Viral Quasispecies Assembly via Maximal Clique Enumeration

Armin Töpfer[1,2], Tobias Marschall[3], Rowena A. Bull[4], Fabio Luciani[4], Alexander Schönhuth[3,⋆], and Niko Beerenwinkel[1,2,⋆,⋆⋆]

[1] Department of Biosystems Science and Engineering, ETH Zurich, Basel, Switzerland
[2] SIB Swiss Institute of Bioinformatics, Switzerland
[3] Centrum Wiskunde & Informatica, Amsterdam, The Netherlands
[4] Inflammation and Infection Research Centre, School of Medical Sciences, UNSW, Sydney, Australia
niko.beerenwinkel@bsse.ethz.ch

Genetic variability of virus populations within individual hosts is a key determinant of pathogenesis, virulence, and treatment outcome. It is of clinical importance to identify and quantify the intra-host ensemble of viral haplotypes, called viral quasispecies. Ultra-deep next-generation sequencing (NGS) of mixed samples is currently the only efficient way to probe genetic diversity of virus populations in greater detail. Major challenges with this bulk sequencing approach are (i) to distinguish genetic diversity from sequencing errors, (ii) to assemble an unknown number of different, unknown, haplotype sequences over a genomic region larger than the average read length, (iii) to estimate their frequency distribution, and (iv) to detect structural variants, such as large insertions and deletions (indels) that are due to erroneous replication or alternative splicing. Even though NGS is currently introduced in clinical diagnostics, the *de-facto* standard procedure to assess the quasispecies structure is still single-nucleotide variant (SNV) calling. Viral phenotypes cannot be predicted solely from individual SNVs, as epistatic interactions are abundant in RNA viruses. Therefore, reconstruction of long-range viral haplotypes has the potential to be adopted, as data is already available.

We present HaploClique, a computational method that combines a probabilistic model of sequence similarity and structural similarity with a graph theoretical method to reconstruct viral quasispecies from NGS paired-end data. We define a read alignment graph, in which nodes correspond to single-end and paired-end alignments (Figure 1A right). We draw an edge between nodes if alignments (i) have sufficient overlap, (ii) are compatible in the insert size (Figure 1A left), defined as the unsequenced fragment between read pairs, and (iii) show that sequences are sufficiently similar. Taken together these criteria ensure that both reads are likely to stem from the same haplotype (Figure 1B). If alignments stem from the same haplotype, their sequences are identical up to sequencing errors in the intervals of overlap. The corresponding probability is computed

⋆ Equal contributions.
⋆⋆ Corresponding author.

R. Sharan (Ed.): RECOMB 2014, LNBI 8394, pp. 309–310, 2014.

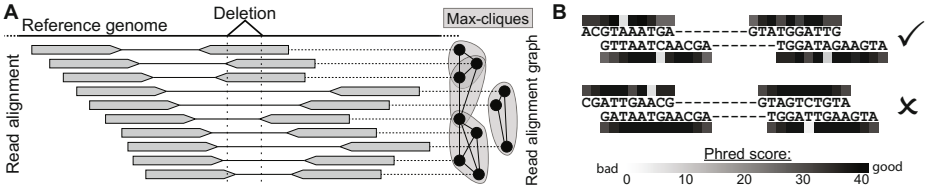

Fig. 1. (A) Paired-end read alignment with a deletion harboring haplotype and the corresponding read alignment graph with max-cliques of minimal size three, based on the insert size compatibility. (B) Phred scores of paired-end reads are used to assess sequence similarity. The unsequenced fragment is indicated by the gap symbol '-'. Top example is sequence compatible, as reads differ only in sequencing errors, bases with low phred scores. Bottom example is not sequence compatible, because reads differ in bases with high pred scores.

using the base calling quality scores (phred scores). In addition, we compute the probability that the non-overlapping alignment sequences are identical.

We develop a maximal clique (max-clique) enumeration approach (Figure 1A right) to cluster NGS reads. Max-cliques are fully connected subgraphs that cannot be extended and consist of reads with mutually compatible alignments. We use max-cliques to reconstruct haplotype sequences and detect indels. In detail, the structural similarity of all reads in a max-clique and its deviation from the empirical insert size distribution allows to detect large indels. The consensus sequence of each max-clique, called super-read, is the predicted error-corrected local haplotype. By iterating read alignment graph construction and max-clique enumeration of super-reads, haplotype fragments grow in length and possibly allow full-length reconstruction. Haplotype abundance estimation is performed by counting original reads that participated in a super-read.

In extensive simulation studies, we benchmarked the accuracy and robustness of estimating haplotype frequencies, the error correction performance, and the minimal distance of two haplotypes to be perfectly distinguishable. We showed that HaploClique outperforms the state-of-the-art tools ShoRAH, PredictHaplo, and QuRe on a simulated dataset of five well known HIV strains with a low coverage of 600x. HaploClique successfully reconstructed one haplotype at its full length and the other strains are covered with reconstructed haplotypes of sizes 5-6 kb, where the original strain lengths are 9-10 kb. The structure prediction accuracy and robustness was assessed for varying deletion sizes between 100 bp and 1 kb. We applied HaploClique to a clinical hepatitis C virus infected sample and detected a novel deletion of size 357±167 bp, which was validated by two independent long-read sequencing experiments. HaploClique is able to predict large indels that cannot be detected by current computational methods and reconstruct full-length haplotypes from low coverage samples. HaploClique's implementation is available at `https://github.com/armintoepfer/haploclique`.