IZSAM G.CAPORALE
TERAMO

Listeria monocytogenes
Laboratorio Nazionale di Riferimento

Draft ISO/CD 23418:
WGS per la tipizzazione e la caratterizzazione del genoma di batteri di origine alimentare – Guida e requisiti generali

Cesare Cammà – Cristina Marfoglia

Workshop LNR per *Listeria monocytogenes*
04 luglio 2019 - Teramo

Premessa

WGS offre **vantaggi** significativi rispetto ai precedenti metodi di tipizzazione

Nei laboratori di **sanità pubblica** viene utilizzato per:

- studiare **focolai d'infezione**
- Identificare **caratteristiche** genetiche di **virulenza** e di potenziale **sopravvivenza** (es. AMR)

Nei laboratori delle **Industrie alimentari** viene utilizzato per:

- **caratterizzare** gli **isolati** batterici da ingredienti e superfici ambientali
- capire meglio la loro **origine** ed **ecologia** (capacità di persistenza e/o resistenza ai disinfettanti)
- aggiornare le procedure e **ridurre** i **rischi**.

Necessità di una Linea guida riconosciuta

- Necessità, da parte dei laboratori, di **sviluppare** i loro **protocolli** tecnici e i **sistemi di gestione di qualità**
- Necessità di **standard** durante il **confronto** dei **dati** generati dai **laboratori locali** con **database esterni** pubblici o privati
- Qualsiasi organizzazione che esegua WGS o analisi bioinformatiche sui dati generati deve **dimostrare** che viene raccolta e mantenuta una **documentazione adeguata** per verificare che i **metodi** siano **idonei** e che vengano **controllati nel tempo**

English

International Organization for Standardization

When the world agrees

ISO/CD 23418

Microbiology of the Food Chain -- Genomic sequencing of foodborne microorganisms -- General requirements and guidance for bacterial genomes

General information

Status : Under development

Edition : 1

Technical Committee : ISO/TC 34/SC 9 Microbiology

ICS : 07.100.30 Food microbiology

Life cycle

A standard is reviewed every 5 years

00 | 10 | 20 | **30.20** Committee | 40 | 50 | 60 | 90 | 95

Revisions / Corrigenda

Now under development

ISO/CD 23418

WGS di batteri di origine alimentare

- Isolamento batterico
- Estrazione del DNA

Laboratorio di microbiologia

- Preparazione delle librerie genomiche
- NGS

Laboratorio di genomica

- Analisi bioinformatiche

Laboratorio di bioinformatica

I laboratori devono implementare **procedure operative standard** (SOP), tenere sotto **controllo** e **registrare** il **flusso di lavoro**, **gestire** i **reagenti** e gestire le **apparecchiature** e i **software**.

WGS: flusso di lavoro

IZSAM G.CAPORALE TERAMO

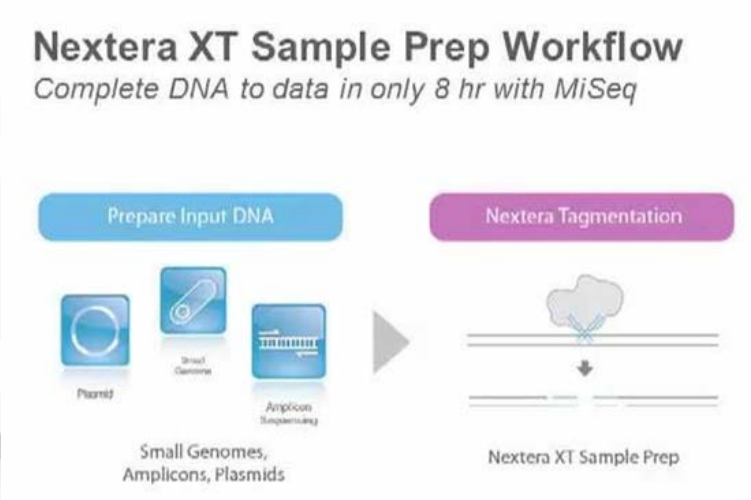
Listeria monocytogenes
Laboratorio Nazionale di Riferimento

Isolamento batterico

Estrazione DNA

Quantificazione DNA

Preparazione librerie genomiche



2 x 150 bp

NGS

Illumina NextSeq 500

Isolamento batterico

Qualsiasi materiale da sequenziare (isolato batterico o DNA genomico estratto) deve essere **manipolato** e **conservato** in modo che **si riduca al minimo** il rischio di **degradazione** del campione, **errata identificazione** e **contaminazione** incrociata.

Isolamento batterico

Le informazioni sugli isolati devono essere inserite nella LIMS e/o in fogli di lavoro

Gli isolati batterici devono essere conservati (es: -80°C in microbank)

Gli isolati batterici dovrebbero essere coltivati e stoccati **riducendo** al minimo il **potenziale** per **l'introduzione** di **cambiamenti genetici** (ad esempio la perdita di plasmidi o polimorfismi introdotti attraverso i passaggi)

- Non più di **5 passaggi** di singole colonie
- In caso di isolati **con plasmidi** che possono essere persi facilmente è consigliabile **collezionare 2/3 replicati**.
- Il laboratorio deve garantire la **purezza** dell'isolato e confermare le **specie**

Estrazione DNA

Estrazione DNA

Quantificazione DNA

La procedura di estrazione dovrebbe essere **appropriata** per gli organismi e la piattaforma di sequenziamento utilizzata.

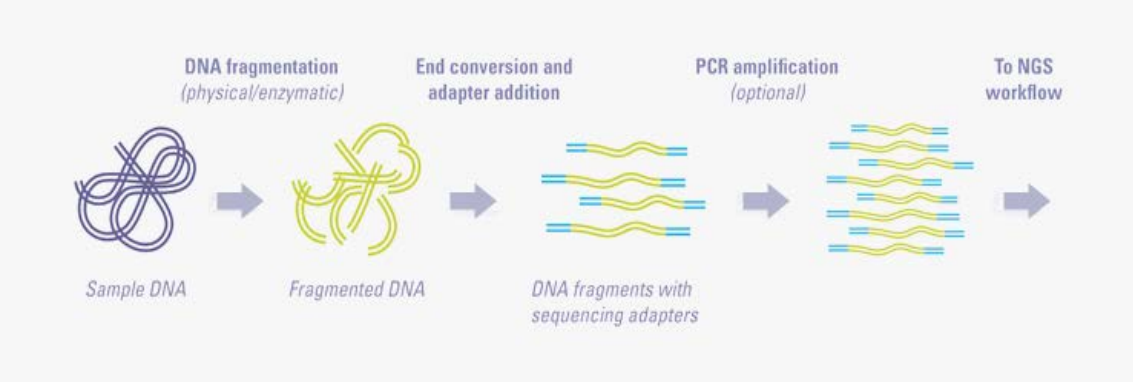
La **quantità** e la **qualità** del DNA dovrebbero essere valutate e documentate.

L'isolamento del DNA batterico è **influenzato** da una serie di fattori

- il tipo di cellula (Gram positivo o negativo)
- la fase di crescita
- il terreno di coltura.

- **Controllo negativo** per valutare la possibile **contaminazione incrociata**: acqua o brodo non inoculato. Da includere in ogni estrazione. Da sequenziare quando ritenuto necessario
- **Controllo positivo** per valutare **l'efficienza** del metodo di estrazione (es. ceppo di riferimento). Non è necessario che venga sequenziato

Preparazione delle librerie genomiche



Fasi

- Frammentazione del DNA
- Aggiunta di indici e adattatori
- Quantificazione, normalizzazione e controllo di qualità delle librerie

Il laboratorio dovrebbe seguire il **protocollo raccomandato** dal produttore. Tutte le **modifiche** devono essere pienamente **documentate** e **validate**.

Durante i protocolli per la preparazione delle librerie che prevedono **PCR**, i passaggi pre- e post-PCR devono essere eseguiti in **aree diverse** del laboratorio per evitare la contaminazione da **carryover**.

Preparazione delle librerie genomiche

- **Controllo negativo** per valutare la possibile **cross-contaminazione** durante la preparazione delle librerie (possibile ma deve essere al di sotto dei limiti stabiliti durante la validazione): **acqua**. Da includere quando ritenuto necessario.
- **Controllo positivo** per **valutare** la **qualità** nella preparazione delle librerie. DNA estratto da un ceppo ben caratterizzato. Utilizzare possibilmente sempre lo stesso. Non necessario in tutte le run.

Il laboratorio deve documentare

- il metodo di **tracciamento** del **campione** (codice a barre o equivalente) per consentire **l'associazione** dei dati di **sequenza** con i **metadati** corretti.
- la **piattaforma** di sequenziamento utilizzata
- **l'operatore** che ha preparato la libreria
- la **data** in cui è stata creata la libreria
- le informazioni sul **lotto** del **kit** utilizzato.

NGS

La strumentazione di sequenziamento deve essere sottoposta a **manutenzione**.

Il **basecalling** deve essere effettuato utilizzando un software specifico per lo strumento e la chimica di sequenziamento.

Le **metriche di qualità** possono essere definite a livello di run e a livello di campione

Metriche di qualità **a livello di run**: specifiche per la piattaforma

- cluster density
- numero di reads
- qualità media delle basi

Devono essere registrate e monitorate per ciascuna corsa di sequenziamento.

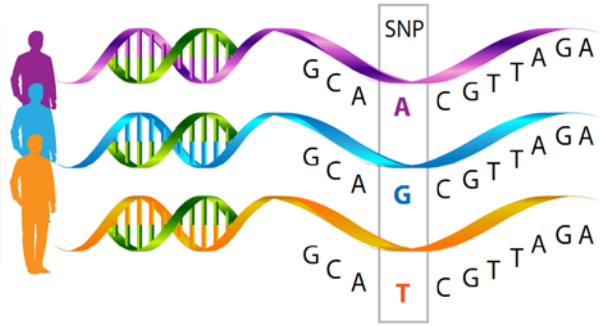
Metriche di qualità **per campione** :

- dimensione media delle reads
- distribuzione della lunghezza della sequenza
- numero di reads
- Coverage verticale e orizzontale

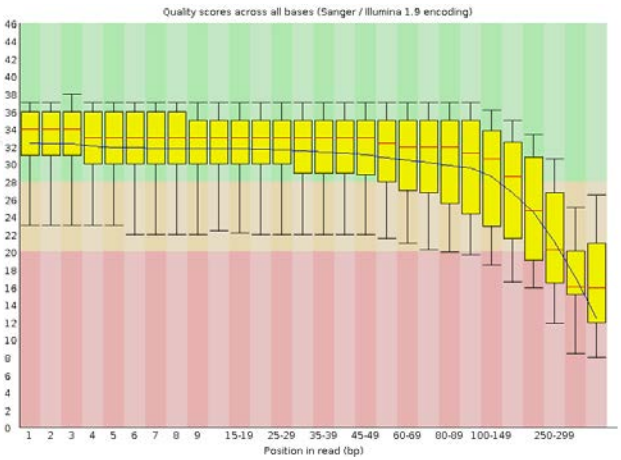
- **Controllo positivo** per **valutare** la **qualità** della run: **libreria** di controllo **PhiX**. Da includere in ogni run. Valutare il tasso di errore.
- **Controllo negativo** per **valutare** la **cross-contaminazione** tra una run e l'altra: indici usati nella run precedente (possibile ma contenuta). Se ritenuto necessario.

Analisi dati NGS

- Le **pipeline bioinformatiche** devono essere validate prima dell'analisi dei dati
- Gli sviluppatori di pipeline dovrebbero distribuire **set di dati - test**
- Gli utenti devono assicurarsi che le pipeline siano installate correttamente analizzando i set di dati del test e controllando che i risultati attesi siano generati
- I **metodi sviluppati internamente** dovrebbero essere pienamente validati
- Tutti i **passaggi** e le analisi dei dati devono essere registrati e documentati
- Gli **aggiornamenti** dei componenti software devono essere documentati e validati



Controllo qualità dati NGS



Mauve assembly metrics

Le metriche di qualità dovrebbero essere specifiche della piattaforma

- Punteggio medio di qualità e numero di basi superiori a una soglia di qualità specifica
- **Numero** di **reads** e punteggio medio di **Phred (Q)** per read
- Test di **contaminazione**
- Qualità degli **assemblaggi**:
  - La profondità di lettura deve essere sufficiente per garantire che le varianti possano essere rilevate in modo affidabile nell'assemblaggio.
  - **Numero** di **contigs**. Per i draft, potrebbe essere necessario rimuovere piccoli contigs o con basso coverage
  - **N50** e/o NG50 e **lunghezza** del contig più lungo
- La **lunghezza totale** di tutti i contigs dovrebbe approssimare la dimensione nota del genoma del microrganismo

Valori di qualità per short-read Illumina

Qualità dei dati di sequenza: i dati grezzi di sequenza devono avere una **qualità**, una **lunghezza media** delle reads e un **coverage** sufficiente per l'uso previsto. Strumento: **FastQC**. Tutte le sequenze devono essere identificate come **warn** o **pass**. Coverage minimo stimato: 20x a 60x.

Parametri di accettabilità della run: specifiche per la piattaforma e la chimica

- Reads passing filter
- % **Q30**
- Tasso di errore del **PhiX**
- Numero di reads del controllo **negativo**

Parametri di accettabilità del campione:

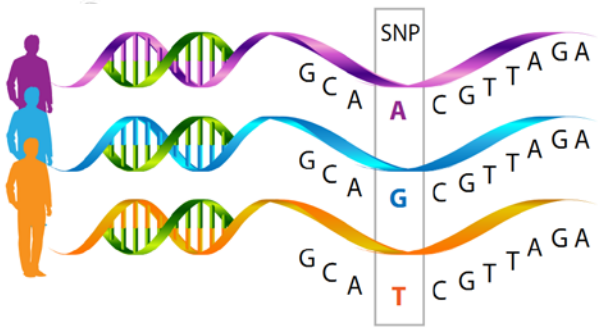
- Coverage stimato: **≥ 20X**
- Q score medio: **≥ 30**
- Contaminazione (verifica della specie attesa e assenza di specie/ceppi non previsti: **<5%** delle reads

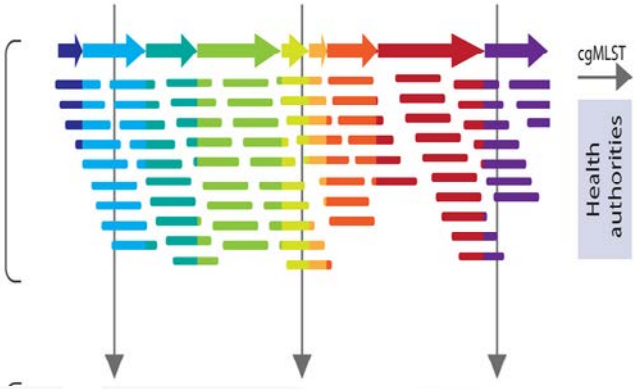
È possibile **scegliere** se considerare i parametri di accettabilità del **singolo campione** invece di considerare quelli della run completa

Informazioni sulla run: specifiche per la piattaforma e la chimica

- Cluster density
- Cluster passing filter
- Qualità media delle basi (Q score)
- Throughput

WGS clustering analysis





Analisi SNP

- Può essere utilizzato come **reference** un draft o una sequenza del genoma completa.
- Il genoma di riferimento dovrebbe essere geneticamente **simile** alle sequenze in analisi.

Analisi MLST (cgMLST e wgMLST)

Se le determinazioni degli alleli sono fatte usando **genomi assemblati**:

- Utilizzare lo stesso metodo di **annotazione** usato durante la generazione dei database
- Devono essere stabiliti gli **standard minimi** di qualità per gli assemblaggi, come la percentuale dei loci rilevati

Se le determinazioni degli alleli vengono eseguite utilizzando il **mapping** delle reads:

- Stabilire I criteri per l'identificazione di SNP e indel
- Stabilire **standard minimi** di qualità (percentuale di loci identificati)

Report e interpretazione dati

U.S. Food and Drug Administration
Protecting and Promoting Public Health
www.fda.gov

WGS USED FOR FIRST TIME AS PART OF REGULATORY ACTION

FDA shuts down Roos Foods cheese plant over listeria outbreak

WGS of *Listeria monocytogenes* strains isolated from Roos Foods cheese products was performed by the FDA and Virginia's Division of Consolidated Laboratory Services.

These strains were found to be highly related by WGS to the *Listeria* strains isolated from patients in this outbreak, adding further confidence that cheese products produced by Roos Foods were a likely source of the outbreak.

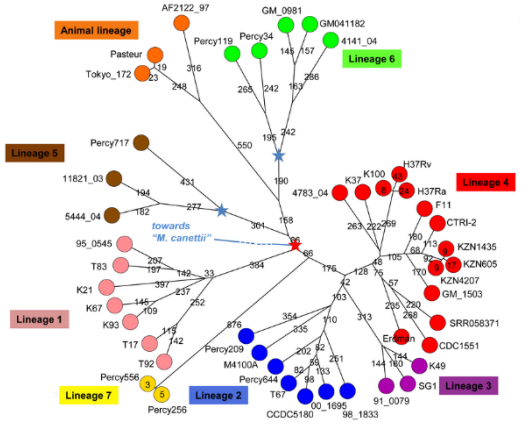
WGS provided genetic information that allowed investigators to rapidly identify differences among isolates.

Report dei risultati dell'analisi genomica

- **versione** della pipeline;
- identificazione dei **dati** di **input**;
- **genoma** di **riferimento** o **database MLST** utilizzato e **versione**
- **parametri** delle analisi (ad es. coverage minimo per chiamata allelica)

Interpretazione e conclusioni dei risultati delle analisi di confronto

- I risultati delle pipeline di bioinformatica dovrebbero essere **interpretati** nel contesto delle informazioni relative ai **metadati** (origine degli isolati e informazioni epidemiologiche)
- Le **soglie** stabilite per un caso non devono essere utilizzate per l'interpretazione di casi diversi (ad es. focolaio di breve durata con unica sorgente, patogeni persistenti o resistenti).



Validazione

- **La validazione del flusso di lavoro WGS può essere eseguita con approccio modulare, separatamente per i diversi componenti del processo di analisi.**
- **Fornirà la prova che il metodo è ripetibile, riproducibile e preciso.**

Validazione delle attività di laboratorio
La validazione delle attività di laboratorio può partire **dalla cultura batterica** fino al **sequenziamento del DNA** e **tutte le fasi** intermedie, a seconda del flusso di lavoro del laboratorio.

Validazione della pipeline bioinformatica
Le pipeline di bioinformatica possono essere valutate e/o validate utilizzando diversi tipi di set di dati. La validazione può includere **dati generati nel laboratorio** utilizzando uno specifico flusso di lavoro WGS insieme a set di **dati standard** e/o a set di **dati simulati**.

Validazione delle diverse fasi di analisi

ISOLAMENTO

Ripetibilità (accuratezza/precisione):

- includere diverse **subculture** dello **stesso giorno** fatte dallo **stesso operatore**

Riproducibilità (accuratezza/precisione):

- Includere diverse **subculture** di **diversi giorni** fatte da **diversi operatori**

Concordanza con altri metodi:

- Includere **ceppi correlati** e **non correlati** (ad es. focolaio e non-focolaio), o ceppi che non contengono il/i geni target

ESTRAZIONE DNA

Ripetibilità (accuratezza/precisione):

- Includere **diverse estrazioni** del DNA fatte dallo **stesso operatore** usando lo **stesso lotto** di reagenti e utilizzando la **stessa subcultura**

Riproducibilità (accuratezza/precisione):

- Includere **diverse estrazioni** del DNA fatte da **diversi operatori** e con **diversi lotti** di reagenti utilizzando **diverse subculture** di **diversi giorni**

Concordanza con altri metodi:

- Includere **DNA estratto** da **ceppi correlati e non correlati** (ad es. focolaio e non-focolaio), o ceppi che non contengono il/i geni target

Validazione delle diverse fasi di analisi

NGS

Ripetibilità (accuratezza/precisione):

- Includere nella **stessa run** librerie (ad es. in **triplicato**) prodotte dallo **stesso operatore** lo **stesso giorno** a partire dallo **stesso ceppo**

Riproducibilità (accuratezza/precisione):

- Includere in **strumenti diversi** e in **giorni diversi** librerie prodotte da **operatori diversi**

Concordanza con altri metodi:

- Includere **librerie** prodotte da **ceppi correlati** e **non correlati** (ad es. focolaio e non-focolaio), o ceppi che non contengono il/i geni target

Pipeline di Bioinformatica

Ripetibilità (accuratezza/precisione):

- Dimostrare **identici risultati** partendo dallo **stesso set di dati** almeno **due volte** sullo **stesso computer/server**, utilizzando la **stessa versione** del software con le **stesse opzioni/parametri**

Riproducibilità (accuratezza/precisione):

- Dimostrare **risultati comparabili** partendo dallo **stesso set di dati** almeno due volte su **diversi computer/servers**, utilizzando la **stessa versione** del software con le **stesse opzioni/parametri**

Concordanza con altri metodi:

- Dimostrare che i **risultati** sono **confrontabili** con **altre pipeline** per la **stessa applicazione** e specificare eventuali differenze note che potrebbero influire sul risultato

Validazione della pipeline bioinformatica

Set di dati di sequenziamento reali

- Tengono in considerazione i **batteri** e le **procedure** microbiologiche **utilizzate** in un particolare laboratorio.
- Vengono generati utilizzando uno **specifico flusso** di lavoro **end-to-end**.
- Derivano in genere da collezioni di **ceppi** di **riferimento** con caratteristiche note o da raccolte di **isolati associati** a un **focolaio** con informazioni epidemiologiche note

Set di dati standard

- **Casi** in cui l'origine, il fenotipo e/o la **relazione epidemiologica** degli isolati sono **noti** e i dati della sequenza sono stati resi **pubblici**.
- I set di dati standard possono essere utili per **confrontare** l'output di **diverse pipeline** bioinformatiche.

Set di dati simulati

- Dei **software** possono essere utilizzati per **generare dati** di reads di sequenze **sintetiche** da **dati** di sequenza di **genomi reali**.
- Possono testare una **gamma** più **ampia** di valori e errori dei parametri.
- Possono essere introdotte **differenze note**.
- A volte possono produrre **risultati non realistici** e le pipeline spesso hanno prestazioni migliori con dati simulati rispetto ai dataset reali

Validazione delle diverse fasi di analisi

Criteri di accettabilità

Ripetibilità (accuratezza/precisione):

- **L'interpretazione** dei risultati **non dovrebbe cambiare**, cioè non dovrebbero essere osservate differenze significative **durante la ripetizione del flusso di lavoro WGS** nello **stesso laboratorio**, con gli **stessi operatori** che usano lo **stesso strumento**.

Riproducibilità (accuratezza/precisione):

- **L'interpretazione** dei risultati **non dovrebbe cambiare**, cioè non dovrebbero essere ottenute differenze significative **durante la riproduzione del flusso di lavoro WGS** in **diversi laboratori**, con **operatori diversi** o strumenti diversi. Sono previste **lievi differenze** (non significative), causate da alcuni passi metodologici.

Concordanza con altri metodi:

- Il flusso di lavoro WGS deve essere in grado di arrivare alle **stesse conclusioni degli altri metodi di tipizzazione e/o caratterizzazione standard** (ad es. concordanza epidemiologica, capacità di differenziare ceppi non correlati e raggruppare isolati strettamente correlati)

Verifica

Verifica delle attività di laboratorio
Nel caso di **metodi riconosciuti/normalizzati** (ad es. metodi che sono stati pubblicati in riviste con peer review), la validazione non è richiesta ma solo la verifica.

Verifica della pipeline bioinformatica
Se vengono utilizzati **software commerciali** o **open source** che sono stati **validati** dai loro sviluppatori, i test di validazione sono **pubblicati** e i set di dati di validazione sono pubblicamente **disponibili**, potrebbe essere necessario solo (parzialmente) **ripetere** il **test** di validazione una volta che il software è stato installato.
I **dati di test** vengono utilizzati come **input** per una pipeline di bioinformatica e **l'output** viene **confrontato** con i risultati previsti.

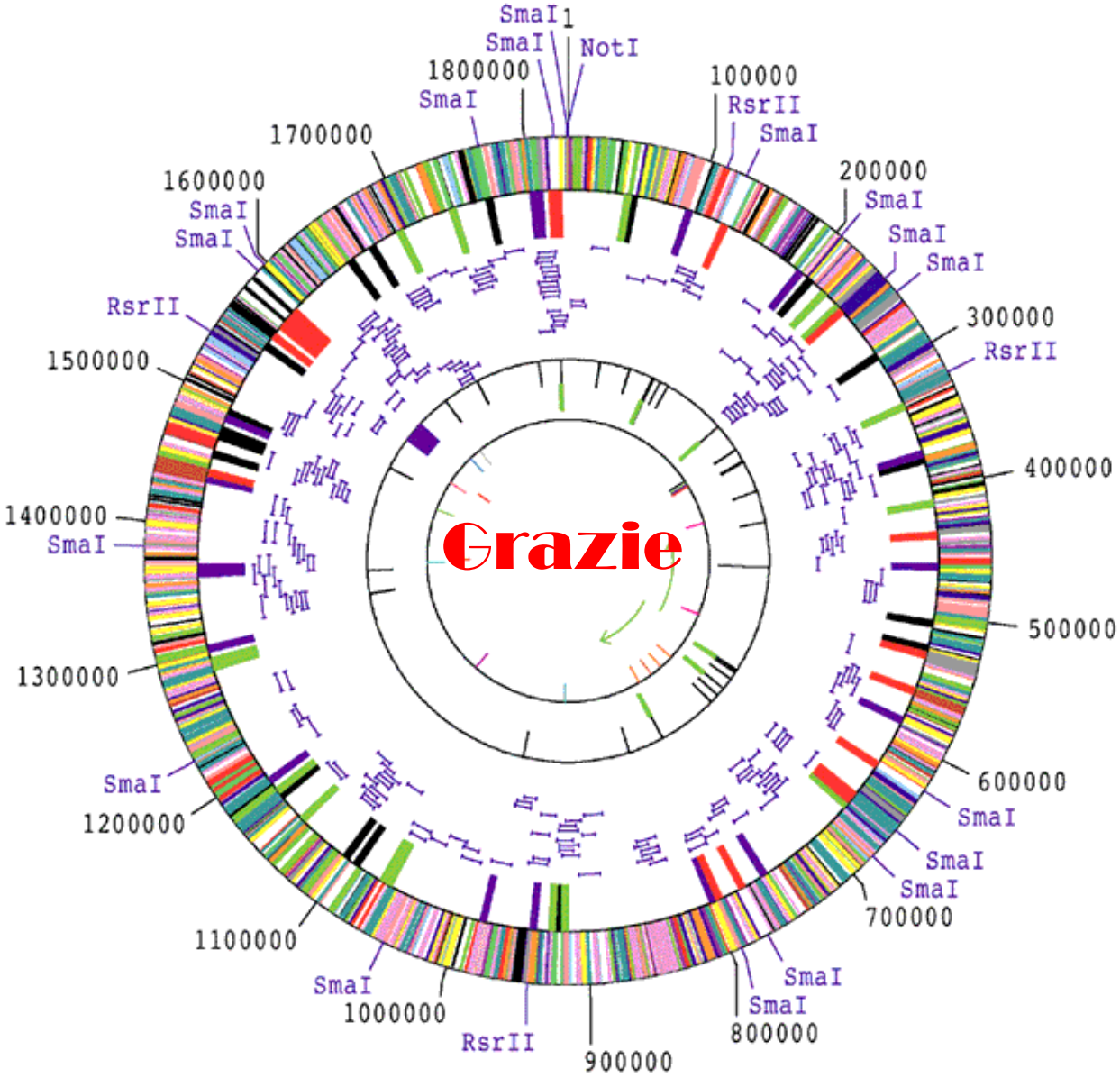
METADATI

I metadati associati alla sequenza genomica memorizzati in archivi privati devono includere informazioni minime che descrivono il **campione**, **l'isolato**, e la **sequenza**

- Nome del Lab. o del contatto che colleziona il campione prima dell'analisi
- Origine geografica del campione
- Tipo di matrice da cui è stato isolato
- Destinazione d'uso della matrice (consumo umano o animale)
- Processo di trasformazione dell'alimento (es. pastorizzazione
- Alimento naturale o trasformato
- Localizzazione ambientale
- Supporto di conservazione
- Metodo di conservazione

- Nome del Lab. o del contatto che ha isolato l'organismo da sequenziare
- Specie isolata
- Identificazione dell'isolato
- Sierotipo
- Terreno usato per l'isolamento
- Numero di passaggi effettuati
- Antibiogramma: metodo e risultati
- Fattori di virulenza: metodo e risultati

- Nome della Struttura o del contatto che esegue il sequenziamento
- Dati di tracciabilità della run di sequenziamento
- Metodo estrazione DNA
- Numero replicati sequenziati
- Metodo di preparazione Librerie
- Nome della Struttura o del contatto che esegue le analisi bioinformatiche
- Data produzione dati grezzi
- Data del filtraggio dati
- Assemblaggio
- Annotazione
- Metriche di qualità dell'assemblaggio


SmaI
1
SmaI
NotI
1800000
100000
SmaI
RsrII
SmaI
1700000
200000
1600000
SmaI
SmaI
SmaI
SmaI
RsrII
300000
RsrII
1500000
400000
1400000
SmaI
Grazie
500000
1300000
600000
SmaI
SmaI
1200000
SmaI
SmaI
700000
1100000
SmaI
SmaI
1000000
SmaI
800000
RsrII
900000