BIOINFORMATICA
II

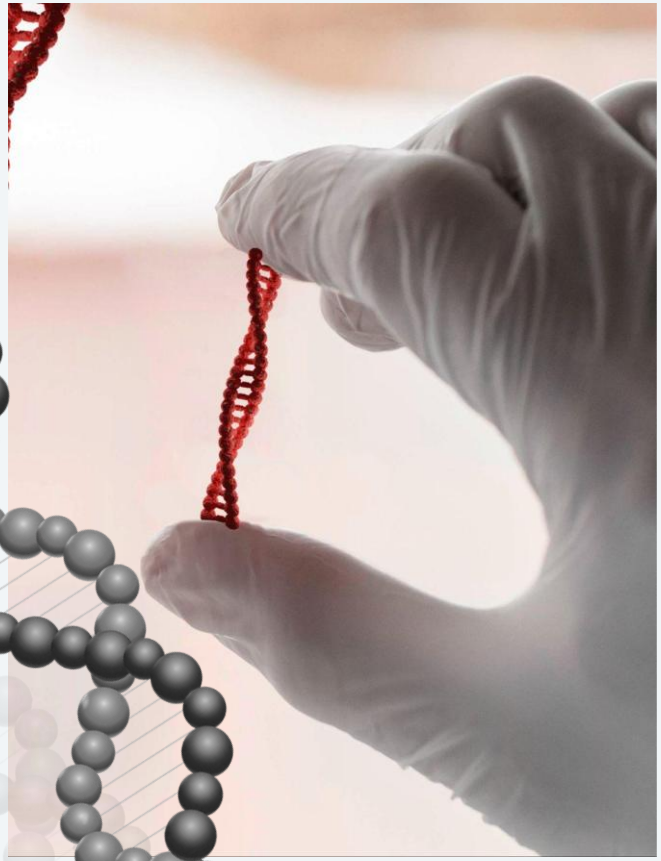
UN APPROCCIO PRATICO ALLA BIOINFORMATICA

ASSEMBLAGGIO TRASCRITTOMA DE NOVO

Argomenti

01 *DE NOVO* ASSEMBLY TRASCRIPTOME

02 WORFLOW



01

DE NOVO ASSEMBLY TrASCRIPTOME

DE NOVO TASCRIPTOME: Generalità

La tecnica di assembly del trascrittoma de novo è un approccio bioinformatico per ricostruire il trascrittoma di un organismo a partire dai dati di sequenziamento RNA-Seq, senza un genoma di riferimento. Questo metodo è particolarmente utile per organismi non modelli o per studi in cui non è disponibile un genoma ben annotato

DE NOVO TASCRIPTOME Obiettivo e uso

Scopo:
Ricostruire tutte le sequenze di RNA (trascritti) presenti in un campione, inclusi mRNA, lncRNA e altri tipi di RNA codificanti e non codificanti.
Identificare nuovi trascritti, isoforme alternative e livelli di espressione genica.

Quando viene utilizzato:
Studi su organismi senza un genoma di riferimento.
Analisi comparativa di trascrittomi tra specie non strettamente correlate.
Rilevamento di trascritti rari o altamente specifici.

DE NOVO TASCRIPTOME: Processo di creazione

1. **Pre-elaborazione dei dati (Data Preprocessing)**: I dati RNA-Seq grezzi sono analizzati e filtrati per rimuovere basi di bassa qualità, adattatori, contaminanti e sequenze troppo corte

2. **Assemblaggio de novo**: Le letture RNA-Seq sono assemblate direttamente senza un riferimento genomico.Gli algoritmi sfruttano grafi di de Bruijn o approcci di overlap per unire le sequenze in contigs.

3. **Valutazione e miglioramento dell'assemblaggio** (uso di metriche di qualità come N50, Lunghezza media del contig, Numero di contigs) e **completezza** (Benchmarking Universal Single-Copy Orthologs,- BUSCO verifica la presenza di geni ortologhi universali per valutare la completezza dell'assemblaggio.

DE NOVO TASCRIPTOME: Processo di creazione

4. **Ridondanza:** Gli assemblaggi possono contenere contigs ridondanti o frammenti, per cui si utilizzano strumenti per ridurre la ridondanza e migliorare la qualità.

5. **Annotazione funzionale:** Gli assemblaggi vengono annotati per identificare le sequenze codificanti (ORFs) e prevedere funzioni.

DE NOVO TASCRIPTOME: Principali problemi

Ridondanza e frammentazione:
Gli assemblaggi de novo possono generare trascritti ridondanti o frammentati.
Bias nell'abbondanza:
I trascritti altamente espressi possono dominare l'assemblaggio, oscurando quelli rari.
Errori nell’assemblaggio:
Errori dovuti a letture di bassa qualità o a trascritti sovrapposti.
Annotazione limitata:
Per specie non modello, l’assenza di banche dati ben curate rende difficile l’annotazione.

DE NOVO TASCRIPTOME: Applicazioni

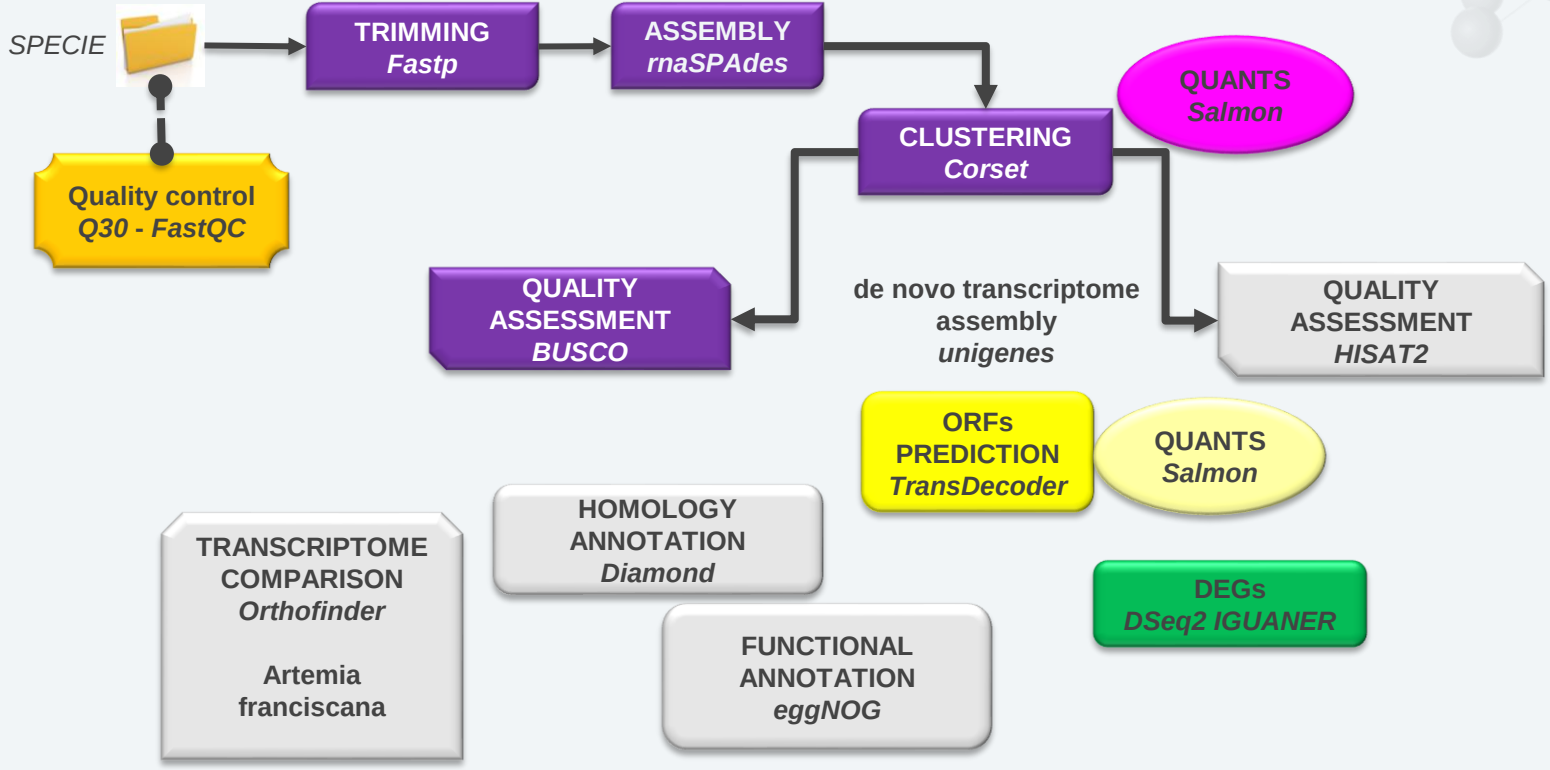
Scoperta di nuovi geni:
Identificazione di trascritti unici per specie non modello.

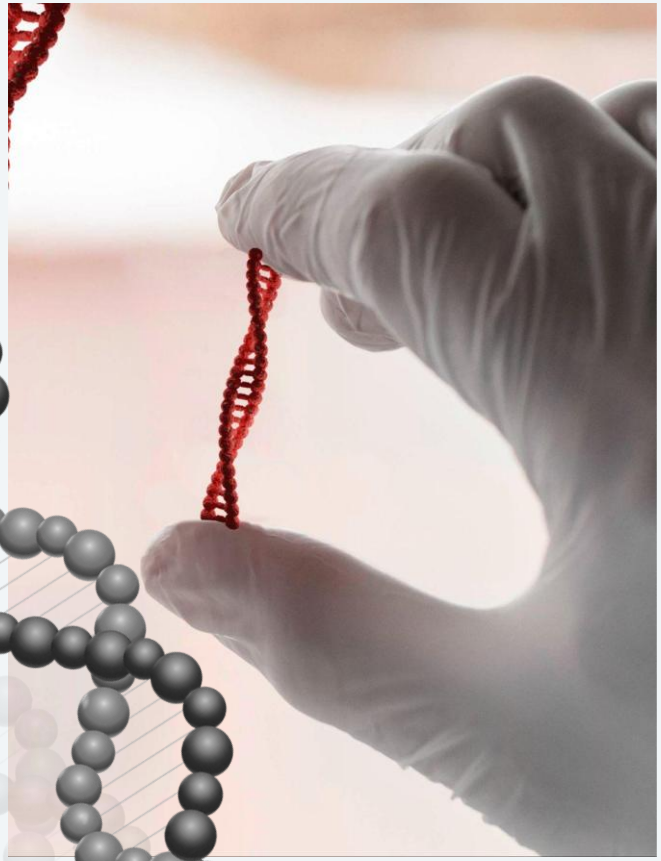
Analisi di isoforme:
Studio di splicing alternativo e varianti di trascritti.

Studio dell'evoluzione:
Confronto di trascrittomi tra specie.

Ricerca applicata:
Identificazione di target terapeutici o biomarcatori in condizioni specifiche.

DE NOVO ASSEMBLY TRANSCRIPTOME





02

WORKFLOW

DE NOVO TASCRIPTOME: 1 Filtered

Rimuovere adattatori e reads di bassa qualità

Strumento: FASTP
Comando:

PAIRED END

fastp -i input_R1.fastq -I input_R2.fastq -o output_R1_trimmed.fastq -O output_R2_trimmed.fastq --html report.html --json report.json --thread 15

SINGLE END

fastp -i input.fastq -o output_trimmed.fastq --html report.html --json report.json --thread 15

15 è il massimo numero di thread

DE NOVO TASCRIPTOME: 1 Filtered

COMANDO SU RAGANELLA

```
./fastp -i /data/ESERCITAZIONE2025/LIBERATI3/CL1_R1.fq.gz  -I /data/ESERCITAZIONE2025/LIBERATI3/CL1_R2.fq.gz
-o /data/ESERCITAZIONE2025/LIBERATI3/TRIMMED/CL1_R1_trimmed.fastq -O /data/ESERCITAZIONE2025/LIBERATI3/TRIMMED/CL1_R2_trimmed.fastq
--html /data/ESERCITAZIONE2025/LIBERATI3/TRIMMED/CL1_report.html
--json /data/ESERCITAZIONE2025/LIBERATI3/TRIMMED/CL1_report.json --thread 15
```

```
./fastp -i /data/ESERCITAZIONE2025/LIBERATI3/CL2_R1.fq.gz  -I /data/ESERCITAZIONE2025/LIBERATI3/CL2_R2.fq.gz
-o /data/ESERCITAZIONE2025/LIBERATI3/TRIMMED/CL2_R1_trimmed.fastq -O /data/ESERCITAZIONE2025/LIBERATI3/TRIMMED/CL2_R2_trimmed.fastq
--html /data/ESERCITAZIONE2025/LIBERATI3/TRIMMED/CL2_report.html
--json /data/ESERCITAZIONE2025/LIBERATI3/TRIMMED/CL2_report.json --thread 15
```

```
./fastp -i /data/ESERCITAZIONE2025/LIBERATI3/LL86_R1.fq.gz  -I /data/ESERCITAZIONE2025/LIBERATI3/LL86_R2.fq.gz
-o /data/ESERCITAZIONE2025/LIBERATI3/TRIMMED/LL86_R1_trimmed.fastq -O /data/ESERCITAZIONE2025/LIBERATI3/TRIMMED/LL86_R2_trimmed.fastq
--html /data/ESERCITAZIONE2025/LIBERATI3/TRIMMED/LL86_report.html
--json /data/ESERCITAZIONE2025/LIBERATI3/TRIMMED/LL86_report.json --thread 15
```

```
./fastp -i /data/ESERCITAZIONE2025/LIBERATI3/LL87_R1.fq.gz  -I /data/ESERCITAZIONE2025/LIBERATI3/LL87_R2.fq.gz  -o
/data/ESERCITAZIONE2025/LIBERATI3/TRIMMED/LL87_R1_trimmed.fastq -O /data/ESERCITAZIONE2025/LIBERATI3/TRIMMED/LL87_R2_trimmed.fastq --
html /data/ESERCITAZIONE2025/LIBERATI3/TRIMMED/LL87_report.html
--json /data/ESERCITAZIONE2025/LIBERATI3/TRIMMED/LL87_report.json --thread 15
```

DE NOVO TASCRIPTOME: 2 Assembly

Assemblaggio

Strumento: SPAdes-4-0-0
Comando:

```
rnaspades.py -t 48
-o ASSEMBLY/SPECIE_spades_k_auto
--dataset specie_dataset.yaml
--tmp-dir /ASSEMBLY/TMP_SPADE/ --only-assembler
```

DE NOVO TASCRIPTOME: 2 Assembly

Assemblaggio: scrittura dataset

Strumento: NOTEPAD/NANO/VI
File da scrivere:

```
[
    {
      orientation: "fr",
      type: "paired-end",
      right reads: [
"campione1trimm_1.fq",
"campione2trimm_1.fq",
…
      ],
      left reads: [
"campione1trimm_2.fq",
"campione2trimm_2.fq",
…
      ]
    }
]
```

DE NOVO TASCRIPTOME: 2 Assembly

COMANDO RAGANELLA

nano dataset.yaml

```
[
    {
      orientation: "fr",
      type: "paired-end",
      right reads: [
"/data/ESERCITAZIONE2025/LIBERATI3/TRIMMED/CL1_R1_trimmed.fastq",
"/data/ESERCITAZIONE2025/LIBERATI3/TRIMMED/CL2_R1_trimmed.fastq",
"/data/ESERCITAZIONE2025/LIBERATI3/TRIMMED/LL86_R1_trimmed.fastq",
"/data/ESERCITAZIONE2025/LIBERATI3/TRIMMED/LL87_R1_trimmed.fastq"
      ],
      left reads: [
"/data/ESERCITAZIONE2025/LIBERATI3/TRIMMED/CL1_R2_trimmed.fastq",
"/data/ESERCITAZIONE2025/LIBERATI3/TRIMMED/CL2_R2_trimmed.fastq",
"/data/ESERCITAZIONE2025/LIBERATI3/TRIMMED/LL86_R2_trimmed.fastq",
"/data/ESERCITAZIONE2025/LIBERATI3/TRIMMED/LL87_R2_trimmed.fastq"
      ]
    }
]
```

python rnaspades.py -t 48 -o /data/ESERCITAZIONE2025/LIBERATI3/ASSEMBLY/SPECIE_spades_k_auto --dataset /data/ESERCITAZIONE2025/LIBERATI3/dataset.yaml --tmp-dir /data/ESERCITAZIONE2025/LIBERATI3//ASSEMBLY/TMP_SPADE/ --only-assembler

DE NOVO TASCRIPTOME: 2 Assembly

COMANDO RAGANELLA

```
python rnaspades.py
-t 48
-o /data/ESERCITAZIONE2025/LIBERATI3/ASSEMBLY/SPECIE_spades_k_auto
--dataset /data/ESERCITAZIONE2025/LIBERATI3/dataset.yaml
--tmp-dir /data/ESERCITAZIONE2025/LIBERATI3//ASSEMBLY/TMP_SPADE/ --
only-assembler
```

DE NOVO TASCRIPTOME: Copia transcritto (opzionale)

COMANDO RAGANELLA

Dalla root di lavoro /data/ESERCITAZIONE2025/LIBERATI2
mkdir TRANSCRIPT

Da /data/ESERCITAZIONE2025/LIBERATI2/ASSEMBLY/
SPECIE_spades_k_auto/
cp transcript.fasta ../../TRANSCRIPT/specie_rnaspades-.transcript.fasta

DE NOVO TASCRIPTOME: 3 Clustering

Clustering

Strumento: CORSET

Comandi:

CREAZIONE INDICE QUANTI

salmon index –index INDEX_SALMON/SPECIE_indici --transcripts /TRANSCRIPT/transcripts.fasta

CREAZIONE QUANTI

salmon quant --index INDEX_SALMON/SPECIE_indici --libType A -1 / TRIMMED/CAMPIONE1_R1_trimmed.fastq -2 TRIMMED/CAMPIONE1_R2_trimmed.fastq --output SALMON/SALMON_CAMPIONE --dumpEq

DE NOVO TASCRIPTOME: 3 Clustering

Clustering

Strumento: CORSET

Comandi:

DECOMPRESSIONE

gunzip -k /SALMON/*/aux_info/eq_classes.txt.gz

CORSET

./corset corset -i salmon_eq_classes /SALMON_*/aux_info/eq_classes.txt -f true

CLUSTERING

python fetchClusterSeqs.py
-i TRANSCRIPT/specie_rnaspades_transcripts.fasta
-o /TRANSCRIPT/specie_corset_transcript.fasta
-c clusters.txt

DE NOVO TASCRIPTOME: 3 Clustering

COMANDI RAGANELLA

salmon index --index /data/ESERCITAZIONE2025/LIBERATI3/INDEX_SALMON/SPECIE_indici --transcripts /data/ESERCITAZIONE2025/LIBERATI3/TRANSCRIPT/specie_rnaspades_transcripts.fasta

salmon quant --index /data/ESERCITAZIONE2025/LIBERATI3/INDEX_SALMON/SPECIE_indici --libType A -1 /data/ESERCITAZIONE2025/LIBERATI3/TRIMMED/CL1_R1_trimmed.fastq -2 /data/ESERCITAZIONE2025/LIBERATI3/TRIMMED/CL1_R2_trimmed.fastq --output /data/ESERCITAZIONE2025/LIBERATI3/SALMON_RNA/CL1 --dumpEq

./salmon quant --index /data/ESERCITAZIONE2025/LIBERATI3/INDEX_SALMON/SPECIE_indici --libType A -1 /data/ESERCITAZIONE2025/LIBERATI3/TRIMMED/LL86_R1_trimmed.fastq -2 /data/ESERCITAZIONE2025/LIBERATI3/TRIMMED/LL86_R2_trimmed.fastq --output /data/ESERCITAZIONE2025/LIBERATI3/SALMON_RNA/LL86 --dumpEq

./salmon quant --index /data/ESERCITAZIONE2025/LIBERATI3/INDEX_SALMON/SPECIE_indici --libType A -1 /data/ESERCITAZIONE2025/LIBERATI3/TRIMMED/CL2_R1_trimmed.fastq -2 /data/ESERCITAZIONE2025/LIBERATI3/TRIMMED/CL2_R2_trimmed.fastq --output /data/ESERCITAZIONE2025/LIBERATI3/SALMON_RNA/CL2 --dumpEq

./salmon quant --index /data/ESERCITAZIONE2025/LIBERATI3/INDEX_SALMON/SPECIE_indici --libType A -1 /data/ESERCITAZIONE2025/LIBERATI3/TRIMMED/LL87_R1_trimmed.fastq -2 /data/ESERCITAZIONE2025/LIBERATI3/TRIMMED/LL87_R2_trimmed.fastq --output /data/ESERCITAZIONE2025/LIBERATI3/SALMON_RNA/LL87 --dumpEq

DE NOVO TASCRIPTOME: 3 Clustering

COMANDI RAGANELLA

gunzip -k /data/ESERCITAZIONE2025/LIBERATI3/SALMON_RNA/*/aux_info/eq_classes.txt.gz

Verifica se ci sono tutti

ls -lh /data/ESERCITAZIONE2025/LIBERATI3/SALMON_RNA/*/aux_info/eq_classes.txt

./corset corset -i salmon_eq_classes /data/ESERCITAZIONE2025/LIBERATI3/SALMON_RNA/*/aux_info/eq_classes.txt -f true

spostamento di Cluster.txt e Count.txt per tenerne traccia nella propria cartella

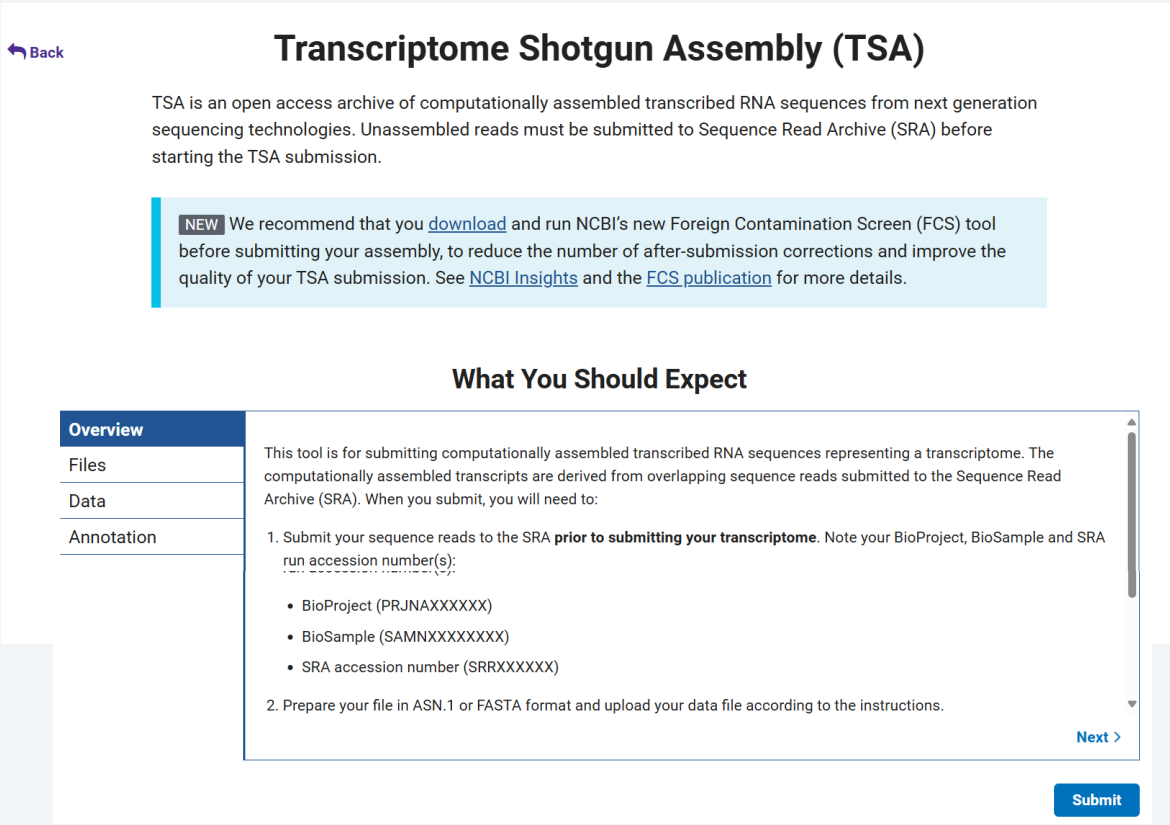
python fetchClusterSeqs.py -i /data/ESERCITAZIONE2025/LIBERATI3/TRANSCRIPT/specie_rnaspades_transcripts.fasta -o /data/ESERCITAZIONE2025/LIBERATI3/TRANSCRIPT/specie_corset_transcripts.fasta -c /data/ESERCITAZIONE2025/LIBERATI3/clusters.txt

DE NOVO TASCRIPTOME: 4 NCBI

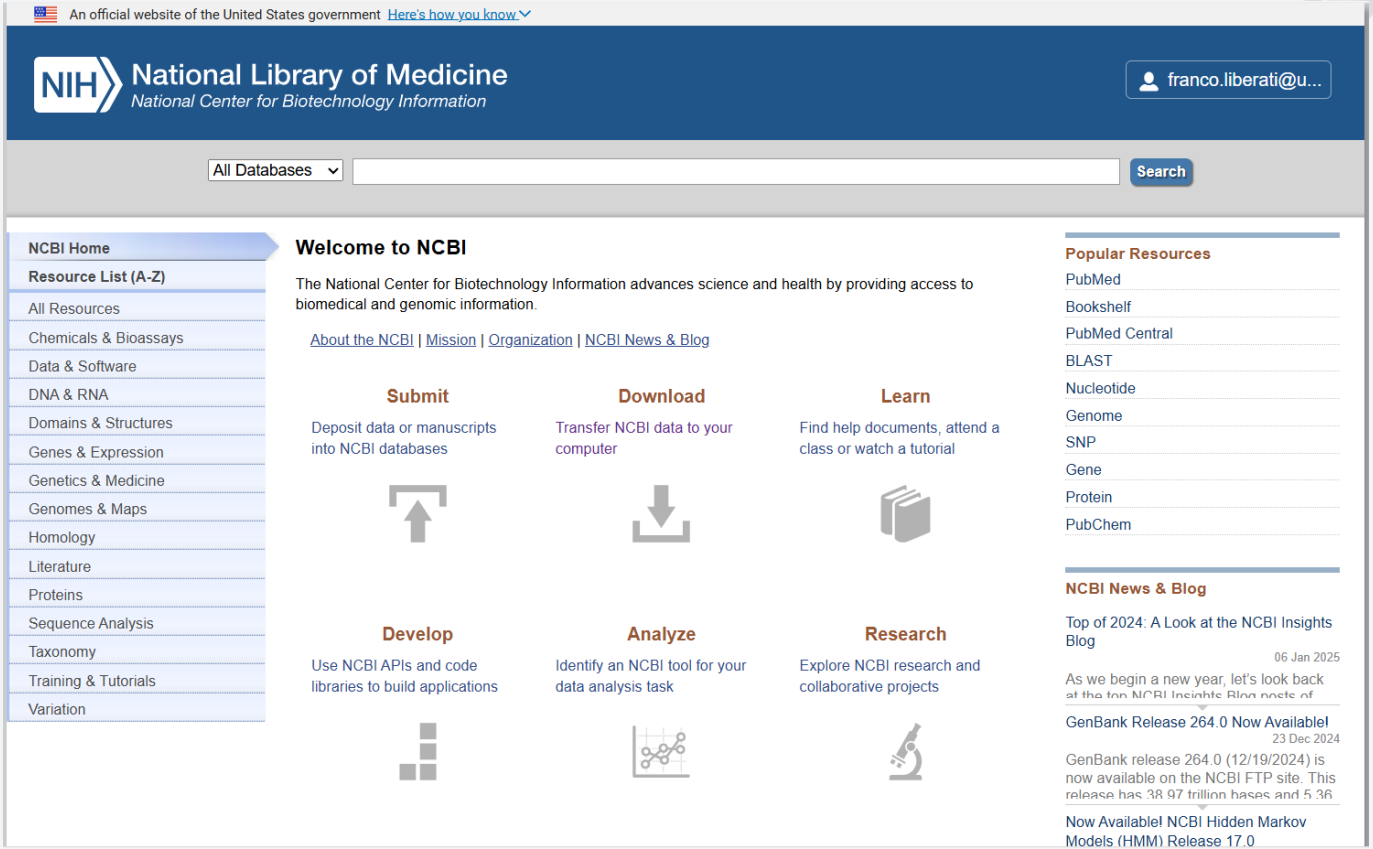
Una volta ottenuto il transcrittoma questo può essere depositato su NCBI dove la sequenza fasta può essere ulteriormente modificata a causa del rilevamento della presenza di batteri, virus, funghi non filtrati nel pre-processamento di **decontaminazione**.

A questo punto sarà il nuovo transcrittoma NCBI a dover essere preso in considerazione anche per garantire la replicabilità dell’esperimento.

DE NOVO TASCRIPTOME: 4 NCBI



DE NOVO TASCRIPTOME: 4 NCBI



DE NOVO TASCRIPTOME: 5 BUSCO

BUSCO (Benchmarking Universal Single-Copy Orthologs) è uno strumento bioinformatico progettato per valutare la completezza degli assemblaggi genomici, trascrittomici o proteici
Si basa su set di geni ortologhi universali (cioè geni che sono conservati in quasi tutti gli organismi di un determinato gruppo tassonomico) e fornisce una metrica quantitativa e qualitativa della qualità e della completezza di un dataset biologico.

Determinare quanto un assemblaggio o un dataset rappresenti accuratamente l'intero contenuto genetico o trascrittomico atteso di un organismo.

DE NOVO TASCRIPTOME: 5 BUSCO

BUSCO utilizza i geni ortologhi conservati come riferimento e fornisce una valutazione basata su quattro categorie principali:

Complete (C): Geni ortologhi trovati nel dataset di input che sono completi e funzionali.

Possono essere ulteriormente classificati in:

Single-Copy (S): Presenti come una singola copia.

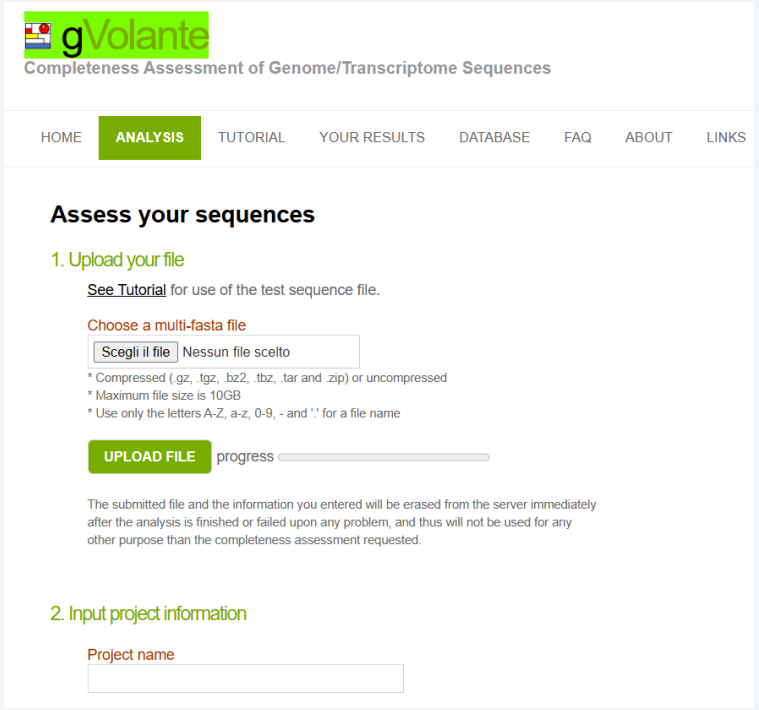
Duplicated (D): Presenti in più copie (indicativo di duplicazione o errori nell'assemblaggio).

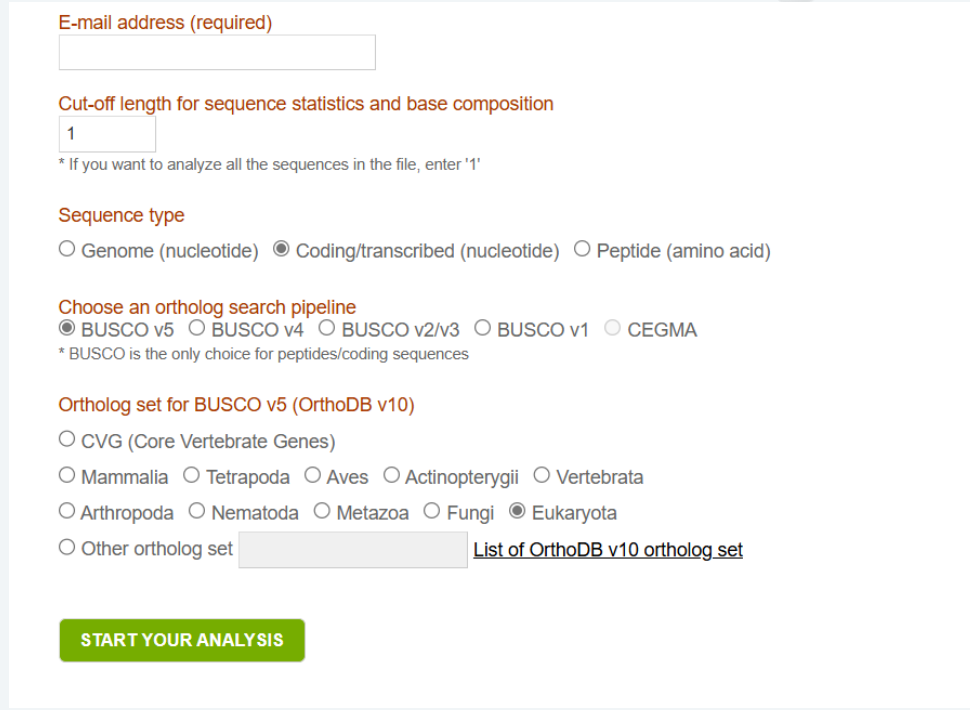
Fragmented (F): Geni ortologhi che sono parzialmente presenti nel dataset, indicando frammentazione dell'assemblaggio.

Missing (M): Geni ortologhi attesi che non sono stati trovati nel dataset, suggerendo incompletezza.

DE NOVO TASCRIPTOME: 5 BUSCO

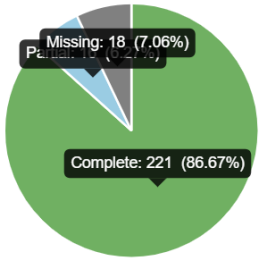
https://gvolante.riken.jp/analysis.html





DE NOVO TASCRIPTOME: 5 BUSCO

RISULTATI



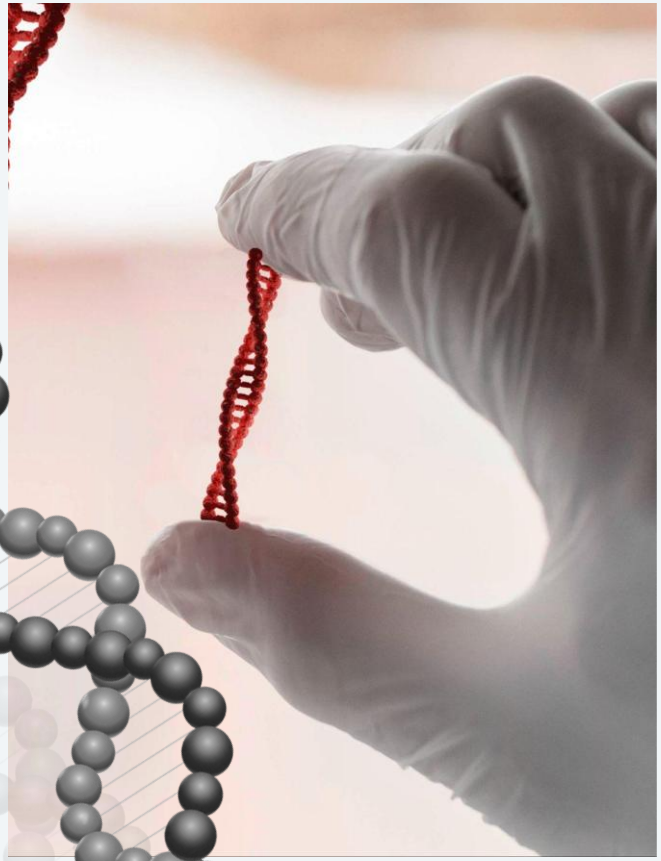
SHOW ORTHOLOG DETAILS

COMPLETENESS ASSESSMENT RESULTS:

Total number of core genes queried	255
Number of core genes detected	
Complete	221 (86.67%)
Complete + Partial	237 (92.94%)
Number of missing core genes	18 (7.06%)
Average number of orthologs per core genes	1.57
% of detected core genes that have more than 1 ortholog	43.89
Scores in BUSCO format	C:86.6%[S:48.6%,D:38.0%],F:6.3%,M:7.1%

LENGTH STATISTICS AND COMPOSITION:

Number of sequences	34710
Total length (nt)	39815585
Longest sequence (nt)	10980
Shortest sequence (nt)	145
Mean sequence length (nt)	1147
Median sequence length (nt)	801
N50 sequence length (nt)	1683
L50 sequence count	7317



02

DEGS

DE NOVO TASCRIPTOME: DEGs

DEGs (Differentially Expressed Genes) sono geni che mostrano un livello di espressione significativamente diverso tra due o più condizioni sperimentali o gruppi (ad esempio, tessuti sani vs. malati, trattati vs. non trattati, condizioni normali vs. stress). L'analisi dei DEGs è fondamentale nella ricerca genomica per identificare geni potenzialmente coinvolti in determinati processi biologici, patologie o risposte a trattamenti.

DE NOVO TASCRIPTOME: DEGs

Dataset di RNA-Seq

I livelli di espressione genica vengono misurati in termini di reads mappati su ciascun gene o trascritto.
Le abbondanze sono normalizzate per tenere conto di fattori come la lunghezza del gene e la profondità di sequenziamento.

Analisi statistica

Confronti tra gruppi sperimentali per identificare geni con variazioni significative nell'espressione

DE NOVO TASCRIPTOME: DEGs

Strumenti comuni

DESeq2: Utilizza modelli statistici per identificare DEGs.
edgeR: Analisi per dati di conteggio RNA-Seq.
limma-voom: Adatta per dati normalizzati.

Criteri comuni per definire i DEGs

Fold Change (FC): Indica quanto aumenta o diminuisce l'espressione di un gene tra due condizioni.
p-value o Adjusted p-value (FDR): Misura la significatività statistica, corretta per test multipli.

DE NOVO TASCRIPTOME: DEGs

Heatmap:

Mostra il livello di espressione dei DEGs in diverse condizioni.

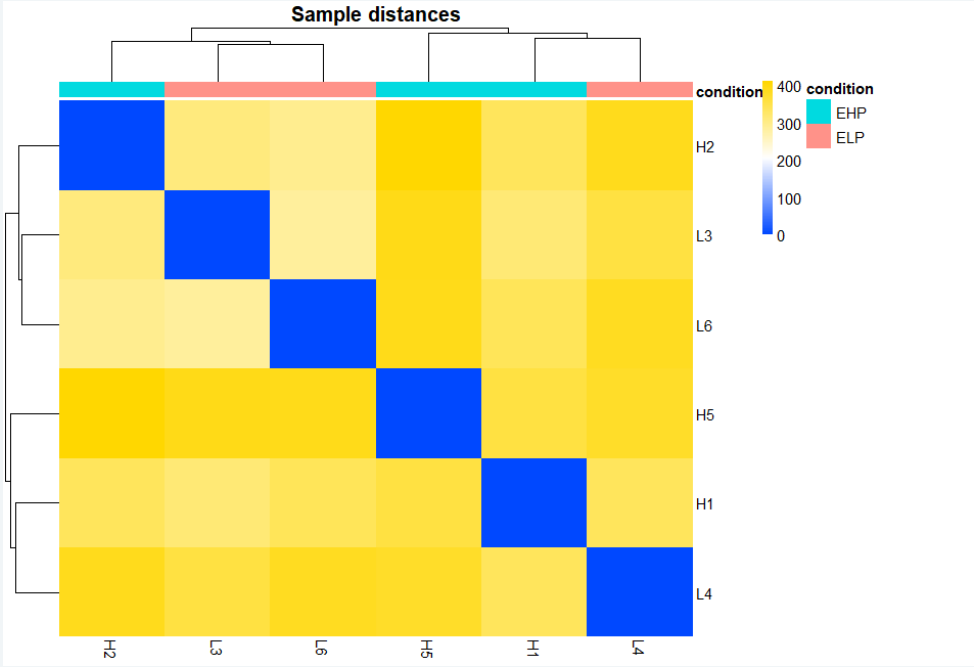
Rappresenta dati come una matrice, dove:

Righe: Geni.

Colonne: Campioni o condizioni.

I colori indicano livelli di espressione (es. blu per espressione bassa, rosso per alta).

Utilizzata per identificare pattern di espressione e cluster di geni o campioni.

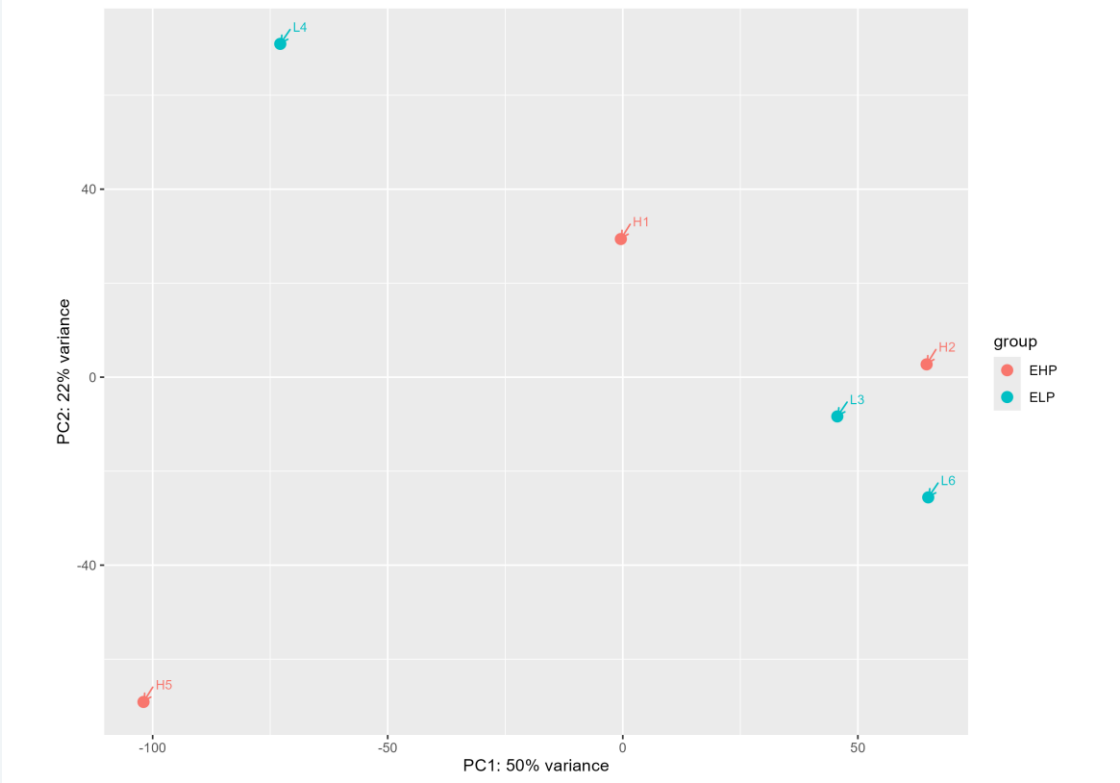


DE NOVO TASCRIPTOME: DEGs

PCA Plot (Principal Component Analysis):

Rappresenta i campioni basandosi sulle principali componenti della variazione nei dati.

Utile per verificare se i campioni si separano chiaramente in base alle condizioni.



DE NOVO TASCRIPTOME: DEGs

Volcano Plot:

Unisce il fold change (log2FC) e il p-value (-log10).

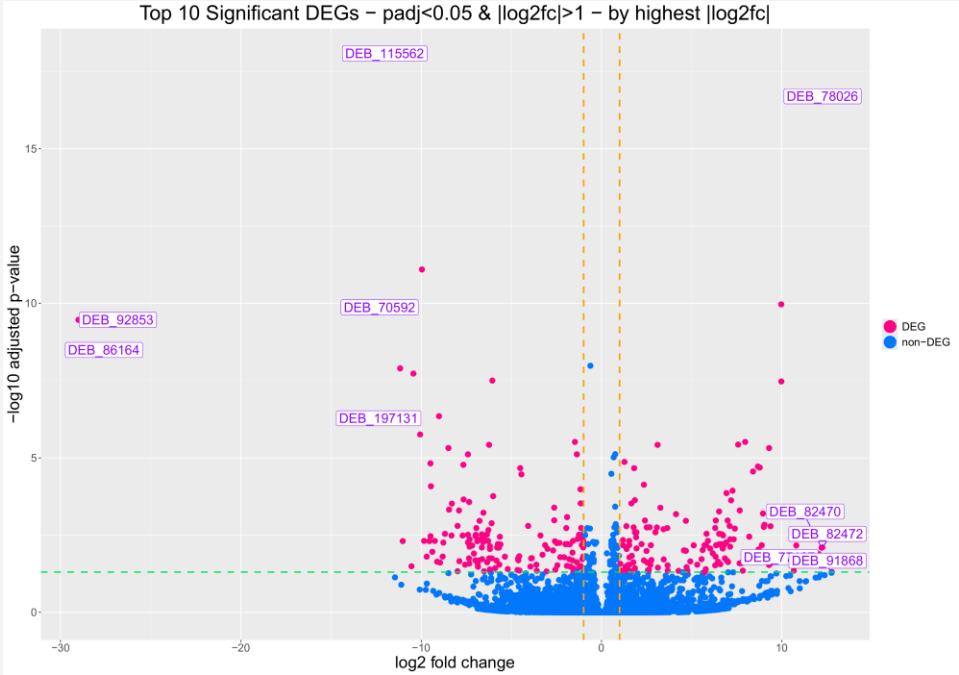
Visualizza quali geni sono significativamente up-regolati o down-regolati.

Esempio:

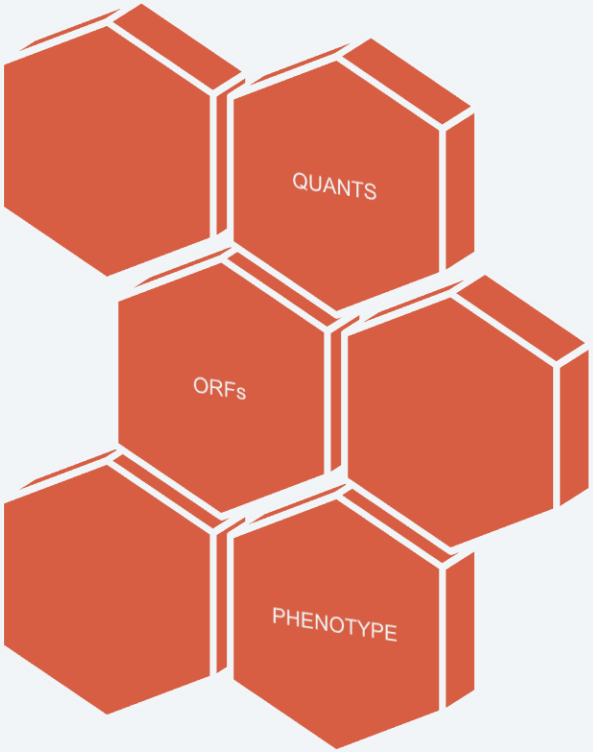
Asse X: log2(Fold Change).

Asse Y: -log10(p-value).

Geni con alta significatività e grandi variazioni di espressione appaiono lontano dal centro (in alto e ai lati)



DEGs: File utili



DEGs: QUANTS definizione

I file quants (prodotti da Salmon) sono il risultato del processo di quantificazione dell'espressione genica o trascrittomica.

Forniscono una stima dell'espressione dei trascritti in un campione, utile per identificare differenze di espressione tra campioni o condizioni.

I dati dei file quants vengono tipicamente utilizzati in software di analisi statistica come DESeq2, edgeR, o Sleuth per condurre analisi differenziali.

I valori come TPM sono utili per confrontare i livelli di espressione tra campioni, correggendo differenze dovute alla lunghezza dei trascritti o al numero totale di letture sequenziate

DEGs: QUANTS formato

I file quants sono in genere file di testo o file tabulari contenenti dati strutturati relativi ai trascritti o ai geni analizzati. I formati più comuni includono TSV o CSV, e possono essere accompagnati da metadati utili per analisi successive.

Campo	Significato
Name	Identificatore del trascritto (ad esempio, un nome come ENST per annotazioni Ensembl)
Length	Lunghezza del trascritto
EffectiveLength	Lunghezza effettiva del trascritto, che tiene conto delle probabilità di mappatura degli spezzoni (frammenti di RNA-Seq)
TPM (Transcripts Per Million)	Una misura normalizzata per la quantificazione dell'espressione, che tiene conto della lunghezza del trascritto e del numero totale di letture nel campione.
NumReads	Numero grezzo di letture mappate su ciascun trascritto
EstimatedCounts	Quantità stimata di letture attribuite a un trascritto, dopo l'elaborazione probabilistica del software

DEGs: QUANTS esempio

Name	Length	EffectiveLength	TPM	NumReads
ENST00000456328	1657	1507.44	50.00	300
ENST00000450305	632	482.33	20.00	100
ENST00000488147	1351	1201.50	10.00	50

DEGs: QUANTS altre informazioni

Oltre ai file quant.sf, Salmon può generare:

Log files: Informazioni sul processo di quantificazione, utili per il debug.

Auxiliary files: File di supporto contenenti informazioni sulle librerie o sugli indici.

DEGs QUANTs (comandi trascrittoma NCBI o corset)

salmon index --index /data/ESERCITAZIONE2025/LIBERATI3/INDEX_CORSET/SPECIE_indici --transcripts
/data/ESERCITAZIONE2025/LIBERATI3/TRANSCRIPT/specie_corset_transcripts.fasta

./salmon quant --index /data/ESERCITAZIONE2025/LIBERATI3/INDEX_CORSET/SPECIE_indici --libType A -1
/data/ESERCITAZIONE2025/LIBERATI3/TRIMMED/CL1_R1_trimmed.fastq -2
/data/ESERCITAZIONE2025/LIBERATI3/TRIMMED/CL1_R2_trimmed.fastq --output
/data/ESERCITAZIONE2025/LIBERATI3/QUANTS_CORSET/CL1 --dumpEq

./salmon quant --index /data/ESERCITAZIONE2025/LIBERATI3/INDEX_CORSET/SPECIE_indici --libType A -1
/data/ESERCITAZIONE2025/LIBERATI3/TRIMMED/CL2_R1_trimmed.fastq -2
/data/ESERCITAZIONE2025/LIBERATI3/TRIMMED/CL2_R2_trimmed.fastq --output
/data/ESERCITAZIONE2025/LIBERATI3/QUANTS_CORSET/CL2 --dumpEq

./salmon quant --index /data/ESERCITAZIONE2025/LIBERATI3/INDEX_CORSET/SPECIE_indici --libType A -1
/data/ESERCITAZIONE2025/LIBERATI3/TRIMMED/LL86_R1_trimmed.fastq -2
/data/ESERCITAZIONE2025/LIBERATI3/TRIMMED/LL86_R2_trimmed.fastq --output
/data/ESERCITAZIONE2025/LIBERATI3/QUANTS_CORSET/LL86 --dumpEq

./salmon quant --index /data/ESERCITAZIONE2025/LIBERATI3/INDEX_CORSET/SPECIE_indici --libType A -1
/data/ESERCITAZIONE2025/LIBERATI3/TRIMMED/LL87_R1_trimmed.fastq -2
/data/ESERCITAZIONE2025/LIBERATI3/TRIMMED/LL87_R2_trimmed.fastq --output
/data/ESERCITAZIONE2025/LIBERATI3/QUANTS_CORSET/LL8L7 --dumpEq

DEGs: ORFs Definizione

I file ORFs (Open Reading Frames), pep, e cds (prodotti da TransDecoder) sono file di output generati dal software durante il processo di predizione delle regioni codificanti (proteine) in trascritti.

DEGs: ORFs - Utilità

Annotazione funzionale:

Le sequenze predette (proteiche o nucleotidiche) possono essere confrontate con database pubblici (come UniProt o Pfam) per attribuire funzioni biologiche ai trascritti.

Conferma di trascritti codificanti:

Permettono di distinguere tra trascritti codificanti proteine e trascritti non codificanti.

Costruzione di database proteici personalizzati:

I file .pep possono essere usati per creare database proteici specifici da analizzare con approcci proteomici (es. spettrometria di massa).

Analisi filogenetiche e comparative:

Le sequenze CDS possono essere utilizzate per analisi evolutive e confronti tra specie.

Previsione di proteine strutturalmente importanti:

Le sequenze di proteine possono essere analizzate per studiare domini, segnali di localizzazione, o altre caratteristiche strutturali.

DEGs: ORFs - PEP

Sono file in formato FASTA che contengono le sequenze amminoacidiche (in formato a una lettera) delle proteine predette (le traduzioni degli ORFs) in un campione di trascritti.

```
>Transcript1|ORF1|123-456|+|protein_id=ORF00001
MSTAGWVLGLL... (sequenza amminoacidica)
>Transcript2|ORF2|100-500|+|protein_id=ORF00002
MKLPLLVIALV...
```

Viene usato per analisi proteomiche, confronti con database proteici (ad esempio, BLASTP), o annotazione funzionale delle proteine predette.

DEGs: ORFs - CDS

Sono file in formato FASTA che contengono le sequenze nucleotidiche (ATGC) dei geni predetti corrispondenti agli ORFs codificanti. Queste rappresentano la porzione codificante (coding sequence, CDS) all'interno dei trascritti.

```
>Transcript1|ORF1|123-456|+|cds_id=ORF00001
ATGGCTACCGT... (sequenza nucleotidica)
>Transcript2|ORF2|100-500|+|cds_id=ORF00002
ATGAAAACCTG...
```

Usato per analisi geniche, costruzione di modelli genici, e come input per strumenti che prevedono l'espressione di geni o annotano funzionalità codificanti.

DEGs ORFs (comandi trascrittoma NCBI o corset)

./TransDecoder.LongOrfs -t /data/ESERCITAZIONE2025/LIBERATI3/TRANSCRIPT/specie_corset_transcripts.fasta -O /data/ESERCITAZIONE2025/LIBERATI3/TRANSDECODER_CORSET/

./TransDecoder.Predict -t /data/ESERCITAZIONE2025/LIBERATI3/TRANSCRIPT/specie_corset_transcripts.fasta -O /data/ESERCITAZIONE2025/LIBERATI3/TRANSDECODER_CORSET/

DEGs: Fenotipo

Un fenotipo è l'insieme delle caratteristiche osservabili o misurabili di un organismo, che risultano dall'interazione tra il suo genotipo (l'insieme delle informazioni genetiche) e l'ambiente. Il fenotipo può includere caratteristiche fisiche, comportamentali, biochimiche o fisiologiche.

Componenti del fenotipo

Genotipo:

Il corredo genetico di un organismo, che fornisce le istruzioni per determinare il fenotipo.

Ambiente:

I fattori esterni (nutrizione, clima, esposizione a sostanze chimiche, ecc.) che influenzano come il genotipo si esprime.

DEGs: Fenotipo esempi

1. Caratteristiche fisiche

Colore degli occhi: Determinato da geni specifici, ma influenzabile da mutazioni genetiche o fattori ambientali.

Altezza: Influenzata dai geni, ma anche dalla nutrizione e dall'ambiente durante la crescita.

Forma e dimensione delle foglie in una pianta.

2. Caratteristiche comportamentali

Abilità di apprendimento negli animali: Dipende da geni che regolano lo sviluppo cerebrale e dall'educazione ricevuta.

Comportamento di nidificazione negli uccelli.

3. Caratteristiche biochimiche

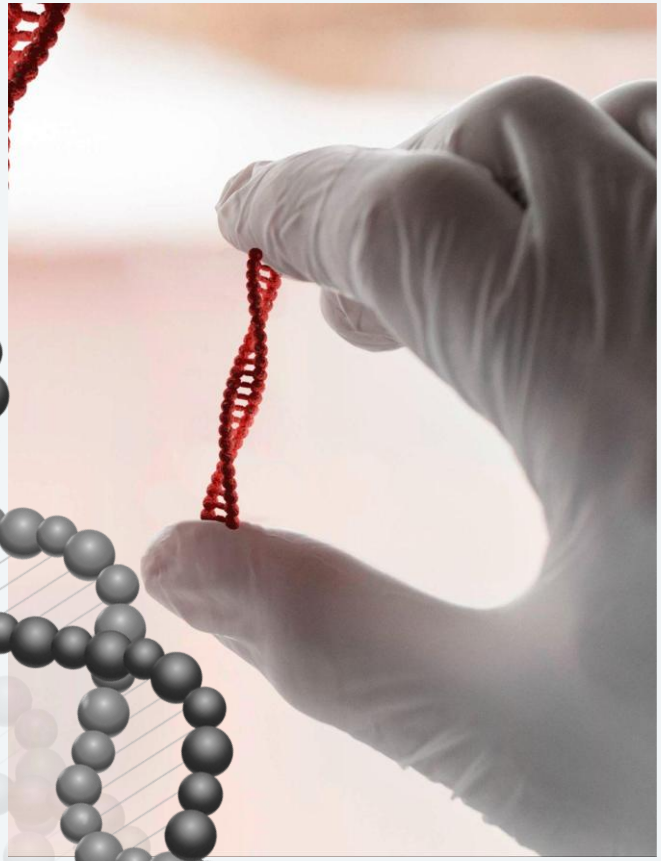
Gruppo sanguigno: A, B, AB, O. È interamente genetico e si basa sulla presenza o assenza di specifici antigeni sulle cellule del sangue.

Tolleranza al lattosio: Dipende dall'espressione del gene che codifica per l'enzima lattasi, ma può essere influenzato dalla dieta.

4. Caratteristiche fisiologiche

Tasso metabolico: Dipende da fattori genetici ma può essere modulato dalla dieta e dall'esercizio fisico.

Resistenza a malattie: Determinata dai geni del sistema immunitario e dalle condizioni ambientali.



03

IGUANER

IGUANER

IGUANER (DIfferential Gene Expression and fUnctionAl aNalyzER) è un software progettato per l'analisi integrata e aggiornata dei dati RNA-Seq provenienti da qualsiasi organismo, indipendentemente dal livello di annotazione genomica.

Il software affronta diverse limitazioni presenti negli strumenti esistenti:

- Analisi isolate: Molti strumenti supportano solo tipi specifici di analisi, complicando l'interpretazione biologica dei risultati.
- Interfacce web: Alcuni strumenti sono eseguibili solo tramite interfacce web, trascurando il parallelismo e l'efficienza offerti dai moderni supercomputer.
- Annotazioni funzionali obsolete: Alcuni strumenti si basano su annotazioni funzionali non aggiornate o supportano solo un numero limitato di organismi con annotazione genomica.
- Confronti limitati: Alcuni strumenti permettono di testare una sola comparazione tra due condizioni sperimentali per ogni esecuzione.

IGUANER

Valentina Pinna, Jessica Di Martino, Franco Liberati, Paolo Bottoni, and Tiziana Castrignanò. 2024. **IGUANER - DIfferential Gene Expression and fUnctionAl aNalyzER**. In Big Data Analytics in Astronomy, Science, and Engineering: 11th International Conference on Big Data Analytics, BDA 2023, Aizu, Japan, December 5–7, 2023, Proceedings. Springer-Verlag, Berlin, Heidelberg, 78–93. https://doi.org/10.1007/978-3-031-58502-9_5

IGUANER: Comando Analisi Differenziale

SENZA GENOMA DI RIFERIMENTO

```
python \code\src\runDGEAnalysis.py--codeFolder "path\IGUANER\code"
    -p python
    -r "C:\Program Files\R\R-4.3.1\ bin\Rscript.exe"
    --analysisFolder "path\analysis_results"
    --phenodata "path\phenodata.txt"
    --salmonFolder "path\quants_Folder"
    --transcriptome "path\transcriptome.fasta"
    --transcriptomeOrfs "path\transcriptome_orfs_transdecoder.pep"
```

IGUANER: Comando Analisi Differenziale

SENZA GENOMA DI RIFERIMENTO

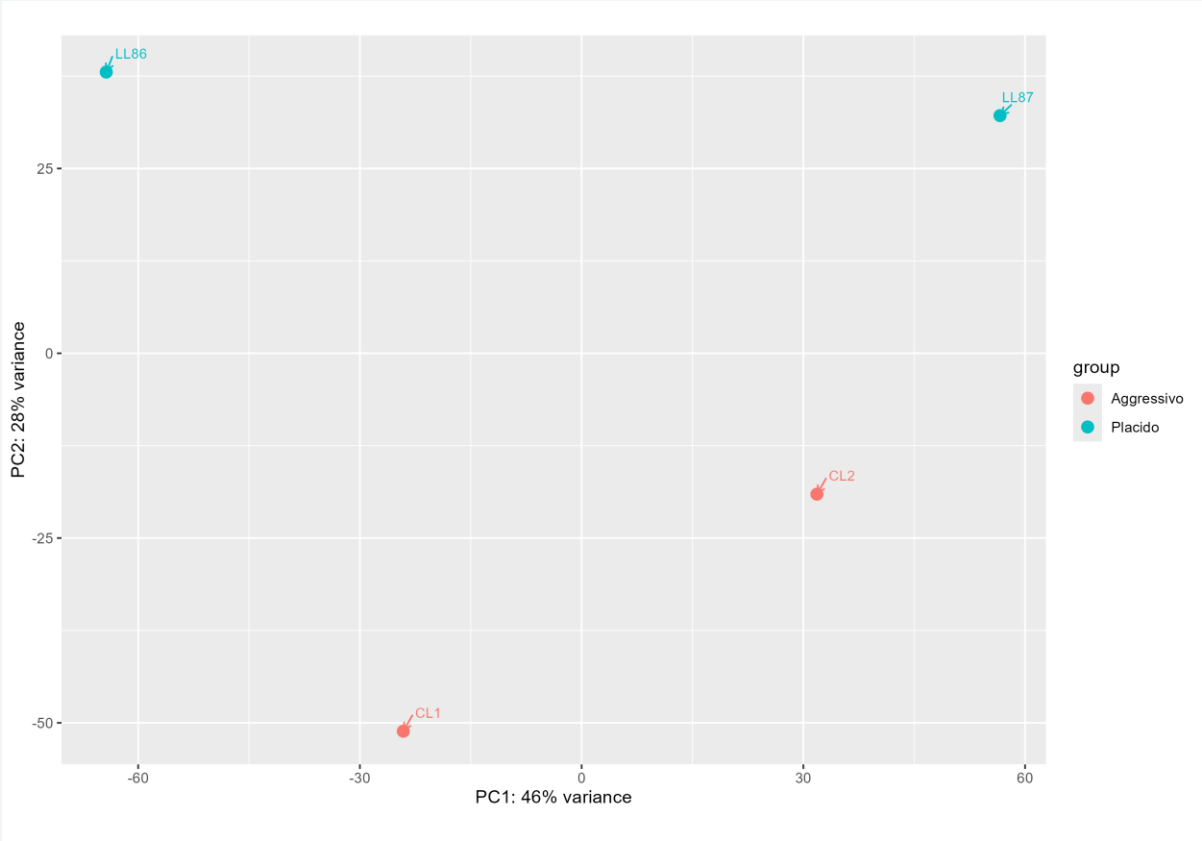
Parametri:

- codeFolder : Il path alla directory code di IGUANER.
- annotationsFolder: DirectorydeirisultatidigetFunctionalAnnotation.py.
- analysisFolder : Directory dei risultati di runDGEAnalysis.py. • organism : Nome dell'organismo.
- p: Il comando per eseguire script python (python o python3).
- r: Il comando per eseguire script R. Nel caso di Linux è solitamente RScript, mentre è il path al file RScript.exe (es. C:\Program Files\R\R-4.3.1\bin\Rscript.exe) nel caso di Windows.

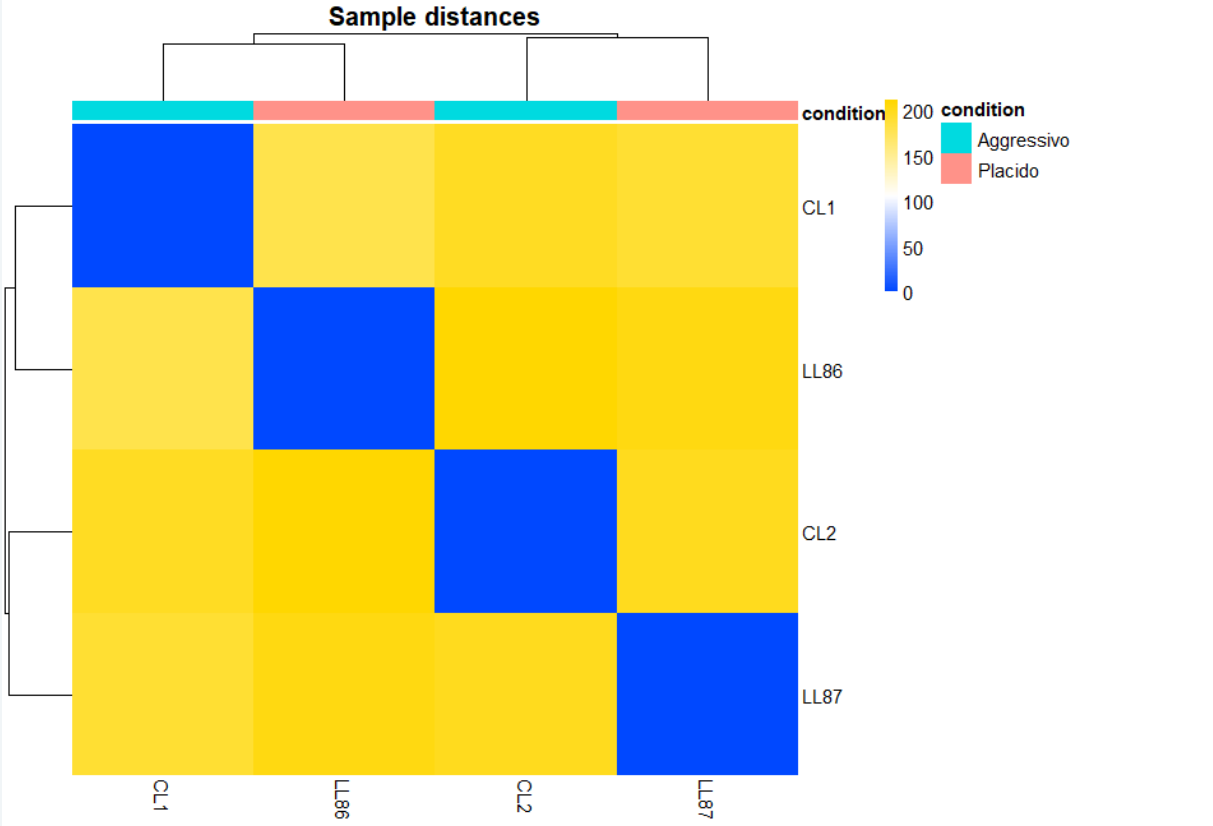
IGUANER: Fenotipo

sample	Aggressivo____Placido
CL1	Aggressivo
LL86	Placido
CL2	Aggressivo
LL87	Placido

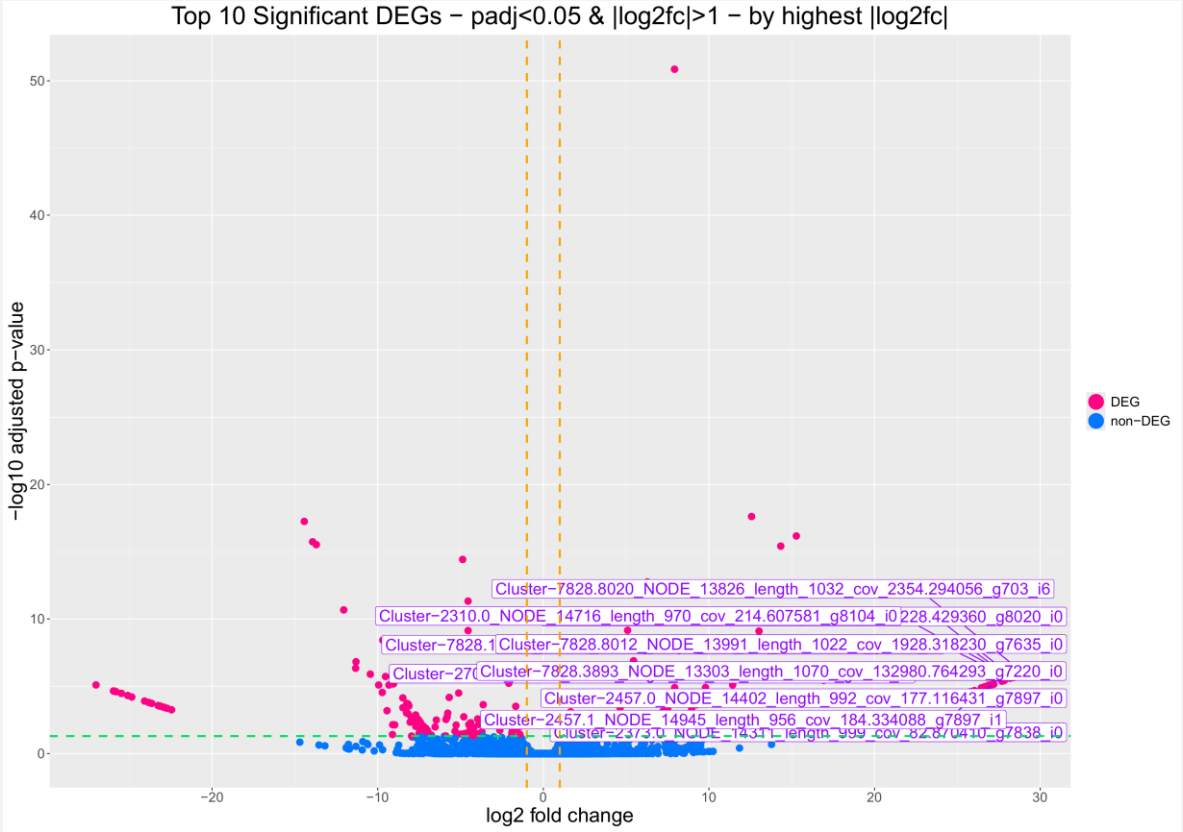
IGUANER: PCA


LL86
LL87
CL2
CL1
25
0
-25
-50
PC2: 28% variance
-60
-30
0
30
60
PC1: 46% variance
group
Aggressivo
Placido

IGUANER: Heatmap



IGUANER: Volcano



IGUANER: UP REGOLATI

	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj
Cluster-11552.0_NODE_22970_length_553_cov_11.214583_g14972_i0	12.3898819566028	8.03834078850378	2.17487877397943	3.69599486862241	0.000219027395888574	0.0255059911875462
Cluster-11692.0_NODE_19145_length_711_cov_97.192790_g11493_i0	85.2665240425568	24.1992439879487	4.78654625623362	5.05567954272531	4.28860942080276e-07	0.000114065871348105
Cluster-12090.0_NODE_27614_length_427_cov_35.338983_g19566_i0	23.3651853459474	8.95665336337767	1.90592260166858	4.69937937434416	2.60953300387864e-06	0.000545339142035046
Cluster-2282.0_NODE_14595_length_979_cov_228.429360_g8020_i0	1850.54165868949	28.407803462937	4.78482910063826	5.93705707465029	2.9018366978772e-09	2.20109687305648e-06
Cluster-2310.0_NODE_14716_length_970_cov_214.607581_g8104_i0	1803.34663652157	28.3727612511849	4.78483127632962	5.92973077055901	3.03431785601199e-09	2.21938677468305e-06
Cluster-2314.0_NODE_38961_length_293_cov_61.245455_g30909_i0	134.133220898299	24.8150425507597	4.78589304492649	5.18503909673996	2.15969554557315e-07	6.44842076766411e-05
Cluster-2334.0_NODE_16977_length_823_cov_19.437333_g9732_i0	132.891246630722	24.8027079057962	4.7859037665997	5.18245019444219	2.18989854416453e-07	6.44842076766411e-05
Cluster-2373.0_NODE_14311_length_999_cov_82.870410_g7838_i0	711.651255321529	27.0973230773605	4.78496215694755	5.66301721697347	1.48734229815738e-08	7.61519256656579e-06
Cluster-2382.0_NODE_15462_length_921_cov_67.172170_g8614_i0	527.839063720157	26.6546672784018	4.78503743761082	5.57041979836851	2.54126277222931e-08	1.08427211615117e-05
Cluster-2457.0_NODE_14402_length_992_cov_177.116431_g7897_i0	817.21906806556	27.2090565730581	4.78493421374454	5.68640139187307	1.29744244369786e-08	6.99253190708739e-06
Cluster-2457.1_NODE_14945_length_956_cov_184.334088_g7897_i1	853.236321825288	27.1699081098126	4.78492630453599	5.67822916813916	1.36096302962243e-08	7.14680072991471e-06
Cluster-2508.0_NODE_16744_length_838_cov_39.073203_g9552_i0	280.686184472366	25.7951361648618	4.7852940971769	5.39050174159196	7.02612347886398e-08	2.52447383942341e-05
Cluster-2517.0_NODE_23160_length_546_cov_19.923890_g15153_i0	115.503606884646	24.435387948108	4.78607790651779	5.10551403996817	3.29896492090887e-07	9.25517829865941e-05
Cluster-2651.0_NODE_15751_length_903_cov_61.424096_g8821_i0	445.868762060085	26.4568953390042	4.78509103169182	5.52902654594851	3.2201262019834e-08	1.29310165914941e-05
Cluster-2700.0_NODE_58124_length_233_cov_50.287500_g50071_i0	79.4863531249177	24.0968693500484	4.78668175601108	5.03414903649855	4.79976154678597e-07	0.000125642918642406
Cluster-2701.0_NODE_7814_length_1612_cov_10.172190_g4170_i0	150.278886376798	24.7962657247629	4.78576992245811	5.18124902085282	2.20405006707269e-07	6.44842076766411e-05
Cluster-2706.0_NODE_16043_length_885_cov_151.243842_g9024_i0	1131.4385577625	27.7051205108729	4.78488193520103	5.79013670265385	7.03291606144151e-09	4.23629767465653e-06
Cluster-2741.1_NODE_16992_length_822_cov_47.441923_g8832_i1	327.881206640285	26.0365140804793	4.78521521047815	5.44103304350186	5.29724864273533e-08	2.08630100390807e-05
Cluster-2811.0_NODE_21181_length_620_cov_14.914077_g13260_i0	16.5323851489537	8.46724072289751	2.10151501018931	4.02911265531945	5.59877848851146e-05	0.00807485798906441
Cluster-3026.0_NODE_21625_length_602_cov_22.385633_g13680_i0	43.0641693274956	9.79430950450462	1.76661941883396	5.54409704777798	2.95474931245926e-08	1.23496461059522e-05
Cluster-3161.0_NODE_30515_length_377_cov_12.434211_g22466_i0	16.1495489291411	8.38073839295606	2.03424483188302	4.11982779142585	3.79155671747797e-05	0.00592756347892739
Cluster-4076.0_NODE_18188_length_757_cov_17.878655_g10688_i0	22.986232576777	8.93861459519702	1.94838078737645	4.5877144001369	4.48125114562461e-06	0.000882461764061463
Cluster-4456.0_NODE_18512_length_740_cov_139.610195_g10963_i0	165.527523394356	25.1078523998186	4.7856742598807	5.24646079870144	1.55048820383409e-07	4.86741269721527e-05
Cluster-6885.0_NODE_15681_length_908_cov_160.166467_g8773_i0	178.645450170134	25.1989611208728	4.78560616206459	5.26557352768067	1.39752370488674e-07	4.47207585563756e-05
Cluster-7828.11331_NODE_18220_length_754_cov_34173.487518_g2683_i7	60.0746885463192	7.93412233631307	1.4326698104601	5.53799785434514	3.0594900930861e-08	1.25316714212807e-05
Cluster-7828.1145_NODE_2269_length_2922_cov_499.254826_g717_i1	242.650268763306	2.91835538167464	0.764737589684042	3.81615265293862	0.000135548663505054	0.0168244644156576

IGUANER: DOWN REGOLATI

	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj
Cluster-10580.0_NODE_1733_length_3197_cov_109.885723_g451_i1	18.5153767219086	-6.76456233223968	1.91445389006044	-3.53341617019887	0.00041022616204584	0.0421060640347541
Cluster-10650.0_NODE_19481_length_695_cov_112.929260_g11773_i0	34.2119879064779	-3.70731401196308	0.997935568722325	-3.71498333976573	0.000203217083092411	0.0241439866775809
Cluster-11404.0_NODE_19831_length_680_cov_69.920923_g12070_i0	52.6858591760992	-4.80356668849804	1.088638235582	-4.41245450645963	1.0220527503142e-05	0.00185888630359512
Cluster-11526.0_NODE_18494_length_741_cov_50.803892_g10947_i0	38.0263935944245	-4.17634264801774	1.17247489284476	-3.56198898032243	0.000368055836219767	0.0384580792131675
Cluster-11647.0_NODE_5036_length_2063_cov_109.485930_g2759_i0	29.2482844926186	-7.42310957938902	1.77941862754589	-4.1716487983644	3.02403480782434e-05	0.00483845569251895
Cluster-11821.0_NODE_1131_length_3632_cov_87.162124_g628_i0	28.255490236015	-7.37578764023174	1.92921644986273	-3.82320378864514	0.000131728818218589	0.0165509582645197
Cluster-1222.0_NODE_17913_length_772_cov_132.896996_g10468_i0	108.99189948008	-9.32083262834815	1.65447351147508	-5.63371523551196	1.76368156159806e-08	8.06133845858429e-06
Cluster-1429.0_NODE_14542_length_982_cov_178.547855_g7985_i0	167.656411631129	-9.94234040105471	1.76564340147245	-5.63100136344823	1.7916631079416e-08	8.06133845858429e-06
Cluster-1614.0_NODE_14199_length_1007_cov_593.659529_g7762_i0	839.619942647531	-25.4821796623487	4.78477042379802	-5.3256849138692	1.00573341381209e-07	3.3766262811265e-05
Cluster-2205.0_NODE_22260_length_578_cov_217.108911_g14286_i0	157.317582167542	-23.1699952390166	4.78491973793036	-4.84229548415338	1.28347710953879e-06	0.000292062346703937
Cluster-3307.0_NODE_26063_length_462_cov_142.735219_g18015_i0	123.863306293088	-3.63132980109249	0.741865308845052	-4.89486401075393	9.83737300162498e-07	0.000234266743108465
Cluster-6500.0_NODE_11048_length_1262_cov_23.878049_g5875_i0	33.9237057324145	-6.46116364816451	1.51145459632165	-4.27479837230223	1.91310346453682e-05	0.00321150483227164
Cluster-6504.0_NODE_5861_length_1893_cov_53.521978_g3173_i0	30.7800842650382	-7.49924455530756	2.04979062851866	-3.65854173151679	0.000253654442156842	0.0290214691361571
Cluster-6576.0_NODE_6799_length_1754_cov_39.827484_g3642_i0	31.5541089824462	-7.53179497888948	1.84509705057875	-4.08205897707494	4.46384648331212e-05	0.00686725952552105
Cluster-6635.0_NODE_8745_length_1490_cov_43.179252_g4483_i1	18.3737658935601	-6.75224211486675	1.933191171286	-3.49279585752247	0.000477991770767037	0.0485715234971969
Cluster-722.0_NODE_8585_length_1509_cov_32.378134_g4562_i0	51.5399634302737	-8.24054944651165	1.66602528800979	-4.94623311291734	7.56633433632644e-07	0.000186697020732489
Cluster-7771.0_NODE_2509_length_2805_cov_283.389092_g1399_i0	35.6775788196693	-7.71064296188904	1.71821643833269	-4.48758537624703	7.2034948042185e-06	0.00135346397789353
Cluster-7828.10108_NODE_13251_length_1074_cov_2804.535465_g4246_i2	95.9685942733392	-4.49523876820378	1.10253566708989	-4.07718217413213	4.55847770535224e-05	0.00691537951152695
Cluster-7828.10543_NODE_8434_length_1526_cov_291.095664_g4488_i0	66.9906126522647	-2.62527344056517	0.692424179919492	-3.79142369186243	0.000149786171888894	0.0183689868280512
Cluster-7828.10676_NODE_10224_length_1338_cov_12172.365217_g2324_i4	1617.02516571606	-1.66494987698239	0.344908956074103	-4.8272155525723	1.38455238734645e-06	0.000308213401009296
Cluster-7828.10682_NODE_3987_length_2301_cov_205.911131_g2197_i0	23.6115686202734	-7.11508191915299	1.82650404320429	-3.89546464220839	9.80106597597131e-05	0.0129500536250253
Cluster-7828.10686_NODE_7503_length_1653_cov_377.526582_g4009_i1	250.475572048735	-1.87312966208841	0.500296406826905	-3.74403980625926	0.000181084945076908	0.021815409853971
Cluster-7828.10687_NODE_1216_length_3563_cov_23.667622_g675_i0	36.9318973101189	-4.19394447211551	1.06685204499013	-3.93113974126966	8.4544110971272e-05	0.011316754200599
Cluster-7828.108_NODE_2186_length_2969_cov_168.373964_g1217_i0	47.371988237458	-4.21067830205491	1.14835028546395	-3.66671942816972	0.000245682105229314	0.0284269464129737
Cluster-7828.10914_NODE_763_length_4032_cov_309.685527_g421_i0	33.8583850917185	-4.38805222894151	1.21152292104993	-3.62193083820381	0.000292412307910429	0.0323708327892193

IGUANER: heatmap Up-Down REGOLATI

Grazie!

Domande?

franco.liberati@unitus.it
deb.scienceontheweb.com