Haplotype assignment of virus NGS data using co-variation of variant frequencies

Richard A. Goldstein[1*] Asif U. Tamuri[2], Sunando Roy[1], Judith Breuer [1,3,4]

1. Division of Infection and Immunity, University College London, UK
2. Research IT Services, University College London, UK
3. Infection, Immunity, Inflammation and Physiological Medicine, Institute of Child Health, University College London, UK
4. Great Ormond Street Hospital for Children, London, UK

*Corresponding author
r.goldstein@ucl.ac.uk

Abstract:

Longitudinal deep sequencing of viruses can provide detailed information about intra-host evolutionary dynamics, including how viruses interact with and transmit between hosts. Many analyses require haplotype reconstruction to identify which variants are co-located on the same genomic element. Most current methods to perform this reconstruction are based on a high density of variants and cannot perform this reconstruction for slowly evolving viruses. We describe a new approach to perform this reconstruction based on identifying co-varying variant frequencies using a probabilistic framework. We test this method with data sets of mixed cytomegalovirus genomes, demonstrating high accuracy when longitudinal samples are available.

The short generational times and high mutation rates characteristic of viruses can result in substantial intra-host evolution during a single infection. This is especially true of RNA viruses such as HIV and influenza, but even ssDNA and dsDNA viruses can undergo substantial evolutionary change if the infection persists. The intra-host evolutionary dynamics of viruses can be monitored through longitudinal deep sequencing, deepening our understanding of their population dynamics, their response to drug therapies, how they interact with their hosts and how they are transmitted between individuals (1-3).

While smaller than bacterial and eukaryotic genomes, virus genomes are still much larger than the individual reads that are obtained through next generation sequencing (NGS). Detailed analyses generally require determining which variants are found together in the same genome or genomic segment, a process known as haplotype reconstruction. This is commonly performed by identifying variants at sites that are close enough to be found on the same reads. If these variants are sufficiently dense, co-localising variants across the genome can be 'stitched together', resulting in the determination of whole genome haplotypes (4). Slowly evolving viruses such as cytomegalovirus (CMV) can have long regions with few segregating sites, making it impossible to connect variants that bridge these regions.

It is increasingly common for multiple samples isolated at different times to be available. Selection and drift result in changes in the relative frequencies of the haplotypes, and thus in the frequencies of the variants that they contain. Co-variation of variants provides an additional source of information for haplotype reconstruction, even when these variants are far apart in the genome. For this reason, we have created a new method for reconstructing whole-genome haplotypes from longitudinal sequence data (HAplotype Reconstruction Of Longitudinal Deep sequencing data, HaROLD).

We start with an assumed number of haplotypes. The likelihood of observing a given number of variants at each position is dependent on a) the frequencies of the haplotypes at each time point, b) the identity of the base found at that position on each of the haplotypes, and c) the probability of making an erroneous measurement at that site. In order to separate these factors, we calculate the likelihood by summing over all possible assignments of the bases observed at a given site to the different haplotypes; if three different bases are observed and our current model involves four haplotypes, we sum over all $3^4 = 81$ possible assignments. In this manner, the composition of the haplotypes represents a nuisance parameter.

The error rate defines the probability $p_{\mathrm{OBS}}(\mathrm{Y}|\mathrm{X})$ that base Y will be observed when base X is present. As this rate will be different at different sites and on different strands, we consider that this probability is drawn from a Dirichlet distribution, and integrate over this distribution. We can use maximum likelihood optimisation to find the optimal set of haplotype frequencies at each time point and parameter values defining the Dirichlet distribution that characterises the error rates.

Following estimation of the parameters in the model, we determine how much each assignment of bases to haplotypes contributes to the likelihood. This allows us to calculate the posterior probability of each given assignment. By summing over these posterior probabilities, we can compute the total posterior probability that a base is found at that site in each of the haplotypes. If these probabilities are sufficiently definitive, an assignment is made.

As increasing the number of haplotypes increases the number of ways of assigning bases to each of the haplotypes, decreasing the prior probability of any given assignment, the log likelihood typically decreases when the number of haplotypes increases beyond that necessary to represent the data. We can perform this calculation for a range of haplotype numbers, and select the number of haplotypes that maximise the log likelihood.

The run time is strongly dependent on the number of haplotypes; run times for the synthetic data described in the Supplementary Material (235kb, 1 to 5 haplotypes) ranged from 6s to 10m on a single 2.4GHz processor. This can be significantly longer in some cases, mostly dominated by the estimation of the error rate parameters. The calculations can, correspondingly, be greatly sped up if these parameters are estimated and fixed. *HaROLD* supports multiple threads.

The software implementing this approach is available at https://github.com/RichardAGoldstein/HaROLD.

Supplementary Material

Theory

The data: The haplotype analysis looks at matching a statistical model to longitudinal data of the form $\{n_{b,s,l,t}\}$ representing the number of reads with base b on strand s at position l at timepoint t. We consider that the reads come from a set of haplotypes where $\{x_{j,1}, x_{j,2}, x_{j,3} \dots x_{j,M}\}$ is the sequence of length M of haplotype j. At time point t, we represent the frequencies of the k different haplotypes by $\{\Pi_{1,t}, \Pi_{2,t}, \Pi_{3,t} \dots \Pi_{k,t}\}$ which obey $\sum_j \Pi_{j,t} = 1$. Multiple haplotypes might share the same base at a given location; the frequencies of base b at site l at timepoint t is equal to the sum of the frequencies of all haplotypes that have that base at that site and is equal for the two strands: $\pi_{b,s,l,t} = \sum_{j \,\forall\, x_{j,l}=b} \Pi_{j,t}$.

Computing likelihoods: We consider that, for a site with an error rate of ϵ there is a probability $1 - 3\,\epsilon$ of actually observing the true base and a probability ϵ of observing one of the three other bases. If we know $\pi_{b,s,l,t}$, then the probability $P_{s,l,t}(b_i)$ of a read at that time, strand and position being observed as a base b_i is given by

$$P_{s,l,t}(b_i) = \pi_{b_i,s,l,t}(1 - 3\,\epsilon) + \epsilon \sum_{b_j \neq b_i} \pi_{b_j,s,l,t} = \pi_{b_i,s,l,t} + \left(1 - 4\,\pi_{b_i,s,l,t}\right)\epsilon \tag{1}$$

The probability of observing $\{n_{b,s,l,t}\}_{s,l,t}$, the reads for a specified strand, site, and time, is then given by the multinomial distribution

$$P\left(\{n_{b,s,l,t}\}_{s,l,t} | \{\pi_{b,l,t}\}, \epsilon\right) = \frac{n_{s,l,t}!}{\prod_{b_i} n_{b_i,s,l,t}!} \prod_{b_i} \left(\pi_{b_i,s,l,t} + \left(1 - 4\,\pi_{b_i,s,l,t}\right)\epsilon\right)^{n_{b_i,s,l,t}} \tag{2}$$

We do not know the values of ϵ, and cannot assume that these probabilities are the same for all strands, bases, and time points or bases. Rather, we describe a distribution of probabilities of the bases where Equation (2) is satisfied on average. We do this by constructing a Dirichlet distribution $\mathrm{Dir}_{\{\alpha_{b,s,l,t}\}}(\{P_{s,l,t}(b_i)\})$ where $\alpha_{b_i,s,l,t} = \alpha_0\,\pi_{b_i,l,t} + \alpha_\epsilon$. For such a Dirichlet distribution, the average probability of observing a given base $\langle P_{s,l,t}(b_i)\rangle$ is given by

$$\langle P_{s,l,t}(b_i)\rangle = \frac{\alpha_0\,\pi_{b_i,l,t} + \alpha_\epsilon}{\alpha_0 + 4\alpha_\epsilon} = \pi_{b_i,l,t} + \left(1 - 4\,\pi_{b_i,l,t}\right)\left(\frac{\alpha_\epsilon}{\alpha_0 + 4\alpha_\epsilon}\right) \tag{3}$$

which mirrors Equation (1) when $\epsilon = \frac{\alpha_\epsilon}{\alpha_0 + 4\alpha_\epsilon}$. The advantage of this approach is that we are not specifying an error rate, but allowing the error rate to vary by an amount determined by the parameters in the Dirichlet distribution, α_0 and α_ϵ. We are assuming that the distribution is symmetric with respect to the various bases, but the rates for specific errors need not be the same. We now have a distribution of $\{P_{s,l,t}(b_i)\}$ rather than specific values, so in order to calculate the likelihood of the observed data on that strand at that position and time, we need to integrate over this distribution

$$\Lambda\left(\{n_{b,s,l,t}\}_{s,l,t}|\{\overrightarrow{x_j}\}_l,\{\Pi_{j,t}\},\alpha_0,\alpha_\epsilon\right)$$
$$=\frac{n_{s,l,t}!}{\prod_{b_i} n_{b_i,s,l,t}!}\int\prod_{b_i}\left(P_{s,l,t}(b_i)\right)^{n_{b_i,s,l,t}}\mathrm{Dir}_{\{\alpha_{b,s,l,t}\}}(\{P_{s,l,t}(b_i)\})\ d(\{P_{s,l,t}(b_i)\}) \tag{4}$$

where the integral is over the entire range of the Dirichlet distribution, which is all values of $\{P_{s,l,t}(b_i)\}$ such that $\sum_i P_{s,l,t}(b_i)=1$, and we have explicitly considered that the calculations of $\pi_{b_i,l,t}$, and thus $P_{s,l,t}(b_i)$, rely on $\{\overrightarrow{x_j}\}_l$, the haplotype sequences $\{\overrightarrow{x_j}\}_l$at position l , as well as the haplotype frequencies and the parameters α_0 and α_ϵ. Taking advantage of the properties of the Dirichlet distribution, this integral is quite easy, resulting in

$$\Lambda\left(\{n_{b,s,l,t}\}_{s,l,t}|\{\overrightarrow{x_j}\}_l,\{\Pi_{j,t}\},\alpha_0,\alpha_\epsilon\right)=\frac{n_{s,l,t}!}{\prod_{b_i} n_{b_i,s,l,t}!}\frac{\mathrm{B}(\{\alpha_0\,\pi_{b_i,l,t}+\alpha_\epsilon+n_{b_i,s,l,t}\})}{\mathrm{B}(\{\alpha_0\,\pi_{b_i,l,t}+\alpha_\epsilon\})} \tag{5}$$

where $\mathrm{B}(\{\theta_i\})$ is the multivariate beta function defined as $\mathrm{B}(\{\theta_i\})\equiv\frac{\Gamma(\theta_1)\Gamma(\theta_2)\ldots\Gamma(\theta_n)}{\Gamma(\theta_1+\theta_2+\cdots+\theta_n)}$.

Distributing bases amongst haplotypes: The calculation of $\pi_{b,s,l,t}=\sum_{j\,\forall\,x_{j,l}=b}\Pi_{j,t}$ relies on knowing the haplotype sequences $\{x_{j,1},x_{j,2},x_{j,3}\ldots x_{j,M}\}$. As this information is unavailable *a priori*, we calculate the likelihood for all 4^M possible ways of assigning bases to M haplotypes. The haplotype sequence is assumed to be the same for both strands and all time points; therefore, the sum over possible haplotype sequences is outside the sum over strands and time points, but we can consider each site separately.

$$\Lambda\left(\{n_{b,s,l,t}\}\right)=\prod_l\sum_{\{\overrightarrow{x_j}\}_l}\frac{1}{4^M}\prod_{s,t}\Lambda\left(\{n_{b,s,l,t}\}_{s,l,t}|\{\overrightarrow{x_j}\}_l,\{\Pi_{j,t}\},\alpha_0,\alpha_\epsilon\right) \tag{6}$$

where the sum is over all possible distributions of bases amongst haplotypes at position l.

We first maximise this expression by adjusting the values of $\{\Pi_{j,t}\}$, α_0, and α_ϵ, noting that $\{\Pi_{j,t}\}$ is the same for all locations at each time point. By considering which assignments of bases to haplotypes contribute the most to the likelihood, we are able to calculate posterior probabilities of the arrangement of bases at each position in the sequence, allowing us to calculate the posterior probability of each base at each position on each haplotype.

Evaluating haplotype reconstruction accuracy

In order to evaluate the accuracy of the haplotype reconstruction scheme, we created synthetic data sets consisting of mixtures of two or three full-length CMV database sequences, and evaluated the ability of *HaROLD* as well as other available software to reconstruct these sequences and the relative haplotype frequencies, both for datasets consisting of only one sample and when we had multiple time points representing longitudinal sampling. The various synthetic sets are summarised in Table S1, involving various combinations of KP745665.1, KP745666.1, KP745692.1, KP745694.1, KP745710.1 and KC519321.1 (5).

Sequence 1	Sequence 2	Sequence 3	% Identity	Single or longitudinal	Relative fractions		
					Sample 1	Sample 2	Sample 3
KP745666.1	KC519321.1		96.8%	Single	20/80		
				Single	40/60		
				Longitudinal	0/100	10/90	20/80
				Longitudinal	0/100	20/80	40/60
KP745694.1	KP745710.1		98.6%	Single	20/80		
				Single	40/60		
				Longitudinal	0/100	10/90	20/80
				Longitudinal	0/100	20/80	40/60
KP745665.1	KP745692.1		99.1%	Single	20/80		
				Single	40/60		
				Longitudinal	0/100	10/90	20/80
				Longitudinal	0/100	20/80	40/60
KP745665.1	KP745666.1		99.97%	Single	20/80		
				Single	40/60		
				Longitudinal	0/100	10/90	20/80
				Longitudinal	0/100	20/80	40/60
				Longitudinal	0/100	50/50	100/0
KP745666.1	KC519321.1	KP745694.1	96.8% / 97.2% / 97.1%	Longitudinal	20/50/30	40/40/20	
				Longitudinal	20/50/30	40/40/20	60/30/10
				Longitudinal	80/20/0	60/30/10	40/40/20

Table S1. Summary of the synthetic data sets used to test the accuracy of the haplotype reconstruction methods.

We used *SimSeq* (6) to create ensembles of paired end reads of length 250 corresponding to the database sequences listed in Table S1, including sequencing errors typical of Illumina mySeq sequencing. 10,000-read data sets were then constructed by mixing reads from each ensemble according to the relative fractions listed in Table S1. Reads were trimmed for adapters and low-quality reads and mapped to CMV Merlin (NCBI:txid295027) using CLC Genomics Workbench 10.1.1, requiring 80% of each read to be mapped with 80% identity or higher to the reference sequence. Duplicate reads were removed. BAM files were then input into bam-readcount (https://github.com/genome/bam-readcount) using default parameters to obtain site and strand specific nucleotide counts.

The majority of methods of determining haplotypes contained in a mixed sample consider reads that contain multiple polymorphic sites. Co-localisation of variants at these sites provides evidence that these variants exist on the same haplotype. If the polymorphic sites are sufficiently dense, it is possible to link the variants associated with overlapping reads to generate genome-scale haplotypes (4). These methods require sufficient density of polymorphic sites so that adjacent sites are spaced at distances smaller than the read length. This is, unfortunately, not the case with CMV, especially as much of the observed sequence diversity is confined to short intervals between protein coding regions. As a result, methods such as *Shorah* (7) and *HaploClique* (8) could only reconstruct short segments of haplotypes, inadequate for phylogenetic analysis. *Evohra* (9) uses a combination of overlapping variants as well as variant frequencies, and could potentially be able to resolve haplotypes for samples with less diversity, although it is

not able to incorporate variant frequencies from multiple sampling timepoints. For this reason, we only compare the performance of *HaROLD* and *Evohra*.

The accuracy of the reconstructed haplotypes is shown in Figure S1, for each of the data sets listed in Table S1. When only single samples were available, *HaROLD* generally over-estimated the number of different haplotypes in the sample, providing a number of haplotypes corresponding to each database sequence represented in the sample. In one case (KP745665.1 and KP745692.1, 40/60 mixture) all haplotypes mapped to the same database sequence. This is not surprising, as there is no information available for assigning sequences to haplotypes as the mixture ratio approaches 50/50. Reconstruction accuracy of *HaROLD* for these data sets consistently exceeded 99%. Performance improved considerably when longitudinal samples were available. Under these conditions, *HaROLD* correctly estimated the number of haplotypes, except in the case where the two haplotypes were extremely similar (KP745665.1 and KP745666.1, 99.97% sequence identity). Haplotype reconstruction accuracy was also greatly improved.

Evorha (9), in contrast, generally estimated a larger number of haplotypes than existed in the sample, and consistently yielded haplotypes that most resembled the majority database sequence contained in the dataset. *Evorha* was constructed to model haplotypes of bacteria, and tended to yield haplotypes that closely resembled the reference sequence required by the software. (Best performance was obtained when the consensus sequence of the data set was used as a reference, as was done for the results shown in Figure S1.) It is perhaps not surprising that *Evorha* did not perform well analysing viral sequences, a situation for which it was not designed, and it is possible that *Evorha* could be modified to perform better in these tasks.

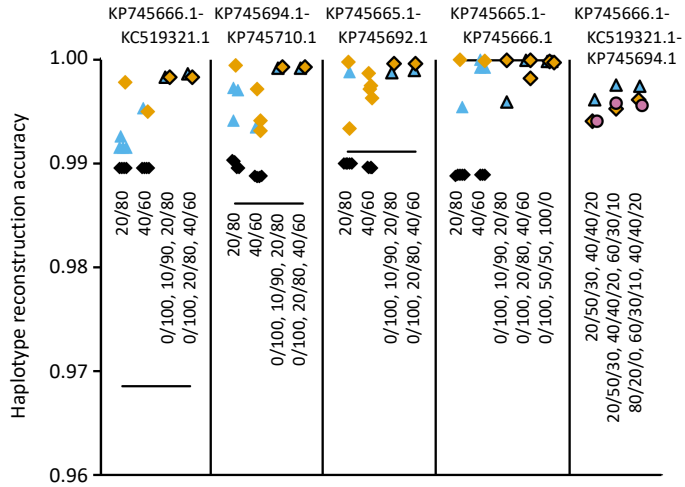

Figure S1. Accuracy of the reconstructed haplotypes. Results for *HaROLD* are shown in cyan triangles (haplotype 1), orange diamonds (haplotype 2), and red circles (haplotype 3). In all cases, comparisons were made between the reconstructed haplotype and the most similar database sequence contained in that data set. Black borders represent results obtained with multiple data sets representing longitudinal sampling. Black lines show the sequence identity for the database sequences used to construct the data sets, for comparison. Performance of *Evohra* is indicated in black diamonds: in all cases, all haplotypes found by *Evorhra* were most similar to the most frequent haplotype, haplotype 2.

The performance of HaROLD in estimating the relative frequencies of the haplotypes in the various data sets is shown in Figure S2. Again, the performance on this reconstruction task is uneven when only a single data set is available, but is highly accurate when analysing longitudinal data, except in the case of nearly identical haplotypes (KP745665.1 and KP745666.1).

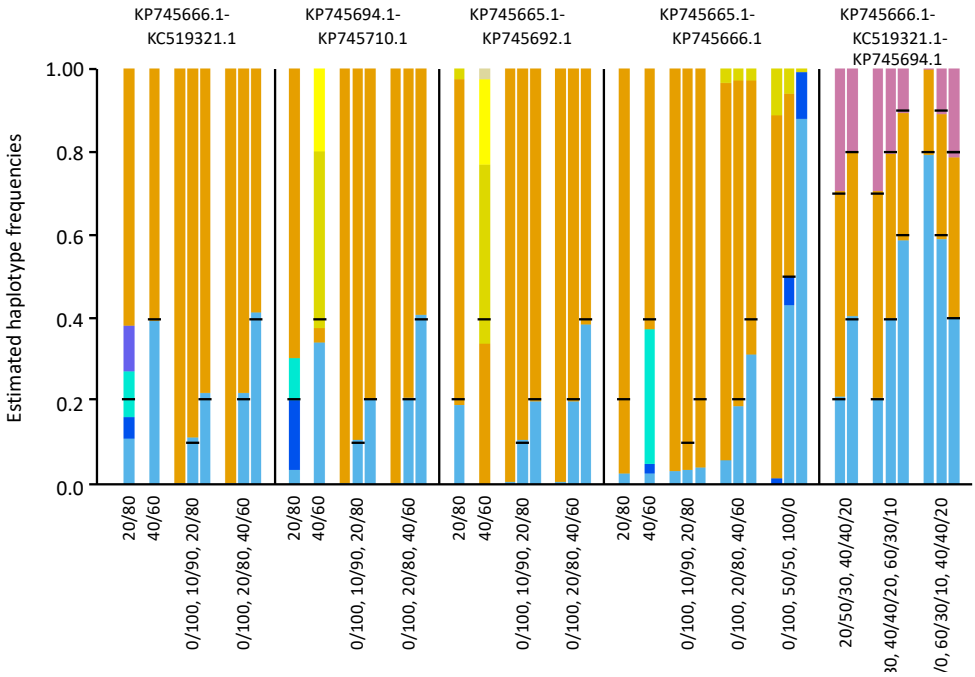

Figure S2 Estimated haplotype frequencies for the synthetic data sets listed in Table S1. Colours are as in Figure S1, except a range of related colours are used to represent the frequencies of different haplotypes that all correspond to the same database sequence: haplotype 1 (cyan and related blues), haplotype 2 (orange and related yellows), haplotype 3 (rose). Horizontal black lines represent the true frequencies of the database proteins represented in the data sets.

1. Moncla LH, Weiler AM, Barry G, Weinfurter JT, Dinis JM, Charlier O, Lauck M, Bailey AL, Wahl-Jensen V, Nelson CW, Johnson JC, Cai Y, Goldberg TL, O'Connor DH, Jahrling PB, Kuhn JH, & Friedrich TC (2017) Within-Host Evolution of Simian Arteriviruses in Crab-Eating Macaques. *J Virol* 91(4).

2. Leung P, Eltahla AA, Lloyd AR, Bull RA, & Luciani F (2017) Understanding the complex evolution of rapidly mutating viruses with deep sequencing: Beyond the analysis of viral diversity. *Virus Res* 239:43-54.

3. Houldcroft CJ, Beale MA, & Breuer J (2017) Clinical and biological insights from viral genome sequencing. *Nat Rev Microbiol* 15(3):183-192.

4. Posada-Cespedes S, Seifert D, & Beerenwinkel N (2017) Recent advances in inferring viral diversity from high-throughput sequencing data. *Virus Res* 239:17-32.

5. Sijmons S, Thys K, Mbong Ngwese M, Van Damme E, Dvorak J, Van Loock M, Li G, Tachezy R, Busson L, Aerssens J, Van Ranst M, & Maes P (2015) High-throughput analysis of human cytomegalovirus genome diversity highlights the widespread occurrence of gene-disrupting mutations and pervasive recombination. *J Virol*.

6. Benidt S & Nettleton D (2015) SimSeq: a nonparametric approach to simulation of RNA-sequence datasets. *Bioinformatics* 31(13):2131-2140.

7. 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.

8. Topfer A, Marschall T, Bull RA, Luciani F, Schonhuth A, & Beerenwinkel N (2014) Viral quasispecies assembly via maximal clique enumeration. *PLoS Comput Biol* 10(3):e1003515.

9. Pulido-Tamayo S, Sanchez-Rodriguez A, Swings T, Van den Bergh B, Dubey A, Steenackers H, Michiels J, Fostier J, & Marchal K (2015) Frequency-based haplotype reconstruction from deep sequencing data of bacterial populations. *Nucleic Acids Res* 43(16):e105.