Evaluation of haplotype callers for next-generation sequencing of viruses

Anton Eliseev[a,b,^], Keylie M. Gibson[c,*,^], Pavel Avdeyev[c,d,#], Dmitry Novik[a,#], , Matthew L. Bendall[c], Marcos Pérez-Losada[c,e,f], Nikita Alexeev[a,‡], and Keith A. Crandall[c,e,‡]

[a] Computer Technologies Department, ITMO University, Saint-Petersburg, Russia
[b] JetBrains Research, Saint-Petersburg, Russia
[c] Computational Biology Institute, Milken Institute School of Public Health, George Washington University, Washington, DC, USA
[d] Department of Mathematics, George Washington University, Washington, DC, USA
[e] Department of Biostatistics and Bioinformatics, Milken Institute School of Public Health, George Washington University, Washington, DC, USA
[f] CIBIO-InBIO, Centro de Investigação em Biodiversidade e Recursos Genéticos, Universidade do Porto, Campus Agrário de Vairão, Vairão, Portugal

^ Co-First Author
# Co-Second Author
‡ Joint Senior Authors

* Corresponding author at: Computational Biology Institute, Milken Institute School of Public Health, George Washington University, Washington, DC, USA
Email address: kmgibson@gwu.edu (K.M. Gibson)

Highlights

- Haplotype callers for NGS data vary greatly in their performance.
- Haplotype callers performance is mainly determined by mutation rate.
- Haplotype callers performance is less sensitive to effective population size.
- Most haplotype callers perform well with low diversity and poorly with high diversity.
- PredictHaplo performs best if genetic diversity is in the range of HIV diversity.

Abbreviations

NGS – Next generation sequencing
HIV – Human immunodeficiency virus
HCV - Hepatitis C virus
HPV - Human papillomavirus

Abstract

Currently, the standard practice for assembling next-generation sequencing (NGS) reads of viral genomes is to summarize thousands of individual short reads into a single consensus sequence, thus confounding useful intra-host diversity information for molecular phylodynamic inference. It is hypothesized that a few viral strains may dominate the intra-host genetic diversity with a variety of lower frequency strains comprising the rest of the population. Several software tools currently exist to convert NGS sequence variants into haplotypes. However, previous studies suggest that current approaches of haplotype reconstruction greatly underestimate intra-host diversity. Here, we tested twelve NGS haplotype reconstruction methods using viral populations simulated under realistic evolutionary dynamics. Parameters for the simulated data spanned known fast evolving viruses (e.g., HIV-1) diversity estimates to test the limits of the haplotype reconstruction methods and ensured coverage of predicted intra-host viral diversity levels. Using those parameters, we simulated HIV-1 viral populations of 216-1,185 haplotypes per host at a frequency <7%. All twelve investigated haplotype callers showed variable performance and produced drastically different results that were mainly driven by differences in mutation rate and, to a lesser extent, in effective population size. Most methods were able to accurately reconstruct haplotypes when genetic diversity was low. However, under higher levels of diversity (e.g., those seen intra-host HIV-1 infections), haplotype reconstruction accuracy was highly variable and, on average, poor. High diversity levels led to severe underestimation of, with a few tools greatly overestimating, the true number of haplotypes. PredictHaplo and PEHaplo produced estimates close to the true number of haplotypes, although their haplotype reconstruction accuracy was worse than that of the other ten tools. We conclude that haplotype reconstruction from NGS short reads is unreliable due to high genetic diversity of fast-evolving viruses. Local haplotype reconstruction of longer reads to phase variants may provide a more reliable estimation of viral variants within a population.

Keywords: fast-evolving viruses, haplotype reconstruction, HIV, intra-host diversity, next-generation sequencing, second generation sequencing

1. Introduction

Next-generation sequencing (NGS) technologies provide novel opportunities to study the evolution of many viruses that impose health issues among humans, such as human immunodeficiency virus (HIV), hepatitis C virus (HCV), human papillomavirus (HPV), and influenza. Such sequencing platforms allow an in-depth characterization of the genetic diversity in a heterogeneous intra-host viral population by sequencing many viral strains directly. Illumina and 454/Roche offered the first round of next-generation sequencing machines, which gradually replaced Sanger sequencing for viral studies. These platforms are able to generate a sufficiently high coverage of the genome, which allows one to detect mutations present in less abundant strains. However, the large number of short reads with a relatively high error rate produced during sequencing poses computational and statistical challenges for reconstructing full-length strain sequences and estimating their frequency. In particular, since abundance rates can be comparable or lower than sequencing error rates, high sequence error rates (≤0.1% for Illumina reads) can interfere with the detection of true mutations that are present at low frequencies. Moreover, short reads length (25 – 400 bp) need to be assembled into an unknown number of contigs. Ultimately, the goal of assembly is to produce contigs that can cover the entire targeted gene region (i.e., targeted amplicon sequencing) or that can be scaffolded together to cover the length of a full genome (i.e., shotgun sequencing). Finally, the large number of sequencing reads (25 – 300 million) requires the development of algorithms capable of processing this large amount of data. The amount of data generated by a single NGS run (1 GB to 1 TB) can be up to a million times greater than that generated by a single Sanger sequencing run (1 MB of data).

Several computational tools have been developed over the last decade to address the challenge of defining sequence variants (haplotypes – sometimes erroneously referred to as 'quasispecies'; see Holmes, 2010) from NGS data (Beerenwinkel and Zagordi, 2011; Di Giallonardo et al., 2014; Pandit and de Boer, 2014; Posada-Cespedes et al., 2016; Schirmer et al., 2014). Different software has been tailored to various sequencing platforms and experimental designs. It is important to note that 454/Roche sequencing reads were the main input data for developers of viral variant assemblers until 2013. This was because 454/Roche was the first widely-used NGS platform and generated longer reads than all other Illumina platforms at the time (Beerenwinkel and Zagordi, 2011; Schirmer et al., 2014). A number of computational methods were proposed for handling the 454/Roche reads, including PredictHaplo (Prabhakaran et al., 2014), ViSpA (Astrovskaya et al., 2011), QuRe (Prosperi and Salemi, 2012), QuasiRecomb (Topfer et al., 2013), VirA (Skums et al., 2013), BIOA (Mancuso et al., 2011), Mutant-Bin (Prabhakara et al., 2013), V-Phaser + V-Profiler[1] (Henn et al., 2012; Macalalad et al., 2012), and ShoRAH (Zagordi et al., 2010). Some of these methods were empirically validated using HIV and HCV data sets with the methods showing little success in estimating reliable sequence variants from NGS data (Prosperi et al., 2013). Later, thanks to the better cost-effectiveness and higher coverage offered by the Illumina sequencing platforms, the main focus migrated towards Illumina technology and has become dominant for developers of viral sequence variant assemblers since then (Posada-cespedes et al., 2016). Following this paradigm shift, several methods such as PredictHaplo, V-Phaser (Yang et al., 2013), and QuasiRecomb were extended to handle Illumina reads, and a number of tools, including VGA (Mangul et al., 2014), HaploClique (Töpfer et al., 2014), QColors (Huang et al., 2011), QSdpR (Barik et al., 2018), and ViQuas (Jayasundara et al., 2015), were developed specifically to handle Illumina reads.

Currently, all state-of-the-art methods for viral variant reconstruction are designed to assemble contigs from Illumina reads and can be divided into two main categories based on their dependency on a reference genome: *reference-based* assemblers and *de novo* assemblers (Fig. 1). In the former category, sequencing reads are aligned to a reference genome and information about the reads positioning and orientation with respect to a reference genome is obtained (Fig. 1 c_1). This information

[1] V-phaser is used for calling variants in a viral sequence sample and V-Profiler is utilized for analyzing and visualizing variants from V-Phaser at the nucleotide, codon, and haplotype levels.

is further used to reconstruct haplotypes (Fig. 1 c_2, c_3, d_1, d_2). *De novo* assemblers, however, do not rely on reference genomes, and haplotype sequences are usually reconstructed directly from the reads (Fig. 1 c_4, d_3). *De novo* assembly often requires more computational resources, but *reference-based* assembly requires a closely related genome, which is not always available in high quality. Each strategy has advantages and disadvantages and have been implemented in several software programs, but the performance of these assembly tools has not been comprehensively examined yet. In this study, we simulated realistic, coalescent based intra-host viral diversity with diversity measurements encompassing known variation from fast-evolving viruses such as HIV-1 for empirical grounding. We then used these simulated populations to assess the performance and accuracy of three *de novo* and nine *reference-based* sequence variant reconstruction tools or haplotype callers.

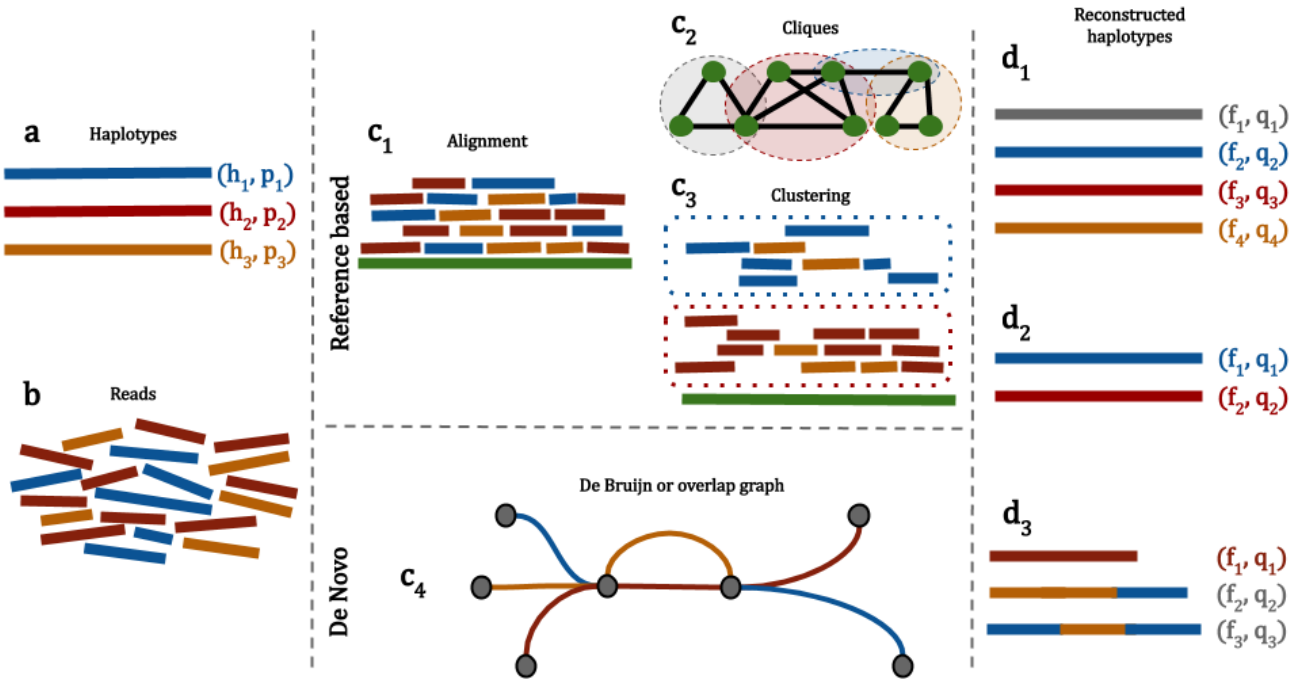

Figure 1. Schematic diagram representing the process of reconstructing haplotypes from next-generation sequencing reads by reference-based and *de novo* methods. (a) A hypothetical virus population consisting of three haplotypes is sequenced by NGS techniques. (b) Reads originating from different haplotypes are identified by distinct colors in the diagram. (c_1) After sequencing, reads are aligned against reference genome (green) as a first step in all reference-based methods. (c_2) Read alignment is used for building a graph and candidate haplotypes are reconstructed as maximal cliques in the graph. (c_3) Read alignment is used for dividing reads into clusters and candidate haplotypes are reconstructed by concatenation of all reads from clusters. (c_4) Alternatively, after sequencing, reads are *de novo* assembled using De Bruijn or overlap graphs and candidate haplotypes are reconstructed as paths by analyzing the graph structure. (d_1) A method based on clique detections overestimates the number of reconstructed haplotypes with relative frequencies. (d_2) A method based on clustering procedure underestimates the number of reconstructed haplotypes with relative frequencies. (d_3) A *de novo* method reconstructs the correct number of haplotypes with frequencies, but one inferred haplotype is smaller than the true haplotype and the other two haplotypes are admixtures of the original haplotypes.

2. Material and Methods

2.1 Viral Sequence Variant Estimators

All *de novo* viral variant reconstruction methods can be further divided into two subcategories: *consensus* and *strain-specific* assemblers. The main goal of consensus-based tools is, generally speaking, to construct a suitable consensus reference genome that may be further used as a template for more fine-grained studies. VICUNA (Yang et al., 2012) and IVA (Hunt et al., 2015) represent this subcategory of methods. VICUNA is the most popular software among them, as it generates full-length consensus and detects polymorphisms. VICUNA merges NGS reads into contigs, and those into a bigger contig, by calculating "good" prefix-suffix overlap between sequences. During this process, contigs are also clustered and validated to reach a better quality of consensus. IVA follows the same approach with only one difference, the tool starts from k-mers that are sorted in decreasing order with respect to their abundance and then extends sequences into a bigger sequence by using reads that have perfect overlap with initial sequences. VICUNA also has an additional option for contig merging if a reference genome exists.

Contrary to *de novo* consensus approaches, *de novo* strain-specific assemblers aim to reconstruct sequences at the strain resolution level (Table 1). It is worth mentioning that the *de novo* viral variant reconstruction problem is quite similar to the assembly effort of multiple genomes in microbial communities at once using shotgun metagenomic reads (e.g., Bishara et al., 2018; Scholz et al., 2016). The arising challenges in the microbial community genome assembly are addressed by metagenome assemblers. Thus, at first glance, applying metagenome assemblers to *de novo* viral variant reconstruction seems very promising. However, SPAdes is the only assembler that was able to identify haplotypes in the case of sufficiently abundant strains (Jasmijn A Baaijens et al., 2017; Bankevich et al., 2012; Nurk et al., 2017). Therefore, the development of specific assemblers for viral sequence variants is required. Currently, there exist three *de novo* strain-specific assemblers, namely MLEHaplo (Malhotra et al., 2015), SAVAGE (Jasmijn A Baaijens et al., 2017), and PEHaplo (Chen et al., 2018) (Table 1). MLEHaplo was the first assembler that truly applied *de novo* viral sequence variant assembly at the strain resolution level. MLEHaplo performs k-mer counting and then filters erroneous k-mers using raw reads and a specified threshold value. Afterwards, the tool builds a De Bruijn graph (see Compeau et al., 2011) based on the set of k-mers obtained in the previous round (Fig. 1 c_4). On the next step, MLEHaplo recovers paths from the De Bruijn graph that may correspond to haplotypes. Finally, the tool chooses correct haplotypes and estimates their frequencies using the maximum likelihood method. PEHaplo follows the same workflow as MLEHaplo. However, PEHaplo constructs an overlap graph instead of creating a De Bruijn graph during the initial steps[2] (Fig. 1 c_4). PEHaplo also has a more careful path finding algorithm based on paired-end connection information. SAVAGE uses overlap graphs as a key data structure, but the pipeline is different from those in PEHaplo and MLEHaplo. After constructing an overlap graph (Fig. 1 c_4), SAVAGE joins overlapped read pairs. At the next step, SAVAGE iteratively merges reads into contigs and contigs into scaffolds using clique enumeration and contig formation. Finally, the tool uses Kallisto (Bray et al., 2016) to estimate frequencies of the resulting haplotypes

[2] One can read the difference between overlap graphs and De Bruijn graphs in a review paper by Li et al. (2012).

Table 1. *De novo* and *reference-based* viral haplotype callers compared in this study.

Software Tool	Published Year	Programming Language	Frequencies	Haplotypes	Type
MLEHaplo	2015	Perl	+	+	*de novo*
Savage	2017	Python 3	-	-	*de novo*
PEHaplo	2018	Python 2.7	+	-	*de novo*
ShoRAH	2011	C++	+	+	*reference-based*
QuRe	2012	Java 6	+	+	*reference-based*
QuasiRecomb	2013	Java 7	+	+	*reference-based*
PredictHaplo	2014	C++	+	+	*reference-based*
HaploClique	2014	C++	+	+	*reference-based*
ViQuas	2015	R, Perl	+	+	*reference-based*
aBayesQR	2017	C++	+	+	*reference-based*
RegressHaplo	2017	R	+	+	*reference-based*
CliqueSNV	2018	Java 6	+	+	*reference-based*

If a haplotype caller produces haplotypes as an output, it given a plus sign. If a haplotype caller reports corresponding frequencies for the sequences produced, it is given a plus sign. Savage and PEHaplo claim they produce contigs not haplotypes, which is why we did not deem that they produce haplotypes.

While the final sequences produced by MLEHaplo, PEHaplo, and SAVAGE are strain-specific, the obtained sequences, in general, do not represent full-length haplotypes (Fig. 1 d_3). Recently, Virus-VG and VG-flow have been developed for completing strain-specific assemblies produced by the aforementioned *de novo* strain-specific assemblers (Baaijens et al., 2018, 2019). Virus-VG and VG-flow try to convert strain-specific contigs into full-length haplotypes taking into account their abundances. The difference between Virus-VG and VG-flow is that the former uses a brute-force exact approach, while the latter utilizes a heuristic algorithm. Therefore, VG-flow is faster than Virus-VG, but less accurate. The main goal for both tools is to find and select maximum-length paths in a variation graph.

The main advantage of *reference-based viral* variant reconstruction methods prior to *de novo* haplotype assemblers is the potential ability to reconstruct full-length haplotypes (Fig. 1 d_1, d_2). However, it was shown in several studies (Jasmijn A Baaijens et al., 2017; Mangul et al., 2014) that the reference genome may bias the reconstruction of haplotypes. An additional disadvantage of using a reference-based tool is the potential lack of a high-quality reference genome of a virus population. In this case, the required reference genome can be potentially assembled from sequencing reads by first using *de novo* consensus assembly tools. Nevertheless, the reference genome is often available for common pathogenic viruses, such as HIV, HCV, polyomavirus or influenza.

Currently there are nine commonly used state-of-the-art reference-based tools (Table 1). All these tools claim to be global haplotype inference methods, i.e., able to infer the sequences and frequencies of the underlying viral strains over a longer region than the average read length. ShoRAH (Short Read Assembly into Haplotypes) is, historically, the first publicly available software (Zagordi et al., 2011). ShoRAH uses a probabilistic clustering algorithm for short haplotype sequence reconstruction (Fig. 1 c_1, c_3, d_2). Then, it computes a minimal set of haplotypes using the principle of parsimony that provides the best explanation for the given a set of error corrected sequencing reads (Eriksson et al., 2008). The tool then uses an expectation minimization algorithm for haplotype frequency estimation.

The next important milestone in the reference-based viral variant reconstruction tool development was the release of QuRe (Prosperi and Salemi, 2012). QuRe uses the combinatorial method proposed in Prosperi and Salemi (2012) for inferring genetic variants in local windows that do not exceed read lengths. After that, the obtained genetic variants are clustered by a probabilistic

algorithm (Zagordi et al., 2010) (Fig. 1 c_1, c_3, d_2). Finally, haplotypes with their frequencies are obtained by utilizing a genome reference and clustered variants. Later, the same probabilistic clustering and combinatorial algorithms were used for developing the reference-assisted assembly pipeline ViQuas (Jayasundara et al., 2015). The main difference between QuRe and ViQuas is that the latter tool assembles reads into contigs using the SSAKE assembler (Warren et al., 2007) and then iteratively extends obtained contigs by connecting overlapping pairs without using any sequence information from the reference.

The next developed software was PredictHaplo (Prabhakaran et al., 2014), HaploClique (Töpfer et al., 2014), and QuasiRecomb (Topfer et al., 2013). All these tools have special features in comparison to the previous generation tools. For example, PredictHaplo was specifically designed for identifying haplotypes in an HIV-1 population. HaploClique allows for detection of point mutations, large insertions and deletions. QuasiRecomb, on the other hand, incorporates the existence of recombination events into the estimated viral evolution. PredictHaplo, HaploClique, and QuasiRecomb are based on different approaches and their applications to the viral variant reconstruction problem were novel at the time. PredictHaplo reformulates the original problem in terms of a non-standard clustering problem, where reads are points in some metric space and haplotypes are clusters (Fig. 1 c_1, c_3, d_2). To take into account an unknown number of variants, the stochastic Dirichlet process and the infinite mixture model were used (Prabhakaran et al., 2010). HaploClique uses the insert size distribution and an iterative enumeration of maximal cliques in a graph to reconstruct super-reads that may represent haplotypes (Fig. 1 c_1, c_2, d_1). Due to the computational complexity of maximal clique enumeration, this tool requires excessive computational resources on data sets with coverage >1,000x. HaploClique provided inspiration for the development of SAVAGE. Finally, QuasiRecomb utilizes data parameters of a hidden Markov model for estimating point mutations and recombination events (David Posada et al., 2002). These parameters allow estimation of the probability of each possible haplotype with respect to the observed read data.

The latest releases of reference-based methods for viral sequence variant reconstruction are aBayesQR (Ahn and Vikalo, 2017), CliqueSNV (Knyazev et al., 2018), and RegressHaplo (Leviyang et al., 2017). CliqueSNV extends the previous approach used in the 2SNV tool (Artyomenko et al., 2017). CliqueSNV constructs a graph based on linkage information between single nucleotide variations and then identifies true viral variants by merging cliques in that graph (Fig. 1 c_1, c_2, d_1). RegressHaplo, in turn, is based on a regression-based approach specifically designed low diversity and convergent evolutions. This tool implements penalized regression to assess the haplotype interactions that belong to different unlinked regions. aBayesQR employs a maximum-likelihood approach to infer viral sequences. The search of most likely viral sequence is conducted on long contigs, which enables identification of closely related haplotypes in a population and provides computational tractability of the Bayesian method. It should be noted that aBayesQR is designed for reconstructing viral haplotypes that are near genetically identical.

Each haplotype reconstruction tool in Table 1 was run on the Colonial One high performance computing cluster at The George Washington University. We used 64 standard CPU nodes featuring dual Intel Xeon E5-2670 2.6GHz 8-core processors with a RAM capacity of 128GB. A single node with a 48-hour time limit was allocated for each run.

2.2 Simulation Data Description

Previous benchmarking of viral haplotype reconstruction programs (Pandit and de Boer, 2014; Prosperi et al., 2013; Schirmer et al., 2014) used simulation scenarios that are useful from a mathematical perspective but do not necessarily reflect viral evolution and epidemiology. For example, PredictHaplo artificially mutated ten haplotypes from a single HIV-1 reference genome at varying proportions (Prabhakaran et al., 2014); HaploClique used an in-house mixture of known HIV-1 strains (number of specific strains unknown) (Töpfer et al., 2014); and SAVAGE simulated their data based on Illumina MiSeq sequencing results from an in-house mixture of five unique strains of HIV-1

subtype B with varying relative abundances (see supplemental methods in (Jasmijn A Baaijens et al., 2017)). In those studies, often the pairwise divergence between the strains used to represent "real" HIV-1 haplotype diversity was either unreported (Prabhakaran et al., 2014) or ranged between 0.05% and 10% (Jasmijn A Baaijens et al., 2017; Töpfer et al., 2014). But realistic intra-host HIV-1 diversity is substantially lower with pairwise divergences ranging between 0.02% and 2%, while inter-host pairwise comparisons of the same viral subtypes can exceed 5% (Kearney et al., 2009; Maldarelli et al., 2013). Furthermore, unless the HIV-1 viral population in an individual was the product of a dual infection (see (van der Kuyl and Cornelissen, 2007) for review of dual infections), these benchmarking methods do not accurately represent the evolution of the virus, where the HIV viral population originated from an infection of one strain. All of these studies conditioned their simulations on HIV-1 data sets, but we also want to explore the general utility across a broader parameter space that encompasses more fast-evolving viral populations.

In our simulations, we used parameters and settings under the coalescent theory (Kingman, 2000, 1982; Rodrigo and Felsenstein, 1999; Rosenberg and Nordborg, 2002) to more accurately reflect viral intra-host diversity and evolution as seen in empirical studies (see (Crandall and Templeton, 1993)). We simulated viral intra-host evolutionary histories and the constituent haplotype sequences (tips) using the coalescent simulator CoalEvol v. 7.3.5 (Arenas and Posada, 2014). We set the mutation rate (μ) between 1e-3 and 5e-8 per-site to span past known viral mutation rates to test the limits of the reconstruction algorithms and number of haplotypes present using the human genome mutation rate as an upper limit and other retroviruses' mutation rates as a lower limit. These parameters encapsulated the empirical mutation rate of 2.5e-5 and 3.4e-5 estimated by Neher and Leitner (2010) and Maldarelli et al. (2013), respectively for HIV-1; HCV with an estimated mutation rate between 2.5e-5 and 1.2e-4 (Echeverría et al., 2015; Ribeiro et al., 2012; Sanjuán et al., 2010); HTLV-1 with an estimated mutation rate between 3.44e-7 and 7e−6 (Mansky, 2000; Nobre et al., 2018); and influenza with an estimated mutation rate of 3e−5 to 4e−5 (McCrone, 2018; McCrone et al., 2018). Although Neher and Leitner (2010) reported that the HIV-1 virus recombines at a rate of 1.4±0.6e−5, we chose to not include recombination in the simulated evolution histories because some of the compared haplotype reconstruction programs do not include recombination events in their reconstruction process. Other parameters that were fixed in the CoalEvol config file included: i) nucleotide frequencies (A=0.37, C=0.16, G=0.23, and T=0.25); ii) the transition/transversion rate (ti/tv = 2.5), as estimated among host diversity from Crandall et al. (1999); and iii) rate heterogeneity among sites (Γ = 0.95) and invariable site rate (I = 0.4) (Posada and Crandall, 2001), which are unique to HIV-1.

Recombination occurs frequently in natural HIV-1 populations (Crandall, 1999; Neher and Leitner, 2010; D Posada et al., 2002) but we chose not to model recombination in our simulations. First, many of the haplotype programs do not account for recombination. Second, we assume that approaches that fail on a simplified model without recombination will not perform well on a more complex that includes recombination.

We chose to use HIV-1 as an empirical viral strain to assess the capabilities of the haplotype reconstruction tools given that most developers validated their programs on this virus and genetic diversity values for this virus are well established. HIV-1 genetic diversity (Watterson's theta) for the polymerase gene (*pol*) has been estimated to fall between 0.067 and 0.09 for subtype B HIV-1 strains in the United States (Gibson et al., 2019; Pérez-Losada et al., 2017, 2010). Boltz et al. (2016) completed single genome sequencing that resulted in 677 – 1,577 sequences per sample for HIV-1, therefore, we limited our sample size to range between 100 and 2,000 with an alignment length of 1,137bp. This length was chosen because we used a section of the polymerase gene (*pol*) from the HXB2 reference sequence (GenBank accession number: K03455; (Ratner et al., 1985)) as the most recent common ancestor (MRCA) for each parameter set (HXB2 numbering: 2,253 – 3,390). It is important to note that CoalEvol is restricted to sample sizes of up to 2,000 haplotypes. Maldarelli et al. (2013) estimated the effective population size (N_e) of intra-host diversity to be between 1,000 and 10,000, so we varied the effective population size between 500 and 10,000. We also denoted the ploidy

as diploid (Coffin, 1992). Wherever possible, we varied the parameters to be above and below estimated HIV-1 estimates to ensure we adequately represented viral intra-host diversity and to examine the limits of the haplotype reconstruction programs. Expanding our parameter set allowed us to gain insights into other viral species with different evolutionary and population characteristics. For example, the N_e for influenza is considerably smaller than HIV-1 at around 20-100 viral sequences (Kim and Kim, 2016; McCrone, 2018; McCrone et al., 2018), while HCV hovers around the lower end of HIV-1 with an N_e of 10-1,000 sequences (Bernini et al., 2011).

Since the Illumina MiSeq platform is the most popular NGS technology currently used for viral amplicon sequencing due to low cost and high throughput, we simulated sequencing reads in the FASTA output (excluding the original HXB2 sequence we deemed as the GMCRA in the coalescence simulation) of CoalEvol using the NGS read simulator ART v. MountRainier-2016-06-05 (Huang et al., 2012). ART mimics real sequencing processes, therefore, we used the built-in sequencing Illumina MiSeq platform (MSv1). We simulated error-free 150 bp paired-end reads with a read count of 100 reads, mean size of 215 bp for DNA fragments, and a standard deviation of 120 bp for DNA fragment size.

The error free output data generated for the haplotype populations with the ART read simulator was processed with HAPHPIPE, a HAplotype reconstruction and PHylodynamics PIPEline for viral NGS sequences (Bendall et al., 2019). By both not simulating recombination and starting with sequencing-error free data, we removed nuisance variables that would impact haplotype reconstruction and could not be handled by some haplotype callers. Briefly, we used HAPHPIPE and its implementation of Trimmomatic v. 0.33 (Bolger et al., 2014) to trim the starting FASTQ files from the output of ART by removing low quality reads, low quality bases, and adapter contamination. We performed *de novo* assembly on the clean reads using Trinity v. 2.5.1 (Grabher et al., 2013) and formed scaffolds with MUMMER 3+ v. 3.23 (Alnafee, 2016). With two iterative refining steps, the cleaned reads were mapped back to the scaffolds with Bowtie2 v. 2.3.4.1 (Langmead and Salzberg, 2013). The BAM file of aligned reads generated as final output by HAPHPIPE and a FASTA file containing the cleaned reads (an intermediate output by HAPHPIPE) were used as input for the haplotype reconstruction algorithms.

2.3 Haplotype Assembly Comparative Indices

In order to evaluate the quality of haplotype assembly provided by different tools, we used common statistical measures of precision and recall, as well as weighted normalized UniFrac distance (Lozupone and Knight, 2005), which is widely used to compare microbial communities. Our simulated data can be represented as $P = \{(h_i, p_i), i = 1,\ 2, \dots\}$ – the ground truth haplotypes h_i and their associated abundances p_i ($\sum p_i = 1$), and $Q = \{(f_i, q_i), i = 1,\ 2, \dots\}$ – the set of predicted haplotypes f_i together with their predicted abundances q_i.

We define precision as $\frac{TP}{(TP+FP)}$ and recall as $\frac{TP}{(TP+FN)}$. Since the length of viral sequences reconstructed by *de novo* tools may differ from actual length of ground truth haplotypes, we define *TP* (true positive) and *FP* (false positive) differently for *reference-based* and *de novo* tools. We define *FN* (false negative) as $1 - TP$, for both assembly strategies equally.

In the case of *reference-based* methods, we define *TP* as the total frequency of those haplotypes h in the ground truth set P which have an accurate enough prediction f in Q (which means that the edit distance $d(h, f)$ is less than some threshold $T = T(\mu)$); we also define *FP* as the total frequency of those haplotypes f in the predicted set Q which do not match any haplotype f from the ground truth set (which means that $d(h, f) \geq T$ for all $f \in P$). We choose the threshold $T = 12$ because 12 bp is about 1% of the haplotypes' length. We consider the haplotype $h \in P$ to be reconstructed correctly if there exists a haplotype $f \in Q$ such that the edit distance between them $d(h, f) \leq 12$.

For *de novo* methods, we define *TP* as follows: We say that a contig f from Q is *proper* if there exists such a ground truth haplotype h and its substring $s \in h$ so that the edit distance between f and

s is small (less than 1% of f 's length). Then, for each ground truth haplotype h_i, we define c_i – the proportion of its part which is *properly* covered by contigs from Q. We then define *TP* as a weighted total frequency of properly predicted haplotypes $\sum_{h_i \in P} c_i f_i$. It is important to note that the definition of *TP* is a generalization of the *TP* definition for reference-based tools. Indeed, in the latter case, all the c_i are equal to either 0 or 1. We define *FP* as the total frequency of non-proper contigs in Q.

While these measures are standard and they show how good the haplotype reconstruction is, they are not very sensitive to the errors in frequency prediction. In order to address this issue, we also computed the UniFrac distance $EMD(P,Q)$ using the EMDUniFrac algorithm (McClelland and Koslicki, 2018). The UniFrac distance takes into account both the phylogenetic structure of the haplotype set and their frequency distribution, which makes it ideal for incorporating sensitivity to errors in frequency prediction. The UniFrac EMD method makes the following steps:

- construct a tree T with branch length l_e on the set of all haplotypes $h_i \in P$ and $f_i \in Q$
- for each tree branch e and its descendant subtree T_e, estimates the imbalance W_e:

$$W_e := \left| \sum_{i:h_i \in T_e} p_i - \sum_{i:f_i \in T_e} q_i \right|;$$

- evaluate the weighted imbalance with respect to the branch lengths

$$EMD := \sum_{e \in T} l_e W_e.$$

As a baseline for the UniFrac EMD comparison, we evaluate the UniFrac distance between reference or, more formally, a set of haplotypes Q containing only one haplotype – the reference at a frequency of 1.

3. Results and Discussion

3.1 True haplotypes from simulated data

All analyses were completed using the simulated dataset developed under the coalescent framework. For each mutation rate $\mu \in$ {1e-8, 3e-8, 5e-8, 1e-7, 5e-7, 1e-6, 5e-6, 1e-5, 3e-5, 5e-5, 1e-4, 3e-4, 5e-4, 5e-3, 5e-8} and effective population size $N_e = \{500, 1000, 2500, 5000, 7500, 10000\}$, there were five simulated haplotype populations $P = \{(h_i, p_i), i = 1, 2, \ldots\}$ used as replicates for each parameter set. Under the coalescent model, the number of true haplotypes ranged from 1 to 1,993 with a median of 342 haplotypes for a parameter set (Fig. 2). Unlike previous attempts to represent intra-host HIV-1 diversity levels – often five haplotypes at varying abundances (J A Baaijens et al., 2017; Prabhakaran et al., 2014; Töpfer et al., 2014), our intra-host populations have 216-1,185 haplotypes per host at a frequency <7%, with a median of 525 haplotypes. Therefore, the number of haplotypes at high diversity levels may actually be even higher, but we primarily focused on the diversity levels of intra-host HIV-1 populations. Additionally, the number of haplotypes at smaller diversity levels, such as those seen in influenza, are likely to be smaller than ours.

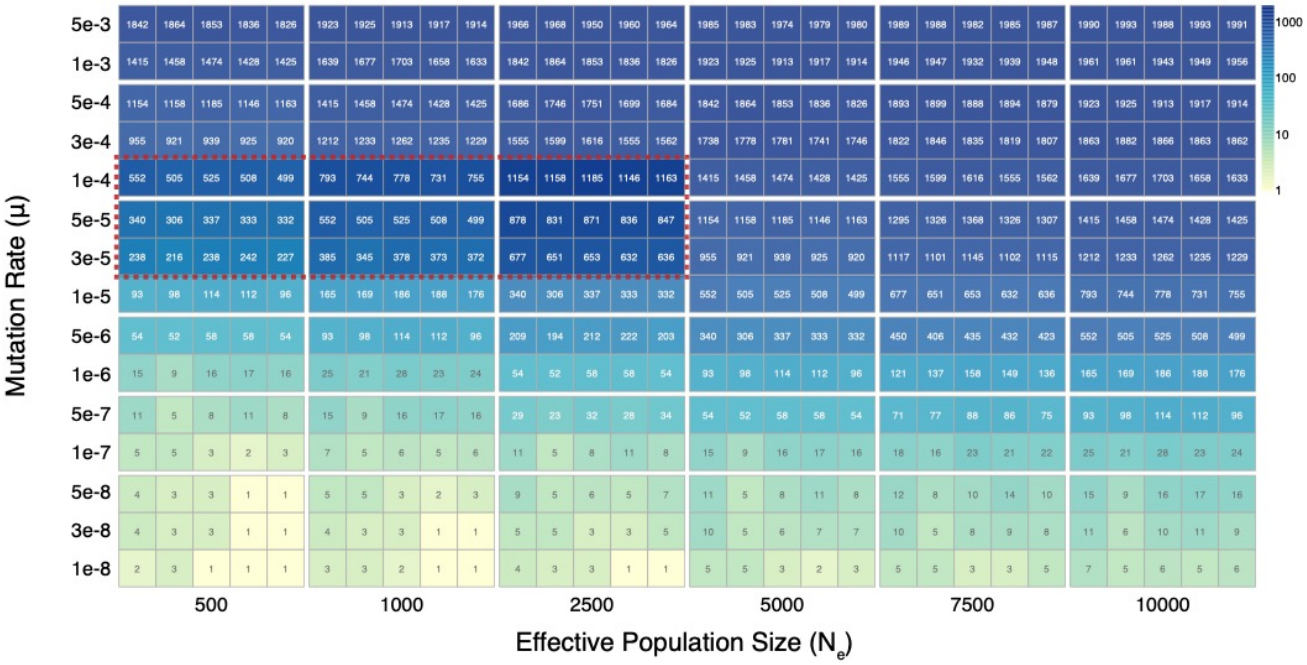

Figure 2. Simulated population parameters with the haplotype count in each parameter box. All five population replicates are displayed. The color darkens as the number of haplotypes increases. The red dashed box indicates expected HIV-1 mutation rates and effective population sizes.

3.2 Haplotype caller performance

HIV-1 intra-patient populations exhibit levels of diversity that exceed the limitations of all twelve haplotype callers we compared in this study, regardless of the assembly approach used (*de novo* or *reference-based*). However, because HCV and influenza both have lower mutation rates and effective population sizes, they may fall within the limitations of some of the compared haplotype reconstruction approaches. The haplotype callers varied drastically in their haplotype reconstruction accuracy (precision, recall, UniFrac, and number of reconstructed haplotypes), with most tools performing well with low genetic diversity and poorly with high genetic diversity. Since HIV-1 diversity is very high, all haplotype reconstruction tools seemed to have difficulties either producing output (i.e., predicted haplotypes) or reconstructing haplotypes that reflect the true haplotypes. Furthermore, haplotype reconstruction accuracy was more sensitive to the mutation rate of the virus than to its effective population size. Although, the opposite was true for PEHaplo, where N_e seemed to play a major role in the quality of predicted haplotypes. Fortunately, we often know, or have better *a priori* estimates for, the mutation rate of a virus than for the effective population size of an intra-host population. Furthermore, the effective population size changes over time during infection, whereas the mutation rate remains relatively constant (Maldarelli et al., 2013), unless there are pressures from antiretroviral treatment. However, as theta is estimated, effective population size and mutation rate are indeed coundfounded. Below, we discuss the current results in more detail.

MLEHaplo and ViQuas did not produce any valid results within the given time limit, whereas QuRe crashed in all analyses because of memory limitations. While HaploClique produced results within our time limit (Fig. S1), we excluded this tool from final comparisons because the length of the reconstructed viral sequences was always significantly shorter than the length of the ground truth haplotypes (Fig. S2). Such behavior is atypical among reference-based methods. Moreover, since SAVAGE can be considered as the next installment of HaploClique, it provides an additional argument for excluding HaploClique from our comparison.

In addition to the two *de novo* tools assessed (i.e., SAVAGE and PEHaplo), we also ran the VG-flow tool to complete contigs produced by those methods. We selected VG-flow over Virus-VG because VG-flow is faster and almost as accurate (Baaijens et al., 2018, 2019). Despite the claim that

VG-flow improves assemblies from any *de novo* tool (Baaijens et al., 2018, 2019), we also included the output from PEHaplo in our comparison since VG-flow was tested on the SAVAGE output only in the original paper (Baaijens et al., 2018, 2019).

Datasets completed within our time limits varied across reconstruction tools; in general, datasets with higher mutation rates (μ) and effective population sizes (N_e) represented challenges for almost all the tools. For example, RegressHaplo and PredictHaplo did not produce any output if both N_e and μ were high but performed well otherwise. For low values of μ ($\mu \leq 1e-5$), all the callers except ShoRAH produced some output. aBayesQR, SAVAGE, and RegressHaplo had problems reconstructing haplotypes for datasets with low μ and N_e values.

For HIV-1 μ estimates ($3e-5 \leq \mu \leq 1e-4$), CliqueSNV and QuasiRecomb did not produce any valid output; aBayesQR, PredictHaplo, SAVAGE, and RegressHaplo produced some outputs; and ShoRAH, and PEHaplo produced output for all the datasets. It is also important to note that only PEHaplo was able to solve all datasets within the given time limit. For all datasets, where haplotype callers performed successfully, we measured their results in terms of precision, recall, and the Unifrac distance EMD. Below we present and discuss the behavior of the *reference-based* tools and the *de novo* tools.

3.2.1 Reference-based Program Performance

We evaluated results from six *reference-based* haplotype callers: aBayesQR, RegressHaplo, CliqueSNV, ShoRAH, PredictHaplo, and QuasiRecomb. Precision (Fig. 3) and recall (Fig. 4) values were calculated for each tool. The quality of obtained results did not seem to depend much on the effective population size (N_e). This is a positive finding, as determining the effective population size for intra-host viral infections is often difficult and can vary between studies. All the tools, except ShoRAH, performed very well (i.e., both precision and recall are close to one) if the mutation rate was relatively small ($\mu \leq 1e-5$), which is an estimated mutation rate for influenza. For higher values of μ ($\mu \geq 1e-4$), such as those seen in HCV and HIV-1, all the tools performed poorly (i.e., both precision and recall were close to zero). For the values of μ seen in HIV-1 ($3e-5 \leq \mu \leq 1e-4$), PredictHaplo was able to produce better results than the other tools; PredictHaplo's precision and recall decreased with $\mu \in (3e{-}5, 1e{-}4)$ but stayed positive. It also should be noted that CliqueSNV outperformed all other tools for $\mu = 1e-6$, but did not produce any results for $\mu \in (1e-5, 1e-4)$. Such behavior looks promising and it is possible that in future releases, if run-time is increased, CliqueSNV will exceed PredictHaplo in precision and recall performance.

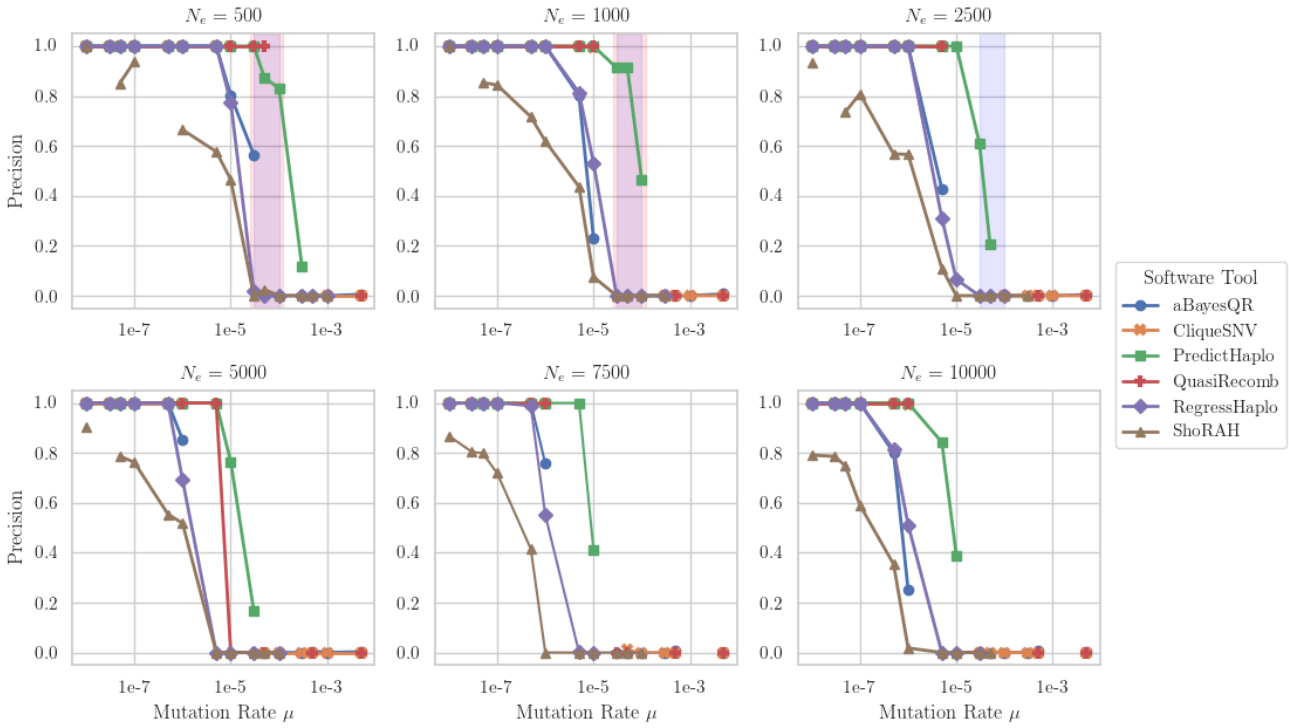

Figure 3. Reference-based haplotype callers: variation of precision values with mutation rate (log-scaled) for all considered N_e. The shaded light blue and shaded light red regions correspond to HIV-1 and HCV diversity levels, respectively. For all pairs of parameters μ and N_e, we report the mean estimates of precision over all valid outputs produced by each software tool for the five haplotype populations. If a tool did not produce any output for any pair of parameters, we included a gap in the corresponding plot.

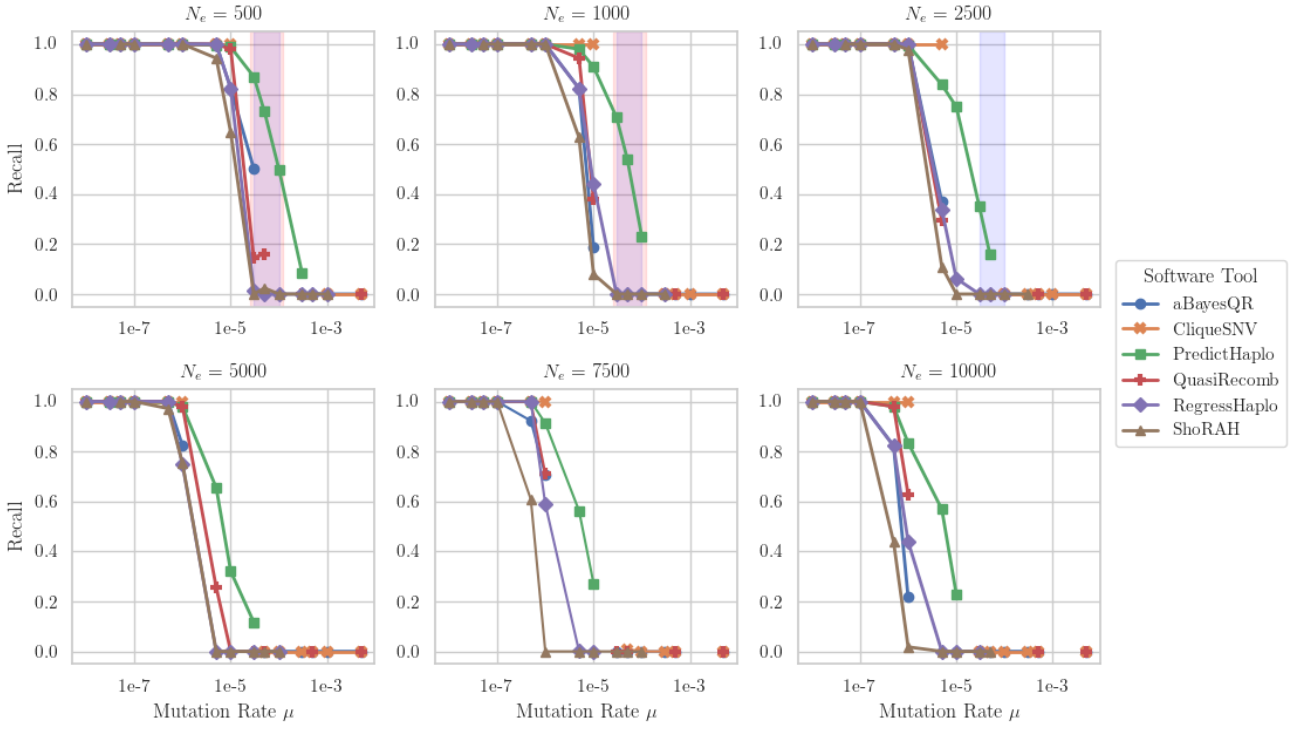

Figure 4. Reference-based haplotype callers: variation of recall values with mutation rate (log-scaled) for all considered N_e. The shaded light blue and shaded light red regions correspond to HIV-1 and HCV diversity levels, respectively. For all pairs of parameters μ and N_e, we report the mean estimates of recall over all valid outputs produced by each software tool for five haplotype populations. If a tool did not produce any output for some pair of parameters, we included a gap in the corresponding plot.

We calculated UniFrac distance values for the aforementioned tools (Fig. 5). The UniFrac distance further supported the previous observation that the quality of obtained results does not depend much on the effective population size (N_e). Comparisons using the UniFrac distance also showed that all the tools, except ShoRAH, performed well if $\mu \leq 1e-5$; the UniFrac distance between the ground truth sets of haplotypes and those predicted by the tool sets are all close to zero. With increasing μ values, UniFrac distances also increased. For HIV-1 mutation rates, PredictHaplo showed the best performance since it produced outputs for almost all pairs of parameters and the sets of predicted haplotypes were the closest to the correct haplotypes. Again, CliqueSNV outperformed all other methods for $\mu = 1e-6$, which further supports our previous observation.

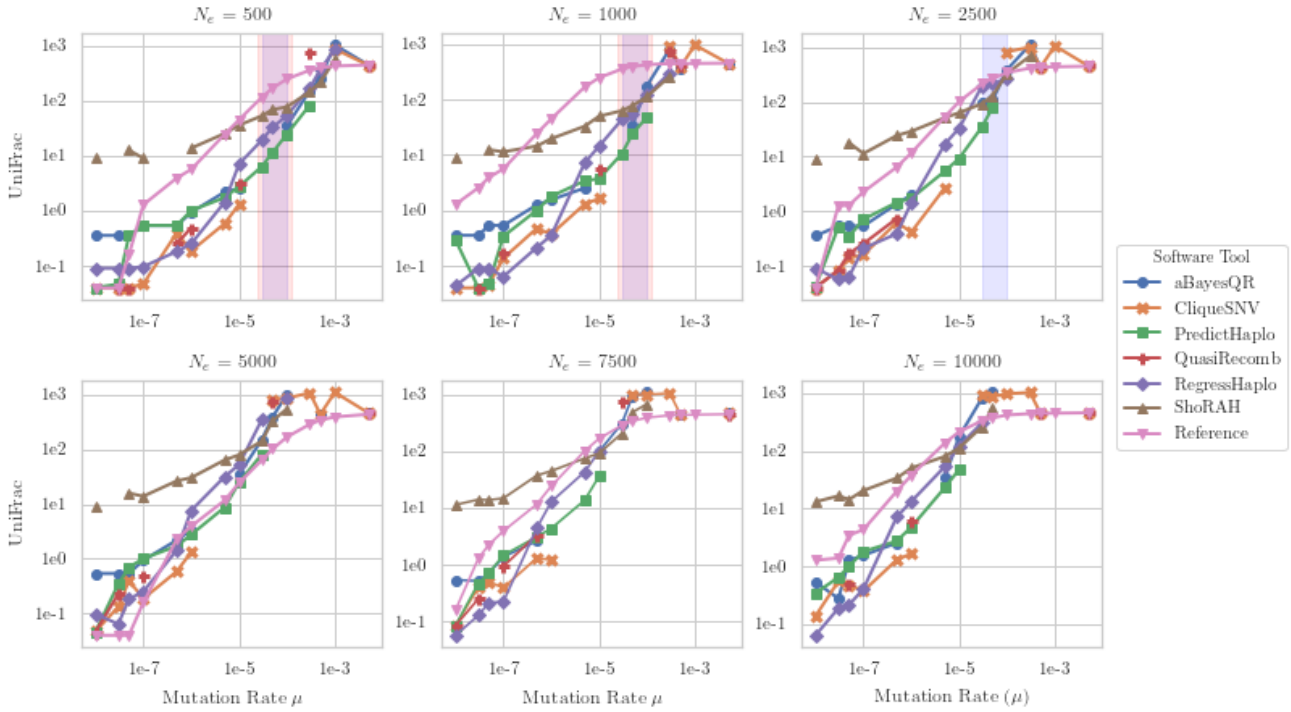

Figure 5. Reference-based haplotype callers: variation of UniFrac distances (EMD) with mutation rate (log-scaled) for all considered N_e. The shaded light blue and shaded light red regions correspond to HIV-1 and HCV diversity levels, respectively. For all pairs of parameters μ and N_e, we report the mean estimates of UniFrac distances over all valid outputs produced by each software tool for five haplotype populations. If a tool did not produce any output for some pair of parameters, we included a gap in the corresponding plot.

For large values of μ ($\mu \geq 1e-4$), ShoRAH, QuasiRecomb, RegressHaplo and PredictHaplo rarely produced a valid output within the given time limit. aBayesQR and CliqueSNV produced results that were worse than or comparable to the baseline. For large values of the effective population size ($N_e \geq 5000$) and low values of μ, all the tools except ShoRAH showed better results than the baseline. However, for mutation values larger than $5e-4$, none of the tools made a better prediction of the set of haplotypes than just a reference. It is important to note that HCV, HIV, and influenza do not have

N_e close to 5,000 (Bernini et al., 2011; Kim and Kim, 2016; Maldarelli et al., 2013; McCrone, 2018; McCrone et al., 2018). Most methods severely underestimated the true number of haplotypes in a population at high genetic diversity levels or overestimated it at low genetic diversity levels (Fig. 6), compared to the true number of haplotypes across the same levels of underlying genetic diversity obtained from the simulated datasets (Fig. S3). PredictHaplo, RegressHaplo, aBayesQR, and CliqueSNV underestimated haplotype numbers in the HIV intra-host diversity range (shaded in yellow). HaploClique and QuasiRecomb, on the other hand, overestimated haplotype numbers, whereas ShoRAH provided the closest estimate to the true number of haplotypes in the HIV-1 diversity range. aBayesQR and CliqueSNV did not produce results for any dataset within the HIV-1 diversity range.

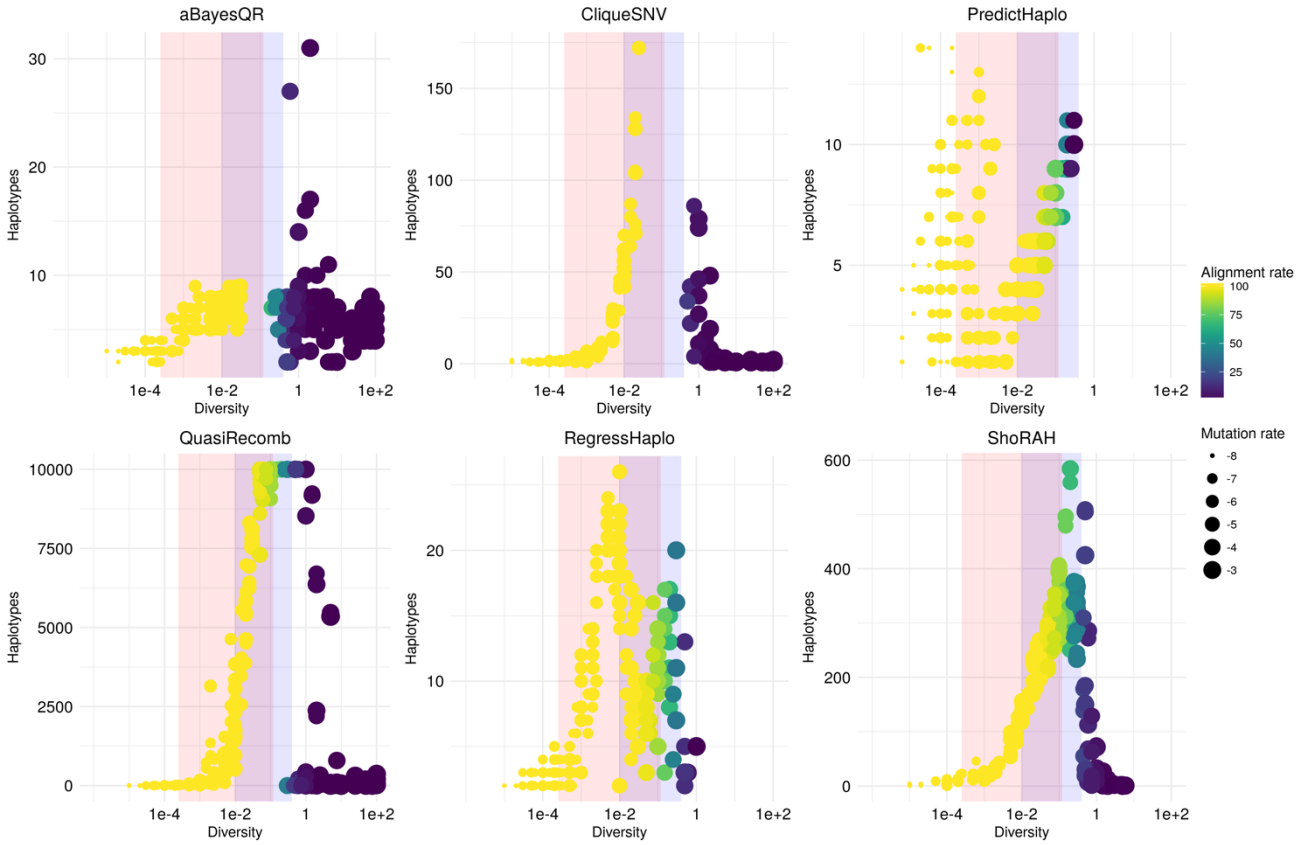

Figure 6. Reference-based haplotype callers: number of predicted haplotypes across levels of underlying genetic diversity. Intra-host HIV-1 and HCV diversity levels are highlighted shaded light blue and shaded light red regions, respectively. If a software tool did not complete haplotype reconstruction within the given time frame, we included a gap in the corresponding plot (see Fig. S1 for more information on dataset completions).

3.2.2 De Novo Program Performances

We analyzed the behavior of two *de novo* haplotype callers: SAVAGE and PEHaplo. The output of both tools usually contained shorter contigs, so we completed the assembly using the VG-flow tool. PEHaplo itself produced valid output for all the datasets, while SAVAGE or PEHaplo with VG-flow failed to produce results for some datasets (Fig. S1). Moreover, the length of PEHaplo output haplotypes was usually closer to the ground truth haplotype length, while the SAVAGE+VG-flow produced shorter contigs (see N50 statistics plot on Fig. S4). Thus, we only further considered PEHaplo, PEHaplo + VG-flow and SAVAGE + VG-flow.

We compared the *de novo* tools using our modified versions of precision and recall (Fig. 7 and Fig. 8). VG-flow usually improved slightly the performance of PEHaplo, while PEHaplo usually performed better than SAVAGE+VG-flow. Although the quality of results of SAVAGE+VG-flow did not seem to depend on the effective population size, N_e played a role in the quality of obtained results by PEHaplo. For example, both precision and recall were close to zero for $\mu = 1e-8$ and $N_e \in \{500, 1000, 2500\}$, but significantly higher for $N_e \in \{5000, 7500, 10000\}$ and $\mu = 1e-8$. It is also important to note the behavior of recall values for the obtained results in PEHaplo; those values, in general, were close to one for small μ values, close to zero for μ values near $1e-5$, and stayed positive for higher μ values.

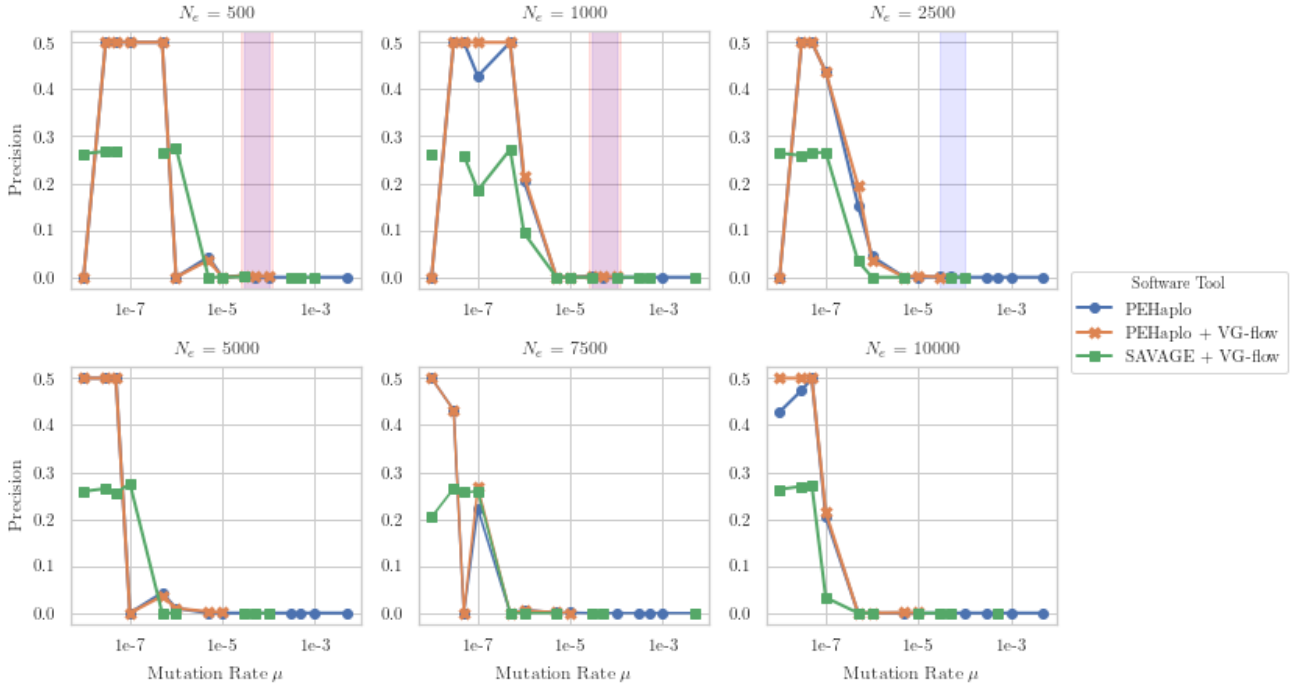

Figure 7. *De novo* haplotype callers: variation of precision values with mutation rate (log-scaled) for all considered N_e. The shaded light blue and shaded light red regions correspond to HIV-1 and HCV diversity levels, respectively. For all pairs of parameters μ and N_e, we report the mean estimates of precision over all valid outputs produced by each software tool for five haplotype populations. If a tool did not produce any output for some pair of parameters, we included a gap in the corresponding plot.

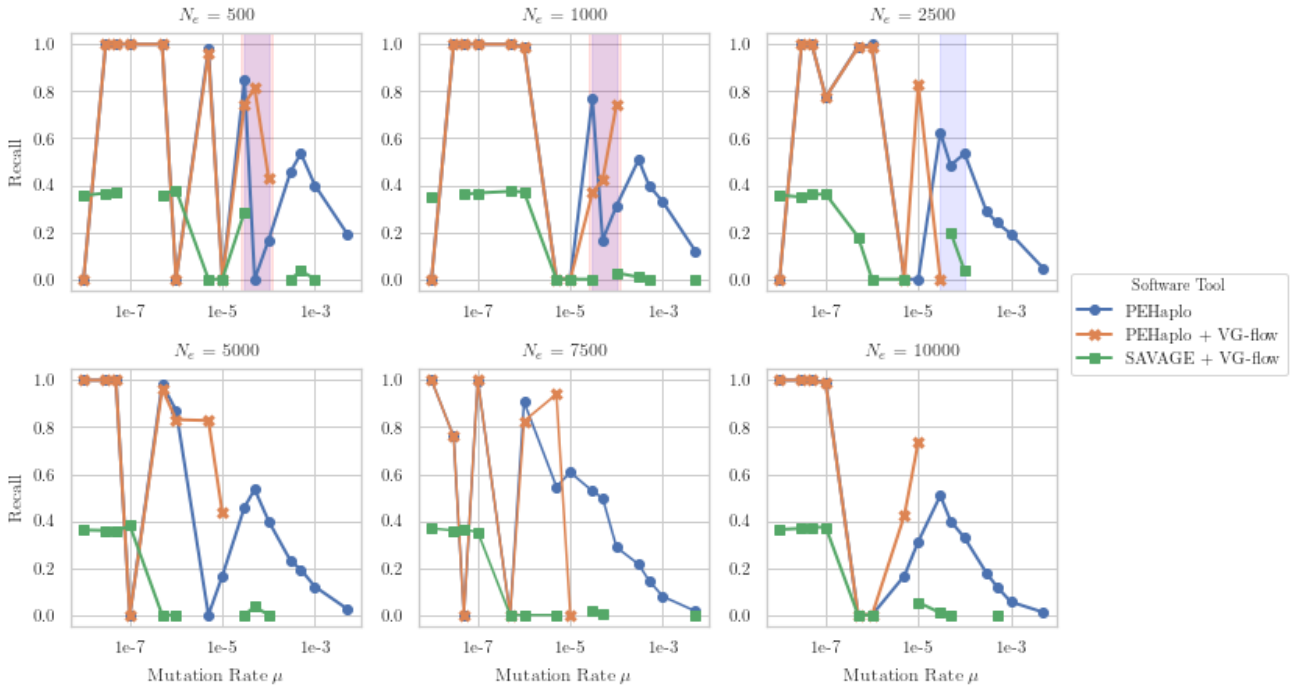

Figure 8. *De novo* haplotype callers: variation of recall values with the mutation rate (log-scaled) for all considered N_e. The shaded light blue and shaded light red regions correspond to HIV-1 and HCV diversity levels, respectively. For all pairs of parameters μ and N_e, we report the mean estimates of recall over all valid outputs produced by each software tool for five haplotype populations. If a tool did not produce any output for some pair of parameters, we included a gap in the corresponding plot.

De novo tools performed very well, in both precision and recall values, if the mutation rate was less than $1e-6$ (in contrast to $\mu \leq 1e-5$ for *reference-based* tools). Additionally, recall values for PEHaplo when $\mu \geq 1e-4$ were usually better than those seen for any *reference-based* approaches. *De novo* tools did not produce results with a positive precision for HIV-1 and HCV mutation rates. The UniFrac distance further confirmed our previous observation that VG-flow slightly improved the performance of PEHaplo (Fig. 9). Moreover, the performance of SAVAGE + VG-flow did not depend on the mutation rate or the effective population size N_e. It is important to note that all UniFrac distance values were, in general, higher than baseline values. We also compared UniFrac distances between both categories of assemblers (Fig. S5); as we expected, *reference-based* tools largely outperformed *de novo* tools. At the same time, PEHaplo performed better than ShoRAH for some datasets. Moreover, SAVAGE + VG-flow showed the worst performance based on UniFrac distances.

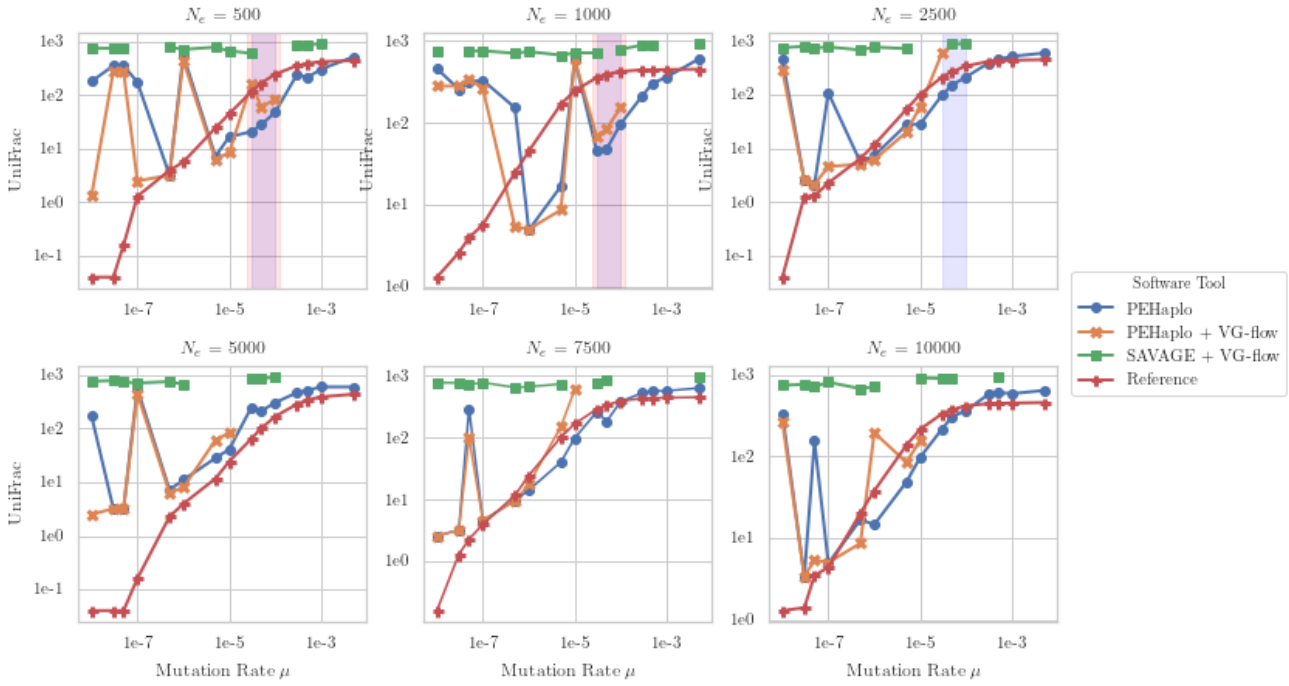

Figure 9. *De novo* haplotype callers: variation of UniFrac distance (EMD) with mutation rate (log-scaled) for all considered N_e. The shaded light blue and shaded light red regions correspond to HIV-1 and HCV diversity levels, respectively. For all pairs of parameters μ and N_e, we report the mean estimates of UniFrac distances over all valid outputs produced by each software tool for five haplotype populations. If a tool did not produce any output for some pair of parameters, we included a gap in the corresponding plot.

Although the *de novo* methods produced more haplotypes in the HIV-1 diversity range compared to *reference-based* methods, they all still underestimated the true number of haplotypes in a population at higher diversity levels. They also overestimated true haplotype numbers at lower genetic diversity levels (Fig. 10) compared to the true number of haplotypes from the simulated datasets (Fig. S3). When extending the contigs into scaffolds with VG-flow, the number of haplotypes reconstructed decreased considerably and remained below the number of true haplotypes estimated for the varying genetic diversity levels. PEHaplo reconstructed the lower limit of the true number of haplotypes within HIV-1 diversity levels, but like other tools, including aBayes, CliqueSNV and QuasiRecomb, PEHaplo and SAVAGE, had trouble reconstructing viral sequences at higher diversity levels.

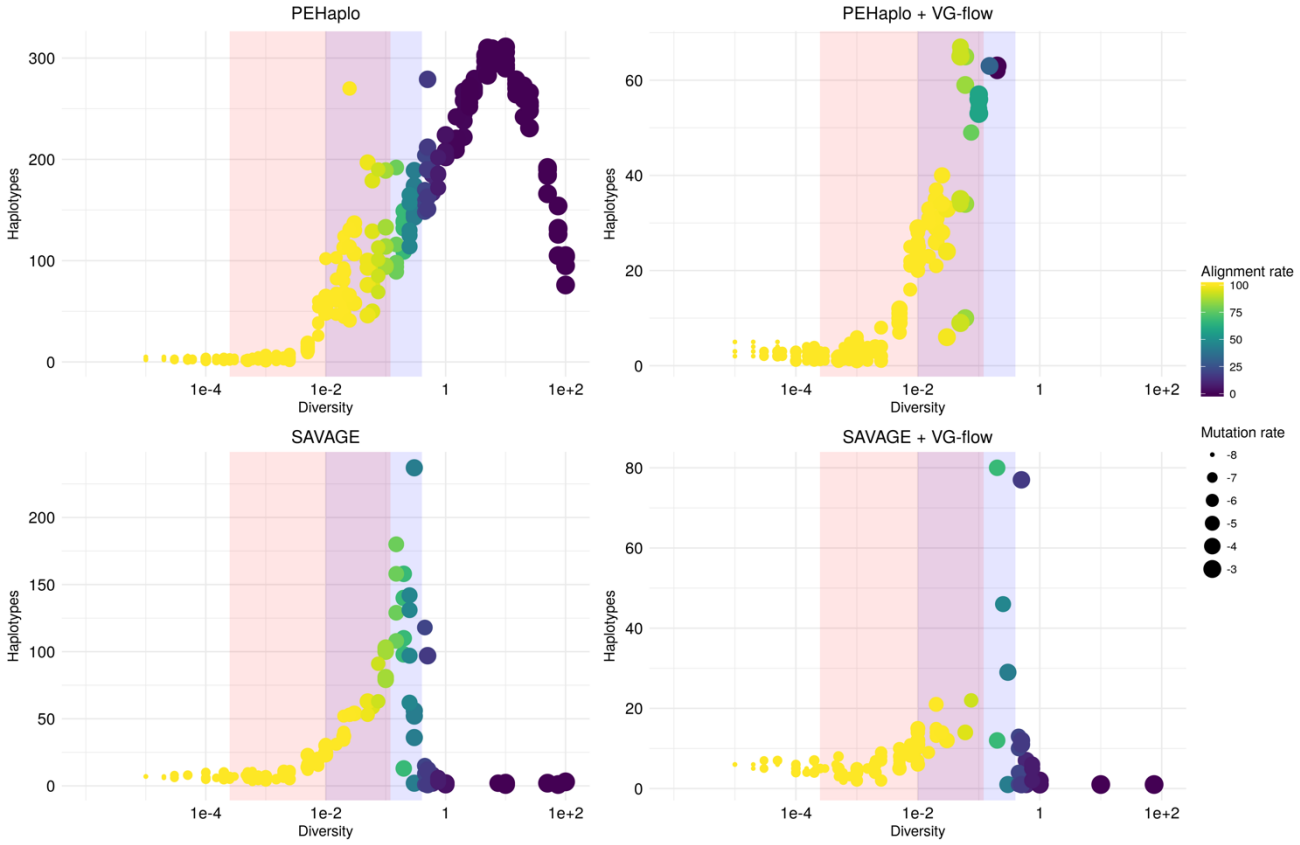

Figure 10. *De novo* haplotype callers: variation in number of predicted haplotypes across levels of underlying genetic diversity. Intra-host HIV-1 and HCV diversity levels are highlighted shaded light blue and shaded light red regions, respectively. If a software tool did not complete haplotype reconstruction within the given time frame, we included a gap in the corresponding plot (see Fig. S1 for more information on dataset completions).

4. Conclusions and Future Directions

We compared twelve of the most commonly used software tools to reconstruct haplotypes from viral NGS data. We simulated coalescent-based populations that spanned past known levels of viral diversity, including mutation rates, sample size, and effective population size. We focused our empirical comparisons on the intra-host diversity levels of fast-evolving RNA viruses such as HIV-1 because parameter value ranges are well established and a better understanding of viral dynamics is important for drug and vaccine development. Additionally, the majority of haplotype tool developers used HIV-1 to validate their own programs. In our analyses of HIV-1 intra-host diversity, we estimated between 216 and 1,185 haplotypes with a <7% frequency for a single haplotype.

Overall, *reference-based* assemblers produced more accurate haplotypes than *de novo*-based assemblers for all performance indices (precision, recall, UniFrac, and number of reconstructed haplotypes) across HIV-1 diversity levels. This performance could be attributed to the availability of high-quality reference sequences for HIV-1, HIV-2, HCV, influenza and other viruses. Furthermore, using a reference sequence reduces the computational time and power needed to reconstruct haplotypes. Reference-based assemblers likely performed better than *de novo* assemblers because of the high variation within viral populations, especially HIV-1, where the reference sequence may have provided needed guidance to orient the highly diverse NGS sequences into a haplotype sequence.

Our results show that PredictHaplo offers the best tradeoff between statistical performance and computational efficiency within HIV-1 diversity ranges. PredictHaplo was found to have the highest precision, recall, and lowest UniFrac distance values. CliqueSNV followed closely and may actually

outperform PredictHaplo if more computational resources were made available. An important caveat for both these approaches, however, is that the number of true haplotypes is greatly underestimated. If it is important to identify the true number of haplotypes (as in rare haplotype discovery) approaches such as ShoRAH or PEHaplo may be more appropriate. The haplotype programs also varied greatly in terms of their ease-of-use. This variation is due to differences in coding language, program dependencies, availability of executable files, absence of comprehensive documentation and lack of example datasets. For example, SAVAGE, PEHaplo, ShoRAH can be easily installed by package managers, and CliqueSNV and QuasiRecomb are distributed as executable files. In contrast, Virus-VG and VG-flow requires installment of proprietary software, which has an academic license. Installation and usage of PredictHaplo is challenging because of the lack of description and instructions regarding the config file. While CliqueSNV is easier to install and use, there are no example datasets.

It is important to note that our study represents an initial attempt of comprehensive comparison of available haplotype reconstruction tools. For example, we focused HIV-1 diversity estimates for the polymerase gene, which is less variable than the envelope gene. Moreover, almost all developers of the aforementioned tools used the polymerase gene as a source of simulating sequencing data for assessing performance of their programs and rarely used the envelope gene for the same purposes. Given the envelope gene has a higher mutation rate and the haplotype reconstruction tools – *de novo* or *reference-based* – seem to be dependent on mutation rate, it is likely that the tools available here would not be successful in reconstructing envelope haplotypes for HIV-1 accurately. However, we chose polymerase as our gene of interest because of research focus on this gene as the target for drug resistance mutations. The same concept of lower mutation rates in conserved genes and higher mutation in less conserved genes can be seen in other fast-evolving viruses. For example, in HCV the core protein is more conserved compared to the E1/E2 region. Thus, users should target methods for haplotype calling that best match the mutation rate of their target gene.

Another limitation of our study is coverage. It is well-known that coverage plays a crucial role in all algorithms for distinguishing between errors and rare sequence variants. We chose 100x coverage because it represents a reasonable amount of data that can be obtained without intensive labor or money consuming procedures. Contrary to our simulations, the developers of haplotype reconstruction tools usually test their methods on datasets with higher coverage than ours. For example, the famous golden-standard benchmark HIV dataset (*labmix* dataset (Di Giallonardo et al., 2014)) on which all tools have been tested by developers, consisted of an average of 20,000x coverage. Thus, our study represents an attempt to measure the performance of the haplotype reconstruction tools on datasets that are more likely to be seen and produced in laboratories. Moreover, according to our results for higher mutation rates, many tools did not produce any results within the time limit. Considering that higher coverage implies a larger amount of data and thus requiring more computational time to process these data, it is expected that the tools available here would require extensive computational resources.

We also considered error-free and recombination-free data in our study. Only a few tools explicitly took into account the presence of errors or recombination in their models (e.g., only QuasiRecomb explicitly assumes the presence of recombination events). By not simulating recombination and sequencing errors, we removed nuisance parameters that would impact haplotype reconstruction. Moreover, since almost all tools have been tested on ultra-deep data, our comparison study by error-free data is giving an advantage to these methods by removing errors in sequence data (i.e., one does not need deep coverage to distinguish between rare variants and errors). Furthermore, Zanini et al. (2015) found evidence that recombination likely interrupts haplotypes, specifically in HIV-1, every 100-200bp, so, the concept of haplotypes in HIV-1, and maybe other fast-evolving viruses with high recombination rates, may not exist or be feasible to study with frequent recombination events. Together these facts imply that the performance of the aforementioned tools would be even worse than observed here.

Overall, results and limitations of our study indicate the importance of creating broad and diverse golden-standard datasets that must include several different genes, diverse parameters of

mutation rates and effective population sizes, different average coverages, presence or absence of recombination events or/and error prone data. Moreover, future simulation studies should address error-prone data using haplotype callers that can handle sequencing errors and investigate the effect of recombination and average coverage on the reconstruction of haplotype. In addition to simulation studies, some theoretical work similar to DNA sequencing theory should be done for laying analytical foundations for determining coverage depending on the mutation rate, effective population size, error-rate of a sequencing technology, and so on. Finally, there are still a lot of opportunities for developing new haplotype callers that can process a wide range of data with different mutation rates, average coverage, and presence or absence of recombination events. Moreover, since the reconstructed haplotypes are often used for reconstructing phylogeny, the future tools may also consider the problem of reconstructing haplotype sequences together with their phylogeny. Considering the possibility that the reconstructing haplotype sequences from short-read sequencing technologies may represent an intractable problem, focusing on reconstructing haplotype phylogeny directly from short-reads may lead to better results after all. In addition to mentioned future directions, the advances and price-decreasing of long-read sequencing technologies (e.g., Nanopore, PacBio, 10X Genomics) poses a whole new set of challenges for haplotype reconstructions including the development of new sequencing protocols and haplotype reconstruction tools. This new technology has the power to sequence long amplicons or even entire viral genomes in a single pass, i.e., no need to assemble sequencing reads. However, this type of data requires development of new methods that can distinguish between rare variants and sequencing errors. Therefore, the application of long-read sequencing technology may be more beneficial for studying global, or entire genome, haplotypes.

Acknowledgements

The authors would like to thank Adam Kai Leung Wong for his invaluable help installing and troubleshooting different software packages in the Colonial One high performance computing cluster at The George Washington University.

Funding

This study was supported by a DC D-CFAR pilot award. MPL was supported by a 2015 HIV Phylodynamics Supplement award from the District of Columbia for AIDS Research, an NIH funded program (AI117970), and NIH grants AI076059 and UL1TR001876. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH. The work of NA is supported by the Government of the Russian Federation through the ITMO Fellowship and Professorship Program.

Authors' contributions

KMG, MLB, and KAC developed project idea. KAC and MP-L acquired funding for this study. KMG and MLB completed the simulation study. AE and DN ran all the haplotype reconstruction tools. AE, PA, and NA performed the statistical analysis of the results. PA and NA supervised AE and DN. PA, NA and KMG developed initial manuscript draft. All authors read and reviewed final draft of manuscript.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interest

The authors declare that they have no competing interests.

References

Ahn, S., Vikalo, H., 2017. aBayesQR: A Bayesian Method for Reconstruction of Viral Populations Characterized by Low Diversity. J. Comput. Biol. 25, 637–648. https://doi.org/10.1089/cmb.2017.0249

Alnafee, K., 2016. Versatile and open software for comparing large genomes 5, 12–23. https://doi.org/10.1186/gb-2004-5-2-r12

Arenas, M., Posada, D., 2014. Simulation of genome-wide evolution under heterogeneous substitution models and complex multispecies coalescent histories. Mol. Biol. Evol. 31, 1295–1301. https://doi.org/10.1093/molbev/msu078

Artyomenko, A., Wu, N.C., Mangul, S., Eskin, E., Sun, R., Zelikovsky, A., 2017. Long Single-Molecule Reads Can Resolve the Complexity of the Influenza Virus Composed of Rare, Closely Related Mutant Variants. J. Comput. Biol. 24, 558–570. https://doi.org/10.1089/cmb.2016.0146

Astrovskaya, I., Tork, B., Mangul, S., Westbrooks, K., Măndoiu, I., Balfe, P., Zelikovsky, A., 2011. Inferring viral quasispecies spectra from 454 pyrosequencing reads. BMC Bioinformatics 12. https://doi.org/10.1186/1471-2105-12-S6-S1

Baaijens, J., Roest, B. Van Der, Koester, J., Stougie, L., Baaijens, J., Roest, B. Van Der, Koester, J., Stougie, L., Schoenhuth, A., Baaijens, J.A., Roest, B. Van Der, Johannes, K., 2018. Full-length de novo viral quasispecies assembly through variation graph construction. bioRxiv.

Baaijens, J A, Aabidine, A.Z.E., Rivals, E., Schonhuth, A., 2017. De novo assembly of viral quasispecies using overlap graphs. Genome Res 27, 835–848. https://doi.org/10.1101/gr.215038.116

Baaijens, J.A., Stougie, L., Schönhuth, A., 2019. Strain-aware assembly of genomes from mixed samples using variation graphs. bioRxiv 645721. https://doi.org/10.1101/645721

Baaijens, Jasmijn A, Zine, A., Aabidine, E., Rivals, E., Schönhuth, A., Wiskunde, C., Amsterdam, X.G., Montpellier, U. De, 2017. De novo assembly of viral quasispecies using overlap graphs. Genome Res. 27, 835–848. https://doi.org/10.1101/gr.215038.116.4

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A.A., Dvorkin, M., Kulikov, A.S., Lesin, V.M., Nikolenko, S.I., Pham, S., Prjibelski, A.D., Pyshkin, A. V, Sirotkin, A. V, Vyahhi, N., Tesler, G., Alekseyev, M.A., Pevzner, P.A., 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J. Comput. Biol. 19, 455–77. https://doi.org/10.1089/cmb.2012.0021

Barik, S., Das, S., Vikalo, H., 2018. QSdpR: Viral quasispecies reconstruction via correlation clustering. Genomics 110, 375–381. https://doi.org/10.1016/j.ygeno.2017.12.007

Beerenwinkel, N., Zagordi, O., 2011. Ultra-deep sequencing for the analysis of viral populations. Curr. Opin. Virol. 1, 413–418. https://doi.org/10.1016/j.coviro.2011.07.008

Bendall, M.L., Gibson, K.M., Steiner, M.C., Pérez-Losada, M., Keith A. Crandall, 2019. HAPHPIPE: Haplotype reconstruction and real-time phylodynamics for deep sequencing of intra-host viral populations. Mol. Biol. Evol.

Bernini, F., Ebranati, E., De Maddalena, C., Shkjezi, R., Milazzo, L., Presti, A. Lo, Ciccozzi, M., Galli, M., Zehender, G., 2011. Within-host dynamics of the hepatitis C virus quasispecies population in HIV-1/HCV coinfected patients. PLoS One 6, 1–11. https://doi.org/10.1371/journal.pone.0016551

Bishara, A., Moss, E.L., Kolmogorov, M., Parada, A.E., Weng, Z., Sidow, A., Dekas, A.E., Batzoglou, S., Bhatt, A.S., 2018. High-quality genome sequences of uncultured microbes by assembly of read clouds. Nat. Biotechnol. 36, 1067–1080. https://doi.org/10.1038/nbt.4266

Bolger, A.M., Lohse, M., Usadel, B., 2014. Trimmomatic: A flexible trimmer for Illumina sequence data. Bioinformatics 30, 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Boltz, V.F., Rausch, J., Shao, W., Hattori, J., Luke, B., Maldarelli, F., Mellors, J.W., Kearney, M.F., Coffin, J.M., 2016. Ultrasensitive single - genome sequencing : accurate , targeted , next

772 generation sequencing of HIV - 1 RNA. Retrovirology 1–17. https://doi.org/10.1186/s12977-
773 016-0321-6
774 Bray, N.L., Pimentel, H., Melsted, P., Pachter, L., 2016. Near-optimal probabilistic RNA-seq
775 quantification. Nat. Biotechnol. 34, 525–7. https://doi.org/10.1038/nbt.3519
776 Chen, J., Zhao, Y., Sun, Y., 2018. De novo haplotype reconstruction in viral quasispecies using
777 paired-end read guided path finding. Bioinformatics 34, 2927–2935.
778 https://doi.org/10.1093/bioinformatics/bty202
779 Coffin, J.M., 1992. Genetic Diversity and Evolution of Retroviruses, in: Holland J.J. (Eds) Genetic
780 Diversity of RNA Viruses. Current Topics in Microbiology and Immunology, Vol 176.
781 Springer, Berlin, Heidelberg, pp. 143–164. https://doi.org/10.1007/978-3-642-77011-1_10
782 Compeau, P.E.C., Pevzner, P.A., Tesler, G., 2011. How to apply de Bruijn graphs to genome
783 assembly. Nat. Biotechnol. 29, 987–991. https://doi.org/10.1038/nbt.2023
784 Crandall, K.A., 1999. The Evolution of HIV. The Johns Hopkins University Press, Baltimore.
785 Crandall, K.A., Kelsey, C.R., Imamichi, H., Lane, H.C., Salzman, N.P., 1999. Parallel evolution of
786 drug resistance in HIV: Failure of nonsynonymous/synonymous substitution rate ratio to detect
787 selection. Mol. Biol. Evol. 16, 372–382. https://doi.org/10.1093/oxfordjournals.molbev.a026118
788 Crandall, K.A., Templeton, A.R., 1993. Empirical tests of some predictions from coalescent theory
789 with applications to intraspecific phylogeny reconstruction. Genetics 134, 959–969.
790 Di Giallonardo, F., Töpfer, A., Rey, M., Prabhakaran, S., Duport, Y., Leemann, C., Schmutz, S.,
791 Campbell, N.K., Joos, B., Lecca, M.R., Patrignani, A., Däumer, M., Beisel, C., Rusert, P.,
792 Trkola, A., Günthard, H.F., Roth, V., Beerenwinkel, N., Metzner, K.J., 2014. Full-length
793 haplotype reconstruction to infer the structure of heterogeneous virus populations. Nucleic
794 Acids Res. 42, e115. https://doi.org/10.1093/nar/gku537
795 Echeverría, N., Moratorio, G., Cristina, J., Moreno, P., 2015. Hepatitis C virus genetic variability and
796 evolution. World J. Hepatol. 7, 831–845. https://doi.org/10.4254/wjh.v7.i6.831
797 Eriksson, N., Pachter, L., Mitsuya, Y., Rhee, S.Y., Wang, C., Gharizadeh, B., Ronaghi, M., Shafer,
798 R.W., Beerenwinkel, N., 2008. Viral population estimation using pyrosequencing. PLoS
799 Comput. Biol. 4. https://doi.org/10.1371/journal.pcbi.1000074
800 Gibson, K.M., Steiner, M.C., Kassaye, S., Maldarelli, F., Grossman, Z., Pérez-Losada, M., Crandall,
801 K.A., 2019. A 28-Year History of HIV-1 Drug Resistance and Transmission in Washington,
802 DC. Front. Microbiol. 10, 1–16. https://doi.org/10.3389/fmicb.2019.00369
803 Grabher, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., Adiconis, X.F.,
804 Raychowdhury, L.R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N.,
805 Palma, F. di, W., B., Friedman, N., Regev, A., 2013. Trinity: reconstructing a full-length
806 transcriptome without a genome from RNA-Seq data. Nat. Biotechnol. 29, 644–652.
807 https://doi.org/10.1038/nbt.1883.Trinity
808 Henn, M.R., Boutwell, C.L., Charlebois, P., Lennon, N.J., Power, K.A., Macalalad, A.R., Berlin,
809 A.M., Malboeuf, C.M., Ryan, E.M., Gnerre, S., Zody, M.C., Erlich, R.L., Green, L.M., Berical,
810 A., Wang, Y., Casali, M., Streeck, H., Bloom, A.K., Dudek, T., Tully, D., Newman, R., Axten,
811 K.L., Gladden, A.D., Battis, L., Kemper, M., Zeng, Q., Shea, T.P., Gujja, S., Zedlack, C.,
812 Gasser, O., Brander, C., Hess, C., Günthard, H.F., Brumme, Z.L., Brumme, C.J., Bazner, S.,
813 Rychert, J., Tinsley, J.P., Mayer, K.H., Rosenberg, E., Pereyra, F., Levin, J.Z., Young, S.K.,
814 Jessen, H., Altfeld, M., Birren, B.W., Walker, B.D., Allen, T.M., 2012. Whole genome deep
815 sequencing of HIV-1 reveals the impact of early minor variants upon immune recognition
816 during acute infection. PLoS Pathog. 8. https://doi.org/10.1371/journal.ppat.1002529
817 Holmes, E.C., 2010. The RNA Virus Quasispecies: Fact or Fiction? J. Mol. Biol. 400, 271–273.
818 https://doi.org/10.1016/j.jmb.2010.05.032
819 Huang, A., Rami Kantor, DeLong, A., Schreier, L., Istrail, S., 2011. QColors: An algorithm for
820 conservative viral quasispecies reconstruction from short and non-contiguous next generation
821 sequencing reads. In Silico Biol. 11, 193–201. https://doi.org/10.3233/ISB-2012-0454

822 Huang, W., Li, L., Myers, J.R., Marth, G.T., 2012. ART: A next-generation sequencing read
823 simulator. Bioinformatics 28, 593–594. https://doi.org/10.1093/bioinformatics/btr708
824 Hunt, M., Gall, A., Ong, S.H., Brener, J., Ferns, B., Goulder, P., Nastouli, E., Keane, J.A., Kellam,
825 P., Otto, T.D., 2015. IVA: Accurate de novo assembly of RNA virus genomes. Bioinformatics
826 31, 2374–2376. https://doi.org/10.1093/bioinformatics/btv120
827 Jayasundara, D., Saeed, I., Maheswararajah, S., Chang, B.C., Tang, S.L., Halgamuge, S.K., 2015.
828 ViQuaS: An improved reconstruction pipeline for viral quasispecies spectra generated by next-
829 generation sequencing. Bioinformatics 31, 886–896.
830 https://doi.org/10.1093/bioinformatics/btu754
831 Kearney, M., Maldarelli, F., Shao, W., Margolick, J.B., Daar, E.S., Mellors, J.W., Rao, V., Coffin,
832 J.M., Palmer, S., 2009. Human Immunodeficiency Virus Type 1 Population Genetics and
833 Adaptation in Newly Infected Individuals. J. Virol. 83, 2715–2727.
834 https://doi.org/10.1128/jvi.01960-08
835 Kim, K., Kim, Y., 2016. Population genetic processes affecting the mode of selective sweeps and
836 effective population size in influenza virus H3N2. BMC Evol. Biol. 16, 1–15.
837 https://doi.org/10.1186/s12862-016-0727-8
838 Kingman, J.F.C., 2000. Origins of the Coalescent: 1974–1982. Genetics 156, 1461–1463.
839 Kingman, J.F.C., 1982. The coalescent. Stoch. Process. their Appl. 13, 235–248.
840 https://doi.org/10.1016/0304-4149(82)90011-4
841 Knyazev, S., Tsyvina, V., Melnyk, A., Malygina, T., Porozov, Y.B., Campbell, E., Switzer, W.M.,
842 Skums, P., Zelikovsky, A., 2018. CliqueSNV : Scalable Reconstruction of Intra-Host Viral
843 Populations from NGS Reads. bioRxiv 1–8. https://doi.org/10.1101/264242
844 Langmead, B., Salzberg, S.L., 2013. Fast gapped-read alignment with Bowtie2. Nat. Methods 9,
845 357–359. https://doi.org/10.1038/nmeth.1923
846 Leviyang, S., Griva, I., Ita, S., Johnson, W.E., 2017. A penalized regression approach to haplotype
847 reconstruction of viral populations arising in early HIV/SIV infection. Bioinformatics 33, 2455–
848 2463. https://doi.org/10.1093/bioinformatics/btx187
849 Li, Z., Chen, Y., Mu, D., Yuan, J., Shi, Y., Zhang, H., Gan, J., Li, N., Hu, X., Liu, B., Yang, B., Fan,
850 W., 2012. Comparison of the two major classes of assembly algorithms: Overlap-layout-
851 consensus and de-bruijn-graph. Brief. Funct. Genomics 11, 25–37.
852 https://doi.org/10.1093/bfgp/elr035
853 Lozupone, C.A., Knight, R., 2005. UniFrac : a New Phylogenetic Method for Comparing Microbial
854 Communities UniFrac : a New Phylogenetic Method for Comparing Microbial Communities
855 [see notes, compare to Bray-Curtis]. Appl. Environ. Microbiol. 71, 8228–8235.
856 https://doi.org/10.1128/AEM.71.12.8228
857 Macalalad, A.R., Zody, M.C., Charlebois, P., Lennon, N.J., Newman, R.M., Malboeuf, C.M., Ryan,
858 E.M., Boutwell, C.L., Power, K.A., Brackney, D.E., Pesko, K.N., Levin, J.Z., Ebel, G.D., Allen,
859 T.M., Birren, B.W., Henn, M.R., 2012. Highly sensitive and specific detection of rare variants
860 in mixed viral populations from massively parallel sequence data. PLoS Comput. Biol. 8,
861 e1002417. https://doi.org/10.1371/journal.pcbi.1002417
862 Maldarelli, F., Kearney, M., Palmer, S., Stephens, R., Mican, J., Polis, M.A., Davey, R.T., Kovacs,
863 J., Shao, W., Rock-Kress, D., Metcalf, J.A., Rehm, C., Greer, S.E., Lucey, D.L., Danley, K.,
864 Alter, H., Mellors, J.W., Coffin, J.M., 2013. HIV populations are large and accumulate high
865 genetic diversity in a nonlinear fashion. J. Virol. 87, 10313–23.
866 https://doi.org/10.1128/JVI.01225-12
867 Malhotra, R., Wu, M.M.S., Rodrigo, A., Poss, M., Acharya, R., 2015. Maximum Likelihood de novo
868 reconstruction of viral populations using paired end sequencing data. arXiv 1–14.
869 Mancuso, N., Tork, B., Skums, P., Ganova-Raeva, L., M˘andoiu, I., Zelikovsky, A., 2011.
870 Reconstructing viral quasispecies from NGS amplicon reads. In Silico Biol. 11, 237–249.
871 Mangul, S., Wu, N.C., Mancuso, N., Zelikovsky, A., Sun, R., Eskin, E., 2014. Accurate viral

872 population assembly from ultra-deep sequencing data. Bioinformatics 30, 329–337.
873 https://doi.org/10.1093/bioinformatics/btu295
874 Mansky, L.M., 2000. In Vivo Analysis of Human T-Cell Leukemia Virus Type 1 Reverse
875 Transcription Accuracy. J. Virol. 74, 9525–9531. https://doi.org/10.1128/jvi.74.20.9525-
876 9531.2000
877 McClelland, J., Koslicki, D., 2018. Emdunifrac: Exact linear time computa-tion of the unifrac metric
878 and identification of differentially abundant organisms. J. Math. Biol. 77, 935–949.
879 McCrone, J., 2018. Influenza Virus Evolution Within and Between Human Hosts.
880 McCrone, J.T., Woods, R.J., Martin, E.T., Malosh, R.E., Monto, A.S., Lauring, A.S., 2018.
881 Stochastic processes constrain the within and between host evolution of influenza virus. Elife 7.
882 https://doi.org/10.7554/eLife.35962
883 Neher, R.A., Leitner, T., 2010. Recombination rate and selection strength in HIV intrapatient
884 evolution. PLoS Comput. Biol. 6. https://doi.org/10.1371/journal.pcbi.1000660
885 Nobre, A.F.S., De Souza Almeida, D., Ferreira, L.C., Ferreira, D.L., Júnior, E.C.S., De Almeida
886 Viana, M.D.N. do S., Silva, I.C., Pinheiro, B.T., Ferrari, S.F., Da Costa Linhares, A., Ishikawa,
887 E.A., Sousa, R.C.M., De Sousa, M.S., 2018. Low genetic diversity of the human T-cell
888 lymphotropic virus (HTLV-1) in an endemic area of the brazilian Amazon basin. PLoS One 13,
889 1–9. https://doi.org/10.1371/journal.pone.0194184
890 Nurk, S., Meleshko, D., Korobeynikov, A., Pevzner, P.A., 2017. metaSPAdes : a new versatile
891 metagenomic assembler 824–834. https://doi.org/10.1101/gr.213959.116.4
892 Pandit, A., de Boer, R.J., 2014. Reliable reconstruction of HIV-1 whole genome haplotypes reveals
893 clonal interference and genetic hitchhiking among immune escape variants. Retrovirology 11,
894 56. https://doi.org/10.1186/1742-4690-11-56
895 Pérez-Losada, M., Castel, A.D., Lewis, B., Kharfen, M., Cartwright, C.P., Huang, B., Maxwell, T.,
896 Greenberg, A.E., Crandall, K.A., on behalf of the DC Cohort Executive Committee, 2017.
897 Characterization of HIV diversity, phylodynamics and drug resistance in Washington, DC.
898 PLoS One 12, e0185644. https://doi.org/10.1371/journal.pone.0185644
899 Pérez-Losada, M., Jobes, D. V., Sinangil, F., Crandall, K.A., Posada, D., Berman, P.W., 2010.
900 Phylodynamics of HIV-1 from a phase-III AIDS vaccine trial in North America. Mol. Biol.
901 Evol. 27, 417–425. https://doi.org/10.1093/molbev/msp254
902 Posada-cespedes, S., Seifert, D., Beerenwinkel, N., 2016. Recent advances in inferring viral diversity
903 from high-throughput sequencing data. Virus Res.
904 https://doi.org/10.1016/j.virusres.2016.09.016
905 Posada, D., Crandall, K.A., 2001. Selecting models of nucleotide substitution: An application to
906 human immunodeficiency virus 1 (HIV-1). Mol. Biol. Evol. 18, 897–906.
907 https://doi.org/10.1093/oxfordjournals.molbev.a003890
908 Posada, David, Crandall, K.A., Holmes, E.C., 2002. Recombination in Evolutionary Genomics.
909 Annu. Rev. Genet. 36, 75–97. https://doi.org/10.1146/annurev.genet.36.040202.111115
910 Posada, D, Crandall, K.A., Holmes, E.C., 2002. Recombination in evolutionary genomics. Annu Rev
911 Genet 36, 75–97. https://doi.org/10.1146/annurev.genet.36.040202.111115
912 Prabhakara, S., Malhotra, R., Acharya, R., Poss, M., 2013. Mutant-Bin: Unsupervised haplotype
913 estimation of viral population diversity without reference genome. ournal Comput. Biol. 20,
914 453–463.
915 Prabhakaran, S., Rey, M., Zagordi, O., Beerenwinkel, N., 2010. HIV-haplotype inference using a
916 constraint-based dirichlet process mixture model. Mach. Learn. Comput. Biol. NIPS Work. 1–4.
917 Prabhakaran, S., Rey, M., Zagordi, O., Beerenwinkel, N., Roth, V., 2014. HIV haplotype inference
918 using a propagating dirichlet process mixture model. IEEE/ACM Trans. Comput. Biol.
919 Bioinforma. 11, 182–191.
920 Prosperi, M.C.F., Salemi, M., 2012. QuRe: Software for viral quasispecies reconstruction from next-
921 generation sequencing data. Bioinformatics 28, 132–133.

https://doi.org/10.1093/bioinformatics/btr627
Prosperi, M.C.F., Yin, L., Nolan, D.J., Lowe, A.D., Goodenow, M.M., Salemi, M., 2013. Empirical validation of viral quasispecies assembly algorithms: state-of-the-art and challenges. Sci. Rep. 3, 2837. https://doi.org/10.1038/srep02837
Ratner, L., Haseltine, W., Patarca, R., Livak, K.J., Starcich, B., Josephs, S.F., Doran, E.R., Rafalski, J.A., Whitehorn, E.A., Baumeister, K., 1985. Complete Nucleotide Sequence of the AIDS Virus, HTLV-III. Nature 313, 277–84. https://doi.org/10.1038/313277a0
Ribeiro, R.M., Li, H., Wang, S., Stoddard, M.B., Learn, G.H., Korber, B.T., Bhattacharya, T., Guedj, J., Parrish, E.H., Hahn, B.H., Shaw, G.M., Perelson, A.S., 2012. Quantifying the Diversification of Hepatitis C Virus (HCV) during Primary Infection: Estimates of the In Vivo Mutation Rate. PLoS Pathog. 8. https://doi.org/10.1371/journal.ppat.1002881
Rodrigo, A.G., Felsenstein, J., 1999. Coalescent Approaches to HIV Population Genetics, in: The Evolution of HIV. The Johns Hopkins University Press, Baltimore, pp. 233–274.
Rosenberg, N.A., Nordborg, M., 2002. Genealogical trees, coalescent theory and the analysis of genetic polymorphisms. Nat. Rev. Genet. 3, 380–390. https://doi.org/10.1038/nrg795
Sanjuán, R., Nebot, M.R., Chirico, N., Louis, M., Belshaw, R., Sanjua, R., Mansky, L.M., 2010. Viral Mutation Rates Viral Mutation Rates. J. Virol. 84, 9733–9748. https://doi.org/10.1128/JVI.00694-10
Schirmer, M., Sloan, W.T., Quince, C., 2014. Benchmarking of viral haplotype reconstruction programmes: An overview of the capacities and limitations of currently available programmes. Brief. Bioinform. 15, 431–442. https://doi.org/10.1093/bib/bbs081
Scholz, M., Ward, D. V., Pasolli, E., Tolio, T., Zolfo, M., Asnicar, F., Truong, D.T., Tett, A., Morrow, A.L., Segata, N., 2016. Strain-level microbial epidemiology and population genomics from shotgun metagenomics. Nat. Methods 13, 435–438. https://doi.org/10.1038/nmeth.3802
Skums, P., Mancuso, N., Artyomenko, A., Tork, B., Mandoiu, I., Khudyakov, Y., Zelikovsky, A., 2013. Reconstruction of viral population structure from next-generation sequencing data using multicommodity flows. BMC Bioinformatics 14, S2. https://doi.org/10.1186/1471-2105-14-S9-S2
Töpfer, A., Marschall, T., Bull, R.A., Luciani, F., Schö Nhuth, A., Beerenwinkel, N., Mchardy, A.C., 2014. Viral Quasispecies Assembly via Maximal Clique Enumeration. PLoS Comput Biol 10. https://doi.org/10.1371/journal.pcbi.1003515
Topfer, A., Zagordi, O., Prabhakaran, S., Roth, V., Halperin, E., Beerenwinkel, N., 2013. Probabilistic Inference of Viral Quasispecies Subject to Recombination. J. Comput. Biol. 20, 113–123. https://doi.org/10.1089/cmb.2012.0232
van der Kuyl, A.C., Cornelissen, M., 2007. Identifying HIV-1 dual infections. Retrovirology 4, 1–12. https://doi.org/10.1186/1742-4690-4-67
Warren, R.L., Sutton, G.G., Jones, S.J.M., Holt, R.A., 2007. Assembling millions of short DNA sequences using SSAKE. Bioinformatics 23, 500–501. https://doi.org/10.1093/bioinformatics/btl629
Yang, X., Charlebois, P., Gnerre, S., Coole, M.G., Lennon, N.J., Levin, J.Z., Qu, J., Ryan, E.M., Zody, M.C., Henn, M.R., 2012. De novo assembly of highly diverse viral populations. BMC Genomics 13. https://doi.org/10.1186/1471-2164-13-475
Yang, X., Charlebois, P., Macalalad, A., Henn, M.R., Zody, M.C., 2013. V-Phaser 2: variant inference for viral populations. BMC Genomics 14, 674. https://doi.org/10.1186/1471-2164-14-674
Zagordi, O., Bhattacharya, A., Eriksson, N., Beerenwinkel, N., 2011. ShoRAH: estimating the genetic diversity of a mixed sample from next-generation sequencing data. BMC Bioinformatics 12, 119. https://doi.org/10.1186/1471-2105-12-119
Zagordi, O., Geyrhofer, L., Roth, V., Beerenwinkel, N., 2010. Deep Sequencing of a Genetically Heterogeneous Sample: Local Haplotype Reconstruction and Read Error Correction. J. Comput.

Biol. 17, 417–428. https://doi.org/10.1089/cmb.2009.0164
Zanini, F., Brodin, J., Thebo, L., Lanz, C., Bratt, G., Albert, J., Neher, R.A., 2015. Population genomics of intrapatient HIV-1 evolution. Elife 4, 1–26. https://doi.org/10.7554/eLife.11282

Appendix

Figure S1. *Plots showing percentage of datasets analyzed by a tool within our time limit. From green to red, full completion of a dataset (green) to no completion (red) is indicated for each set.*

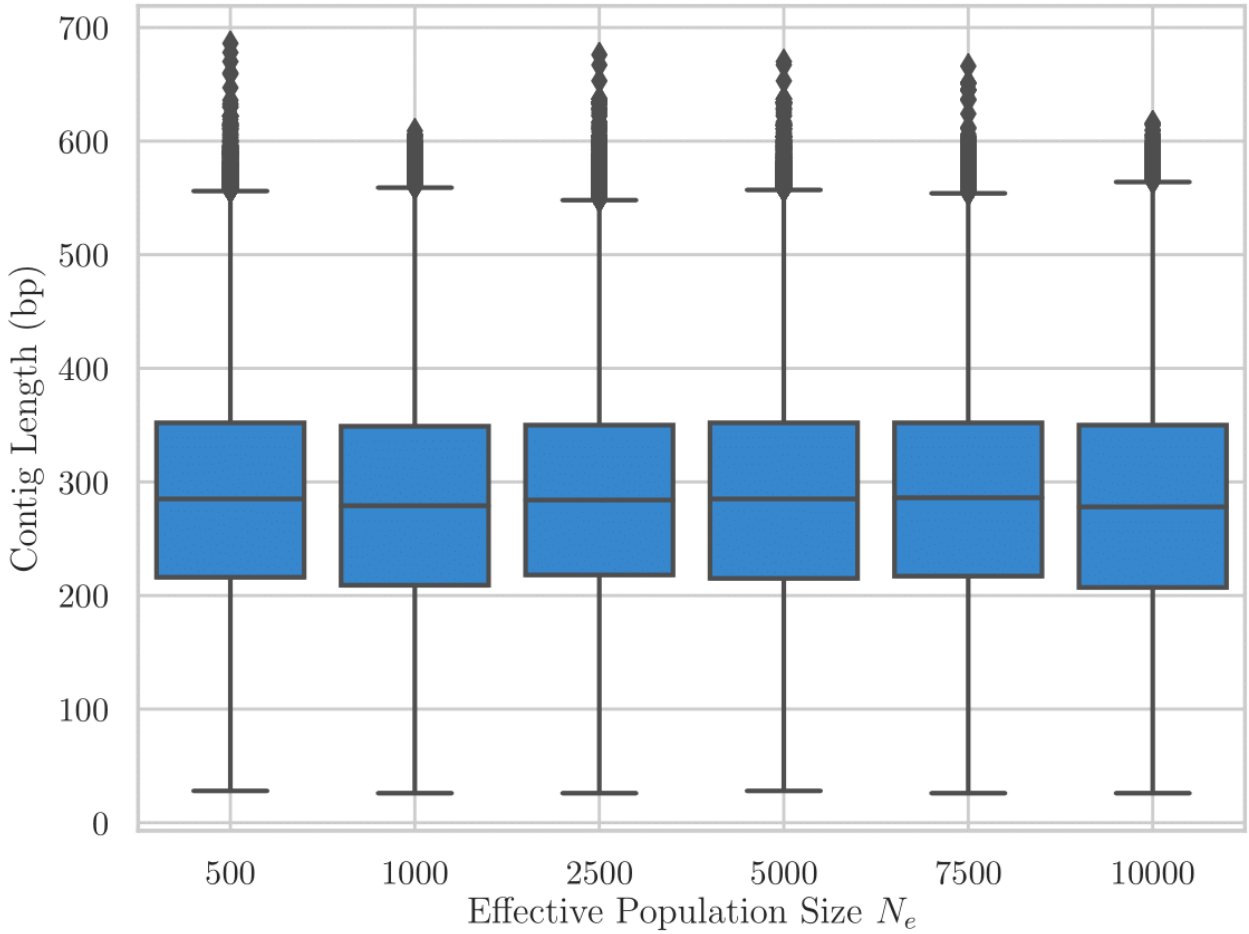

Figure S2. *The length distribution of haplotypes obtained by HaploClique for each effective population size.*

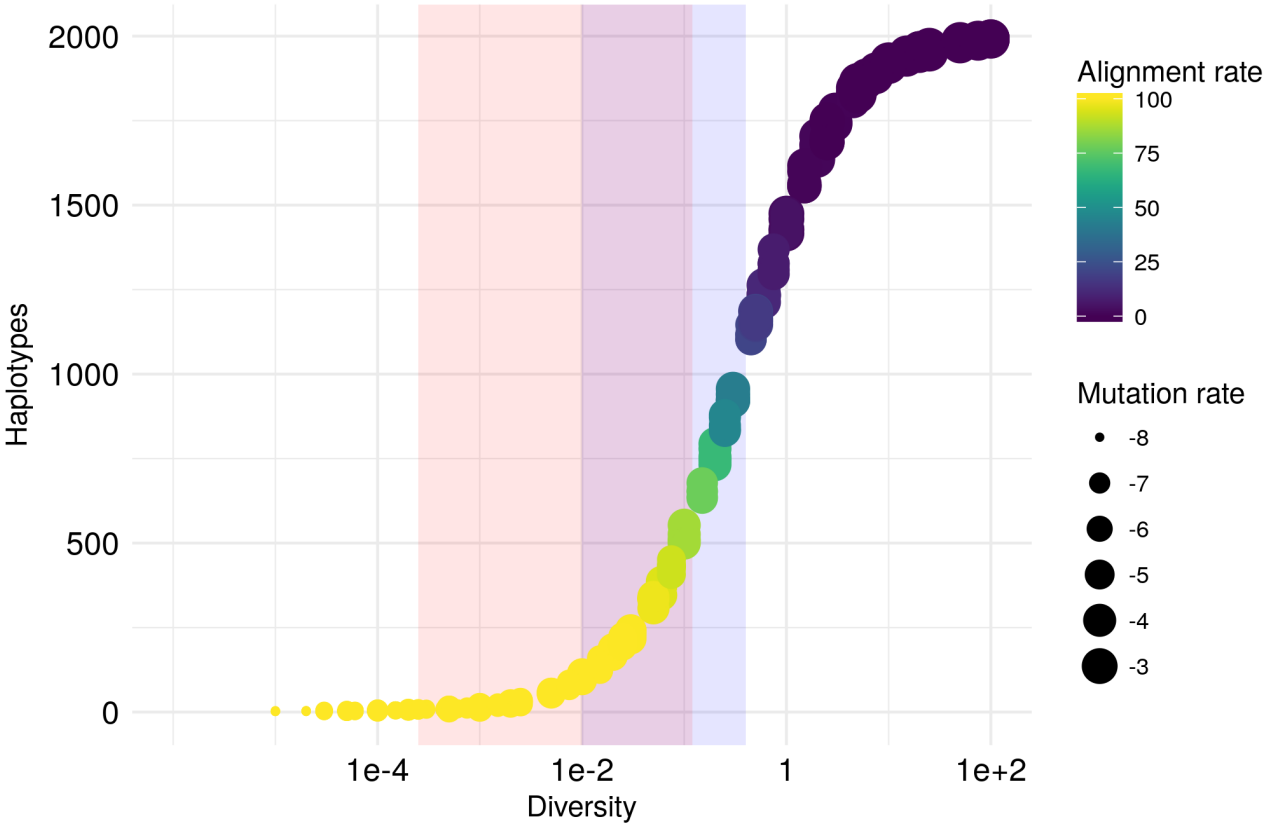

Figure S3. *Number of true haplotypes estimated from the coalescent-based simulated data across levels of underlying intra-patient genetic diversity. Intra-host HIV-1 and HCV diversity levels are highlighted shaded light blue and shaded light red regions, respectively.*

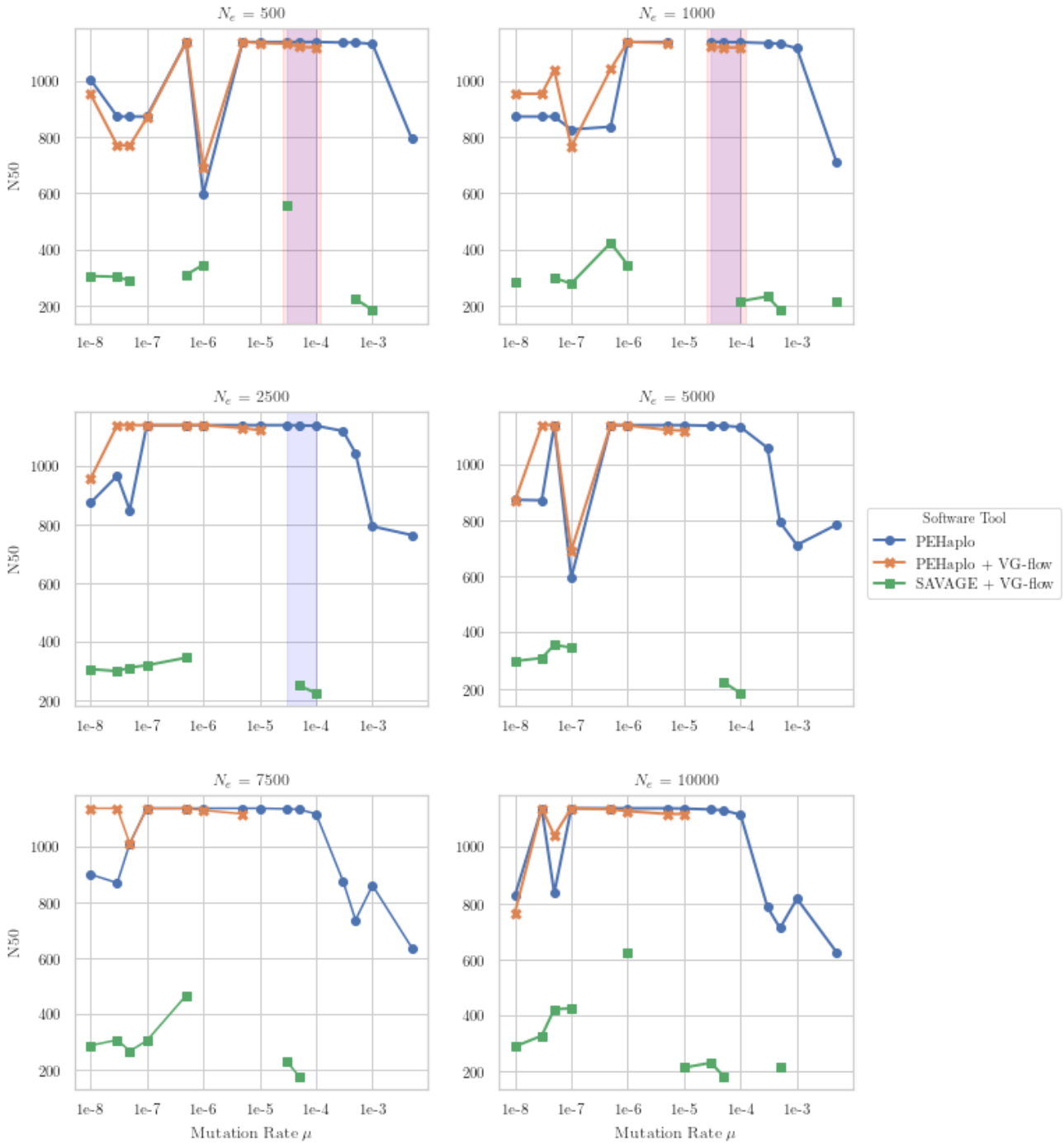

Figure S4. Reference-based and *de novo* haplotype callers: variation of recall values with mutation rate (log-scaled) for all considered N_e. The shaded light blue and red regions correspond to HIV-1 and HCV diversity levels, respectively. For all pairs of parameters μ and N_e, we report the mean estimates of recall over all valid outputs produced by each software tool for five haplotype populations. If a tool did not produce any output for some pair of parameters, we included a gap in the corresponding plot.

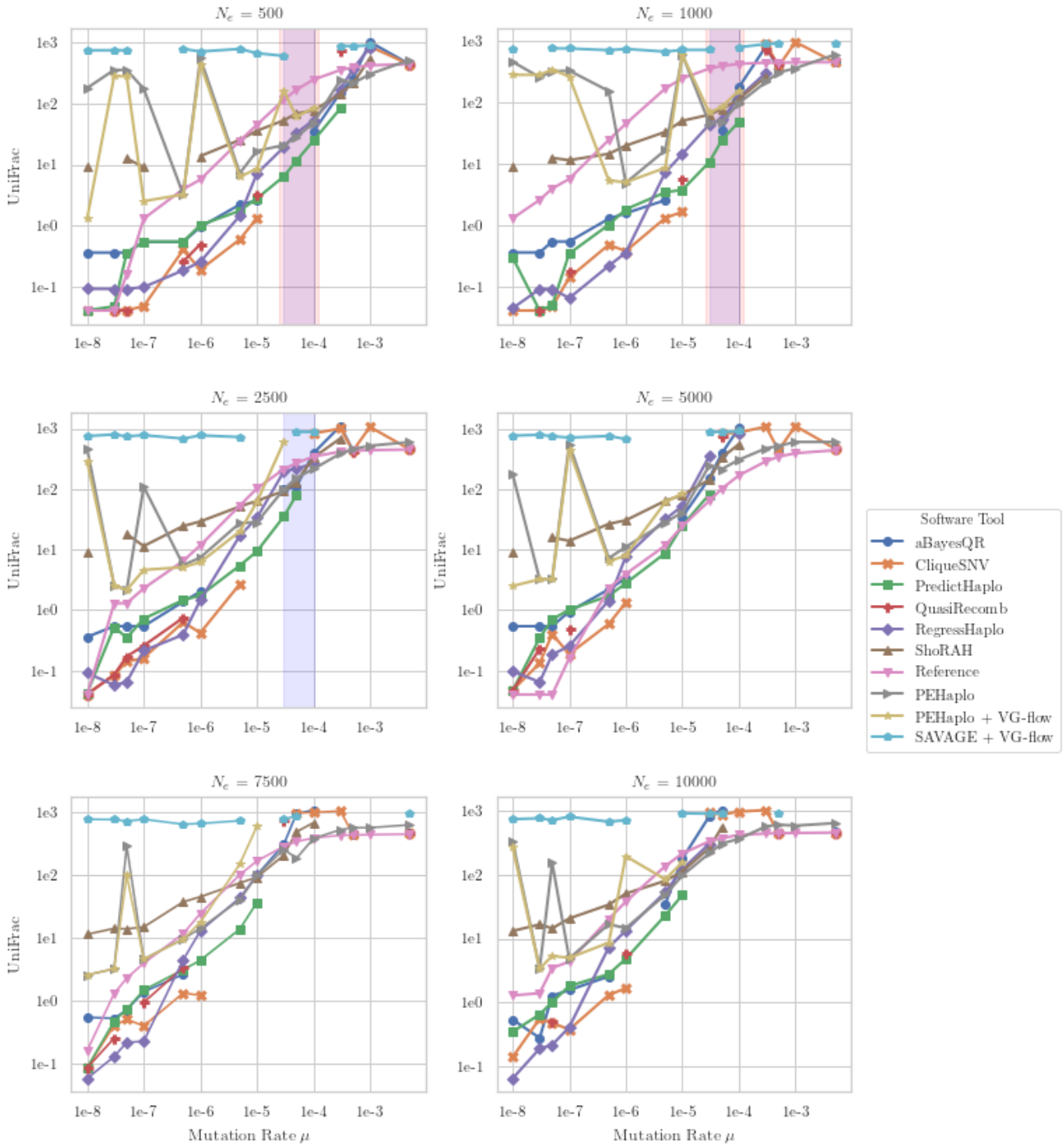

Figure S5. Reference-based and *de novo* haplotype callers: variation of UniFrac distances (EMD) with mutation rate (log-scaled) for all considered N_e. The shaded light blue and red regions correspond to HIV-1 and HCV diversity levels, respectively. For all pairs of parameters μ and N_e, we report the mean estimates of UniFrac distances over all valid outputs produced by each software tool for five haplotype populations. If a tool did not produce any output for some pair of parameters, we included a gap in the corresponding plot. The values for *de novo* tools are usually higher than for *reference-based* tools.