Within-host recombination in structural proteins of the Foot-and-Mouth Disease Virus

Luca Ferretti[1,*], Eva Pérez-Martín[1], Fuquan Zhang[1], François Maree[2,3], Bryan Charleston[1], Paolo Ribeca[1]

[1]The Pirbright Institute, Ash Road, Woking, Surrey, GU24 0NF, United Kingdom

[2]Transboundary Animal Disease Programme, ARC-Onderstepoort Veterinary Institute, Private Bag X05, Onderstepoort 0110, South Africa

[3]South Africa Department of Microbiology and Plant Pathology, University of Pretoria, Pretoria, South Africa

Abstract

It is well known that recombination between different serotypes generates a mosaic structure in the FMDV genome. Such mosaic structure is clearly visible in the sequence of non-structural proteins and at the genomic boundaries between structural and non-structural proteins. Recombination also occurs among structural proteins, but appears to be strongly suppressed at phylogenetic scales, and only a few events have been observed. Here we show that co-inoculation of closely related strains in buffalos results in extensive within-host recombination within structural proteins during the infection process. For the first time, we are able to build a high-resolution map of effective within-host recombination rates in capsid genes. The effective recombination rate in VP1 during the acute infection phase is about 0.1 per base per year, i.e. comparable to the mutation/substitution rate. We find that the linkage disequilibrium pattern inside VP1 points to a mosaic structure with two main genetic blocks. Epistatic interactions between mutations appear to be present both within and between blocks. Overall our findings show that within the host capsid genes recombine at a high rate during FMDV co-infections.

1 Introduction

Foot-and-mouth disease virus (FMDV) is a picornavirus of the genus *Aphtovirus* that causes one of the most economically relevant diseases of livestock and cloven-hoofed animals. Seven different serotypes - A, O, C, Asia1 and SAT1/2/3 - are known, with a distribution spanning from south-eastern Asia to Africa and South America. SAT serotypes are endemic to Africa, where they circulate mostly among African buffalos (*Syncerus caffer*).

The FMDV genome is short (about 8000 nucleotides) and encodes 12 proteins: a leader protease (Lpro), four capsid proteins (1A–1D or VP4, VP2, VP3, VP1) and seven non-structural proteins (2A–2C, 3A–3D). As for most RNA viruses [2], the mutation rate is high due among other things to the lack of proof-reading capabilities of the polymerase, which contributes to the substantial genetic and antigenic variability of the virus.

Recombination is an important mechanism in the evolution of FMDV genomes [8]. Direct evidences of recombination date back to 40 years [12, 7]. More recently, several studies [5, 6] have

*Email: luca.ferretti@gmail.com

found phylogenetic evidences of extensive recombination among non-structural proteins and a small number of recombination events within capsid genes.

While recombination events within capsid genes have been described before [17, 18], they appear to be much rarer than recombination events in non-structural proteins. However, recombination inferred from phylogenetic incompatibilities suffers from a strong detection bias. In fact, only events that occur between sufficiently divergent lineages can be detected, and only events that do not disrupt positive epistatic interactions among variants (i.e. events preserving correlated sets of genomic variants that taken together confer an evolutionary advantage to the virus) can generate viral sequences that are fit enough to be observed in samples [6]. In addition, since the capsid region is the primary target of the immune response, cross-immunity of viruses with similar capsid sequences could reduce co-infections and therefore recombination.

Within-host studies offer the opportunity to observe recombination in action without any of these biases. In this respect, one of the best systems to study are arguably infections in African buffalos, since animals of this species are FMDV carriers: after an initial acute phase of the infection, the virus can persist for years in some tissues, albeit at lower levels of replication [11]. In principle this increases the chances to see recombination events. Moreover, the SAT serotypes of the virus are well-adapted to this host.

In a recent study, we obtained viral sequences generated with both Sanger and high-throughput sequencing technologies from an experiment on African buffalos infected by FMDV [11]. An interesting feature of this experiment is the subsequent discovery of a strong viral structure (Cortey et al, in preparation). It turns out that both the inoculum and the animal samples contain two major viral swarms with moderate sequence divergence between them. Our results (Cortey et al, in preparation) show that recombinants of these swarms were already present in the inoculum - probably due to previous recombination in culture or in buffalo - and the amount of recombination increased both after the acute phase of the infection and during the persistent phase. Thus this experimental system provides an excellent setup to infer the relative and absolute rates of within-host recombination.

In this paper, we present a detailed analysis of the genomic patterns of recombination in the capsid region. In the first section, we provide a brief explanation of the experimental setup and how it allows to detect within-host recombination. In the second section we present some estimates of the absolute recombination rates during the infection process. In the third section, we infer the recombination profile inside the capsid region. In the fourth section, we show how the linkage disequilibrium (LD) patterns of VP1 sequences suggests a mosaic structure with two main blocks, with epistatic interactions both within and between blocks. In the last section, we discuss some of the possible explanations behind the mismatch between our results and the phylogenetic (lack of) evidence of capsid recombination.

2 Results

2.1 Co-infection and recombination of viral swarms

In our experimental setup, several buffalos were infected with buffalo-derived FMDV. Details of the experiment are discussed in [11]. The inoculum contained several serotypes (SAT1, SAT2, SAT3), but only SAT1 was found in infected buffalos one year after infection, hence we focus on this serotype only.

Intra-host FMDV sequences are expected to consist of either clonal sequences or a single quasispecies or viral swarm, i.e. a cloud of similar genotypes differing only by a handful of mutations. This is a typical pattern of genetic variability in organisms with high mutation rates, such as RNA viruses [2].

Interestingly, it turns out that SAT1 viral sequences in the inoculum exhibit a strong multi-swarm structure, with most sequences belonging to one of two major swarms (Cortey et al, in preparation). The VP1 sequences of the two swarms differ by about 3%, much larger than the genetic diversity within each swarm. Hence, the two swarms are clearly genetically distinct and separable. A similar multi-swarm structure is found in viral sequences from micro-dissections of several tissues in buffalos, proving that co-infection occurred in this experiment. A detailed analysis of these swarms and their evolution post inoculation will be presented elsewhere (Cortey et al, in preparation).

Co-infection of these buffalo hosts by different swarms offers an opportunity to observe within-host recombination. In fact, recombination is assumed to occur whenever two viruses co-infect the same cell, but it can only be detected when their sequences are different enough to be clearly separated. This is clearly the case for our experiment. Indeed we observe a large number of recombinants among the Sanger sequences of clones derived from the buffalo tissue micro-dissections. Extensive recombination between sequences belonging to the two initial swarms was also detected from high-throughput sequencing of the inoculum. These observations cannot be explained by sequencing errors, chimeric artefacts or repeated mutations (see Methods).

The amount of recombination in these sequences is surprisingly large: there is at least a recombination event between almost all pairs of SNPs considered. This allows us to apply to this system some classical population genetics approaches which were originally developed for eukaryotic sequences.

In classical population genetics, recombination is inferred from *linkage disequilibrium* (LD), a measure of the correspondence between the genotypes of two closely occurring SNPs [4]. In the absence of recombination (and of recurrent mutations), the physical linkage between alleles along the sequence constrains the possible allelic combinations. As an example, for two SNPs originated by a mutation ...A...G...$\rightarrow$...T...G... in the first site followed by a ...A...G...$\rightarrow$...A...C... mutation in the second site, the only possible allelic combinations without recombination are $\{$A,G$\}$, $\{$T,G$\}$ and $\{$A,C$\}$, i.e. C in the second site would always be found with A in the first, and T in the first site would always be found with G in the second; in addition, $\{$A,C$\}$ would tend to appear at lower frequencies. The effect of recombination events between the two SNPs is to reshuffle these allelic combinations; in our example, recombination occurring between the two SNPs would generate sequences with a $\{$T,C$\}$ genotype and increase the frequency of the $\{$A,C$\}$ genotype. Linkage disequilibrium quantifies the observed extent of reshuffling between genotypes.

We note that linkage disequilibrium is also affected by *epistasis*, i.e. interactions in fitness between genetic mutations – if different combinations of alleles at multiple loci have different fitness, the frequency of favoured combinations of alleles increases and the correspondence between alleles corresponding to these combinations is reinforced. Hence these epistatic interactions often act in opposition to recombination and cause an effective increase in LD [3].

2.2 Absolute recombination rates

Linkage disequilibrium (LD) and recombination rates were inferred for the whole capsid region of the inoculum and for the VP1 sequence of the virus from three buffalos: two sampled at 35 days post inoculation (dpi) and one sampled at 400 dpi. All recombination rates inferred from our sequence data are relative to the time of origin of the swarm structure, which is unknown; hence their absolute values do not have any easy interpretation. However, their differences provide absolute recombination rates per unit time across the acute and persistent phases of the infection. We can estimate these rates only for VP1, since it is the only genomic region for which multiple time-points are available.

Recombination rates were estimated using LD between pairs of variants consistent with the two main sworms of the inoculum. Two approaches were used for inference of recombination rates: the

"local" approach uses only information from consecutive variants, while the "global" approach uses information from all variants. The "local" approach is therefore more noisy, while the "global" one is more precise but could be more sensitive to biases. The two methods are also affected by epistatic interactions, but at different scales.

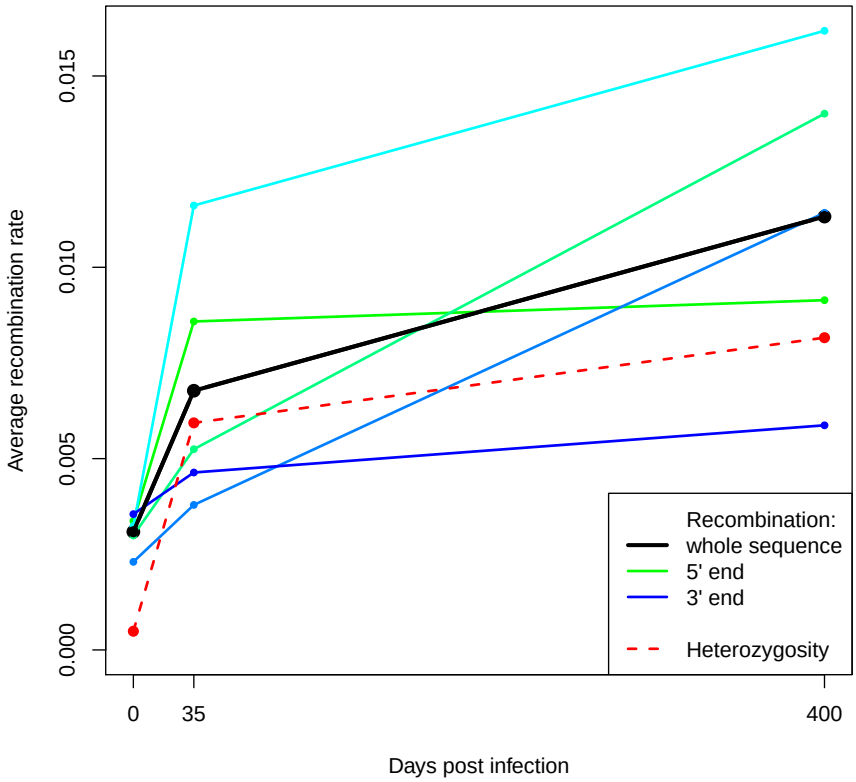


Figure 1: Recombination rate in VP1 sequences from the inoculum and from animals sampled at different time points, inferred from LD. All the rates are defined from the beginning of the experiment to the time the sequences were sampled. Colours indicate different non-overlapping parts of the sequence. The dashed red line shows the heterozygosity per base, computed on the variants unrelated to the main swarm structure.

The average recombination rates per base, estimated using the "global" and "local" approaches, are $R_0 \approx 2.6 - 3 \cdot 10^{-3}$, $R_{35} \approx 4 - 7 \cdot 10^{-3}$ and $R_{400} \approx 8 - 11.7 \cdot 10^{-3}$ respectively. Hence, the rate per year in the first 35 days post inoculation is $r_{0-35} \approx 0.015 - 0.04/bp/y$, while for later times the rate is $r_{35-400} \approx 0.004 - 0.005/bp/y$. Hence, during the first month post-infection, the average recombination rate is higher by a factor $r_{0-35}/r_{35-400} \approx 3.8 - 8.7$. Since the acute phase of the infection lasts for about a week [11], the actual rates can be estimated as

$$r_{acute} \approx 0.6 - 1.9 \cdot 10^{-1}/bp/y$$

$$r_{persistent} \approx 4 - 5 \cdot 10^{-3}/bp/y$$

i.e. the recombination rate during the acute phase is $15 - 40$ times faster than during the persistent phase.

Note that these rates are quite high. They are comparable to the typical mutation rates for FMDV, which are as high as $10^{-2}/bp/y$ due to the error-prone nature of the RNA polymerase.

From the recombination map of the inoculum, it is also possible to obtain an estimate of the relative recombination rates of the other capsid proteins 1A-1C (VP4,2,3) and some non-structural proteins (2A-2B and a fragment of 50 bases at the 3'-end of Lpro) with respect to VP1. The relative rates are

Region	R/R_{VP1}, global	R/R_{VP1}, local
Lpro	6.2	8.7
2A	1.12	2.8
2B	1.04	1.17
2A-B	1.05	1.4
1A	0.93	0.87
1B	1.07	1.27
1C	1.26	1.15
1A-C	1.12	1.16

Note that the rates in the flanking regions of the capsid (Lpro and 2A) are higher than in capsid proteins. However, the rates in 2A-B is of the same order of magnitude as the rate inside the capsid, and the rate in 2B is actually not too dissimilar from the rate in the capsid. Previous results based on phylogenetic evidence inferred high levels of recombination for non-structural proteins [5, 6], but they observed much less recombination in VP1 and in the capsid, in partial contrast to our results above. We will discuss possible explanations in the final section.

2.3 Recombination profile in capsid genes

The previous analysis was based on the normalised linkage disequilibrium D' between pairs of derived swarm-specific variants. The measure D' is defined in the Methods and it takes values $+1$ or -1 in the absence of recombination, while it is close to 0 for strong recombination. Figure 2 presents the LD values for pairs of variants in the region between Lpro and 2B, estimated from the reads sequenced from the inoculum.

The "global" and "local" approaches discussed in the previous section can then be used to build a recombination profile. This recombination map extends almost to the whole sequenced region and the distance between swarm-specific variants (~ 30 bp) determines the resolution of the profile. The final profile is shown in Figure 3.

Recombination rates in the capsid peak strongly at the 3' end of Lpro/5' end of 1A. Aside from that, they show a moderate heterogeneity both between and within genes, with peaks around the middle of 1C (VP3) and 2A-B (VP4-2).

2.4 Mosaic structure and epistasis in VP1

Thanks to the experimental design of our data (described in [11]) recombination profiles for 1D/VP1 can be reconstructed from different individuals and timepoints: from two animals sampled at 35 dpi; an animal sampled at 400 dpi; and the inoculum, as discussed in the previous section. It is therefore interesting to compare the different profiles. The absolute recombination rates inferred from the "local" approach are shown in Figure 4.

On the top of a clear trend of increase of recombination rates with time, which has been discussed in a previous section, we observe an heterogeneity in the recombination rates along the genome, with several peaks found in similar regions across different individuals.

The complete sequences of VP1 at 35 and 400 dpi contain information about the linkage disequilibrium (LD) of all pairs of variants. The corresponding LD maps are shown in Figures 5a-d (lower

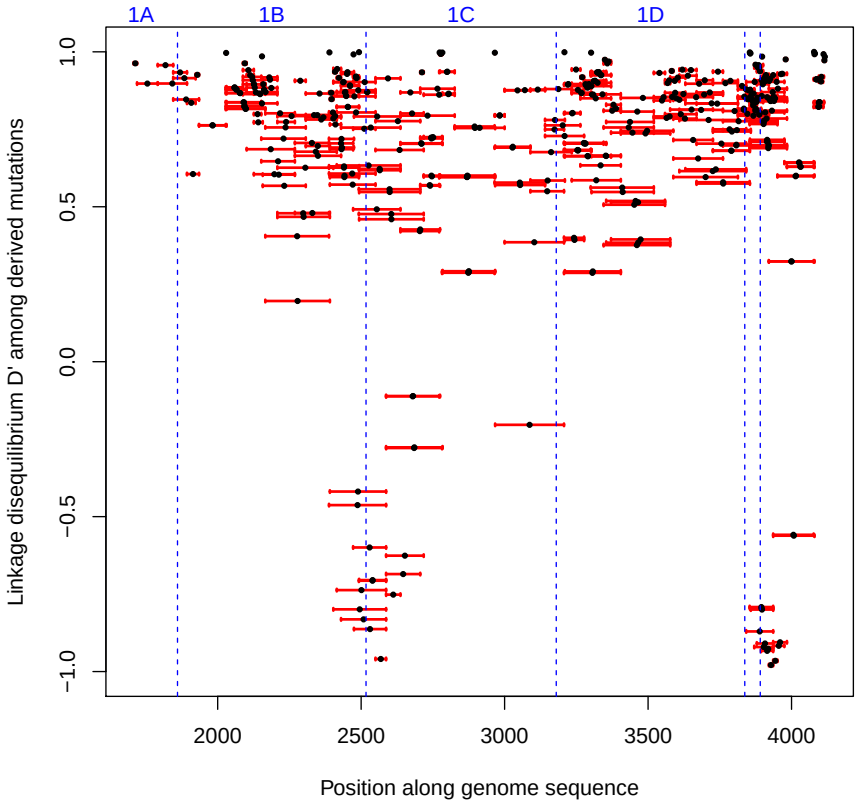


Figure 2: Normalised linkage disequilibrium D' between pairs of derived swarm-specific variants covered by at least 10^4 reads. Red bars illustrate the interval between variants, with a black dot at the mid-point.

triangles). We observe that LD patterns are broadly consistent across different individuals and show two strongly linked blocks, roughly corresponding to the first 200 and last 250 bases of VP1. This suggests that the mosaic structure observed in FMDV genomes [5, 6] extends to VP1: recombination (and possibly epistasis) maintain a modular structure with at least two different genomic blocks inside VP1.

LD between variants in the same protein can be also affected by epistatic interactions. If the original combinations of alleles are fitter than the recombinants, the recombination rate is effectively reduced by selection [4, 13]. The effects of recombination rate and epistasis cannot be separated for pairs of consecutive variants, since this is already the scale of the finest resolution of LD, and the only information at this scale is a single measure of D'.

However, for distant pairs of variants, it is possible to detect footprints of epistatic interactions from excess of LD with respect to the naive value predicted by the "local" approach to recombination rates. We can use the effective suppression of recombination $\Delta R_e = -\log|D'/D'_{predicted}|$ as a measure of the epistatic effects.

As expected, intra-protein epistatic interactions shape the LD structure of VP1. In fact, Figures 5a-d (upper triangles) show signatures of epistasis inside both genomic blocks in VP1. Interestingly, the strongest signatures of epistasis are found between the two blocks. This suggests that even if recombination tends to decouple the two blocks, linkage equilibrium is prevented by epistatic interactions between the blocks. Overall, these results imply that epistatic interactions are widespread

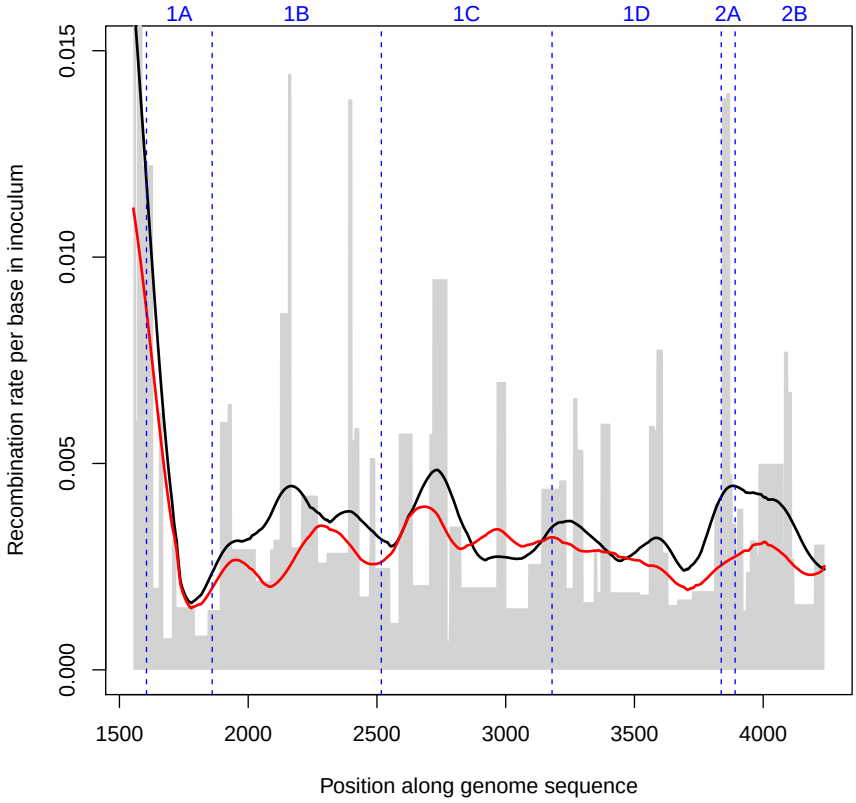


Figure 3: Effective recombination rate per base for the inoculum (in grey), inferred by the "local" approach from pairs of SNPs covered by at least 10^3 reads. The lines indicates the average local rates (black) and global rates (red) Gaussian-smoothed over a 150 bp window.

inside VP1. These interactions reduce the recombination observed in our experiment by up to an order of magnitude.

2.5 Why is recombination suppressed in structural proteins?

Recombination can also be inferred from sequences collected from different animals and locations. In fact, in the presence of recombination different regions of the genome can have different genealogical trees. This signal can be detected in phylogenetic analyses. Such analyses were performed in the past to infer FMDV recombination from phylogenetic incompatibilities [5, 6].

There is a clear mismatch between our results on within-host recombination rates and the much lower recombination rates inferred from previous phylogenetic analyses. As an example, the number of recombination events in the capsid region inferred in [5] for the whole FMDV phylogeny is similar to the number of intra-host events that we observe after one year of persistent infection. Furthermore, while our findings imply that structural proteins recombine less than non-structural ones located in flanking regions in the genome, this difference in recombination rates appears to be much less extreme than the one observed on a phylogenetic scale.

However, previous phylogenetic studies focused on recombination between highly divergent sequences. In fact, only inter-serotype recombination was studied in [6], while in [5], all recombination events were considered, but 70% of the recombination events found were between different

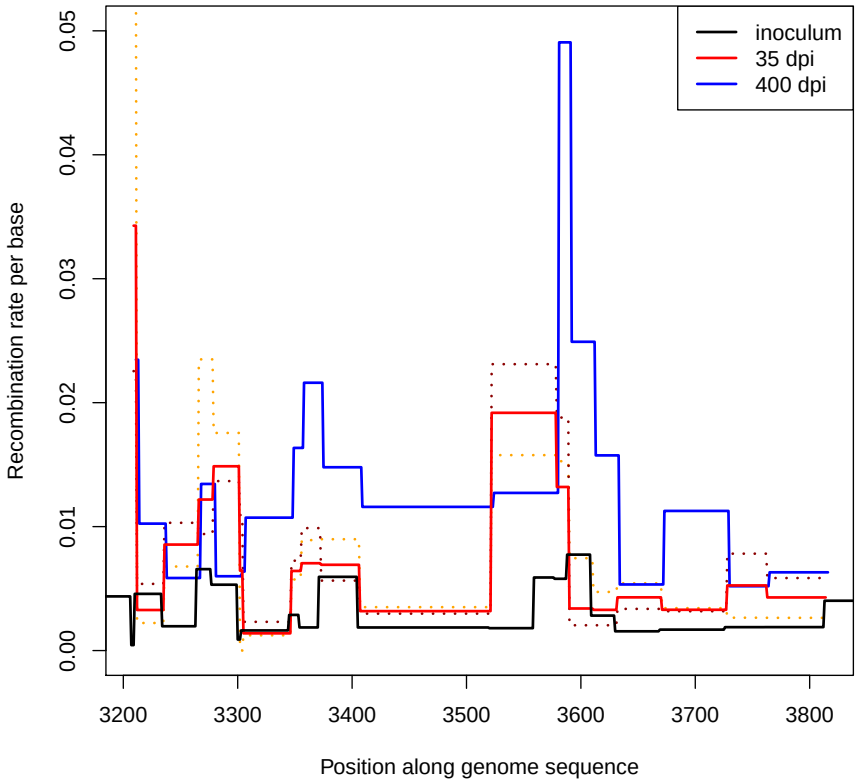


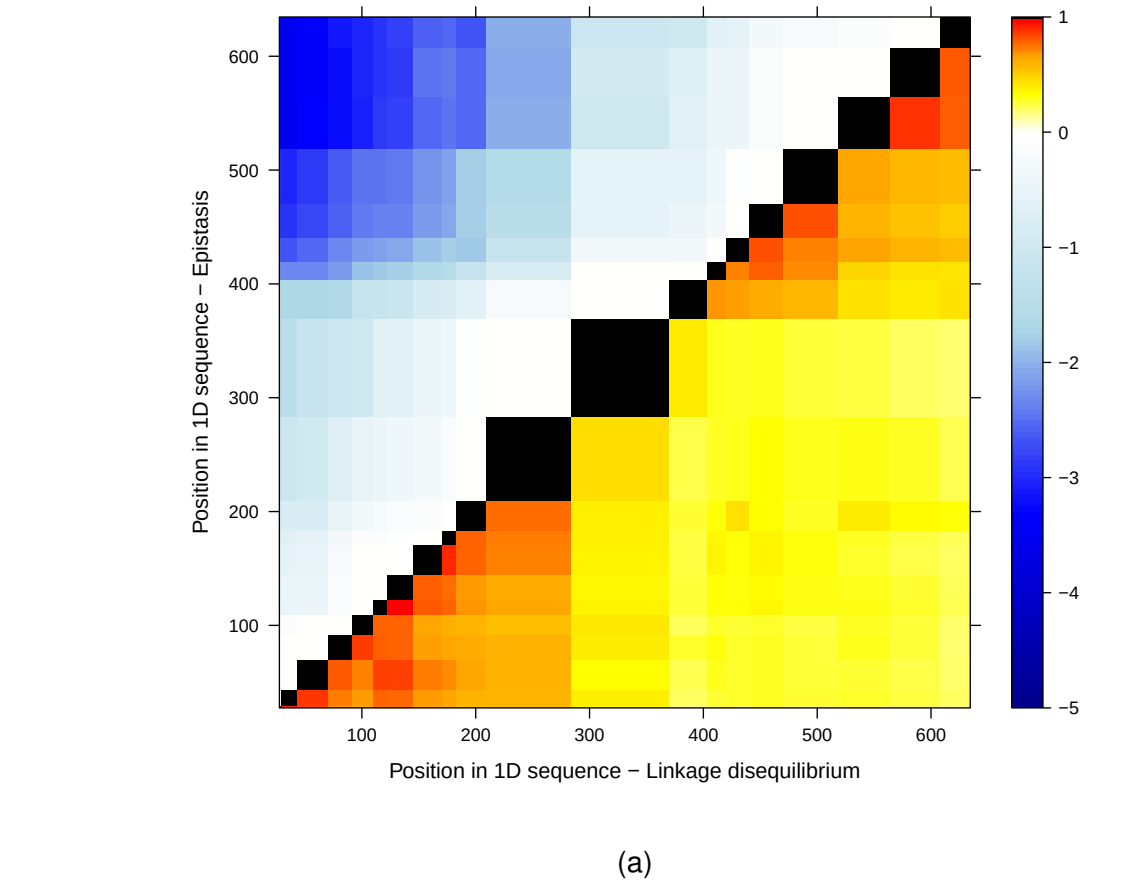
Figure 4: Effective recombination rate per base along the sequence of VP1, measured from the beginning of the experiment to sampling times: 0 days post inoculation (dpi) i.e. inoculum; 35 dpi; and 400 dpi. To illustrate the heterogeneity in inferred recombination rates between individuals, two separate dashed lines drawn in different colours are shown for the two individuals sampled at 35 dpi.

serotypes. In contrast, the two main swarms studied here are very similar and belong to the same topotype. This suggests that divergence-dependent effects that suppress recombination could be responsible for the mismatch.

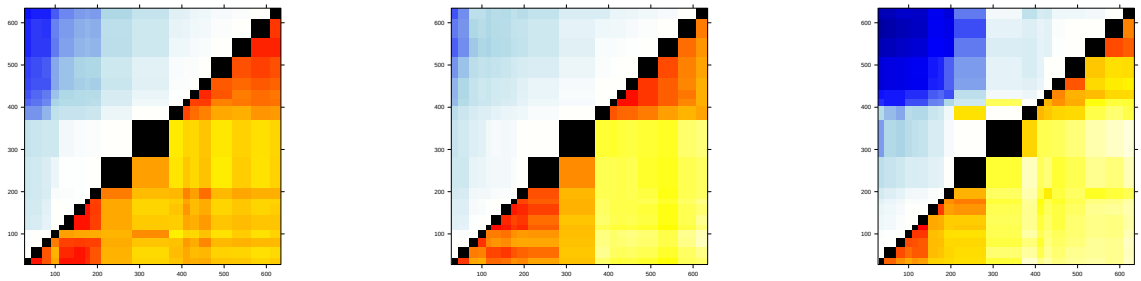
There are several factors that can suppress FMDV recombination in endemic and epidemic contexts, and some of these factors could have a stronger impact on structural proteins. We discuss these factors, ordered by increasing effect on structural versus non-structural proteins.

Biased detection: inference of recombination by phylogenetic methods depends on the resolution of the tree and the similarity of recombining sequences. Recombination between similar sequences is difficult to infer, since the trees generated by these events are very close to each other. In particular, recombination between very close sequences in the phylogenetic tree is almost undetectable. This affects structural and non-structural proteins in a similar way, since it depends only on the local molecular clock.

Cross-immunity: the effective rate of recombination is proportional to the rate of co-infections, since co-infection of the same animal/cell by two different strains is a necessary condition for recombination to occur via template switching. The probability of co-infections depend on the ability of the second strain to escape the immune response induced by the first strain, i.e. on the cross-immunity between strains. Cross-immunity depends only on capsid proteins (which are exposed to



(a)



(b) (c) (d)

Figure 5: LD and epistasis across (a) all individuals sampled at 35 and 400 dpi, (b,c) the two individuals sampled at 35 dpi and (d) the individual sampled at 400 dpi. Lower triangles: map of pairwise LD between variants in VP1, estimated as $|D'|$ (stronger LD in red). Upper triangles: signatures of epistasis in terms of effective suppression of recombination ΔR_e (stronger epistasis in blue).

the immune system) and it decreases with increasing divergence between the capsid sequences of the two strains, hence suppressing recombination between closely related strains only.

Co-circulation of lineages: cross-immunity affects also the spatial co-existence of different FMDV lineages. In fact, because of competition for hosts, cross-immunity could reduce the spread of closely related lineages in the same area, hence reducing the probability of co-infection. However, spatial patterns of FMDV lineages are complex and depend on the endemic/epidemic system considered, so the importance of this effect is difficult to estimate.

Epistasis: as discussed in the previous section, the selective pressure due to epistatic interac-

tions between positively selected alleles can reduce the effective rate of recombination. Recombination between different strains could easily generate combinations of alleles with suboptimal fitness for within-host growth or inter-host transmission; that would then reduce the probability of observing these recombinants in other hosts. This effect increases with the amount and strength of epistatic interactions. While these interactions in fitness for infectivity and inter-host transmissibility would be present among all FMDV proteins, they are expected to be stronger in the capsid due to the amount of structural interactions between capsid proteins and the interplay between the opposite selective pressures of protein stability and immune escape.

In order to better understand the shape of these constraints, we built a simple model for the suppression of recombination. Following an approach used for influenza [9], we model cross-immunity as an exponentially decreasing function of the divergence d between strains. Assuming that cross-immunity acts between lineages with divergence less than d_{ci}, the reduction in recombination corresponds to the reduction in co-infection $1 - e^{-d/d_{ci}}$. Epistasis is modelled after [14] as a decrease in viral growth rate proportional to the number of pairwise interactions disrupted by recombination, i.e. to the square of the divergence, leading to a recombination suppression factor $e^{-s_e^2 d^2}$ where the epistatic coefficient s_e is related to the strength and number of epistatic interactions. Assuming that the observed epistasis between the two blocks of VP1 (which causes a sizeable reduction in recombination) is comparable to the one among different capsid proteins, we can estimate a coefficient $s_e \sim 50$.

Results from this model are shown in Figure 6. For all reasonable values of cross-immunity and epistatic parameters, the model predicts a strong suppression of the recombination rate in structural proteins. The suppression can be of several orders of magnitude, which would explain the near absence of recombination events on a phylogenetic scale.

3 Discussion

In this paper we present the first inference of within-host recombination rates for structural proteins of FMDV. The recombination rates during the acute and persistent phases of the infection are about $r_{acute} \sim 0.2/bp/y$ and $r_{persistent} \sim 0.005/bp/y$. We also provide high-resolution maps of recombination at the scale of the capsid and flanking proteins, showing that recombination is a pervasive phenomenon in the FMDV genome.

These results have been inferred for infections of a single serotype (SAT1) in a single host (buffalo). The details of the recombination map might differ between serotypes; the absolute rates and the amount of epistasis are expected to depend on the within-host infection dynamics and the immune response of the host, hence they could vary when considering infections in cattle or other species. However, we do not expect the qualitative picture for other host species and serotypes to be essentially different.

The inferred recombination rates are related to the amount of co-infections of the same cells in the host, hence they depend intrinsically on within-host dynamics. This is unavoidable in all studies of viral recombination in vivo. In fact, it is better to consider our results as “effective recombination rates” which already take already into account the effect of within-host evolution and infection dynamics.

Recombination rates depend on epistasis as well. This is also unavoidable when considering high-resolution maps such as ours, since we have no way to account for the effect of local epistatic interactions. However, we were able to detect epistasis on the scale of the VP1 sequence and discover that epistatic interactions are widespread. This also means that “local” estimates, which are affected by local epistasis only, should be more reliable than “global” estimates, which are likely to systematically underestimate the real rates due to the additional effect of longer-range epistasis.

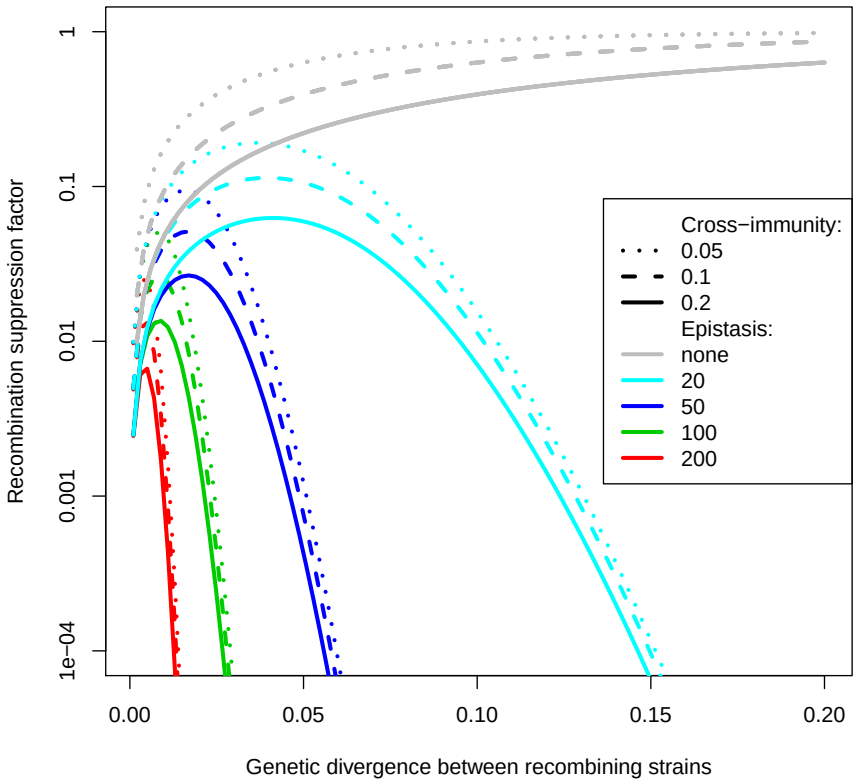


Figure 6: Predicted reduction in recombination in structural proteins, due to cross-immunity and epistatic interactions, shown as a function of the genetic divergence (Hamming distance per base) between recombining sequences.

In fact, such underestimation appears clearly in our results.

The high recombination rates in structural proteins between genetically close lineages represent an important finding of this work, with potentially relevant implications. In fact such recombination events can possibly generate new genetic diversity in capsid proteins, provided that they are not suppressed by lineage competition or other effects. Further studies are needed to understand which phenomena suppress recombination on epidemiological scales, and if their effects extend to genetically close viruses.

Finally, it might be possible to study recombination in a more controlled setup. In vitro and in vivo co-infection experiments guided by our findings represent a promising avenue to a more systematic inference of recombination rates.

4 Material and methods

4.1 Sequencing

The samples from micro-dissections of infected buffalos (pharyngeal/palatine tonsils and dorsal soft palate) were sequenced by Sanger technology. RNA was extracted from LMD material using RNeasy Micro kit (Qiagen), followed by cDNA synthesis using TaqMan RT reagents (Agilent) and random hexamers, then the VP1 region of SAT1 was amplified using Platinum Taq Hi-Fidelity (Invitro-

gen) and the following primer pair: 5'-AGTGCTGGACCCGACTTCGA-3' and 5'-TGTAGCGATCCTTGCCACC 3' and the VP1 fragment was cloned into a TOPO®TA vector (Life Technologies). After colony picking and plasmid purification, the fragments were Sanger sequenced using BigDye terminator v3.1 (Applied Biosystems) and M13 primers. More details on these sequences will be provided in a separate publication (Cortey et al, in preparation).

The inoculum used for the experiment and a further sample from the tonsil swab of one animal at 400 days post inoculation were sequenced at high throughput. RNA was extracted using RNeasy mini Kit (Qiagen), followed by cDNA synthesis using SuperScript^{TM} III First-Strand Synthesis System (Life Technologies), amplification of the capsid region using Platinum Taq Hi-Fidelity (Invitrogen) and the primer pair 1A1F/2B2R (sequences available on request). Libraries were constructed using Nextera XT DNA Sample Preparation Kit (Illumina) and deep sequenced on a MiSeq system using 300 cycle version 2 reagent cartridges (Illumina) to produce paired end reads of approximately 150 bp each.

A reference sequence for the inoculum was assembled using a sensitive in-house pipeline (Bolt, Ribeca et al, in preparation) based on SPAdes [1]. Reads were aligned to this sequence using the GEM mapper [10] version 3. More than 99.9% of the reads mapped to the assembly, with a mean read depth of about 30000.

4.2 Multi-swarm structure

SNP variants in the inoculum were called by a in-house pipeline using an approximation of the snape-pooled algorithm [15] and selecting only SNPs with $p < 0.05$. The sequence of SAT1/KNP/196/91 was used to infer the ancestral allele for each SNP. We considered biallelic variants only. The derived frequency distribution in the inoculum is clearly bimodal with a gap between 0.15 and 0.20 (Figure 11A). That makes easy to separate all SNPs in two classes: common nucleotide variants ($0.20 < f < 0.55$) and the low-frequency variants ($f < 0.15$).

We estimated the linkage disequilibrium (LD) by the normalised measure D' among all pairs of common variants covered by at least 10^4 reads. This measure is defined as $D' = D/D_{max}$ if $D > 0$ and $D' = D/|D_{min}|$ if $D < 0$, where D is the classical linkage disequilibrium $D = f(A_1A_2) - f(A_1)f(A_2)$ for two SNPs with ancestral alleles A_1 and A_2, while D_{max} and D_{min} are its maximum and minimum possible value given the frequencies of the variants [4].

The local haplotype structure of the multi-swarm, with two haplotypes containing ancestral and derived SNP alleles, is clearly illustrated by the concentration of allele frequencies around a value of 0.4 (Figure 11B) and by the high LD between consecutive common variants. In fact, almost all values of $|D'|$ are between 0.75 and 1 (Figure 11C). Note that a few mutations (in red in Figure 11B) have $D' \approx -1$, suggesting an erroneous inference of their ancestral state.

4.3 Evidence of within-host recombination

The main evidence of recombination in the capsid region comes from the LD data from the inoculum in Figure 2. There are clearly many pairs of mutations with low linkage disequilibrium ($-1 \ll D' \ll 1$), which is a characteristic signature of recombination. Low values of LD can be due to recombination or other factors:

- sequencing errors;
- chimeric reads or similar artefacts of sequencing protocols;
- multiple mutations/backmutations in mutation hotspots.

However, none of these other factors can explain the patterns in our data:

(i) Sequencing errors cannot explain the presence of two alleles per site. Unless the error rates are extremely skewed towards the same pair of alleles that is already present in a site (which is highly improbable), if LD would be caused by sequencing errors, they should contribute a number of other variants with frequencies similar to the minimum frequency among the four possible pairs of alleles. Instead, Figure 7 shows the negligible contribution of sequencing errors compared to the one predicted if they would explain our data.

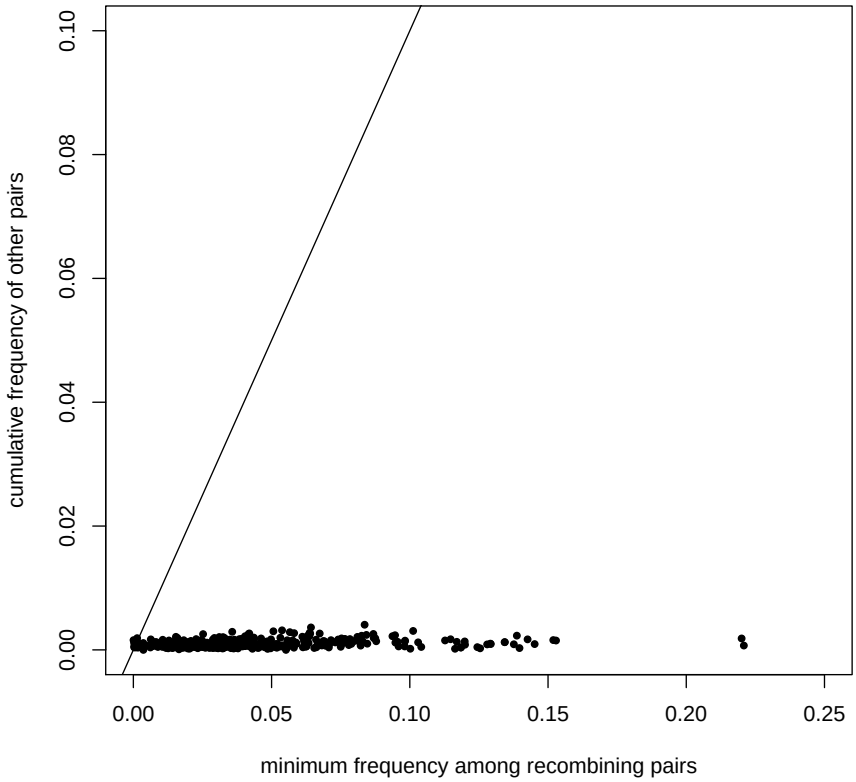


Figure 7: Frequency of the least frequent pair of recombinant alleles $f_{r,min}$ versus the cumulative frequency f_o of all other pairs. For example, if the two main alleles are (C,T) in the first site and (A,G) in the second, then the recombinants are CA, TG, CG, TA; if $f_{CA} = 0.5$, $f_{TG} = 0.2$, $f_{CG} = 0.1$, $f_{TA} = 0.15$, $f_{CC} = 0.03$, $f_{TC} = 0.02$, then $f_{r,min} = 0.1$ and $f_o = 0.05$. Data points are shown for all pairs of polymorphic alleles of frequency > 0.25 covered by at least 10^4 reads. The black line corresponds to the case of equal frequency.

(ii) A similar argument suggests that mutation hotspots represent an unlikely explanation for the data unless all mutations in these hotspots show a very strong bias towards the two alleles observed at these sites (i.e. all mutation are back-and-forth mutations between these two alleles).

(iii) Mutation hotspots and sequencing errors cannot explain a pattern of decay of LD with distance along the genome, since they occur at each site independently. But a clear decay of LD with distance is precisely what is observed in the data (see Figure 8 for the inoculum, and Figure 5 for sequences from buffalo tissues), ruling out these explanations.

(iv) Recombination implies that the mean LD between positions i and j satisfies $D'_{ij} \sim e^{-R_{ij}}$ where $R_{ij} = \sum_{x=i}^{j} R(x)$ in terms of the recombination rate per base $R(x)$. This implies the approxi-

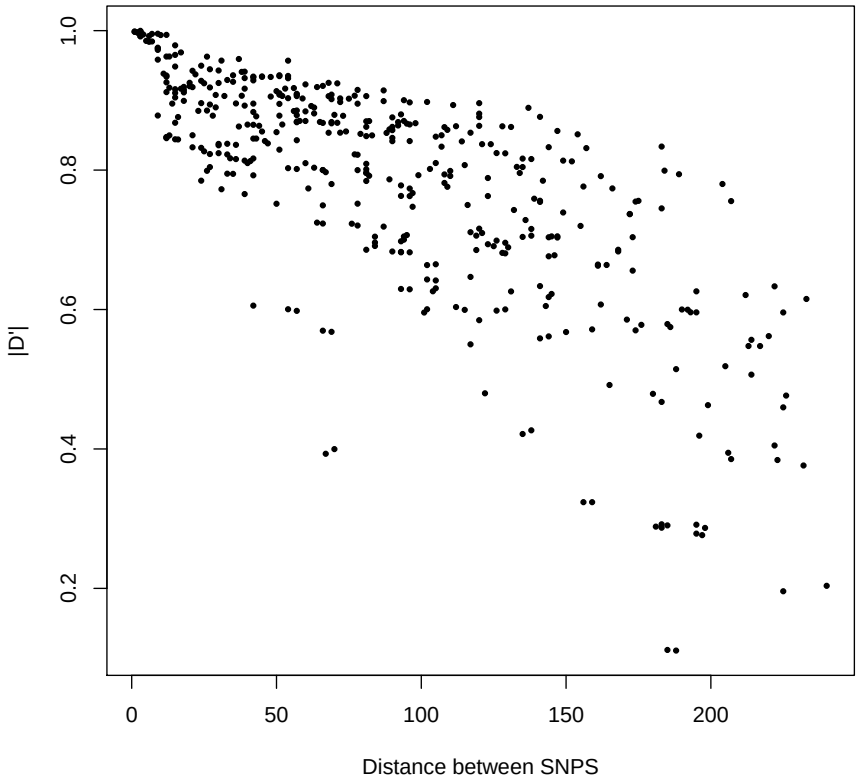


Figure 8: Decay of LD measure $|D'|$ with distance between SNPs. Data points are shown for all pairs of polymorphic alleles of frequency > 0.25 covered by at least 10^4 reads.

mate prediction $D'_{ik} \approx D'_{ij}D'_{jk}$ for $i < j < k$. The results and the predictions for next-to-nearest and next-next-to-nearest SNPs are shown in Figures 9 and 10 respectively. As for the decay of LD, these patterns cannot be replicated by sequencing errors or by back-and-forth mutations.

(v) Chimeric reads and sequencing errors could not cause the decrease in LD (i.e. the increase in overall recombination) over time in Figure 1, since they do not depend on time points but on protocols only.

(vi) Furthermore, recombination was observed both in Sanger- and High-Throughput-sequenced samples. It is extremely unlikely that all these different protocols would generate chimeric reads with similar profiles.

(vii) Finally, the sample from the tonsil swab of one of the animals in the persistent phase shows little internal variability. At the consensus level, its sequence is a complex recombinant of the two initial swarms, with 1A-1B (VP4-2) mostly derived from the major swarm in the inoculum and 1C-1D (VP3-1) from the minor one. This consensus-level evidence rules out chimeric sequences or sequencing errors.

4.4 Linkage disequilibrium and recombination rates

The high levels of positive LD between all the SNPs of the swarms supports a recent origin for the mixture of swarms. Based on this, we can infer the overall rate of recombination since the origin of the swarms using the classical equations for the decay of LD with time: $D = D_0 e^{-r \cdot t}$, where r is the

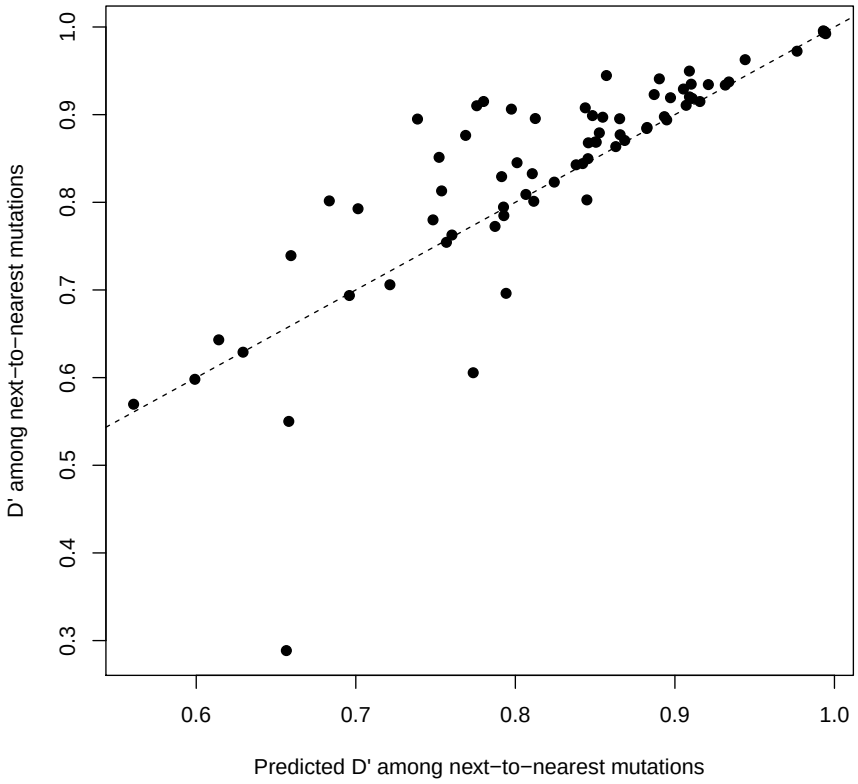


Figure 9: D' among next-to-nearest SNPs versus the predicted value from $D'_{ik} = D'_{ij}D'_{jk}$. The dashed line corresponds to equality between predicted and estimated value.

recombination rate per generation and t is the time in number of viral generations [4]. The overall recombination rate $R = r \cdot t$ for the genomic region between two variants can then be inferred as

$$\hat{R} = -\ln(D'). \tag{1}$$

We apply two different statistical approaches to infer the recombination rate for each variant-free interval. The first ("local" approach) is based on the above estimate (1) for consecutive variants only. The second ("global" approach) is the weighted least squares [16] estimate R^{wls} from all variants, defined by the equations:

$$\sum_j \sum_{I \supset i,j} R_j^{wls} / \mathrm{Var}(\hat{R}_I) = \sum_{I \supset i} \hat{R}_I / \mathrm{Var}(\hat{R}_I) \tag{2}$$

where i, j denote intervals between consecutive variants and I intervals between any pair of variants. We use the approximate form for the variance: $\mathrm{Var}(\hat{R}) = (D')^{-2}/c + \hat{R}$, where c is the number of reads covering both variants in the pair; the first term comes from the delta method applied to the variance of binomial sampling of c sequences (assuming low recombination and similar frequencies for all SNPs), the second from the Poisson noise of the random recombination events. To get comparable results between Sanger and short-reads data, only intervals of length less than 200 bp are used for the "global' estimate for analyses involving both approaches.

Data from Sanger sequencing of viruses from micro-dissections reveals only weak differentiation between tissues from the same animal. The average estimate of $\hat{R}$ across tissues and the joint

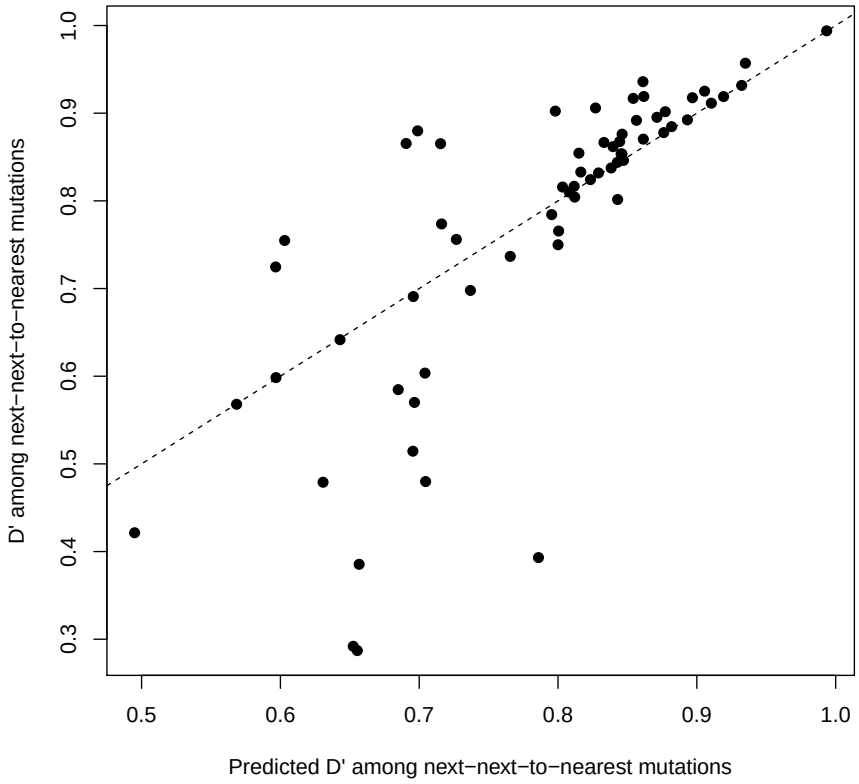


Figure 10: D' among next-next-to-nearest SNPs versus the predicted value from $D'_{ik} = \max_{i<j<k}(D'_{ij}D'_{jk})$. The dashed line corresponds to equality between predicted and estimated value.

estimate from all tissues differ by less than 10%, hence we neglect differences across tissues and compute D' from the pooled set of all sequences from a given animal.

References

[1] Anton Bankevich, Sergey Nurk, Dmitry Antipov, Alexey A Gurevich, Mikhail Dvorkin, Alexander S Kulikov, Valery M Lesin, Sergey I Nikolenko, Son Pham, Andrey D Prjibelski, et al. Spades: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of computational biology*, 19(5):455–477, 2012.

[2] EJJH Domingo and JJ Holland. Rna virus mutations and fitness for survival. *Annual Reviews in Microbiology*, 51(1):151–178, 1997.

[3] Ian Franklin and RC Lewontin. Is the gene the unit of selection? *Genetics*, 65(4):707–734, 1970.

[4] Daniel L Hartl, Andrew G Clark, and Andrew G Clark. *Principles of population genetics*, volume 116. Sinauer associates Sunderland, 1997.

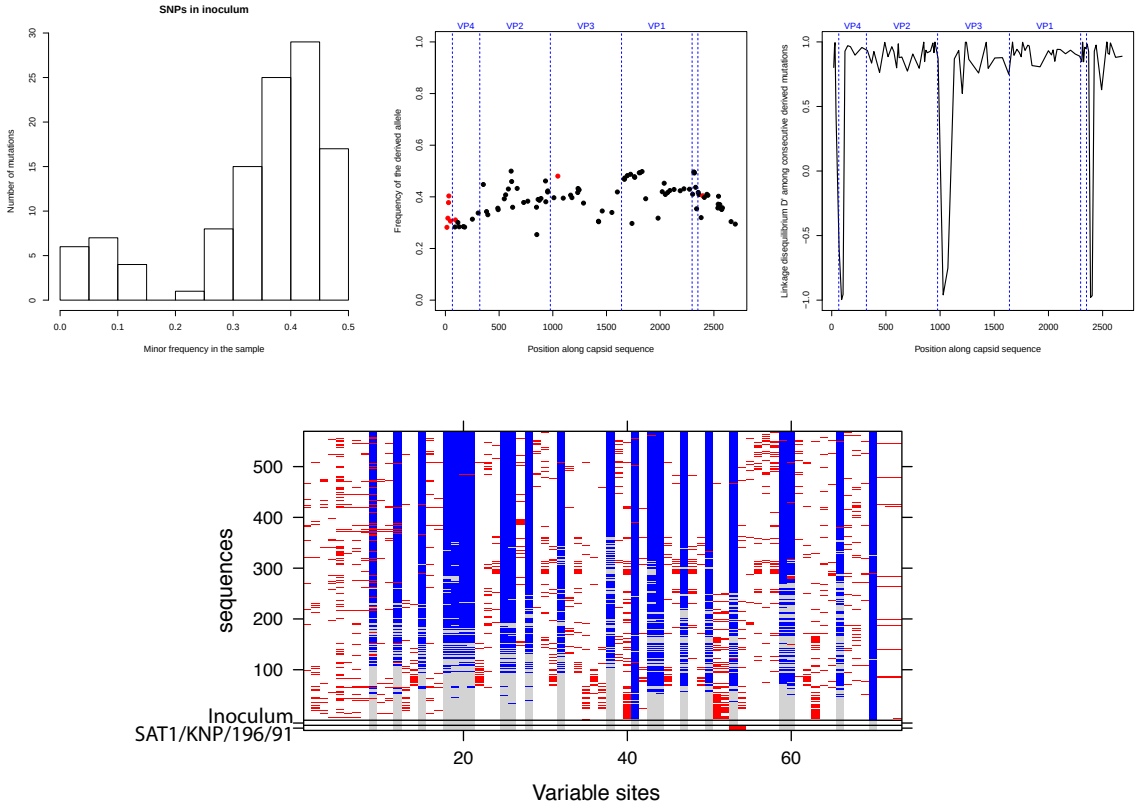


Figure 11: (A) Distribution of minor SNP frequencies in the reads from the inoculum. (B) Location and frequency of SNPs with minor frequency > 0.2. (C) Linkage disequilibrium D' between pairs of consecutive derived variants. (D) Illustration of the content of the sequences from buffalo tissues. (New mutations are in red, mutations characterising the two initial swarms are in grey and blue).

[5] Livio Heath, Eric Van Der Walt, Arvind Varsani, and Darren P Martin. Recombination patterns in aphthoviruses mirror those found in other picornaviruses. *Journal of Virology*, 80(23):11827–11832, 2006.

[6] AL Jackson, H O'neill, F Maree, B Blignaut, C Carrillo, L Rodriguez, and DT Haydon. Mosaic structure of foot-and-mouth disease virus genomes. *Journal of General Virology*, 88(2):487–492, 2007.

[7] AM King, WR Slade, JW Newman, and D McCahon. Temperature-sensitive mutants of foot-and-mouth disease virus with altered structural polypeptides. ii. comparison of recombination and biochemical maps. *Journal of virology*, 34(1):67–72, 1980.

[8] Andrew MQ King. Genetic recombination in positive strand rna viruses. In *RNA Genetics, Volume II, Retroviruses, viroids, and RNA recombination*, pages 149–165. CRC Press Albany, NY, 1988.

[9] Marta Łuksza and Michael Lässig. A predictive fitness model for influenza. *Nature*, 507(7490):57, 2014.

[10] Santiago Marco-Sola, Michael Sammeth, Roderic Guigó, and Paolo Ribeca. The gem mapper: fast, accurate and versatile alignment by filtration. *Nature methods*, 9(12):1185, 2012.

[11] Francois Maree, Lin-Mari de Klerk-Lorist, Simon Gubbins, Fuquan Zhang, Julian Seago, Eva Pérez-Martín, Liz Reid, Katherine Scott, Louis van Schalkwyk, Roy Bengis, et al. Differential persistence of foot-and-mouth disease virus in african buffalo is related to virus virulence. *Journal of virology*, 90(10):5132–5140, 2016.

[12] D McCahon, WR Slade, RAJ Priston, and JR Lake. An extended genetic recombination map for foot-and-mouth disease virus. *Journal of General Virology*, 35(3):555–565, 1977.

[13] Richard A. Neher and Boris I. Shraiman. Competition between recombination and epistasis can cause a transition from allele to genotype selection. *Proceedings of the National Academy of Sciences*, 106(16):6866–6871, 2009.

[14] H Allen Orr. The population genetics of speciation: the evolution of hybrid incompatibilities. *Genetics*, 139(4):1805–1813, 1995.

[15] Emanuele Raineri, Luca Ferretti, Anna Esteve-Codina, Bruno Nevado, Simon Heath, and Miguel Pérez-Enciso. Snp calling by sequencing pooled samples. *BMC bioinformatics*, 13(1):239, 2012.

[16] Tilo Strutz. *Data fitting and uncertainty: A practical introduction to weighted least squares and beyond*. Vieweg and Teubner, 2010.

[17] Chakradhar Tosh, Divakar Hemadri, and Aniket Sanyal. Evidence of recombination in the capsid-coding region of type a foot-and-mouth disease virus. *Journal of general virology*, 83(10):2455–2460, 2002.

[18] Chakradhar Tosh, Aniket Sanyal, and Divakar Hemadri. Genetic and antigenic analysis of a recombinant foot-and-mouth disease virus. *Current Science*, pages 1016–1019, 2002.