

Quantitative differences between intra-host HCV populations from persons with recently established and persistent infections

Pelin Icer Baykal¹, James Lara², Alex Zelikovsky¹, Yury Khudyakov², and Pavel Skums^{1,3}

¹Department of Computer Science, Georgia State University, Atlanta, GA, USA

²Division of Viral Hepatitis, Centers for Disease Control and Prevention, Atlanta, GA, USA

³Email: pskums@gsu.edu, Phone number: 404 413-6654

Summary

Background Detection of incident hepatitis C virus (HCV) infections is crucial for identification of outbreaks and development of public health interventions. However, there is no single diagnostic assay for distinguishing recent and persistent HCV infections. HCV exists in each infected host as a heterogeneous population of genomic variants, whose evolutionary dynamics remain incompletely understood. Genetic analysis of such viral populations can be applied to the detection of incident HCV infections and used to understand intra-host viral evolution.

Methods We studied intra-host HCV populations sampled using next-generation sequencing from 98 recently and 256 persistently infected individuals. Genetic structure of the populations was evaluated using 245,878 viral sequences from these individuals and a set of selected parameters measuring their diversity, topological structure, complexity, strength of selection, epistasis, evolutionary dynamics, and physico-chemical properties.

Findings Distributions of the viral population parameters differ significantly between recent and persistent infections. A general increase in viral genetic diversity from recent to persistent infections is frequently accompanied by decline in genomic complexity and increase in structuredness of the HCV population, likely reflecting a high level of intra-host adaptation at later stages of infection. Using these findings, we developed a Machine Learning classifier for the infection staging, which yielded a detection accuracy of 95.22%, thus providing a higher accuracy than other genomic-based models.

Interpretation The detection of a strong association between several HCV genetic factors and stages of infection suggests that intra-host HCV population develops in a complex but regular and predictable manner in the course of infection. The proposed models may serve as a foundation of cyber-molecular assays for staging infection, that could potentially complement and/or substitute standard laboratory assays.

Funding AZ and PS were supported by NIH grant 1R01EB025022. PIB was supported by GSU MBD fellowship.

1. Introduction

Hepatitis C virus (HCV) infection remains a major cause of morbidity and mortality, with an estimated 70 million people being HCV infected worldwide in 2015¹. HCV infection is the leading cause of chronic liver diseases and hepatocellular carcinoma worldwide, contributing to the death of more than 350,000 people in 2015¹. Hepatitis C outbreaks continue to occur, posing a serious challenge to public health². HCV is highly mutable. As a result, each infected individual hosts a heterogeneous population of genetically related HCV variants or *quasispecies*³. Substantial diversity of intra-host viral populations plays a crucial role in disease progression and epidemic spread⁴⁻⁶. However, intra-host dynamics of HCV and other RNA viruses remain poorly understood. One of the most important questions is the relative contribution of random and deterministic evolutionary factors in disease progression, or, using the metaphor of S.J. Gould⁷, whether is it possible to “replay the tape of life” for the virus evolution inside a host. This question is of high importance for biomedical research, as predictability of viral evolution potentially implies the power to understand and control the disease^{8,9}, which may result in advanced diagnostic and treatment strategies.

In this paper, we study evolutionary factors associated with the transition between HCV infection stages. In more than 50% of cases untreated HCV infection proceeds to the chronic phase, which can lead to the development of liver cirrhosis and/or hepatocellular carcinoma⁸. Accurate recent or persistent staging of HCV infection is important for biomedical applications. In clinical settings, it may inform the patient management and treatment strategy. In epidemiology, identification of acute cases allows for detection and investigation of recent transmissions and outbreaks and provides information on disease incidence. Understanding of changes in intra-host HCV populations at different stages of infection would constitute a large step towards reliable forecasting of viral evolutionary dynamics.

Recent HCV infection is usually accessed using clinical symptoms and time since seroconversion. HCV infection may, however, remain asymptomatic for years while seroconversion is not frequently detected, preventing accurate identification of infection stages. Several laboratory methods have been reported for distinguishing acute and chronic stages of infection^{10,11}. Detection of HCV RNA in the absence of anti-HCV activity in serum specimens was used as an indication of recent HCV infection¹². Although a strong marker, it has a very short duration and cannot be used for reliable detection of acute infections.

Advent of next-generation sequencing (NGS) presented an opportunity to sample and analyze unprecedented large numbers of intra-host viral variants from numerous infected individuals. HCV variants sampled by NGS have been used to detect stages of HCV infection^{13,14}. The stage detection methods are generally based on the assumption that intra-host viral evolution is driven by the continuous immune escape resulting in genetic diversification. Consequently, quantitative measures of genetic diversity of intra-host viral variants are assumed to be most useful for staging. However, several recent reports contested the veracity of this assumption. In particular, after initial diversification, intra-host HCV populations may actually lose heterogeneity and stop diversifying at later stages of infection^{5,15} with certain viral variants persisting in infected hosts for years^{5,16}. Furthermore, this process is accompanied by increase of negative selection over the course of HCV infection^{5,15,17,18}. These findings suggest a high level of intra-host adaptation at late stages of infection⁴ and indicate that genetic heterogeneity is not a reliable marker for infection staging, and more elaborate metrics are needed to understand HCV evolution and to accurately classify recent and persistent HCV infection.

Here, we present a new approach for staging HCV infection using quantitative genomic measures to evaluate diversity, information content, effective dimensionality, topological structure, evolutionary dynamics and physical-chemical properties of intra-host HCV variants and populations. Analysis of parameters' distributions at early and late stages of infection suggests that intra-host HCV populations evolve in a complex but regular and predictable manner. Based on these findings, we propose a multi-parameter machine learning classifier for staging HCV infection. The model allows for more accurate detection of recent HCV infection than models based only on population diversity and provides new insights into mechanisms of infection progression.

2. Materials and Methods

2.1 Data Collection and Preprocessing

We analyzed intra-host HCV populations sampled from recently (N=98) and persistently (N=256) infected persons collected as described in³⁵. The E1/E2 junction of the HCV genome ($L = 246\text{nt}$), which contains the hypervariable region 1 (HVR1), was sequenced using the GS FLX System and the GS FLX Titanium Sequencing Kit (454 Life Sciences, Roche, Branford, CT). Obtained sequences were processed using the error correction and haplotyping algorithm KEC¹⁹, which produced 245,878 unique viral haplotypes with frequencies.

2.2 Parameters Calculation

The analyzed parameters could be loosely split into four groups: genomic parameters, complexity parameters, network parameters and biochemical parameters. We assumed that a given intra-host population contains n unique haplotypes with frequencies f_1, \dots, f_n . Sixteen parameters corresponding to this population constitutes its *feature vector*.

Genomic Parameters

These parameters are obtained by direct comparison of sequences from each population. *Distance-based parameters* include *mean and standard deviation* of pairwise hamming distance distribution, and the *conservation score* of the population consensus sequence calculated with the NUC44 scoring matrix. We also utilized the so-called *mutation frequency* parameter,¹³ which is defined as the mean distance between all haplotypes and the most frequent haplotype. All four parameters measure the population diversity.

Diversity was also quantified using three *entropy-based parameters*. For a genomic position i , its positional k -entropy is defined as the entropy of the frequency distribution of k -mers (subsequences of length k) starting at i . An *average positional k -mer entropy* E_k is the mean of positional k -entropies over all positions:

$$E_k = \frac{1}{L - k + 1} \sum_{i=1}^{L-k+1} \sum_{h \in H_i} -f_i(h) \log_2(f_i(h)) \quad (1)$$

Here h is a k -mer, H_i is the set of k -mers starting from the i -th position and $f_i(h)$ refers to the relative frequency of h inside H_i . For $k = L$ the parameter E_L is an entropy of observed haplotype frequencies, while for $k = 1$ it is an average position-wise nucleotide entropy. In our model we used entropies E_1 , E_L and E_{10} .

Next, we estimated the frequency of *transversions* (mutations between purines and pyrimidines) among all observed mutations within the population. This parameter is suggested by previous studies²⁰ that reported higher frequencies of transitions over transversions in HCV populations. *Selective Pressure* was measured using the DN/DS ratio, which has been calculated as the ratio of rates of non-synonymous (DN) and synonymous (DS) substitutions with respect to the most frequent genomic variant.

2.3 Complexity Parameters

PCA complexity is derived from principal component analysis (PCA). For each population, its alignment is transformed into $n \times L$ numerical matrix, and the complexity is defined as the percentage of principal components required to explain at least $\alpha = 50\%$ percent of the observed genetic variance. PCA complexity measures the effective dimensionality of the population as the multidimensional system.

Kolmogorov complexity is the classical concept of information theory, which quantifies the descriptive/information complexity of a string over a finite alphabet. Informally it is defined as the highest possible degree of compression of a given string without loss of information. Although the exact value of Kolmogorov complexity is algorithmically incomputable, it can be efficiently approximated using data compression techniques. In our case, each viral sequence has been transformed into a binary string, the strings have been concatenated, and Kolmogorov complexity of the resulting string has been estimated by a variant of Lempel-Ziv algorithm²¹.

2.3.1 Network Parameters

This group of parameters is derived from the analysis of *genetic networks* of HCV populations, that represent a *sequence space*²⁰ of a virus. Formally, for each patient its genetic network $G_N = (V, E)$ is a graph, whose vertices V represent sampled viral haplotypes, and edges E connect variants which differ by at most T mutations (by default $T = 1$) (see Fig. 1). With each vertex we associate the frequency of the corresponding haplotype. In the case of a large population size accompanied by a high mutation rate and a fast reproduction time, genetic networks constructed using NGS data represent population structures significantly more accurately than phylogenetic trees¹⁷. Their structure is shaped by various factors, such as epistasis, founder effects, and selection pressures that affect the virus over the course of infection^{5,22}. For each network, the following four parameters have been calculated.

Robustness/selection balance has been measured by the correlation between vectors of vertex frequencies and eigenvector centralities. The latter is the principal eigenvector of the adjacency matrix of G_N . In the classical quasispecies model, vertex centralities are indicative of the mutational robustness of corresponding viral variants²³, while a high frequency may be indicative of a higher fitness.

Topological structures of genetic networks have been assessed using two parameters. The first of them is *s-metric*²⁴ $s(G_N) = \sum_{(i,j) \in E} d_i d_j$, which measures how close a network is to being scale-free (here, d_i is a degree (number of neighbors) of a vertex i). Scale-free networks are ubiquitous

in biological and social systems and share specific properties such as a power-law degree distribution, small diameter and presence of hubs. To account for variable sample sizes, $\langle \text{deg} \rangle$ is normalized by the factor $\langle \text{deg} \rangle^2$ (the order of magnitude of the maximum ℓ_1 -metric for n -vertex network).

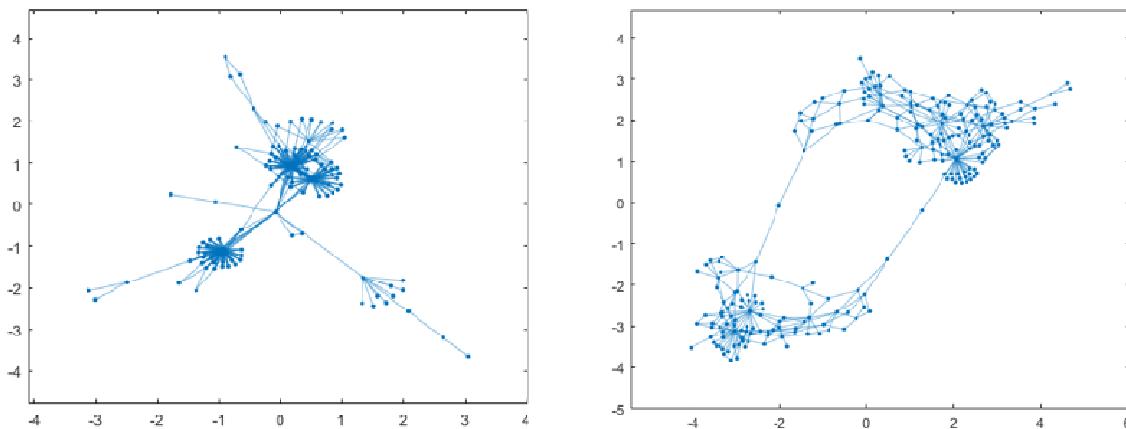


Figure 1: Examples of genetic viral networks for a recently infected (left) and a persistently infected (right) individual. The viral network of the recently infected host has the structural properties typical for scale-free networks.

The second network structural parameter is the clustering coefficient, which measures the degree to which network vertices tend to cluster together. It is defined as the probability that a random connected vertex triplet is complete (i.e. every pair of vertices is connected by an edge).

Evolutionary dynamics parameter estimates an age of the genetic network using an evolutionary model. Given n viral variants, we simulate their frequencies $\{f_i\}$ using a system of ordinary differential equations (S1)–(S3), which describes the interaction of the viral population with the host’s immune system (Supplemental Section S1). We classify populations as recent or persistent based on the qualitative behavior of the function describing the deviation of simulated and observed frequencies over time. Formally, we define an estimated population age as the time τ , when simulated viral frequencies $\{f_i\}$ achieve the best agreement with observed frequencies, i.e., $\text{JSD}(\{f_i\}, \{g_i\})$ is a Jensen-Shannon divergence between distributions $\{f_i\}$ and $\{g_i\}$. Owing to the inherent uncertainty of the quantitative parameters of the model, rather than using τ as a prediction variable we utilize qualitative characteristics of the divergence function $\text{JSD}(\{f_i\}, \{g_i\})$. Namely, persistent and recent populations are characterized by divergence functions with descending and ascending trends, respectively (Fig. S1, see Supplemental Section S1). The classification is performed separately for each connected component of the genetic network, and the patient is classified as persistently infected (the parameter $\tau > 0$), if at least one of the components is persistent, and as recently infected ($\tau \leq 0$), otherwise.

2.3.2 Biochemical parameter

For each viral sequence, we assess whether this sequence has physico-chemical properties associated with recent or persistent infection. The biochemical index of an intra-host population is thus defined as the sum of frequencies of variants identified as having a physico-chemical profile pointing to persistent infection.

The method for evaluation of the properties of a given viral haplotype is described in detail in Supplemental Section S2. Briefly, for a given sequence we construct its biochemical feature profile using the following physico-chemical indexes of DNA dimers: the thermodynamic indexes (Breslauer-dH and Breslauer-dG), structural indexes (twist-tilt, slide-rise, protein-DNA twist, slide-2, twist-1), the nucleotide composition index (G-content) and the energy indexes of DNA (stabilizing energy of Z DNA and enthalpy)²⁵. Such set of parameters can accurately measure changes in structure-function relationships and can be used to predict a broad range of biological and biochemical properties of DNA/RNA biomolecules²⁵. The constructed set of features is processed by the problem-specific dimensionality reduction and feature selection pipeline, and binarized. The obtained binary feature vector representations of intra-host HCV variants were used as input data to train a stochastic gradient descent (SGD) classifier²⁶. The SGD classifier implements regularized linear models with stochastic gradient descent (SGD) learning and is a very efficient approach, with linear training cost, which can easily be scaled to big data problems. Selection and tuning of the hyperparameters of the SGD classifier was done using a balanced training set (1,968 and 1,965 feature vectors for sequences sampled from recently and persistently infected hosts) and assessed by five-fold cross-validation.

2.4 Machine Learning Classifier

Feature vectors of recently and chronically infected hosts were used to train machine learning classifiers for infection stage prediction. Given a labeled training set comprising feature vectors together with their class labels (recent or persistent), each classifier is fitted to the training data by adjusting its model parameters and assigns labels for unlabeled feature vectors using the trained model. In this study, we used Support Vector Machines with linear and polynomial kernel and Logistic Regression. Both approaches are classical supervised learning methods that construct a hyperplane in the multidimensional Euclidean space, which serves as a separator for feature vectors from classes of recently and persistently infected hosts.

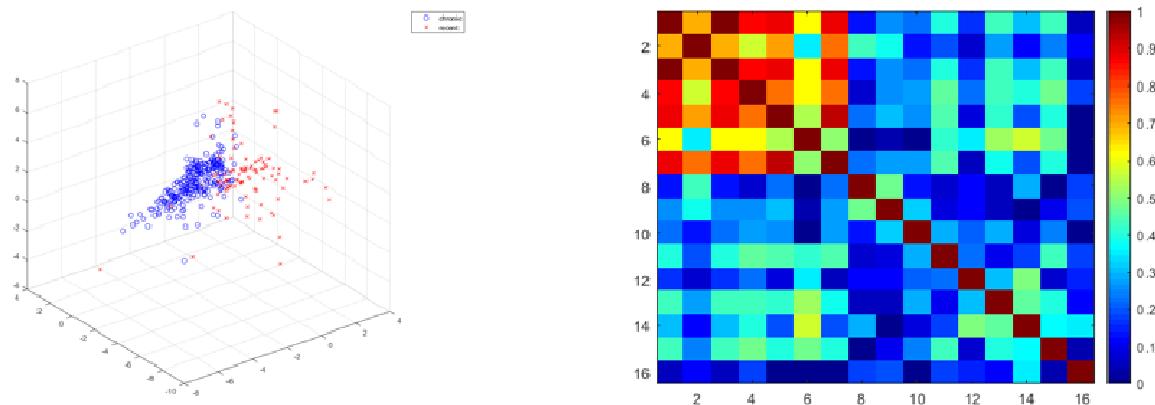


Figure 2: Left: 3-D projection of feature vectors of recently and persistently infected hosts (with highly correlated features removed) constructed by multidimensional scaling. Right: heatmap of absolute values of pairwise correlations between parameters

3 Results

3.1 Stage-specific distributions of parameters.

Except for several diversity measures (k -entropy, site entropy, mean distance, conservation score and mutation frequency), there is a small-to-medium correlation between the parameters (Fig. 2), demonstrating that they reflect different properties of intra-host viral populations.

Parameters	p-value	Persistently infected		Recently infected	
		Mean	95% CI	Mean	95% CI
1. Mean distance	1·41E-24	0·034	(0·031, 0·037)	0·015	(0·012, 0·017)
2. Std distance	4·10E-14	0·019	(0·018, 0·021)	0·010	(0·007, 0·013)
3. Conservation score	5·11E-25	0·422	(0·393, 0·451)	0·188	(0·155, 0·221)
4. Mutation frequency	3·07E-16	0·023	(0·021, 0·026)	0·010	(0·007, 0·013)
5. k -entropy	7·00E-23	0·630	(0·602, 0·658)	0·357	(0·323, 0·392)
6. Frequency entropy	4·56E-06	0·668	(0·648, 0·689)	0·567	(0·527, 0·607)
7. SNV entropy	1·14E-21	0·084	(0·079, 0·089)	0·043	(0·037, 0·048)
8. Transversion mutation	1·06E-07	0·061	(0·054, 0·068)	0·032	(0·027, 0·038)
9. DN/DS	5·39E-10	0·713	(0·626, 0·779)	1·330	(1·096, 1·565)
10. PCA complexity	4·91E-04	0·014	(0·012, 0·026)	0·034	(0·023, 0·045)
11. Kolmogorov complexity	1·55E-11	0·041	(0·040, 0·043)	0·052	(0·048, 0·056)
12. Robustness/Selection	3·66E-15	0·628	(0·608, 0·647)	0·386	(0·329, 0·442)
13. s-metric	1·93E-20	0·001	(0, 0·002)	0·044	(0·023, 0·065)
14. Clustering coefficient	2·08E-13	0·082	(0·064, 0·100)	0·356	(0·292, 0·420)
15. ODE	2·42E-06	-0·270	(-0·378, -0·162)	0·224	(0·055, 0·394)
16. Biochemical parameter	2·92E-35	0·628	(0·614, 0·642)	0·379	(0·354, 0·403)

Table 1: Parameters with a significant association to the infection stage. The columns contain p-values of U-test, mean values, and 95% confidence intervals for viral populations among persistently and recently infected persons.

Feature vectors of recent and persistent populations are separable from each other (Fig. 2). For each parameter, Mann-Whitney U-test suggests statistically significant difference between recent and persistent intra-host populations (Table 1).

As expected, diversities are on average higher for persistent than recent populations (p -values between $1.41 \cdot 10^{-24}$ and $4.56 \cdot 10^{-6}$; Fig. 3 (1-7)). Higher genetic diversity of persistent populations is accompanied by significantly lower PCA and Kolmogorov complexities ($p = 4.91 \cdot 10^{-4}$ and $p = 1.55 \cdot 10^{-11}$; Fig. 3 (10-11)). This could be explained by the role of intra-host adaptation during the later stage of infection, when genomes are highly specific to the environment and SNVs selected over the course of intra-host evolution are highly interdependent, thus reducing the effective dimensionality of the population. It is known that high Kolmogorov complexity indicates high level of randomness of a sequence, while low complexity implies the presence of specific structural patterns inside a sequence. Thus, lower Kolmogorov complexity at later stages of disease suggests the increase in strength of epistatic connectivity among nucleotide positions during intra-host evolution and points to a higher level of adaptation and specialization of members of intra-host populations. At the earlier stages of infection, nucleotide changes are seemingly more random, resulting in populations with higher dimensionality. Increase in negative selection additionally contributes to the reduction of dimensionality and complexity at later stages of HCV infection ($p = 5.39 \cdot 10^{-10}$; Fig. 3 (9)).

Transition mutations were overwhelmingly more frequent than transversion mutations for both classes of samples. This fact agrees with the previously published results²⁰, although the magnitude of difference vary along the genome: HVR1 transitions are ~ 18 times more frequent than transversions, while a 75-fold difference was reported for NS5B²⁰. Prevalence of transversions was ~ 2 times higher in persistent populations ($p = 1.06 \cdot 10^{-7}$; Fig. 3 (8)). This phenomenon could be interpreted as another reflection of increasing intra-host adaptation over the course of infection. Indeed, transversions represent a higher genetic barrier for the selection of escape mutants from HCV-specific immune responses²⁰. Thus, growth of transversion frequencies at later evolutionary stages may mark a declining role of immune escape and a growing role of other evolutionary mechanisms such as adaptation by antigenic cooperation⁴.

Genetic networks of recent and persistent intra-host populations possess different structural properties. Networks of recent populations have significantly higher s -metrics and clustering coefficients ($p = 1.93 \cdot 10^{-20}$ and $p = 2.08 \cdot 10^{-13}$; Fig. 3 (13-14)). It indicates that, in contrast to the persistent populations, they tend to have structural properties more typical for scale-free networks, including the power-law degree distribution with clearly manifested hubs (high-degree vertices), with their vertices having propensity to cluster (Fig.1). This observation can be explained by the role of founder viral variants at the earlier stage of infection. A significantly higher correlation between frequencies and network centralities of variants in persistent populations ($p = 3.66 \cdot 10^{-15}$; Fig. 3 (12)) indicates that the population structure at later stages is significantly influenced by mutational robustness, while at earlier stages it is basically defined by founders. Recent and persistent HCV populations are also separable by an evolutionary dynamic parameter c_{ODE} ($p = 2.42 \cdot 10^{-6}$; Fig. 3 (15)).

Finally, individual sequences of recent and persistent populations have distinct physico-chemical properties ($p = 2.92 \cdot 10^{-35}$; Fig. 3 (16)). It suggests that the physico-chemical property of HVR1 is influenced by, and is responsive to, within-host environmental factors specific to the recent and persistent stages of HCV infection.

3.2 Machine Learning Classification.

Mutation frequency, k -entropy and frequency entropy have been excluded from the prediction model as they are highly correlated with other parameters. The remaining 13 parameters were used to train Support Vector Machines (SVM) and Logistic Regression classifiers for binary classification of intra-host viral populations labelled as "persistent" and "recent". Accuracy of classifiers has been assessed using a two-step cross-validation. First, to account for the bias associated with unequal numbers of cases with persistent ($n = 256$) and recent ($n = 98$) infection, repeated random subsampling of 98 populations from the persistent sample dataset was performed. For each of the balanced training sets 10-fold cross-validation was carried out.

The average prediction accuracies are reported in Table 2. Classification performance evaluation of all methods indicates a high accuracy of infection stage inference, with SVM with quadratic kernel demonstrating the highest accuracy of 95.22%.

Method	Prediction Accuracy
SVM – linear kernel	95.17%
SVM – quadratic kernel	95.22%
Logistic regression	93.17%

Table 2: HCV infection stage prediction accuracies of machine learning methods

SVM classifier with quadratic kernel has been compared to the previously published HCV infection staging models¹³ which classify intra-host viral populations as recent or persistent using frequency entropy, SNV entropy or mutation frequency. The ROC curves of the classifiers are shown in Fig. 4. Previously proposed methods ($AUROC = 0.81, 0.66$ and 0.78 , respectively) were less accurate in comparison with the SVM classifier ($AUROC = 0.99$), thus suggesting that diversity parameters alone are not sufficient for accurate distinction between recent and persistent cases. SVM classifier performed at the expected lower accuracy on randomly labelled datasets (average $AUROC = 0.4966$), thus indicating that the associations between parameter distributions and infection stages are likely due to the structural and evolutionary factors rather than to random statistical correlations in the data.

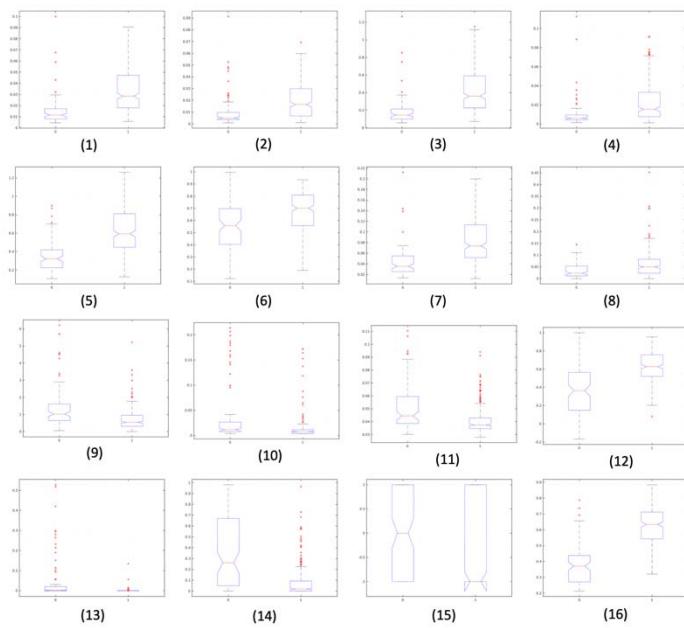


Figure 3: Box plots of parameter distributions for recent (left box plot on each graph) and persistent (right box plot on each graph) intra-host HCV populations. The plots are in the same order as in Table 1.

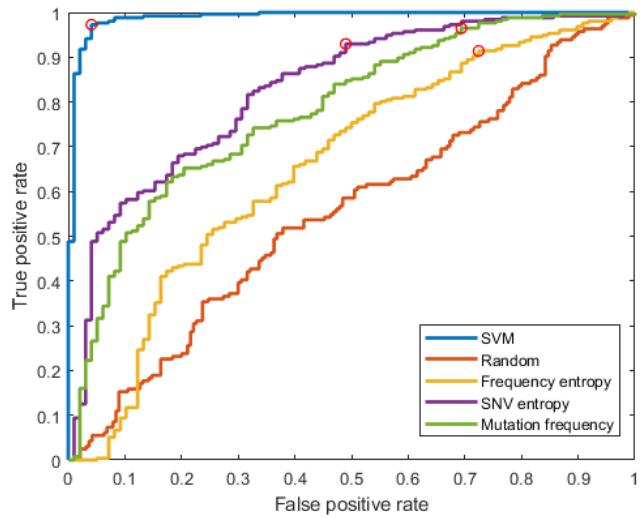


Figure 4: ROC Curves of classification models

4 Discussion

We present the results of comprehensive analyses of the structure of intra-host viral populations using a large set of samples from individuals with recent and persistent infection, which significantly exceeds data sets used in earlier studies¹³. Amplicons covering HCV HVR1 have been sequenced by NGS. Intrinsically disordered regions (IDR) of proteins like HVR1 seem to be most useful for application in models to identify viral clinical properties from sequences. It has an extensive epistatic connectivity across the entire HCV polyprotein²⁷, and is associated with immune escape²⁸, drug resistance^{27,29} and virulence³⁰. Consequently, IDRs play an

important role in viral adaptation to the host environment, making regions like HVR1 sensitive "sensors" that accurately reflect intra-host biological changes during the infection process.

We identified a set of quantitative characteristics of intra-host HCV populations strongly associated with stages of infection. Our results indicate significant differences in the structure of HCV populations sampled from recently and persistently infected hosts. Models constructed using these parameters allowed us to train machine learning classifiers capable of inferring infection stage from HCV sequence data with accuracies as high as 95%. Our study confirms a previously established positive correlation between infection stage and intra-host viral diversity^{13,10,31}. However, because of complexities in the structural development of intra-host populations affected by bouts of selective sweeps and negative selection during chronic infection^{4,32}, simple metrics of genetic heterogeneity are insufficient for the accurate staging of HCV infections. High accuracy could be achieved by using a combination of parameters measuring different structural and evolutionary properties of viral populations.

The proposed prediction models may serve as *cyber-molecular assays* for staging infection, that could potentially complement and substitute standard laboratory assays. In particular, the proposed models are currently being incorporated into Global Hepatitis Outbreak and Surveillance Technology (GHOST)³³ — a web-based molecular surveillance system developed and maintained by CDC. They could also be applicable to other highly mutable viruses, such as HIV.

The detection of a strong association between several HCV genetic factors and stages of infection suggests that intra-host HCV populations develop in a complex but regular and predictable manner during the course of infection. Decline in dN/dS, increase in abundance of transversion mutations and decline in information complexity of HCV population progressing from the recent to persistent state is consistent with an orderly process of HCV population development during infection as was suggested earlier^{4,5} and is different from a model of an "arms race" predicting a continuous genetic diversification. These observations support the hypothesis that intra-host viral populations may evolve as quasi-social systems by complementary specialization of viral variants engaged in a certain type of cooperation^{4,34}. Such specialization enables HCV populations to adapt to an intra-host environment as a group of cooperators rather than independent variants.

Ethics approval and consent to participate

Research was conducted as approved by the Institutional Review Board of the Centers for Disease Control and Prevention, Atlanta, GA (protocol 7270.0).

Disclaimer

The findings and conclusions in this report are those of the authors and do not necessarily represent the official position of the Centers for Disease Control and Prevention and Georgia State University.

Availability of data and material

The proposed method's scripts are available in the following GitHub repository <https://github.com/compbel/recentvschronic>. The data can be requested from the CDC.

Competing interests

The authors declare they don't have any competing interests.

References

1. Polaris Observatory HCV Collaborators. Global prevalence and genotype distribution of hepatitis C virus infection in 2015: a modelling study. *Lancet Gastroenterol Hepatol*. 2017

Mar;2(3):161–76.

2. Zibbell JE, Iqbal K, Patel RC, Suryaprasad A, Sanders KJ, Moore-Moravian L, et al. Increases in hepatitis C virus infection related to injection drug use among persons aged ≤30 years - Kentucky, Tennessee, Virginia, and West Virginia, 2006-2012. *MMWR Morb Mortal Wkly Rep.* 2015 May 8;64(17):453–8.
3. Domingo E, Sheldon J, Perales C. Viral quasispecies evolution. *Microbiol Mol Biol Rev.* 2012 Jun;76(2):159–216.
4. Skums P, Bumimovich L, Khudyakov Y. Antigenic cooperation among intrahost HCV variants organized into a complex network of cross-immunoreactivity. *Proc Natl Acad Sci U S A.* 2015 May 26;112(21):6653–8.
5. Ramachandran S, Campo DS, Dimitrova ZE, Xia G-L, Purdy MA, Khudyakov YE. Temporal variations in the hepatitis C virus intrahost population during chronic infection. *J Virol.* 2011 Jul;85(13):6369–80.
6. Campo DS, Xia G-L, Dimitrova Z, Lin Y, Forbi JC, Ganova-Raeva L, et al. Accurate Genetic Detection of Hepatitis C Virus Transmissions in Outbreak Settings. *J Infect Dis.* 2016 Mar 15;213(6):957–65.
7. Gould SJ. *Wonderful Life: The Burgess Shale and the Nature of History.* W. W. Norton & Company; 1990. 352 p.
8. Seo S, Silverberg MJ, Hurley LB, Ready J, Saxena V, Witt D, et al. Prevalence of spontaneous clearance of hepatitis C virus infection doubled from 1998 to 2017. *Clin Gastroenterol Hepatol* 2020;18:511–513.
9. Lässig M, Mustonen V, Walczak AM. Predicting evolution. *Nat Ecol Evol.* 2017 Feb 21;1(3):77.
10. Araujo AC, Astrakhantseva IV, Fields HA, Kamili S. Distinguishing Acute from Chronic Hepatitis C Virus (HCV) Infection Based on Antibody Reactivities to Specific HCV Structural and Nonstructural Proteins [Internet]. Vol. 49, *Journal of Clinical Microbiology.* 2011. p. 54–7. Available from: <http://dx.doi.org/10.1128/jcm.01064-10>
11. Bowen DG, Walker CM. Adaptive immune responses in acute and chronic hepatitis C virus infection. *Nature.* 2005 Aug 18;436(7053):946–52.
12. Tservadze T, Sharvadze L, Chkhartishvili N, Dziga L, Karchava M, Gatserelia L, et al. The natural history of recent hepatitis C virus infection among blood donors and injection drug users in the country of Georgia. *Virol J.* 2016 Feb 3;13:22.
13. Montoya V, Olmstead AD, Janjua NZ, Tang P, Grebely J, Cook D, et al. Differentiation of acute from chronic hepatitis C virus infection by nonstructural 5B deep sequencing: a population-level tool for incidence estimation. *Hepatology.* 2015 Jun;61(6):1842–50.
14. Astrakhantseva IV, Campo DS, Araujo A, Teo C-G, Khudyakov Y, Kamili S. Differences in variability of hypervariable region 1 of hepatitis C virus (HCV) between acute and chronic stages of HCV infection. *In Silico Biol.* 2011;11(5-6):163–73.
15. Gismondi MI, Carrasco JMD, Valva P, Becker PD, Guzmán CA, Campos RH, et al.

Dynamic changes in viral population structure and compartmentalization during chronic hepatitis C virus infection in children [Internet]. Vol. 447, *Virology*. 2013. p. 187–96. Available from: <http://dx.doi.org/10.1016/j.virol.2013.09.002>

16. Palmer BA, Dimitrova Z, Skums P, Crosbie O, Kenny-Walsh E, Fanning LJ. Analysis of the evolution and structure of a complex intrahost viral population in chronic hepatitis C virus mapped by ultradeep pyrosequencing. *J Virol*. 2014 Dec;88(23):13709–21.
17. Campo DS, Dimitrova Z, Yamasaki L, Skums P, Lau DT, Vaughan G, et al. Next-generation sequencing reveals large connected networks of intra-host HCV variants. *BMC Genomics*. 2014 Jul 14;15 Suppl 5:S4.
18. Lu L, Tatsunori N, Li C, Waheed S, Gao F, Robertson BH. HCV selection and HVR1 evolution in a chimpanzee chronically infected with HCV-1 over 12 years [Internet]. Vol. 38, *Hepatology Research*. 2008. p. 704–16. Available from: <http://dx.doi.org/10.1111/j.1872-034x.2008.00320.x>
19. Skums P, Dimitrova Z, Campo DS, Vaughan G, Rossi L, Forbi JC, et al. Efficient error correction for next-generation sequencing of viral amplicons. *BMC Bioinformatics*. 2012 Jun 25;13 Suppl 10:S6.
20. Powdrill MH, Tchesnokov EP, Kozak RA, Russell RS, Martin R, Svarovskaia ES, et al. Contribution of a mutational bias in hepatitis C virus replication to the genetic barrier in the development of drug resistance [Internet]. Vol. 108, *Proceedings of the National Academy of Sciences*. 2011. p. 20509–13. Available from: <http://dx.doi.org/10.1073/pnas.1105797108>
21. Kaspar F, Schuster HG. Easily calculable measure for the complexity of spatiotemporal patterns. *Phys Rev A Gen Phys*. 1987 Jul 15;36(2):842–8.
22. Schaper S, Johnston IG, Louis AA. Epistasis can lead to fragmented neutral spaces and contingency in evolution. *Proc Biol Sci*. 2012 May 7;279(1734):1777–83.
23. van Nimwegen E, Crutchfield JP, Huynen M. Neutral evolution of mutational robustness. *Proc Natl Acad Sci U S A*. 1999 Aug 17;96(17):9716–20.
24. Li L, Alderson D, Doyle JC, Willinger W. Towards a Theory of Scale-Free Graphs: Definition, Properties, and Implications [Internet]. Vol. 2, *Internet Mathematics*. 2005. p. 431–523. Available from: <http://dx.doi.org/10.1080/15427951.2005.10129111>
25. Chen W, Lei T-Y, Jin D-C, Lin H, Chou K-C. PseKNC: a flexible web server for generating pseudo K-tuple nucleotide composition. *Anal Biochem*. 2014 Jul 1;456:53–60.
26. Zhang T. Solving large scale linear prediction problems using stochastic gradient descent algorithms [Internet]. Twenty-first international conference on Machine learning - ICML '04. 2004. Available from: <http://dx.doi.org/10.1145/1015330.1015332>
27. Lara J, Tavis JE, Donlin MJ, Lee WM, Yuan H-J, Pearlman BL, et al. Coordinated evolution among hepatitis C virus genomic sites is coupled to host factors and resistance to interferon. *In Silico Biol*. 2011;11(5-6):213–24.
28. Law JLM, Logan M, Wong J, Kundu J, Hockman D, Landi A, et al. Role of the E2

Hypervariable Region (HVR1) in the Immunogenicity of a Recombinant Hepatitis C Virus Vaccine. *J Virol* [Internet]. 2018 Jun 1;92(11). Available from: <http://dx.doi.org/10.1128/JVI.02141-17>

29. Aurora R, Donlin MJ, Cannon NA, Tavis JE. Genome-wide hepatitis C virus amino acid covariance networks can predict response to antiviral therapy in humans. *J Clin Invest*. 2009 Jan;119(1):225–36.
30. Lara J, Khudyakov Y. Epistatic connectivity among HCV genomic sites as a genetic marker of interferon resistance. *Antivir Ther*. 2012 Dec 7;17(7 Pt B):1471–5.
31. Shen C, Gupta P, Xu X, Sanyal A, Rinaldo C, Seaberg E, et al. Transmission and evolution of hepatitis C virus in HCV seroconverters in HIV infected subjects. *Virology*. 2014 Jan 20;449:339–49.
32. Raghwani J, Rose R, Sheridan I, Lemey P, Suchard MA, Santantonio T, et al. Exceptional Heterogeneity in Viral Evolutionary Dynamics Characterises Chronic Hepatitis C Virus Infection. *PLoS Pathog*. 2016 Sep;12(9):e1005894.
33. Longmire AG, Sims S, Rytsareva I, Campo DS, Skums P, Dimitrova Z, et al. GHOST: global hepatitis outbreak and surveillance technology. *BMC Genomics*. 2017 Dec 6;18(Suppl 10):916.
34. Domingo-Calap P, Segredo-Otero E, Durán-Moreno M, Sanjuán R. Social evolution of innate immunity evasion in a virus [Internet]. Vol. 4, *Nature Microbiology*. 2019. p.1006–13. Available from: <http://dx.doi.org/10.1038/s41564-019-0379-8>
35. Lara J, Teka M, and Khudyakov Y. Identification of recent cases of hepatitis C virus infection using physical-chemical properties of hypervariable region 1 and a radial basis function neural network classifier. *BMC Genomics*. 2017; 18(Suppl 10):880.