JOURNAL OF VIROLOGY, Mar. 2011, p. 2266–2275 Vol. 85, No. 5
0022-538X/11/$12.00 doi:10.1128/JVI.01396-10

Beyond the Consensus: Dissecting Within-Host Viral Population Diversity of Foot-and-Mouth Disease Virus by Using Next-Generation Genome Sequencing$^{\nabla}$‡

Caroline F. Wright,[1,2]† Marco J. Morelli,[2]†* Gaël Thébaud,[3] Nick J. Knowles,[1] Pawel Herzyk,[4] David J. Paton,[1] Daniel T. Haydon,[2] and Donald P. King[1]

Institute for Animal Health, Ash Road, Pirbright, Woking, Surrey GU24 0NF, United Kingdom[1]; MRC, University of Glasgow Centre for Virus Research, Institute of Biodiversity, Animal Health and Comparative Medicine, College of Medical, Veterinary and Life Sciences, University of Glasgow, Glasgow G12 8QQ, United Kingdom[2]; Institut National de la Recherche Agronomique (INRA), UMR BGPI, Cirad TA A-54/K, Campus de Baillarguet, 34938 Montpellier Cedex 5, France[3]; and The Sir Henry Wellcome Functional Genomics Facility, Faculty of Biomedical and Life Sciences, University of Glasgow, Glasgow G12 8QQ, United Kingdom[4]

Received 5 July 2010/Accepted 29 November 2010

The diverse sequences of viral populations within individual hosts are the starting material for selection and subsequent evolution of RNA viruses such as foot-and-mouth disease virus (FMDV). Using next-generation sequencing (NGS) performed on a Genome Analyzer platform (Illumina), this study compared the viral populations within two bovine epithelial samples (foot lesions) from a single animal with the inoculum used to initiate experimental infection. Genomic sequences were determined in duplicate sequencing runs, and the consensus sequence of the inoculum determined by NGS was identical to that previously determined using the Sanger method. However, NGS revealed the fine polymorphic substructure of the viral population, from nucleotide variants present at just below 50% frequency to those present at fractions of 1%. Some of the higher-frequency polymorphisms identified encoded changes within codons associated with heparan sulfate binding and were present in both foot lesions, revealing intermediate stages in the evolution of a tissue culture-adapted virus replicating within a mammalian host. We identified 2,622, 1,434, and 1,703 polymorphisms in the inoculum and in the two foot lesions, respectively: most of the substitutions occurred in only a small fraction of the population and represented the progeny from recent cellular replication prior to onset of any selective pressures. We estimated the upper limit for the genome-wide mutation rate of the virus within a cell to be 7.8 × 10^{-4} per nucleotide. The greater depth of detection achieved by NGS demonstrates that this method is a powerful and valuable tool for the dissection of FMDV populations within hosts.

RNA viruses evolve rapidly due to their large population size, their high replication rate, and the poor proofreading ability of their RNA-dependent RNA polymerase. These viruses exist as heterogeneous and complex populations comprising similar but nonidentical genomes, but the evolutionary importance of this phenomenon remains unclear (15, 16, 25). Consensus sequencing identifies the predominant or major viral sequence present in a sample but is uninformative about minority variants that are present. Evidence for population heterogeneity, where individual sequences differ from the consensus sequence, has been obtained routinely using cloning approaches (1, 8), providing insights into the evolutionary processes that shape viral populations. Unfortunately, these cloning processes are laborious and usually provide only a limited resolution of the mutant spectrum within a sample.

Next-generation sequencing (NGS) techniques offer an unprecedented "step-change" increase in the amount of sequence data that can be generated from a sample. Albeit used mostly for *de novo* sequencing of large genomes, NGS can be applied to resequence short viral genomes to obtain ultradeep coverage. Therefore, NGS has the potential to provide information beyond the consensus for a viral sample by revealing nucleotide substitutions present in only a small fraction of the population. Several studies have previously used the 454 pyrosequencing platform (Roche Applied Science) to detect minority sequence variants for human viruses such as HIV-1 (17, 22, 29, 36, 42, 45, 47), hepatitis B virus (32, 43), hepatitis C virus (48), and attenuated virus (46). A promising alternative to 454 pyrosequencing is the reversible terminator-based sequencing chemistry utilized by the Illumina sequencing platform (Genome Analyzer II). The lower costs of the runs and the higher throughput of this NGS approach are likely to make it widely used for deep-sequencing genomic investigations in the future (41). Illumina sequencing was recently used to obtain sequences of West Nile virus (through the use of virus-derived small interfering RNA [siRNA]) (3), mutants of severe acute respiratory syndrome coronavirus (14), and human rhinovirus (7).

The aim of this study was to explore the extent to which the Illumina sequencing platform can be used to characterize and monitor changes in viral sequence diversity that occur during

* Corresponding author. Mailing address: MRC, University of Glasgow Centre for Virus Research, Graham Kerr Building, Glasgow G12 8QQ, United Kingdom. Phone: 44 141 330 6638. Fax: 44 141 330 5971. E-mail: Marco.Morelli@glasgow.ac.uk.

‡ Supplemental material for this article may be found at http://jvi.asm.org/.

† C.F.W. and M.J.M. contributed equally to this work.

$^{\nabla}$ Published ahead of print on 15 December 2010.

TABLE 1. Oligonucleotide primers used for amplification of two large overlapping FMDV genome fragments for both the first and second runs[c]

PCR set	Primer[a]	Primer sequence (5′ to 3′)	Location on genome[b]	Amplicon size (bp)
1	BFS-370F	CCCCCCCCCCCCCTAAG	351–366	4557
	BFS-4926R	AAGTCCTTGCCGTCAGGGT	4891–4909	4557
2	BFS-3876F	AAATTGTGGCACCGGTGA	3859–3876	4317
	BFS-8193R	TTTTTTTTTTTTTTGATTAAGG	8155–8176	4317
Both	UKFMD/Rev6	GGCGGCCGCTTTTTTTTTTTTTTT	Poly(A) tail	

[a] The last letter indicates whether the primer is a forward (F) or reverse (R) primer.
[b] Numbering is according to the sequence under GenBank accession no. EU448369.
[c] The fragments have the S fragment omitted, up to and including the poly(C) tract, and overlap by 1,051 bp.

replication of a positive-strand RNA virus within a host. This study used NGS to dissect foot-and-mouth disease virus (FMDV) within-host population structure at a depth unobtainable by previous cloning techniques. FMDV belongs to the *Picornaviridae* family and is highly infectious, causing vesicular lesions in the mouth and on the feet of cloven-hoofed animals. The samples analyzed here were collected during an infection experiment in which a bovine host was inoculated with FMDV. We developed a protocol that enabled identification of artifacts introduced during amplification and sequencing which was used to validate and quantify the minority sequence variants that were detected. In particular, we expected to see evidence for reversion of capsid amino acid residues responsible for heparan sulfate (HS) binding associated with replication of a cell culture-adapted strain of FMDV in a mammalian host (20, 37). Although this study was conducted using FMDV, we anticipate that the features we observed may be broadly representative of populations found in samples obtained for other positive-strand RNA viruses.

MATERIALS AND METHODS

Sample preparation and genome amplification. The samples analyzed were collected during an infection experiment in which a single bovine host was inoculated intradermolingually with a dose of $10^{5.7}$ 50% tissue culture infective doses ($TCID_{50}$) of FMDV (O1/BFS 1860). The full-length FMDV genome sequence of this sample had been determined previously using Sanger sequencing (GenBank accession no. EU448369) and was used as a reference genome in this study. The inoculum was derived from a bovine tongue vesicle specimen that had been passaged extensively in cell culture (9).

Total RNA was extracted by use of TRIzol (Invitrogen, Paisley, United Kingdom) from a sample of the inoculum as well as from two 10% tissue suspensions prepared from epithelial lesions (front left foot [FLF] and back right foot [BRF]) collected from the animal at 2 days postinoculation. Reverse transcription (RT) was performed using an enzyme with high specificity (Superscript III reverse transcriptase; Invitrogen) and an oligo(dT) primer (Table 1). For each sample, two PCRs generating long overlapping fragments (4,557 bp and 4,317 bp) were carried out using a proofreading enzyme mixture (Platinum *Taq* Hi-Fidelity; Invitrogen). For biosecurity reasons, these individual fragments together comprised <80% of the complete FMDV genome and corresponded to nucleotides (nt) 351 to 4909 and 3859 to 8176 of the EU448369 sequence. This enabled the amplified DNA to be transported outside the high-containment FMD laboratory for sequencing at The Sir Henry Wellcome Functional Genomics Facility (University of Glasgow). The samples were amplified using the following cycling program: 94°C for 5 min, followed by cycles of 94°C for 30 s, 55°C for 30 s, and 70°C for 4 min, with a final step of 72°C for 7 min. For each RNA sample, the number of PCR cycles used was optimized (using parallel reactions undertaken using Picogreen) such that products were collected from the exponential part of the amplification curve prior to the plateau phase. Once established for each sample, the same optimized cycle number was used for both runs. Individual PCR products were visualized using agarose gel electrophoresis and then quantified (Nanodrop spectrophotometer; Labtech), after which the concentrations of each PCR fragment were adjusted to equimolar ratios for each of the three samples prior to sequence analysis. We repeated the PCR with the original reverse-transcribed sample in order to obtain an independent replica of the amplified sample. The number of viral RNA copies put into the initial PCR mix was established by quantitative PCR for each of the samples (4).

Next-generation sequencing. Sequencing was carried out on a Genome Analyzer II platform (Illumina). Briefly, DNA was fragmented using sonication and the resultant fragment distribution assessed by an Agilent BioAnalyzer 2100 platform. After size selection of fragments of between 300 and 400 bp, a library of purified genomic DNAs was prepared by ligating adapters to the fragment ends to generate flow-cell-suitable templates. A unique 6-nt sequence index or "tag," for identification during analysis, was added to each sample by PCR. Once the adapter/index-modified fragments were pooled and attached to the flow cell by complementary surface-bound primers, isothermal bridging amplification formed multiple DNA clusters for reversible terminator sequencing, yielding reads of 50 nucleotides. We conducted two sequencing runs. In the first, we sequenced the three amplified viral populations (inoculum, FLF sample, and BRF sample) in a single lane after tagging. The second run was performed on a different flow cell: again, we sequenced the same populations in a single lane, using second, independent amplifications of the three original cDNAs. The second run was performed after the Illumina Genome Analyzer instrument went through an upgrade and was able to deliver longer reads of 70 nucleotides.

Data filtering. In order to make direct comparisons between the two runs, we trimmed reads from the second run to 50 nt. Typically, quality scores decreased along a read, as the reliability of the sequencing process decreased with the number of cycles of the sequencing platform. The second run yielded reads of much better quality following an upgrade of the Illumina platform. For both runs, reads with average errors per nt below a fixed threshold ($\theta = 0.2\%$) were discarded to generate a flatter error profile along the read (see Appendix and Fig. A1). The first and last 5 nt of each aligned read were removed from the analysis because they showed larger numbers of mismatches to the reference sequence due to insertions or deletions close to the edges of the reads (see Fig. A2). More details can be found in the appendix.

Validation and analysis of sequence diversity in the samples. The frequency of site-specific polymorphisms was estimated from the frequency of mismatches of the aligned reads to the reference genome. A proportion of these mismatches were expected to be artifacts arising from a base miscalling in the sequencing process or from a PCR error in the amplification of the sample. In order to identify polymorphisms arising from possible base miscalls in the sequencing reaction analysis, we used the quality score of each nucleotide read to compute the average probability of a sequencing error, p_i, at each site i. Typical values of p_i are around 0.1%. Assuming sequencing errors to be independent, we computed the expected number of such errors as the mean of the binomial distribution $B(x; p_i, n_i)$, where n_i is the coverage of site i. If the observed number of mismatches exceeded this expected number of errors in both runs, then we excluded the possibility of a sequencing error. On the other hand, we hypothesized that the probability that PCR errors in both runs independently generated identical base changes at the same site was very low. Based on values quoted for the enzymes used, we estimated that the error rate for the combined RT-PCR amplification process was 7.7×10^{-6} per base pair copied (2, 31, 38). We therefore defined polymorphic sites that could not be attributed to sequencing errors and at which both the most-common and second-most-common nucleotides were the same between the two runs as qualitatively validated sites. For each site in the set of qualitatively validated polymorphisms, we computed the 95% confidence interval (95% CI) for the polymorphism frequencies, using the binomial distribution described above. If the 95% confidence intervals from the runs overlapped, we defined the polymorphism frequency estimates from the two runs to be in quantitative agreement.

We assessed the quantitative repeatability of site-specific polymorphism frequency estimates by calculating Spearman rank correlation coefficients between polymorphism frequencies in the samples within each run and between polymorphism frequencies from runs 1 and 2.

We counted the numbers of transitions (Ts) and transversions (Tv) observed at qualitatively validated sites across the genome and computed the term $\kappa = 2\mathrm{Ts}/\mathrm{Tv}$ and the relative distribution of mutations across the 1st, 2nd, and 3rd codon positions across the open reading frame (ORF). We obtained an estimate for the ratio of nonsynonymous to synonymous evolutionary changes (dN/dS) as follows: for each codon of the reference ORF, we computed the expected number of synonymous (s_i) and nonsynonymous (n_i) sites, and for each read j spanning that codon, we computed the number of observed synonymous ($s^{r_{ij}}$) and nonsynonymous ($n^{r_{ij}}$) substitutions. Using all codons where s_i was >0, we then obtained an estimate of the number of synonymous substitutions per synonymous site (p_S) and nonsynonymous substitutions per nonsynonymous site (p_N), using the following equation:

$$p_S = \frac{1}{n_{\mathrm{cod}}} \sum_{i=1}^{n_{\mathrm{cod}}} \frac{1}{m_i} \sum_{j=1}^{n_j} \frac{s^r_{ij}}{S_i}$$

(and analogously for p_N), where m_i is the number of reads covering codon i and n_{cod} is the total number of codons in the ORF. From p_N and p_S, we obtained dN/dS as described previously (34).

We calculated the number of validated sites at which stop codons were observed within the reading frame and used these counts to estimate an upper limit on the mutation rate. Let n_i be the coverage at the ith nucleotide position, and let $x_{i,\mathrm{obs}}$ be the number of reads indicating a mutation generating a stop codon at the ith position. Assuming independence, the probability density function describing the number of mutations, x_i, that might be observed at site i is the binomial $B(x_i;\lambda,n_i)$, where λ is the mutation frequency corresponding to mutations accumulated at a site during a single cellular passage. The maximum likelihood estimate of λ is $\sum_i x_{i,\mathrm{obs}} / \sum_i n_i$ (18). Using a flat conjugate prior distribution (beta function with shape parameters set to 1), we obtained confidence intervals for λ from the corresponding posterior distribution (beta function with parameters $1 + \sum_i x_{i,\mathrm{obs}}$ and $1 + \sum_i (n_i - x_{i,\mathrm{obs}})$ [21]). Assuming an equal probability for each mutation, λ is related to the mutation rate μ (per nucleotide, per single copying event) via the relationship $\lambda = 2ga\mu$ (44), where g is the number of transcription generations (positive → negative → positive) that the virus underwent in the cell. Here we assumed that $g = 1$, which corresponds to a stamping machine replication strategy and therefore to the minimum number of copying events in a cell. a is a factor weighting the fraction of mutations generating a stop codon among all the possible changes that could arise at a single nucleotide position: we considered only sites whose mutation could lead to a stop codon. Among the 18 codons that are one mutation away from a stop codon, 5 of them (UCA, UUA, UAC, UAU, and UGG) can become a stop codon through either two different mutations to the same position or a single mutation to one of two different positions. Assuming the same probability for each of the 3 nucleotide mutations, we obtained the following: $a = (4 \times 2 + 15 \times 1)/(3 \times 19) = 0.4035$.

Randomizations were conducted whereby we assembled putative clones from the read data by sampling nucleotides randomly from (qualitatively validated) nucleotide frequencies observed at each site along the genome. We computed the median number of observed nucleotide substitutions (those differing from the consensus for the resampled clones) in sets of 26 such independently assembled clones, and these numbers were compared with equivalent numbers from real clones obtained from an individual cow naturally infected with FMDV (8).

The complexity of the viral populations was characterized by computing the entropy of the viral populations as follows:

$$S = -\frac{1}{N} \sum_{i=1}^{N} \sum_{j \in \{A,C,G,T\}} p_{ij} \ln p_{ij}$$

where N is the number of sites and p_{iX} is the fraction of reads bearing nt X at site i. The entropy measures the amount of "disorder" in the population, and it is maximal at a site where all four bases are equally represented.

Online data repository accession numbers. Sequence read data from this study have been deposited in the NCBI Sequence Read Archive (SRA) under accession numbers ERA015837 and ERA015838.

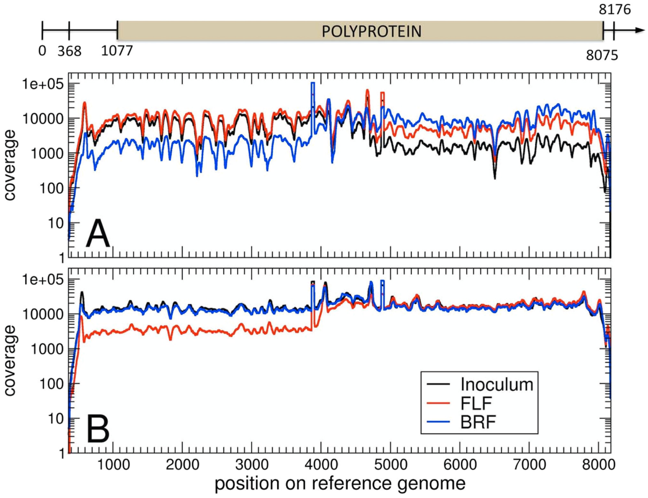

FIG. 1. Coverage of the reference genome obtained with filtered, trimmed reads. (A) First data set (run), (B) second data set (run). The three samples (inoculum, front left foot, and back right foot) received generous coverage from both runs, while fluctuations were higher for the first data set. Average coverage values were ×4,873 (inoculum), ×8,665 (FLF sample), and ×6,594 (BRF sample) for the first data set and ×16,827 (inoculum), ×11,924 (FLF sample), and ×15,945 (BRF sample) for the second data set. At the top of the figure, the sequenced fraction of the genome (nt 368 to 8176) is represented, together with the position of the polyprotein.

RESULTS

In this section, we discuss the results of Illumina sequencing of three FMDV populations: the inoculum (field sample O1/BFS1860/UK/67, used to artificially infect a bovine host) and two lesions developed on two different feet of the host, obtained 2 days after inoculation.

Description and filtering of Illumina data. Sequences from the Illumina Genome Analyzer platform consist of a collection of several million short reads. Sequencing was repeated following independent amplification of cDNAs generated through PCR. In the first run, ~8% of the reads were discarded because of unresolved nucleotides or corrupted tags. In the second run, ~3% of the reads were discarded. Each nucleotide of each read is characterized by a quality score, which quantifies the reliability of the base-calling process during sequencing. Only reads whose average error per nt was below 0.2% (66% of reads for the first run and 95% of reads for the second run) were considered for this analysis. Further details about the reads and the filtering process can be found in the appendix.

Coverage and consensus genomes. Reads that passed the quality test were aligned to the consensus genome sequence of the starting material from which the inoculum was prepared (see Appendix). The mean coverage of the reference genome in the first run was ×4,863 for the inoculum, ×8,665 for the FLF sample, and ×6,594 for the BRF sample, while for the second run it was ×16,827 for the inoculum, ×11,924 for the FLF sample, and ×15,945 for the BRF sample (Fig. 1A and B). For some samples (inoculum and BRF samples in first run and the FLF sample in second run), the levels of coverage for the two PCR fragments composing the viral genome were not

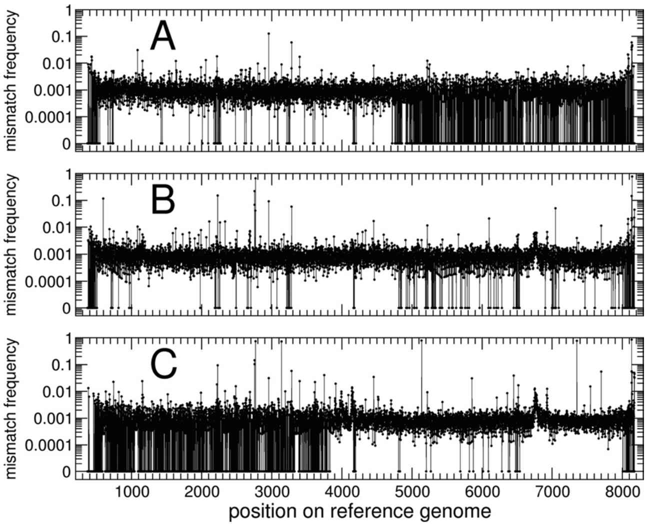

FIG. 2. Frequencies of mismatches (first data set) obtained by aligning the reads to the reference genome. (A) Inoculum; (B) FLF sample; (C) BRF sample. The average mismatch frequency lay around 0.1% for all three samples. At a few sites, the mismatch frequency was higher; as expected, the number of these peaks was larger for the FLF and BRF samples than for the inoculum. A small fraction of sites showed perfect agreement of all reads with the reference genome (mismatch frequency = 0).

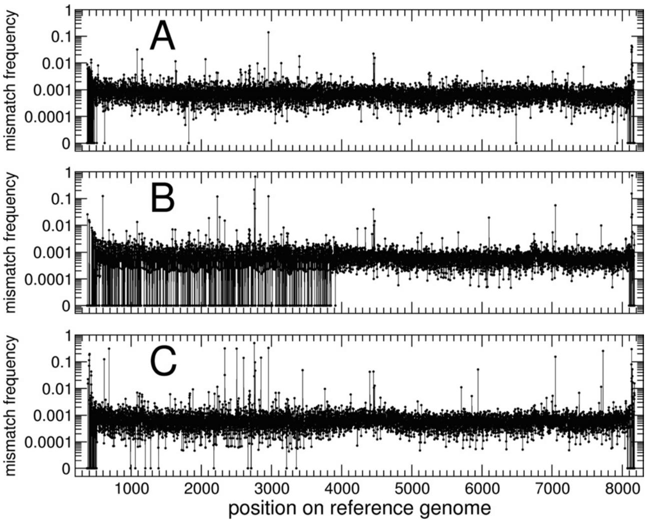

FIG. 3. Frequencies of mismatches (second data set) obtained by aligning the reads to the reference genome. (A) Inoculum; (B) FLF sample; (C) BRF sample. The second data set had a higher level of coverage than the first one, with a smaller fraction of sites with no mismatches. The average mismatch frequency is very similar to that of the first data set.

equal. More details on the statistics of the Illumina yield can be found in the appendix.

We obtained consensus genomes for each sample by identifying, site by site, the most abundant nucleotide in the aligned reads. As expected, the consensus for the inoculum exactly matched the reference genome at all sites. For the FLF sample, both runs indicated two substitutions (at nt 2767, G→A; and at nt 8140, G→T). For the BRF sample, the two runs suggested slightly different consensus sequences: the first run revealed five substitutions (at nt 2767, G→A; at nt 3138, G→A; at nt 5138, T→C; at nt 7354, C→T; and at nt 8134, C→T), whereas the second run had none. However, at position 8134, about 30% of the reads in the second run showed a T in place of a C, and at position 2767, 5% of the reads had an A in place of a T. For the remaining 3 sites, the second run had a small number of reads confirming the polymorphism found in the first run. This result indicates that the same pattern of variation was present in both runs, although the frequencies of the mutations were not in quantitative agreement across the two runs for the BRF sample. Finally, the second run showed an almost-consensus substitution in 49.9% of the reads (at nt 2754 [C→T]) which was present at a 10% frequency in the first run.

Validation of polymorphic sites. Mismatch frequencies, obtained by determining the fraction of reads differing from the consensus genome site by site, are shown in Fig. 2 (first run) and Fig. 3 (second run). An evident correlation was present between the regions of the sample genomes receiving low coverage and those with the largest fraction of sites showing no variation (Fig. 2A, second half, Fig. 2C, first half, and Fig. 3B, first half). Using these raw data, and considering only sites receiving coverage of ×100 or more, we found polymorphisms at 7,755 sites for the inoculum, 7,730 sites for the FLF sample, and 7,710 sites for the BRF sample, out of the 7,825 nt sequenced. While a few sites exhibited higher levels of polymorphism, the vast majority of sites displayed a mismatch frequency of around 0.1%.

After screening for possible PCR and sequencing artifacts, we found that qualitatively validated polymorphisms were present at 2,622 sites for the inoculum, 1,434 sites for the FLF sample, and 1,703 sites for the BRF sample. The different consensus genomes obtained in the two runs for the BRF sample can be reconciled in part by noting that all six substitutions observed (at nt 2754, 2767, 3138, 5138, 7354, and 8134) were validated qualitatively in each run. We observed 2,469 quantitatively validated sites for the inoculum (94% qualitatively validated sites), 1,303 sites for the FLF sample (91% qualitatively validated sites), and 1,528 sites for the BRF sample (90% qualitatively validated sites).

Site-specific polymorphism (SSP) frequencies at qualitatively validated sites were correlated between the two runs for each of the three samples (Fig. 4). The intrarun correlation for run 1 (Spearman rank correlation values, 0.64 [inoculum-FLF sample], 0.55 [inoculum-BRF sample], and 0.60 [FLF sample-BRF sample]) was higher than that for run 2 (Spearman rank correlation values, 0.40 [inoculum-FLF sample], 0.43 [inoculum-BRF sample], and 0.42 [FLF sample-BRF sample]). The reason for the poor intrarun correlation for run 2 is unclear. The numbers of viral RNA copies present in the initial PCRs were found to be large (3.2×10^9 for the inoculum, 6.4×10^8 for the FLF sample, and 2.4×10^8 for the BRF sample): assuming that the PCR process amplifies all genomes with the same probability, the probability of resequencing the same genome is exceedingly low ($<10^{-5}$), thus excluding the possibility of biases due to low viral load in the RNA. However, relative to run 1, run 2 yielded consistently smaller amounts of DNA library concentrations per sample prior to sequencing (3.4 versus 4.9 ng/μl, 3.7 versus 10.6 ng/μl, and 3.4 versus 9.5

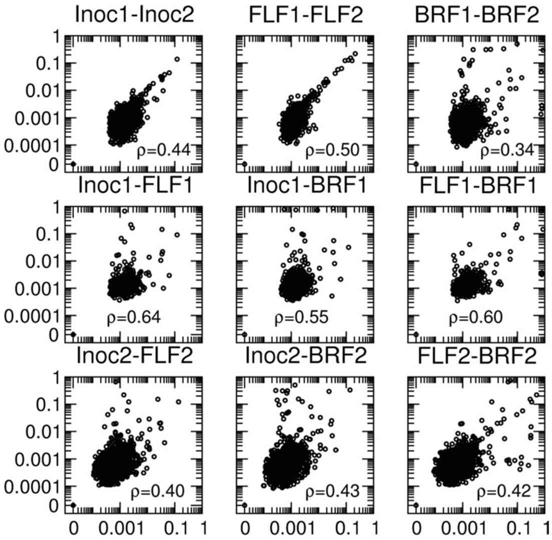

FIG. 4. Correlations of polymorphism frequencies in viral populations. Correlations were computed between the two runs (first row) and within each run (second and third rows). The Spearman rank correlation ρ is indicated for each pair of data sets. Only data for qualitatively validated SSPs receiving coverage above ×100 in both runs are shown. The correlation coefficients between the two runs for the inoculum and the FLF sample are similar, while they are lower for the BRF sample. The remaining panels show that the first run was more correlated than the second run.

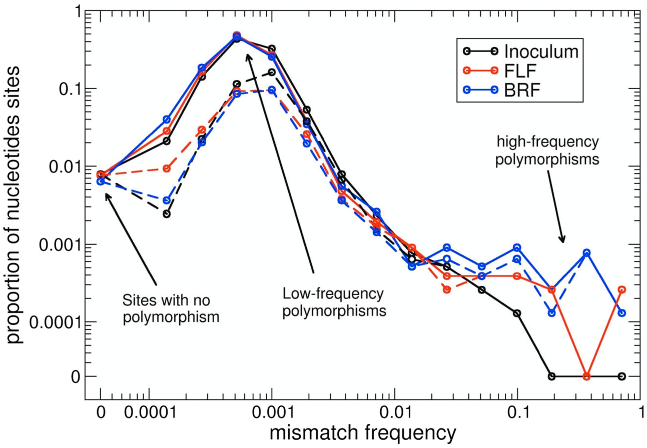

FIG. 5. Variability in viral populations. Frequency distributions of the weighted averaged mismatch frequencies between the two runs are shown for the three samples (the ordinate represents the frequencies of sites showing each fraction of mismatches). Solid lines, all sites receiving minimum coverage of ×100 in both runs (7,755 sites for inoculum, 7,730 for FLF sample, and 7,710 sites for BRF sample); dashed lines, sites receiving coverage of ×100 or more in both runs and classified as validated SSPs (2,622 sites for inoculum, 1,434 for FLF sample, and 1,703 for BRF sample). All lines show similar trends: a small fraction of the sites (<1%) display no variability for both runs, most of the sites show a very small amount of polymorphism in the viral population (between 0.01% and 1%), and a very small fraction of the sites (0.14% for inoculum, 0.22% for FLF sample, and 0.39% for BRF sample) present variation at a level above 1%.

ng/μl for the inoculum and the FLF and BRF samples, respectively), and this may have contributed to the differences between the two consensus sequences. Despite the discrepancies in frequencies of these mutations between runs, the fact that the same mutations were present at the same qualitatively validated sites in both runs provides confidence that the mutations are genuine and not artifacts. The intrarun correlation, together with the high fractions of quantitative validation among the qualitatively validated SSPs, provides sound evidence that nt changes are linked between the different samples. The interrun correlation between the samples (Spearman rank correlation, 0.34 versus 0.44 and 0.50) indicates that validated polymorphisms are unlikely to be artifacts.

Distribution of polymorphisms across the genome. There were 12 SSPs for the inoculum whose average frequencies between the two runs were above 1%, with 19 for the FLF sample and 25 for the BRF sample (see Table S1 in the supplemental material). Some of these were clustered in the capsid protein region (beginning of protein VP3) (1 for the inoculum, 4 for the FLF sample, and 5 for the BRF sample) and in the 3′-untranslated region (3′UTR) (6 for the inoculum, 5 for the FLF sample, and 6 for the BRF sample). Where single reads spanning these sites within the VP3 or 3′UTR were available, there was no evidence that that these mutations were linked together on individual FMDV genomes. In particular, the first cluster was shared between the two foot samples and corresponded to changes encoding amino acid residues associated with HS binding. The inoculum used in this experiment had undergone extensive cell culture passages and, in common with other *in vitro*-adapted viruses, utilizes HS as a cellular receptor (27). Subsequent replication in mammalian hosts drives the reversion of positively charged amino acid residues at specific sites in the viral capsid (20, 37). A consensus-level substitution (>50%) compared to the reference sequence existed for both foot samples in run 1 (see above and Table S1). This polymorphism corresponded to a change within the 60th codon of the VP3 protein (VP3^{60}). Although below the level of the consensus sequence, additional qualitatively validated SSPs that were present in both foot samples were detected at four further sites (VP2^{134}, two codon positions within VP3^{56}, and VP3^{59}) that impact the ability of FMDV to bind HS. All but one of the mutations clustered within the 3′UTRs of the three samples were located within the first four RNA-RNA pairings either side of the apex of a conserved stem-loop. This structure, one of two stem-loops previously predicted for FMDV and other picornaviruses (6, 33), is thought to generate long-distance RNA-RNA interactions that may impact viral replication (40). The presence of shared mutations between the two foot samples suggests a common history for the viruses, with the viruses arising as a result of the shared route of intrahost transmission from initial replication sites in the tongue to epithelial sites in the feet via the blood. However, an alternative explanation—that viruses were subject to a common selection pressure at both sites—cannot be ruled out.

Frequencies of site-specific polymorphisms. Some variability was present almost everywhere in the genome. Above the minimum coverage level of ×100, only 61 sites exhibited no polymorphism (0.79%) for the inoculum, with 59 (0.76%) sites for the FLF sample and 49 (0.64%) sites for the BRF sample. These sites received relatively low coverage levels, suggesting that the absence of observed genetic variability may have been due to a lack of power to detect it. By grouping the site-specific polymorphism frequencies into discrete bins, we examined the

TABLE 2. Statistics for polymorphic sites[a]

Sample	No. of sites	No. of SSPs	No. of Ts	No. of Tv	κ	*dN*/*dS*	Proportion of mutations at codon position		
							1	2	3
Inoculum	7,755	2,622	2,562	60	85.40	0.651	0.288	0.286	0.427
FLF sample	7,730	1,434	1,400	34	82.36	1.065	0.326	0.333	0.341
BRF sample	7,710	1,703	1,649	54	61.08	0.680	0.334	0.307	0.359

[a] General statistics are shown for qualitatively and quantitatively validated SSPs receiving coverage of >×100. Ts, transitions in SSPs; Tv, transversions in SSPs; κ = 2Ts/Tv; *dN*, nonsynonymous mutations in the ORF; *dS*, synonymous mutations in the ORF.

proportions of sites experiencing different polymorphic frequencies and thereby obtained a comprehensive picture of the heterogeneity in the viral populations (Fig. 5). Across the three samples, most sites exhibited a range of low-frequency SSPs from 0.01% to 1%. Only a few sites showed higher-frequency polymorphisms, and these sites were more numerous for the samples from the feet than for the inoculum, indicating the generation of new high-frequency substitutions during passage in the host. The dashed lines in Fig. 5 correspond to the same analysis restricted to qualitatively validated sites and reveal a similar pattern.

Statistics of polymorphic sites. NGS provided sufficient resolution to detect polymorphisms where two alternative substitutions were present simultaneously. The secondary substitutions (the third most abundant nucleotides in the reads at any particular site) that would have been qualitatively valid even in the absence of the second most abundant nucleotide substitution were present at 67 sites for the inoculum, 15 sites for the FLF sample, and 41 sites for the BRF sample. Secondary substitutions typically appeared at frequencies below 0.5%, confirming the large amount of low-frequency variability in the samples.

Table 2 shows that transversions are rare among the validated mutations, and thus κ (defined as 2Ts/Tv) is high (similar values are reported in reference 8). The ratio of nonsynonymous to synonymous substitutions in the open reading frame (*dN*/*dS*) was higher for the FLF sample than for the other two samples because of the presence in a large number of reads of nonsynonymous mutations at positions 2754 and 2767, associated with heparan sulfate-binding amino acid reversions within $VP3^{56}$ and $VP3^{60}$, respectively. The mutation frequency at the third codon position was only marginally higher than those for the first and second positions. Taken together, these observations suggest that the observed polymorphisms are dominated by mutations arising during the last round of intracellular replication that have not been subjected to extensive purifying selection. Further evidence of this lack of selective pressure is provided by the presence of validated polymorphisms generating stop codons within the ORF. These mutations are clearly lethal for the virus and would therefore be removed from the population during infection of another cell. They must therefore have arisen during the most recent rounds of viral replication. Stop codons were found at 24 sites for the inoculum, 9 sites for the FLF sample, and 21 sites for the BRF sample, mostly at frequencies of around 0.1% (with a single exception for the BRF sample, where a mutation generating a stop codon was present in 0.7% of the reads).

The presence of stop codons can be used to obtain an upper limit on the mutation rate (number of mutations per nucleotide per transcription event) of this virus. We hypothesized that these mutations are lethal and are therefore generated in the last round of cellular replication. Moreover, by assuming a replication strategy involving the minimum number of copying events in the cell (the "stamping machine" strategy [44]), we obtained an upper bound for the mutation rate (μ) of 7.8×10^{-4} per nucleotide per transcription event (95% CI, 7.4×10^{-4} to 8.3×10^{-4}), in line with previous estimates (e.g., see references 12, 13, and 39).

Finally, we asked whether these results are broadly consistent with those acquired from cloning studies. Cottam et al. (8) generated 26 viral capsid clones from an FMDV sample taken from a single lesion of a bovine host. We simulated 10,000 sets of 26 viral capsid clones, essentially bootstrapping from the nucleotide frequencies revealed by the NGS alignments to be present at each site within the capsid genes. For these 26 clones, the median number of sequences in each of the 10,000 simulated data sets that were identical to the consensus was 12 (95% CI, 5 to 17), compared to the previously observed value of 15 (8). The median numbers of simulated clones containing 1, 2, 3, and 4 differences compared to the consensus were 9 (95% CI, 4 to 14), 3 (95% CI, 1 to 7), 1 (95% CI, 0 to 3), and 0 (95% CI, 0 to 1), respectively. These numbers correspond well with those obtained by Cottam et al. (8), which were 6, 3, 2, and 0, respectively.

Complexity of viral populations. In the host, the viral population evolves via extensive replication, mutation, and selection. The result of these combined processes can be quantified by computing how much diversity is present within the three samples, using the entropy-like measure S that, site by site, takes a maximum value when all nt are present in the same proportion. The entropy of the three populations, computed for the qualitatively validated sites, showed higher values for the foot samples than for the inoculum ($S = 0.01138$ for the FLF sample, $S = 0.01198$ for the BRF sample, and $S = 0.00841$ for the inoculum), suggesting that repeated cycles of cellular replication during passage in the host do result in greater viral population diversity than that in the inoculum.

DISCUSSION

This study describes a novel use of Illumina NGS to investigate the population genetic structure of a positive-strand RNA virus causing an acute-acting disease in hosts. These experiments generated an unprecedented amount of sequence data and required a new systematic approach to confidently distinguish sequences that were actually present in the samples from artifacts introduced during the amplification and sequencing steps of sample processing. The results obtained here were consistent with the findings of previous investigations, providing validation of the use of NGS in the study of FMDV evolution within a host. Carrillo et al. (5) reported an average of 1 to 5 substitutions per animal passage during an infection experiment with pigs, in line with the 2 substitutions we found in the FLF sample. However, the case of the BRF sample points out a more complex scenario that could not have been observed with consensus sequences only: the drift of mutations above and below the threshold needed to appear in the consensus. Apparent loss and subsequent regain of mutations during the transmission of infection across hosts (5) can be explained by this mechanism, which is made more accessible to study by NGS. Moreover, the statistical characteristics of the SSPs we identified (κ and *dN*/*dS*) were very similar to those found previously (8), further corroborating the validity of our results. Finally, randomizations of the diversity measured in the capsid region allowed us to obtain simulated clones whose characteristics in terms of mutation were analogous to those found previously (8). We concluded that NGS data can be used to examine the nucleotide diversity of each genome position at unprecedented resolution. Observing the mutant spectrum of the viral population at a fine resolution will provide a more sophisticated understanding of evolutionary processes shaping its variability.

Comparisons between the sequences recovered from the inoculum and from clinical lesions provided new insights into the impact of early replication events on viral evolution within a host. This study revealed that only a few sites displayed mutations present in a large fraction of the population, i.e., high-frequency polymorphisms (>1%), while the vast majority of the polymorphisms were present at lower frequencies. We hypothesize that the high-frequency polymorphisms were selected over multiple rounds of replication within cells and that the low-frequency polymorphisms most likely directly reflect the high rate of mutation experienced by these viruses, as suggested by our estimate of the upper limit of the genome-wide mutation rate. In this study, we used a cell culture-adapted virus as the inoculum, which gave us the opportunity to monitor changes at specific loci associated with the HS binding site that were under selection pressure during initial replication in a mammalian host. Examination of these sites (collated in Table S1 in the supplemental material) revealed, for the first time, the presence of intermediate stages in the evolution of the viral population between a tissue culture-adapted genome and a host-adapted genome.

Cordey et al. (7) investigated the dynamics of human rhinovirus (HRV) during an infection experiment and in HeLa cells and found results similar to ours in terms of the number of mutations fixed at the consensus level. However, while their approach identified hot and cold spots in the HRV ORF, with some minority variants, the resolution was not sufficient to observe the microevolutionary processes whose signature lies in small fractions of the viral population (<2%). Moreover, their estimation of the substitution rate during the infection was based solely on the count of the nucleotides changed among those analyzed, and although the value is compatible with our genome-wide mutation rate, we believe that considering the cellular process of viral replication (and specifically assuming the minimum number of copying events in a cell) allowed us to gain a better insight into the process generating variation in the viral population and to obtain a more stringent upper bound.

Figure 5 reveals that the viral population sequences are highly heterogeneous, supporting the findings of previous studies that have used cloning approaches (10, 28). However, the massively increased coverage enabled by NGS permits the nature of this heterogeneity to be established at much greater resolution. This is important for understanding viral evolutionary processes because heterogeneity is a necessary but not sufficient condition (23, 24) for the dominance of quasispecies dynamics (see references 10 and 15 and references therein). For quasispecies dynamics to dominate the microevolutionary process, the frequency of the master sequence must be maintained primarily by the back mutation or recombination of closely related genetic variants rather than the faithful replication of any single genome. This requires a balance of two qualities: genetic variants closely related to the dominant sequence must be maintained at sufficiently high prevalence, and the mutation and recombination rates must be sufficiently high to generate the observed prevalence of the dominant sequence among these variants. Previous studies have examined this question empirically and concluded that these conditions are indeed met in many RNA viruses, mostly through studies of mutational robustness as a selectable trait ("survival of the flattest" effect, in which selection acts not on the dominant sequence but on the swarm of viruses immediately mutationally adjacent to the dominant sequence) (11, 35; reviewed in reference 19, with particular focus on hepatitis C virus). However, taking FMDV as an example, given that there are ~25,000 one-step mutant variants to any one sequence (3 alternative nucleotides at each position of the ~8,300-nucleotide genome), NGS approaches are clearly a powerful tool for examining directly whether viral populations are structured in a way that is consistent with quasispecies dynamics.

NGS data can be coupled to evolutionary models to estimate parameters such as the genome-wide mutation rate of FMDV. Here we computed this number by hypothesizing that the viral replication strategy follows the so-called "stamping machine" mode of replication, where all viral genomes leaving the cells are obtained as copies of "first-generation" negative-strand genomes, which in turn are direct copies of the genomic RNA originally infecting the cells. For this reason, the estimate of 7.8×10^{-4} mutation per genome per transcription round should be considered an upper bound on the mutation rate, which is a tighter estimate than previous figures obtained for other RNA viruses (12, 13) as a result of the deep coverage that NGS generates. Were the replication strategy "geometric" (i.e., including the possibility of several rounds of positive/negative-strand copying before exiting the cell), the mutation rate would be severalfold (perhaps 3 to 6 times) lower (44). The assumptions that all nucleotide mutations at a site are equally likely and that all stop codons are generated by a *de novo* mutation are also likely to lead to an overestimation of the mutation rate.

Until the present, analysis of the amount of complexity carried by a genome has coincided mostly with information-theoretical measures aimed to quantify the entropy and the frequency distributions of short oligomers (26, 30). This approach looks at the "horizontal" complexity along a genome; with NGS, we are now able to obtain closely related sequences for a whole viral population in a single experiment, thus enabling us to look at the "vertical" complexity of the viral variants, i.e., the amount of variability present in the population at each site.

A viral population within a host undergoes complex processes, including the onset of infection, cellular replication, selection, and migration to different tissues. In particular, it is not clear how the diversity generated within a cell propagates through a host to give rise to the amount of diversity we observe. The data collected in studies such as the present study can be used to build models aimed at understanding the link between the microevolution of FMDV at the cellular scale and the population heterogeneity at the host scale. We anticipate that a model of viral replication across several cell generations within a host will produce a more stringent upper bound to the genome-wide mutation rate.

Although further work is required, these findings strongly suggest that data generated through the use of this methodology can provide novel insights into viral evolutionary dynamics at a greater resolution than that previously achieved for a positive-strand virus such as FMDV. In particular, the genome-wide assessment of polymorphic frequencies is likely to be an important asset in the parameterization of models that can evaluate the role of quasispecies dynamics in RNA virus evolution.

APPENDIX

Basic statistics of Illumina yield. The reads obtained with the Illumina Genome Analyzer were collected in fastq files. The first run consisted of a total of 7,190,884 reads of 57 nt each. The last 7 nt of each read defined the sequence tag and were used to assign individual reads to each sample. Reads containing at least one unresolved nt (387,809; 5.55% of the total) and reads having a corrupted tag (207,749; 2.89% of the total) were removed from the analysis. The 6,595,326 remaining 50-nt reads were assigned to the three samples as follows: 1,723,151 (26.1%) had the first tag (corresponding to the inoculum), 2,751,260 (41.7%) had the second tag (corresponding to the lesion on the front left foot [FLF sample]), and 2,112,932 (32.0%) had the third tag (corresponding to the lesion on the back right foot [BRF sample]).

The second run yielded 10,116,147 79-nt long reads, with the last 9 nt containing the sequence tag. A total of 26,428 (0.27%) reads contained at least one unresolved nucleotide, and 288,230 (2.85%) reads had a corrupted tag and were removed from the analysis. Among the remaining 9,801,489 70-nt reads, 3,775,685 (38.5%) belonged to the inoculum, 2,542,913 (25.9%) to the FLF sample, and 3,482,891 (35.5%) to the BRF sample.

Data filtering. The quality scores associated with each nucleotide were lower on the first run and decreased toward the end of reads (Fig. A1). In order to make direct comparisons between the two runs, we trimmed reads from the second run to 50 nt. Typically, quality scores decreased along a read, as the reliability of the sequencing process decreased with the number of cycles of the sequencing platform. As Fig. A1 shows, a tradeoff was present between the number of reads kept and their quality. For both runs, we discarded reads with an average error per nt (θ) below 0.2%, resulting in a flatter error profile along the read.

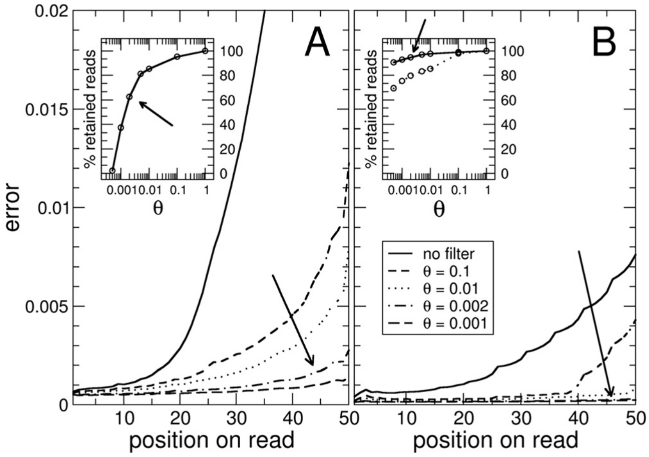

FIG. A1. Average errors in reads, computed with base qualities. (A) First data set; (B) second data set. The average error increased greatly towards the ends of the reads (solid lines). The second data set was less noisy. Different filtering strategies were tested, and only reads whose average error was below a threshold θ value were accepted. More-stringent thresholds decreased the errors in the reads (small dashed, dotted, dotted-dashed, and dashed lines). The insets show the fractions of reads retained after the filtering process (using a threshold θ value of 0.2%), including 66% of the reads in the first data set and 95% of the reads in the second data set.

With this choice of threshold, 66% of reads were retained from the first run (a total of 4,361,101 reads, with 1,060,906 for the inoculum, 1,736,381 for the FLF sample, and 1,328,588 for the BRF sample), and 95% of reads were retained from the second run (a total of 9,277,876 reads, with 3,567,541 for the inoculum, 2,412,897 for the FLF sample, and 3,303,438 for the BRF sample). The better performance of the second data set has to be attributed to an upgrade of the Illumina platform.

Read alignment and trimming. The vast majority of the filtered 50-nt reads aligned unambiguously, with fewer than 5 mismatches, to the reference inoculum genome (GenBank accession no. EU448369), previously established using conventional Sanger sequencing (9) (for run 1, 92.5% of reads for the inoculum, 98.9% for the FLF sample, and 97.8% for the BRF sample; and for run 2, 95.8% of reads for the inoculum, 98.4% for the FLF sample, and 96.2% for the BRF sample). The remaining reads either were ambiguously aligned reads or contained a large number (>4) of mismatches to the reference sequence and were discarded from the analysis. For each sample, almost equal numbers of reads were derived from positive and negative strands of the viral cDNA.

Further filtering of the data was undertaken after alignment of the reads. Within the aligned reads, mismatches occurred with similar frequencies at each of the 50 nt of the reads, except for the edges, where larger numbers of mismatches were observed (Fig. A2). Presumably, these peaks were due to a small number of sequences with insertions or deletions close

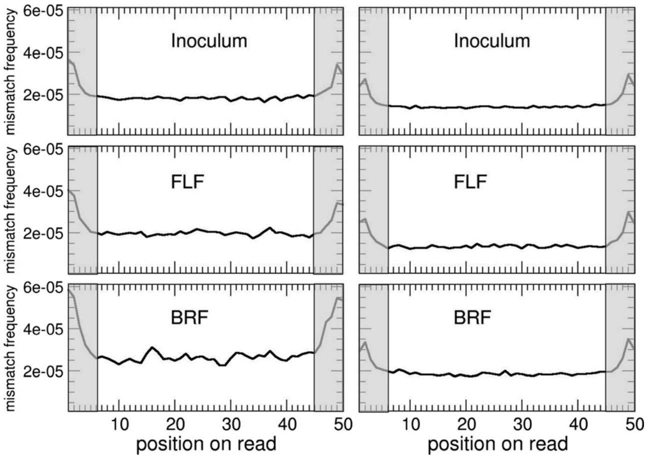

FIG. A2. Distribution of mismatches to the reference genome on the reads after alignment. (Left) First data set; (right) second data set. The curves are largely flat, indicating an even distribution of mismatches over the reads, apart for a mild increase towards the edges of the reads, possibly due to reads containing insertions and deletions. We kept only data coming from the flat region of the curve, i.e., nucleotides 5 to 45 in each aligned read.

to the ends of the reads: for subsequent analysis, we trimmed away the first and last 5 nt of each aligned read, leaving only the 40 central nucleotides for which the mismatch curve was flat.

Data handling. All data handling was performed with parsing scripts, written in C language, acting on plain text files.

ACKNOWLEDGMENTS

This work was supported by the Biotechnology and Biological Sciences Research Council, United Kingdom, via a DTA Ph.D. studentship (project BB/E018505/1), a SYSBIO postdoctoral grant (project BB/F005733/1), and the IAH's Institute Strategic Programme Grant on FMD.

We thank N. Juleff for providing the samples used in the study and J. Galbraith at the Sir Henry Wellcome Functional Genomics Facility, University of Glasgow, for sequencing the samples and providing advice. We also thank C. Randall, L. Fitzpatrick, and M. Jenkins for their care of the animals used in the experimental infection study and Derek Gatherer and Gareth Elvidge for useful discussions.

REFERENCES

1. **Airaksinen, A., N. Pariente, L. Menendez-Arias, and E. Domingo.** 2003. Curing of foot-and-mouth disease virus from persistently infected cells by ribavirin involves enhanced mutagenesis. Virology **311:**339–349.
2. **Arezi, B., and H. H. Hogrefe.** 2007. Escherichia coli DNA polymerase III epsilon subunit increases Moloney murine leukemia virus reverse transcriptase fidelity and accuracy of RT-PCR procedures. Anal. Biochem. **360:**84–91.
3. **Brackney, D. E., J. E. Beane, and G. D. Ebel.** 2009. RNAi targeting of West Nile virus in mosquito midguts promotes virus diversification. PLoS Pathog. **5:**e1000502.
4. **Callahan, J. D., et al.** 2002. Use of a portable real-time reverse transcriptase-polymerase chain reaction assay for rapid detection of foot-and-mouth disease virus. J. Am. Vet. Med. Assoc. **220:**1636–1642.
5. **Carrillo, C., et al.** 2007. Genetic and phenotypic variation of foot-and-mouth disease virus during serial passages in a natural host. J. Virol. **81:**11341–11351.
6. **Carrillo, C., et al.** 2005. Comparative genomics of foot-and-mouth disease virus. J. Virol. **79:**6487–6504.
7. **Cordey, S., et al.** 2010. Rhinovirus genome evolution during experimental human infection. PLoS One **5:**e10588.
8. **Cottam, E. M., D. P. King, A. Wilson, D. J. Paton, and D. T. Haydon.** 2009. Analysis of foot-and-mouth disease virus nucleotide sequence variation within naturally infected epithelium. Virus Res. **140:**199–204.
9. **Cottam, E. M., et al.** 2008. Transmission pathways of foot-and-mouth disease virus in the United Kingdom in 2007. PLoS Pathog. **4:**e1000050.
10. **Domingo, E., et al.** 2006. Viruses as quasispecies: biological implications. Curr. Top. Microbiol. Immunol. **299:**51–82.
11. **Domingo, E., D. Sabo, T. Taniguchi, and C. Weissmann.** 1978. Nucleotide sequence heterogeneity of an RNA phage population. Cell **13:**735–744.
12. **Drake, J. W.** 1993. Rates of spontaneous mutation among RNA viruses. Proc. Natl. Acad. Sci. U. S. A. **90:**4171–4175.
13. **Drake, J. W., and J. J. Holland.** 1999. Mutation rates among RNA viruses. Proc. Natl. Acad. Sci. U. S. A. **96:**13910–13913.
14. **Eckerle, L. D., et al.** 2010. Infidelity of SARS-CoV Nsp14-exonuclease mutant virus replication is revealed by complete genome sequencing. PLoS Pathog. **6:**e1000896.
15. **Eigen, M.** 1971. Selforganization of matter and evolution of biological macromolecules. Naturwissenschaften **58:**465–523.
16. **Eigen, M., and P. Schuster.** 1978. Hypercycle—principle of natural self-organization. B. Abstract hypercycle. Naturwissenschaften **65:**7–41.
17. **Eriksson, N., et al.** 2008. Viral population estimation using pyrosequencing. PLoS Comput. Biol. **4:**e1000074.
18. **Evans, M., N. A. J. Hastings, and J. B. Peacock.** 2000. Statistical distributions, 3rd ed. Wiley, New York, NY.
19. **Fishman, S. L., and A. D. Branch.** 2009. The quasispecies nature and biological implications of the hepatitis C virus. Infect. Genet. Evol. **9:**1158–1167.
20. **Fry, E. E., et al.** 1999. The structure and function of a foot-and-mouth disease virus-oligosaccharide receptor complex. EMBO J. **18:**543–554.
21. **Gelman, A.** 2004. Bayesian data analysis, 2nd ed. Chapman & Hall/CRC, Boca Raton, FL.
22. **Hoffmann, C., et al.** 2007. DNA bar coding and pyrosequencing to identify rare HIV drug resistance mutations. Nucleic Acids Res. **35:**e91.
23. **Holmes, E. C.** 2010. Does hepatitis C virus really form quasispecies? Infect. Genet. Evol. **10:**431–432.
24. **Holmes, E. C.** 2010. The RNA virus quasispecies: fact or fiction? J. Mol. Biol. **400:**271–273.
25. **Holmes, E. C., and A. Moya.** 2002. Is the quasispecies concept relevant to RNA viruses? J. Virol. **76:**460–465.
26. **Holste, D., I. Grosse, and H. Herzel.** 2001. Statistical analysis of the DNA sequence of human chromosome 22. Phys. Rev. E Stat. Nonlin. Soft Matter Phys. **64:**041917.
27. **Jackson, T., et al.** 1996. Efficient infection of cells in culture by type O foot-and-mouth disease virus requires binding to cell surface heparan sulfate. J. Virol. **70:**5282–5287.
28. **Jridi, C., J. F. Martin, V. Marie-Jeanne, G. Labonne, and S. Blanc.** 2006. Distinct viral populations differentiate and evolve independently in a single perennial host plant. J. Virol. **80:**2349–2357.
29. **Le, T., et al.** 2009. Low-abundance HIV drug-resistant viral variants in treatment-experienced persons correlate with historical antiretroviral use. PLoS One **4:**e6079.
30. **Liu, Z., S. S. Venkatesh, and C. C. Maley.** 2008. Sequence space coverage, entropy of genomes and the potential to detect non-human DNA in human samples. BMC Genomics **9:**509.
31. **Malet, I., M. Belnard, H. Agut, and A. Cahour.** 2003. From RNA to quasispecies: a DNA polymerase with proofreading activity is highly recommended for accurate assessment of viral diversity. J. Virol. Methods **109:**161–170.
32. **Margeridon-Thermet, S., et al.** 2009. Ultra-deep pyrosequencing of hepatitis B virus quasispecies from nucleoside and nucleotide reverse-transcriptase inhibitor (NRTI)-treated patients and NRTI-naive patients. J. Infect. Dis. **199:**1275–1285.
33. **Melchers, W. J., et al.** 2000. Cross-talk between orientation-dependent recognition determinants of a complex control RNA element, the enterovirus oriR. RNA **6:**976–987.
34. **Nei, M., and T. Gojobori.** 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol. Biol. Evol. **3:**418–426.
35. **Pfeiffer, J. K., and K. Kirkegaard.** 2005. Increased fidelity reduces poliovirus fitness and virulence under selective pressure in mice. PLoS Pathog. **1:**e11.
36. **Rozera, G., et al.** 2009. Massively parallel pyrosequencing highlights minority variants in the HIV-1 env quasispecies deriving from lymphomonocyte subpopulations. Retrovirology **6:**15.
37. **Sa-Carvalho, D., et al.** 1997. Tissue culture adaptation of foot-and-mouth disease virus selects viruses that bind to heparin and are attenuated in cattle. J. Virol. **71:**5115–5123.
38. **Salazar-Gonzalez, J. F., et al.** 2008. Deciphering human immunodeficiency virus type 1 transmission and early envelope diversification by single-genome amplification and sequencing. J. Virol. **82:**3952–3970.
39. **Schrag, S. J., P. A. Rota, and W. J. Bellini.** 1999. Spontaneous mutation rate of measles virus: direct estimation based on mutations conferring monoclonal antibody resistance. J. Virol. **73:**51–54.
40. **Serrano, P., M. R. Pulido, M. Saiz, and E. Martinez-Salas.** 2006. The 3′ end of the foot-and-mouth disease virus genome establishes two distinct long-range RNA-RNA interactions with the 5′ end region. J. Gen. Virol. **87:**3013–3022.
41. **Shendure, J., and H. Ji.** 2008. Next-generation DNA sequencing. Nat. Biotechnol. **26:**1135–1145.

42. **Simen, B. B., et al.** 2009. Low-abundance drug-resistant viral variants in chronically HIV-infected, antiretroviral treatment-naive patients significantly impact treatment outcomes. J. Infect. Dis. **199:**693–701.

43. **Solmone, M., et al.** 2009. Use of massively parallel ultradeep pyrosequencing to characterize the genetic diversity of hepatitis B virus in drug-resistant and drug-naive patients and to detect minor variants in reverse transcriptase and hepatitis B S antigen. J. Virol. **83:**1718–1726.

44. **Thebaud, G., J. Chadoeuf, M. J. Morelli, J. W. McCauley, and D. T. Haydon.** 2010. The relationship between mutation frequency and replication strategy in positive-sense single-stranded RNA viruses. Proc. Biol. Sci. **277:**809–817.

45. **Tsibris, A. M., et al.** 2009. Quantitative deep sequencing reveals dynamic HIV-1 escape and large population shifts during CCR5 antagonist therapy in vivo. PLoS One **4:**e5683.

46. **Victoria, J. G., et al.** 2010. Viral nucleic acids in live-attenuated vaccines: detection of minority variants and an adventitious virus. J. Virol. **84:**6033–6040.

47. **Wang, C., Y. Mitsuya, B. Gharizadeh, M. Ronaghi, and R. W. Shafer.** 2007. Characterization of mutation spectra with ultra-deep pyrosequencing: application to HIV-1 drug resistance. Genome Res. **17:**1195–1201.

48. **Wang, G. P., S. A. Sherrill-Mix, K. M. Chang, C. Quince, and F. D. Bushman.** 2010. Hepatitis C virus transmission bottlenecks analyzed by deep sequencing. J. Virol. **84:**6218–6228.