V.I.P. and V.I.P. Validator: A diagnostic amplicon sequencing Variant Interpretation Pipeline and variant detection validator

Joachim De Schrijver (joachim.deschrijver@ugent.be) & Prof. Dr. Wim Van Criekinge (wim.vancriekinge@ugent.be)

Laboratory for Bioinformatics and Computational Genomics (BioBix), Department of Molecular Biotechnology, Faculty of Bioscience Engineering, Ghent University, Belgium – on behalf of NXTGNT

Introduction

High throughput amplicon sequencing allows rapid genetic diagnostics for a series of patients in multiple genes. Since amplicon sequencing is PCR dependant, PCR optimization (reduction of byproducts and improved equimolarity) can improve the cost-effectiveness of sequencing based diagnostics. Also the variation detection software must be capable of verifying if the set of known variations can be detected using the analysis pipeline. At the moment no out-of-the-box solution exists that fill all the needs for a diagnostic environment, especially the efficient detection of deletions and insertions, and certainly combinations of both, remains a problem. Comparing different sequencing runs or analyzing several runs at the same time also remains a problem because there are no standard database solutions available.

We developed the Variant Interpretation Pipeline (V.I.P.), an integrated software pipeline that enables variation detection, pre-sequencing PCR efficiency optimalisation, variation detection validation, and data storage in a database.

Methods

The pipeline was developed using Perl, BLAT and a MySQL database.

The Variant Interpretation Pipeline

The Variant Interpretation Pipeline (V.I.P.) can analyze data generated in a single GS-FLX sequencing run in a short time (<1h on a single-node server). The pipeline is developed to process multiplexed DNA pools with sample specific tag sequences (MID sequences). Prior to storing the sequences in the database, tag sequences are removed to improve mapping efficiency (Fig. 1b).

The sequences are mapped on a reference sequence using BLAT. The results of this mapping are processed in such a way that correctly mapped amplicons are separated from priming mismatches and/or primer dimmers. Both the ‘correct’ and ‘erroneous’ sequences are stored in a database (Fig. 1a). By analyzing these ‘erroneous’ sequences, PCR reactions can be optimized and sequencing efficiency improved. The results of a prior sequencing run can thus be used to optimize the PCR reactions of a future sequencing experiment, which is of great interest in a diagnostic setting.

After analysing the data, V.I.P. generates several reports of your choice (Fig. 1b). Because all the data is stored in a database, custom reports can be generated in a short time.

The V.I.P. Validator

The validator uses 500 real sequences of a certain amplicon that are stored in the database and generates random or selected variations in all the 500 sequences (or a desired fraction). The altered sequences are then reanalyzed by the analysis pipeline (Fig. 1a). The detection frequency is the frequency with which an introduced error at a certain position in an amplicon is discovered by the analysis pipeline at that correct position. The average detection frequencies across 50 amplicons (of 500 sequences each) are given in the table below (Fig. 2).

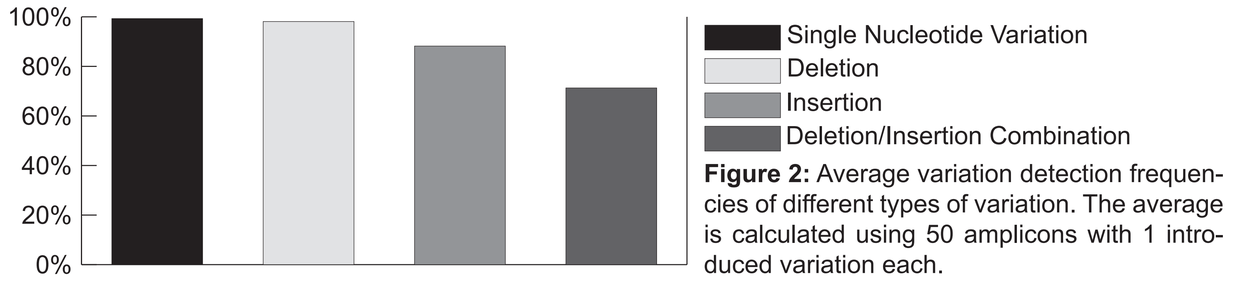

Figure 2: Average variation detection frequencies of different types of variation. The average is calculated using 50 amplicons with 1 introduced variation each.

This validation pipeline has several purposes. It can be used to optimize the analysis pipeline and the variant detection algorithms. Or, more important and of great interest in a diagnostic lab, it allows validation of the analysis software with respect to detection of certain variations, prior to starting a diagnostic screening.

Conclusion

We have developed a pipeline that is of great interest for diagnostic resequencing projects. A sequencing run is processed, mapped on a reference, analyzed and stored in a database. The pipeline reports variations (single nucleotide variations, deletions, insertions and combinations). Analyzing the ‘erroneous’ data in the database, such as primer dimers or undesired sequences allows optimizing the PCR reactions.

The validator allows, prior to detecting a certain known variation via sequencing technology, virtual experiments to be carried out to verify if the analysis pipeline would be capable of detecting the variation.

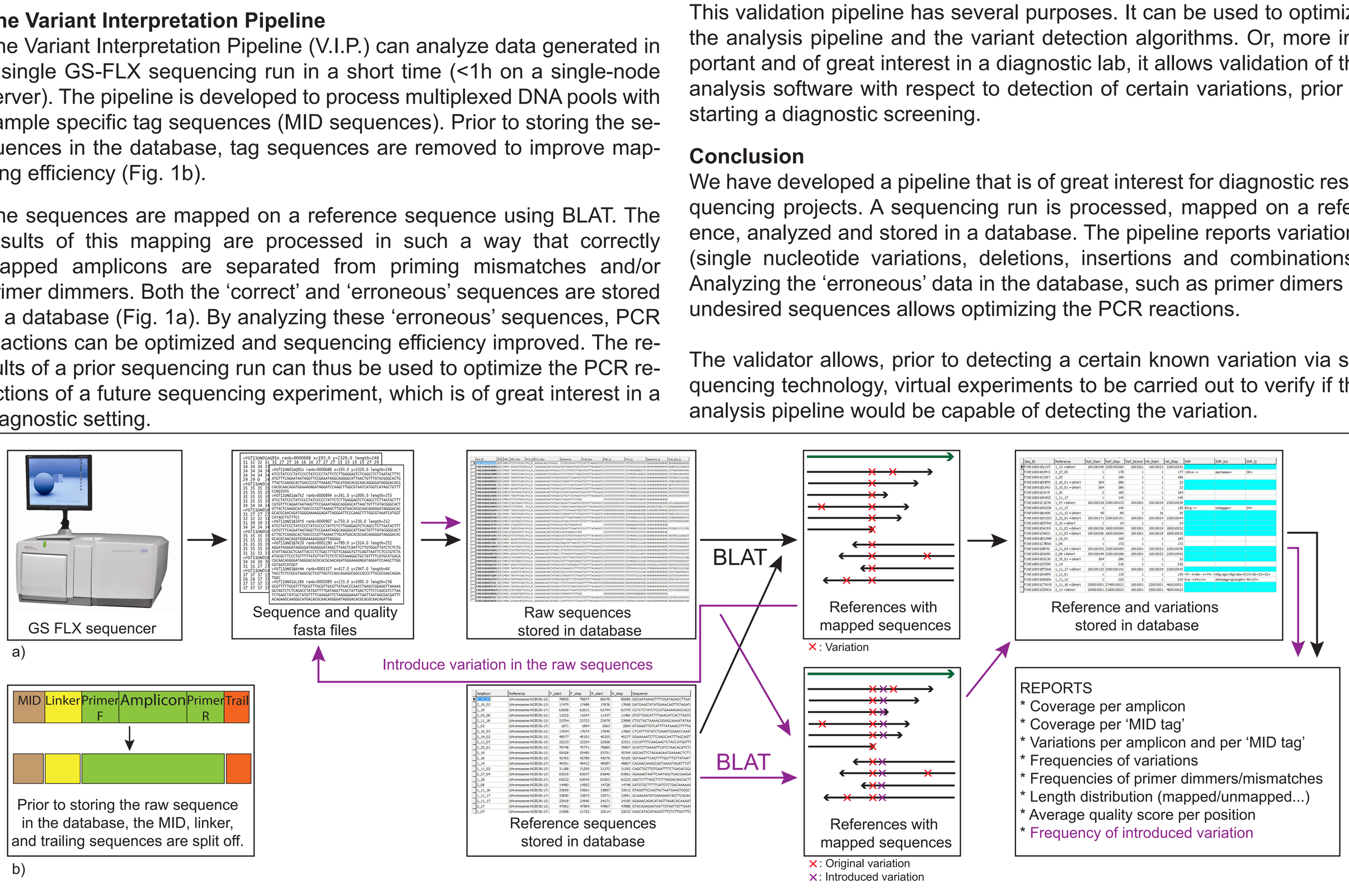

Figure 1: a) An overview of the V.I.P. pipeline (black arrows) and the V.I.P. validator pipeline (magenta arrows). The validator reuses the original VIP pipeline and database, which reduces complexity. b) Prior to storing a raw sequence, the sequence is split using an algorithm that allows one mismatch in the MID and the linker sequence. The trail (reverse complements of parts of the MID and/or linker sequence) is also removed from the real sequence.