V-pipe: a computational pipeline for assessing viral genetic diversity from high throughput data.

Abstract: **Motivation**: High-throughput sequencing technologies are used in viral genomics ,These technologies facilitate the assessment of the genetic diversity, which affects transmission of viral infections , but there are two challenges in analysing viral diversity: First: amplification and sequencing. second: the large data volumes represent computational limitations .

Results : we developed V-pipe, a bioinformatics pipeline that integrates various computaional tools for the analysis of viral high throughput sequencing data. It supports quality control, read mapping and alignment, low-frequency mutation calling and inference of viral haplotypes. V pipe also includes benchmarking functionality ,We demonstrate this capability by assessing the impact of three different read aligners (Bowtie 2, BWA MEM, ngshmmalign) and two different variant callers (LoFreq, ShoRAH) on the perform once of calling single-nucleotide variants in intra-host virus populations .

Introduction: we will use Rna data because they exhibit short generation times and higher mutation rates compared to cellular organisms .

-High throughput sequencing (HTS) technologies have opened up new possibilities for in-depth characterization of the genetic diversity of virus samples However, analysing viral HTS data is complicated because of large volumes of data and short length of the sequencing reads.

the analysis of HTS data involves additional steps implemented by separate tools, quality control and read alignment .

A prerequisite for incorporating HTS data into routine diagnostics is to standardize the processing steps end-to-end, and several bioinformatics pipelines have been developed for that.

-An important step in inferring viral genetic diversity is the alignment of HTS reads. There are two strategies for read alignment, is reference-based approaches and de novo assembly.

Shortcomings of the former are the introduction of biases due to low similarity between the reference genome and viral haplotypes , so we have developed a read aligner, called **ngshmmalign**: The aligner borrows idea from the alignment of protein families to align HTS reads from small and highly diverse genomes -before that we have developed **V-pipe**, a flexible bioinformatics pipeline integrating several tools for analysing viral HTS data. V-pipe allows for assessing viral diversity at the level of SNVs, short variant sequences and long□range haplotypes .

To this end: it contains modules to generate synthetic data and to assess the accuracy of the computational inference. While our focus here is on SNV calling, the performance of various methods for viral haplotype reconstruction We validate V-pipe using sequencing data from a control sample composed of five well defined HIV-1 strains .