Supplementary Material: Full-length de novo viral quasispecies assembly through variation graph construction

Jasmijn A. Baaijens*, Bastiaan Van der Roest†, Johannes Köster‡§, Leen Stougie*¶‖, Alexander Schönhuth**‖††

A Parameters

Our method requires to manually set three parameters, the minimal node abundance α, the minimal haplotype abundance γ, and the maximal trim length τ. The minimal node abundance α refers to removing mismatches when concatenating paths, see 'Correcting errors in paths $p \in P'$' in the main manuscript. As a general guideline, increasing α leads to increasing numbers of candidate paths, hence an increasing number of variables in the minimization problem. The greater the number of variables, the greater the chance to pick up low abundance paths, while at the same time the greater the risk to pick up haplotype artifacts.

The minimal haplotype abundance γ refers to selecting haplotypes after having solved the minimization problem in Section 2.2.2. Any haplotype $p \in P$ where $a(p) < \gamma$ will be discarded from the output.

The trim length τ refers to 'Trimming paths $p \in P'$' in Section 2.2.1. Increasing τ leads to less concatenations of paths from P', hence to less candidate paths in general, at the risk of not concatenating correctly joining contigs.

We recommend setting α and γ to 0.5% and 1.0% of the total sequencing depth of the original data set, respectively, and $\tau = 10$. These default settings were chosen according to the quality of the input contigs [1]. Given that the data sets considered here have a total sequencing depth of 20,000x, all experiments were run with $\alpha = 100, \gamma = 200, \tau = 10bp$.

B Runtime and memory usage

Since our method takes as input a set of pre-assembled contigs, the most time-consuming and memory-expensive step in viral quasispecies assembly has already been accomplished. By their worst-case runtime complexity, both candidate path generation and minimizing for selecting optimal sets of haplotypes minimization steps reflect exponential procedures. However, in practice, the runtime of the algorithm is hardly affected by these procedures. Instead, it is clearly dominated by the read mapping step for computing $a' : V' \to \mathbb{R}$ when constructing the contig variation graph. This step took 7.5 CPU hours for the HCV data, 19 CPU hours for the ZIKV data, and 3.2 CPU

*Centrum Wiskunde & Informatica, Amsterdam, Netherlands
†University Medical Center Utrecht, Utrecht, Netherlands
‡University of Duisberg-Essen, Essen, Germany
§Dana Farber Cancer Institute, Harvard Medical School, Boston, United States
¶Vrije Universiteit, Amsterdam, Netherlands
‖INRIA-Erable
**Utrecht University, Utrecht, Netherlands
††Corresponding author (alexander.schoenhuth@cwi.nl)

hours for the Polio data, with a peak memory usage of 0.6GB, 0.9GB, and 0.6GB, respectively. Given that the read mapping step is perfectly parallelizable, these computations completed in less than an hour on a 24-core machine. For a comparison with the other methods evaluated see Table 1.

	HCV		ZIKV		Poliovirus	
	CPU time (hours)	Peak memory usage (GB)	CPU time (hours)	Peak memory usage (GB)	CPU time (hours)	Peak memory usage (GB)
SAVAGE	55	0.8	61	0.8	72	3.4
Virus-VG	7.5	0.6	19	0.9	3.2	0.6
PredictHaplo	2.7	1.1	7.4	1.1	2.0	0.8
ShoRAH	509	8.9	814	10	-	-

Table 1: Runtime and -space comparison between Virus-VG, SAVAGE, and the state-of-the-art methods PredictHaplo and ShoRAH. ShoRAH could not process the Poliovirus data.

C Simulated Poliovirus data

We extracted sequences for 6 closely related poliovirus strains from the NCBI nucleotide database, accession numbers MG212475.1, MG212489.1, MG212484.1, MG21469.1, MG212490.1, and MG212491.1. Two of these sequences (MG212476.1 and MG21484.1) show a big deletion (larger than 1000bp) compared to the Sabin2 reference strain.

D Detailed results

Abundance estimation errors are presented in Table 2. More detailed assembly statistics per data set for each of the methods (SAVAGE, Virus-VG, PredictHaplo, ShoRAH) can be found in Table 3.

	HCV		ZIKV		Poliovirus	
	absolute error (%)	relative error (%)	absolute error (%)	relative error (%)	absolute error (%)	relative error (%)
Virus-VG	0.1	0.9	0.3	6.0	0.6	12.8
PredictHaplo	0.9	11.3	4.9	69	10.6	10.6
ShoRAH	8.5	64	39	229	-	-

Table 2: Absolute and relative abundance estimation errors per method. For each data set, we present the average error over all assembled strains. Note that ShoRAH was unable to process the Poliovirus data and PredictHaplo only found one of the six virus strains in this data set.

	# strains*	target (%)	N50	NA50	NGA50	N-rate (%)	MR (%)	IR (%)	unaligned (bp)
SAVAGE	26	99.4	8964	8964	8964	0	0.001	0	0
Virus-VG	10	99.3	9281	9281	9203	0	0.001	0	0
PredictHaplo	9	73.8	7636	7622	7608	0.006	0.053	0	0
ShoRAH	639	56.9	7570	7570	7570	0	4.283	0.011	60560

(a) 10-strain HCV mixture (simulated Illumina MiSeq)

	# strains*	target (%)	N50	NA50	NGA50	N-rate (%)	MR (%)	IR (%)	unaligned (bp)
SAVAGE	100	98.8	2954	2954	3801	0.002	0.021	0	0
Virus-VG	20	92.8	10202	10200	10210	0.003	0.106	0.006	0
PredictHaplo	8	53.3	10270	10269	10267	0.001	0.121	0.004	0
ShoRAH	493	26.3	10117	10117	10117	0.0	4.381	0.011	91053

(b) 15-strain ZIKV mixture (simulated Illumina MiSeq)

	# strains*	target (%)	N50	NA50	NGA50	N-rate (%)	MR (%)	IR (%)	unaligned (bp)
SAVAGE	59	83.7	1089	1089	1643	0.006	0.013	0	0
Virus-VG	14	80.7	7316	7316	7428	0	0.064	0	0
PredictHaplo	3	16.6	7461	7434	-	0.009	1.816	0	7461

(c) 6-strain Poliovirus mixture (simulated Illumina MiSeq)

	# strains*	target (%)	N50	NA50	NGA50	N-rate (%)	MR (%)	IR (%)	unaligned (bp)
SAVAGE	68	97.9	1026	1026	1450	0	0.027	0.039	0
Virus-VG	23	90.6	2130	2130	4642	0	0.282	0.042	0
PredictHaplo	6	100.0	8825	8825	8825	0.215	0.673	0.178	0
ShoRAH	250	100.0	8775	8775	8775	0.507	3.228	0.175	26631

(d) Real 5-strain HIV mixture (Illumina MiSeq)

Table 3: Assembly results per dataset. We evaluate the number of strains assembled, the fraction of the target haplotypes reconstructed, the N50, NA50, and NGA50 measures, the fraction of 'N's (uncalled bases), the mismatch rate (MR), the indel rate (IR), and the number of unaligned bases. ShoRAH could not process the Poliovirus data. *For SAVAGE, this column indicates the number of contigs in the assembly; since these are not full-length, this does not reflect the number of strains.