2-Methods and materials:

1.2Computational pipeline

V-pipe uses the Snakemake workflow management system , which enables the well-controlled and scalable execution of the pipeline in local as well as in high-performance computing environments. The pipeline integrates various open-source software packages developed for analysing virus samples. In addition to the alignment file, ngshmmalign outputs two types of consensus sequences constructed from the aligned reads, namely by using a majority vote at each position, and incorporating ambiguous bases. Vpipe produces these consensus sequences in case an alternative aligner is chosen.

We use lowercase characters in the consensus sequences to mark positions with read counts below 50 reads, by default.

2.2 ngshmmalign: read aligner

:To do so, reads are aligned to a panel of reference sequences containing suspected contaminants. The code developed to support this functionality, as well as various other steps of our pipeline, are maintained as an independent Python package called smallgenomeutilities.

A multiple sequence alignment of the reads is performed independently for each of the windows by employing the L-INS-i iterative refinement approach implemented by MAFFT. In the third step, the final read alignment is obtained by re-aligning all reads to the profile HMM . To standardize the position numbering, V-pipe performs a lift-over to report the alignment relative the user-specified reference sequence. To do so, we construct and use a multiple sequence alignment, containing the reference sequence and all consensus sequences from datasets included in the data analysis.

In the second mode, the simulated haplotype sequences are sampled from a perfect binary tree. In the third mode, haplotype sequences are defined a priori and given as input to the pipeline, e. based on other models of viral evolution or on known viral sequences. After haplotype sequences have been generated, we use the ART software to simulate either single-end or paired-end reads with configurable read length.
We simulate reads from every individual haplotype with read coverage proportional to its relative abundance.

3.2Simulated datasets

The datasets are based on HIV-1 or HCV sequences. Emulating populations using sequences derived from plasma samples of individual patients allows us to mimic the structure of viral populations more faithfully.

4.2 Control sample for pipeline validation on real

datasetsThe control sample consists of an in vitro mixture of five known
HIV-1 strains mixed at equal proportions. Four sequencing experiments were carried out starting with approximately 104 and 105 HIV-1 RNA copies. The protocol described by Di Giallonardo et al. was employed for the amplification and sequencing. Two types of sequencing

experiments were carried out: one including all five amplicons and another one using only amplicon B .

Results:

Simulation studies

We align simulated reads using ngshmmalign and compare it against two widely used read aligners, namely BWA MEM and Bowtie 2. To investigate potential read alignment bias due to differences in sequence similarity to the reference sequence, we report the fraction of aligned reads and aligned bases per haplotype. For the HCVbased datasets, sequences exhibit a broad range of divergence from the reference strain . BWA MEM aligns most of the reads regardless of the divergence from the reference, but a large fraction of the bases are softclipped, whereas ngshmmalign aligns a higher fraction of the bases for all haplotypes .

Since the main focus of V-pipe is to infer viral genetic diversity, we evaluate the accuracy of the read aligners based on the F1 score of detecting SNVs using ShoRAH. We evaluate two read coverages and three strategies to generate the underlying haplotype abundances.

In addition to the position-wise performance, we also account for the length of individual deletion events and again find ngshmmalign to perform best .

Next, we focus on mutation calling and compare the accuracy of

SNVs obtained by using ShoRAH versus LoFreq, while fixing ngshmmalign for the read alignment. Although aligning reads with ngshmmalign and performing mutation calling with ShoRAH resulted in better F1 scores in most cases, we note that there is a trade-off between accuracy and computational resources .

Validation of V-Pipe on a mixture of five HIV-1 strains

We employ V-pipe using ngshmmalign with a de novo constructed reference sequence, and ShoRAH for SNV calling. In all cases, Vpipe detected more than 86% of the expected SNVs, with almost perfect specificity . The missed variants are predominately located at the genome termini, which correspond to regions of lower coverage. We also observe a decrease in precision for datasets A-100k, B-10k and B-100k. When inspecting the precision of the mutation calls as a function of the variant frequencies, we find that a precision higher than 98% can be attained for SNVs with frequencies at least% for all the analysed datasets .

V-pipe 5

We visualise the reconstructed local haplotypes by representing pairwise Hamming distances in a two-dimensional plane . Sequences of haplotypes recovered after 561 and 1255 days of infection are situated closer to the initial haplotype sequence, whereas sequences corresponding to later time points are further away.