Tracy: basecalling, alignment, assembly and deconvolution of sanger chromatogram trace files Abstract:

- DNA sequencing is at the core of many molecular biology laboratories. Despite its long history, there is a lack of user-friendly Sanger sequencing data analysis tools that can be run interactively as a web application or at large-scale in batch from the command-line.
- (Tracy) : an efficient and versatile command-line application that enables basecalling, alignment, assembly and deconvolution of sequencing chromatogram files. Its companion web applications make all functionality of Tracy easily accessible using standard web browser technologies and interactive graphical user interfaces.
- (Tracy) : can be easily integrated in large-scale pipelines and high-throughput settings, and it uses state-of-the-art file formats such as JSON for reporting chromatogram sequencing results.
- Tracy can be routinely applied in large-scale validation efforts conducted in clinical genomics studies as well as for high-throughput genome editing techniques that require a fast and rapid method to confirm discovered variants or engineered mutations.

Introduction:

- Sanger sequencing has a long history in molecular biology and it remains indispensable for many routine tasks like the sequencing of single genes, cloned plasmids, expression constructs or PCR products.
- Automatisation of these standard tasks avoids misinterpretation of mutations and aids the researchers to focus on the critical mutations instead of inspecting hundreds of chromatogram peaks by eye.
- For large sequencing projects that aim at cataloging the human genetic variation or the mutation spectrum present in diseases such as cancer it is important to accurately estimate a false discovery rate of their respective call sets or to validate actionable mutations.
- Most of the available trace analysis software aims at analyzing one trace at a time in an interactive, often proprietary and licensed trace analysis viewer that lacks support for standard file formats such as VCF/BCF , the predominant variant calling reporting format in NGS studies.
- This often demands a deconvolution of Sanger chromatogram traces into its constituting alleles, which is non-trivial for mixed chromatogram traces that involve heterozygous insertions or deletions.
- These routine chromatogram evaluation tasks that require a graphical trace analysis application, large-scale genome editing and clinical resequencing projects demand a flexible and scalable command-line application that can be integrated into automated workflows.

Related Work :

- Base calling : is the process of assigning nucleobases to chromatogram peaks or electrical current changes resulting from nucleotides passing through a nanopore.

- Tracy also supports re-estimating basecalling qualities. The output of Tracy can be in JSON, FASTA, FASTQ or TSV format. The JSON and TSV formats output the trace at every sampling position. These formats also list the basecalling positions,

- Sequence alignment or sequence comparison lies at heart of the bioinformatics, which describes the way of arrangement of DNA/RNA or protein sequences, in order to identify the regions of similarity among them.

- In illumina : During the alignment step, the banded Smith-Waterman algorithm aligns clusters from each sample against amplicon sequences specified in the manifest file.
- The banded Smith-Waterman algorithm performs local sequence alignments to determine similar regions between 2 sequences. Instead of comparing the total sequence, the Smith-Waterman algorithm compares segments of all possible lengths. Local alignments are useful for dissimilar sequences that are suspected to contain regions of similarity within the larger sequence. This process allows alignment across small amplicon targets, often less than 10 bp

- In Nanopore sequencing: has emerged as a major sequencing technology and many long-read aligners have been designed for aligning nanopore reads. However, the high error rate makes accurate and efficient alignment difficult. Utilizing the noise and error characteristics inherent in the sequencing process properly can play a vital role in constructing a robust aligner.

- Assembly: Sequence assembly is the initial step towards downstream data analysis of the sequencing data.
- Comparative Assembly : reference based assembly or mapping to genome of a closely related species .
- De Novo Assembly : assembly in the strict sense . No or little information about the genome , transcriptome or proteins.
- -Falcon assembler
- The Falcon assembler developed by Jason Chin from PacBio is another pipeline adopting the strategy of HGAP. It shares many features with PBcR, such as raw reads overlapping for base error correction using DALIGNER and overlap filtering. The major difference lies in its contig consensus generation

-Miniasm assembler

- Read error correction is the most CPU-intensive stage of the whole assembly process, and assemblies on gigabase-sized genomes are still out of reach for many projects due to high sequencing costs and large computational requirements. The Miniasm assembler developed by Heng Li takes a different approach to deal with noisy long reads by skipping the step of read error correction completely.

- -Hybrid assemblers

- The NGS platforms such as Illumina’s HiSeq and MiSeq have played a dominant role in genomic research and applications. It is foreseeable that short read data will continue to be a very important part of data sources for years to come. Different algorithms have been explored for genome assembly and many pipelines have been developed for various applications.