**Project 1 Report - Group 2**

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

**INTRODUCTION**

Microsatellites, or short tandem repeats (STRs), are polymorphic nucleotide sequences that range from 1 to 6 bp and can be repeated several times in the genome. STRs are present in many eukaryotes and prokaryotes and are distributed fairly evenly throughout the human genome(Fan, H., & Chu, J. Y., 2007).

Over the past few decades, it has become evident that many human diseases are directly caused by STRs. The majority of diseases referred to as repeat expansion mainly impact the nervous system (Paulson H *et al.* 2018*)*.

**PROJECT DESCRIPTION**

The Neurology Department at a teaching hospital in South Africa recently acquired a seed grant from a new biotechnology start-up company.They carried out Whole Genome Sequencing on 10 patient genomes 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.

The WGS data did not find any pathogenic variants when screening single nucleotide polymorphisms (SNPs) and short insertions and deletions (in/dels) in the patients samples. The neurologists felt that these patients may have a possible genetic etiology for their symptoms. Given the prominent role of repeat expansions in neurological disorders.

The Neurology Department has approached our 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.

**Expected Results**

On successful implementation, the Neurology Department expects that we should:

i) Provide a genetic diagnosis for one patient by identifying a disease-associated repeat expansion in the pathogenic range.

ii) Identify 3 patients with intermediate range disease-associated repeat expansions.

iii) Provide a visualization of all 4 of the identified disease-associated REs clearly showing the read support for each allele.

iv) Provide some feedback for the Neurology Department regarding the following queries:

**METHODOLOGY**

**Dataset**

We procured Illumina PCR-free whole genome sequencing data with a 30X coverage for a cohort of 10 individuals. This high-coverage sequencing aimed to provide a comprehensive view of the genome, minimizing the gaps and ensuring accuracy in variant calling. The reference genome used for alignment was GRch37, a well-established human genome build that facilitates comparative analysis and interpretation of genomic data.

**Tandem Repeat (TR) Genotyping Tools**

Tandem Repeat (TR) Genotyping Tools Employed For the genotyping of tandem repeats within the dataset, we utilized a suite of specialized bioinformatics tools:

* **ExpansionHunter**: This tool was pivotal in identifying and characterizing expansions of short tandem repeats (STRs) that are implicated in various genetic disorders.
* **REViewer**: A complementary tool to ExpansionHunter, REViewer was employed for the graphical visualization of the STR expansions.
* **Samtools**: An essential utility for manipulating alignments in the Sequence Alignment/Map (SAM) format, which includes sorting and indexing BAM files.

# **ExpansionHunter**

**Operational Workflow:** Each of the 10 genomic samples underwent a rigorous analysis with ExpansionHunter. The workflow was as follows:

1. **Alignment**: Samples were meticulously aligned to the GRCh37 reference genome.
2. **Output Generation**: ExpansionHunter produced three distinct outputs for each sample:
   * **VCF Format**: Containing the variant calls in a standardized format.
   * **JSON Format**: Offering a structured and detailed report of the repeat expansions.
   * **BAM Format**: Providing read realignments at the repeat regions, which is crucial for downstream analysis.

# **Visualizing data for expanded repeats using REViewer**

Following the generation of these outputs, we proceeded to sort and index the BAM files, which is a critical step to organize the data and facilitate efficient access and analysis.

To further delve into the data, we utilized **REViewer** for an in-depth examination of the read alignments at the STRs that were genotyped by **ExpansionHunter.** This step was particularly focused on *loci* that exhibited pathogenic or intermediate range expansions associated with disease. The output from REViewer was an ***SVG file***, which provided a scalable and high-resolution visual representation of the repeat expansions.

**RESULTS**

1. The comprehensive analysis yielded significant findings. A patient, identified by the sample number ***ERR1955473***, exhibited repeat expansions within the C9ORF72 gene repeat unit 'GGCCCC' in the intron. These expansions fell into the pathogenic range and are known to be associated with amyotrophic lateral sclerosis (ALS), a severe neurodegenerative disorder. This discovery underscores the potential of our methodology in uncovering clinically relevant genetic variations.
2. There were three patients with disease associated repeat expansions in the intermediate regions. They were:
   1. A Patient identified by the sample number ***ERR1955473***, exhibited repeat expansions within the DIP2B gene with repeat unit ‘GGC’. These expansions fell into the intermediate range with 41 repeats compared to the normal range which is between 6 to 23 repeats.
   2. A Patient identified by the sample number ***ERR1955482***, exhibited repeat expansions within the ATXN1 gene with repeat unit ‘CAG’ in the coding region. These expansions fell into the intermediate range with 41 repeats compared to the normal range which is between 6 to 23 repeats.
3. **Status**: PASS

* **End Position of the STR in the Genome**: 71,652,202
* **Reference Repeat Count**: 25
* **Repeat Length**: 25
* **Repeat Unit**: Adenine (A)
* **Variant ID**: FXN\_A

This discovery underscores the potential of our methodology in uncovering clinically relevant genetic variations that could have significant implications for patient diagnosis and treatment strategies.

**REFERENCES**

Fan, H., & Chu, J. Y. (2007). A brief review of short tandem repeat mutation. *Genomics, proteomics & bioinformatics*, *5*(1), 7–14. <https://doi.org/10.1016/S1672-0229(07)60009-6>

Paulson H. (2018). Repeat expansion diseases. *Handbook of clinical neurology*, *147*, 105–123. https://doi.org/10.1016/B978-0-444-63233-3.00009-9