PROJECT1-INSTRUCTIONS

May 11, 2024

1 Identification of pathogenic disease-associated short tandem repeats (STRs) in clinical samples

1.1 1. Overview

The Neurology Department at a teaching hospital in South Africa recently acquired a seed grant from a new biotechnology start-up company. This company plans to bring affordable next-generation sequencing to the South African market for genetic diagnosis- a space which currently has a very limited service offering. The Neurology Department jumped at the opportunity to have 10 patient genomes sequenced for free and tasked a bioinformatics student with the analysis of the whole genome sequencing (WGS) data. The 10 selected patients presented with "progressive adult-onset muscle weakness" with a diverse but variable range of symptoms including slurred speech, muscle cramps and twitches, muscle wasting and even cognitive decline in some patients. Although the neurologists felt that these patients may have a possible genetic aetiology for their symptoms, the bioinformatics student who analysed the data did not find any pathogenic variants when screening single nucleotide polymorphisms (SNPs) and short insertions and deletions (in/dels) in the WGS data.

Given the prominent role of repeat expansions in neurological disorders, the Neurology Department has approached your group to ask for assistance to screen the WGS data from their 10 patients. They are concerned that they may be missing disease-associated short tandem repeats (STRs) amongst this group which were not picked up by the standard variant calling pipeline which focuses on SNPs and in/dels.

1.1.1 Dataset

Illumina PCR-free 30X coverage whole genome sequencing data from 10 individuals

1.1.2 TR Genotyping Tools

ExpansionHunter, REViewer, Samtools

1.1.3 Expected Results

On successful implementation of the project, participants should be able to:

- Provide a genetic diagnosis for one patient by identifying a disease-associated repeat expansion in the pathogenic range.
- Identify 3 patients with intermediate range disease-associated repeat expansions.

- Provide a visualization of all 4 of the identified disease-associated REs clearly showing the read support for each allele.
- Provide some feedback for the Neurology Department regarding the following queries:
  - Are you sure these REs are real? Do you suggest we validate them by any other method?
  - How confident are you in the size of the expanded repeats? Is your analysis accurate?
  - Do you think we could offer RE testing as a diagnostic service in our clinic using whole genome sequencing data? Can you foresee any challenges or important things we need to consider?
  - Where did you get your reference ranges from? Are these ranges valid in African populations?
  - Can we use whole exome sequencing (WES) data to analyse REs? The start-up company has offered WES to us at a very good price but WGS is unfortunately not within our budget.

1.2 2. Getting started

Follow the instructions in `PROJECT1-SETUP.ipynb` to install the required tools for this project.

We will be working with the following datasets in the `~/public/project1-expansions` directory:

- BAM files and indices for samples aligned to the GRCh37 reference genome

```
ERR1955514.bam
ERR1955514.bam.bai
ERR1955398.bam
ERR1955398.bam.bai
ERR1955482.bam
ERR1955482.bam.bai
ERR1955462.bam
ERR1955462.bam.bai
ERR1955424.bam
ERR1955424.bam.bai
ERR1955504.bam
ERR1955504.bam.bai
ERR1955527.bam
ERR1955527.bam.bai
ERR1955415.bam
ERR1955415.bam.bai
ERR1955531.bam
ERR1955531.bam.bai
ERR1955473.bam
ERR1955473.bam.bai
```

- ExpansionHunter variant catalogs:

```
# Each folder contains the catalog for different
# reference genome builds
# grch37, grch38, hg19, hg38
~/public/project1-expansions/variant_catalog/
```

- You can also find reference genomes at:

```
~/public/genomes/GRCh37.fa
~/public/genomes/Homo_sapiens_assembly38.fasta
```

2 3. Running ExpansionHunter

Run ExpansionHunter on each of the 10 samples. Samples are aligned to the GRCh37 reference genome. An example ExpansionHunter command is below:

```
ExpansionHunter \
    --reads ~/public/project1-expansions/ERR1955514.bam \
    --reference ~/public/genomes/GRCh37.fa \
    --variant-catalog ~/public/project1-expansions/variant_catalog/grch37/variant_catalog.json
    --output-prefix ERR1955514
```

This command will output: * `ERR1955514.vcf` which contains genotypes in VCF format * `ERR1955514.json` which contains genotypes in JSON format * `ERR1955514_realigned.bam` "bamlet" with read realignments at repeat regions

Take a look at the output and try to answer the questions above in "Expected Results". Do any samples have repeats that have expanded beyond the normal range?

Before running REViewer, we'll have to sort and index the output BAMs. See example commands below:

```
samtools sort -o ERR1955514_realigned.sorted.bam ERR1955514_realigned.bam
samtools index ERR1955514_realigned.bam # creates ERR1955514_realigned.sorted.bam.bai
```

3 4. Visualizing data for expanded repeats using REViewer

You can use REViewer to inspect read alignments at STRs genotyped by ExpansionHunter. Below shows an example command:

```
REViewer \
    --reads ERR1955514_realigned.sorted.bam \
    --vcf ERR1955514.vcf \
    --reference ~/public/genomes/GRCh37.fa \
    --catalog ~/public/project1-expansions/variant_catalog/grch37/variant_catalog.json \
    --locus C9ORF72 \
    --output-prefix ERR1955514
```

You can change which samples and/or loci are visualized by changing input options in the command above. This command will output a file `ERR1955514.C9ORF72.svg`. Note to visualize SVG files you can download the it from datahub and open it in your web browser.

4 5. Write a report summarizing your results

Each group should submit a written report and presentation to: https://bit.ly/3WvnUqW