anche inferiori a 100 euro per campione. Il sequenziamento del genoma completo, in questo caso di batteri, è definito Whole Genome Sequencing (WGS). La possibilità di ottenere con facilità le sequenze WGS di batteri patogeni offre l'opportunità di caratterizzare in modo completo e fine i ceppi batterici di interesse, utilizzando un'unica metodica. A partire dai dati WGS, infatti, è possibile applicare strumenti bioinformatici in grado di estrarre informazioni relative alla presenza di geni di virulenza nel genoma del ceppo batterico analizzato, geni associati a marcatori fenotipici, come per esempio i sierogruppi, o geni di antibiotico-resistenza, o anche eseguire tipizzazioni fini volte a identificare correlazioni tra più ceppi, molto utili per indagini epidemiologiche.
L'analisi di sequenze WGS può essere eseguita attraverso numerosi strumenti bioinformatici ad oggi disponibili. Alcuni strumenti analitici sono disponibili gratuitamente, ma richiedono la capacità di saper operare con il computer attraverso comandi testuali e senza utilizzare una semplice interfaccia grafica. Altri strumenti più semplici utilizzano invece interfacce grafiche, ma in qualche caso sono software a pagamento che richiedono, oltre all'acquisto del software stesso, anche l'utilizzo di un computer ad elevate prestazioni analitiche e con ampio spazio di memoria per il salvataggio dei file.

Esistono tuttavia soluzioni diverse, che sfruttano la potenza di calcolo di potenti server accessibili tramite siti internet dedicati che possono essere utilizzati gratuitamente e attraverso semplici interfacce grafiche. Un esempio di tali siti internet è rappresentato dalla piattaforma ARIES (Advanced Research Infrastructure for Experimentation in genomicS) (4), che sfrutta il sistema Galaxy di interfaccia web per l'analisi di dati ad alta intensità, sviluppato dalla Penn State University, USA (5). La piattaforma ARIES è stata sviluppata dal Laboratorio Nazionale ed Europeo di Riferimento per *E. coli* presso l'Istituto Superiore di Sanità, aggiungendo ai pacchetti disponibili nella versione di base degli strumenti dedicati all'analisi di sequenze WGS di *E. coli*, ma anche di altri microrganismi patogeni e per l'analisi di sequenze di campioni di metagenomica. ARIES sfrutta la potenza analitica di server situati presso l'ISS, rendendo l'utilizzo della piattaforma sicuro a livello nazionale in termini di sicurezza dei dati trasferiti. Qualsiasi utente interessato può creare un proprio account gratuitamente sulla piattaforma ARIES (https://w3.iss.it/site/aries/) e usufruire degli strumenti integrati per eseguire analisi di sequenze WGS a propria disposizione.

Per servire esigenze di sorveglianza nazionale, negli ultimi anni la piattaforma ARIES è stata collegata ad una istanza della piattaforma IRIDA (6) ottimale per la raccolta di dati descrittivi epidemiologici dei campioni da cui sono stati isolati i ceppi batterici analizzati. La piattaforma di sorveglianza genomica che ne è derivata è stata denominata appunto IRIDA-ARIES (https://irida.iss.it). Anche questa, come ARIES, opera sui server dell'ISS ed è stata sviluppata presso il Dipartimento di Sicurezza Alimentare, Nutrizione e Sanità Pubblica Veterinaria a supporto della sorveglianza della listeriosi e delle infezioni da STEC. L'analisi avviene in modo del tutto automatizzato e fornisce come risultato una caratterizzazione completa del ceppo. Nel caso di sequenze di ceppi STEC, vengono identificati il sierogruppo, i geni di virulenza ed eventuali geni di resistenza a sostanze antimicrobiche attraverso confronti delle sequenze con banche dati di sequenze di riferimento precompilate. Inoltre i ceppi vengono tipizzati a livello di Multi Locus Sequence Typing (MLST) e di core genome

MLST (cgMLST) per permettere la tipizzazione fine e il confronto tra sequenze di più ceppi al fine di identificare eventuali correlazioni.
La piattaforma è costruita per permettere una identificazione rapida ed automatica di eventuali correlazioni genomiche tra ceppi, consentendo di identificare e contrastare prontamente eventuali focolai di infezione a livello nazionale. La piattaforma è stata studiata per proteggere la privacy dei dati, rendendo visibili le informazioni sensibili, come ad esempio la data di isolamento del ceppo e la residenza del paziente, solamente agli utenti della Regione di competenza e ai gestori della piattaforma presso l'ISS. D'altra parte la piattaforma consente a tutti gli utenti delle diverse Regioni di ricevere una notifica automatica qualora la sequenza di un ceppo di propria competenza avesse un'alta correlazione genomica con quella di un altro ceppo presente nella banca dati. Tale organizzazione permette un rapido scambio di informazioni e una tempestiva risposta in caso di focolaio epidemico.
I passaggi principali dell'analisi di sequenze WGS di ceppi STEC sono illustrati brevemente in appendice a questo capitolo. Questi possono essere eseguiti manualmente attraverso strumenti di propria scelta, come ad esempio quelli disponibili sulla piattaforma ARIES, o automaticamente, caricando i dati di sequenziamento e i dati descrittivi del campione sulla piattaforma IRIDA-ARIES. I passaggi preliminari di analisi della qualità, pulizia, assemblaggio e verifica della qualità dell'assemblaggio possono essere utilizzati anche per analizzare sequenze ottenute da ceppi di altre specie, mentre i passaggi successivi sono dedicati alla caratterizzazione di ceppi STEC.

## Studio della correlazione tra ceppi STEC attraverso confronti genomici

Gli approcci di tipizzazione basati sul sequenziamento dell'intero genoma consentono il confronto altamente discriminatorio dei genomi batterici. Tra questi, le analisi filogenetiche vengono utilizzate per diverse applicazioni, tra cui l'identificazione di link epidemiologici in caso di focolai, l'identificazione di veicoli di infezione e la sorveglianza di agenti patogeni batterici.

Esistono due approcci principali di tipizzazione basati sul confronto delle sequenze genomiche complete: l'analisi dei polimorfismi a singolo nucleotide (Single Nucleotide Polymorphisms, SNPs) e il confronto "gene-by-gene", che può essere applicato al genoma completo (wgMLST, whole-genome MultiLocus Sequecene Typing) o solamente ai geni *core*, cioè conservati nella specie (cgMLST, core-genome MultiLocus Sequecene Typing).
In particolare, tra gli approcci "gene-by-gene", l'analisi cgMLST è la più utilizzata per la sorveglianza genomica delle infezioni batteriche, in particolare per quelle causate da STEC. Si tratta di una sorta di "estensione" dell'analisi classica di Multi-Locus Sequence Typing (MLST). Infatti, così come l'analisi MLST convenzionale si basa sull'identificazione della combinazione di alleli di un pannello di sette geni "housekeeping", l'analisi cgMLST si basa sull'identificazione della combinazione di alleli di uno schema (numero) più esteso di geni, rendendo l'analisi maggiormente discriminante. Questi geni sono i cosiddetti geni "core" ovvero geni normalmente presenti nella maggior parte (almeno il 95%) dei ceppi appartenenti allo stesso genere, in questo caso il genere *Escherichia*.
Per la specie *E. coli* sono comunemente utilizzati due diversi schemi: lo schema sviluppato da EnteroBase che comprende 2513 *loci* genici e la versione di questo stesso database rivisitata e perfezionata nell'ambito del progetto INNUENDO (7), finanziato dalla European Food Safety Authority (EFSA), che comprende 2360 *loci*. Ad ogni allele di ogni gene presente nello schema viene assegnato un numero identificativo e quindi per ogni genoma batterico analizzato viene estrapolata una stringa di caratteri numerici identificativi degli alleli identificati per ogni gene analizzato. È possibile a questo punto confrontare le stringhe alleliche ottenute per ceppi diversi e contare in modo automatizzato il numero di alleli di differenza, definito distanza allelica. Quando le analisi comparative sono eseguite su più campioni, le distanze alleliche tra tutti i genomi studiati vengono riportate in una cosiddetta matrice di distanza, che riporta le distanze alleliche osservate mediante confronto incrociato tra ogni coppia di campioni. Questo permette di valutare il grado di correlazione

tra i ceppi. Distanza alleliche basse indicano un'elevata similarità dei ceppi cioè una probabile correlazione.

A causa dell'elevata variabilità riscontrata tra popolazioni diverse di ceppi STEC, non è stata ancora stabilita un esatto numero di differenze alleliche da utilizzare come soglia per valutare l'esistenza di una effettiva correlazione tra i ceppi. Nonostante questo, per un'analisi preliminare in caso di indagine epidemiologica può essere usata una soglia di 15 alleli di differenza, per identificare ceppi probabilmente correlati, da investigare meglio tenendo conto dei dati epidemiologici descrittivi dei campioni (come ad esempio la data di isolamento e origine geografica del campione analizzato).

Le stringhe alleliche ottenute per ogni ceppo dall'analisi cgMLST possono essere utilizzate per analisi di *clustering* che permettono di visualizzare le relazioni tra i ceppi batterici analizzati sotto forma di alberi filogenetici. In particolare, possono essere prodotti alberi filogenetici con la tecnica dei "Minimum Spanning Tree" (MST), in cui i diversi profili allelici identificati vengono rappresentati come cerchi collegati tra loro da rami la cui lunghezza rispecchia il numero di differenze alleliche identificate. I cerchi hanno invece un'area di dimensione proporzionale al numero di ceppi che condividono la stessa stringa allelica (Figura 1). Queste rappresentazioni sono molto utili per descrivere la struttura genetica delle popolazioni batteriche e illustrare la correlazione tra diverse popolazioni di ceppi STEC.

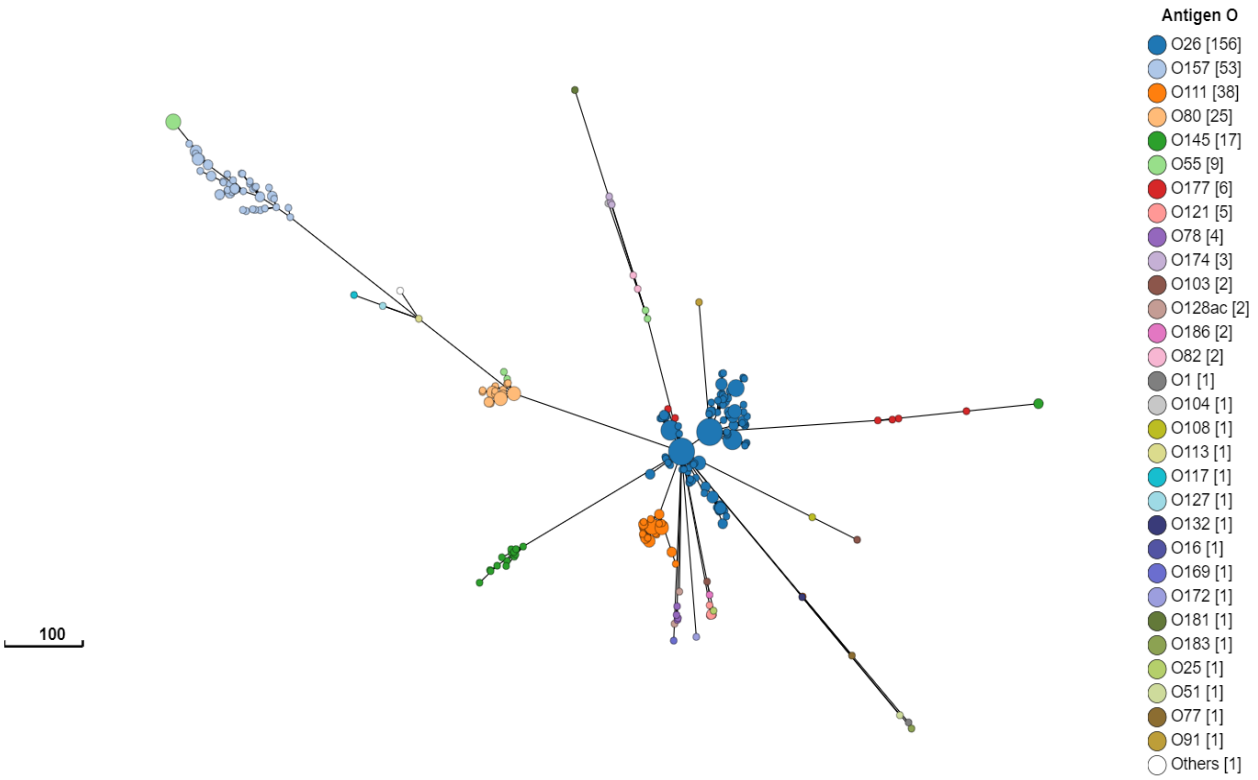


**Figura 1.** Esempio di albero filogenetico costruito con la tecnica del "Minimum Spanning Tree" sulla base di un confronto genomico tra ceppi STEC appartenenti a diversi sierogruppi, eseguito tramite cgMLST.

La piattaforma di sorveglianza genomica IRIDA-ARIES esegue l'analisi cgMLST, in aggiunta ad analisi di caratterizzazione genomica, in modo automatico a seguito dell'introduzione di una nuova sequenza genomica nel sistema. Lo schema utilizzato da IRIDA-ARIES è quello sviluppato dal progetto INNUENDO. Come risultato, la piattaforma IRIDA-ARIES fornisce automaticamente non solo un profilo completo di caratterizzazione molecolare dei ceppi analizzati, ma anche l'analisi filogenetica eseguita attraverso il cgMLST in confronto a tutte le sequenze dei ceppi inseriti nella banca dati. Questo consente di ottenere in automatico informazioni relative ad una eventuale correlazione con altri ceppi analizzati su scala nazionale, facilitando l'indagine dei focolai epidemici.

## Esempio di analisi di cluster finalizzato all'indagine di un focolaio di infezione da STEC

L'analisi genomica eseguita tramite cgMLST permette, come detto sopra, di identificare correlazioni tra ceppi diversi, svolgendo quindi un ruolo molto importante nella sorveglianza genomica. Può essere applicata a ceppi isolati da diversi casi di infezione, allo scopo di identificare eventuali *cluster* genomici suggestivi di possibili focolai, ma anche a ceppi di origine animale, alimentare o ambientale per un confronto con i ceppi umani o per studiare la dinamica di contaminazione da STEC di una filiera produttiva alimentare. La piattaforma IRIDA-ARIES, ad esempio, permette di analizzare in modo congiunto le sequenze di ceppi isolati da casi di infezione da STEC su scala nazionale e ceppi isolati da sorgenti animali, alimentare o ambientale, parte delle collezioni del Laboratorio Nazionale di Riferimento per *Escherichia coli*, rappresentando quindi un esempio pratico dell'approccio "One-Health". Eseguire analisi di *cluster* tra ceppi di origine diversa può consentire, quindi, di identificare le catene di trasmissione dell'infezione e di collegare tra loro casi che possono verificarsi anche a grandi distanze (per esempio ad un focolaio epidemico transregionale o transfrontaliero, nel caso di database sovranazionali). Presentiamo di seguito un esempio di analisi di *cluster* realizzato solo a scopo didattico. Nelle figure 2 e 3 è mostrato un Minimum Spanning Tree, in cui ogni cerchio corrisponde ad un ceppo STEC e i rami hanno una lunghezza proporzionale alla distanza allelica rilevata tra i ceppi che collegano.

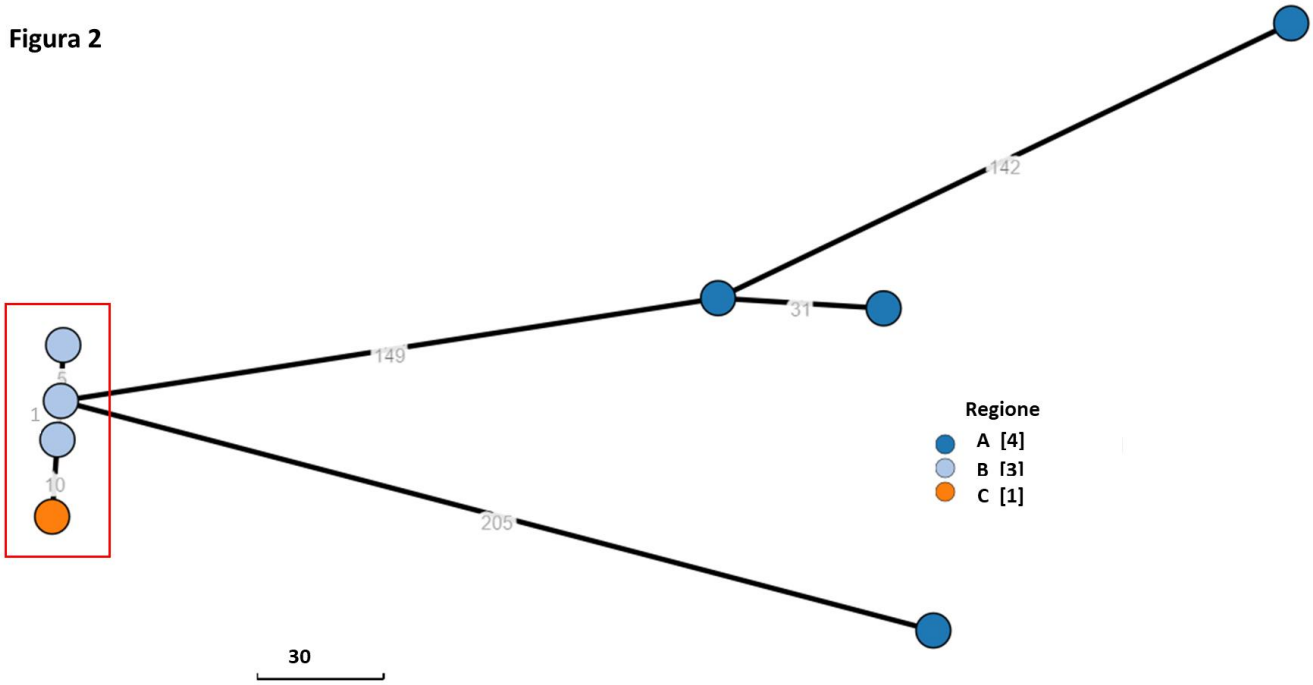


**Figura 2.** Analisi di *cluster* mediante cgMLST eseguita su otto ceppi STEC. Come indicato in legenda, i colori dei cerchi rappresentano la Regione di provenienza dei casi da cui sono stati isolati i ceppi. I ceppi che formano un *cluster* con un numero minore o uguale a 15 AD (differenze alleliche) sono evidenziati con un rettangolo.

La Figura 2 mostra un *cluster* di 4 ceppi STEC provenienti da due Regioni diverse (Regioni B e C), identificato a seguito di analisi cgMLST eseguita sulle sequenze genomiche di 8 ceppi STEC isolati in 3 Regioni diverse. I ceppi appartenenti al *cluster* differiscono per un numero di differenze alleliche compreso tra 0 e 10.

La Figura 3 mostra lo stesso Minimum Spanning Tree utilizzando un'annotazione di colore basata invece sulla sorgente di isolamento. Questo permette di identificare rapidamente che uno dei ceppi parte del *cluster* è di origine alimentare, mentre gli altri tre sono di origine umana.

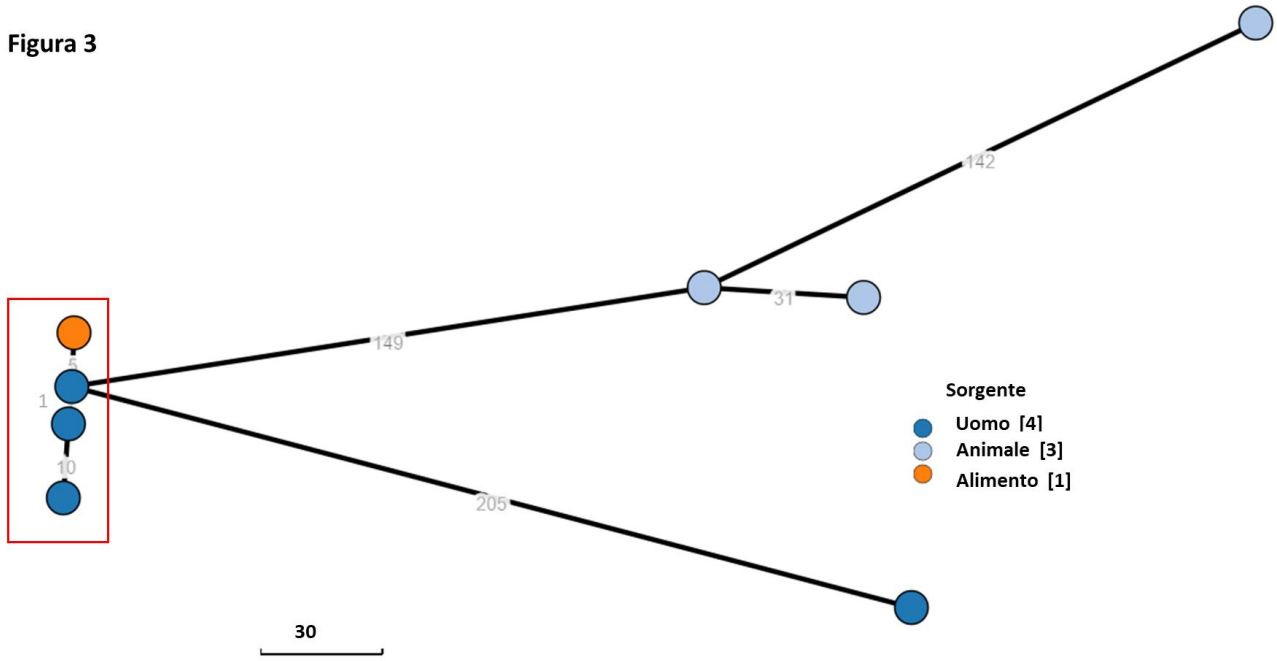


**Figura 3.** Analisi di *cluster* mediante cgMLST eseguita su otto ceppi STEC. Come indicato in legenda, i colori dei cerchi rappresentano la sorgente di isolamento. I ceppi che formano un *cluster* con un numero minore o uguale a 15 AD (differenze alleliche) sono evidenziati con un rettangolo.

In questo esempio è stato possibile identificare un *cluster* di casi sovraregionale, comprendente cioè pazienti di Regioni diverse (B e C). Inoltre è stato possibile identificare come parte del *cluster* anche un ceppo di origine alimentare. Questo tipo di analisi, congiuntamente alle analisi epidemiologiche, può consentire di ipotizzare o identificare le sorgenti di infezione, costituendo uno strumento fondamentale per il contrasto alla diffusione delle malattie a diffusione alimentare, in questo caso dovute ad infezioni da ceppi STEC.

# Appendice

## Cenni sui passaggi principali delle analisi genomiche

Analisi della qualità: questo passaggio ha lo scopo di eseguire un'indagine preliminare della qualità generale della sequenza ottenuta. Lo strumento

generalmente utilizzato a questo scopo è *FastQC*, disponibile anche su ARIES.
Pulizia (o *trimming*): questo passaggio mira a rimuovere le sequenze degli adattatori utilizzati per i passaggi di amplificazione necessari per il processo di sequenziamento NGS e a eliminare le sequenze di bassa qualità. Lo strumento più utilizzato a questo scopo è *Trimmomatic*, ma è possibile utilizzare, tra gli altri, anche lo strumento denominato *FastQ Positional and Quality Trimming*. Entrambi sono disponibili su ARIES.
Assemblaggio in contigs: questo passaggio prevede l'identificazione automatica di regioni di sovrapposizione tra le brevi letture di sequenza prodotte con la tecnologia NGS allo scopo di ridurre la molteplicità producendo sequenze più lunghe (*contigs*) rappresentative di regioni genomiche contigue. Lo strumento *SPAdes* è tra i più utilizzati a questo scopo ed è disponibile su ARIES.
Analisi della qualità dei contigs: questo passaggio è necessario per controllare che il processo di assemblaggio abbia prodotto un numero di contigs adeguato al genoma analizzato, che la lunghezza dei contigs sia sufficientemente elevata e che la lunghezza totale della sequenza assemblata tra tutti i contigs sia compatibile con la lunghezza attesa del genoma (circa 5 Mbp per gli STEC). Lo strumento più utilizzato per quest'analisi è *Quast*, disponibile su ARIES.
Identificazione del sierotipo: consente di identificare e tipizzare i geni associati all'antigene O e all'antigene H presenti nella sequenza WGS del ceppo analizzato, attraverso il confronto con una banca dati di sequenze di riferimento precompilata. Il risultato è l'identificazione del sierotipo del ceppo analizzato, quale combinazione degli antigeni O e H identificati, espresso come O:H (es. O157:H7). Lo strumento disponibile su ARIES a questo scopo è denominato *E. coli Serotyper*.
Determinazione del virulotipo: questo passaggio consente l'identificazione di geni di virulenza nel genoma del ceppo analizzato, attraverso il confronto con una banca dati di sequenze di riferimento precompilata. Tale banca dati comprende non solo i principali geni di virulenza degli STEC, e cioè *stx1*, *stx2* ed *eae*, ma anche numerosi altri geni accessori noti per svolgere un ruolo nel processo patogenetico degli STEC e che comprendono altre tossine e fattori di

colonizzazione, ma anche immunomodulatori e geni in grado di promuovere la sopravvivenza del batterio durante l'infezione. Lo strumento disponibile su ARIES a questo scopo è *E. coli Virulotyper*.

Identificazione del sottotipo dei geni *stx*: questa analisi permette la sottotipizzazione dei geni *stx*, se presenti nella sequenza genomica in studio, attraverso il confronto automatico con una banca dati precompilata contenente sequenze di riferimento di tutti i sottotipi *stx*. Lo strumento disponibile su ARIES a tale scopo è *E. coli Shiga toxin typer*.

Multi Locus Sequence Typing (MLST): questo tipo di analisi permette l'assegnazione di un codice, definito Sequence Type (ST) in base alla combinazione allelica di sette geni *housekeeping,* sempre presenti nei genomi di ceppi di *E. coli* perché convolti nel mantenimento delle funzioni di base del batterio. Poiché tali geni sono sotto forte pressione selettiva, questo approccio è molto robusto per identificare ceppi appartenenti a ST diversi e quindi evolutivamente distanti tra loro. Su ARIES sono disponibili diversi strumenti per MLST, tra cui *MLST* (strumento sviluppato da T. Seeman, da preferire agli altri per velocità di esecuzione), *SRST2* e *MentaLIST*.

Core Genome MLST (cgMLST): analisi di un pannello di geni conservati nella specie, definiti geni *core*, volta all'identificazione degli alleli di ciascun gene per confronto con una banca dati di sequenze alleliche di riferimento. Questo approccio permette di costruire alberi filogenetici e di contare il numero di differenze alleliche tra coppie di ceppi, allo scopo di stabilirne il grado di correlazione. L'analisi cgMLST può essere eseguita con il tool *chewBBACA* disponibile sul server pubblico ARIES. Il risultato prodotto che contiene le stringhe alleliche dei ceppi analizzati può essere utilizzato per calcolare automaticamente le distanze alleliche tra i ceppi attraverso lo strumento *chewTree*, anch'esso disponibile su ARIES.

## Bibliografia

1. McDaniel TK, Jarvis KG, Donnenberg MS, Kaper, JB. 1995. A genetic locus of enterocyte effacement conserved among diverse enterobacterial pathogens. Proc. Natl. Acad. Sci. USA 92, 1664–1668

2. EFSA BIOHAZ Panel, Koutsoumanis K, Allende A, Alvarez-Ordo~nez 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 SilvaFelıcio MT, Messens W and Bolton D, 2020. Scientific Opinion on the pathogenicity assessment of Shigatoxin-producing *Escherichia coli* (STEC) and the public health risk posed by contamination of food withSTEC. EFSA Journal 2020;18(1):5967, 105 pp
3. World Health Organization & Food and Agriculture Organization of the United Nations. (2018). Shiga toxin-producing *Escherichia coli* (STEC) and food: attribution, characterization, and monitoring: report. World Health Organization. https://apps.who.int/iris/handle/10665/272871
4. Knijn A., Michelacci V, Orsini M, Morabito S. Advanced Research Infrastructure for Experimentation in genomicS (ARIES): a lustrum of Galaxy experience bioRxiv 2020.05.14.095901
5. Giardine B, Riemer C, Hardison RC, Burhans R, Elnitski L, Shah P, Zhang Y, Blankenberg D, Albert I, Taylor J, Miller W, Kent WJ, Nekrutenko A Galaxy: a platform for interactive large-scale genome analysis Genome Res. 2005 Oct;15(10):1451-5
6. Matthews TC, Bristow FR, Griffiths EJ, Petkau A, Adam J, Dooley D, Kruczkiewicz P, Curatcha J, Cabral J, Fornika D, Winsor GL, et al. 2018. The Integrated Rapid Infectious Disease Analysis (IRIDA) Platform. bioRxiv 381830
7. Llarena A. K., Ribeiro-Gonçalves B. F., Silva D. N., Halkilahti J., Machado M. P., Da Silva M. S., et al. (2018). INNUENDO: A Cross-Sectoral Platform for the Integration Ofgenomics in the Surveillance of Food-Borne Pathogens Vol. 15 (EFSA Supporting Publications, EXTERNAL SCIENTIFIC REPORT)