Viral quasispecies assembly via maximal clique enumeration

Armin Töpfer[1,2], Tobias Marschall[3], Rowena A. Bull[4], Fabio Luciani[4], Alexander Schönhuth[3,†], Niko Beerenwinkel[1,2,†,∗]
1 Department of Biosystems Science and Engineering, ETH Zurich, Basel, Switzerland
2 SIB Swiss Institute of Bioinformatics, Basel, Switzerland
3 Centrum Wiskunde & Informatica, Amsterdam, The Netherlands
4 Inflammation and Infection Research Centre, School of Medical Sciences, UNSW, Sydney, Australia
† equal contributions
∗ E-mail: niko.beerenwinkel@bsse.ethz.ch

Abstract

Virus populations can display high genetic diversity within individual hosts. The intra-host collection of viral haplotypes, called viral quasispecies, is an important determinant of virulence, pathogenesis, and treatment outcome. We present HaploClique, a computational approach to reconstruct the structure of a viral quasispecies from next-generation sequencing data as obtained from bulk sequencing of mixed virus samples. We develop a statistical model for paired-end reads accounting for mutations, insertions, and deletions. Using an iterative maximal clique enumeration approach, read pairs are assembled into haplotypes of increasing length, eventually enabling global haplotype assembly. The performance of our quasispecies assembly method is assessed on simulated data for varying population characteristics and sequencing technology parameters. Owing to its paired-end handling, HaploClique compares favorably to state-of-the-art haplotype inference methods. It can reconstruct error-free full-length haplotypes from low coverage samples and detect large insertions and deletions at low frequencies. We applied HaploClique to sequencing data derived from a clinical hepatitis C virus population of an infected patient and discovered a novel deletion of length 357 ± 167 bp that was validated by two independent long-read sequencing experiments. HaploClique is available at `https://github.com/armintoepfer/haploclique`.

Author Summary

Humans infected with a virus, such as the human immunodeficiency virus (HIV-1) or hepatitis C virus (HCV), host a population of billions of virus particles. Among these, there is an unknown number of genetically different strains, some of which can harbor drug resistance and immune escape mutations. It is of clinical importance to know the DNA sequences and abundances of these variants, as they can affect treatment outcome. Here, we present HaploClique, a computational approach to reconstruct these sequences and to predict large insertions and deletions from paired-end next-generation sequencing data. Using simulations, we demonstrate that HaploClique can reconstruct full-length HIV-1 variants from low-coverage samples. Using real-world clinical data, we predict a novel deletion of 357 ± 167 bp in a HCV patient sample that has been validated by two independent long-read sequencing experiments.

Introduction

Genetic diversity is an important characteristic of evolving populations and it affects the chances of survival in changing environments. Assessing the genetic diversity of a population experimentally is generally labor-intensive and difficult. Populations of individual cells or viruses, however, can be analyzed efficiently using next-generation sequencing (NGS). Although single-cell approaches are still immature, direct NGS of mixed samples at deep coverage allows for probing populations in great detail. The challenges with this bulk sequencing approach are (i) to separate sequencing errors from genetic variation,

(ii) to assemble the short NGS reads into an unknown number of different, unknown, longer haplotype sequences, and (iii) to estimate their frequency distribution.

Viruses such as human immunodeficiency virus (HIV-1) and hepatitis C virus (HCV) populate their hosts as swarms of related but genetically different mutant strains, each defined by its haplotype sequence. The structure of such a mutant cloud, which is often referred to as a viral quasispecies [1], is of clinical importance, because it has been shown to affect virulence [2] and pathogenesis [3]. In addition, low-frequency genetic variants may harbor resistance mutations that are capable of evolutionary escape from the selective pressure of host immune responses [4] and of medical interventions, such as anti-viral drug treatment [5]. NGS is currently introduced into clinical diagnostics, but the *de facto* standard procedure for assessing the quasispecies structure is simply based on single-nucleotide variant (SNV) calling. This approach allows only for estimating the per-site allele frequency spectrum of the virus population and it ignores patterns of co-occurrence among mutations. This limitation is critical, because epistatic interactions are abundant in RNA viruses [6]. Hence, one cannot predict viral phenotypes without knowing the underlying mix of haplotypes. Here, we address this challenge and present a computational approach for the viral quasispecies assembly problem.

The viral haplotype reconstruction problem is related to the human haplotype reconstruction problem, but it differs in several key aspects and faces different challenges. First, the number of unique haplotypes in a viral quasispecies is unknown unlike in the case of human diploid genomes. Second, viral populations typically exhibit more than two variants at each polymorphic locus and often all four different nucleotides. Hence, viral haplotypes cannot be described by binary sequences. Third, in a viral quasispecies, low-frequency variants are abundant and of clinical importance, yet they are difficult to distinguish from technical sequencing errors. Finally, RNA virus genomes are orders of magnitude shorter than the human genome, but exhibit more diversity within one host than the $\sim 0.1\%$ diversity between the two parental human haplotypes [7].

Several methods for viral haplotype reconstruction have been developed in recent years, specialized for different NGS technologies, experimental designs, and quasispecies structures. In general, reconstruction can be performed either locally, in a genomic region that can be covered by the average read length, or globally, over longer regions such that overlapping reads are necessary for assembly. Local reconstruction means estimating the number of locally unique haplotype sequences and, at the same time, correcting sequencing errors. Probabilistic clustering [8–11] and k-mer statistics [12] have been proposed for this task. Global reconstruction is more challenging, as it requires computational solutions for assembling NGS reads, which has proven itself to be demanding even in settings without poly-ploidy [13].

For quasispecies assembly, approaches from different domains have been developed: (i) probabilistic mixture models [14], (ii) hidden Markov models [15], (iii) sampling schemes [16], (iv) combinatorial approaches based on analyzing the read overlap graph [8, 17–19], (v) coloring of overlap and conflict graphs by constraint programming [20], and (vi) exploiting the "identical by descent" information [21] in the HapCompass framework [22], originally designed for diploid single nucleotide polymorphism data.

The performance of global haplotype reconstruction depends on several factors, including the true underlying diversity of the population, the distribution of amplification and sequencing errors, the read length, and the distribution of the read coverage along the genome [23–25]. A major shortcoming of all existing methods is that they are unable to handle large insertions or deletions (indels). For example, large deletions can result from erroneous replication or, as observed recently in HIV-1, they may occur as alternative splice variants [26]. In the context of analyzing structural variation in the human genome, such as indels of varying sizes, the use of paired-end reads has been instrumental. For viral haplotype reconstruction, however, approaches that systematically exploit paired-end information are lacking.

In this paper, we present a new quasipecies assembly method for paired-end reads, called HaploClique, based on enumeration of maximal cliques (max-cliques) as a general approach to clustering NGS paired-end reads. Although, in general, the runtime of enumerating all max-cliques in a graph is exponential, it has recently been shown that the graphs induced by overlapping NGS reads can be handled efficiently [27, 28]. Here, we exploit this fact for the quasispecies assembly problem and develop a probabilistic model of sequence and structural similarity between reads.

Using max-clique enumeration for reference-based read assembly is orthogonal to combinatorial approaches for *de novo* assembly that rely on path finding in de Bruijn or similar graphs [29–31]. Instead of computing paths, we iteratively transform max-cliques into super-reads and then seek max-cliques of super-reads, thereby obtaining haplotype segments of increasing length. The haplotype segments can eventually be extended to global haplotypes if the degree of heterogeneity of the viral quasispecies is high enough. HaploClique is related to max-cut-driven approaches in human haplotype reconstruction [32], but the computational complexity of those approaches is prohibitive for virus populations of high and unknown ploidy. While HaploCliques enumerates all max-cliques, a max-cut approach seeks an optimal cut of the overlap graph.

HaploClique explicitly incorporates paired-end information for assembling viral haplotypes. We define the insert as the unsequenced fragment between the two ends of a paired-end read. We use linkage information among variant alleles in the distant pairs to identify reads that stem from the same haplotypes and generate error-corrected paired-end super-reads. Paired-end reads allow to bridge homogeneous, and hence ambiguous, genomic regions if the insert size is sufficiently large. They also increase the statistical power to distinguish local haplotypes from sequencing errors in homogeneous regions if the paired read is located in a more heterogeneous region. Employing our iterative clique enumeration procedure, we show that error-free full-length HIV-1 viral haplotypes can be reconstructed in a heterogeneous mix of five viral strains *in silico* from a data set with mean coverage of 600x. Furthermore, we demonstrate that, unlike existing methods, HaploClique can detect large indels in mixed virus populations *in silico* and *in vivo*. Finally, we apply HaploClique to a HCV Illumina paired-end NGS data set and predict a novel deletion of length 357 ± 167 bp that has been confirmed independently by two long-read NGS platforms.

Results

We developed and implemented HaploClique, a computational viral quasispecies assembly method for paired-end NGS data. HaploClique defines a read alignment graph, in which each node corresponds to a single-end or paired-end alignment (Figure 1A). We draw edges between two nodes if the two corresponding alignments have sufficient overlap and are likely to stem from the same haplotype (Figure 1B). Each max-clique in this graph consists of a large number of reads from the presumed same haplotype segment. Thus, the consensus sequence of all reads in a max-clique is a prediction of a local haplotype sequence. We refer to such a consensus sequence as a super-read. This consensus sequence also serves to correct errors in the reads that participate in the super-read by replacing the sequence of the original reads with the consensus. This form of error correction benefits from phasing sequential variants through super-read construction. Paired-end reads are particularly helpful, as they allow to also phase distantly co-occurring variant alleles.

HaploClique proceeds by iterating (i) (super-)read alignment graph construction, (ii) max-clique enumeration, and (iii) super-read construction, until convergence. The lengths of the super-reads increase while iterating and convergence is established when super-reads have reached their maximum length. If the mixed sample is sufficiently heterogeneous, super-reads will eventually represent haplotypes of full length. Because we process paired-end reads and incorporate insert-size compatibility into our edge definition, we can also identify max-cliques that indicate larger insertions and deletions. These structural variations are recognized by too small or too large insert sizes among the alignments of the reads that participate in a max-clique. We analyzed HaploClique's performance on simulated data and demonstrate its use on *in vivo* HCV quasispecies sequencing data.

Simulation studies

HaploClique integrates paired-end and base quality information for improved sequencing error correction and haplotype frequency estimation, which we assess first. Second, we evaluate HaploClique's behaviour when confronted with low heterogeneity among the different haplotype strains. Third, we demonstrate HaploClique's ability to detect large insertions and deletions in the quasispecies by making use of paired-end information. Fourth, we evaluate the quality of the local and global haplotypes that HaploClique predicts. Lastly, we compare HaploClique to state-of-the-art tools ShoRAH [33], PredictHaplo [14], and QuRe [16] in quasispecies reconstruction of a simulated five virus mix of well-known HIV-1 lab-strains.

In all of the following experiments, we simulated Illumina 2x250 bp paired-end reads using SimSeq [34] with fragment size 600 bp. To make the simulated data as realistic as possible, we estimated the required error profiles from an in-house MiSeq data set of a mixture of known HIV-1 strains. The average error-rate was 0.33% per base.

Error correction and frequency estimation. We assessed HaploClique's performance in error correction and frequency estimation, and their dependency on coverage and relative haplotype abundance. We generated ten HIV-1 strains by substituting ten percent of the nucleotides of strain $\mathrm{HIV}_{\mathrm{HXB2}}$. Positions to be substituted were sampled uniformly, separately for each of the ten strains. We sampled reads from these ten strains at coverage rates that resulted in abundance levels of 0.6%, 1.2%, 2.4%, 3.6%, 4.8%, 6.0%, 9%, 12.1%, 24.1%, and 36.2%.

We measured the accuracy of estimating relative haplotype frequencies for different true abundances and coverages (Figure 2A). For each coverage, we repeated the simulation ten times and depict the mean deviances. For example, a mean deviance of +1% for a strain of frequency 4.8% translates to an estimate of 5.8%. The estimated frequencies approach the true ones for increasing coverage, as indicated by approaching the dashed line of 0% deviance. HaploClique tended to underestimate frequencies above and slightly overestimated frequencies below a true frequency of ten percent. For 1600x coverage, the absolute deviation was always below one percent. We additionally measures the frequency estimation robustness by computing the standard deviation for each true frequency and coverage (Figure S1). We observed increased robustness with increasing coverage, except for the cases where the super-reads do not fully cover the genome.

We assessed HaploClique's paired-end error correction capabilities using the same simulated data set. All super-reads, regardless of frequency, perfectly agreed with the respective true haplotype sequences. For frequency-coverage combinations of (3.6%, 400x), (2.4%, 800x), (1.2%, 1600x) and higher, the genomes were fully covered by super-reads (Figure 2A). For lower frequencies, reconstructed local haplotypes do not fully cover the genome in all ten data sets.

Minimal variant heterogeneity. To investigate the minimal heterogeneity necessary to distinguish between different haplotypes, we simulated datasets of two different haplotypes, varying coverage rates between 25 and 800x and pairwise distance from 0.05% to 10%.

We measured the false positive rates, percentages of reconstructed haplotypes with at least one false nucleotide, and percentages of perfectly reconstructed haplotypes that map uniquely to exactly one true variant (Figure 2B). For distances of 1% and above, reconstruction was perfect at all coverage rates. For distances of 0.75% and below, the vast majority of the reconstructed haplotypes still perfectly match a true haplotype. In general, the false positive rate decreases with higher coverage. However, with decreased distance between true haplotypes, many regions are conserved and cannot be assigned uniquely to a single true haplotype.

Large deletion prediction. The edge definition of HaploClique's read alignment graph allows for identifying large indels in the haplotypes (Methods). Thanks to the insert size criterion, one can predict indels from paired-end read information despite the lack of alignments of read ends that directly cover the deletion breakpoint. Such alignments may be lacking in practice, because even the most advanced read mappers have difficulties aligning reads across long indels.

In this simulation, we benchmarked the false negative rate of the predicted deletions and the deviations of the estimated to the true deletion lengths (Figure 3). Aiming at a simple, yet instructive benchmark data set, we created a new haplotype by randomly placing three deletions of sizes 100, 500, and 1000 bp into the HIV_{HXB2} genome. We simulated reads from the reference haplotype with a mean coverage of 100x and from the deletion-harboring haplotype with coverages of 5, 12, 24, 48, 96, and 144x. We sampled 100 data sets for each coverage to account for variability in the sampling process. Comparing coverages of 5 and 144x, we observed a more reliable size estimate with increasing coverage, as the standard deviation decreased by a factor of up to two. Independent of the true deletion size, the estimated median size approached the true size up to 6 bp. For the true deletion sizes of 100, 500, and 1000 bp the number of false negatives decreased to zero for a coverage of 24, 12, and 12x, respectively. The median size deviation to the true deletion size was always below the standard deviation of the insert fragment size if there was sufficient coverage, i.e., no false negatives.

Global haplotype assembly. We benchmarked HaploClique's performance in terms of genome coverage, as well as numbers and lengths of constructed super-reads. We simulated a heterogeneous population, henceforth referred to as *lab-mix*, of five lab strains, namely HIV_{HXB2}, $\text{HIV}_{\text{NL4-3}}$, HIV_{YU2}, $\text{HIV}_{89.6}$, and $\text{HIV}_{\text{JR-CSF}}$. Among these strains, pairwise distances vary per gene and along the entire genome between 1 and 16% and between 2 and 6%, respectively. We sampled reads uniformly with a mean coverage of 600x, i.e., 120x per strain. This low coverage corresponds to very low-frequency variants in datasets of higher coverage (commonly 10,000 to 100,000x), which are of interest in diagnostic applications.

HaploClique was able to reconstruct haplotype HIV_{HXB2} at its full length, without a single error (Table 1). The haplotypes of the other four variants grew up to a maximal size of 5-6 kb for the longest super-reads, where full length of the five true strains varied between 9 and 10 kb. For all strains, the reconstructed haplotypes covered the genome at its full length. The false positive rate (the rate of not perfectly matching super-reads) was 0.3%. The maximal read length increased monotonically for each iteration from 250 bp up to full-length of $\sim$ 10 kb (Figure 4). During the first two iterations, the number of super-reads increased, but from the third iteration decreased, converging to a number of 56 super-reads (Figure 4).

Comparisons. We performed haplotype reconstruction for the lab-mix using the tools ShoRAH [33], PredictHaplo [14], and QuRe [16], and compared the results to those of HaploClique (Table 1). Hapler v1.60 [19] did not accept NGS alignments with insertions, and hence was not applicable to this heterogeneous virus populations. PredictHaplo v0.4 with paired-end option reconstructed local or global haplotypes with an average error-rate of 1.65%. ShoRAH does not take paired-end information into account and reconstructed 196 full-length genomes, 16 over 1% estimated frequency, with an average error rate of 1.99%. QuRe v0.99971 performed its own single-end alignment and reconstructed global haplotypes with an average error rate of 2.95%, the highest among all methods. The wall-clock time and maximal memory consumption, respectively, was 10 min and 12 mb for PredictHaplo, 30 min and 150 mb for HaploClique, 1.5 hr and 680 mb for ShoRAH, and 2 hr and 33 gb for QuRe on a 80-core server with 1 tb memory. PredictHaplo and HaploClique used a single thread, QuRe and ShoRAH are partially multithreaded. The results of QuRe were not reproducible, because new instances of QuRe on the same data set did not finish within one day. HaploClique performed best in maximal reconstructed haplotype length that was error-free, and it had the closest estimated haplotype distribution and overall the highest precision.

Patient sample

For the application of HaploClique to a clinical sample, HCV RNA was extracted from the plasma collected from a subject isolated 135 days post infection and the NS5 region RT-PCR amplified as previously described [35]. In this subject there was experimental evidence of antigen-specific CD8+ T cell responses targeting two epitopes in the NS5 region (K2629SKRTPMGF and W2820LGNIIMFA). The NS5B region encodes for the RNA-dependent RNA polymerase and is essential for the replication of the virus.

This amplicon was sequenced to a coverage of 80,000x on a MiSeq instrument using a 2x250 bp read kit. The resulting reads were aligned using BWA-MEM [36]. We found the insert size distribution to have a mean of 155 bp and a standard deviation of 167 as estimated by HaploClique. Despite this large standard deviation, HaploClique was able to discover a 357 ± 167 bp deletion. No other indels were reported by HaploClique.

In two independent sequencing runs of the same amplicon, once on a 454/Roche GS FLX+ system and once on a PacBio instrument, the presence of the deletion was confirmed. Both technologies yield longer reads than MiSeq that could successfully be aligned across the deletion breakpoint, allowing to determine breakpoint coordinates at base-pair resolution.

For the alignment of the longer reads, extreme affine gap costs have been used to find the deletion. In general, this leads to alignment artifacts in other regions, causing false positive haplotype calls. With a read length of 250 bp, we did not succeed to align reads across the large deletion.

Comparing coordinates, we found that the start positions predicted by HaploClique was 15 bp off the true position and the true length amounted to 444 bp. That is, the length difference between true and predicted deletion amounted to 87 bp, or 0.52 standard deviations.

Discussion

We have presented HaploClique, a method for local haplotype reconstruction, structural variant detection of large insertions and deletions, and global haplotype assembly, which represents a principled approach to viral quasispecies assembly from NGS paired-end reads. HaploClique builds on a read alignment graph as underlying combinatorial model, where nodes correspond to single-end or paired-end alignments of reads. Edges are modeled in a probabilistic fashion.

They are based on sequence similarity of the read overlap by incorporating phred-style quality scores in combination with a position-wise prior for the non-overlapping parts of the reads, and on a criterion that measures insert size compatibility of the two alignments. While the sequence similarity criterion accounts for correct assembly of reads, the insert size criterion allows for detecting insertions and deletions in viral haplotypes that cannot be detected from single-end read alignments alone. We suggest a model that unifies sequencing error correction, clustering reads into haplotype groups, as well as assembling reads into longer fragments, all of which naturally emerge from the model.

In the read alignment graph, max-cliques represent maximal read sets that overlap and represent (locally) identical haplotype sequence. The advantage of the max-clique computation is twofold. First, it clusters reads, thereby separating reads stemming from different haplotypes.

Second, it enables sequencing error correction in a way that can make full use of co-occurrence, that is, statistical correlation of variant alleles within reach of the reads participating in a max-clique. In particular, the error correction exploits paired-end information if provided. The improved error correction is important, as it gives rise to improved frequency estimates and allows for distinguishing between haplotypes whose pairwise distance is below 1%.

HaploClique allows for reconstructing full-length global haplotypes using a read assembly procedure that is orthogonal to all existing assembly methods. In our iterative approach, we alternate between transforming max-cliques into super-reads, which form the nodes of a new alignment graph, and finding max-cliques in the new graph. We repeat this process until convergence, which is established when super-reads do not grow any longer.

HaploClique depends on three parameters to be adjusted manually: minimal read overlap, D, a threshold for the probability that two overlapping reads stem from locally identical haplotypes, δ, and the minimal coverage to call the super-read sequence κ. In general, if one of the parameters is decreased, the number and size of cliques will increase. If D and δ are too small, the purity of cliques will decrease, meaning reads from different but very similar haplotypes cluster. If D is too large, cliques will grow slower and less frequent haplotypes may be missed. If δ is too large, reads are more likely to cluster not only if they stem from the same haplotype but also if they have technical errors in common; this leads to lower error correction efficiency. If κ is too small, there might not be enough statistical power to correct for sequencing errors. If κ is too large, the false negative rate will rise, as low-frequency haplotypes do not provide enough reads to form cliques. We used two different parameter sets for HaploClique. In the first iteration, local haplotype reconstruction with error correction is performed and we chose $D = 0.8, \delta = 0.95$, and $\kappa = 5$. In practice, the results are insensitive to the parameter choice (Figure S2). For the following iterations, the quasispecies assembly, we assume that haplotypes are error-corrected and must match perfectly. We set $\delta = 0.99$ to account only for the stochasticity of the Phred scores.

We evaluated HaploClique by extensive simulation studies. The simulated haplotypes were well-known and much analyzed HIV-1 virus strains. We kept coverage in the simulation study rather low, so as to evaluate our tool in the presence of only weak signals. We did this also in comparison with extant state-of-the-art tools. We demonstrated that our approach has superior error correction capabilities. This, in turn, yields accurate haplotype frequency estimates, even at the rather low coverage of 120x per haplotype. The tools we compared to were not able to provide similarly accurate frequency estimates. HaploClique proved to be insensitive to a coverage reduction of one order of magnitude with respect to prevalent sequencing experiments, which commonly operate at 5000x or higher.

Beyond improved frequency estimates, we also improve haplotype sequence reconstruction. In all experiments, more than 99% of the haplotype segments we predict perfectly matched true haplotype sequences. None of the other tools generated even only one such perfectly matching segment, possibly because they require much higher coverage. This improvement in terms of accuracy may be due to the probabilistic model that treats error correction and assembly within one unifying framework. Our simulations also indicated that the degree of heterogeneity required in order to reconstruct large enough haplotype segments can be lower than 1%.

We also ran HaploClique on a real, Illumina MiSeq dataset of coverage 80,000x, which was found to consist of two HCV strains one of which had a frequency of only approximately 3% and contained a deletion of size 444 bp, as conformed by independent 454/Roche and PacBio sequencing experiments. In the MiSeq dataset, the deletion in question could not be detected by state-of-the-art read alignment tools. HaploClique successfully predicts this deletion, despite the large standard deviation of the fragment size distribution (≈ 167 bp). These experiments document that our method can detect large deletions also in Illumina paired-end datasets that otherwise would be difficult to identify.

Despite these improvements over previous methods, there are limitations of this approach. For example, the runtime of HaploClique is exponential in the read coverage. This feature is critical in the first two iterations of the procedure, before the number of reads is decreased. We observed that, approximately, the runtime doubles for each additional 250 reads of coverage. The baseline runtime was ~ 4 minutes for a data set with coverage 1000x, on a single 3GHz core. To overcome this computational bottleneck, one may perform the first iterations of haplotype reconstruction on subsets of the data and then assemble the merged results. Another extension that may decrease the runtime is to employ improved clustering techniques [37].

In the future, we also plan to explore on human whole-genome data, including polyploid cancer genomes, to perform error correction of the paired-end reads by local haplotype reconstruction and to assemble diploid haplotypes. This problem is more challenging due to the larger genome size and smaller levels of diversity, but several ideas presented here and implemented in HaploClique may prove useful for this task.

Methods

HaploClique performs paired-end error correction, local and global haplotype reconstruction, and structural variant calling based on enumerating cliques in the read alignment graph. We first explain the graph construction and then how super-reads are built, how global haplotypes are iteratively constructed, and how haplotype abundancies are estimated.

Read Alignment Graph

Let $\mathcal{R}$ be the set of all reads from a viral quasispecies sequencing experiment and $\mathcal{A} := \mathcal{A}(\mathcal{R})$ the set of their alignments to a reference genome as computed by a read aligner. In this paper, we assume that each read can be uniquely mapped, which is a reasonable assumption for short, non-repetitive viral genomes. In our experiments, we use the Illumina MiSeq technology for sequencing and BWA-MEM [36] as a read aligner. However, HaploClique depends on the sequencing technology and read mapper only insofar that it expects reads to be equipped with quality scores and that the reads can be properly aligned.

We construct a graph $G = (\mathcal{A}, E)$ where the read alignments $\mathcal{A}$ are the vertices. An edge $(A, B) \in E$ indicates that the two alignments A and B overlap sufficiently and that the corresponding reads are likely to originate from (locally) identical haplotypes. More precisely, we draw an edge between A and B if they satisfy two criteria based on sequence similarity and insert sizes, respectively. While the sequence-based criterion ensures that the reads R_A and R_B do not exhibit mutually contradictory sequences, the insert size-based criterion guarantees that A and B do not contradict each other in terms of their fragment sizes. We allow alignments A and B to be any combination of single- and paired-end reads, but the size-based criterion applies only if both alignment are based on paired-end reads.

Sequence similarity criterion. We define pairwise sequence distance as Hamming distance. We assume if reads stem from the same haplotype, their sequences are identical up to sequencing errors in the intervals of overlapping alignments.

The reference alignments A and B induce a direct read-to-read alignment of R_A and R_B. Let $s \in \{1, ..., m\}$ index the positions of the resulting read-to-read alignment (Figure 1). We consider the subset of positions $U_{A \cap B} \subset \{1, ..., m\}$ that are covered by both reads R_A and R_B. For $s \in U_{A \cap B}$, let $R_A[s]$ and $R_B[s]$ be the corresponding nucleotides or gap symbols ("$-$"). We construct the induced read-to-read alignment, so that no column contains gap symbols in both rows. If there is a gap $R_A[s] = -$ or $R_B[s] = -$ for some $s \in U_{A \cap B}$, then we consider A and B *incompatible* and do not connect them by an edge.

In the following, we assume a gapless read-to-read alignment, $R_A[s] \neq -$ and $R_B[s] \neq -$ for all $s \in U_{A \cap B}$. Each nucleotide in each read comes with a base calling quality score (phred score) determined by the sequencer. Let $q_A[s]$ and $q_B[s]$ be the corresponding probabilities that $R_A[s]$ and $R_B[s]$, respectively, was sequenced erroneously. For $X \in \{A, C, G, T\}$, we define

$$Q_A[s][X] := \begin{cases} 1 - q_A[s] & \text{if } X = R_A[s] \\ q_A[s]/3 & \text{otherwise} \end{cases} \qquad (1)$$

and compute

$$\mathbf{P}_M(A, B) := \prod_{s \in U_{A \cap B}} \sum_{X \in \{A,C,G,T\}} Q_A[s][X] \cdot Q_B[s][X]. \qquad (2)$$

$\mathbf{P}_M(A,B)$ is the probability that the underlying DNA sequences of R_A and R_B are identical on the overlap $U_{A\cap B}$. The actual reads R_A and R_B might differ on $U_{A\cap B}$ due to sequencing errors.

Let $T_{A\setminus B}$ denote all reference positions covered by A but not by B. Let $q_0[t]$ be the probability that the nucleotides of two randomly drawn reads coincide when being aligned with reference position t. This quantity can be estimated based on the empirical allele frequency distribution at position t, denoted $\rho[t]$. If $\rho[t][X]$ denotes the probability of observing nucleotide X at position t, then $q_0[t] = \sum_{X\in\{A,C,G,T\}} \rho[t][X]^2$. The probability that two randomly drawn alignments that, in contrast to A and B, both cover $T_{A\setminus B}$ and $T_{B\setminus A}$, exhibit identical nucleotides at all these positions is

$$\mathbf{P}_0(A,B) := \prod_{t\in T_{A\setminus B}\cup T_{B\setminus A}} q_0[t]. \tag{3}$$

To finally decide whether two reads R_A and R_B are likely to originate from (locally) identical haplotypes, we consider

$$\mathbf{P}(A,B) := \mathbf{P}_M(A,B)\cdot\mathbf{P}_0(A,B), \tag{4}$$

and say that two alignments A and B satisfy the *sequence similarity criterion* if

1. they do not contain incompatible gaps: $R_A[s] \neq -$ and $R_B[s] \neq -$ for all $s \in U_{A\cap B}$,
2. there is sufficient overlap: $|U_{A\cap B}| > D \cdot \min\{|R_A|, |R_B|\}$, and
3. the probability that the two reads were sampled from the same haplotype is sufficiently large: $\mathbf{P}(A,B)^{\frac{1}{|T_{A\setminus B}|+|U_{A\cap B}|+|T_{B\setminus A}|}} > \delta$, where the exponent ensures proper length normalization.

Insert size criterion. If a deletion or insertion is present, the distance between the reference alignments of the two ends increases (for deletions) or decreases (for insertions) compared to the situation without indels. Such insert size discrepancies can thus indicate the presence of indels as well as the (in)compatibility of two read pairs. Following [27], we define an insert size criterion based on this observation. This criterion only applies to aligments A and B of two paired-end reads R_A and R_B.

Let x_A and y_A be the rightmost position of the left end and the leftmost position of the right end, respectively, of alignment A. Let $I(A) := y_A - x_A - 1$ be the alignment interval length. For two overlapping alignments A and B, let $O(A,B) := \min(y_A, y_B) - \max(x_A, x_B) - 1$ be the length of the overlap of the alignment intervals and $\bar{I}(A,B) := (I(A) + I(B))/2$ the mean interval length (Figure S3). Let Z be $\mathrm{Norm}(0,1)$-distributed and μ and σ be mean and variance, respectively, of the empirical insert size distribution. The *insert* criterion is satisfied if

$$\mathbf{P}\left(|Z| \geq \frac{1}{\sqrt{2}}\frac{|I(A)-I(B)|}{\sigma}\right) \leq 0.05 \quad \text{and} \quad \mathbf{P}\left(Z \geq \sqrt{2}\frac{\bar{I}(A,B)-O(A,B)-\mu)}{\sigma}\right) \leq 0.05,$$

that is, if the alignments have similar interval lengths and sufficient overlap [27] (Figure S4).

Edge Definition. HaploClique has two modes. When the input consists solely of paired-end reads with a known insert size distribution, then both the sequence similarity criterion and the insert size criterion can be used, and an edge is drawn if both criteria are satisfied. If this is not the case, only the sequence similarity criterion applies. In the former mode, the statistical power to distingiush different haplotypes is larger due to employing two independent criteria. This mode also allows for calling indels based on the average insert size of alignments in a clique. Applying only the sequence similarity criterion, on the other hand, comes with the flexibility of mixing paired-end reads, single-end reads, and super-reads.

Max-clique enumeration

Cliques $C \subset \mathcal{A}$ are fully connected subgraphs of the read alignment graph $G = (\mathcal{A}, E)$. They indicate groups of reads that are all likely to stem from locally identical haplotypes. Hence max-cliques form maximal groups of reads that originate from locally identical haplotypes.

The algorithm proceeds by first sorting all nodes $\mathcal{A} = \{A_1, ..., A_m\}$ from left to right, in ascending order of the alignment coordinates such that A_t starts left of A_{t+1}. The algorithm then computes maximal cliques by processing all nodes $\mathcal{A} = \{A_1, ..., A_m\}$ in this order. Let G_t be the induced subgraph of G with vertices $\mathcal{A}_t = \{A_1, ..., A_t\}$ and let $\mathcal{C}_t$ be all max-cliques in G_t. For a node A, let $N(A)$ be all nodes that are connected to A by an edge $e \in E$. If the rightmost coordinates of all alignments in a clique $C \in \mathcal{C}_t$ are smaller than the leftmost coordinate of A_{t+1}, this clique cannot be further affected by any node $A \in \{A_{t+1}, ..., A_m\}$ — such cliques are maximal in G and can be output if only nodes $A \in \{A_{t+1}, ..., A_m\}$ are left to be considered. After having processed all nodes $\{A_1, ..., A_t\}$, we declare all cliques $C \in \mathcal{C}_t$ that can still be affected by nodes $A \in \{A_{t+1}, ..., A_m\}$ to be *active*.

When processing node A_{t+1}, we compute its neighborhood $N(A_{t+1})$ and add a new clique $C := \{A_{t+1}\}$ if intersecting $N(A_{t+1})$ with each active clique yields the empty set. Otherwise, for each active clique C, we set $C := C \cup \{A_{t+1}\}$ if $C = C \cap N(A_{t+1})$, and we add a new clique $C \cap N(A_{t+1}) \cup \{A_{t+1}\}$ if $\emptyset \neq C \cap N(A_{t+1}) \neq N(A_{t+1})$. Among all new cliques to be added, we eliminate duplicates.

Max-clique enumeration is related to the problem of finding a minimum clique cover [38], where a minimal set of non-overlapping max-cliques is sought, whose vertices cover the graph.

Runtime Analysis. Let c be an upper bound on the number of active cliques and a be an upper bound on the number of active nodes, where we call a node active if it is part of at least one active clique. Since no two distinct active cliques can be extended to the same max-clique, we have $c \leq s$, where s is the number of max-cliques in G. Let m be the number of alignments $\mathcal{A}$. First, sorting all nodes requires $O(m \log m)$ time. Processing each node requires $O(ca)$ time (when doing the duplicate removal by radix sorting bit vectors representing the new cliques). Therefore, the total runtime is $O(m(\log m + ca) + s)$. Since $a \leq m$ and $c \leq s$, the algorithm has *output polynomial* runtime, that is, it is polynomial in the size of the input plus the size of the output.

In general graphs and in the worst case, the number of max-cliques is exponential in the number of nodes [39]. Thus, our algorithm is not necessarily polynomial in the input alone. In interval graphs, however, the number of cliques grows at most linearly and simple sweep algorithms can enumerate all max-cliques in linear time. In these graphs, each vertex is an interval on the real line and an edge is drawn between two vertices whenever the corresponding intervals overlap. The read alignment graph as defined above can be regarded as an interval graph with removed edges. Indeed, that two alignments overlap is necessary but not sufficient for drawing an edge. Although, formally, any graph can be constructed by removing edges from an interval graph, being close to an interval graph with bounded coverage is a property that one can exploit. Due to the "banded" shape of the graph (Figure 1), the number of maximal cliques grows linearly in the genome length (and thus in the number of nodes) in practice (Figure S5B). Under this assumption, the runtime of the clique finder we use [27] is also linear in the number of nodes once they have been sorted. This linear behavior was indeed observed in computational experiments using viral genome lengths differing by several orders of magnitude (Figure S5). We optimized the implementation of the algorithm by exploiting bit-parallelism whenever possible. Being linear in the genome length, in practice, is crucial for analyzing the haplotype structure of longer viral or bacterial, or eventually also cancer genomes. It is for this reason, that we have selected this algorithm and prefer it over fast general purpose algorithms for max-clique enumeration in sparse graphs [40,41], which generally scale quadratically with genome length.

Super-read assembly

Each max-clique is a set of reads with mutually compatible alignments. Therefore, we can construct consensus sequences for max-cliques, which we refer to as super-reads. The purpose of super-read construction is two-fold. First, super-reads represent haplotype segments. Second, super-reads can be used as input to further iterations of max-clique enumeration, with the goal of global haplotype reconstruction, which is discussed in the next section.

To construct super-reads, let $A_1, ..., A_k$ be the k alignments participating in a max-clique and let $U^{\kappa}_{A_1,...,A_k}$ be the set of positions where at least κ of these alignments contain non-gap characters. We recall that $Q_A[s][X]$ denotes the probability that nucleotide X gave rise to position s in read R_A, although $R_A[s]$ might differ from X due to sequencing errors.

We determine the nucleotide sequence of the super-read $\bar{R}(R_{A_1}, ..., R_{A_k})$ by means of a weighted position-wise majority vote. We set $\bar{R}(R_{A_1}, ..., R_{A_k})[s] = \text{argmax}_{X\in\{A,C,G,T\}} \sum_{i=1}^{k} Q_{A_i}[s][X]$, for each position $s \in U^{\kappa}_{A_1,...,A_k}$, where $Q_{A_i}[s][X]$ is defined to be zero when A_i does not cover position s. The parameter κ ensures that the super-reads have sufficient coverage of high quality. For later frequency estimation, we keep track of which original reads gave rise to which super-read.

Global assembly strategy

For global haplotype reconstruction, or quasispecies assembly, we iterate the clique enumeration procedure. We align reads against the reference sequence, construct the read alignment graph, find max-cliques, and merge them into larger super-reads with updated phred scores that reflect corrected error profiles. In the next iteration, we use these super-reads as reads and restart the procedure until number and length of super-reads have converged. For the assembly step, we assume that reads have already been error-corrected in the first iteration and set $\delta = 0.99$. Reads have to match perfectly and δ only allows for stochasticity of the Phred scores. We start the iterations with a relative overlap of $D = 0.8$. Once the length and number of super-reads converged, we decrease D by 0.1 down to a minimum of $D = 0.5$.

Haplotype abundance estimation

We estimate haplotype abundance by counting the number of (original) reads that participate in the super-reads giving rise to the haplotypes. Original reads may participate in several super-reads and thereby contribute to abundance counts for several haplotypes. We resolve this issue by keeping track of the original read in each iteration, such that each read can be assigned to the final haplotypes after convergence. Reads contributing to several haplotypes abundances are then taken into account by weighting them accordingly.

Data

The MiSeq raw read data set is available through the Sequence Read Archive under the BioProject accession number `SRP034655`. The MiSeq 2x250 bp error profiles for SimSeq [34] used in the simulations are available at `https://github.com/armintoepfer/haploclique` under `data`.

Acknowledgments

References

1. Domingo E, Sheldon J, Perales C (2012) Viral quasispecies evolution. Microbiol Mol Biol Rev 76: 159–216.

2. Töpfer A, Höper D, Blome S, Beer M, Beerenwinkel N, et al. (2013) Sequencing approach to analyze the role of quasispecies for classical swine fever. Virology 438: 14–19.

3. Vignuzzi M, Stone JK, Arnold JJ, Cameron CE, Andino R (2006) Quasispecies diversity determines pathogenesis through cooperative interactions in a viral population. Nature 439: 344–348.

4. Henn MR, Boutwell CL, Charlebois P, Lennon NJ, Power KA, et al. (2012) Whole genome deep sequencing of HIV-1 reveals the impact of early minor variants upon immune recognition during acute infection. PLoS pathogens 8: e1002529.

5. Beerenwinkel N, Günthard HF, Roth V, Metzner KJ (2012) Challenges and opportunities in estimating viral genetic diversity from next-generation sequencing data. Front Microbio 3: 239.

6. Hinkley T, Martins J, Chappey C, Haddad M, Stawiski E, et al. (2011) A systems analysis of mutational effects in HIV-1 protease and reverse transcriptase. Nat Genet 43: 487–489.

7. Jorde LB, Wooding SP (2004) Genetic variation, classification and 'race'. Nature genetics 36: S28–S33.

8. Eriksson N, Pachter L, Mitsuya Y, Rhee SY, Wang C, et al. (2008) Viral population estimation using pyrosequencing. PLoS Comput Biol 4: e1000074.

9. Macalalad AR, Zody MC, Charlebois P, Lennon NJ, Newman RM, et al. (2012) Highly sensitive and specific detection of rare variants in mixed viral populations from massively parallel sequence data. PLoS Comput Biol 8: e1002417.

10. Quince C, Lanzen A, Davenport RJ, Turnbaugh PJ (2011) Removing noise from pyrosequenced amplicons. BMC Bioinformatics 12: 38.

11. 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.

12. Skums P, Dimitrova Z, Campo DS, Vaughan G, Rossi L, et al. (2012) Efficient error correction for next-generation sequencing of viral amplicons. BMC Bioinformatics 13 Suppl 10: S6.

13. Salzberg S, Phillippy A, Zimin A, Pulu D, Magoc T, et al. (2012) Gage: A critical evaluation of genome assemblies and assembly algorithms. Genome Research 22: 557-567.

14. Prabhakaran S, Rey M, Zagordi O, Beerenwinkel N, Roth V (2010) HIV haplotype inference using a constraint-based dirichlet process mixture model. In: NIPS Workshop on Machine Learning in Computational Biology.

15. Töpfer A, Zagordi O, Prabhakaran S, Roth V, Halperin E, et al. (2013) Probabilistic inference of viral quasispecies subject to recombination. Journal of Computational Biology 20: 113–123.

16. Prosperi MCF, Salemi M (2012) Qure: software for viral quasispecies reconstruction from next-generation sequencing data. Bioinformatics 28: 132–133.

17. Astrovskaya I, Tork B, Mangul S, Westbrooks K, Mandoiu I, et al. (2011) Inferring viral quasispecies spectra from 454 pyrosequencing reads. BMC Bioinformatics 12 Suppl 6: S1.

18. Mancuso N, Tork B, Mandoiu I, Skums P, Zelikovsky A (2011) Viral quasispecies reconstruction from amplicon 454 pyrosequencing reads. In: Proc. 1st Workshop on Computational Advances in Molecular Epidemiology. pp. 94-101. URL `http://dx.doi.org/10.1109/BIBMW.2011.6112360`.

19. O'Neil ST, Emrich SJ (2012) Haplotype and minimum-chimerism consesus determination using short sequence data. BMC Genomics 13: S4.

20. Huang A, Kantor R, DeLong A, Schreier L, Istrail S (2011) Qcolors: An algorithm for conservative viral quasispecies reconstruction from short and non-contiguous next generation sequencing reads. In: IEEE International Conference on Bioinformatics and Biomedicine Workshops. pp. 130 -136. doi:10.1109/BIBMW.2011.6112365.

21. Aguiar D, Istrail S (2013) Haplotype assembly in polyploid genomes and identical by descent shared tracts. Bioinformatics 29: i352–i360.

22. Aguiar D, Istrail S (2012) Hapcompass: a fast cycle basis algorithm for accurate haplotype assembly of sequence data. Journal of Computational Biology 19: 577–590.

23. Schirmer M, Sloan WT, Quince C (2012) Benchmarking of viral haplotype reconstruction programmes: an overview of the capacities and limitations of currently available programmes. Brief Bioinform .

24. Zagordi O, Däumer M, Beisel C, Beerenwinkel N (2012) Read length versus depth of coverage for viral quasispecies reconstruction. PLOS ONE 7: v.

25. Prosperi MC, Yin L, Nolan DJ, Lowe AD, Goodenow MM, et al. (2013) Empirical validation of viral quasispecies assembly algorithms: state-of-the-art and challenges. Scientific reports 3.

26. Ocwieja KE, Sherrill-Mix S, Mukherjee R, Custers-Allen R, David P, et al. (2012) Dynamic regulation of HIV-1 mrna populations analyzed by single-molecule enrichment and long-read sequencing. Nucleic acids research 40: 10345–10355.

27. Marschall T, Costa IG, Canzar S, Bauer M, Klau GW, et al. (2012) Clever: clique-enumerating variant finder. Bioinformatics 28: 2875–2882.

28. Marschall T, Hajirasouliha I, Schönhuth A (2013) Mate-clever: Mendelian-inheritance-aware discovery and genotyping of midsize and long indels. Bioinformatics .

29. Gnerre S, et al. (2011) High-quality draft assemblies of mammalian genomes from massively parallel sequence data. Proc Natl Acad Sci USA 108: 1513-1518.

30. Li B, Ruotti V, Stewart RM, Thomson JA, Dewey CN (2010) Rna-seq gene expression estimation with read mapping uncertainty. Bioinformatics 26: 493–500.

31. Zerbino DR, Birney E (2008) Velvet: algorithms for de novo short read assembly using de bruijn graphs. Genome Res 18: 821–829.

32. Bansal V, Bafna V (2008) Hapcut: an efficient and accurate algorithm for the haplotype assembly problem. Bioinformatics 24: i153–i159.

33. 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.

34. Earl DA, Bradnam K, St John J, Darling A, Lin D, et al. (2011) Assemblathon 1: A competitive assessment of de novo short read assembly methods. Genome Research .

35. Bull RA, Luciani F, McElroy K, Gaudieri S, Pham ST, et al. (2011) Sequential bottlenecks drive viral evolution in early acute hepatitis c virus infection. PLoS Pathog 7: e1002243.

36. Li H (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv .

37. Bellitto T, Marschall T, Klau G, Schönhuth A (2013) Next-generation cluster editing. Technical Report 1310.3353, ArXiV.

38. Karp R (1972) Reducibility among combinatorial problems. In: Miller R, Thatcher J, editors, Proceedings of a Symposium on the Complexity of Computer Computations. Plenum Press, pp. 85-103.

39. Moon JW, Moser L (1965) On cliques in graphs. Israel Journal of Mathematics 3: 23–28.

40. Eppstein D, Löffler M, Strash D (2010) Listing all maximal cliques in sparse graphs in near-optimal time. In: Cheong O, Chwa KY, Park K, editors, Algorithms and Computation, Springer Berlin Heidelberg, volume 6506 of *Lecture Notes in Computer Science*. pp. 403–414.

41. Eppstein D, Strash D (2011) Listing all maximal cliques in large sparse real-world graphs. CoRR abs/1103.0318.

Figure Legends

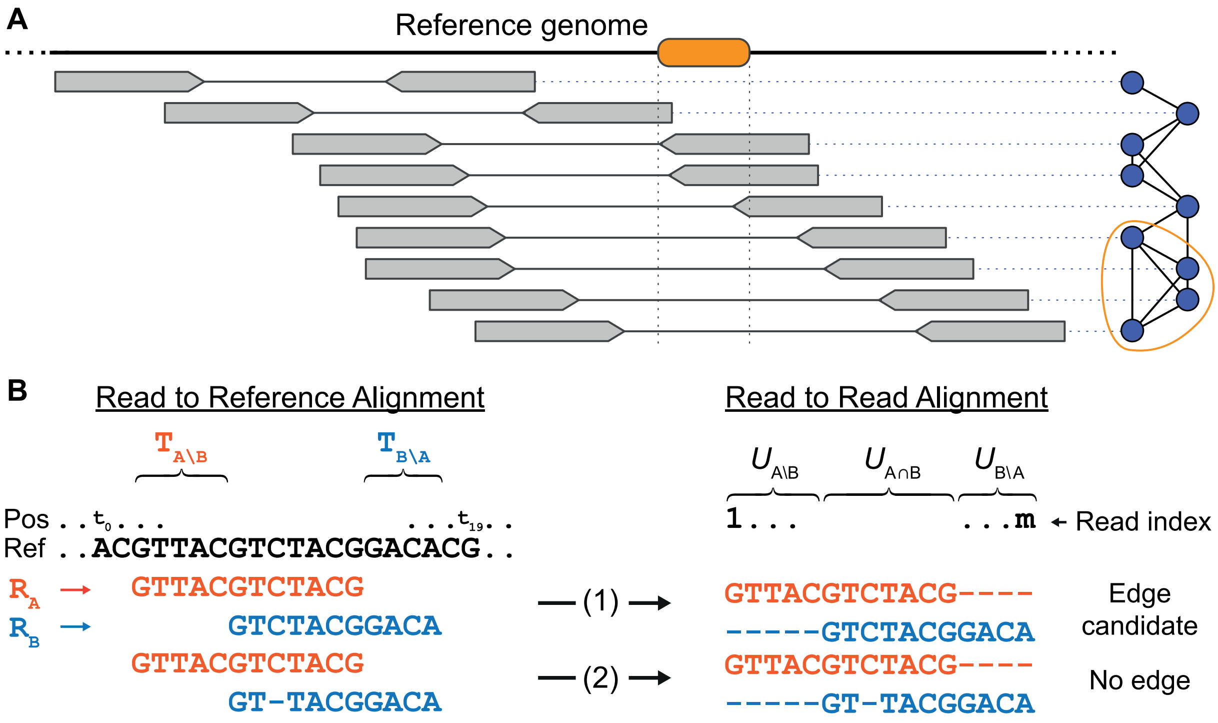
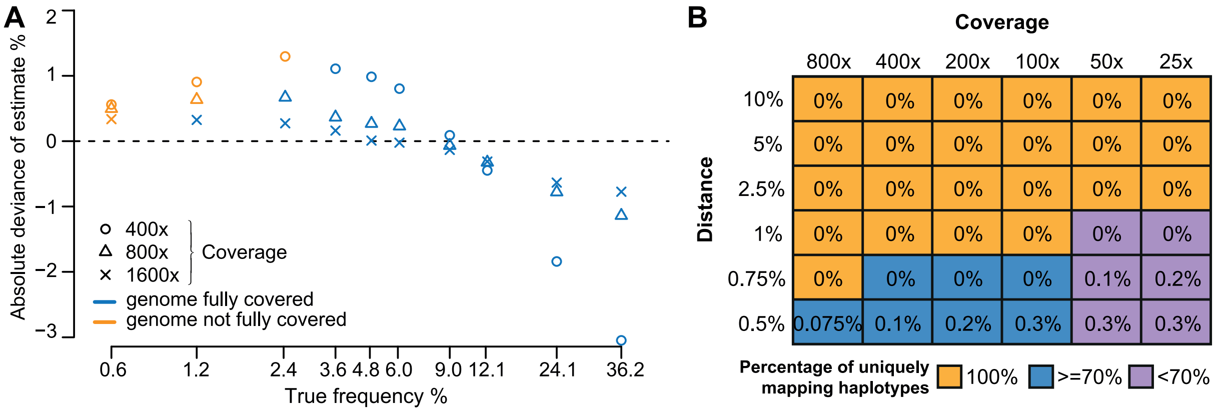


Figure 1. Max-clique enumeration and edge definitions. (A) Example of a read alignment graph based on the insert size criterion. Alignments of read pairs are shown in gray and the corresponding nodes in the graph representation are depicted in blue. The four bottom-most alignment pairs stem from a haplotype harboring a deletion (shown in orange in the reference genome) and therefore display a larger insert size than the remaining alignment pairs. Note that the four deletion-indicating alignment pairs form a max-clique (circled in orange). (B) Illustration of the compatible gaps condition of the sequence similarity criterion. Two reads R_A and R_B are aligned against the reference (left). This induces a direct read-to-read alignment of R_A and R_B (right). Case (1): No gaps in the reference alignments lead to a gapless read-to-read alignment, which renders the pair of reads an edge candidate. Case (2): Gaps in the reference alignment lead to gaps in the read-to-read alignment, excluding the possibility of an edge. See also Figure S6 in the appendix for more complicated cases involving gaps.



Distance	800x	400x	200x	100x	50x	25x
10%	🟧 0%	🟧 0%	🟧 0%	🟧 0%	🟧 0%	🟧 0%
5%	🟧 0%	🟧 0%	🟧 0%	🟧 0%	🟧 0%	🟧 0%
2.5%	🟧 0%	🟧 0%	🟧 0%	🟧 0%	🟧 0%	🟧 0%
1%	🟧 0%	🟧 0%	🟧 0%	🟧 0%	🟪 0%	🟪 0%
0.75%	🟧 0%	🟦 0%	🟦 0%	🟦 0%	🟪 0.1%	🟪 0.2%
0.5%	🟦 0.075%	🟦 0.1%	🟦 0.2%	🟦 0.3%	🟪 0.3%	🟪 0.3%

Percentage of uniquely mapping haplotypes: 🟧 100% 🟦 >=70% 🟪 <70%

Figure 2. Performance in (A) frequency estimation and (B) distinguishing reconstructed local haplotypes. (A) Ten haplotypes were sampled with different frequencies (x-axis, logarithmic scale), and the mean deviations of the estimated to the true frequencies are reported for ten repetitions of the simulation (y-axis). The different symbols represent data sets with coverages 400x, 800x, and 1600x. Color indicates whether the genome was fully covered by predicted haplotypes (blue) or not (orange). (B) Performance in distinguishing reconstructed local haplotypes, depending on pairwise distance and coverage. The displayed percentages are the fractions of super-reads that do not match any true haplotype without error. Color-coded is the fraction of super-reads that match exactly one true haplotype (100%, orange; $\geq 70\%$, blue; $< 70\%$, violet).

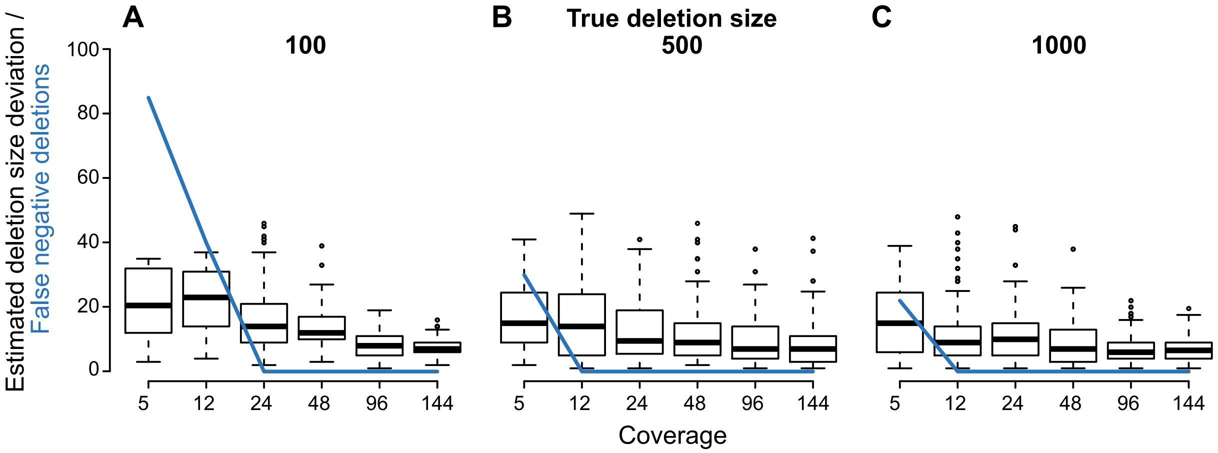


Figure 3. Large deletion estimates. Estimated deletion size deviation and false negative rate for different true deletion sizes of (A) 100, (B) 500, and (C) 1000 bp. For each deletion length and each coverage of 5, 12, 24, 48, 96, and 144x, a boxplot summarizes the deviations of the estimated to the true deletion size in 100 simulated samples. The blue line represents the number of false negative predicted deletions in each of the 100 samples.

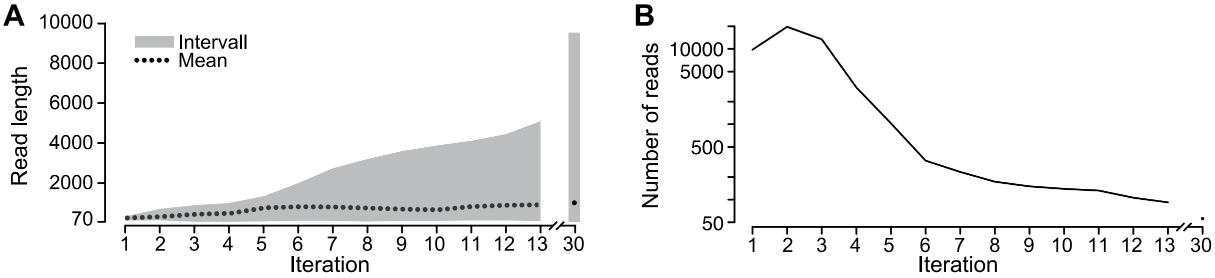


Figure 4. Global haplotype assembly results. Minimum, maximum, and mean read lengths (A) and the total number of reads (B) for the global haplotype assembly of the lab-mix, for the first 13 and the last iteration (30).

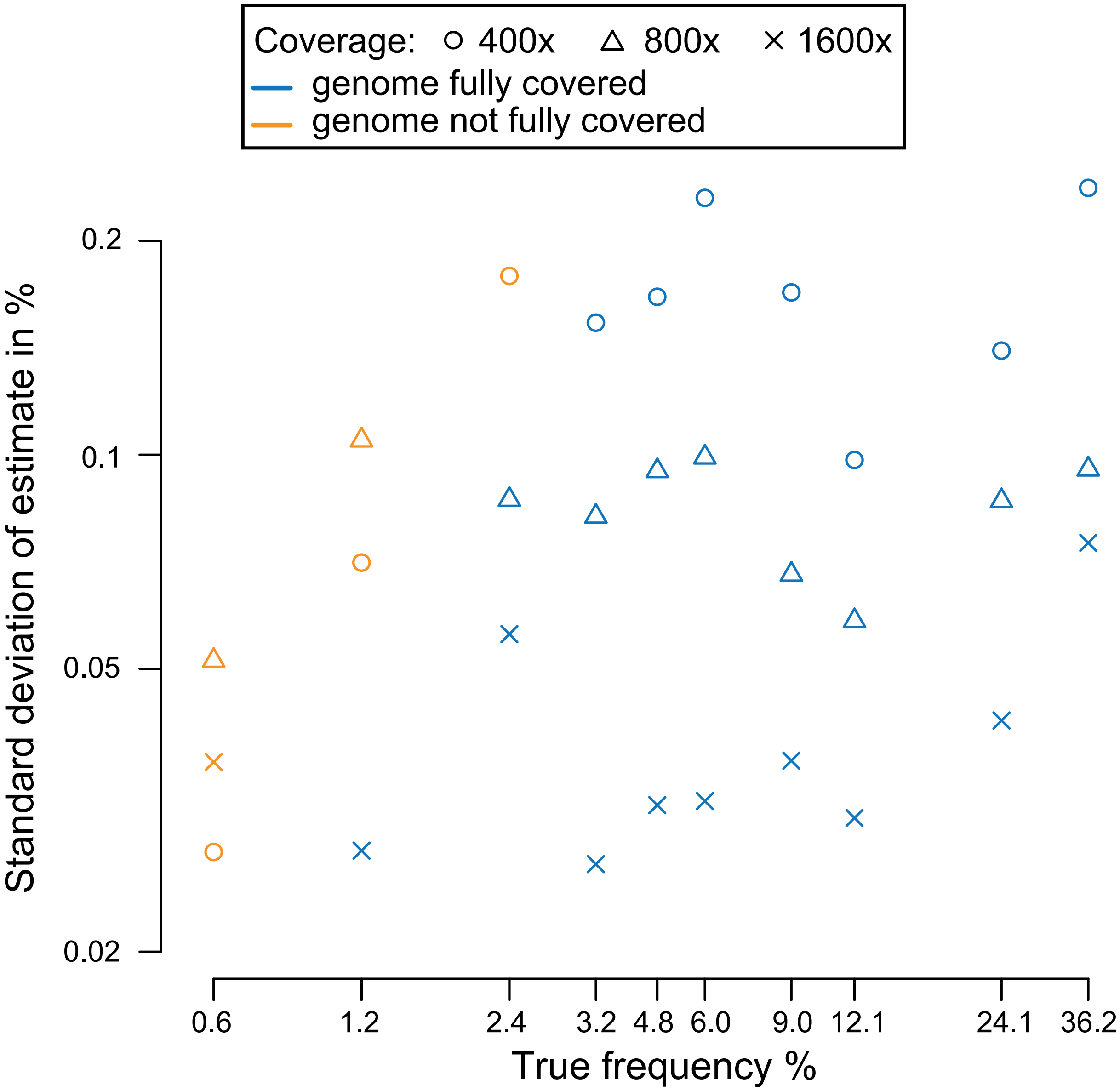


Figure S1. Standard deviation of ten haplotype frequency estimates for different coverages. Ten haplotypes were sampled with different frequencies (x-axis, logarithmic scale), and the standard deviations of the frequency estimates are reported for ten repetitions of the simulation (y-axis, logarithmic scale). The different symbols represent data sets with coverages of 400x, 800x, and 1600x. Color indicates whether the genome was fully covered by predicted haplotypes (blue) or not (orange).

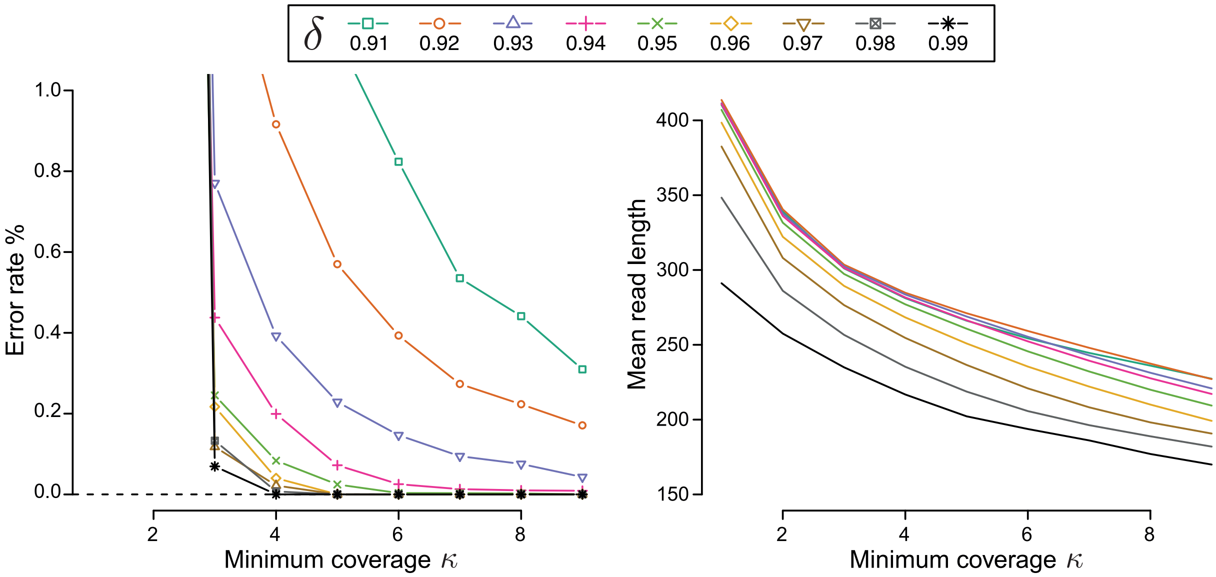


Figure S2. Empirical studies for parameters κ and δ. Using the lab-mix, as described in the Results section, we benchmarked the reconstruction performance, with respect to error rate (left) and mean read length (right), for varying levels of κ, the number of reads required for initial super-read construction. Only sequence fragments that are supported by at least κ original reads within one max-clique are turned into super-reads of the first generation (iteration). Performance depends on the parameters δ, the threshold for the probability that both reads stem from the same haplotype and the minimal coverage κ to create a consensus sequence of the super-read. We varied δ between 0.91 and 0.99, and κ between 1 and 9.

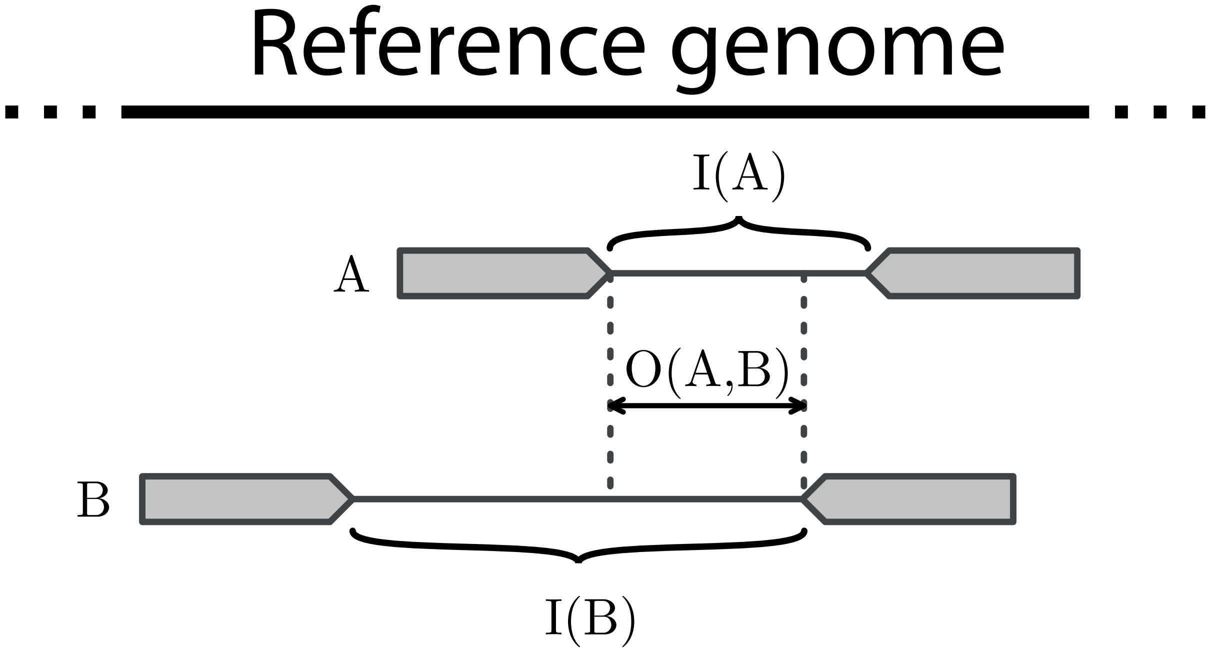


Figure S3. Overlapping inserts. Two alignment pairs A and B along with their insert sizes $I(A)$ and $I(B)$ and their overlap $O(A, B)$ are shown.

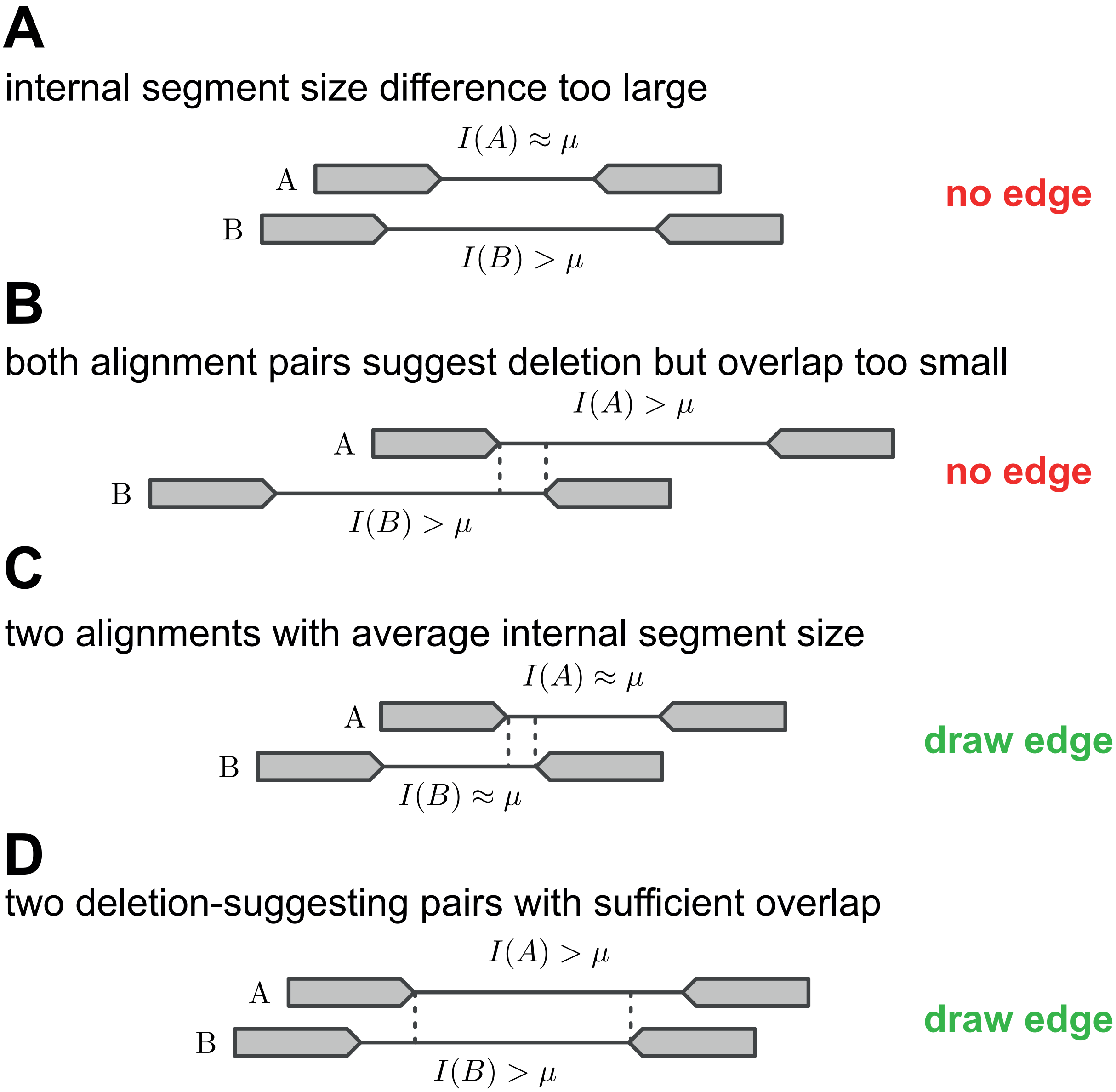

A
internal segment size difference too large
$I(A) \approx \mu$
A
B
$I(B) > \mu$
no edge
B
both alignment pairs suggest deletion but overlap too small
$I(A) > \mu$
A
B
$I(B) > \mu$
no edge
C
two alignments with average internal segment size
$I(A) \approx \mu$
A
B
$I(B) \approx \mu$
draw edge
D
two deletion-suggesting pairs with sufficient overlap
$I(A) > \mu$
A
B
$I(B) > \mu$
draw edge

Figure S4. Insert edge definitions. The different scenarios (A)–(D) of the insert size criterion are shown.

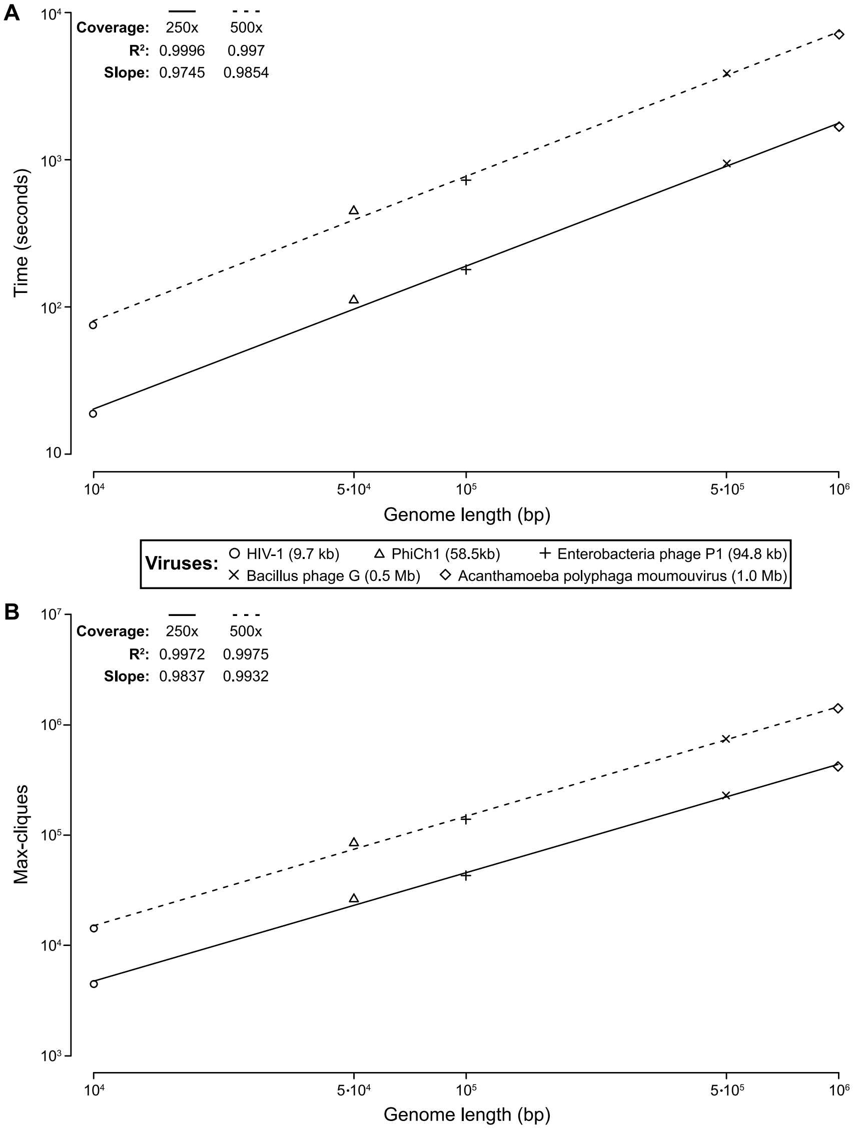


Figure S5. Runtime analysis. (A) Run time and (B) number of max-cliques for varying genome lengths between 10 kb and 1 Mb and for coverages of 250x and 500x. Dots represent observed runtime in seconds (A) and number of max-cliques in the corresponding alignment graph (B). Lines represent linear regressions after log-log transformation. The slopes of approximately one indicate linear relationships. R^2 is the fraction of variance explained by the log-log linear model. For each of the five viruses HIV-1, PhiCh1, enterobacteria phage P1, Bacillus phage G, and Acanthamoeba polyphage moumouvirus, we simulated three haplotypes with a distance of five percent to the reference genome. For each virus, we generated data sets with mean coverage 250x and 500x.

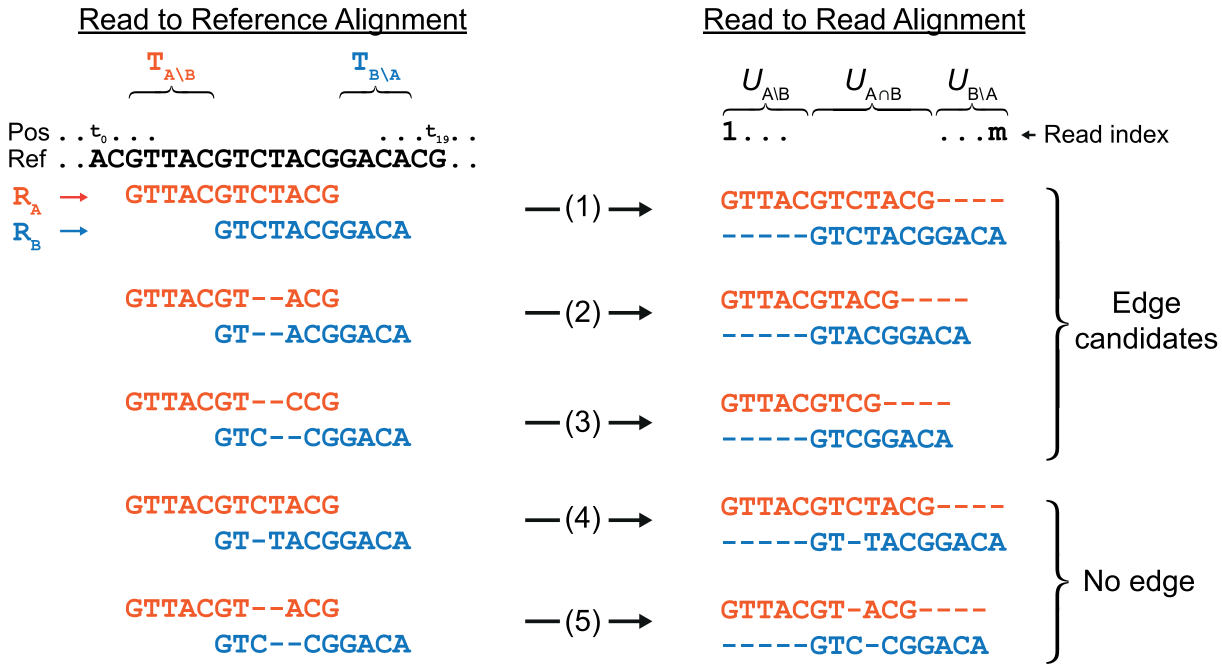


Figure S6. Edge definitions. Illustration of the compatible gaps condition of the sequence similarity criterion. Two reads R_A and R_B are aligned against the reference (left). This induces a direct read-to-read alignment of R_A and R_B (right). Case (1): No gaps in the reference alignments lead to a gapless read-to-read alignment, which renders the pair of reads an edge candidate. Case (2): Gaps in the reference alignment become eliminated in the direct read-to-read alignment implying an edge candidate. Case (3): The reference alignment leads to aligning '-C' against 'C-', which we interpret as aligning C against C, that is, virtually case (2) is in effect. Case (4): Gaps in the reference alignment that lead to gaps in the read-to-read alignment exclude the possibility of edges. Case (5): Similar to (3), but we interpret '-A' against 'C-' as gap implying that no edge is possible.

Tables

Table 1. Global haplotype assembly comparison

	Estimated HIV-1 strain frequency (max. rel. haplotype length)							
Method	HXB2	NL4-3	YU2	89.6	JR-CSF	Error rate	# haplotypes	Precision
HaploClique	22% (100%)	21% (61%)	19% (59%)	18% (61%)	20% (57%)	0.012%	56	99.7%
ShoRAH	29% (97%)	20% (97%)	20% (97%)	9% (97%)	20% (99%)	1.99%	196	0%
PredictHaplo	91% (99%)	0% (0%)	1% (3%)	6% (12%)	2% (18%)	1.65%	95	0%
QuRe	46% (91%)	7% (91%)	0% (0%)	27% (91%)	20% (93%)	2.95%	15	0%

Global haplotype assembly comparison of HaploClique with the software packages ShoRAH [33], PredictHaplo [14], and QuRe [16]. We report the estimated variant frequencies and, in parenthesis, the maximal length of the reconstructed haplotypes relative to the genome length, for each of the five variants. In the remaining columns, the average error rate (computed as the number of mistaken nucleotides, divided by the length of the haplotype computed), the total number of reconstructed haplotypes, and the precision (percentage of perfectly reconstructed haplotypes weighted by the respective estimated frequency) are reported. See Methods for more details on frequency estimation.