SevenBridges

RNA-Seq kvantifikacija

Kroz ovu vezbu isprobaćete neke od dostupnih Common Workflow Language (CWL) alata na CGC platformi. Za početak je neophodno imati CGC nalog - otvorite ga koristeći fakultetski email.

Koraci:

1. Kreirajte novi projekat i nazovite ga GI RNA-Seq Ime Prezime Indeks.

2. Idite na Files > Add Files i iz Public Files (tab koji ce biti otvoren po defaultu) izaberite i importujte G20479.HCC1143.2.converted.pe_1_1Mreads.fastq i G20479.HCC1143.2.converted.pe_2_1Mreads.fastq . Ovo dva fajla su paired end RNA-Seq sekvence jednog uzorka iz Cancer Cell Line Encyclopedia (CCLE) data seta. Možete pogledati kako podaci izgledaju klikom na fajl i izborom Raw View taba. Na prvom tabu (Metadata) možete videti neke osnovne podatke kao i koji tip kancera je u pitanju.

3. Vratite se u svoj projekat. Klik na Apps > Add App i pronađite RNA-seq Alignment - TopHat. Iskopirajte ovaj workflow u svoj projekat (2x klik na Copy).

4. Ponovo se vratite u projekat i u Apps tabu kliknite Run pored Tophat workflow-a.

5. Biće vam ponuđeno automatsko importovanje preporučenih referentnih fajlova (Bowtie indeks i genske anotacije). Prihvatite (“Copy”).

6. Input “reads” su FASTQ readovi koje smo iskopirali u koraku #2. Izaberite ove fajlove i pokrenite task (klikom na Run gore levo).

7. Dok se task izvršava (~45min), proučite parametre i njihove default vrednosti.

8. Picard summary metrics fajl (*.summary_metrics.txt) sadrži informacije o procentu poravnatih ridova i druge *alignment* metrike. Ako su metrike zadovoljavajuće (npr. procenat ridova poravnatih u paru veći od 80%) pređite na sledeći korak - kvantifikaciju. Slično kao u koraku #3 kada ste u projekat importovali Tophat workflow, sada to uradite sa alatom HTSeq-count. HTSeq-count kvantifikuje gensku ekspresiju prostim prebrojavanjem ridova mapiranih na regionu svakog gena.

9. HTSeq-count prima dva inputa - alignment fajl (BAM) i genske anotacije (GTF). Prvi

ste kreirali pokretanjem Tophat workflow-a, a drugi je već korišćen kao input za Tophat. **Vodite računa da umesto poravnatih ridova ne odaberete nemapirani BAM fajl**. Eksperimentisanje sa parametrima je dozvoljeno (ali nije neophodno) osim što **ne treba dirati *ID attribute.***

10. Nakon što se task završi, proverite da li ste na outputu dobili TXT fajl sa 2 kolone - ID gena i broj ridova mapiranih na taj gen.

11. Pokrenite interaktivnu analizu (gore desno - Interactive Analysis tab > Data Cruncher > Create your first analysis). Odaberite JupyterLab, pa kada se analiza startuje i otvorite je - Python 3 kernel. Sad ste u standardnom notebook okruženju.

12. Path do željenog fajla u projektu biće */sbgenomics/project-files/ime_fajla.ext* , ali u kodu dole su već uneseni najverovatniji pathovi (kakvi će biti ako ste pratili uputstva do sada).

13. Učitajte rezultate kvantikacije sa:

```
import pandas as pd
counts =
pd.read_csv('/sbgenomics/project-files/G20479.HCC1143.2.converted.pe_accepted_hits.table.txt',
                names = ['gene_id', 'raw_count'], sep='\t')
counts = counts[:-5]
```

14. Dužinu gena izračunaćete uz pomoć anotacionog fajla i GTFtools alata:

```
!wget http://www.genemine.org/codes/GTFtools_0.6.9.zip
!unzip GTFtools_0.6.9.zip
!/sbgenomics/workspace/GTFtools_0.6.9/gtftools.py -l gene_length.bed
/sbgenomics/project-files/Homo_sapiens.GRCh37.75.gtf
gene_lengths = pd.read_csv('gene_length.bed', sep = '\t')
```

15. gene_lengths data frame ima više kolona sa različitim interpretacijama dužine gena - koristite *median*.

16. Normalizujte ekspresiju (counts) koristeći dužine gena iz gene_lengths i formule/kod sa predavanja. Svejedno je da li ćete koristiti TPM ili FPKM jedinice.

17. Izdvojiti 20 najviše eksprimovanih gena (nakon normalizacije).

18. Dodajte usera marko_zecevic u projekat (Dashboard tab u projektu > Invite new members).