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.