

1     **Characterization of HCV envelope diversification from acute to chronic infection using**  
2             **SMRT sequencing within a sexually-transmitted hepatitis C virus cluster**

3

4     Cynthia K.Y Ho<sup>1</sup>, Jayna Raghvani<sup>2</sup>, Sylvie Koekkoek<sup>1</sup>, Richard H Liang<sup>3</sup>, Jan TM Van der  
5     Meer<sup>4</sup>, Marc Van Der Valk<sup>4</sup>, Menno De Jong<sup>1</sup>, Oliver G Pybus<sup>2</sup>, Janke Schinkel<sup>1§</sup> and  
6     Richard Molenkamp<sup>1§#</sup>-- on behalf of the MOSAIC (MSM Observational Study of Acute  
7     Infection with hepatitis C) study group

8

9     <sup>1</sup>Department of Medical Microbiology, Academic Medical Center, Amsterdam, the  
10    Netherlands

11    <sup>2</sup>Department of Zoology, University of Oxford, Oxford, United Kingdom

12    <sup>3</sup>BC Centre for Excellence in HIV/AIDS, Vancouver, BC, Canada

13    <sup>4</sup>Department of Infectious Diseases, Academic Medical Center, Amsterdam, the  
14    Netherlands

15    § These authors contributed equally

16

17    **RUNNING TITLE:** SMRT sequencing of serial sampled HCV

18

19    # Address correspondence to Richard Molenkamp, r.molenkamp@amc.uva.nl.

20

21    **Word count abstract:** 224 (max 250)

22    **Word count importance:** 147 (max 150)

23 **Word count text:** 6204

24 **ABSTRACT**

25 In contrast to other available next generation sequencing platforms, Pacbio Single  
26 Molecule, Real-Time (SMRT) sequencing has the advantage of generating long reads,  
27 albeit with a relatively higher error rate in unprocessed data. Using this platform we  
28 longitudinally sampled and sequenced the hepatitis C virus (HCV) envelope genome  
29 region (1680 nt) from individuals belonging to a cluster of sexually-transmitted cases.  
30 All five subjects were HIV-1 coinfecting and infected with a closely related strain of HCV  
31 genotype 4d. In total 50 samples were analyzed using SMRT sequencing. By using 7  
32 passes of circular consensus sequencing the error rate was reduced to 0.37% and the  
33 median number of sequences was 612 per sample. Further reduction of insertions was  
34 achieved by aligning against a sample-specific reference sequence. However, in vitro  
35 recombination during PCR amplification could not be excluded. Phylogenetic analysis  
36 supported close relationships among HCV sequences from the four male subjects and  
37 the subsequent transmission from one subject to his female partner. Transmission was  
38 characterized by a strong genetic bottleneck. Viral genetic diversity was low during  
39 acute infection, increased upon progression to chronicity, but subsequently fluctuated  
40 during chronic infection, caused by alternate detection of distinct co-existing lineages.  
41 SMRT sequencing combines long reads with sufficient depth for many phylogenetic  
42 analyses, and can therefore provide insights into within-host HCV evolutionary dynamics  
43 without the need for haplotype reconstruction using statistical algorithms.

44

## **IMPORTANCE**

Next generation sequencing has revolutionized the study of genetically variable RNA virus populations, but for phylogenetic and evolutionary analyses longer sequences than generated by most available platforms are desired, while minimizing the intrinsic error rate. Here, we demonstrate for the first time that Pacbio SMRT sequencing technology can be used to generate full-length HCV envelope sequences at the single molecule level, providing a dataset with large sequencing depth for characterization of intra-host viral dynamics. Selecting consensus reads derived from at least 7 full circular consensus sequencing rounds significantly reduced the intrinsic high error rate of this method. We used this method to genetically characterize a unique transmission cluster of sexually transmitted HCV infections, providing insight in the distinct evolutionary pathways in each patient over time, identifying the transmission-associated genetic bottleneck, as well as fluctuations in viral genetic diversity over time, accompanied by dynamic shifts in viral subpopulations.

## INTRODUCTION

Phylogenetic analysis plays a central role in the molecular epidemiology of rapid evolving RNA viruses, such as HIV, HCV and influenza A viruses (1–3) and requires sequencing methods that can accurately capture the genetic diversity of RNA virus genomes. Past studies have relied heavily on Sanger based sequencing, which is still widely used today (4). This first generation sequencing technology is comparatively labour intensive and costly, and provides a relatively a low number of sequences.

Advances in high throughput sequencing technologies during the last decade have substantially increased the number of pathogen gene and genome sequences available. In particular, for rapidly evolving RNA viruses, next generation sequencing (NGS) technologies offer new and more detailed insights into the highly heterogeneous nature of viral populations (5). For instance, NGS technologies have been used to investigate the co-receptor usage of HIV (6) and transmission bottlenecks in HCV infection (7). Viral genome sequencing by NGS has been dominated by the 454 Sequencing and Illumina technologies. One disadvantage of these technologies, in comparison to first generation Sanger sequencing, is the short sequence lengths obtained and the higher sequencing error rate. The generally accepted error rate for these NGS technologies is 1% and sequence lengths are up to 300 nt long for Illumina (8) and 700 nt for 454 Sequencing (9). It is difficult to reconstruct virus phylogenetic relationships from short sequences because they lack complete information about the genetic linkage of polymorphic sites, which reduces the statistical power of many methods for inferring evolutionary processes from genetic data. High error rates, which vary among different NGS

platforms, may lead to biases in estimates of evolutionary parameters because true biological variation cannot always be easily distinguished from sequencing errors (10).

Bioinformatics algorithms can identify and correct for NGS errors to a certain extent and it is common for some form of data pre-processing to be applied for error reduction prior to analysis (11). Longer sequences can be estimated by clustering short overlapping reads into haplotypes (12). However, this haplotype reconstruction from genetically diverse short read sequences is a difficult statistical problem and is sensitive to false positives and may fail to reconstruct minor viral variants, which are especially abundant in large and heterogeneous virus populations (13, 14)

Pacbio Single Molecule Real-Time (SMRT) sequencing, which records the incorporation of nucleotides into a single DNA template molecule by an immobilized DNA polymerase, offers an alternative to other NGS technologies due to its long read length with a mean of 2-3 kb (15, 16). However, the throughput of SMRT sequencing is lower compared to other NGS platforms and the error rate of raw, unprocessed sequences is reported to be 11% -14% (17–19). Read accuracy can be improved greatly by circular consensus sequencing (CCS) in which the DNA template is circularized, allowing multiple sequencing passes and resulting in multiple reads for a single template (18). A consensus sequence can be generated from these multiple reads that eliminates most of the sequencing errors, since these are randomly distributed along the read sequence (17).

Although the combination of long reads and reduction of the error rate by using CCS has great potential for studies addressing the variation of highly heterogeneous RNA

genomes, very few studies have been published utilizing SMRT sequencing for this purpose since its commercial release in 2011 (20–23). Here, we present an analysis of HCV full-length envelope (E1E2) gene sequences from clinical longitudinal samples obtained with SMRT sequencing and CCS. We investigated the variability of the HCV envelope gene through time, from acute infection to chronicity, in four men and the female partner of one of the subjects, all of who acquired HCV through sexual transmission. All subjects were infected with a genetically closely related genotype 4d strain and were co-infected with HIV-1 (24). This study provides a unique opportunity to study diversification of HCV in individuals as they progress from early acute infection to more than a decade of chronic infection.

## **MATERIAL AND METHODS**

### **Subjects and sample selection**

The four male subjects were selected on the basis of a close genetic relationship between their HCV genotype 4d viruses, as observed in a previous study (24). The fifth female subject was enrolled because personal communication with her physician revealed that she and subject 004 were a couple and formed a possible transmission pair. Two subjects were participants of the MSM Observational Study of Acute Infection with hepatitis C (MOSAIC) study and one was a participant of the Amsterdam Cohort Studies (ACS) (25, 26). Ethical approval was granted through these cohorts. The study was performed according to the Dutch FEDERA code of conduct for responsible use of human tissue and medical research 2011. Samples were selected as follows; within the

first year of infection at least 4 samples were selected, starting with the first available HCV RNA+ sample. After the first year, one sample for each year of infection was selected. The date of infection was estimated by calculating the mid point between the date of last HCV RNA- and the first RNA+ sample.

### **RT-PCR and SMRT sequencing**

HCV RNA was extracted with QIAamp viral RNA minikit (Qiagen) from 140 µl plasma or serum. Reverse transcription (RT) reaction was performed with poly-A primers to target the poly-U stretch in the HCV 3'UTR and Superscript III (Invitrogen). Target specific PCR primers forward 840F: 5'-acactgacgacatggttctacaGGTTGCTCYTTYTCTATCTTCC-3' with in lower case the Pacbio tail adaptors and reverse 2579R: 5'-tacggtagcagagacttggtctGCYTCGACCTGCGMAACCA-3' and 2579R(2): 5'-tacggtagcagagacttggtctCGCCTCRACYTGASNYACCA-3' were used for the amplification using FastStart High Fidelity PCR System (Roche Applied Science). The final amplification product was 1778 nt and spanned the full length of the E1E2 region. In cases of insufficient amplification, a nested PCR was performed with outer primer set forward primer GEN1.F1: 5'- CAAGACTGCTAGCCGAGTAGTGTGGGTCG-3' and reverse primer GEN4.R1: 5'-TCGGGCAYGRGACAYGCTGTGATAAATG-3' and the same inner primer set as described above. Amplification products were purified from 0.8% agarose gel.

A control plasmid was constructed from one genotype 4d clone using the same primer set as that described above. One aliquot of this amplification product was directly sequenced using standard techniques on a capillary DNA-sequencing instrument (Applied Biosystems). The other aliquot was prepared for SMRT sequencing.

150

151 **Single molecule real time sequencing**

152 The purified amplification products were sent to KeyGene N.V., Wageningen, The  
153 Netherlands for SMRTbell library preparation and sequencing according to PacBio RS II  
154 manufacturer's protocols (Pacific Biosciences). Briefly, barcode sequences were  
155 incorporated by PCR and the products were AMPure (Beckman Coulter) purified. End  
156 repair and ligation of universal hairpin adaptors was performed. Four SMRT cells were  
157 used to sequence 50 amplicon libraries from the subjects and in each SMRT cell the  
158 amplification product of the control plasmid was included. SMRT sequencing was  
159 performed using the P4-C2 chemistry. The raw reads were separated by barcode and  
160 CCS reads from 1 to 7 full passes were generated using the SMRT analysis tools v 2.2.0.

161

162 **Analysis of *in vitro* recombination**

163 To investigate the potential for recombination during reverse-transcription and PCR, we  
164 constructed two plasmids with 5 nt marker mutations consisting of two substitutions  
165 and three insertions at either the 5' (original sequence CC---; mutant sequence AATGT)  
166 or 3' (original sequence AAGTG---; mutant sequence CCGTGGAT ) end of an otherwise  
167 identical E1E2 sequence (1687 nt). In vitro RNA was synthesized by T7 transcription  
168 using the Megascript T7 kit (Thermo Fisher) according to the instructions of the  
169 manufacturer. In vitro transcripts were DNaseI treated, purified and quantified by  
170 absorbance spectrophotometry. A mixture of  $1 \times 10^6$  RNA copies of each mutant was  
171 subjected to essentially the same (first round) amplification protocol as RNA derived



from patient plasma samples, with the exception of the use of the HCV-specific 2579R primer (5'-TGCTCGACCTGCGMAACCA-3') for reverse-transcription. The amplicons with the marker mutations were SMRT sequenced in a separate run using identical parameters as used for the patient's samples. Recombination was inferred if a sequence combined both the 3'- and the 5'- marker mutations or both the original sequences. The marker mutations were detected using bash scripts.

#### **Circular consensus read processing**

Sequencing error rates (insertions, deletions and mismatches) per CCS full pass were calculated by comparing the CCS reads of the control plasmid with the sequence from Sanger sequencing. Because insertions distorted the nucleotide sequences, we developed a pipeline to remove these. The processing pipeline was constructed using Python scripts, Biopython and the R Bioconstructor package Biostrings (27, 28) and is available from the authors on request. Briefly, CCS reads with at least 7 full passes in the reverse-complement orientation were converted to the forward orientation. Then a sample specific reference (SSR) sequence was generated by aligning the reads of each sample using MUSCLE (29) and columns containing >50% gaps were deleted. Consensus sequences were generated from this alignment and, if needed, manually edited to obtain the correct reading frame. Each read from a given sample was then pairwise aligned against its corresponding SSR and insertions with respect to this SSR were removed. The resulting reads were 1680 nt long (primers trimmed and in frame) and spanned nucleotide positions 876-2558, relative to H77 reference strain (Genbank

accession number NC\_004102). Translated amino acid sequences were manually inspected using Sea View (30). Reads with a remaining distorted open reading frame were removed (these comprised fewer than 9 reads per sample). The read frequencies of each unique nucleotide sequence, which we will refer to as read variants, were calculated and stored in the read name.

### **Phylogenetic methods**

Neighbor Joining (NJ) trees were reconstructed using the Kimura 2-parameter model plus gamma rate heterogeneity as implemented in MEGA v6 (31). Subsets of the full data set were created by randomly sampling 100 CCS reads without replacement from each time point from each subject. For the among-host NJ phylogeny (Fig. 2) all subsets were combined to create a dataset containing 5000 CCS reads. Statistical support for phylogenetic topology was assessed using 100 NJ bootstrap replicates. For each subject a separate NJ phylogeny was reconstructed from sequences from all time points; if CCS reads from the same time point were identical in sequence then those were collapsed into a unique sequence. Phylogenetic trees were annotated using FigTree v1.4 (<http://tree.bio.ed.ac.uk>) and custom-made Java scripts.

### **Pairwise Distance methods**

To assess the nucleotide diversity at each time point for each subject, the pairwise distance among reads was calculated using the R package ape (32) under the Kimura 2-parameter substitution model.

217

## 218 **AVAILABILITY OF DATA AND MATERIALS**

219 The E1E2 CCS reads have been submitted to the Sequence Read Archive database  
220 (<http://www.ncbi.nlm.nih.gov/sra/>) and can be found under accession numbers xx  
221 through xx. *Will be available once the manuscript is accepted.*

222

## 223 **RESULTS**

### 224 **Subjects and samples characteristics**

225 The study subjects included four men who have sex with men (MSM) and one female, all  
226 of whom were infected with HIV-1 prior to acquiring infection with HCV subtype 4d  
227 (Table 1). The ages of the subjects ranged from 49 to 61 years. All had documented  
228 acute HCV infection, as the duration between the last RNA- and the first RNA+ sample  
229 was <7 months for each. In addition, subjects 004, 037 and 061 seroconverted within  
230 the period of HCV acquisition, or after being tested positive for HCV RNA, indicating that  
231 these are primary infections. The seroconversion date for subject 147 was unknown,  
232 however that subject's anti-HCV test was negative at the date of the first HCV RNA  
233 positive test. For subject 053, the seroconversion interval was almost 5 years, as a more  
234 recent sample to determine the last negative anti-HCV status was not tested. Estimated  
235 infection dates, calculated as the midpoint between the last RNA- and the first RNA+  
236 test, were between 2001 and 2002 for the four male subjects, and was earliest for  
237 subject 053, followed by subjects 061 and 037, and lastly by subject 004. The men  
238 developed chronic HCV infection and had a follow-up period ranging from 7 to 11 years.

239 Subjects 053 and 061 received treatment during chronic infection with pegylated  
240 interferon and ribavirin; subject 053 cleared the infection after 48 weeks of treatment  
241 whilst subject 061 was a non-responder and discontinued treatment after 28 weeks.  
242 Subject 147 is the female partner of subject 004 and was diagnosed with acute HCV  
243 infection in 2013; she cleared the infection within 6 months following treatment with  
244 pegylated interferon and ribavirin.

245 A total of 63 stored sera, collected between 2001 and 2014, were selected from  
246 these five subjects. PCR amplification of the envelope gene was not successful in 13  
247 samples. For the remaining 50 samples a fragment of 1680 nt spanning the full length  
248 HCV envelope was analyzed using SMRT sequencing (see Methods).

249

#### 250 **Sequencing error rate and sequencing depth**

251 The error rate of the SMRT sequencing process was averaged over the four control  
252 plasmids that were processed in parallel with the subjects' samples. Error rates were  
253 calculated for each CCS full pass (1-7) (Fig. 1a). One full pass is defined as completion of  
254 sequencing of at least one DNA template strand and two full passes is defined as  
255 sequencing of at least one round of the double stranded template. As expected,  
256 sequencing errors were efficiently reduced with each increase in the number of CCS full  
257 passes. With  $\geq 1$  CCS full pass the mean error rate was 1.57% (SD = 0.016%). This  
258 decreased fivefold to 0.37% (SD = 0.023%) at  $\geq 7$  CCS full passes, with insertions being  
259 the main mode of error (0.24%, SD = 0.036%), followed by deletions (0.11%, SD =

0.022%). Mismatch rates after 7 CCS full passes were surprisingly rare (0.02%, SD = 0.005%).

The median sequencing depth across all 50 samples decreased approximately linearly with each increase in the number of CCS full passes. The median sequencing depth decreased from 2025 reads per sample (interquartile range (IQR): 1582 - 2380) for one CCS full pass to 612 (IQR: 498 - 739) for seven CCS full passes (Fig. 1b). Because further increases in the number of full passes would not lead to a significant reduction of the error rate (Fig 1a), subsequent analyses were undertaken on the CCS reads from full pass 7.

Insertions were randomly distributed across the sequence. The insertion rate after 7 CCS full passes was 0.24%, which resulted in ~4 insertions per read within the 1680 nt long sequences being generated. These insertions caused numerous frameshifts and stop codon errors in the open reading frame but were easily identified and removed by the bioinformatics pipeline (see Methods). After removal of insertions, the proportion of gaps and stop codons in the amino acid alignment was 0.31% (SD=0.068%) and 0.006% (SD=0.0036%), respectively. The majority of these were found in the hypervariable 1 (HVR1) region.

### **Among-host phylogenies of the five subjects**

The phylogenetic relationships among the virus sequences obtained from the five subjects was evaluated from a Neighbor Joining (NJ) phylogeny reconstructed from a random selection of 100 reads per sample (Fig. 2). The results were consistent with prior

expectations, as each of the four males were infected with genetically closely related viruses and the virus population in each subject diverged from its common ancestor over the course of chronic infection (24). The CCS reads clustered by subject, however, only subjects 147 and 037 form monophyletic clusters. Reads from subject 004 were non-monophyletic with respect to those from subject 147, indicating direct (or indirect, via an unsampled individual) transmission of HCV from the former to the latter (Fig. 2). Reads from subject 037 formed a monophyletic ingroup within the diversity of reads from subjects 061 and 053, suggesting direct or indirect transmission to the former from one of the latter (Fig. 2). Reads from subjects 061 and 053 were intermingled and therefore the direction of transmission between these two subjects (directly, or via one or more unsampled individuals) cannot be resolved.

#### **Within host phylogenies of each subject**

Figure 3 shows the within-host NJ phylogeny of subject 061, using all CCS reads for this subject obtained across 18 time points. NJ trees for the four other subjects are presented in supplemental material Fig. S1-4. Interestingly, in all five subjects, the viral populations at the first RNA+ time point were genetically homogeneous and exhibited a star-like phylogeny (see Fig. S5 in the supplemental material), consistent with rapid viral population growth from a single successful transmitted virion (7, 33–35). In general, progression towards chronicity was characterized by the replacement of early time point lineages by new variants, such that after 2 to 3 years the former could not be detected anymore. Further, chronic infection was characterized by temporary co-

circulation of multiple different viral lineages. Such lineages may remain undetected for a few time points, only to re-emerge at later time points (36–41). This complex dynamic of re-emerging lineages is represented by the presence of long internal branches and was especially seen at the later time points of subjects 004, 037 and 061 (see Fig. S1-3 in the supplemental material and Fig. 3).

The viral population of the non-responder subject 061 around the time of anti-HCV treatment is of particular interest. At the start of treatment (time point 2009.07) at least six distinct lineages were present. After two months of treatment (time point 2009.25), a new lineage replaced the population and in the proceeding months, more lineages emerged that were not observed before. Since HCV RNA was still detectable after six months, treatment was discontinued, resulting in the restoration of pre-treatment lineages (time point 2008.95). These lineages co-existed with those lineages that emerged during treatment for the remaining time points (Fig. 3). In contrast to subject 061, subject 053 successfully cleared the virus with anti-HCV treatment and his population at later time points showed less diversification, indicated by shorter internal branch lengths. Furthermore, only one lineage was detected per time point (see Fig. S3 in the supplemental material).

### **Genetic diversity during infection**

The pairwise genetic distances between the CCS reads at each time point were used to measure the genetic diversity of the within-subject HCV population over time (Fig. 4a-e). HCV genetic diversity varied among subjects and fluctuated among time points within

subjects. The median genetic diversity per time point for HCV from subject 053 ranged from 0.00067 to 0.0035 substitutions per site. By contrast, HCV from subject 061 (treatment non-responder) had high median genetic diversity values, ranging from 0.00069 to 0.014 substitutions per site and that from subject 004 ranged from 0.0007 to 0.015 substitutions per site. The median diversity for subject 037 ranged from 0.00068 to 0.0085 substitutions per site.

One pattern shared by all subjects was a low genetic diversity during early acute infection, as exemplified by subject 147 (Fig. 4c). The three time points analyzed for this subject were within 1, 5 and 7 months after the estimated infection date, and median HCV genetic diversity during this time ranged from 0.00069 to 0.002 substitutions per site. Similar values were observed for other subjects during acute infection. Thus, the transition to chronic infection among the four MSM was accompanied by an increase in viral genetic diversity.

Another common pattern among subjects was the fluctuation in HCV genetic diversity during chronic infection, which has been observed in other longitudinally sampled HCV patients (36, 40). To verify that procedural artifacts did not cause these fluctuations, clonal sequences of an E2 560 nt fragment of three samples were analyzed (data not shown). The IQR of the genetic diversity from the clonal sequences overlapped with those from the CCS reads, indicating that these fluctuations were not artifacts.

#### ***In vitro* recombination analysis**



Homology-mediated recombination during reverse transcription and PCR has the potential to influence the results of gene sequence analysis. This has become evident with the large number of sequences generated by NGS techniques (22). In an effort to estimate the level of homologous recombination in a similar experimental setup, we performed a marker-exchange experiment. A recombination event in this experiment could result in a chimeric sequence containing either both 5'- and 3'- marker mutations or both 5'- and 3'- unmutated sequences. In total 10,541 reads with at least 7 full passes and the correct 5'- and 3'- sequences were retrieved. In 3,137 reads (29.8 %) evidence for recombination was observed. In 11.9% of the reads the two marker mutations were combined and in 17.9% of the reads the two original sequences were combined. Despite using identical parameters as with the patient's samples datasets, the error rate was slightly higher (1.3%) underlining the run to run variation in NGS experiments.

## **DISCUSSION**

Longer read lengths, greater sequencing depth and lower error rates increase the potential of studies addressing the variation of highly heterogeneous RNA viral populations. This study demonstrates the feasibility and suitability of Pacbio SMRT sequencing for generating full length HCV envelope sequences. The high intrinsic error rate of the platform can be reduced by including only consensus sequences inferred using 7 or more CCS full passes, by applying a downstream bioinformatics pipeline that corrects for insertions, and by taking potential high in vitro recombination frequency into consideration. Phylogenetic analysis confirmed a close genetic relationship among

the genotype 4d viruses isolated from the five studied subjects. In addition, in each subject we observed a star-like phylogeny shortly after transmission, low genetic diversity during acute infection, and fluctuations in sample genetic diversity and transient lineages co-existence during chronic infection

The E1E2 viral envelope, harboring the HVR1, is the most genetically variable region of the HCV genome (42). The high error rate of SMRT sequencing makes it especially challenging to distinguish true biological variation from sequencing errors in this region, even with sophisticated bioinformatics algorithms (10). Multiple passes of CCS effectively reduced the error rate, but came at the cost of lower numbers of complete sequences. For the downstream analysis, we used CCS reads inferred from at least 7 CCS full passes, which generated a median of 612 reads per sample. Despite this depth being in the lower range of most NGS technologies, it was sufficient to capture the within host dynamics of viral genetic diversity and divergence in great detail. Moreover, with this number of full passes, the error rate was 0.37%, which is well beneath the general accepted error rate of 1% for NGS technologies (19).

Insertions were the main mode of error in all CCS full passes, followed by deletions and mismatches, respectively. A similar error profile has been reported by others using the SMRT sequencing technology (17, 20). Mismatches were our main concern, as these could lead to an overestimation of the number of genetic differences among sequences. However, with 7 CCS full passes the mismatch rate was 0.02%, implying that in theory only 2 out of 10000 sequenced nucleotides would be incorrect. Even though the insertion rate was low (0.24%) at 7 or more CCS full passes, parts of the unprocessed

alignments were distorted, which would have obstructed the downstream analysis. It was therefore necessary to correct for these insertions, which we did by using a sample-specific reference, a well-established method utilized by others (17, 20, 22, 43).

Despite using a sample-specific reference sequence and exploring several alignment algorithms and scoring parameters (data not shown), we recognize that errors may still be introduced into the CCS read sequence if the alignment algorithm mistakes a real SNP for an insertion, resulting in removal of the wrong base, a problem also described by Carneiro et al. (17). We investigated the scope of this error by inspecting the amino acid alignment and found that the majority of gaps and stop codons in the amino acid alignments were within the HVR1 region. Due to the extensive variation in the HVR1, combined with the amount of insertions, it might not be feasible to increase the accuracy of the alignment for this region (44). However, frequencies of the gaps and stop codons were extremely low (0.31% and 0.006%, respectively) and therefore unlikely to have meaningfully affected downstream analysis.

Although a recent study which sequenced the envelope gene of two different HIV clones estimated that PCR recombination rate was only 0.46-0.85% (45), in vitro RT and PCR recombination rates of ~ 30% have been reported previously (22, 46). In our experimental setup we found a similarly high recombination rate. The experimental conditions in the recombination control experiment were similar to those for the patient sample dataset, but there were small differences that may have influenced the recombination efficiency. First, the primer used for the reverse transcription reaction was different, since the primer used for the patients samples was located in the extreme

3'-end of the viral genome, a region not present in our control plasmids. A second difference was that in the control experiment there were only two almost identical sequences with marker mutations in the extreme 5'- and 3'-ends. This 100% identity over a length of 1687 nucleotides in all available RNA templates could be more favorable for homology-mediated recombination.

Whatever the true rate of recombination was during our sequencing of patient samples, it does not appear to have significantly influenced the resulting phylogenetic and evolutionary conclusions. Importantly, we observed key patterns of HCV within-host evolution that have been noted by previous studies using different sequencing platforms. This includes the detection of multiple lineages during chronic infection, which are intermittently observed. Moreover, the within-host phylogenies and the fluctuations in genetic diversity in each patient closely agree with a recent comparative analysis of within-host HCV and HIV evolution (36). Also, the known transmission events are inferred correctly and the TMRCAs match the known dates of infection well (data not shown). The genetic variation in each sample might be overestimated due to artificially generated recombinant sequences, but the overall conclusion that shortly after transmission there is close genetic relationship between the viruses isolated from these five subjects and the diversification of the E1E2 region during progression to chronicity remains the same.

The among-host phylogeny (Fig. 3) supports our previous findings, observed using direct sequencing of a smaller fragment of the E2 gene, of a cluster of closely related viruses among the four MSM subjects (24). The additional time points analyzed here

demonstrate the diversification of the E1E2 region in each of the four subjects, reflecting the selective pressure of host immune responses on the envelope sequences (47). Further, the intermingling of sequences from subjects 053 and 061 during early infection time points was not previously detected with direct sequencing of E2 gene fragments. Although the branch support values for clustering was low, likely caused by the genetic similarity of reads during early time points, this intermingling was not observed for other subjects and other time points and supports the hypothesis that subjects 053 and 061 form a transmission pair. Despite the fact that the NJ phylogeny on its own cannot resolve the direction of transmission, nor account for participation of unsampled subjects (48), the transmission event from the male subject 004 to his female partner 147 was confirmed by the physician's report. Although documented heterosexual transmission of HCV is rare (49), transmission among these two subjects may have been facilitated by the fact that the couple was HIV-1 positive and reported anal intercourse. These factors may have increased the risk of heterosexual transmission of HCV.

Compared to HIV-1, HCV transmitted founder viruses have been less intensively investigated but, similar to HIV-1, they may have important implications for vaccine design and therapeutic interventions (33, 35, 50). A star-like phylogeny and low genetic diversity among strains sampled early in infection indicate a transmission bottleneck and that one transmitted founder virus established the infection (43). Other studies have reported either low numbers of transmitted founder viruses (between 1 and 6; (7, 34, 51, 52)) or high numbers (up to 37; (33)). Limited sequencing resolution (i.e. low

variability in the sequenced region, or short sequences) could potentially underestimate the number of transmitted founder viruses. Here, even with comparatively long HCV envelope sequences and greater sequencing depth, our results suggest that in all five subjects only one transmitted founder virus established infection.

In general, progression towards chronic infection was characterized by an increase in genetic diversity, divergence and the number of co-circulating lineages, but these values varied widely among subjects. Early lineages that emerged shortly after transmission were replaced within 2-3 years, which may be explained by immune pressure (e.g. neutralizing antibodies) causing dynamic changes in the viral population (53, 54). New lineages that emerged at different time points, a pattern that was observed throughout chronic infection, could reflect ongoing viral escape from neutralizing antibodies (47, 55, 56). However, some lineages persisted but were undetected for certain periods in time. Cell-to-cell transmission would allow HCV to spread without being exposed to humoral immune responses and this provides a possible explanation why lineages co-exists during HCV infection without going extinct (57, 58). The presence of these lineages resulted in a structured phylogeny containing long internal branches. This complex pattern of within-host envelope gene evolution during chronic HCV infection supports the findings of previous studies (40, 59, 60) that have used sequencing technologies with a lower depth. Here we show that even with a greater sequencing depth that not all HCV lineages are detected at all times which is consistent with the hypothesis of compartmentalization in the liver or in extra-hepatic tissues, giving rise to distinct HCV subpopulations (61–65).

479

## 480 **CONCLUSIONS**

481 Greater numbers of sequences and longer read lengths can significantly increase the  
482 resolution and statistical power of evolutionary and molecular epidemiological analyses  
483 of highly variable RNA virus genomes, such as those of HCV. When appropriate  
484 bioinformatics quality control is used and potential *in vitro* recombination is taken into  
485 account, SMRT sequencing, can provide both and enhances insight into the within- and  
486 among-host dynamics of HCV.

487

## 488 **FUNDING INFORMATION**

489 This work was supported by the Virgo consortium, funded by the Dutch government  
490 (project number 050-060-452). This work was funded by the European Research Council  
491 under the European Union's Seventh Framework Programme (FP7/2007-2013)/ERC  
492 grant agreement no. 614725-PATHPHYLODYN.

493

## 494 **ACKNOWLEDGEMENTS**

495 We kindly thank Alexander Wittenberg Keygene (Keygene B.V) and Christoph Koenig  
496 (Pacific Biosciences) for their help during the SMRT sequencing process. We  
497 acknowledge the Amsterdam Cohort Studies on HIV infection and AIDS, a collaboration  
498 between the Public Health Service of Amsterdam, the Academic Medical Center of the  
499 University of Amsterdam, Sanquin Blood Supply Foundation, Medical Center Jan van  
500 Goyen and the HIV Focus Center of the DC-Clinics, are part of the Netherlands HIV

501 Monitoring Foundation and financially supported by the Center for Infectious Disease  
502 Control of the Netherlands National Institute for Public Health and the Environment. In  
503 addition, we would like to thank the MOSAIC participants.

504



## REFERENCES

1. **Ou CY, Ciesielski C a, Myers G, Bandea CI, Luo CC, Korber BT, Mullins JI, Schochetman G, Berkelman RL, Economou a N.** 1992. Molecular epidemiology of HIV transmission in a dental practice. *Science* **256**:1165–1171.
2. **Bhatt S, Lam TT, Lycett SJ, Brown AJL, Bowden TA, Holmes EC, Guan Y, Wood JLN, Brown IH, Kellam P, Swine C, Consortium I, Pybus OG, Aj LB, Ta B, Ec H, Jln W, Ih B.** 2013. The evolutionary dynamics of influenza A virus adaptation to mammalian hosts.
3. **Pybus OG, Charleston M a, Gupta S, Rambaut a, Holmes EC, Harvey PH.** 2001. The epidemic behavior of the hepatitis C virus. *Science* **292**:2323–5.
4. **Sanger F, Nicklen S, Coulson a R.** 1977. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A* **74**:5463–5467.
5. **Metzker ML.** 2010. Sequencing technologies - the next generation. *Nat Rev Genet* **11**:31–46.
6. **Archer J, Rambaut A, Taillon BE, Harrigan PR, Lewis M, Robertson DL.** 2010. The Evolutionary Analysis of Emerging Low Frequency HIV-1 CXCR4 Using Variants through Time—An Ultra-Deep Approach. *PLoS Comput Biol* **6**:e1001022.
7. **Bull R a., Luciani F, McElroy K, Gaudieri S, Pham ST, Chopra A, Cameron B, Maher L, Dore GJ, White P a., Lloyd AR.** 2011. Sequential Bottlenecks Drive Viral Evolution in Early Acute Hepatitis C Virus Infection. *PLoS Pathog.*
8. **Bennett S.** 2004. Solexa Ltd. *Pharmacogenomics* **5**:433–8.
9. **Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J,**

527 **Braverman MS, Chen Y-J, Chen Z, Dewell SB, Du L, Fierro JM, Gomes X V, Godwin**  
 528 **BC, He W, Helgesen S, Ho CH, Irzyk GP, Jando SC, Alenquer MLI, Jarvie TP, Jirage**  
 529 **KB, Kim J-B, Knight JR, Lanza JR, Leamon JH, Lefkowitz SM, Lei M, Li J, Lohman**  
 530 **KL, Lu H, Makhijani VB, McDade KE, McKenna MP, Myers EW, Nickerson E,**  
 531 **Nobile JR, Plant R, Puc BP, Ronan MT, Roth GT, Sarkis GJ, Simons JF, Simpson**  
 532 **JW, Srinivasan M, Tartaro KR, Tomasz A, Vogt KA, Volkmer GA, Wang SH, Wang**  
 533 **Y, Weiner MP, Yu P, Begley RF, Rothberg JM.** 2005. Genome sequencing in  
 534 microfabricated high-density picolitre reactors. *Nature* **437**:376–80.

535 10. **Kunin V, Engelbrektson A, Ochman H, Hugenholtz P.** 2010. Wrinkles in the rare  
 536 biosphere: pyrosequencing errors can lead to artificial inflation of diversity  
 537 estimates. *Environ Microbiol* **12**:118–23.

538 11. **Quince C, Lanzen A, Davenport RJ, Turnbaugh PJ.** 2011. Removing Noise From  
 539 Pyrosequenced Amplicons. *BMC Bioinformatics* **12**:38.

540 12. **Zagordi O, Bhattacharya A, Eriksson N, Beerenwinkel N.** 2011. ShoRAH:  
 541 estimating the genetic diversity of a mixed sample from next-generation  
 542 sequencing data. *BMC Bioinformatics* **12**:119.

543 13. **Schirmer M, Sloan WT, Quince C.** 2014. Benchmarking of viral haplotype  
 544 reconstruction programmes: an overview of the capacities and limitations of  
 545 currently available programmes. *Brief Bioinform* **15**:431–42.

546 14. **Ho CKY, Welkers MRA, Thomas X V., Sullivan JC, Kieffer TL, Reesink HW, Rebers**  
 547 **SPH, de Jong MD, Schinkel J, Molenkamp R.** 2015. A comparison of 454  
 548 sequencing and clonal sequencing for the characterization of hepatitis C virus NS3

549 variants. *J Virol Methods* **219**:28–37.

550 15. **Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, Peluso P, Rank D, Baybayan P,**  
551 **Bettman B, Bibillo A, Bjornson K, Chaudhuri B, Christians F, Cicero R, Clark S,**  
552 **Dalal R, DeWinter A, Dixon J, Foquet M, Gaertner A, Hardenbol P, Heiner C,**  
553 **Hester K, Holden D, Kearns G, Kong X, Kuse R, Lacroix Y, Lin S, Lundquist P, Ma C,**  
554 **Marks P, Maxham M, Murphy D, Park I, Pham T, Phillips M, Roy J, Sebra R, Shen**  
555 **G, Sorenson J, Tomaney A, Travers K, Trulson M, Vieceli J, Wegener J, Wu D,**  
556 **Yang A, Zaccarin D, Zhao P, Zhong F, Korlach J, Turner S.** 2009. Real-Time DNA  
557 Sequencing from Single Polymerase Molecules. *Science* (80- ) **323**:133–138.

558 16. **Koren S, Harhay GP, Smith TPL, Bono JL, Harhay DM, Mcvey SD, Radune D,**  
559 **Bergman NH, Phillippy AM.** 2013. Reducing assembly complexity of microbial  
560 genomes with single-molecule sequencing. *Genome Biol* **14**:R101.

561 17. **Carneiro MO, Russ C, Ross MG, Gabriel SB, Nusbaum C, DePristo M a.** 2012.  
562 Pacific biosciences sequencing technology for genotyping and variation discovery  
563 in human data. *BMC Genomics* **13**:375.

564 18. **Travers KJ, Chin C-S, Rank DR, Eid JS, Turner SW.** 2010. A flexible and efficient  
565 template format for circular consensus sequencing and SNP detection. *Nucleic*  
566 *Acids Res* **38**:e159.

567 19. **Quiñones-Mateu ME, Avila S, Reyes-Teran G, Martinez M a.** 2014. Deep  
568 sequencing: Becoming a critical tool in clinical virology. *J Clin Virol* **61**:9–19.

569 20. **Archer J, Baillie G, Watson SJ, Kellam P, Rambaut A, Robertson DL.** 2012.  
570 Analysis of high-depth sequence data for studying viral diversity: a comparison of

571 next generation sequencing platforms using Segminator II. BMC Bioinformatics  
572 **13:47.**

573 21. **Brinzevich D, Young GR, Sebra R, Ayllon J, Maio SM, Deikus G, Chen BK,**  
574 **Fernandez-Sesma A, Simon V, Mulder LCF.** 2014. HIV-1 Interacts with Human  
575 Endogenous Retrovirus K (HML-2) Envelopes Derived from Human Primary  
576 Lymphocytes. J Virol **88:6213–23.**

577 22. **Giallonardo F Di, Töpfer A, Rey M, Prabhakaran S, Duport Y, Leemann C,**  
578 **Schmutz S, Campbell NK, Joos B, Lecca MR, Patrignani A, Däumer M, Beisel C,**  
579 **Rusert P, Trkola A, Günthard HF, Roth V, Beerenwinkel N, Metzner KJ.** 2014. Full-  
580 length haplotype reconstruction to infer the structure of heterogeneous virus  
581 populations. Nucleic Acids Res **42.**

582 23. **Ocwieja KE, Sherrill-Mix S, Mukherjee R, Custers-Allen R, David P, Brown M,**  
583 **Wang S, Link DR, Olson J, Travers K, Schadt E, Bushman FD.** 2012. Dynamic  
584 regulation of HIV-1 mRNA populations analyzed by single-molecule enrichment  
585 and long-read sequencing. Nucleic Acids Res **40:10345–55.**

586 24. **Thomas X V, Grady BPX, Van Der Meer JTM, Ho CK, Vanhommerig JW, Rebers**  
587 **SP, De Jong MD, Van Der Valk M, Prins M, Molenkamp R, Schinkel J.** 2015.  
588 Genetic characterization of multiple hepatitis C virus infections following acute  
589 infection in HIV-infected MSM. Aids **29:2287–9.**

590 25. **van den Hoek JA, Coutinho RA, van Haastrecht HJ, van Zadelhoff AW, Goudsmit**  
591 **J.** 1988. Prevalence and risk factors of HIV infections among drug users and drug-  
592 using prostitutes in Amsterdam. AIDS **2:55–60.**

- 593 26. **Lambers F.** 2011. Treatment of acute hepatitis C virus infection in HIV-infected  
594 MSM: the effect of treatment duration 1333–1336.
- 595 27. **Cock PJ a, Antao T, Chang JT, Chapman B a, Cox CJ, Dalke A, Friedberg I,**  
596 **Hamelryck T, Kauff F, Wilczynski B, de Hoon MJL.** 2009. Biopython: freely  
597 available Python tools for computational molecular biology and bioinformatics.  
598 *Bioinformatics* **25**:1422–3.
- 599 28. **Pages H, Aboyoun P, Gentleman R, DebRoy S.** Biostrings: String objects  
600 representing biological sequences, and matching algorithms.
- 601 29. **Edgar RC.** 2004. MUSCLE: multiple sequence alignment with high accuracy and  
602 high throughput. *Nucleic Acids Res* **32**:1792–7.
- 603 30. **Galtier N, Gout M, Galtier C.** 1996. SEA VIEW and PHYLO\_ WIN: two graphic tools  
604 for sequence alignment and molecular phylogeny. *Comput Applic Biosci* **12**:543–  
605 548.
- 606 31. **Tamura K, Stecher G, Peterson D, Filipinski A, Kumar S.** 2013. MEGA6: Molecular  
607 evolutionary genetics analysis version 6.0. *Mol Biol Evol* **30**:2725–2729.
- 608 32. **Paradis E, Claude J, Strimmer K.** 2004. APE: Analyses of phylogenetics and  
609 evolution in R language. *Bioinformatics* **20**:289–290.
- 610 33. **Li H, Stoddard MB, Wang S, Blair LM, Giorgi EE, Parrish EH, Learn GH, Hraber P,**  
611 **Goepfert P a, Saag MS, Denny TN, Haynes BF, Hahn BH, Ribeiro RM, Perelson AS,**  
612 **Korber BT, Bhattacharya T, Shaw GM.** 2012. Elucidation of hepatitis C virus  
613 transmission and early diversification by single genome sequencing. *PLoS Pathog*  
614 **8**:e1002880.

- 615 34. **Wang GP, Sherrill-Mix S a, Chang K-M, Quince C, Bushman FD.** 2010. Hepatitis C  
616 virus transmission bottlenecks analyzed by deep sequencing. *J Virol* **84**:6218–28.
- 617 35. **Li H, Stoddard M, Wang S, Giorgi EE, Blair L, Learn G, Hahn B, Alter H, Busch M,**  
618 **Fierer D, Ribeiro RM, Perelson a. S, Bhattacharya T, Shaw GM.** 2015. Single  
619 Genome Sequencing of Hepatitis C Virus in Donor-Recipient Pairs Distinguishes  
620 Modes and Models of Virus Transmission and Early Diversification. *J Virol*  
621 *JVI.02156-15.*
- 622 36. **Raghwani J, Rose R, Sheridan I, Lemey P, Suchard MA, Santantonio T, Farci P,**  
623 **Klenerman P, Pybus OG.** 2016. Exceptional Heterogeneity in Viral Evolutionary  
624 Dynamics Characterises Chronic Hepatitis C Virus Infection. *PLOS Pathog*  
625 **12**:e1005894.
- 626 37. **Gray RR, Strickland SL, Veras NM, Goodenow MM, Pybus OG, Lemon SM, Fried**  
627 **MW, Nelson DR, Salemi M.** 2012. Unexpected maintenance of hepatitis C viral  
628 diversity following liver transplantation. *J Virol* **86**:8432–9.
- 629 38. **Alfonso V, Mbayed VA, Sookoian S, Campos RH.** 2005. Intra-host evolutionary  
630 dynamics of hepatitis C virus E2 in treated patients. *J Gen Virol* **86**:2781–6.
- 631 39. **Smith J a, Aberle JH, Fleming VM, Ferenci P, Thomson EC, Karayiannis P, McLean**  
632 **AR, Holzmann H, Klenerman P.** 2010. Dynamic coinfection with multiple viral  
633 subtypes in acute hepatitis C. *J Infect Dis* **202**:1770–9.
- 634 40. **Ramachandran S, Campo DS, Dimitrova ZE, Xia G, Purdy M a, Khudyakov YE.**  
635 2011. Temporal variations in the hepatitis C virus intrahost population during  
636 chronic infection. *J Virol* **85**:6369–80.

- 637 41. **Culasso a. C a, Baré P, Aloisi N, Monzani MC, Corti M, Campos RH.** 2014. Intra-  
638 host evolution of multiple genotypes of hepatitis C virus in a chronically infected  
639 patient with HIV along a 13-year follow-up period. *Virology* **449**:317–327.
- 640 42. **Kato N, Ootsuyama Y, Tanaka T.** 1992. Marked in the of hepatitis C viruses  
641 **22**:107–123.
- 642 43. **McCloskey RM, Liang RH, Harrigan PR, Brumme ZL, Poon AFY.** 2014. An  
643 Evaluation of Phylogenetic Methods for Reconstructing Transmitted HIV Variants  
644 using Longitudinal Clonal HIV Sequence Data. *J Virol* **88**:6181–94.
- 645 44. **Gray RR, Parker J, Lemey P, Salemi M, Katzourakis A, Pybus OG.** 2011. The mode  
646 and tempo of hepatitis C virus evolution within and among hosts. *BMC Evol Biol*  
647 **11**:131.
- 648 45. **Laird Smith M, Murrell B, Eren K, Ignacio C, Landais E, Weaver S, Phung P, Ludka**  
649 **C, Hepler L, Caballero G, Pollner T, Guo Y, Richman D, Pognard P, Paxinos EE,**  
650 **Kosakovsky Pond SL, Smith DM.** 2016. Rapid Sequencing of Complete *env* Genes  
651 from Primary HIV-1 Samples. *Virus Evol* **2**:vew018.
- 652 46. **Cronn R, Cedroni M, Haselkorn T, Grover C, Wendel JF.** 2002. PCR-mediated  
653 recombination in amplification products derived from polyploid cotton. *Theor*  
654 *Appl Genet* **104**:482–489.
- 655 47. **von Hahn T, Yoon JC, Alter H, Rice CM, Rehmann B, Balfe P, McKeating J a.**  
656 2007. Hepatitis C virus continuously escapes from neutralizing antibody and T-cell  
657 responses during chronic infection in vivo. *Gastroenterology* **132**:667–78.
- 658 48. **Vandamme A-M, Pybus OG.** 2013. Viral phylogeny in court: the unusual case of

659 the Valencian anesthetist. BMC Biol **11**:83.

660 49. **Tohme R a, Holmberg SD.** 2010. Is sexual contact a major mode of hepatitis C  
661 virus transmission? Hepatology **52**:1497–505.

662 50. **Keele BF, Derdeyn C a.** 2009. Genetic and antigenic features of the transmitted  
663 virus. Curr Opin HIV AIDS **4**:352–7.

664 51. **D'Arienzo V, Moreau A, D'Alteroche L, Gissot V, Blanchard E, Gaudy-Graffin C,**  
665 **Roch E, Dubois F, Giraudeau B, Plantier J-C, Goudeau A, Roingeard P, Brand D.**  
666 2013. Sequence and functional analysis of the envelope glycoproteins of hepatitis  
667 C virus variants selectively transmitted to a new host. J Virol **87**:13609–18.

668 52. **Shen C, Gupta P, Xu X, Sanyal A, Rinaldo C, Seaberg E, Margolick JB, Martinez-**  
669 **Maza O, Chen Y.** 2014. Transmission and evolution of hepatitis C virus in HCV  
670 seroconverters in HIV infected subjects. Virology **449**:339–49.

671 53. **Pestka JM, Zeisel MB, Bläser E, Schürmann P, Bartosch B, Cosset F-L, Patel AH,**  
672 **Meisel H, Baumert J, Viazov S, Rispeter K, Blum HE, Roggendorf M, Baumert TF.**  
673 2007. Rapid induction of virus-neutralizing antibodies and viral clearance in a  
674 single-source outbreak of hepatitis C. Proc Natl Acad Sci U S A **104**:6025–30.

675 54. **Dowd K a, Netski DM, Wang X-H, Cox AL, Ray SC.** 2009. Selection pressure from  
676 neutralizing antibodies drives sequence evolution during acute infection with  
677 hepatitis C virus. Gastroenterology **136**:2377–86.

678 55. **Bowen DG, Walker CM.** 2005. Adaptive immune responses in acute and chronic  
679 hepatitis C virus infection. Nature **436**:946–952.

680 56. **Farci P.** 2000. The Outcome of Acute Hepatitis C Predicted by the Evolution of the



681           Viral Quasispecies. *Science* (80- ) **288**:339–344.

682   57.   **Xiao F, Fofana I, Heydmann L, Barth H, Soulier E, Habersetzer F, Doffoël M, Bukh**  
683           **J, Patel AH, Zeisel MB, Baumert TF.** 2014. Hepatitis C virus cell-cell transmission  
684           and resistance to direct-acting antiviral agents. *PLoS Pathog* **10**:e1004128.

685   58.   **Timpe JM, Stamatakis Z, Jennings A, Hu K, Farquhar MJ, Harris HJ, Schwarz A,**  
686           **Desombere I, Roels GL, Balfe P, McKeating J a.** 2007. Hepatitis C virus cell-cell  
687           transmission in hepatoma cells in the presence of neutralizing antibodies.  
688           *Hepatology* **47**:17–24.

689   59.   **Sede M, Roberto L, Moretti F, Laufer N, Quarleri J.** 2014. Inter and intra-host  
690           variability of hepatitis C virus genotype 1a hypervariable envelope coding  
691           domains followed for a 4 – 11 year of human immunodeficiency virus coinfection  
692           and highly active antiretroviral therapy. *Virology* **471–473**:19–28.

693   60.   **Löve A, Molnégren V, Månsson AS, Smáradóttir A, Thorsteinsson SB, Widell A.**  
694           2004. Evolution of hepatitis C virus variants following blood transfusion from one  
695           infected donor to several recipients: A long term follow-up. *J Gen Virol* **85**:441–  
696           450.

697   61.   **Farci P.** 2011. New insights into the HCV quasispecies and compartmentalization.  
698           *Semin Liver Dis* **31**:356–374.

699   62.   **Gray RR, Salemi M, Klenerman P, Pybus OG.** 2012. A new evolutionary model for  
700           hepatitis C virus chronic infection. *PLoS Pathog* **8**:e1002656.

701   63.   **Gismondi MI, Díaz Carrasco JM, Valva P, Becker PD, Guzmán CA, Campos RH,**  
702           **Preciado MV.** 2013. Dynamic changes in viral population structure and

compartmentalization during chronic hepatitis C virus infection in children.  
Virology **447**:187–96.

64. **Laskus T, Radkowski M, Piasek a, Nowicki M, Horban a, Cianciara J, Rakela J.**  
2000. Hepatitis C virus in lymphoid cells of patients coinfectd with human  
immunodeficiency virus type 1: evidence of active replication in  
monocytes/macrophages and lymphocytes. J Infect Dis **181**:442–448.

65. **Blackard JT, Ma G, Sengupta S, Martin CM, Powell E a, Shata MT, Sherman KE.**  
2014. Evidence of distinct populations of hepatitis C virus in the liver and plasma  
of patients co-infected with HIV and HCV. J Med Virol **86**:1332–41.

713 **FIGURE LEGENDS**

714 **Figure 1. Error rate and sequencing depth of each CCS full pass (FP).** A) Error rate is  
715 divided into insertions, deletions and mismatches. The rates are averaged over the  
716 four control plasmids with the bars representing the SD. B) The sequencing depth as  
717 the number of reads per sample. The median from 50 subject samples is shown with  
718 the bars representing the IQR.

719

720 **Figure 2. NJ phylogeny of the five subjects.** The NJ tree of the HCV E1E2 genomic  
721 region (1680 nt). Orange branches indicate the CCS reads of subject 037, blue:  
722 subject 053, green: subject 061, purple: subject 004 and red: subject 147. Bootstrap  
723 analysis was performed with 100 replicates, however are not displayed because the  
724 great majority was below the 70% threshold. The scale bar indicates the nucleotide  
725 substitutions per site.

726

727 **Figure 3. NJ phylogeny of subject 061.** The NJ tree of the HCV E1E2 genomic region  
728 (1680 nt) of subject 061, reconstructed with all CCS reads from all 18 time points.  
729 The asterisk (\*) indicates the time points under treatment.

730

731 **Figure 4 a-e. Genetic diversity during the course of infection.** The median genetic  
732 diversity with the bars representing the IQR. The grey shaded area indicates the  
733 interval between the last negative and first RNA positive time point and the blue  
734 area represents the treatment period. Note the difference in y-axis scale.

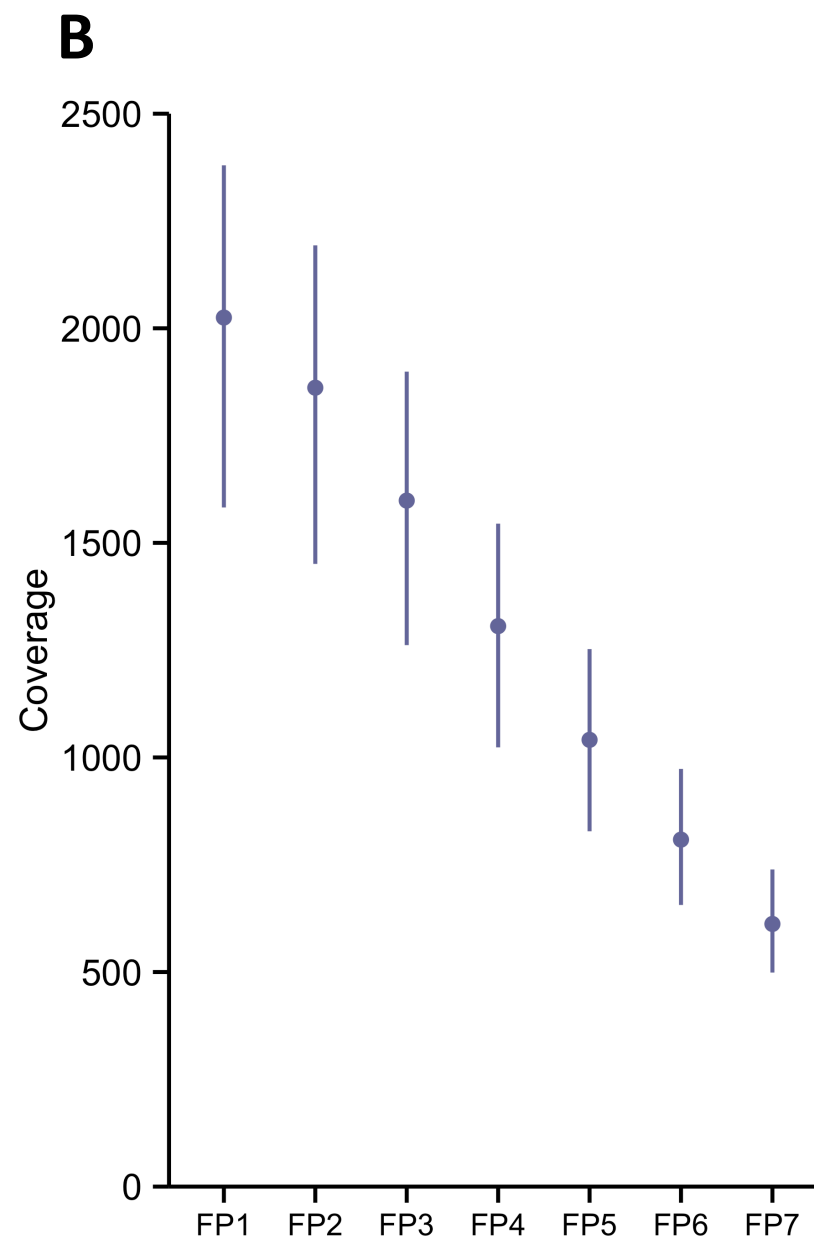
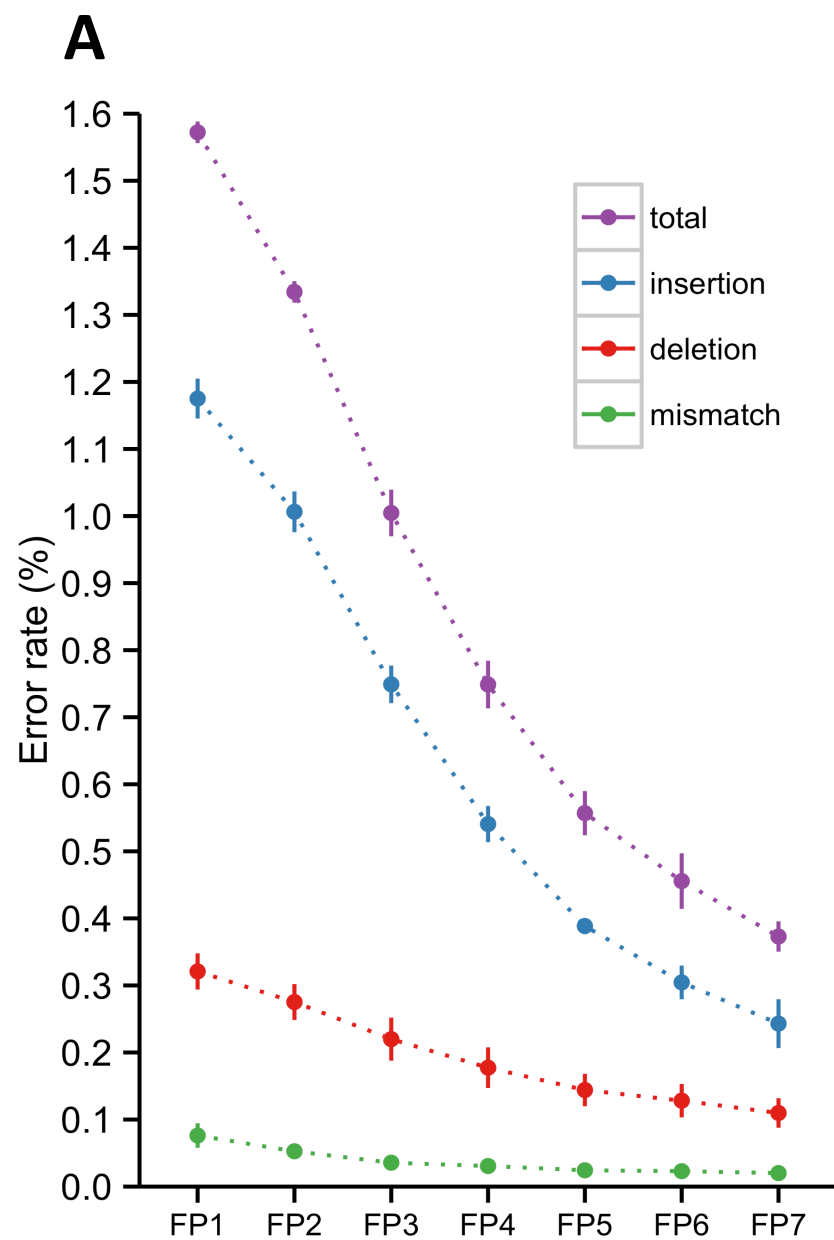
735

736

737 **TABLES**738 **Table 1. Subjects and samples characteristics**

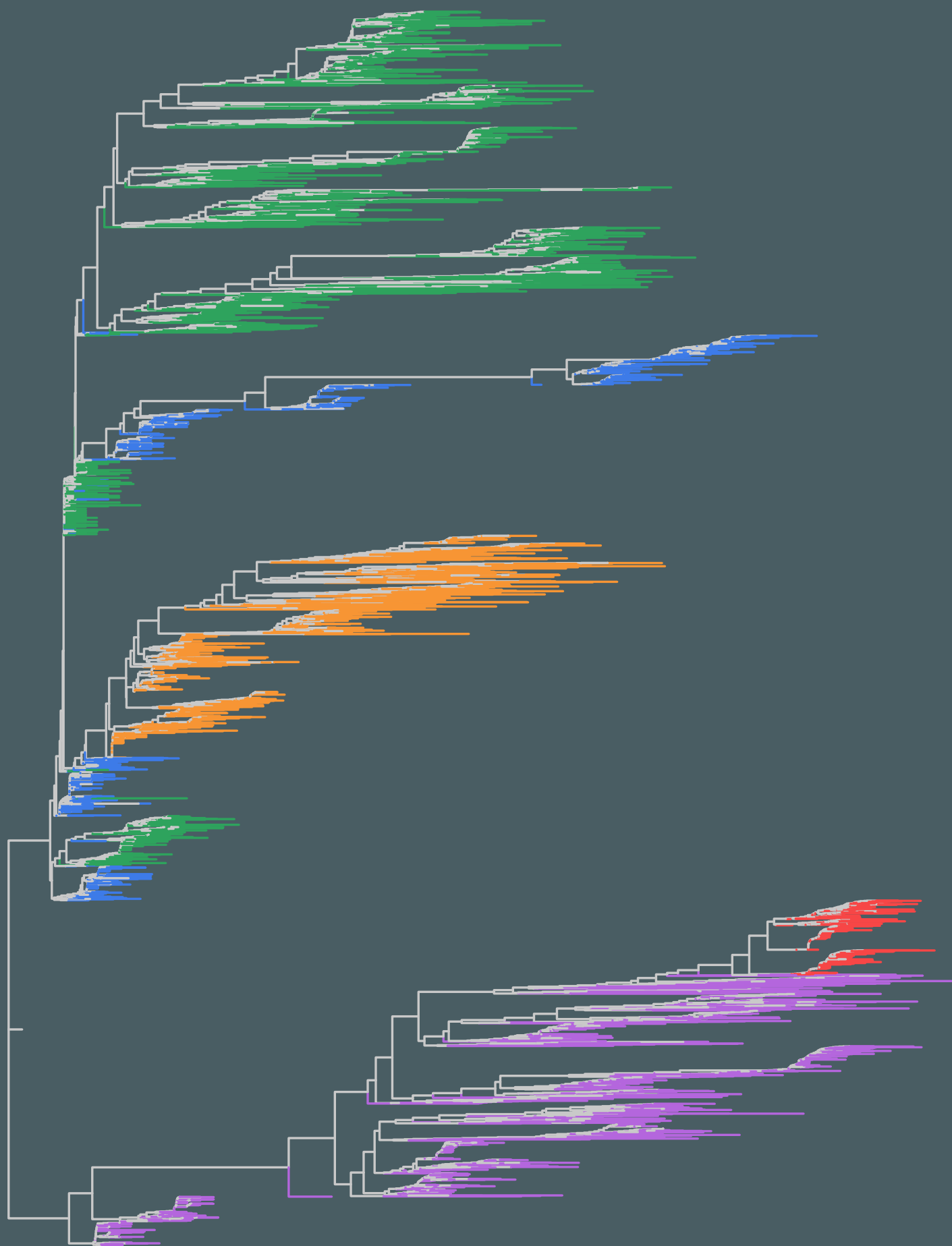
Subject	Sex	Age	Last HCV RNA- sample	First HCV RNA+ sample	Estimated infection date	Last seronegative sample	First seropositive sample	Follow- up (yr)	No. of samples
004	M	61	8-Feb-02	5-Jun-02	9-Apr-02	8-Feb-02	5-Jun-02	10	11
037	M	55	26-Oct-01	1-Feb-02	14-Dec-01	26-Oct-01	1-Feb-02	7	9
053	M	61	12-Dec-00	3-Jul-01	24-Mar-01	5-Aug-98	25-Apr-03	8	9
061	M	60	14-Sep-01	7-Jan-02	11-Nov-01	14-Sep-01	19-Aug-02	11	18
147	F	49	21-Dec-12	1-Feb-13	11-Jan-13	1-Feb13	Not available	0.5	3

739



# Subject

- 004
- 037
- 053
- 061
- 147



0.007

0.005

### Time

- 2002.02
- 2002.39
- 2002.63
- 2003.10
- 2003.76
- 2004.67
- 2005.01
- 2006.02
- 2007.03
- 2008.29
- 2008.95
- 2009.07
- \* 2009.25
- \* 2009.54
- 2010.38
- 2012.19
- 2012.81
- 2013.64

