Ghid de proceduri bioinformatice de utilizare a suitelor specifice platformei Epi2me Oxford Nanopore Technologies

Workflow-ul wf-basecalling

Workflow-ul utilizează `nextflow` pentru a gestiona resursele de calcul și software, motiv pentru care este necesar a fi instalat înainte de a merge mai departe la următorii pași. Pentru a cumula toate software-urile necesare pentru rulare într-un singur loc, workflow-ul se folosește în asociere cu `Docker` sau `Singularity`. Basecaller-ul standard folosit de către acest workflow este `Dorado`. Nu este necesară descărcarea depozitului `git` pentru a rula workflow-ul.

După instalarea `nextflow`, workflow-ul wf-basecalling se inițializează astfel:

```
nextflow run epi2me-labs/wf-basecalling –help
```

Există mai multe opțiuni disponibile de rulare, de exemplu:

```
nextflow run epi2me-labs/wf-basecalling \
    -profile singularity
    --input /path/to/my/fast5 \
    --dorado_ext fast5 \
    --ref /path/to/my/ref.fa \
    --out_dir /path/to/my/outputs \
    --basecaller_cfg "dna_r10.4.1_e8.2_400bps_hac@v4.0.0" \
    --basecaller_basemod_threads 2 \
    --remora_cfg dna_r10.4.1_e8.2_400bps_hac@v4.0.0_5mCG@v2
```

Se recomandă actualizarea periodică a acestui work-flow, folosind următoarea comandă:

```
nextflow pull epi2me-labs/wf-basecalling
```

Rezultatele principale ale workflow-ul wf-basecalling sunt două fișiere sortate și indexate de tip CRAM, aliniate la referința furnizată, având citirile separate de scorul lor de calitate.

- `<sample_name>.pass.cram` – conține citiri cu qscore >= pragul stabilit
- `<nume_probă>.fail.cram` – conține citiri cu qscore < pragul stabilit

Workflow-ul wf-basecalling acceptă duplex basecalling, prin activarea opțiunii --duplex. Această optiune se poate folosi dacă s-a folosit o chimie și flowcell-uri care acceptă citirile duplex. Există câteva avertismente la apelurile duplex, și anume:

- nu poate fi utilizat pentru citirea bazelor azotate modificate
- necesită convertirea datelor într-un format de tip POD5

Dacă apelarea duplex este activată, workflow-ul va genera în schimb patru CRAM-uri:

- `<nume_probă>.pass.duplex.cram` conține citiri duplex + citiri simplex care nu aparțin unei perechi cu qscore >= pragul stabilit

- `<nume_probă>.fail.duplex.cram` conține citiri duplex + citiri simplex care nu aparțin unei perechi cu qscore < pragul stabilit
- `<nume_probă>.pass.simplex.cram` conține citiri simplex aparținând unei perechi cu qscore >= pragul stabilit
- `<nume_probă>.fail.simplex.cram` conține citiri simplex aparținând unei perechi cu qscore < pragul stabilit

Basecalling-ul cu Dorado necesită un GPU NVIDIA cu arhitectura Pascal sau mai nouă, și cel puțin 8 GB de vRAM.

În mod implicit, workflow-urile sunt configurate pentru a rula sarcini GPU în serie, însemnând că o singură sarcină de apelare de bază va fi rulată la un moment dat. Acest lucru este implementt astfel pentru a preveni ca GPU-ul să rămână fără memorie la execuția locală.

Workflow-ul wf-alignment

Acest workflow oferă o modalitate ușoară de a alinia citirile Oxford Nanopore și de a aduna statistici de mapping fie local (pentru cantități mici de date), fie la scară largă (într-un mediu distribuit, cum ar fi un cluster).

După instalarea `nextflow`, workflow-ul wf-alignment se inițializează astfel:

```
nextflow run epi2me-labs/wf-alignment –help
```

Rezultatele principale ale fluxului de lucru includ:
- Un fișier BAM sortat, indexat, care conține alinierile
- Un CSV care conține diverse statistici de mapping
- Un raport HTML cu vizualizări ale statisticilor de mapping

Workflow-ul wf-bacterial-genomes

Dacă nu este inclusă nicio referință, asamblarea va fi realizată folosind `flye` și îmbunătățită cu `medaka`. Dacă este furnizată o referință, alinierea se va face cu `mini_align` și numirea variantei se va realiza folosind medaka. Workflow-ul are câteva opțiuni suplimentare: poate rula `prokka` pentru a adnota secvența de consens rezultată și `ResFinder` pentru a analiza secvența folosind o bază de date cu gene de rezistență la antibiotice.

După instalarea `nextflow`, workflow-ul wf-bacterial genomes se inițializează astfel:

```
nextflow run epi2me-labs/wf-bacterial-genomes –help
```

Rezultatele principale ale fluxului de lucru includ:
- o secvență consens FASTA construită dintr-o secvență de referință furnizată
- un fișier VCF care conține variante din probă comparate cu referința (dacă aceasta este furnizată)
- un raport HTML care detaliază valorile QC
- (opțional) o adnotare a secvenței consens folosind prokka
- (opțional) un director de ieșire ResFinder per eșantion cu rezultate diverse

Workflow-ul wf-metagenomics

wf-metagenomics este un flux de lucru Nextflow pentru identificarea originii citirilor unice atât din secvențierea metagenomică vizată de amplicon, cât și din secvențierea metagenomică. Fluxul de lucru are două moduri de operare, poate folosi fie `kraken2` sau `minimap2` pentru a determina originea citirilor.

Modul `kraken2` poate fi utilizat în timp real, permițând fluxului de lucru să ruleze continuu alături de o rulare de secvențiere în curs de desfășurare, deoarece datele citite sunt produse de instrumentul de secvențiere al Oxford Nanopore Technologies. Utilizatorul poate vizualiza clasificarea citirilor și abundența speciilor într-un raport de actualizare în timp real.

Se pot folosi doua abordari:

- Kraken2 - Implicit

`Kraken2` este utilizat împreună cu `Kraken2-server` pentru a oferi cea mai rapidă metodă de clasificare a citirilor.

`Braken` este apoi folosit pentru a oferi o estimare bună a abundenței la nivel de specii din eșantion, care poate fi vizualizată în raport. Modul de flux de lucru Kraken2 poate fi rulat în timp real.

- Minimap2

`Minimap2` oferă cea mai bună analiză de rezoluție, dar, în funcție de baza de date de referință utilizată, în detrimentul unui timp de calcul semnificativ mai mare. În prezent, modul minimap2 nu acceptă în timp real.

Fluxul de lucru wf-metagenomics utilizează în mod implicit baza de date NCBI 16S + 18S rRNA, care va fi descărcată la începutul unei analize, există opțiuni extinse ale bazei de date metagenomice disponibile cu parametrul `--database_set`, dar fluxul de lucru nu este legat de această bază de date și poate poate fi folosit și cu baze de date personalizate, după cum este necesar.

După instalarea `nextflow`, workflow-ul wf-bacterial genomes se inițializează astfel:

```
nextflow run epi2me-labs/wf-metagenomics --help
```

Rezultatele principale ale fluxului de lucru includ:

Rezultatele pipeline-ului wf-metagenomics este `wf-metagenomics-report.html` care pot fi găsite în directorul de ieșire. Conține un rezumat al statisticilor citite, compoziția taxonomică a comunității și unele metrici de diversitate. De asemenea, se poate utiliza `--abundance_threshold` pentru a elimina din tabelul de abundență toți taxonii aflati sub limita inferioara. În plus, `--n_taxa_barplot` controlează numărul de taxoni afișați în graficul cu bare și îi grupează pe restul în categoria „Altele”.

Există și alte foldere în folderul de ieșire care conțin alte fișiere de ieșire din pipeline, cum ar fi rapoartele kraken și bracken. În plus, „species-abundance.tsv” este un tabel cu numărul

diferiților taxoni per eșantion. Puteți utiliza marcajul `--include_kraken2_assignments` pentru a include un fișier TSV per eșantion care indică modul în care a fost clasificată fiecare secvență de intrare, precum și taxonul care a fost atribuit fiecărei citiri. Acest fișier TSV va fi scos numai la finalizarea fluxului de lucru și, prin urmare, deloc dacă se utilizează opțiunea în timp real în timp ce rulează pe termen nelimitat. Această opțiune este disponibilă în pipeline-ul kraken2.

Workflow-ul wf-transcriptomes

Acest flux de lucru identifică izoformele de ARN folosind citirile Oxford Nanopore fie ADNc, fie ARN direct (ARNd).

- Preprocesare

Citirile cADN sunt inițial preprocesate de `pychopper` pentru identificarea citirilor de lungime completă, precum și pentru tăierea și corectarea orientării (Acest pas este omis pentru citirile directe de ARN).

- Asamblare transcripte

Abordarea asamblarii transcripției asistate de referințe
- Citirile de lungime completă sunt mapate la un genom de referință furnizat folosind `minimap2`
- Transcriptele sunt asamblate de `stringtie` în modul de citire lungă (cu sau fără o anotare de referință de ghid) pentru a genera adnotarea GFF.
- Anotarea generată de pipeline este comparată cu anotarea de referință folosind `gffcompare`

- Detectarea genelor de fuziune

Detectarea genei de fuziune se realizează folosind `JAFFA`, cu extensia JAFFAL pentru utilizare cu citiri lungi ONT.

- Analiza expresiei diferențiale

Analizele privind expresia diferenţială a genelor (DGE) şi utilizarea diferenţială a transcripţiei (DTU) urmăresc identificarea genelor şi/sau a transcriptelor care prezintă modele de expresie modificate statistic într-un sistem biologic studiat. Rezultatele analizelor diferențiale sunt prezentate într-un format cantitativ și, prin urmare, gradul de schimbare (reglare în sus sau în jos) între condițiile experimentale poate fi calculat pentru fiecare genă identificată.

După instalarea `nextflow`, workflow-ul wf-bacterial genomes se inițializează astfel:

```
nextflow run epi2me-labs/wf-transcriptomes --help
```

Exemplu de execuție a unui flux de lucru pentru asamblarea transcripției bazate pe referințe și detectarea fuziunii:

```
OUTPUT=~/output;
nextflow run epi2me-labs/wf-transcriptomes \
  --fastq ERR6053095_chr20.fastq \
  --ref_genome chr20/hg38_chr20.fa \
```

```
--ref_annotation chr20/gencode.v22.annotation.chr20.gtf \
--jaffal_refBase chr20/ \
--jaffal_genome hg38_chr20 \
--jaffal_annotation "genCode22" \
--out_dir outdir -w workspace_dir
```

Rezultatele principale ale fluxului de lucru includ:
Un document de raport HTML care detaliază principalele constatări ale fluxului de lucru. Pentru fiecare mostră:

- `gffcompare` directoarele de ieșire
- read_aln_stats.tsv - statistici rezumate de aliniere
- transcriptome.fas - transcriptomul asamblat
- merged_transcritptome.fas - transcriptom adnotat, asamblat
- `jaffal` directoare de ieșire

Workflow-ul wf-tb-amr

wf-tb-amr este un flux de lucru pentru determinarea rezistenței la antibiotice a probelor de secvențiere vizate de *Mycobacterium tuberculosis*. Fluxul de lucru gestionează secvențierea multiplexată și oferă rapoarte clare și simple care rezumă profilul de rezistență prezis al fiecărei probe în funcție de variantele genetice descoperite.

Fluxul de lucru poate fi rulat în prezent folosind fie `Docker` sau `Singularity` https://docs.sylabs.io/guides/latest/user-guide/ pentru a asigura izolarea software-ului necesar. Ambele metode sunt automatizate din momentul în care sunt instalate fie docker, fie singularity.

După instalarea `nextflow`, workflow-ul wf-bacterial genomes se inițializează astfel:

```
nextflow run epi2me-labs/wf-tb-amr --help
```

Rezultatele principale ale fluxului de lucru includ:

- un document HTML care detaliază valorile QC și principalele constatări ale fluxului de lucru.
- un fișier CSV care conține o versiune a rezultatelor care poate fi citită de mașină.
- Un fișier VCF pentru fiecare probă.
- Un fișier BAM pentru fiecare probă.