# STR Expansion Analysis of Pacific Biosciences Data Using NextGENe®LR Software

November 2020

Megan McCluskey, Jacie Wu, Lidong Luo, Jonathan Liu

## Introduction

Short tandem repeats (STRs) are found scattered throughout the human genome. Certain STRs, often trinucleotide repeat expansion (TRE) variants, are known to be related to heritable disease (1). The long repetitive sequences present a challenge for many sequencing technologies. Pacific Biosciences SMRT (single molecule real time) sequencing provides a solution for this challenge with the ability to sequence reads up to dozens of kilobases in length. This enables the sequencing of the full repeat lengths, for up to hundreds of repeats, within a single read (2).

NextGENe®LR software can be used for STR Expansion analysis from long read sequencing data such as data from Pacific BioSciences systems. This provides an accurate reporting of STR lengths to identify STR loci with lengths associated with disease.

## Method

NextGENeLR aligns the long sequence reads to the reference sequence using a specialized alignment algorithm. Alignment results are saved to a BAM file which can be loaded in the built-in Alignment Viewer for visualization and reporting. When the STR Expansion Report is opened, the STR regions to be evaluated are defined. A default list of STR regions associated with disease (3) is provided. This list of regions can be modified by removing regions, adding regions, editing regions, or uploading a text file containing a list of regions to be added. For the STR regions of interest, the number of reads with each repeat length is counted.

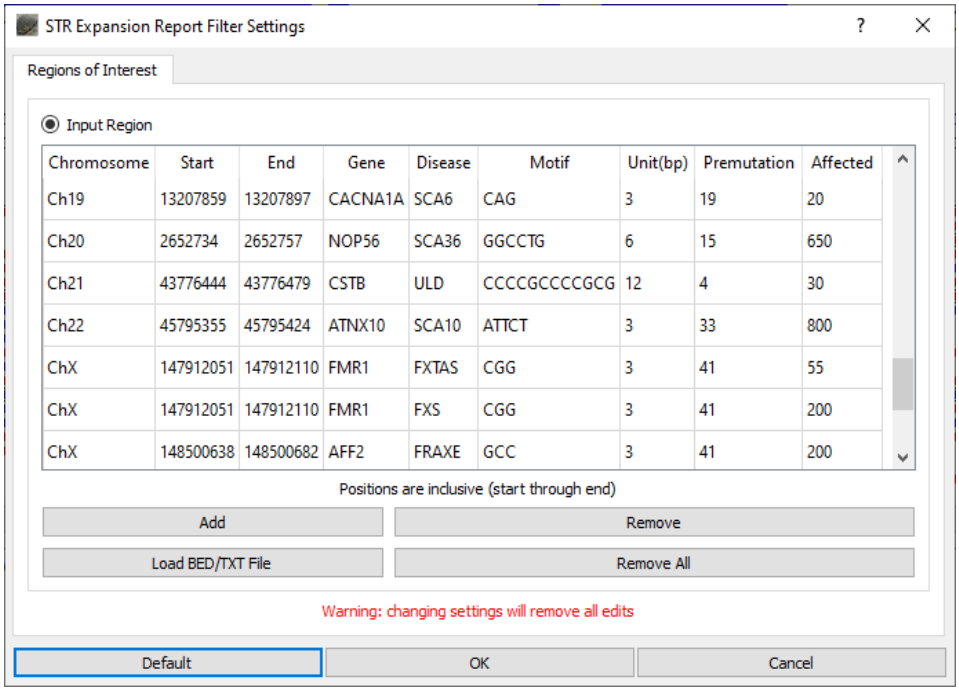


*Figure 1: STR Expansion Regions of Interest can be defined. The default list can be used as is, regions can be removed and/or added, or a custom list of regions can be loaded.*

## Results

Repeat lengths for all included regions of interest are displayed in the graphical STR Expansion Report. The normal range for the repeat length is displayed with green. Gray denotes the premutation range and red indicates the repeat length associated with disease. Blue boxes are placed at the repeat length(s) detected for the sample. The top number for each blue box indicates the repeat length and the bottom number indicates the number of reads with this length

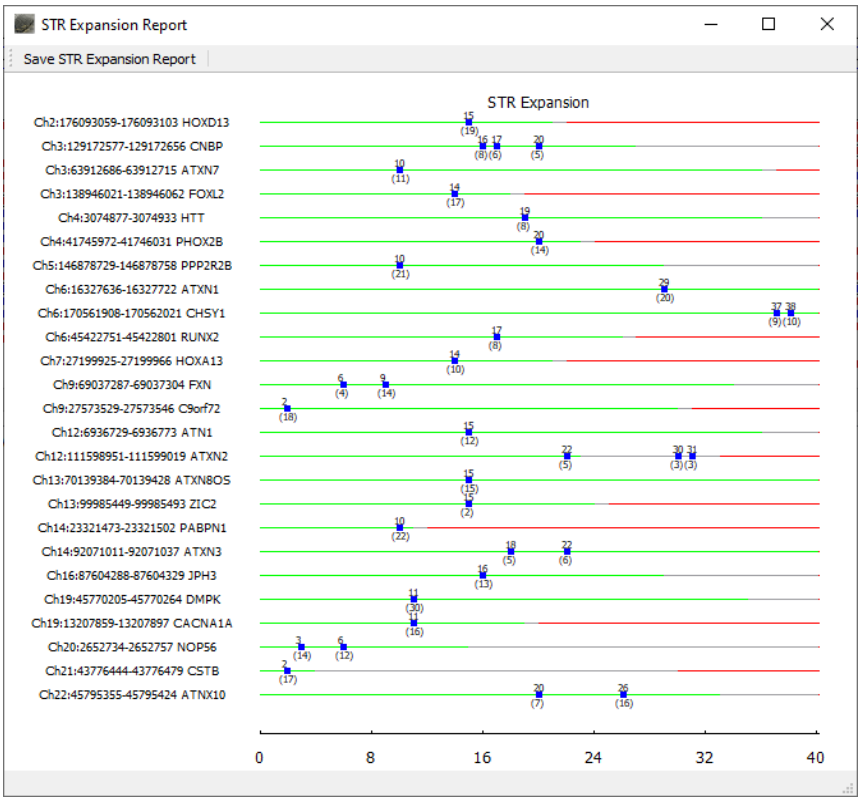


*Figure 2: The NextGENeLR STR Expansion Report displays the repeat lengths detected for all STR regions of interest. Green indicates the normal STR length range, gray for premutation, and red for affected. In this case, the ATXN2 STR region shows some reads in the premutation range.*

## Discussion

In conjunction with the long reads lengths of the PacBio systems, NextGENeLR provides a powerful tool for STR repeat expansion analysis to assess disease associations. STR lengths can quickly and accurately be assessed for multiple loci in a single, easy to interpret report. Multiple samples can also be loaded in batch and automatically processed consecutively.

NextGENeLR also includes tools for structural variation detection, SNV and small indel detection, and whole genome mitochondrial DNA analysis, including haplotyping of mixed samples.

## References


1. de Leeuw, R.H., Garnier, D., Kroon, R.M.J.M. *et al.* Diagnostics of short tandem repeat expansion variants using massively parallel sequencing and componential tools. *Eur J Hum Genet* **27,** 400–407 (2019).

2. Loomis EW, Eid JS, Peluso P, et al. Sequencing the unsequenceable: expanded CGG-repeat alleles of the fragile X gene. *Genome Res*. 2013;23(1):121-128.

3. Tang H, Kirkness EF, Lippert C, et al. Profiling of Short-Tandem-Repeat Disease Alleles in 12,632 Human Whole Genomes. *Am J Hum Genet*. 2017 Nov 2;101(5):700-715.