ESERCITAZIONE 2 – PROGRAMMARE UN PROGETTO DI SEQUENZIAMENTO GENOMICO

Questa esercitazione sarà organizzata per punti e ci consentirà di affrontare il problema relativo alla programmazione di un progetto di sequenziamento genomico che riguarda una specie animale non modello, per la quale non sono disponibili dati molecolari pregressi.

Il consiglio è quello di provare a lavorare autonomamente punto per punto, facendo riferimento a quanto visto nelle lezioni teoriche, eventualmente riprendendo le slides in questione. Durante la lezione in diretta avrete alcuni minuti di tempo per provare a trovare una soluzione in autonomia, ai quali seguirà un confronto con il docente, per verificare la correttezza di quanto fatto. Naturalmente la stessa procedura potrà essere seguita in fase di studio in un secondo momento, seguendo la video lezione caricata su Stream.

L'intenzione è quella di:

1. Ripassare quanto visto nelle lezioni teoriche, permettendo di fissare meglio alcuni concetti in vista dell'esame
2. Mettervi in prima persona davanti a problemi molto pratici che chi si occupa di genomica funzionale si trova ad affrontare quasi ogni giorno

Troverete alcuni files accessori che saranno di grande utilità per lo svolgimento di alcuni dei task di questa esercitazione nella cartella apposita, chiamata "Esercitazione 2", su Moodle.

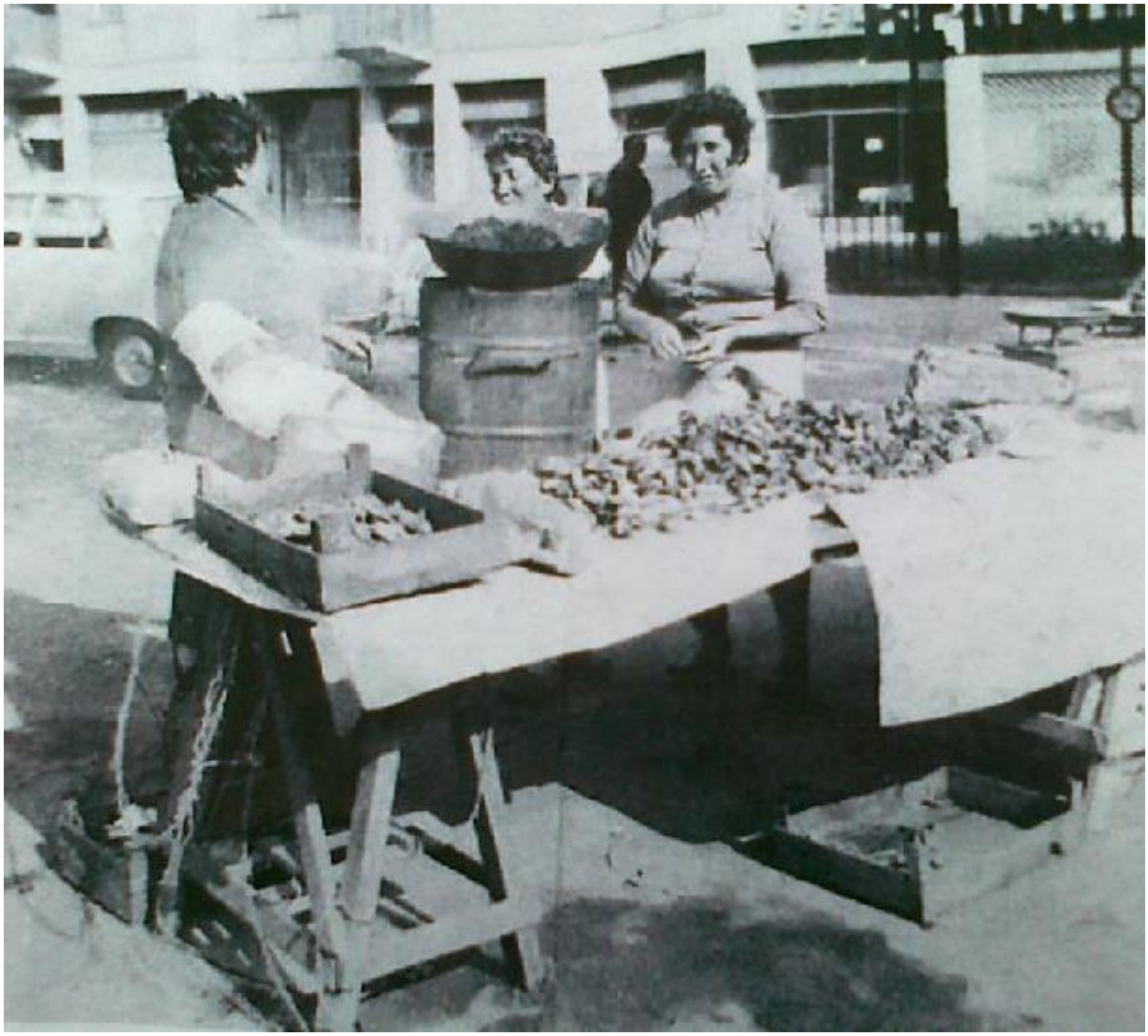
IL PROGETTO IN ESAME

Il mollusco bivalve *Arca noae*, detto comunemente "mussolo", è un organismo che un tempo si consumava largamente in tutte le località costiere del Mar Adriatico. Rappresentava anche un classico "cibo povero" da osteria, nonché da strada, in quanto veniva fornito cotto al vapore su banchetti agli angoli delle strade del centro città, spesso e volentieri da signore che erano colloquialmente chiamate "mussolere".

Tuttavia, a partire dai primi anni del '900 le popolazioni locali di questo bivalve, un tempo molto abbondanti, sono andate incontro ad un importante declino, legato a cause ignote ma probabilmente dovuto all'arrivo di un patogeno che ha decimato i banchi che si trovavano lungo tutte le scogliere della costiera triestina.

Questo, unito ad un probabile sovrasfruttamento delle risorse disponibili, ha determinato un progressivo abbandono di questa attività e ben presto i banchi delle mussolere hanno iniziato a sparire dalle strade. Oggi

la pesca di *A. noae* è strettamente regolamentata e non riguarda grossi numeri, anche perché la loro raccolta avviene a mano da parte di sub. Tuttavia ci sarebbe potenzialmente la possibilità di riscoprire questo antico piatto di mare, magari con il coinvolgimento di aziende già coinvolte nella mitilocoltura, che abbiano quindi il know-how per garantire l'allevamento di questi organisimi senza impattare in modo significativo l'integrità delle popolazioni locali.



Il nostro obiettivo in questo caso potrebbe essere quello di sequenziare il genoma di questa specie, puntando ad ottenere un assemblaggio genomico di alta qualità, in modo da generare una risorsa che possa fungere da riferimento per ulteriori studi. Gli obiettivi principali di questo progetto saranno dunque:

-ottenere un assemblaggio de novo "chromosome-scale" del genoma di Arca noae

-garantire una buona annotazione dei geni codificanti

-studiare la struttura delle popolazioni naturali esistenti analizzando i polimorfismi genetici ad esse associate

La cosa più importante da tenere in considerazione è però il costo che un progetto di questo tipo potrebbe avere. E' nostro interesse ottenere dei finanziamenti da parte degli enti locali, che potrebbero essere interessati ad un progetto di valorizzazione di una attività locale nell'ambito della pesca. Tuttavia dobbiamo proporre un progetto che sia realistico, da un lato per massimizzare la possibilità di ricevere un finanziamento

(ricordiamoci che i progetti vengono valutati da esperti del settore e non possiamo sperare di ricevere una buona valutazione se pianifichiamo un progetto irrealizzabile!), e dall'altro per non ritrovarci in condizioni di ristrettezze economiche nel momento in cui il progetto sia stato finanziato con stime troppo ottimistiche.

Task 1

Teniamo in considerazione che non abbiamo nessuna indicazione a riguardo del genoma di riferimento. In particolare dobbiamo stimare una dimensione approssimativa per calcolare il costo ipotetico del progetto.

Chiediamoci dunque quale potrebbe essere una dimensione ipotetica del genoma di riferimento. Abbiamo due possibili opzioni:

1) Verificare quali sono le dimensioni tipiche dei genomi di altri bivalvi filogeneticamente vicini ad *A. noae*
2) Verificare se esistono stime della dimensione del genoma tramite approcci diversi da quello di sequenziamento, ovvero legate a tecniche citogenetiche

Aiutiamoci con la scheda del World Register of Marine Species (WoRMS): https://www.marinespecies.org/aphia.php?p=taxdetails&id=138788

Anche senza essere bisogno di essere degli esperti di tassonomia, questa ci permette di capire a quale genere, famiglia, ordine, ecc. appartenga la nostra specie di interesse.

Utilizziamo questa informazione per effettuare una ricerca nella sezione del portale bioinformatico NCBI che raccoglie i dati di assemblaggio genomico, accessibile a questo indirizzo: https://www.ncbi.nlm.nih.gov/assembly/

Provate ora a cercare se ci sono genomi già sequenziati disponibili per specie evolutivamente affini ad *A. noae*, specificando nel campo di ricerca il nome del gruppo tassonomico di interesse (può trattarsi di qualsiasi livello, scegliete quello che vi sembra più opportuno, seguito da "[organism]"

Ad esempio, se voleste fare una ricerca degli assemblaggi disponibili per il genere Homo dovreste digitare bel campo di ricerca: "Homo [organism]

Osservate i risultati, facendo molta attenzione nelle tabelle che ritroverete al campo "Total sequence length".

Task 2

Tuttavia una stima basata su una sola specie potrebbe non essere sufficientemente accurata, anche in virtù del fatto che le informazioni che abbiamo trovato non appartengono a specie particolarmente vicine a quella di nostro interesse. Proviamo a ricavare qualche utile informazione in più da un altro database, ovvero Animal Genome Size Database. Potete effettuare una ricerca da questo indirizzo, come fatto sopra: http://www.genomesize.com

Annotate anche in questo caso i risultati ottenuti. In quale range può essere secondo voi stimata a questo punto la dimensione del genoma di *A. noae*?

Task 3

Il modo però più accurato per stimare la complessità del nostro genoma bersaglio potrebbe essere quello di fare un sequenziamento "esplorativo" a bassa copertura, che potrebbe garantirci di valutare anche altri aspetti importanti oltre a quello della dimensione, come ad esempio l'eterozigosità o il contenuto di repeats.

Scaricate dunque il file Excel **"costi_di_sequenziamento.xlsx"** da Moodle, nel quale troverete un listino prezzi di vari servizi di sequenziamento secondo una media di quanto offre ad oggi il mercato, assieme ad alcune specifiche tecniche, che ci potrebbero tornare utili in seguito.

La stima a cui siamo interessati può essere effettuata con un sequenziamento Illumina paired-end. Chiedetevi quale potrebbe essere la copertura di sequenziamento richiesta e quale metodo bisognerebbe utilizzare per ottenere una stima attendibile della dimensione del genoma e del livello di eterozigosità.

Fatto questo, andate sul secondo foglio del file excel, intitolato "stima costi progetto" e provate a fare un calcolo di quale potrebbe essere il costo di questa analisi preliminare. Inserite questa prima cifra in una casella apposita, che andrà poi sommata ai costi relativi alle analisi che dovremmo svolgere successivamente.

Task 4

Immaginiamo a questo punto di aver ottenuto dei dati di sequenziamento Illumina paired-end in modalità 2x100bp con questa analisi preliminare e di voler dunque stimare le dimensioni, l'eterozigosità ed il contenuto di repeats del genoma. Nella cartella su Moodle troverete un file intitolato "**arca_noae_jellyfish.txt**": scaricatelo.

Si tratta del file di output generato da un'analisi condotta con Jellyfish, che ci permetterà di effettuare queste stime tramite la distribuzione dei k-meri ottenuti dai dati di sequenziamento.

Potremo analizzarlo tramite GenomeScope, uno strumento online disponibile a questo indirizzo: http://qb.cshl.edu/genomescope/

Trascinate il file su "Click or drop .histo file here to upload", lasciate tutti gli altri parametri così come sono e cliccate su submit.

Provate ad interpretare i risultati, facendo riferimento alle slide viste a lezione se necessario.

In particolare chiediamoci:

-quale è la dimensione stimata del genoma?

-quale copertura di sequenziamento abbiamo ottenuto con questo approccio?

-quale è la stima del livello di eterozigosità?

-Quale è la stima del livello di repeats presenti nel genoma?

-A che cosa corrispondono i picchi che osservate e a quale copertura corrispondono?

-Come mai si nota un picco a zero?

Task 5

Sulla base di quanto visto nel punto precedente, dovreste avere a disposizione tutti i dati necessari per stimare il costo di sequenziamento del genoma di riferimento. E per pianificare il campionamento.

Chiedetevi dunque quale potrebbe essere la strategia più appropriata, e che livello di copertura di sequenziamento utilizzereste. Provate a fare un po' di conti utilizzando il foglio Excel ed annotate i costi previsti per i diversi tipi di librerie di cui pensate di dover avere bisogno.

Alla luce di quanto avete pensato, quale sarà la strategia di campionamento e gestione del campione biologico più appropriata? Di quali reagenti e tecniche dovreste eventualmente servirvi?

Task 6

Il contenuto di repeats di questo genoma sembra essere piuttosto alto, così come il tasso di eterozigosità. Non abbiamo a disposizione un sequenziatore ONT tra i servizi offerti e pertanto non abbiamo garanzie di poter risolvere questo problema con le sole long reads PacBio. Vi viene in mente una strategia alternativa per riuscire a superare questo problema, basandoci sulle piattaforme di sequenziamento disponibili nel file excel? Pensate ai vari tipi di librerie di sequenziamento di cui abbiamo parlato a lezione e valutate che cosa andrebbe aggiunto ai costi.

Task 7

Ora che avete un piano di sequenziamento completo con tanto di budget, per curiosità provate a chiedervi quale sarebbe il costo di sequenziamento richiesto per il medesimo progetto se dovesse essere completato con una copertura simile a quella ipotizzata per le long reads che avete scelto, ma utilizzando esclusivamente sequenziamento Sanger, come fatto per il genoma umano.

Task 8

Anche le tempistiche sono importanti! Provate a chiedervi in quanto tempo potrebbe essere possibile generare i dati di sequenziamento di vostro interesse con le piattaforme selezionate. Quanto tempo sarebbe invece richiesto per generare la stessa mole di dati con metodiche Sanger, avendo a disposizione un singolo sequenziatore?

Task 9

Abbiamo bisogno di un'ultima cosa per garantirci una annotazione di buona qualità… di che cosa si tratta? Quale strategia di sequenziamento utilizzereste? Approssimativamente, di quante reads potremmo aver bisogno?

Tenete in considerazione per questo punto che i tessuti di maggior rilievo per i bivalvi sono: emolinfa, ghiandola digestiva, mantello, branchie, piede e muscolo adduttore.

Dal punto di vista del campionamento, di che cosa ci dovremmo preoccupare in modo particolare? Eventualmente, di quali reagenti o espedienti dovremmo fare utilizzo?

Task 10

Immaginiamo a questo punto di aver ottenuto i primi risultati del nostro sequenziamento Illumina. Scaricate da Moodle il file "**read_quality_pretrim.html**" ed apritelo. Si tratta di un file che presenta diversi grafici che analizzano la qualità delle reads ottenute. Anche se molti punti riguardano elementi di cui non abbiamo

parlato a lezione, il report dovrebbe essere piuttosto intuitivo. Provate a dare uno sguardo a tutte le voci e a dare una vostra interpretazione (in questo file sono riportati 16 campioni, immaginate pure che quello di nostro interesse sia uno qualsiasi di questi).

In particolare soffermatevi sulla sezione "Per Base Sequence Content", dalla quale potete cliccare sulle singole barre per visualizzare i dati nei dettaglio. Che cosa notate e come lo potete spiegare?

Sulla base di quanto avete appena visto, come impostereste il trimming delle reads?

Task 11

A questo punto scaricate ed aprite il file "**read_quality_posttrim.html**". Vi sembra che i problemi evidenziati nella precedente fase siano stati risolti?

Task 12

Verificata dunque la buona qualità delle reads ottenute (immaginiamo di averlo fatto per tutte le librerie sequenziate) non ci resta che procedere con l'assemblaggio *de novo* del genoma. Sulla base delle reads ottenute, che tipo di strategia di assemblaggio sceglierestе di utilizzare?

Task 13

Ipotizziamo a questo di voler valutare la qualità del genoma assemblato ed annotato ottenuto. Scaricate il file "**BUSCO_report.txt**" e provate ad interpretare i risultati sulla base di quanto visto a lezione.

Oltre ad un'analisi di questo tipo, quali altre metriche potrebbe essere utile analizzare per fornire un report più completo?

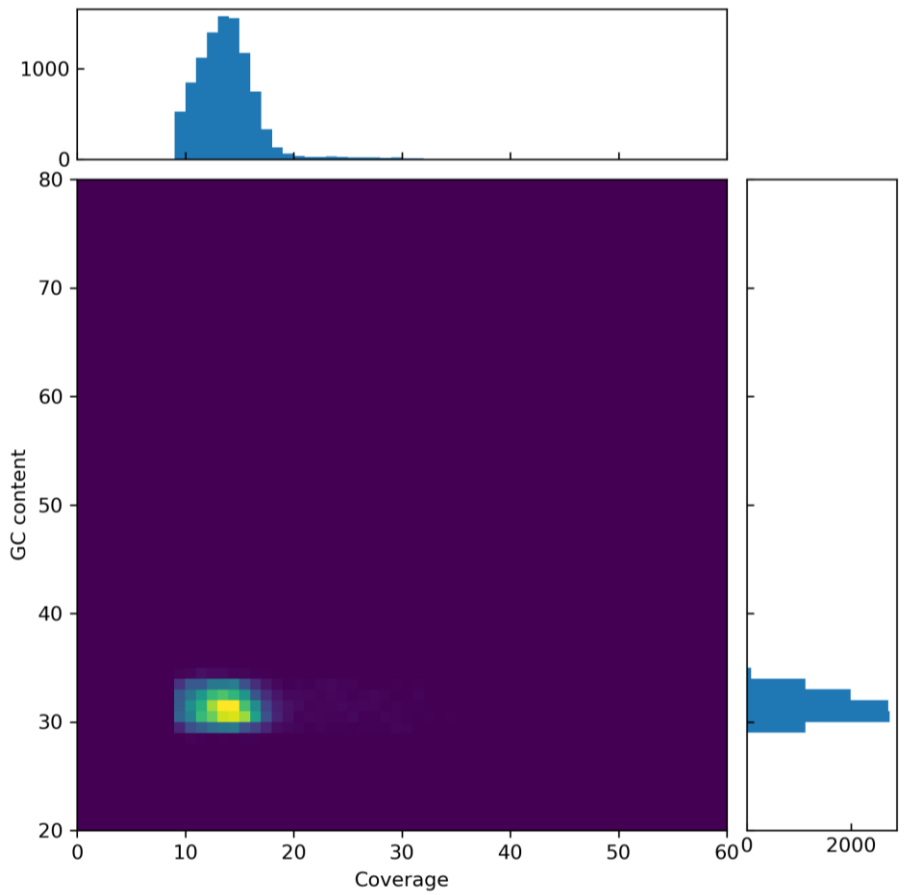
Task 14

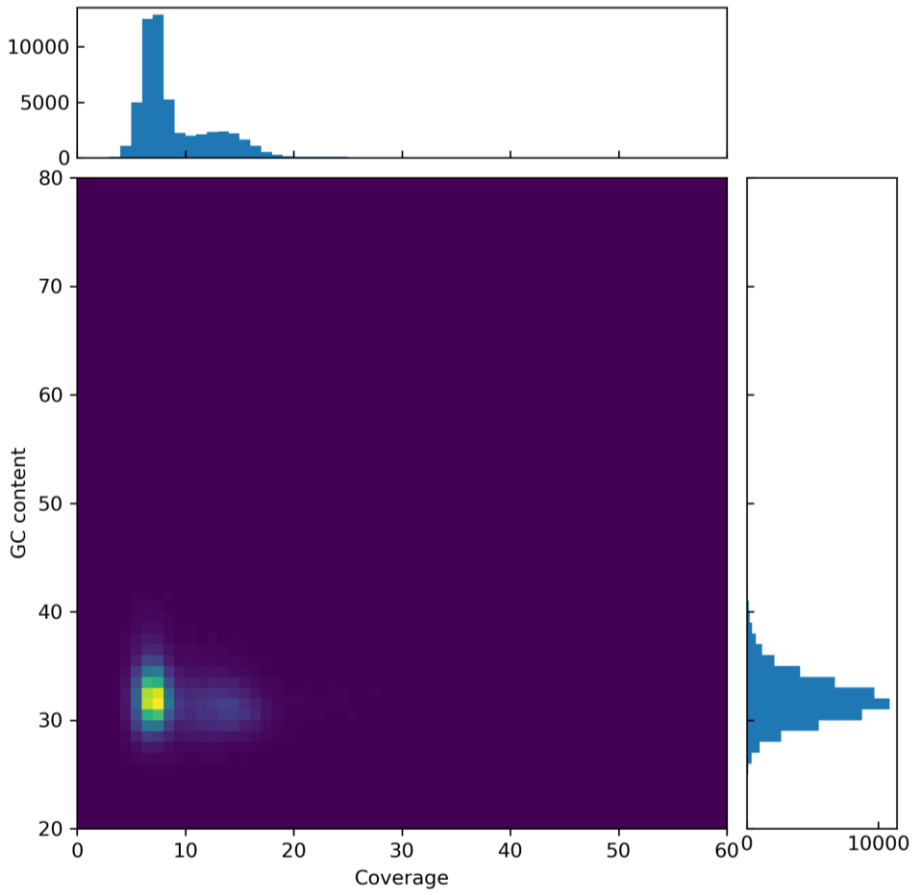
Il nostro controllo qualità però non è ancora completo. In un organismo di questo tipo, che è un filtratore, è possibile ottenere una contaminazione anche piuttosto importante da parte di DNA genomico esogeno, derivante ad esempio da microalghe, zooplankton, batteri o altri piccoli organismi che si trovano dispersi nell'acqua marina. Tutto ciò senza contare la possibilità che siano presenti dei parassiti nei tessuti da noi prelevati! Osservate dunque la figura allegata sotto, che rappresenta un plot di densità che mette in relazione la profondità di sequenziamento con il contenuto in GC dei contig che abbiamo ottenuto dall'assemblaggio.

Il primo grafico rappresenta quello che ci attenderemmo in assenza di contaminazione, mentre il secondo rappresenta quello che notiamo nel nostro campione. Ricordate che:

1) Il contenuto in GC dei contig che abbiamo ottenuto dovrebbe rispecchiare il contenuto in GC a livello dell'intero genoma e pertanto i contigs derivanti da un medesimo organismo dovrebbero formare una curva gaussiana centrata attorno a questo valore.
2) La copertura di sequenziamento dei singoli contigs dovrebbe allinearsi con la profondità del sequenziamento del genoma, dal momento che le reads derivano dalla frammentazione del DNA genomico

Tenendo in considerazione queste due affermazioni, osservate i grafici sottostanti e provate a darne una interpretazione.


1000
0
80
70
60
50
40
30
20
GC content
0
10
20
30
40
50
60
Coverage
0
2000


10000
5000
0
80
70
60
50
40
30
20
GC content
0
10
20
30
40
50
60
Coverage
0
10000

Task 15

Riflettendo su quanto discusso a lezione riguardo alle strategie utilizzate per individuare la presenza (e stimare l'abbondanza) delle singole specie in una comunità microbica o di piccoli eucarioti, riuscite a pensare ad una strategia che potrebbe essere valida per identificare quale possa essere la fonte di questa contaminazione?

Task 16

Diamo una breve dimostrazione del funzionamento di BLAST. Nonostante non sia un argomento trattato a lezione (sebbene sia stato citato molte volte), è importante in questa sede comprendere le sue potenzialità utilizzandolo nella pratica. Immaginiamo dunque di aver ottenuto, tramite alcune ricerche di similarità effettuate a partire da un marker molecolare di interesse, la sequenza che trovate incollata qui sotto in formato FASTA, che presumibilmente, non corrispondendo a quella attesa di *A. noae*, appartiene al contaminante che abbiamo identificato dai grafici sopra.

Recuperate questa sequenza, copiatela, e collegatevi al seguente indirizzo web: https://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE=Nucleotides&PROGRAM=blastn

Da qui sarà possibile effettuare un confronto tra la nostra sequenza ignota ed un database di sequenze nucleotidiche note precedentemente depositate in un database pubblicamente accessibile da altri ricercatori, tramite il programma BLASTN.

Il nostro interesse è verificare se la sequenza che abbiamo identificato trova una corrispondenza con una o più sequenze già presenti nel database, il che ci permetterebbe di individuare la nostra contaminazione esogena.

Incollate la sequenza nella finestra "Enter accession number(s), gi(s), or FASTA sequence(s)", scendete in fondo alla pagina e cliccate su "BLAST"

```
>contaminazione_ignota
GGTTTTTAATCATATTTGATAATAACTTTTAAATAAAACTAAAAGGACGCGGTAACCTTGACCGTGATAA
CGTAGCATAATCACTCGCCATTTAATTGATGGATAGTATGAATGGTTGAACGAATATTCCACTGTCTTAG
AGAGAATCAAAAAAAATTGAAATAGTAGTTAAGATGCTATTTAAAATTGTAAGACGAAAAGACCCTATAG
AGCTTCACTATACACTCTTTATAACGAACGAAACCTATTTTTTAAATTAAGGAGTATGGTAGTTTAGTTG
GGGCGACTACTTTCTAAATCTAACGAAAGCAAGCAATGTTAATGATAAATTTACTGTATAACTAAATTTT
TTAACAGTTATTAATATAGGCCATAATGACCCGTTGTGTTTTCAGAATTAAACACAACGATCAATTGATA
AAAGCTACCTTAGGGATAACAGGATAATTTTATTTTAGAGTTCTTATCGAAAATAAAGTTTGTCACCTCT
ATGTTGAATTAA
```

Task 17

Attendete alcune dozzine di secondi per ottenere i risultati, che saranno visualizzati sotto forma di una tabella che riporterà una lista di sequenze che mostrano elevata similarità con la nostra, associate al nome scientifico della specie, ad un accession ID univoco delle sequenze e ad alcune statistiche (ad esempio la % di identità tra la nostra sequenza di interesse (la sequenza query) e quella ritrovata all'interno del database (la sequenza subject). L'allineamento tra le due sequenze è visualizzabile nel dettaglio cliccando su "Alignments". Provate a dare uno sguardo ai risultati e ad interpretarli. Vi sembra sia possibile stabilire con certezza l'origine di questa contaminazione?

Task 18

Verifichiamo se questa contaminazione possa avere un senso e chiediamoci anche quale possa essere stata la sua origine. Fa una grande differenza sapere se la contaminazione è stata originata dalla filtrazione, come detto in precedenza, da un parassita, oppure da una errata manipolazione dei campioni biologici durante le fasi di estrazione degli acidi nucleici, preparazione delle librerie oppure sequenziamento. Collegatevi dunque a Pubmed, importante database di letteratura scientifica, e cercate gli eventuali articoli pubblicati che abbiano qualcosa a che vedere con la specie di cui abbiamo appena trovato evidenza. Che cosa vi sembra di poter concludere a riguardo?

Task 19

Non abbiamo ancora completato il nostro piano finanziario… ora che siamo sicuri di avere a disposizione un genoma di riferimento di buona qualità e privo di contaminazioni, per sostenere un progetto di produzione nel medio-lungo periodo sarà opportuno valutare la diversità genetica tra le popolazioni attualmente esistenti, anche per porre le basi per comprendere meglio le eventuali differenze qualitative esistenti tra popolazioni residuali che non sono attualmente connesse. Immaginiamo un progetto internazionale che voglia coinvolgere Italia e Croazia e che punti dunque a risequenziare il genoma di una dozzina di individui appartenenti a due popolazioni distinte: quella dell'alto Adriatico che si trova sulle coste italiane, e quella del basso Adriatico che si trova sulle isole del sud della Dalmazia.

Che tecnica di sequenziamento sarebbe maggiormente opportuna? Quale profondità di sequenziamento potrebbe essere indicata per riuscire a mappare in modo efficiente i polimorfismi, tenendo in considerazione che si tratta di una specie diploide? Aggiungete ai calcoli fatti anche questa ipotesi di costi per completare il piano finanziario del progetto.