Identification of Pathogenic Disease-Associated Short Tandem Repeats (STRs) in Clinical Samples

James Osei-Mensa Data Repository

1. Overview

The Neurology Department at a teaching hospital in South Africa acquired a seed grant from a biotechnology start-up aiming to introduce affordable next-generation sequencing for genetic diagnostics in the region. The department sequenced the genomes of 10 patients presenting with progressive adult-onset muscle weakness, with symptoms including slurred speech, muscle cramps, muscle wasting, and cognitive decline. Initial analysis did not find pathogenic variants in SNPs or short in/dels, leading the department to suspect disease-associated STRs not detected by standard pipelines.

2. Dataset

Source: Illumina PCR-free 30X coverage whole genome sequencing (WGS) data from 10 individuals. Files: BAM files aligned to the GRCh37 reference genome. Variant catalogs for multiple reference genomes. Reference genomes: GRCh37.fa and Homo_sapiens_assembly38.fasta.

3. Methodology

Tools Used:

ExpansionHunter: To identify and genotype STRs REViewer: To visualise the called STRs. Samtools: To sort and indexing BAM files. stri.py: To annotate the called STRs [STRIpy database] (https://stripy.org/database#)

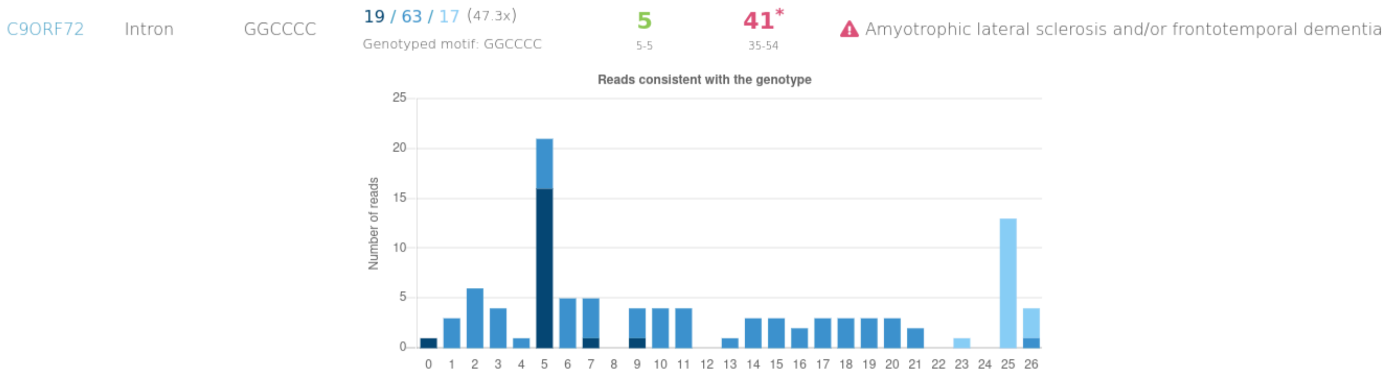
1. Provide a genetic diagnosis for one patient by identifying a disease-associated repeat expansion in the pathogenic range.

ERR1955473 (Pathogenic C9ORF72, Amyotrophic lateral sclerosis and/or frontotemporal dementia)

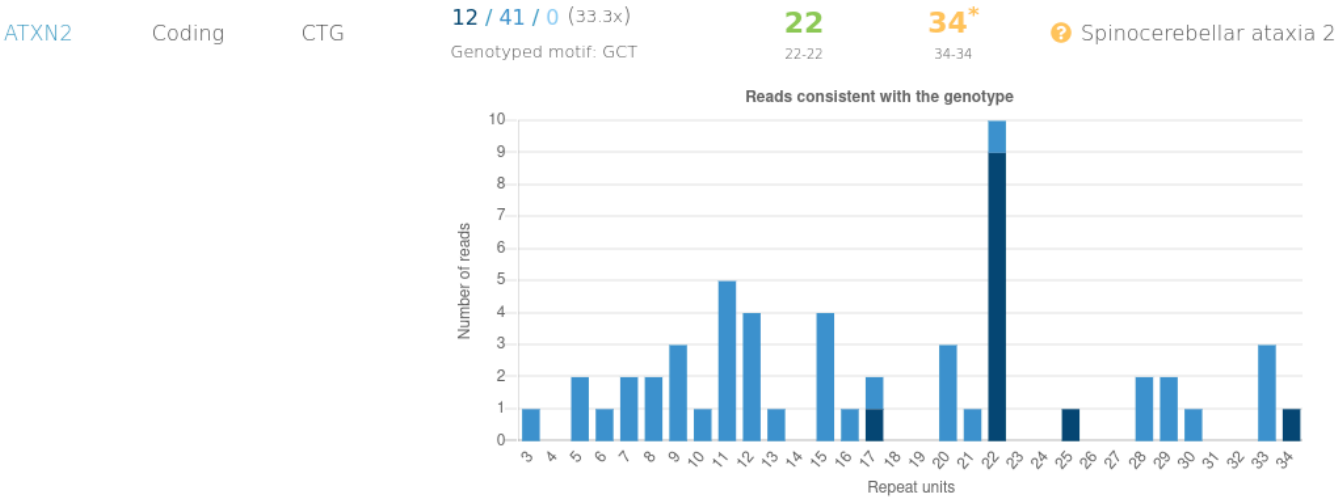
2. Identify 3 patients with intermediate range disease-associated repeat expansions.

ERR1955504 (Intermediate ATXN2 , Spinocerebellar ataxia 2)

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



C9ORF72	Intron	GGCCCC	19 / 63 / 17 (47.3x) Genotyped motif: GGCCCC	5 5-5	41* 35-54	⚠ Amyotrophic lateral sclerosis and/or frontotemporal dementia



ATXN2	Coding	CTG	12 / 41 / 0 (33.3x) Genotyped motif: GCT	22 22-22	34* 34-34	? Spinocerebellar ataxia 2

4. Provide some feedback for the Neurology Department regarding the following queries: a. Are you sure these REs are real? Do you suggest we validate them by any other method? b. How confident are you in the size of the expanded repeats? Is your analysis accurate?
5. 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?
6. Where did you get your reference ranges from? Are these ranges valid in African populations?

STRipy database

7. 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.

WES can be used to analyse REs if we are looking at a specific RE that is known to be in the exomic region.

- ERR1955473 is in the pathogenic range for C9ORF72 which is associated with Amyotrophic lateral sclerosis and/or frontotemporal dementia). The locus had a coverage of 47.2703 and confidence interval of 5-5/35-54
- ERR1955504 is in the intermediate range for ATXN2 which is associated with Spinocerebellar ataxia 2). The locus had a coverage of 33.3243 and confidence interval of 22-22/34-34
- RFC1 is associated with cerebellar ataxia, neuropathy, vestibular areflexia syndrome (CANVAS) due to a bialleic expansion https://www.ncbi.nlm.nih.gov/books/NBK564656/
- ERR1955462 had a biallelic expansion `<STR8>`,`<STR60>`