Bpexa/b project Bio-informatica Hogeschool Leiden 2021

Snakemake Implementatie van een Meta-transcriptomic Pijplijn

Begeleiding - Evelien Jongepier (bio-informaticus; e.jongepier@uva.nl).

Biologische achtergrond - Meta-transcriptomics is de studie van gen expressie van complete microbiële gemeenschappen. Waar meta-genomics gericht is op het characteriseren van de microbiële diversiteit in een bepaald milieu, kijkt meta-transciptomics juist naar de diversiteit aan actieve genen. Het voordeel is dat meta-trancriptomics inzichten kan verschaffen in verschillen in actieve functies tussen milieus of experimentele condities. Het gebruik van meta-transciptomics technieken wordt steeds populairder binnen het Instituut voor Biodiversiteit en Ecosysteem Dynamiek aan de Universiteit van Amsterdam. Om dit onderzoek te faciliteren is er vraag naar een gebruikersvriendelijke, flexibele, open source meta-transcriptomics pijplijn.

Project doelen - In dit project werken de studenten aan een `Snakemake -use-conda` implementatie van een meta-transcriptomic workflow. De pipeline zal bestaan uit enkele kern onderdelen waar de studenten gezamenlijk aan werken met behulp van `github`. Daarnaast zijn er enkele uitbreidingsmogelijkheden waar de studenten elk individueel aan werken binnen de gezamenlijke pijplijn. Hierbij is het belangrijk de eindgebruiker flexibiliteit te bieden in welke optionele modules worden gebruikt.

Beschikbare middelen - De meta-transcriptomics workflow is al getest en een walkthrough met alle commands is voorhanden, alleen de `Snakemake` implementatie ontbreekt. Een test dataset zal worden verstrekt.

Pijplijn beschrijving - De kern pijplijn zal de volgende onderdelen bevatten:

- Qualiteitsfiltering en -trimming van ruwe RNA-seq paired-end data met behulp van `Trimomatic`.
- Meta-transcriptome assemblering met behulp van `TRINITY`.
- Transcript alignment met `RSEM` en `bowtie2` geïmplementeerd in `TRINITY`.
- Differentiële gen expressie analyze met behulp van `TRINITY`.

Optioneel is dit project uit te breiden met de volgende individuele taken:

- Identificatie en filtering van ribosomale reads met behulp van `SortMeRNA`.
- Transcript functionele annotatie met behulp van `Diamond BLAST`.
- Transcript taxonomische classificatie met behulp van `MEGAN`.
- Visualisatie van de resultaten middels bijv. kronas/heatmaps/pcas.

Supervisie - Studenten werken vanaf september, gedurende 20 weken, voor 1 dag per week aan de pijplijn. De studenten zullen wekelijks (via Zoom) met de begeleider afspreken om de vorderingen, volgende stappen en uitdagingen te bespreken. Voorafgaand aan deze meeting zullen de studenten een agenda opstellen en delen met daarin de punten die besproken moeten worden, inclusief relevante resultaten.

Infrastructuur - De studenten werken vanuit de Hogeschool Leiden aan hun eigen laptop. Mocht het noodzakelijk zijn bepaalde stappen op een computational cluster te runnen dan zal dat samen met de begeleider gedaan worden.