Ghid de proceduri bioinformatice de analizã comparativa a amestecurilor bacteriene complexe folosind MEGAN6

MEGAN6 foloseste o abordare prin care se realizeaza analiza taxonomica și funcționala a secvențelor microbiomului se bazează pe omologia proteinelor. În această abordare, secvențele sunt mai întâi aliniate la o bază de date de referință a secvențelor de proteine cu identitate taxonomică și funcțională cunoscută, iar apoi secventele aliniate rezultate sunt utilizate pentru a atribui secvențele în unitati taxonomice și funcționale.

Analiza de bază "DIAMOND + MEGAN" a seturilor de date de secvențiere a microbiomului (atât citiri scurte, cât și citiri lungi asamblate) constă în trei etape ulterioare, și anume:

1. Alinierea tuturor citirilor cu o bază de date de referință proteică folosind DIAMOND

2. Analiza taxonomică și funcțională a alinierii rezultate, susținând un program denumit MEGANIZER (part a pachetului MEGAN), sau MEGAN

3. Explorarea si analiză interactivă a rezlutatelor folosind MEGAN

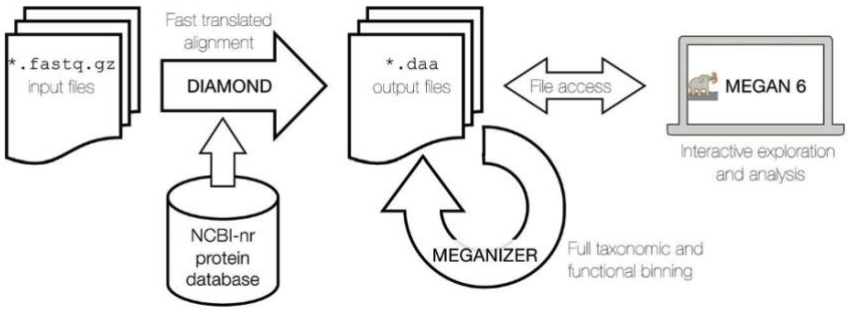

Figura 1 Conducta de analiză a microbiomului Core Diamond + MEGAN. Pe un server, fișierele de intrare comprimate sunt furnizate programului DIAMOND ca intrare, care le compară cu baza de date de referință a proteinelor NCBI-nr. Aliniamentele rezultate sunt scrise în fișiere de ieșire în format DAA (sufix .daa). Instrumentul MEGAN cu tool-ul MEGA-NIZER este aplicat tuturor fișierelor DAA, astfel încât să efectueze legarea taxonomică și funcțională completă a citirilor pe baza alinierilor lor. Fișierele DAA "meganizate" rezultate pot fi apoi explorate și analizate interactiv folosind MEGAN.

Hardware necesar

Analiza computationala initiala a acestor seturi de date ar trebui efectuata pe un server cu un numar bun de nuclee, ~24 sau mai mult, cel putin 64 GB de memorie principala si spatiu suficient pe disc (TB multipli). Odată ce toate probele au fost procesate, fișierele rezultate pot fi descărcate pe desktop sau laptop pentru explorare și analiză interactivă. Acest computer ar trebui să fie un model recent cu un procesor rapid, cel puțin 16 GB de memorie și un disc SSD mare.

Instalarea software-ului

- DIAMOND poate fi obținut de la http://www.diamondsearch.org.; poate fi, de asemenea, instalat în conda, folosind comannda `conda install -c bioconda diamond`.
- MEGAN pot fi obținuți de la http://megan.husonlab.org.
- Pentru procesarea citirilor lungi, este nevoie de instalarea instrumentului de asamblare a citirilor lungi Unicicler, care utilizează miniasm și Racon pe un server. Programul poate fi instalat local pe un server folosind comenzile:

```
git clone https://github.com/rrwick/Unicycler.git
cd Unicycler
make
```

- Programul este apoi lansat tastand: `unicycler`
- Alternativ, Unicycler poate fi instalat și în conda, folosind comanda:
  `conda install -c bioconda unicycler`
- În plus, se va folosi instrumentul de corecție a asamblării citirilor lungi medaka (ONT, 2020), care poate fi instalat în Conda folosind comanda:
  `conda install -c bioconda medaka`

Analiza propriu-zisa

1. Construirea unui indice DIAMOND

În pregătirea utilizării DIAMOND pentru alinierea secvențelor, trebuie mai întâi să construiți un indice DIAMOND.

Din linia de comandă, descărcați cea mai recentă bază de date NCBI-nr după cum urmează:

```
wget https://ftp.ncbi.nlm.nih.gov/blast/db/FASTA/nr.gz
```

Aceasta va crea un fișier `nr.dmnd` care conține indicele DIAMOND.

Parametrii furnizați aici către DIAMOND sunt comanda *`makedb`* care solicită construirea unui index, urmată de *`--in nr.gz`*, specificând fișierul de intrare și *`-db nr`*, specificând numele fișierului index (sau bază de date), în acest caz *`nr`* (fișierul rezultat se va termina pe *suffix* **.** *`dmnd`*).

2. Asamblarea

Unicycler este un pipeline de asamblare a genomului bacterian care poate rula pe seturi de date numai cu citire lungă, precum și pe seturi de date hibride (citire lungă și citire scurtă). Aici descriem utilizarea modului numai pentru citiri lungi. În modul long-reads-only, conducta Unicycler constă din minimap2 și miniasm pentru asamblare, racon pentru construirea consensului și tBLASTn pentru detectarea originii replicării în genomul circular.

Unicycler necesită doar doi parametri în modul de citire lungă: `-l` pentru citirile de intrare (fișier fastq, poate fi gzipped) și `-o` pentru directorul de ieșire. Parametrii suplimentari pe care ii folosim sunt `-t` pentru setarea numărului de fire CPU și `--keep 3` pentru a păstra fișierele intermediare generate de conductă, care pot fi utile pentru inspectarea ulterioară a ansamblului. Se execută urmatoarea comanda:

`Unicycler -l XXX.fastq.gz -o unicycler_asm -t` ***(threads)*** `--keep 3`

Unicycler plasează toate ieșirile în directorul de ieșire specificat, care este `unicycler_asm` în comanda exemplu. Asamblarea finală este scrisă într-un fișier numit `assembly.fasta.` Alte fișiere produse includ `assembly.gfa, care conține graficul de asamblare și` 001_string_graph.gfa, care conține graficul string brut. Astfel de fișiere GFA pot fi vizualizate și explorate folosind `programul Bandage`

3. Corectarea erorilor

Racon, un algoritm de consens rapid, este unul dintre instrumentele de corectare a erorilor utilizate în mod obișnuit după asamblarea seturilor de date cu citire lungă predispuse la erori. Unicycler, pipeline-ul de asamblare folosită aici, efectuează deja trei runde de corectare a erorilor de către Racon. Pentru a îmbunătăți în continuare calitatea secvențelor, aplicăm instrumentul medaka, care este un algoritm de corecție bazat pe rețele neuronale conceput pentru secvențe obținute folosind un dispozitiv ONT.

Executabilul `medaka_consensus`, care rulează pipeline-ul medaka, ia ca intrare proiectul de asamblare (`-d`) (assembly.fasta `în directory-ul de ieșire Unicycler),` `citirile brute (-i)` și un nume de model care descrie atât celula de flux utilizată, cât și versiunea algoritmului de apelare de bază (-m). Este nevoie de specificarea unui director de ieșire (`-o`). Software-ul se execută după cum urmează:

`medaka_consensus -i XXX.fastq.gz -d assembly.fasta -o unicycler_medaka -t`***(threads)*** `-m r941_min_high_g330`

Fișierul in care vor fi exportate rezultatele asamblarii Medaka este `consensus.fasta` în directorul de ieșire specificat.

4. Comparație DIAMOND

În ciuda faptului că au fost corectate de mai multe ori cu mai multe metode, ansamblurile numai pentru citire lungă conțin încă erori, mai ales inserții și deletii de baze unice sau de câteva baze. Acest lucru cauzează probleme pentru algoritmii de aliniere traduși, cum ar fi BLAST×, deoarece inserțiile și ștergerile eronate cauzează deplasări de cadre care întrerup alinierile, astfel se poate folosi DIAMOND care realizeza un mod de aliniere conștient de deplasarea cadrelor, care este activată prin specificarea unei penalizări de deplasare a cadrelor, folosind opțiunea `-F`. Folosim o penalizare de 15, care pare să atingă un echilibru bun între producerea aliniamentelor lungi fără comutarea excesivă a cadrelor.

Folosind fișierul final de asamblare ca fișier de intrare (`consensus.fasta` în directorul de ieșire medaka), rulăm DIAMOND:

```
Diamond BlastX -q consens.fasta -d nr.dmnd -o
XXX_unicycler.daa -F 15 -f 100
 --range-culling --top 10 -p (fire)
```

Fișierul de ieșire trebuie să aibă extensia de fișier `.daa`.

Un fișier DAA produs de DIAMOND poate fi meganizat fie folosind instrumentul de linie de comandă `daa-meganizer, fie` folosind dialogul de meganize al MEGAN

Pentru a meganiza fișierul DAA din secțiunea anterioară, utilizați:

```
daa-meganizer -i XXX_unicycler.daa -mdb
 megan-map-Nov2023.db --longReads
```

Fișierele Meganized DAA pot fi inspectate și analizate.

5. Inspectarea citirilor lungi

MEGAN oferă un *inspector de citiri lungi* care poate fi folosit pentru a explora aliniamentele la proteinele de referință de-a lungul unei citiri lungi sau contig. Pentru a deschide acest vizualizator pentru un anumit taxon sau clasă funcțională, selectați nodul din vizualizatorul corespunzător și alegeți `Inspect Long Reads...`

Inspectorul de citire lungă afișează fiecare secvență atribuită unui nod dat într-un rând separat, listând numele secvenței, lungimea, atribuirea taxonomică, acoperirea procentuală și numărul de alinieri, împreună cu o vizualizare generală a secvenței și a alinierilor sale la proteinele de referință. Aliniamentele sunt reprezentate de săgeți în direcția de translație.

6. Comparatie fisiere DAA

Majoritatea proiectelor de microbiom implică mai multe probe, iar acestea trebuie analizate într-o manieră comparativă. Pentru a aborda acest lucru, MEGAN utilizează conceptul de *comparison document* care conține rezultatul analizei taxonomice și funcționale a mai multor eșantioane individuale. Se poate folosi dinn bara de optiuni File → Compare...

Diversitatea alfa, sau diversitatea în cadrul eșantionului, poate fi calculată în MEGAN selectând fie elementul de meniu Options → Shannon-Weaver Index, fie Op- tions→Simpson-Reciprocal Indexmenu. Oricare măsură va fi calculată pe setul de noduri selectate în prezent, iar rezultatele vor fi scrise în fereastra mesajului.

Diversitatea beta sau diversitatea dintre eșantioane poate fi calculată utilizând o serie de măsuri diferite. Pentru a calcula astfel de măsuri pe un document de comparație în MEGAN, deschideți vizualizatorul de analiză cluster utilizând Window → Cluster analysis... element de meniu.

MEGAN oferă implementări ale indicelui ecologic Bray-Curtis, divergenței Jensen-Shannon, distanțelor euclidiene și o serie de alte măsuri, care pot fi selectate din meniul Opțiuni al vizualizatorului de analiză cluster.

Figura 2. Ploturi de diversitate taxonomică. (A) O diagramă care afișează procentajul atribuit rangului taxonomic al clasei. (B) Diversitatea alfa calculată pentru rangul speciilor. (C) PCoA trasează rangul speciilor, utilizând distanțele Bray-Curtis. (D) O aaliza de corelatie care corelează vârsta, înălțimea și greutatea subiectului cu atribuirile la rangul taxonomic al clasei.

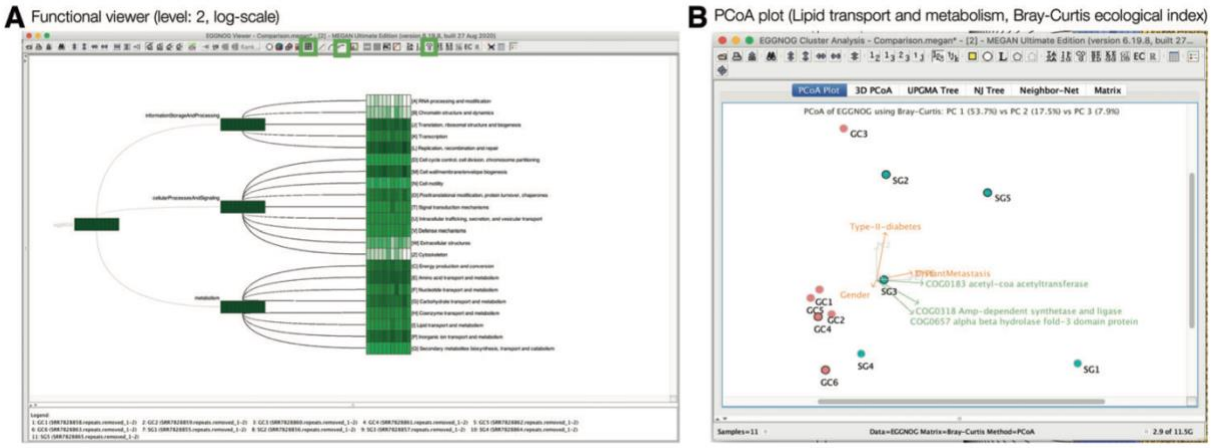

Figura 3. Comparație funcțională (A) Un vizualizator funcțional (eggNOG) pentru un document de comparație, folosind hărți termice pentru a indica diferite eșantioane. (B) Grafic PCoA pentru un vizualizator funcțional (eggNOG), folosind distanțele Bray-Curtis.

5. Exportarea datelor

MEGAN permite utilizatorului să exporte toate citirile în fișiere taxonomice sau funcționale specifice clasei. Pentru a utiliza această caracteristică, selectați toate nodurile de interes într-un vizualizator taxonomic sau funcțional și apoi `File → Extract Reads...` element de meniu pentru a deschide un dialog nou pentru a specifica numele fișierului de utilizat pentru ieșire.

Ambele proceduri de export descrise mai sus pot fi efectuate și pe linia de comandă, folosind programul `read-extractor` care poate fi găsit în subdirectorul `de instrumente` MEGAN. Programul se execută astfel:

```
read-extractor -i XXX_unicycler.daa -o %t_%i.fasta -cclassification-fsc
```

Aici, *clasificarea* poate fi fie taxonomie, fie GTDB, iar fișierele de ieșire vor fi scrise în directorul curent cu formatul descris mai sus - folosind substituenții. Ieșirea poate fi, de asemenea, scrisă într-un director, cum ar fi `-o bins/%t_%i.fasta.gz`. Opțiunea `-gz` poate fi utilizată pentru a comprima fișierele de ieșire.