viruses

MDPI

Article

HCV Genetic Diversity Can Be Used to Infer Infection Recency and Time since Infection

Louisa A. Carlisle [1,2], Teja Turk [1,2], Karin J. Metzner [1,2], Herbert A. Mbunkah [1,2], Cyril Shah [2], Jürg Böni [2,3], Michael Huber [2,3], Dominique L. Braun [1,2], Jan Fehr [1,4], Luisa Salazar-Vizcaya [5], Andri Rauch [5], Sabine Yerly [6], Aude Nguyen [6], Matthias Cavassini [7], Marcel Stoeckle [8], Pietro Vernazza [9], Enos Bernasconi [10], Huldrych F. Günthard [1,2,*] and Roger D. Kouyos [1,2,*]

1 Division of Infectious Diseases and Hospital Epidemiology, University Hospital Zurich, CH-8091 Zurich, Switzerland; louisa.anja@hotmail.co.uk (L.A.C.); turk.teja@virology.uzh.ch (T.T.); Karin.Metzner@usz.ch (K.J.M.); afegenwimbunkah@gmail.com (H.A.M.); dominique.braun@usz.ch (D.L.B.); Jan.Fehr@usz.ch (J.F.)
2 Institute of Medical Virology, University of Zurich, CH-8057 Zurich, Switzerland; shah.cyril@virology.uzh.ch (C.S.); boeni.juerg@virology.uzh.ch (J.B.); huber.michael@virology.uzh.ch (M.H.)
3 Swiss National Reference Center for Retroviruses, University of Zurich, CH-8057 Zurich, Switzerland
4 Department of Public Health, Epidemiology Biostatistics and Prevention Institute, University of Zurich, CH-8001 Zurich, Switzerland
5 Department of Infectious Diseases, Bern University Hospital, University of Bern, CH-3010 Bern, Switzerland; luisapaola.salazarvizcaya@insel.ch (L.S.-V.); Andri.Rauch@insel.ch (A.R.)
6 Laboratory of Virology, Division of Infectious Diseases, Geneva University Hospital, University of Geneva, CH-1205 Geneva, Switzerland; Sabine.Yerly@hcuge.ch (S.Y.); aude.nguyen@hcuge.ch (A.N.)
7 Division of Infectious Diseases, Lausanne University Hospital, CH-1011 Lausanne, Switzerland; Matthias.Cavassini@chuv.ch
8 Division of Infectious Diseases and Hospital Epidemiology, University Hospital Basel, University of Basel, CH-4031 Basel, Switzerland; Marcel.Stoeckle@usb.ch
9 Division of Infectious Diseases, Cantonal Hospital St Gallen, CH-9007 St. Gallen, Switzerland; pietro.vernazza@kssg.ch
10 Division of Infectious Diseases, Regional Hospital Lugano, CH-6900 Lugano, Switzerland; Enos.bernasconi@eoc.ch
* Correspondence: huldrych.guenthard@usz.ch (H.F.G.); roger.kouyos@uzh.ch (R.D.K.); Tel.: +41-44-255-34-50 (H.F.G.); +41-44-255-36-10 (R.D.K.)

Received: 1 July 2020; Accepted: 27 October 2020; Published: 31 October 2020

Abstract: HIV-1 genetic diversity can be used to infer time since infection (TSI) and infection recency. We adapted this approach for HCV and identified genomic regions with informative diversity. We included 72 HCV/HIV-1 coinfected participants of the Swiss HIV Cohort Study, for whom reliable estimates of infection date and viral sequences were available. Average pairwise diversity (APD) was calculated over each codon position for the entire open reading frame of HCV. Utilizing cross validation, we evaluated the correlation of APD with TSI, and its ability to infer TSI via a linear model. We additionally studied the ability of diversity to classify infections as recent (infected for <1 year) or chronic, using receiver-operator-characteristic area under the curve (ROC-AUC) in 50 patients whose infection could be unambiguously classified as either recent or chronic. Measuring HCV diversity over third or all codon positions gave similar performances, and notable improvement over first or second codon positions. APD calculated over the entire genome enabled classification of infection recency (ROC-AUC = 0.76). Additionally, APD correlated with TSI (R^2 = 0.33) and could predict TSI (mean absolute error = 1.67 years). Restricting the region over which APD was calculated to *E2-NS2* further improved accuracy (ROC-AUC = 0.85, R^2 = 0.54, mean absolute error = 1.38 years). Genetic diversity in HCV correlates with TSI and is a proxy for infection recency and TSI, even several years post-infection.

Keywords: hepatitis C virus infection; infection recency; genetic variation; sequence analysis; viral genomics

1. Introduction

Inferring the duration of infection is of key importance for understanding both the epidemiology and pathogenesis of hepatitis C virus (HCV) infections. From an epidemiological perspective, the time of infection can inform incidence assays, phylogenetic studies, and prediction of future chronic liver disease burdens. In particular, it could be vital for monitoring public health progress in the context of elimination [1], as it enables the identification of ongoing transmission. From an individual-patient perspective, this information could contribute to knowledge of disease progression and assessment.

The nature of HCV transmission and its mostly asymptomatic acute infection means that the date of infection is rarely known, and there is a lack of known biomarkers available from which this information could be estimated. Accordingly, studies typically have to rely on some combination of cohort data and mathematical modelling to infer infection dates [2–13], which remain highly uncertain for most HCV-infected individuals. In the present study, we took advantage of the unique opportunity of a cohort with annual HCV screening, detailed clinical characteristics, and sampling.

A similar problem exists for human immunodeficiency virus-1 (HIV-1), which is also an RNA virus of comparable genome size that chronically infects patients. For HIV-1, it has been shown that viral diversity can be used to infer infection recency [14–16] and time since infection [17], and that diversity derived from next-generation sequencing (NGS) sequences is more accurate than diversity derived as the fraction of ambiguous nucleotides from Sanger sequences [18].

In this study, we aimed to investigate whether the same NGS-derived-diversity method can be applied to HCV, and to identify the region of the genome over which it is most informative to measure diversity.

2. Materials and Methods

2.1. Patients

We included 72 HCV-HIV coinfected patients from the Swiss HIV Cohort Study (SHCS), for all of whom an NGS-sequenced HCV sample was available. For patients enrolled in the SHCS, HCV-serologies are determined since 2000 every 12–24 months. Therefore, the patients considered in our study had a date of HCV infection known to within 24 months. This date of infection was calculated as the midpoint between the date of the most recent negative (RNA or antibody) test prior to the sample date, and the earliest date of either a positive HCV (RNA or antibody) test result, or the date at which the sample was taken. Patient characteristics are summarized in Table 1.

Table 1. Patient characteristics.

Total Number		**72**
Gender, *n* (%)	Female	2 (3)
	Male	70 (97)
Age when sample taken (years), median (IQR)		45 (39, 52)
Ethnicity, *n* (%)	Asian	2 (3)
	Black	2 (3)
	Hispanic	4 (4)
	White	64 (89)
HIV transmission group, *n* (%)	HET	3 (4)
	MSM	67 (93)
	IDU	1 (1)
	Unclear/unknown	1 (1)
Recorded history of intravenous drug use ever, *n* (%)	Yes	9 (13)
	No	63 (88)
HCV Viral subtype, *n* (%)	1A	50 (69)
	1B	5 (7)
	2C	1 (1)
	3A	2 (3)
	4D	14 (19)
Time since HCV infection (years), median (IQR)		0.82 (0.47, 2.5)
Clearly recent or chronic [a], *n* (%)	True	50 (69)
	False	22 (31)
Full coverage of gene at all codon positions, *n* (%)	*C*	69 (96)
	E1	69 (96)
	E2	70 (97)
	p7	70 (97)
	NS2	70 (97)
	NS3	71 (99)
	NS4A	71 (99)
	NS4B	70 (97)
	NS5A	69 (96)
	NS5B	65 (90)

IQR = Interquartile range, HET = heterosexual contacts, MSM = men who have sex with men, IDU = injection drug use. [a] Sample collection less than 12 months after last negative HCV test or more than 12 months after first positive HCV test.

2.2. Sequencing

Samples were sequenced in the context of two different projects, resulting in differing, although similar, protocols. Briefly, the following sequencing protocols were used:

Project 1, as part of the Swiss-HCVree-trial (NCT02785666) [19] within the SHCS, 53 samples (see Supplementary Material S1 for the description of near full-length genome sequencing): cDNA was synthesized and amplified using a one-step reverse transcription and PCR kit, or in two individual steps. Some samples then underwent a second, nested-PCR (Supplementary Table S1). Samples were pooled for sequencing, and libraries were run using MiSeq (Illumina, San Diego, CA, USA) 1 × 150 cycles.

Project 2, within the SHCS, 19 samples: Viral RNA was extracted from plasma stored in the SHCS biobank by the Nucleospin RNA Virus Kit (Macherey-Nagel, Düren, Germany). RNA was then amplified by RT-PCR in a two-step process, and some samples underwent a second, nested-PCR if necessary. HCV RNA genome sequences were generated by amplification of almost full-length HCV RNA followed by massive parallel sequencing. A MiSeq (Illumina, San Diego, CA, USA) instrument was used for sequencing with 2 × 250 bp.

Sequences were processed using MinVar version 2.2.1 (https://github.com/medvir/MinVar) [20], which filters and aligns reads before returning sequence variants. A slightly modified script was used in order to output sequence position information at the nucleotide level.

2.3. Diversity Score Calculation

Average pairwise diversity (APD) was calculated from nucleotide minority variant frequencies using Equation (1) [17], explanation as in [18]. This first determines whether a position has minor variants above a threshold, and subsequently sums the diversity contribution of all variants at that position. Finally, diversity across all positions is averaged. This value is functionally equivalent to the average proportion of positions at which two randomly selected sequences differ.

$$\text{APD} = \frac{1}{L}\sum_{i=1}^{L}\Theta\left(1 - x_i^m - x_c\right)\left[\sum_{\alpha} x_{i\alpha}(1 - x_{i\alpha})\right] \quad (1)$$

where

L = length of analysed sequence.

i = sequence position.

$x_i{}^m$ = frequency of major variant m at position i.

x_c = 0.01.

$\Theta(x) = \begin{cases} 1, & x > x_c \\ 0, & x \leq x_c \end{cases}$

$\alpha \in$ {A,C,G,T,deletion}.

$x_{i\alpha}$ = frequency of variant α at position i.

x_c is a frequency cut-off, below which variants are considered to be indistinguishable from those generated by PCR or sequencing errors. It was set to 1% as this is the approximate detection limit of a high-throughput sequencing workflow (Illumina) [21,22].

We calculated average pairwise diversity over various subsections of the HCV genome open reading frame:

– All codon positions, and only the first/second/third codon positions in turn.

– Whole open reading frame, individual genes, and 11 overlapping equal regions that the genome was split into (length = 502 or 501 amino acid codons).

2.4. Data Analysis

All results were analysed in R version 3.5.1 [23], using packages data.table [24], pROC [25], readstata13 [26], and RColorBrewer [27].

We examined how well average pairwise diversity correlated with time since infection (TSI) using linear regression. This linear model (Equation (2)) was then validated using leave-one-out cross-validation, with mean absolute error as the primary outcome. This was conducted by calculating the model coefficients using all but one sample, and then applying this model to the remaining "test" sample. The absolute value of the difference between the estimated and actual time since infection was then calculated for this test sample. This procedure was repeated for all samples in turn, and the mean value of all the absolute errors was taken, providing a simple measure by which to compare average pairwise diversity calculated across different regions.

$$\text{Estimated TSI} = \beta\,\text{APD} + \alpha \quad (2)$$

We studied how well average pairwise diversity could be used to infer infection recency using receiver operator characteristics (ROC) analyses, with recent infection (<1 year post-infection) defined as the positive outcome. We restricted these recency analyses to the 50 patients who could be unambiguously classified as recent (21 patients) or chronic (29 patients) due to the sample having been

collected less than 12 months after the last negative HCV test or more than 12 months after the first positive HCV test.

A sensitivity analysis was conducted to assess variability between viral subtypes, as levels of diversity have been seen to vary between subtypes [28]. We therefore repeated our analyses using just those patients infected with subtype 1A, and 4D; insufficient number of samples were available from other viral subtypes for this analysis to be meaningful for them.

2.5. Ethics Approval and Consent to Participate

This analysis was conducted in the context of the SHCS which was approved by the ethics committees of the participating institutions (http://www.shcs.ch/userfiles/file/ethics_committee_approval_and_informed_consent.pdf, Kantonale Ethikkommission Bern, Ethikkommission des Kantons St. Gallen, Comité Départemental d'Éthique des Spécialités Médicales et de Médicine Communataire et de Premier Recours, Kantonale Ethikkommission Zürich, Repubblica et Cantone Ticino–Comitato Ethico Cantonale, Commission Cantonale d'Éthique de la Recherche sur l'Être Humain, Ethikkommission beider Basel), and written informed consent was obtained from all participants. The study protocol conforms to the ethical guidelines of the 1975 Declaration of Helsinki.

3. Results

3.1. Codon Position Makes a Notable Difference in Average Pairwise Diversity Scores

We found for the first and second codon positions a substantial correlation with time since infection but also that diversity remains low even several years post-infection (Figure 1). Conversely, average pairwise diversity over the third codon position shows a large and steady increase with time since infection, and should therefore be more informative when inferring time since infection and infection recency. However, calculating average pairwise diversity over all three positions yields similar results to considering only third-codon positions (Figure 1). This suggests that average pairwise diversity calculated over either third or all codon positions may be capable of inferring time since infection with reasonable precision.

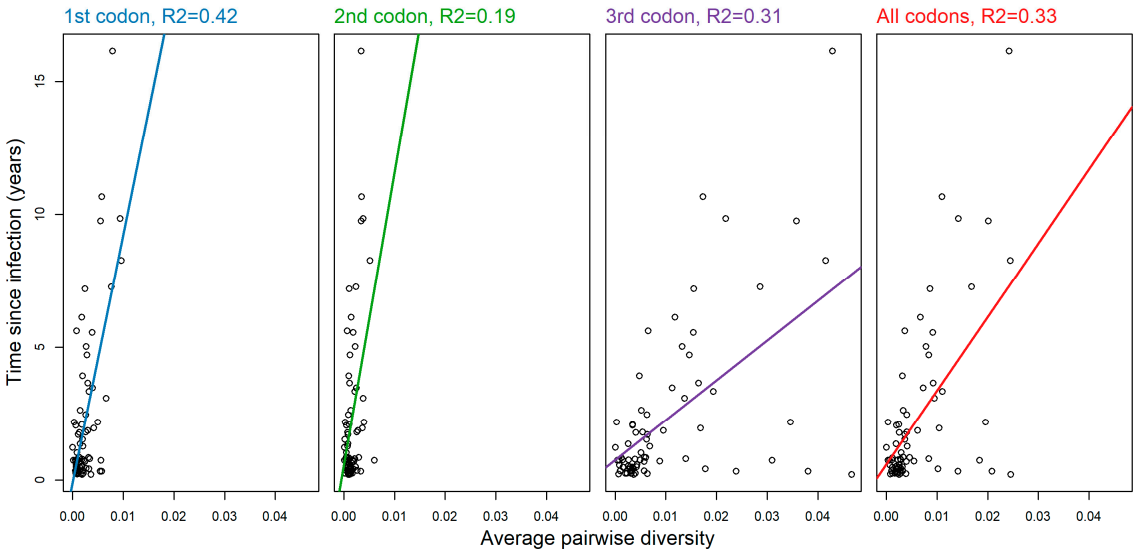

Figure 1. Average pairwise diversity (APD) against time since infection for APD calculated over each of the codon positions in turn, and over all three codon positions. Linear regression models are shown as solid lines.

We compared how well average pairwise diversity over individual and all codon positions could infer infection recency using ROC analyses (Figure 2). This analysis, was restricted to the 50 patients

whose infections could be unambiguously classified as either recent (time since infection <1 year) or chronic (time since infection >1 year). In line with Figure 1, using third and all codon positions outperforms using first and second codon positions, supporting the notion that average pairwise diversity calculated over these positions is more informative. Due to the similarity between using the third codon and all codon positions, we continued our analyses using average pairwise diversity calculated over all codon positions, this being the easier measure to calculate as it does not require identification of the reading frame (results over the third codon position are provided in Supplementary Figure S1 for comparison and give very similar outcomes).

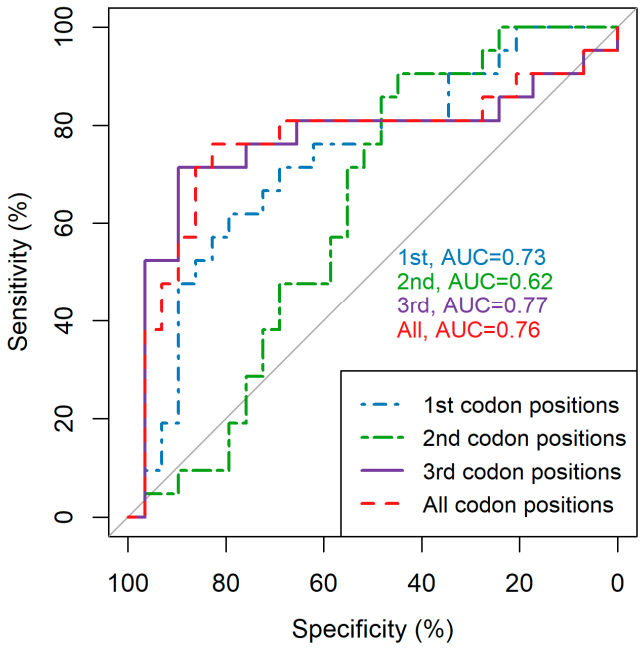

Figure 2. Receiver operator characteristics (ROC) curves comparing the ability of average pairwise diversity (APD) calculated over each and all codon positions to infer whether infections are recent (<1 year post-infection) or chronic. APD was calculated across the whole HCV open reading frame. All 50 patients who could be clearly classified as recent or chronic are included. Recent infection is taken as the positive outcome. AUC = area under the ROC curve.

3.2. Restricting Average Pairwise Diversity to Certain Regions of the Genome Improves Inference of Time since Infection and Infection Recency

We calculated average pairwise diversity over the entire open reading frame of HCV, each gene individually, and for regions derived by dividing the open reading frame into eleven equal overlapping regions. We then assessed how well each average pairwise diversity scores could be used to infer infection recency (via ROC analyses) and time since infection (via linear regression and cross-validation). When comparing the three outcomes (area under the ROC curve, R^2, and mean absolute error) for all regions across the open reading frame (Figure 3), we find that the region covering amino acid codons 503-1004– provides the best score for all three outcomes. This region spans the major 3' part of *E2*, the entirety of *p7*, and the majority of the *NS2* genes. The regions covering the last 1000 amino acid codons of the open reading frame also perform well, but less consistently across outcomes (Figure 3). The figure also shows the HCV genes colour-coded by a composite score of all three outcomes, from which a similar pattern emerges, with *E2* and *NS2* individually performing especially well. We therefore additionally calculated average pairwise diversity over the three genes *E2*, *p7* and *NS2* combined and found similarly strong results. Supplementary Table S2 provides the exact scores for all the regions tested. Supplementary Figure S2 shows the same analysis repeated for larger regions, which provides a similar outcome.

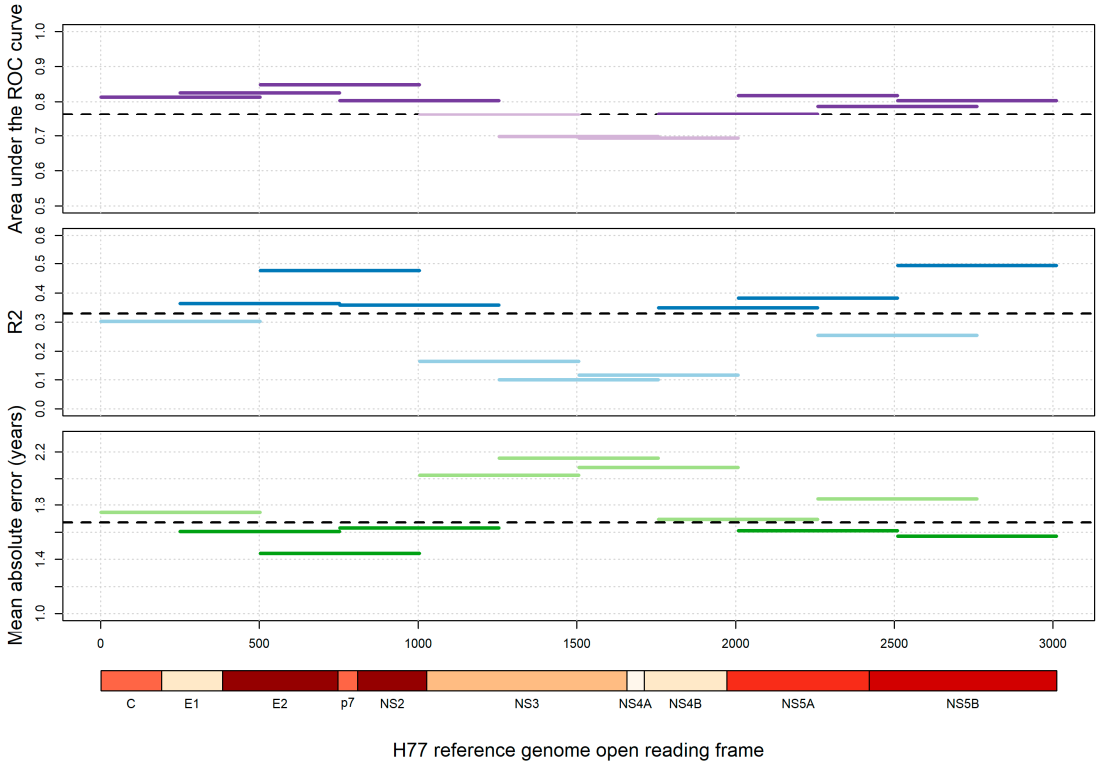

Figure 3. Area under the ROC curve, R^2, and mean absolute error across the HCV open reading frame, all codon positions. The HCV open reading frame was split into 11 overlapping regions of approximately 500 amino acid codons, and average pairwise diversity (APD) was calculated over individual regions, using all codon positions. Regions were tested for their ability to categorize infection as recent (<1 year) or chronic (top), their correlation with time since infection (middle), and their ability to infer time since infection (bottom). Black dashed lines show the respective values for APD calculated over the whole open reading frame. A similar analysis was performed with diversity calculated over each gene in turn. The HCV genome is shown along the *x*-axis, with genes colour-coded for a composite (z-score sum, see Supplementary Equation S1) of all three outcome scores. Darker red indicates a better overall performance. Numbers along the *x*-axis refer to amino acid positions of the H77 reference genome.

Based on these results, we recommend calculating average pairwise diversity using all codon positions, over the region of amino acid codons 503-1004– of the open reading frame, or using the entirety of genes *E2*, *p7* and *NS2* together. Time since infection can be estimated from these scores using Equation (2), with recommended coefficients alpha = 0.42, beta = 353.70 for average pairwise diversity over the region 503-1004–, or alpha = 0.39, beta = 320.38 for average pairwise diversity over *E2*, *p7* and *NS2*. A full list of suggested coefficients for all regions tested is available in Supplementary Table S2. Figure 4 shows the time since infection against estimated time since infection for the recommended regions.

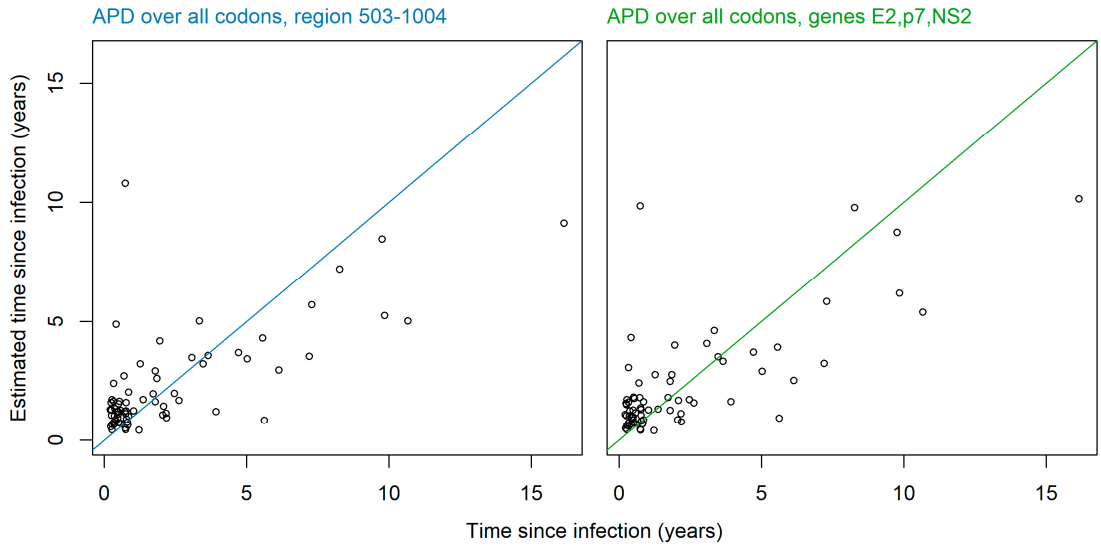

Figure 4. Time since infection against estimated time since infection as calculated from average pairwise diversity (APD). APD calculated over the recommended region of amino acid codons 503–1004 (**left**), and the recommended genes *E2*, *p7*, and *NS2* (**right**).

3.3. Sensitivity Analysis Shows Minor Variation between Viral Subtypes

We conducted a sub-analysis including only patients infected with viral subtype 1A (50 patients, of which 33 could be included in the ROC analysis), and 4D (14 patients, of which 10 could be included in the ROC analysis), using average pairwise diversity calculated over our recommended regions. For diversity calculated over amino acid codons 503–1004, the area under the ROC curve was the same as the overall score of 0.85 for genotype 1A (0.85) and slightly lower for genotype 4D (0.67). Conversely, R^2 was slightly higher than the overall score of 0.48 for genotype 4D (0.56), and lower for genotype 1A (0.43), whilst mean absolute error was worse than the overall 1.44 years for both genotype 1A (1.64 years) and 4D (1.52 years). A similar pattern was seen for average pairwise diversity over the genes *E2*, *p7*, and *NS2* (overall: area under the ROC curve = 0.85, R^2 = 0.54, mean absolute error = 1.38 years. 1A: area under the ROC curve = 0.86, R^2 = 0.51, mean absolute error = 1.51 years. 4D: area under the ROC curve = 0.62, R^2 = 0.56, mean absolute error = 1.46 years).

4. Discussion

Overall, this study shows that HCV genetic diversity as calculated from nucleotide frequencies derived from NGS deep sequencing correlates with, and can be used to infer, the time since infection and infection recency. A wide range of samples of well-defined infection timepoints with differing infection durations was included, showing that this method could be employed for samples up to 16 years post-infection. With a mean absolute error of less than 1.5 years, this method can provide insight which may be particularly helpful for public health monitoring in the context of HCV elimination strategies, and the identification of recent versus chronic infections.

We additionally conducted analyses focusing on individual regions across the entire open reading frame, and have shown that the pattern of diversity differs across the genome, with some regions and codons experiencing differing levels of diversification over time. The contrast in diversity calculated over codons 1 or 2 compared to codon 3 highlights the strong effect of selection against non-synonymous mutations. It should be noted, however, that even at third-codon positions substitutions will not be completely neutral, as they can for example affect RNA secondary structure. Even though such effects may restrict the accumulation of viral diversity, they do not seem to have had a major effect on the association between time since infection and diversity as indicated by similarly strong correlations observed over the entire HCV genome (Figure 3). For example, we observed one of the strongest

associations for the NS5B gene, which (in its second half) exhibits a particularly high density of regions encoding for RNA secondary structure [29,30]. Surprisingly, the most informative region was found to include the *E2* gene. As this encodes one of the two viral envelope proteins, a generally high level of diversification is unsurprising, but it was expected that selective sweeps due to immune system pressure would result in an unsteady increase in diversity unsuitable for inferring time since infection, as has been seen for the HIV envelope gene [17]. This finding defies that and suggests that other factors may be additionally influencing the accumulation of diversity within the HCV genome, causing a steadier increase.

As some variation between diversity within different viral subtypes has been reported [28], we conducted a sensitivity analysis with individual subtypes to evaluate the impact of viral subtype on our analyses. Whilst some variation was observed, overall patterns were robust across subtypes Furthermore, the largest deviation from the overall results was found for the subtype 4D ROC analyses which were conducted with only 10 samples. Our results, hence suggest that this method is applicable across different viral subtypes; however, further analysis with more samples of different subtypes is required to confirm this.

This study was limited by the number of samples available, which were all from patients already infected with HIV-1 at the time of HCV sampling, and primarily from men who have sex with men (MSM) with no recorded history of intravenous drug use (60/72 patients). The small sample size (and especially the small number of patients with large infection times) prevented the examination of non-linear effects, such as a potential saturation of diversity at large infection times, which has been observed for HIV-1 [14,18]. Furthermore, the presence of HIV-1 may affect the diversification of HCV over time, given the known impact of HIV-1 infection on HCV infection [31–33], and the increased virus load of HCV in the presence of HIV-1 [34–36]. The HCV epidemic typically consists primarily of people who inject drugs, but there is growing incidence of HCV infections among HIV-infected MSM [11,12,37–41]. As our study population consists almost exclusively of MSM, the validity of our results in people who inject drugs needs to be examined in future studies. Furthermore, our samples were derived from two different projects and therefore prepared and sequenced by slightly differing protocols. Whilst this variation in preparation and sequencing methods may remove some uniformity from our sample set, it more accurately reflects the scenario should the method be applied to HCV samples that have been collected and sequenced for a variety of purposes and can therefore be seen as a strength of this study.

As drug resistance testing is not routinely performed for HCV-infected patients, the availability of NGS sequences for this technique may be limited. However, as the price of NGS sequencing methods is decreasing, they are likely to be increasingly employed for surveillance purposes and gaining epidemiological understanding. Particularly in the context of elimination strategies, sequence-based information such as viral phylogenies and origins may be very desirable. Our study found a substantial correlation between time since infection and average pairwise diversity. Based on this, we could show that infection recency and time since infection can be accurately predicted by using average pairwise diversity. Thus, in the cases where samples from HCV-infected persons are sequenced, this method provides recency and date of infection information as free by-products of standard NGS sequencing, which can add to such monitoring studies.

Supplementary Materials: The following are available online at http://www.mdpi.com/1999-4915/12/11/1241/s1. Supplementary material S1: Full protocol for preparation and sequencing of samples for study 1; Table S1: Summary of preparations methods for sequencing for study 1; Table S2: Scores for average pairwise diversity calculated over all codon positions, and over various regions of the open reading frame, with associated coefficients from linear regression (AUC = area under the ROC curve, adj. R^2 = adjusted R^2, MAE = mean absolute error); Figure S1: Area under the ROC curve, R^2, and mean absolute error across the HCV open reading frame, third codon positions. The HCV open reading frame was split into 11 overlapping regions of approximately 500 amino acid codons, and average pairwise diversity (APD) was calculated over individual regions, using the third codon positions. Regions were tested for their ability to categorise infection as recent (<1 year) or chronic (top), their correlation with time since infection (middle), and their ability to infer time since infection (bottom). Black dashed lines show the respective values for APD calculated over the whole open reading frame. A similar analysis was

performed with diversity calculated over each gene in turn. The HCV genome is shown along the x-axis, with genes colour-coded for a composite (z-score sum) of all three outcome scores. Darker red indicates a better overall performance. Numbers along the x-axis refer to amino acid positions of the H77 reference genome; Figure S2: Area under the ROC curve, R^2, and mean absolute error across the HCV open reading frame, all codon positions. The HCV open reading frame was split into 5 overlapping regions of approximately 1000 amino acid codons, and average pairwise diversity (APD) was calculated over individual regions, using all codon positions. Regions were tested for their ability to categorise infection as recent (<1 year) or chronic (top), their correlation with time since infection (middle), and their ability to infer time since infection (bottom). Black dashed lines show the respective values for APD calculated over the whole open reading frame. A similar analysis was performed with diversity calculated over each gene in turn. The HCV genome is shown along the x-axis, with genes colour-coded for a composite (z-score sum) of all three outcome scores. Darker red indicates a better overall performance. Numbers along the x-axis refer to amino acid positions of the H77 reference genome; Equation S1: Z-score for converting outcome scores to be summed into a combined single score.

Author Contributions: L.A.C., H.F.G., and R.D.K. conceived and designed the study. K.J.M., H.A.M., C.S., J.B., M.H., D.L.B., J.F., L.S.-V., A.R., S.Y., A.N., M.C., M.S., P.V., and E.B. were responsible for the data acquisition and sequencing. L.A.C. and T.T. analysed the data. All authors interpreted the data. L.A.C. and R.D.K. prepared the paper. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Swiss National Science Foundation [grant number BSSGI0_155851]. HFG was supported by SNF grant 179571 and the University Research Priority Program "Evolution in Action" from the University of Zurich. Furthermore, this study has been financed within the framework of the Swiss HIV Cohort Study, supported by the Swiss National Science Foundation [grant number 177499], and by the Swiss HIV Cohort Study research foundation by the Yvonne Jacob Foundation (to HFG). HFG has received an unrestricted research grant from Gilead to the Swiss HIV Cohort Study Research Foundation. Viral sequencing was supported by Swiss HIV Cohort Study projects 688 and 726, and by the Swiss National Science Foundation [grant numbers 324730_179567 and 324730_146143] to AR.

Acknowledgments: We thank the patients who participate in the Swiss HIV Cohort Study; the physicians and study nurses, for excellent patient care; the resistance laboratories, for high-quality genotyping drug resistance testing; SmartGene (Zug, Switzerland), for technical support; Alexandra Scherrer, Susanne Wild, Anna Traytel from the SHCS data center for data management, Danièle Perraudin, Mirjam Minichiello and Marianne Amstutz for administration. The members of the Swiss HIV Cohort Study include the following: Anagnostopoulos, A., Battegay, M., Bernasconi, E., Böni, J., Braun, D.L., Bucher, H.C., Calmy, A., Cavassini, M., Ciuffi, A., Dollenmaier, G., Egger, M., Elzi, L., Fehr, J., Fellay, J., Furrer, H. (Chairman of the Clinical and Laboratory Committee), Fux, C.A., Günthard, H.F. (President of the SHCS), Haerry, D. (deputy of "Positive Council"), Hasse, B., Hirsch, H.H., Hoffmann, M., Hösli, I., Huber, M., Kahlert, C., Kaiser, L., Keiser, O., Klimkait, T., Kouyos, R.D., Kovari, H., Ledergerber, B., Martinetti, G., Martinez de Tejada, B., Marzolini, C., Metzner, K.J., Müller, N., Nicca, D., Paioni, P., Pantaleo, G., Perreau, M., Rauch, A. (Chairman of the Scientific Board), Rudin, C. (Chairman of the Mother & Child Substudy), Scherrer, A.U. (Head of Data Centre), Schmid, P., Speck, R., Stöckle, M., Tarr, P., Trkola, A., Vernazza, P., Wandeler, G., Weber, R., and Yerly, S.

Conflicts of Interest: HFG has received unrestricted research grants from Gilead Sciences and Roche; fees for data and safety monitoring board membership from Merck; and consulting/advisory board membership fees from Gilead Sciences, Sandoz and Mepha. EB has received fees for his institution for participation to advisory board from MSD, Gilead Sciences, ViiV Healthcare, Sandoz, Pfizer, Abbvie and Janssen. MC has received research and travel grants for his institution from ViiV and Gilead. AR reports support to his institution for advisory boards and/or travel grants from Janssen-Cilag, MSD, Gilead Sciences, Abbvie, and Bristol-Myers Squibb, and an unrestricted research grant from Gilead Sciences. All remuneration went to his home institution and not to AR personally, and all remuneration was provided outside the submitted work. KJM has received travel grants and honoraria from Gilead Sciences, Roche Diagnostics, GlaxoSmithKline, Merck Sharp & Dohme, Bristol-Myers Squibb, ViiV and Abbott; and the University of Zurich received research grants from Gilead Science, Roche, and Merck Sharp & Dohme for studies that Metzner serves as principal investigator, and advisory board honoraria from Gilead Sciences. RDK has received honoraria from Gilead Sciences (unrelated to the current work). JB received fees for his institution from Roche Glycart AG for consulting unrelated to the current work. DLB has received consulting/advisory board honoraria from Gilead Sciences, ViiV, and Merck Sharp & Dohme. MS received educational grants from Janssen-Cilag, MSD and Gilead, and Advisory Board fees from MSD, Gilead, AbbVie, ViiV, Sandoz and Mepha.

References

1. Wedemeyer, H.; Duberg, A.S.; Buti, M.; Rosenberg, W.M.; Frankova, S.; Esmat, G.; Ormeci, N.; Van Vlierberghe, H.; Gschwantler, M.; Akarca, U.; et al. Strategies to manage hepatitis C virus (HCV) disease burden. *J. Viral Hepat.* **2014**, *21*, 60–89. [CrossRef]

2. Armstrong, G.L.; Alter, M.J.; McQuillan, G.M.; Margolis, H.S. The past incidence of hepatitis C virus infection: Implications for the future burden of chronic liver disease in the United States. *Hepatology* **2000**, *31*, 777–782. [CrossRef] [PubMed]
3. Cotte, L.; Cua, E.; Reynes, J.; Raffi, F.; Rey, D.; Delobel, P.; Gagneux-Brunon, A.; Jacomet, C.; Palich, R.; Laroche, H.; et al. Hepatitis C virus incidence in HIV-infected and in preexposure prophylaxis (PrEP)-using men having sex with men. *Liver Int.* **2018**, *38*, 1736–1740. [CrossRef] [PubMed]
4. Ghosn, J.; Deveau, C.; Goujard, C.; Garrigue, I.; Saichi, N.; Galimand, J.; Nagy, Z.; Rouzioux, C.; Meyer, L.; Chaix, M.-L.; et al. Increase in hepatitis C virus incidence in HIV-1-infected patients followed up since primary infection. *Sex. Transm. Infect.* **2006**, *82*, 458–460. [CrossRef]
5. Giuliani, M.; Caprilli, F.; Gentili, G.; Maini, A.; Lepri, A.C.; Prignano, G.; Palamara, G.; Giglio, A.; Crescimbeni, E.; Rezza, G. Incidence and Determinants of Hepatitis C Virus Infection Among Individuals at Risk of Sexually Transmitted Diseases Attending a Human Immunodeficiency Virus Type 1 Testing Program. *Sex. Transm. Dis.* **1997**, *24*, 533–537. [CrossRef] [PubMed]
6. Jin, F.; Prestage, G.P.; Matthews, G.; Zablotska, I.; Rawstorne, P.; Kippax, S.C.; Kaldor, J.M.; Grulich, A.E. Prevalence, incidence and risk factors for hepatitis C in homosexual men: Data from two cohorts of HIV-negative and HIV-positive men in Sydney, Australia. *Sex. Transm. Infect.* **2009**, *86*, 25–28. [CrossRef] [PubMed]
7. Jordan, A.E.; Jarlais, D.C.D.; Arasteh, K.; McKnight, C.; Nash, D.; Perlman, D.C. Incidence and prevalence of hepatitis c virus infection among persons who inject drugs in New York City: 2006–2013. *Drug Alcohol Depend.* **2015**, *152*, 194–200. [CrossRef]
8. Law, M.G.; Dore, G.J.; Bath, N.; Thompson, S.; Crofts, N.; Dolan, K.; Giles, W.; Gow, P.; Kaldor, J.; Loveday, S.; et al. Modelling hepatitis C virus incidence, prevalence and long-term sequelae in Australia, 2001. *Int. J. Epidemiol.* **2003**, *32*, 717–724. [CrossRef]
9. Patrick, D.M.; Tyndall, M.W.; Cornelisse, P.G.; Li, K.; Sherlock, C.H.; Rekart, M.L.; Strathdee, S.A.; Currie, S.L.; Schechter, M.T.; O'Shaughnessy, M.V. Incidence of hepatitis C virus infection among injection drug users during an outbreak of HIV infection. *Can. Med. Assoc. J.* **2001**, *165*, 889–895.
10. Prevost, T.C.; Presanis, A.M.; Taylor, A.; Goldberg, D.J.; Hutchinson, S.J.; De Angelis, D. Estimating the number of people with hepatitis C virus who have ever injected drugs and have yet to be diagnosed: an evidence synthesis approach for Scotland. *Addiction* **2015**, *110*, 1287–1300. [CrossRef]
11. Rauch, A.; Martin, M.; Weber, R.; Hirschel, B.; Tarr, P.E.; Bucher, H.C.; Vernazza, P.; Bernasconi, E.; Zinkernagel, A.S.; Evison, J.; et al. Unsafe Sex and Increased Incidence of Hepatitis C Virus Infection among HIV-Infected Men Who Have Sex with Men: The Swiss HIV Cohort Study. *Clin. Infect. Dis.* **2005**, *41*, 395–402. [CrossRef]
12. Van De Laar, T.J.W.; Van Der Bij, A.K.; Prins, M.; Bruisten, S.M.; Brinkman, K.; Ruys, T.A.; Van Der Meer, J.T.M.; De Vries, H.J.C.; Mulder, J.; Van Agtmael, M.; et al. Increase in HCV Incidence among Men Who Have Sex with Men in Amsterdam Most Likely Caused by Sexual Transmission. *J. Infect. Dis.* **2007**, *196*, 230–238. [CrossRef] [PubMed]
13. Berg, C.V.D.; Smit, C.; Bakker, M.; Geskus, R.B.; Berkhout, B.; Jurriaans, S.; Coutinho, R.A.; Wolthers, K.C.; Prins, M. Major decline of hepatitis C virus incidence rate over two decades in a cohort of drug users. *Eur. J. Epidemiol.* **2007**, *22*, 183–193. [CrossRef] [PubMed]
14. Kouyos, R.D.; Von Wyl, V.; Yerly, S.; Böni, J.; Rieder, P.; Joos, B.; Taffé, P.; Shah, C.; Bürgisser, P.; Klimkait, T.; et al. Ambiguous Nucleotide Calls From Population-based Sequencing of HIV-1 are a Marker for Viral Diversity and the Age of Infection. *Clin. Infect. Dis.* **2011**, *52*, 532–539. [CrossRef] [PubMed]
15. Ragonnet-Cronin, M.; Aris-Brosou, S.; Joanisse, I.; Merks, H.; Vallée, D.; Caminiti, K.; Rekart, M.; Krajden, M.; Cook, D.; Kim, J.; et al. Genetic Diversity as a Marker for Timing Infection in HIV-Infected Patients: Evaluation of a 6-Month Window and Comparison With BED. *J. Infect. Dis.* **2012**, *206*, 756–764. [CrossRef]
16. Andersson, E.; Shao, W.; Bontell, I.; Cham, F.; Do, C.D.; Wondwossen, A.; Morris, L.; Hunt, G.; Sönnerborg, A.; Bertagnolio, S.; et al. Evaluation of sequence ambiguities of the HIV-1 pol gene as a method to identify recent HIV-1 infection in transmitted drug resistance surveys. *Infect. Genet. Evol.* **2013**, *18*, 125–131. [CrossRef]
17. Puller, V.I.; A Neher, R.; Albert, J. Estimating time of HIV-1 infection from next-generation sequence diversity. *PLoS Comput. Biol.* **2017**, *13*, e1005775. [CrossRef]

18. Carlisle, L.A.; Turk, T.; Kusejko, K.; Leemann, C.; Schenkel, C.D.; Bachmann, N.; Posada, S.; Beerenwinkel, N.; Anagnostopoulos, A.; Bucher, H.C.; et al. Viral Diversity Based on Next-Generation Sequencing of HIV-1 Provides Precise Estimates of Infection Recency and Time Since Infection. *J. Infect. Dis.* **2019**, *220*, 254–265. [CrossRef]
19. Braun, D.L.; Hampel, B.; Kouyos, R.; Nguyen, H.; Shah, C.; Flepp, M.; Stöckle, M.; Conen, A.; Béguelin, C.; Künzler-Heule, P.; et al. High Cure Rates With Grazoprevir-Elbasvir With or Without Ribavirin Guided by Genotypic Resistance Testing Among Human Immunodeficiency Virus/Hepatitis C Virus-coinfected Men Who Have Sex With Men. *Clin. Infect. Dis.* **2018**, *68*, 569–576. [CrossRef]
20. Huber, M.; Metzner, K.J.; Geissberger, F.D.; Shah, C.; Leemann, C.; Klimkait, T.; Böni, J.; Trkola, A.; Zagordi, O. MinVar: A rapid and versatile tool for HIV-1 drug resistance genotyping by deep sequencing. *J. Virol. Methods* **2017**, *240*, 7–13. [CrossRef]
21. Schirmer, M.; Ijaz, U.Z.; D'Amore, R.; Hall, N.; Sloan, W.T.; Quince, C. Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform. *Nucleic Acids Res.* **2015**, *43*, e37. [CrossRef]
22. A Quail, M.; Smith, M.; Coupland, P.; Otto, T.D.; Harris, S.R.; Connor, T.R.; Bertoni, A.; Swerdlow, H.; Gu, Y. A tale of three next generation sequencing platforms: Comparison of Ion torrent, pacific biosciences and illumina MiSeq sequencers. *BMC Genom.* **2012**, *13*, 341. [CrossRef] [PubMed]
23. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2018; Available online: https://www.R-project.org/ (accessed on 1 December 2018).
24. Dowle, M.; Srinivasan, A. Data.Table: Extension of 'Data.Frame'. 2017. Available online: https://CRAN.R-project.org/package=data.table (accessed on 1 December 2018).
25. Robin, X.A.; Turck, N.; Hainard, A.; Tiberti, N.; Lisacek, F.; Sanchez, J.-C.; Muller, M.J. pROC: An open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinform.* **2011**, *12*, 77. [CrossRef]
26. Garbuszus, J.M.; Jeworutzki, S. Readstata13: Import "Stata" Data Files. 2017. Available online: https://CRAN.R-project.org/package=readstata13 (accessed on 1 December 2018).
27. Neuwirth, E. RColorBrewer: ColorBrewer Palettes. 2014. Available online: https://CRAN.R-project.org/package=RColorBrewer (accessed on 1 December 2018).
28. Gaudieri, S.; Rauch, A.; Pfafferott, K.; Barnes, E.; Cheng, W.; McCaughan, G.; Shackel, N.; Jeffrey, G.P.; Mollison, L.; Baker, R.; et al. Hepatitis C virus drug resistance and immune-driven adaptations: Relevance to new antiviral therapy. *Hepatology* **2008**, *49*, 1069–1082. [CrossRef]
29. Stewart, H.; Bingham, R.; White, S.J.; Dykeman, E.C.; Zothner, C.; Tuplin, A.K.; Stockley, P.G.; Twarock, R.; Harris, M. Identification of novel RNA secondary structures within the hepatitis C virus genome reveals a cooperative involvement in genome packaging. *Sci. Rep.* **2016**, *6*, 22952. [CrossRef]
30. Fricke, M.; Gerst, R.; Ibrahim, B.; Niepmann, M.; Marz, M. Global importance of RNA secondary structures in protein-coding sequences. *Bioinformatics* **2019**, *35*, 579–583. [CrossRef] [PubMed]
31. Mastroianni, C.M.; Lichtner, M.; Mascia, C.; Zuccalà, P.; Vullo, V. Molecular Mechanisms of Liver Fibrosis in HIV/HCV Coinfection. *Int. J. Mol. Sci.* **2014**, *15*, 9184–9208. [CrossRef]
32. Hernandez, M.D.; Sherman, K.E. HIV/HCV coinfection natural history and disease progression, a review of the most recent literature. *Curr. Opin. HIV AIDS* **2011**, *6*, 478. [CrossRef] [PubMed]
33. Benhamou, Y.; Bochet, M.; Di Martino, V.; Charlotte, F.; Azria, F.; Coutellier, A.; Vidaud, M.; Opolon, P.; Katlama, C.; Poynard, T.; et al. Liver fibrosis progression in human immunodeficiency virus and hepatitis C virus coinfected patients. *Hepatology* **1999**, *30*, 1054–1058. [CrossRef]
34. Briat, A.; Dulioust, E.; Galimand, J.; Fontaine, H.; Chaix, M.-L.; Letur-Könirsch, H.; Pol, S.; Jouannet, P.; Rouzioux, C.; Leruez-Ville, M. Hepatitis C virus in the semen of men coinfected with HIV-1: Prevalence and origin. *Aids* **2005**, *19*, 1827–1835. [CrossRef]
35. Hsieh, M.-H.; Tsai, J.-J.; Hsieh, M.-Y.; Huang, C.-F.; Yeh, M.-L.; Yang, J.-F.; Chang, K.; Lin, W.-R.; Lin, C.-Y.; Chen, T.-C.; et al. Hepatitis C Virus Infection among Injection Drug Users with and without Human Immunodeficiency Virus Co-Infection. *PLoS ONE* **2014**, *9*, e94791. [CrossRef]
36. Hagan, H.; Jordan, A.E.; Neurer, J.; Cleland, C.M. Incidence of sexually transmitted hepatitis C virus infection in HIV-positive men who have sex with men: A systematic review and meta-analysis. *AIDS* **2015**, *29*, 2335–2345. [CrossRef] [PubMed]
37. Bottieau, E.; Apers, L.; Van Esbroeck, M.; Vandenbruaene, M.; Florence, E. Hepatitis C virus infection in HIV-infected men who have sex with men: Sustained rising incidence in Antwerp, Belgium, 2001–2009. *Eurosurveillance* **2010**, *15*, 19673. [PubMed]

38. Sanchez, C.; Plaza, Z.; Vispo, E.; De Mendoza, C.; Barreiro, P.; Fernández-Montero, J.V.; Labarga, P.; Poveda, E.; Soriano, V. Scaling up epidemics of acute hepatitis C and syphilis in HIV-infected men who have sex with men in Spain. *Liver Int.* **2013**, *33*, 1357–1362. [CrossRef]
39. Van De Laar, T.J.W.; Matthews, G.V.; Prins, M.; Danta, M. Acute hepatitis C in HIV-infected men who have sex with men: An emerging sexually transmitted infection. *Aids* **2010**, *24*, 1799–1812. [CrossRef]
40. Yaphe, S.; Bozinoff, N.; Kyle, R.; Shivkumar, S.; Pai, N.P.; Klein, M. Incidence of acute hepatitis C virus infection among men who have sex with men with and without HIV infection: A systematic review. *Sex. Transm. Infect.* **2012**, *88*, 558–564. [CrossRef]
41. Wandeler, G.; Gsponer, T.; Bregenzer, A.; Günthard, H.F.; Clerc, O.; Calmy, A.; Stöckle, M.; Bernasconi, E.; Furrer, H.; Rauch, A. Hepatitis C Virus Infections in the Swiss HIV Cohort Study: A Rapidly Evolving Epidemic. *Clin. Infect. Dis.* **2012**, *55*, 1408–1416. [CrossRef]