

Fighting Outbreaks with LiveTools: A Live Toolbox for DNA Sequencing

Motivation

Next-Generation Sequencing (NGS) is a high-throughput technology to analyze the DNA, i.e. the genetic information, of a biological sample. NGS has a crucial importance in the clinical context. It can, for example, be used to identify the organism(s) that cause the disease of a patient. In the recent large outbreak of Ebola and in 2014, NGS was a crucial tool in effectively combating the disease. It helped researchers to identify the related pathogens, trace back their origin, and describe possible transmission paths. Outbreak situations are highly time-critical, as already the delay of a single day can mean that the disease affects hundreds of more people. Currently, the analysis of NGS data can only start after the sequencing has finished. This can still take hours or days, which is – in spite of the advances NGS technology has made in the last years – not fast enough for outbreak situations.

Setting

The Robert Koch Institute (RKI) is working on tools to reduce the current delay in sequencing and subsequent analysis. They have developed tools that perform important analysis steps when the NGS machine is running, i.e. during data creation. By this, researchers get first analysis results after fractions of sequencing time. These tools are the first approaches that perform such live analyses for NGS data and have already been applied to clinical samples and open-view outbreak diagnostics. These autonomous tools shall now be combined in a unified toolbox.

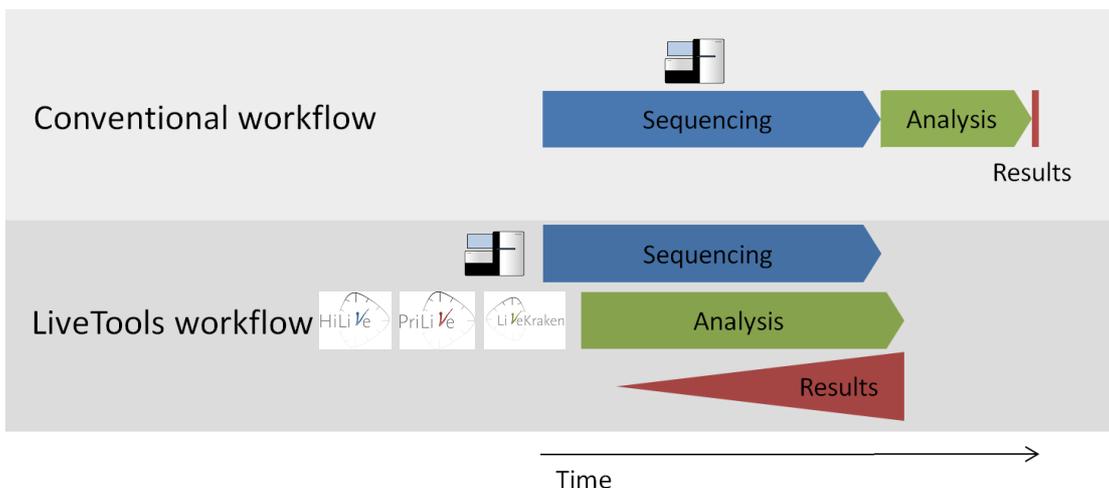


Figure 1 Comparison of NGS workflows with conventional analysis tools and RKI's real-time LiveTools

Project Goals

The overall aim of this Bachelor's project will be to bundle the power of RKI's software by combining their existing tools in a live toolbox for NGS, called LiveTools. Additionally, there are several smaller and more specific tasks, e.g. data conversion, to make it applicable to a plethora of additional use cases. From the technical side, the main task will be the design and implementation of a common framework that allows for integration of existing and new tools. The existing C++ code will be adapted and integrated into this new framework. The project is planned in a very flexible way to allow for realizing your own ideas. These may, for example, include algorithmic optimizations as alternative indexing structures, additional functionality as possible sequencing abort criteria or the implementation of graphical output or a Graphical User Interface (GUI).

Technology and Skills

The existing tools are built in C++. Some of them use the SeqAn library for efficient algorithms and data structures for the analysis of sequences. Key aspects of RKI's live approaches include the efficient use of multithreading, data (de-)serialization and templates. Software compilation is managed with CMake.

Profound experience with C++ will be helpful, but is not mandatory as the project offers enough other possibilities for contribution, e.g. regarding the UI. Biological expertise is not required; you will learn about the necessary biological details during the project. At the end of the project, you will have an in-depth understanding of technologies, standards, data formats, and algorithms used during sequencing and subsequent analysis.

Contact

You are welcome to visit us in our respective offices or reach out to us via mail:

Cindy Perscheid (cindy.perscheid@hpi.de), Campus II, V-1.19

Milena Kraus (milena.kraus@hpi.de), Campus III, G-2.2.16

Prof. Dr. med. Erwin Böttinger (office-boettinger@hpi.de), Campus III, G-2.2.23