5-4-2021

Bioinformatics Methods For Studying Intra-Host and Inter-Host Evolution Of Highly Mutable Viruses

Pelin Burcak Icer

BIONFORMATICS METHODS FOR STUDYING INTRA-HOST AND INTER-HOST EVOLUTION OF HIGHLY MUTABLE VIRUSES

by

PELIN B. ICER BAYKAL

Under the Direction of Pavel Skums, PhD

ABSTRACT

Understanding viral disease progression is vital to the detection of outbreaks and subsequent planning for public health actions. Bioinformatics methods are extremely useful for this purpose through a range of applications among which the analysis of viral next-generation sequencing (NGS) data, tracing virus evolution and reconstruction of transmission networks have been explored in this research.

The first part of this research focuses on the processing of NGS data where quantification methods are proposed to describe the robustness and reproducibility of the output of bioinformatics tools. This research shows the importance of assessing the reliability of genomic tools. The second part of this study is the application of processed NGS data to investigate the intra-host evolution of Hepatitis C Virus (HCV) to diagnose and detect new and incident HCV cases. A computational method based on Machine Learning algorithms is proposed to solve this problem. This genomic multi feature-based model not only aims to predict the stage of infection but also aims to understand the evolution of HCV and its underlying complex mechanism. The third part of this research aims to reconstruct transmission networks for new cases which were identified during the aforementioned research. In this part the inter-host evolution of highly mutable viruses is studied. A Maximum Likelihood approach consisting of Uncapacitated Facility Location Algorithm is proposed to solve this problem. Finally, the last part of this dissertation focuses on the inference of the global transmission network of SARS-CoV-2 prior to the pandemic state.

INDEX WORDS: Robustness, Quasispecies, Transmission Network, Machine Learning, Hepatitis C Virus, SARS-CoV-2

BIONFORMATICS METHODS FOR STUDYING INTRA-HOST AND INTER-HOST EVOLUTION OF HIGHLY MUTABLE VIRUSES

by

PELIN B. ICER BAYKAL

A Dissertation Submitted in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy
in the College of Arts and Sciences
Georgia State University
2021

BIONFORMATICS METHODS FOR STUDYING INTRA-HOST AND INTER-HOST EVOLUTION OF HIGHLY MUTABLE VIRUSES

and

ANALYSES OF READ ALIGNMENT TECHNOLOGIES

by

PELIN B. ICER BAYKAL

Committee Chair:	Pavel Skums
Committee:	Alex Zelikovsky
	Zhipeng Cai
	Yury Khudyakov

Electronic Version Approved:

Office of Graduate Studies
College of Arts and Sciences
Georgia State University
May 2021

DEDICATION

To my parents Alaaddin Icer and Aysun Icer, my sister Anatolya Icer, my husband Aydin Baykal and all my loved ones whom I lost during my five years of PhD study.

ACKNOWLEDGEMENTS

I would like to thank my advisor Dr. Pavel Skums. Without his guidance, encouragement, help and patience I would not have been able to achieve my goals. His deep knowledge and research ideas always helped me to get better throughout my PhD study. I also would like to thank Dr. Alexander Zelikovsky and Dr. Serghei Mangul for giving me the opportunity to collaborate with their research projects.

I would like to thank my parents who raised me with infinite love and care and supported me in every situation, my little sister for cheering me up and standing by my side even from long distances, and my dear husband who believed in me from the beginning and encouraged me for every step that I have been taking throughout my PhD.

I would like to express my appreciation to all of my lab-mates who became my friends. I thoroughly enjoyed working with them and being part of this wonderful research group.

Finally, I would like to thank the Molecular Basis of Disease (MBD) fellowship for financial support during my time preparing this dissertation.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

- BAM - Binary Alignment Map
- GHOST - Global Hepatitis Outbreak and Surveillance Technology
- HCV - Hepatitis C Virus
- LS - Local Search
- MCMC - Markov Chain Monte Carlo
- NCBI - National Center for Biotechnology Informatio
- NGS - Next Generation Sequencing
- SA - Simulated Annealing
- SV - Structural Variant
- TPM - Transcripts Per Million
- VCF - Variant Call Format
- WHO - Whorld Health Organization

PART 1

ROBUSTNESS AND REPRODUCIBILITY OF COMPUTATIONAL GENOMIC TOOLS

1.1 Abstract

Reproducibility and robustness of genomic tools are two important factors to assess the reliability of bioinformatics analysis. Such assessment based on these criteria requires repetition of experiments across lab facilities which is usually costly and time consuming. In this study we propose methods that are able to generate computational replicates, allowing the assessment of the reproducibility of genomic tools. We analyzed three different groups of genomic tools: DNA-seq read alignment tools, structural variant (SV) detection tools and RNA-seq gene expression quantification tools. We tested these tools with different technical replicate data. We observed that while some tools were impacted by the technical replicate data some remained robust. We observed the importance of the choice of read alignment tools for SV detection as well. On the other hand, we found out that the RNA-seq quantification tools (*Kallisto* and *Salmon*) that we chose were not affected by the shuffled data but were affected by reverse complement data. Using these findings, our proposed method here may help biomedical communities to advice on the robustness and reproducibility factors of genomic tools and help them to choose the most appropriate tools in terms of their needs. Furthermore, this study will give an insight to genomic tool developers about the importance of a good balance between technical improvements and reliable results.

1.2 Introduction

Bioinformatics is a rapidly evolving area that reflects extremely fast developments of biotechnological platforms. Development of tools has several difficulties that could be tackled due to the constant evolution of bioinformatics algorithms. The success of these algorithms is often determined by how well it captures the biological problem and accounts for technical aspects of the biotechnological protocol. However, sometimes there is a trade-off between overcoming a technical difficulty, such as lowering the running time, and producing less reliable results.

Perturbation of experimental results may cause unexpected effects on bioinformatics results [1]. The higher are these effects the more deviated the results would be, which will put the reliability of the results in question. Therefore, the reproducibility of results obtained using bioinformatics analysis is important in terms of reliability, especially if such results are used in clinical settings [2].

In order to assess the reproducibility of the sequencing protocol, one would need to repeat the same experiment multiple times across multiple laboratory sites and directly compare the bioinformatics results. However, such an approach carries an additional cost and is often infeasible, sometimes such an approach is even impossible (if, for example, clinical rather than laboratory data is analyzed).

Our approach is capable of generating technical replicates computationally based on a single sample and assessing reproducibility of genomic tools [3]. Here we report the development of this computational approach which is capable of altering the priorities of the dataset and check how consistent the results of bioinformatics tools are across different technical replicates of input data. The goal of our approach is not to replace reproducibility measures based on multiple technical replicates generated by multiple laboratory sites, instead, provide a more scalable and affordable approach to assess the reproducibility of genomics tools [3]. Here we report a computational approach able to generate technical replicates for the purpose of evaluating tool robustness and reproducibility.

Our study is the first one to systematically compare the reproducibility of genomics tools across diverse biology domains. Our analysis will inform the broad biomedical community about robustness and reproducibility of genomic tools. The proposed metrics of reproducibility can be used in addition to current used metrics which could potentially help to build reliable and robust genomics tools that maintain reproducibility across replicates and diverse laboratory sites.

1.3 Materials and Methods

1.3.1 Data collection and preprocessing

Reference genome

We used the *GRCh38* reference genome from NCBI (National Center for Biotechnology Information).

DNA-seq and RNA-seq data

We randomly selected ten samples for DNA-seq data (Table 1.1) and RNA-seq data (Table 1.2) data from 1000 Genomes Project. The samples are paired-end short reads contained in FASTQ files.

Table (1.1) DNA-seq samples randomly selected from 1000 Genomes Project

Sample ID	Number of reads	Read length
ERR009308	$14,587,316$	108bp
ERR009309	$12,118,654$	108bp
ERR009332	$13,024,918$	108bp
ERR009345	$17,763,608$	108bp
ERR013101	$10,640,477$	108bp
ERR013103	$18,297,585$	108bp
ERR013104	$14,645,438$	108bp
ERR013105	$12,246,126$	108bp
ERR013106	$12,235,123$	108bp
ERR015522	$11,577,876$	108bp

Table (1.2) RNA-seq samples randomly selected from 1000 Genomes Project

Sample ID	Number of reads	Read length
ERR188040	$27,256,165$	75bp
ERR188071	$20,054,899$	75bp
ERR188166	$40,548,322$	75bp
ERR188167	$27,795,111$	75bp
ERR188247	$26,894,803$	75bp
ERR188293	$27,915,228$	75bp
ERR188315	$13,960,716$	75bp
ERR188328	$17,669,529$	75bp
ERR188428	$39,687,917$	75bp
ERR188465	$25,841,778$	75bp

1.3.2 Generation of technical replicates

We used two computational techniques to modify the properties of the genomic data. Each FASTQ file was modified (and subsequently saved as a different FASTQ file) by random shuffling of the reads. Additionally, we replaced each paired-end reads by its reverse complement. Each sample produces two types of files: randomly shuffled reads ($n = 3$ times) and reverse complemented reads.

1.3.3 Selection of genomic tools

We selected nine DNA-seq alignment tools available on Bioconda environment, three structural variant (SV) detection tools, and two recent tools for quantifying expression of transcripts (Table 1.3).

Table (1.3) Genomic tools

Tool type	**Tool name**	**Version**
Read alignment	Bowtie2 [4]	2.4.1
	BWA-MEM [5]	0.7.17
	BWA-MEM2 [6]	2.1
	HISAT2 [7]	2.2.1
	Minimap2 [8]	2.17
	NextGenMap [9]	0.5.5
	SMALT [10]	0.7.6
	SNAP [11]	1
	Subread [12]	2.0.1
Structural variant	BreakDancer [13]	1.3.6
	Delly [14]	0.7.3
	PopDel [15]	1.5.0
Transcript quantification	Kallisto [16]	1.4.0
	Salmon [17]	0.46.1

1.3.4 Definition and extraction of unambiguous reads

In order to be able to compare replicates of reads fairly we only considered mapped reads marked as 'primary alignment' in the BAM (Binary Alignment Map) file. In this research we will refer to these type of mapped reads as *unambiguous reads*. The definition of secondary alignment read could be different across tools, it either can refer to a multi-mapped read or another alignment with a worse score than the best alignment. Here, we assumed that all such are reads are multi-mapped and we discarded them all. Therefore, for each read (unambiguous read) there is only one alignment, which is the best alignment reported by the corresponding read alignment tool.

1.3.5 Consistency estimation of read alignment across technical replicates and original samples

To investigate the consistency of global alignment and local alignment across original and technical replicates, we investigated different metrics (position, edit distance) and categorized them as follows (color codes of metrics are represented in Table 1.4:

- **Identical:** common unambiguous reads regardless of other metrics.
- **Consistent global and inconsistent local (CG - IL):** common unambiguous reads mapped to the same position with different edit distance.
- **Inconsistent global (IG):** common unambiguous reads mapped to the different position with different edit distance.
- **Multi-mapped:** common unambiguous reads mapped to different position with same edit distance.
- **Inconsistent type 1 (IT1):** unambiguous reads mapped only with original data.
- **Inconsistent type 2 (IT2):** unambiguous reads mapped only with replicated data.

1.3.6 Consistency of detecting structural variants across technical replicates and original samples

We investigated the ability of the tools to maintain the coordinates of reported SVs across original and replicated read alignment files.

Table (1.4) This table represents comparison output of different properties (reads (R), position (P), edit distance (E). Commonality (green), difference (red), only present in original output (blue), only present in replicate output (orange). Grey fields are not applicable.

Categories	**R**	**P**	**E**
Identical	🟩	🟩	🟩
Consistent global - inconsistent local	🟩	🟩	🟥
Inconsistent global	🟩	🟥	🟥
Multi-mapped	🟩	🟥	🟩
Inconsistent 1	🟦	▨	▨
Inconsistent 2	🟧	▨	▨

For this we chose three SV tools (Table 1.3). We compared the coordinates of the **POS** fields and **END** fields within a VCF (Variant Call Format) file. The **POS** field refers to the reference position which is the first base having position 1 and the **END** refers to the end reference position which indicates the variants interval positions (**POS**-**END**) on reference.

In our analysis we set four different thresholds ($T = 0, 10, 100, 1000$) to compare the percentage of matching coordinates of deletion variants between replicate and original data. If the absolute difference between **POS** and **END** coordinates were smaller or equal to the threshold we considered an exact match of coordinates of deletion between replicated and original data.

1.3.7 Consistency of quantifying gene and transcript expression levels across altered and original samples

We checked the consistency of transcripts counts across replicated and original data. We compared the TPM (Trancripts Per Million) variation across output generated from original data and replicate data. In order to assess the performance of *Kallisto* and *Salmon* without input protocol violation, we used single end data (first end of paired data only) for this process.

1.4 Results

1.4.1 Read alignment tools

In terms of the mapping percentage of *unambiguous reads* we observed that only reverse complement data makes a difference. As shown in Fig. 1.1 the most affected tool in terms of mapping percentage is *Bowtie2* with approximately a 14% of mapping difference between output produced by original data and output produced by reverse complement data. On the other hand, *BWA-MEM*, *BWA-MEM2*, *NextGenmap*, and *Subread* are the tools producing almost the same mapping percentage with both types of replicate data.

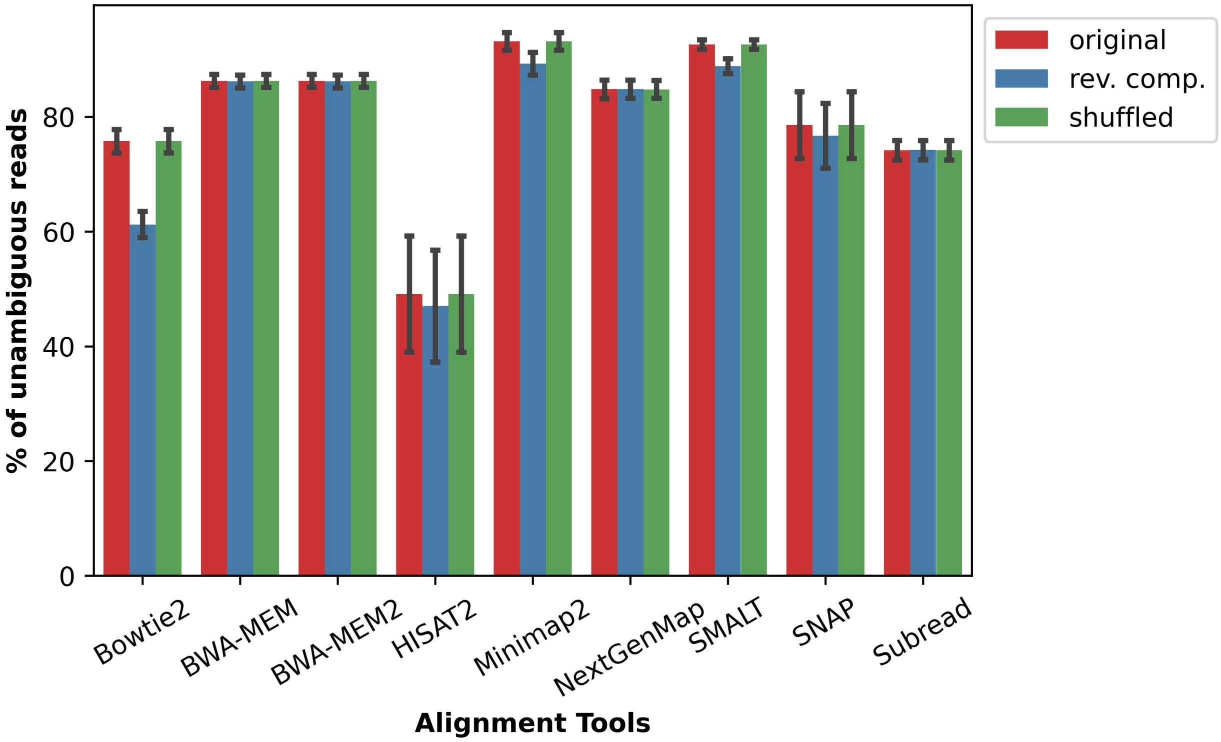


Figure (1.1) Percentage of unambiguous reads.

The breakdown of reads to different categories as in Fig. 1.2 show different behaviors of tools. We noticed that *Bowtie2* produces the least robust results with reverse complemented data, only 78.71% (Table A.1) of reads are identical to the reads produced by original data. The

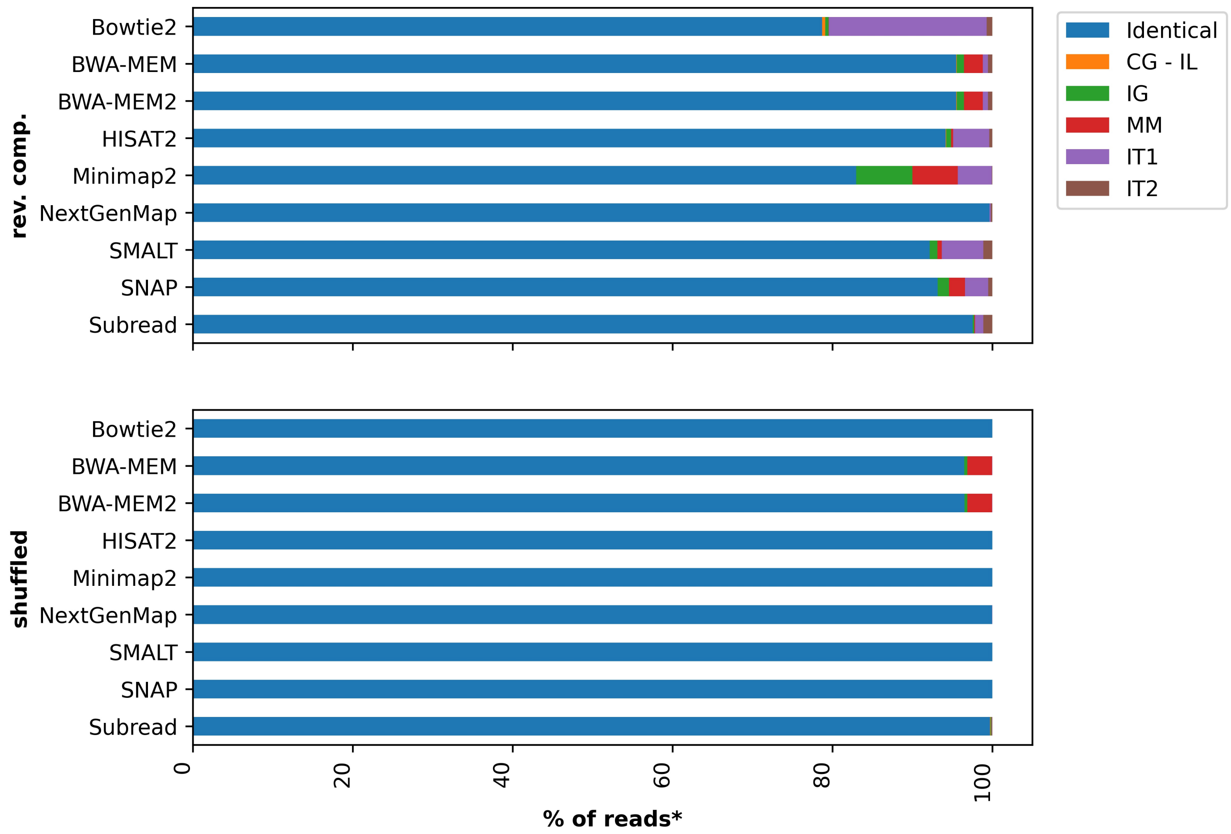


Figure (1.2) Breakdown of reads according to Table 1.4.

difference comes from the inconsistent type 1 reads (unambiguous reads available for original data but missing for reverse complemented data). In particular, although *Bowtie2* has the less identical read percentage between original and reverse complemented output, with other tools we observed differences between output generated with original data and replicate data (Table A.1). For instance, in terms of multi-mapping *Minimap2* has the highest difference, $\sim 7\%$ of unambiguous reads become multi-mapped with reverse complemented data. Again, in terms of inconsistent global alignment *Minimap2* has the highest difference which is $\sim 5.7\%$. *BWA-MEM* and *BWA-MEM2* results are almost identical which suggests there is no significant difference in terms of robustness between these two different versions. Overall, *NextGenmap* performs the best with reverse complemented data with a $\sim 99.7\%$ of identical reads. Concerning shuffled data we can infer from Table A.1 that results do not change that much. The least good performing tools with shuffled data here are *BWA-MEM* and *BWA-MEM2* with $\sim 3.1\%$ of reads aligned globally in an inconsistent way. Tools that produced exactly the same output with shuffled data are in majority (*Bowtie2*, *HISAT2*, *Minimap2*, *SMALT* and *Subread*). However, *NextGenMap* performed the best overall with $\sim 99.7\%$ and 99.97% of similar reads with each technical replicate category.

We checked the percentage of unambiguous reads mapped to different position regardless of other metrics mentioned in Table 1.4. According to Fig. 1.3 *Minimap2* tool has the highest $(\sim 12\%)$ percentage of reads mapped to different position with reverse complemented data. On the other hand, with average of $\sim \%3$ *BWA-MEM*, *BWA-MEM2* and *SNAP* has the highest difference with shuffled data.

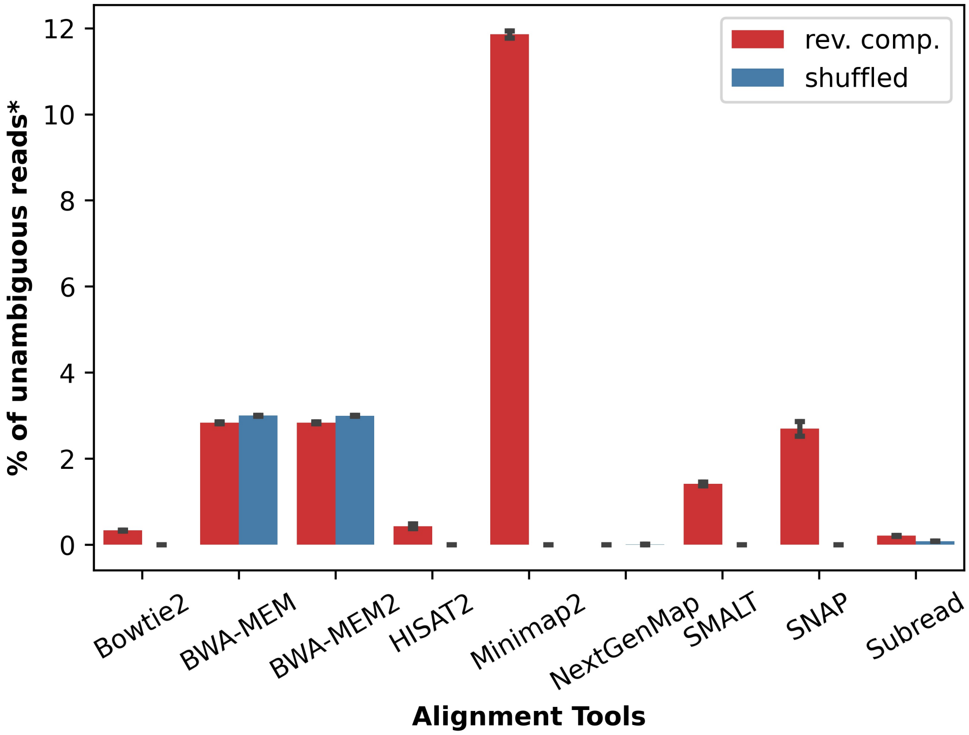


Figure (1.3) Percentage of unambiguous reads mapped to different position regardless of the chromosome.

1.4.2 Structural variant (SV) tools

Our results show that the deletion coordinates change based on the type of the aligner tool and the SV tool.

For this analysis we chose three samples ($ERR009308$, $ERR009309$ and $ERR009332$) and five read alignment tools (*BWA-MEM*, *Bowtie2*, *Minimap2*, *NextGenMap* and *SMALT*) to generate input BAM files for the SV tools (*BreakDancer*, *DELLY*, *PopDel*).

According to Fig. 1.4, with threshold $T = 0$, *PopDel* results obtained by *NextGenMap* BAM file input produces the lowest percentage (89.3%) of matching coordinates between original and shuffled data. In this scenario, the distribution of this percentage across the three different samples is the highest ranging from 91.55% to 99%. Although, with *BreakDancer* and *DELLY* we observe higher matching percentage there is still a variation across different samples. Results obtained from *BWA-MEM* BAM input file also follows the same behavior as results obtained from *NextGenMap*. With input BAM files from *Bowtie2* and *Minimap2*, *BreakDancer* and *PopDel* report exact matching of coordinates and *DELLY* reports a matching of more than 98% between original and all three version of shuffled data. Finally, with *SMALT Breakdancer* again reports an exact matching of coordinates and we can observe a very minor difference of matching reported by *DELLY* and *PopDel*.

We incremented the threshold to see if results would change and we observed that tools overall reported higher percentage of matching coordinates. With *NextGenMap* we observed that the difference across samples decreased significantly and with *SMALT* all SV tools produced an exact matching of coordinates with a threshold of $T = 1000$.

We observed the difference of deletion coordinates varies across the three version of BAM files produced with FASTQ files with randomly shuffled reads. In Fig. 1.5 with *NextGenMap PopDel* reports the highest variation of difference of deletion coordinates between the three different shuffled data and original data. We observed these differences for all three SV tools with *BWA-MEM* and *NextGenMap*. Increasing the threshold reduces the difference, but we can still observe the variation across the three shuffled data. We suggest that this difference is a result of eliminating some alignments once enough coverage is obtained by the respective SV

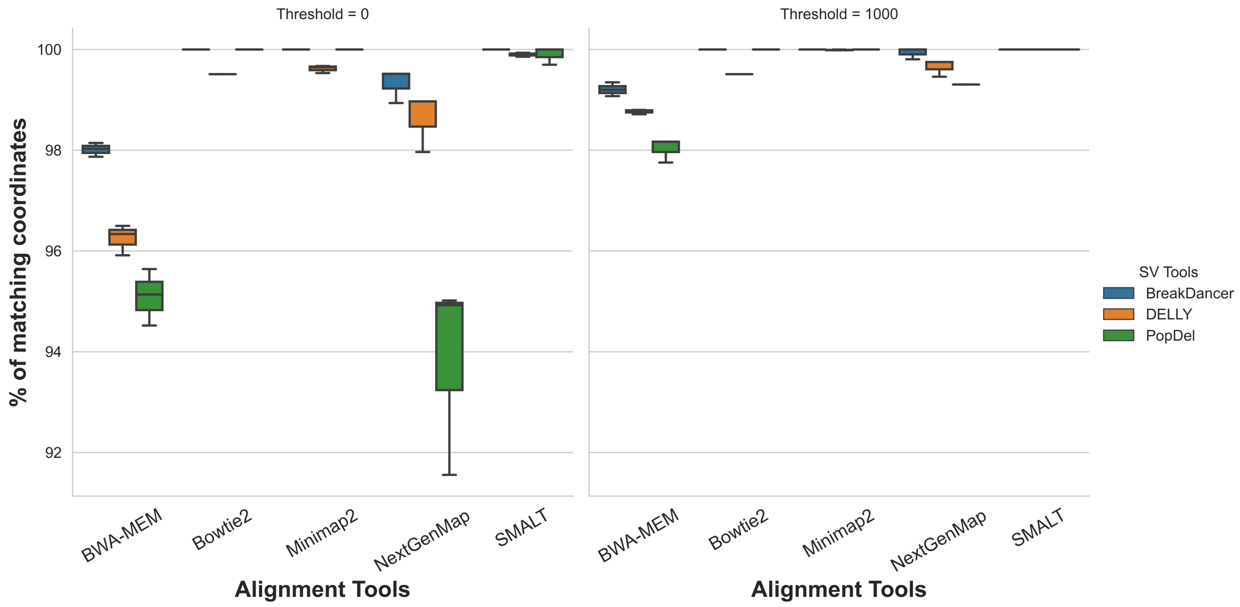


Figure (1.4) Percentage of matching coordinates of shuffled data and original data with two different thresholds. Boxplots represent the difference among the three different samples of shuffled data.

tool. Therefore, the random order of reads will affect the aligned reads present on the coverage produced by the SV tool to detect the structural variant.

1.4.3 Gene expression quantification tools

We ran *Kallisto* and *Salmon* with ten RNA-seq samples. Fig. 1.6 shows that the tools have similar robustness when using shuffled data while *Salmon* has a tendency to show lower deviation with reverse complemented data.

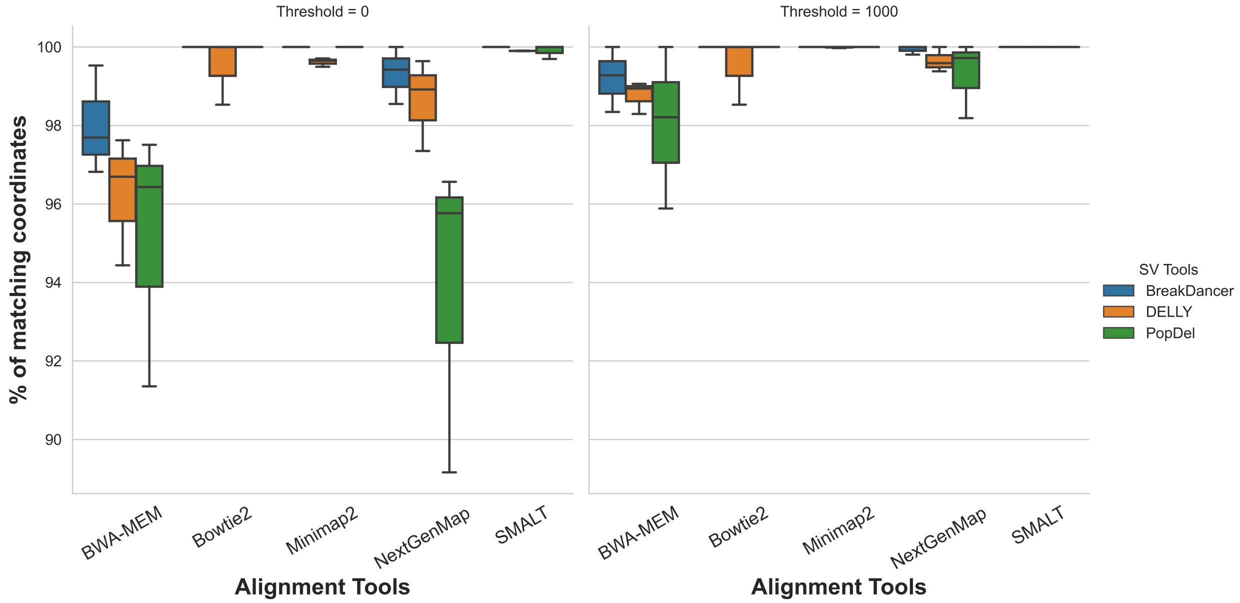


Figure (1.5) Percentage of matching coordinates of shuffled data and original data with two different thresholds. Boxplots represent the difference among the three different version of shuffled data.

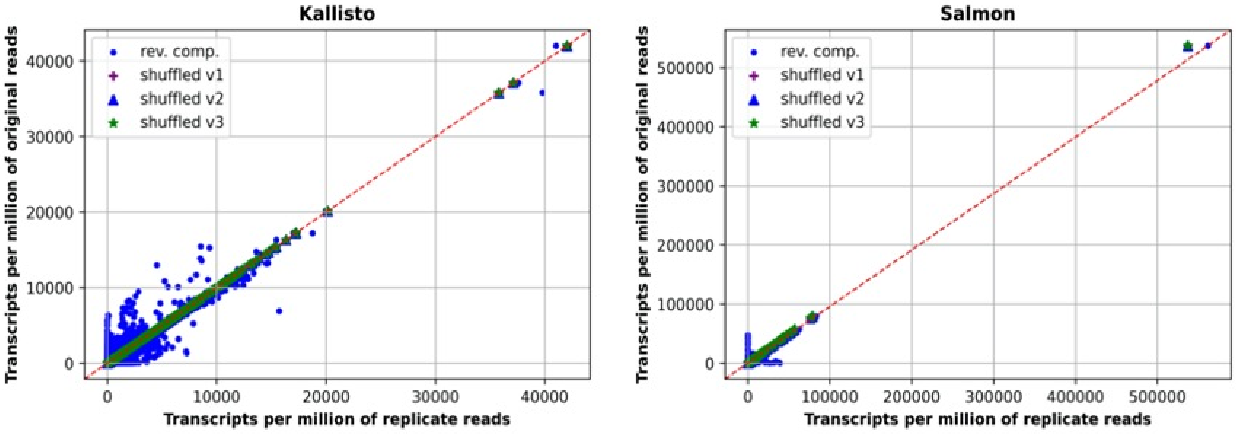


Figure (1.6) TPM of original reads against TPM of replicate reads per transcripts for Kallisto and Salmon tools.

1.5 Discussion

We present a study of reproducibility and robustness of genomic tools which comprises the analysis of a broad range of genomic tools. We introduced the notion of technical replicate data in order to allow the testing of reproducibility of these genomic tools.

Our results indicate different behaviors in terms of *robustness* and *reproducibility*. We observed that some DNA-seq tools such as *BWA-MEM* and *BWA-MEM2* are more sensitive to both types of technical replicate data. In order to validate the effect of replicate data we ran DNA-seq alignment tools a couple of times with exact data and we observed that output remained the same across these different runs. Concerning SV tools the changes were basically related at the same time to the choice of read alignment tools and the order of reads within the BAM file provided by the respective alignment tools. This indicates the impact of results of the read alignment tool on the result produced by the SV tool. On the other hand, some trade-off between output consistency and technical issues seems to also be creating an impact on the robustness of SV tools. As mentioned in the results section of SV tools, *PopDel* relies on the order of the reads within a BAM file to reduce running time which causes variation between different versions of BAM files (output obtained with randomly shuffled reads). Regarding RNA-seq quantification tools, we observed no difference between output obtained from original data and output obtained from randomly shuffled data for both *Kallisto* and *Salmon*. However, higher deviation was observed with *Kallisto* when reverse complement input was used.

Our current analysis clearly suggests the importance of testing reproducibility and robustness of genomic tools. Since genomic tools have protocols for paired-end data we were not able to use the reverse complement data for SV tools and RNA-seq quantification tools in order to not to violate any rules for input data requirement. However, for RNA-seq quantification tools we dropped the second-end of the paired-end data and created replicates based on only first-end data and ran the tools with single-end data. With this approach we were able to compare the tools without violating any input protocols specific to the tools.

Based on our analysis, we believe our study will inform the broad biomedical community whether or not their genomics tools of choice are able to maintain consistent results across replicate data. The ultimate results presented in our study will help to build reliable and robust genomics tools that maintain reproducibility across replicates and diverse laboratory sites. We hope that the proposed reproducibility metrics will be adopted by current bioinformatics studies in addition to standard methods of accuracy, running times and memory usage.

PART 2

QUANTITATIVE DIFFERENCE BETWEEN INTRA-HOST HCV POPULATION FROM PERSONS WITH RECENTLY ESTABLISHED AND PERSISTENT INFECTION

2.1 Abstract

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. 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 features measuring their diversity, topological structure, complexity, strength of selection, epistasis, evolutionary dynamics, and physico-chemical properties. Distributions of the viral population features 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 per cent, thus providing a higher accuracy than other genomic-based models. 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, which could potentially

complement and/or substitute standard laboratory assays.

2.2 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 2015 [18].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 2015 [18]. Hepatitis C outbreaks continue to occur, posing a serious challenge to public health [19]. HCV is highly mutable. As a result, each infected individual hosts a heterogeneous population of genetically related HCV variants or 'quasispecies' [20]. Substantial diversity of intra-host viral populations plays a crucial role in disease progression and epidemic spread [21, 22, 23]. 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 [24], whether it is 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 [25, 26], which may result in advanced diagnostic and treatment strategies.

In this article, we study evolutionary factors associated with the transition between HCV infection stages. In more than 50 per cent of cases untreated HCV infection proceeds to the chronic phase, which can lead to the development of liver cirrhosis and/or hepatocellular carcinoma [26]. 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 assessed 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 [27, 28]. Detection of HCV RNA in the absence of anti-HCV activity in serum specimens was used as an indication of recent HCV infection [29]. 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 analyse 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 [30, 31]. 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 diverting at later stages of infection [21, 32], with certain viral variants persisting in infected hosts for years [21, 32]. Furthermore, this process is accompanied by an increase of negative selection over the course of HCV infection [33, 21, 32, 34]. These findings suggest a high level of intra-host adaptation at late stages of infection [22] 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 physico-chemical properties of intra-host HCV variants and populations. Analysis of fea- tures' 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-feature 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.

2.3 Materials and Methods

2.3.1 Data collection and preprocessing

We analysed intra-host HCV populations sampled from recently (N1/498) and persistently (N1/4256) infected persons collected as described in [35]. The E1/E2 junction of the HCV genome (L = 246nt), which contains the hyper- variable 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 K-mer Error Correction (KEC) [36], which produced 245,878 unique viral haplotypes with frequencies. Sequences of each population were aligned using MUSCLE [37]. Since obtained average numbers of sequences for recent and persistent populations were different ($\sim n = 295$ and $\sim n = 846$, respectively), the features studied in this paper were normalized, when appropriate.

2.3.2 Features calculation

The analysed features could be loosely split into four groups: genomic features, complexity features, network features, and bio-chemical features. The features are summarized in Table B.1. We assumed that a given intrahost population contains n unique haplotypes with frequencies $f_1, ..., f_n$. Sixteen features corresponding to this population constitute its 'feature vector'.

Genomic features

These features are obtained by direct comparison of sequences from each population.

'Distance-based' features include 'mean and SD' of pairwise hamming distance distribution ('Features 1 and 2'), and the 'conservation score (Feature 3)' of the population consensus sequence calculated with the NUC44 scoring matrix [38]. We also used the so-called 'mutation frequency' feature (Feature 4) [31], which is defined as the mean distance between all haplotypes and the most frequent haplotype. All four features measure the population diversity.

Diversity was also quantified using three 'entropy-based' features. Suppose that the intrahost population. $S = \{s^1, ..., s^n\}$ is fixed. For the genomic position i, let $H_{i,k} = \{(s^i_j, ..., s^j_{i+k-1}) :$

$j = 1, ..., n\}$ be a collection of k-mers (subsequencess of length k) of all haplotypes starting at that position.The positional k-entropy $E_{k,i}$ is defined as the entropy of the frequency distribution of k-mers starting at i:

$$E_{k,i} = -\sum_{h \in U(H_{i,k})} f_{i,k}(h) \log_2(f_{i,k}(h)) \tag{2.1}$$

Here $U(H_{i,k})$ is the set of unique elements of $H_{i,k}$, h is a k-mer, and $f_{i,k}(h)$ refers to the relative frequency of h inside $H_{i,k}$. An 'average positional k-mer entropy' E_k (Feature 5) is the mean of positional k-entropies over all positions. For $k = L$, the feature E_L (Feature 6) is an entropy of observed haplotype frequencies, while for $k = 1$, it is an average position-wise single nucleotide variant (SNV) entropy (Feature 7). In our model, we used entropies E_1 (Feature 7), E_L (Feature 6), and E_{10} (Feature 5).

Next, we estimated the frequency of 'transversions (Feature 8)' (mutations between purines and pyrimidines) among all observed mutations within the population. This feature is suggested by previous studies [39] that reported higher frequencies of transitions over transversions in HCV populations.

'Selective pressure' was measured using the DN/DS ratio (Feature 9), 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.

Complexity features

'PCA complexity (Feature 10)' is derived from principal component analysis (PCA). PCA has been widely used to quantify patterns of genomic population structure, and the idea to use the number of principal components that explain a given portion of variation to estimate the effective number of subpopulations inside a population has been described and justified in [40] and [41]. In our case, for each population, we transform its alignment into $n \times 4L$ numerical matrix M by transforming nucleotides as A=0001, C=0010, T=0100, G=1000. The complexity $f_P(v)$, $0 \leq v \leq 1$, is then defined as the percentage of principal components required to explain at least v per cent of the observed genetic variance, that is $f_p(v) = min\Big\{k/4L : \frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{4L} \lambda i} \geq v\Big\}$, where $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_{4L}$ are the eigenvalues of a covariance matrix of M. In this study, we used the complexity $f_p(v)$ with $v = 0.5$. This can be justified by several arguments. First, the value $v = 0.5$ is large enough to reflect the significant portion of the population's genomic composition. Second, each population in general can be characterized by the are under curve $f_p(v)$, which takes into account all possible values of v. However, if M is non-singular, then given that $f_p(v)$ is monotonically increasing, $f_p(v) \to 0$ as $v \to 0$ and $f_p(v) \to 1$ as $v \to 1$, this area is accurately enough reflected by the value $f_p(0.5)$ at the middle-point of the curve's domain.

'Kolmogorov complexity (feature 11)' is the classical concept of information theory, which quantifies the descriptive/information complexity of a string over a finite alphabet [42]. 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 (feature 11) 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 (as described above), and these strings have been concatenated into a string s of length —s—. Following (Kaspar and Schuster 1987), the normalized Kolmogorov complexity KC (feature 11) was estimated as $KC = (c(s))/(b(s))$. Here $c(s)$ is a Lempel-Ziv complexity of s defined as the number of unique substrings encountered by a computational process that constructs s by adding its symbols one by one; and $b(s) = |s|/log_2(|s|)$ is the asymptotical expected Lempel-Ziv complexity of a random string of $|s|$ (for more detailed description see [43, 44]. Under this definition, "simpler strings" (i.e., strings with more regular structure) have lower complexity.

Network features

This group of features is derived from the analysis of genetic networks of HCV populations, that represent a sequence space [44] 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) Fig. 2.1. 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 trees [34]. Their structure is shaped by various factors, such as epistasis, founder effects, and selection pressures that affect the virus over the course of infection [21, 45]. For each network, the following four features have been calculated.

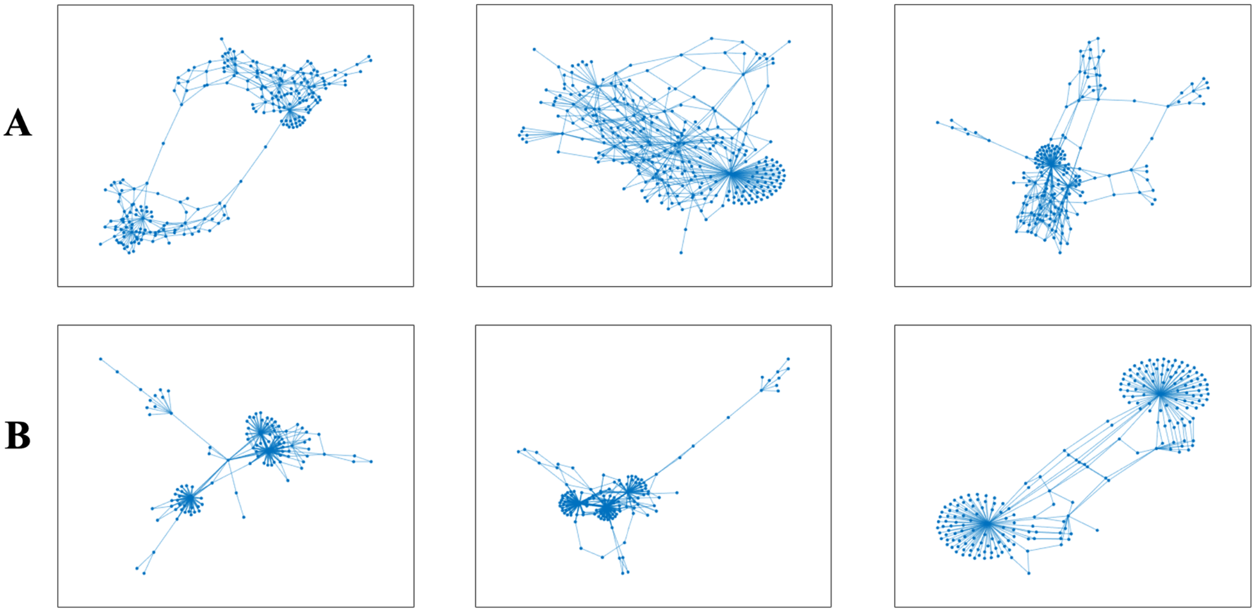


Figure (2.1) Examples of genetic viral networks for a **(A)** persistently infected individual and **(B)** for a recently infected. The viral network of the recently infected host has the structural properties typical for scale-free network.

‘Robustness/selection balance (feature 12)’ has been measured by the correlation between vectors of vertex frequencies and eigenvector centralities. The latter is the principal eigenvector of the adjacency matrix A of G_N. In the classical quasispecies model, vertex centralities are indicative of the mutational robustness of corresponding viral variants, while a high frequency may be indicative of a higher fitness or a higher mutational robustness [46].

Topological structures of genetic networks have been assessed using two features. The first of them is a normalized s-metric (feature 13) [47] $s(G_N) = \sum_{(i,j)\in E(G_N)} d_i d_j$, which measures how close a network is to being scale-free. Here, d_i is a degree (number of neighbors) of a vertex i, and the normalization factor n^4 represents the order of magnitude of the maximum non-normalized s-metric for n-vertex network. 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.

The second network structural feature is the average clustering coefficient C (feature 14), which measures the degree to which network vertices tend to cluster together. It reflects the probability that a random connected vertex triplet is complete (i.e., every pair of vertices is connected by an edge) and is calculated as $C = \frac{2}{n}\sum_{i=1}^{n} \frac{\sum_{j,k \in V} A_{ij}A_{ik}A_{jk}}{d_i(d_i-1)}$.

Evolutionary dynamics (feature 15) feature estimates an age of the genetic network using a dynamic evolutionary model. The general idea is to simulate variant frequencies using a system of ordinary differential equations (ODE) (feature 15) and estimate an age of the network as the time when simulated frequencies achieve the best agreement with observed frequencies. Owing to the inherent uncertainty of quantitative features of the model, we do not use the actual estimated age as a feature for infection staging, rather we classify patients based on the qualitative behavior of the function describing the deviation of simulated and observed frequencies over time.

Formally, we separate each intra-host viral population into subpopulations corresponding to connected components of the genetic network. For each subpopulation, viral evolution is described by the dynamical system (2.2)-(2.5), partially inspired by the ideas from [48, 49].

$$\dot{c} = \alpha + \rho(1 - \frac{c + \sum_{i=1}^{n} c_i}{C^*}) - \theta c - \beta c \sum_{i=1}^{n} v_i \tag{2.2}$$

$$\dot{c_i} = \beta c v_i - \delta h c_i \tag{2.3}$$

$$\dot{v_i} = p \sum_{ij \in E} q_{ji} c_j - \lambda r_i v_i \tag{2.4}$$

$$\dot{r_i} = \phi v_i - \sigma r_i \tag{2.5}$$

In this study we utilized the values of model parameters estimated in [50, 49].

Variable c represents numbers of uninfected target cells, variables c_i, represents cells infected by virions with ith genome, variables v_i represents virions with ith genome in the host's serum, and variables r_i represents B-cell antibodies targeting the ith genome. The constants $\alpha, \rho, \theta, \beta, \delta, p, \lambda, \phi$ and σ are model parameters representing various rates. In this study we utilized the values of these parameters estimated in [50, 49]. According to the model, uninfected hepatocytes are produced by differentiation of precursor cells at a constant rate α, proliferate by a logistic growth law with a rate ρ due to the limited liver cell carrying capacity and die at rate θ. Cells are infected by a variant i at a rate βv_i and, after being infected, are eliminated by the cross-immunoreactive CTLs at rate δ. The virions with the ith genome are introduced to the blood by the cells infected by the variant j at the rate pq_{ji}, where $q_{ji} = (\epsilon/3)^{d_{ij}}(1-\epsilon)^{L-d_{ij}}$is a probability of mutation between variants i and j, d_{ij} is the Hamming distance between genomes i and j, L is the length of the genomes and ϵ is the mutation rate. Specific B-cell antibodies r_i eliminate the corresponding virions at a rate λr_i. They are stimulated by the corresponding viral variants at the rate ϕ and decays at a rate σ in the absence of stimulation.

An estimated age of the network could be defined as the time T^* when simulated frequencies $g_i(t) = \frac{v_i(t)}{\sum_{i=1}^{n} v_i(t)}$ achieve the best agreement with observed frequencies, i.e., $T^* = argmin(JS(g(t), f))$, where $JS(g, f)$ is a Jensen-Shannon divergence between distributions $g(t)$ and f. However, numerical features in equations (2.2)-(2.5) could be uncertain or patient specific. As a result, the estimated numerical values of T^* for different patients cannot be directly compared and used for the classification. More reliable classification can be achieved using qualitative characteristics of the deviation function $\varphi(t) = JS(g(t), f)$, which are not significantly affected by the parameters' variation. Intuitively, if the model (2.2)-(2.5)adequately describes the evolution of viral populations, then the model for recent populations achieves the best agreement between model-based and observed frequencies for earlier populations, and the agreement deteriorates with time. Similarly, for older populations the agreement is low at earlier times and increases till it reaches a maximum at some late time point. Thus, in general, recent and old populations are expected to be characterized by deviation functions $\varphi(t)$ with descending and ascending trends, respectively. Based on this idea, we classify viral subpopulation as recent or old based on the coefficients of the approximation of $\varphi(t)$ by the exponential function $a + be^{ct}$.

It is known that some patients have mixtures of components under different types of selection [34]. In particular, chronically infected hosts may have components of different ages, and the overall age of infection should be defined by the oldest component. Given this, the classification of a patient was performed separately for each connected component of the genetic network. The patient is classified as persistently infected, if at least one of the connected components is old, and as recently infected, if all of them are recent. The prediction age variable c_{ODE} (feature 15) is set to be equal to -1 in the first case and 1 in the second case.

Biochemical feature

This feature (feature 16) assesses the physico-chemical properties of sequences from viral populations. In brief, a large set of physico-chemical parameters is considered for each sequence from a given population. The goal is to synthesize the information from all sequences into a single population feature that is added to the other features for the final analysis. This is done first by applying a prediction model to assess whether a given sequence has a physico-chemical profile associated with recent or persistent infection, using the method described below. The biochemical index of an entire population is then defined as the probability that a random sequence from this population has a profile pointing to persistent infection. This part of the analysis was carried out by our collaborator Dr. James Lara from the Centers for Disease Control and Prevention, and it is described in detail in [51].

2.3.3 Machine learning classifier

Feature vectors of recently and chronically infected hosts were used to train machine learning classifiers for infection stage prediction. Given a labelled 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 features and assigns labels for unlabelled feature vectors using the trained model. In this study, we used Support Vector Machines (SVM) 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.

2.4 Results

2.4.1 Stage-specific distributions of features

Our model consists of 16 diverse features categorized in four groups: genomic features, complexity features, network features and biochemical features. The 3 features (k-entropy, SNV entropy, and conservation score; features 3, 5, and 7 respectively) were found to be in high correlation with feature 1 and with each other Fig. 2.2. For the remaining 13 features there is a small-to-medium correlation between them Fig. 2.2, demonstrating that they reflect different properties of intra-host viral populations.

Feature vectors of recent and persistent populations are separable from each other Fig. 2.2. For the features, Mann-Whitney U-test suggests statistically significant difference between recent and persistent intra-host populations ($p < 0.001$, Table B.1).

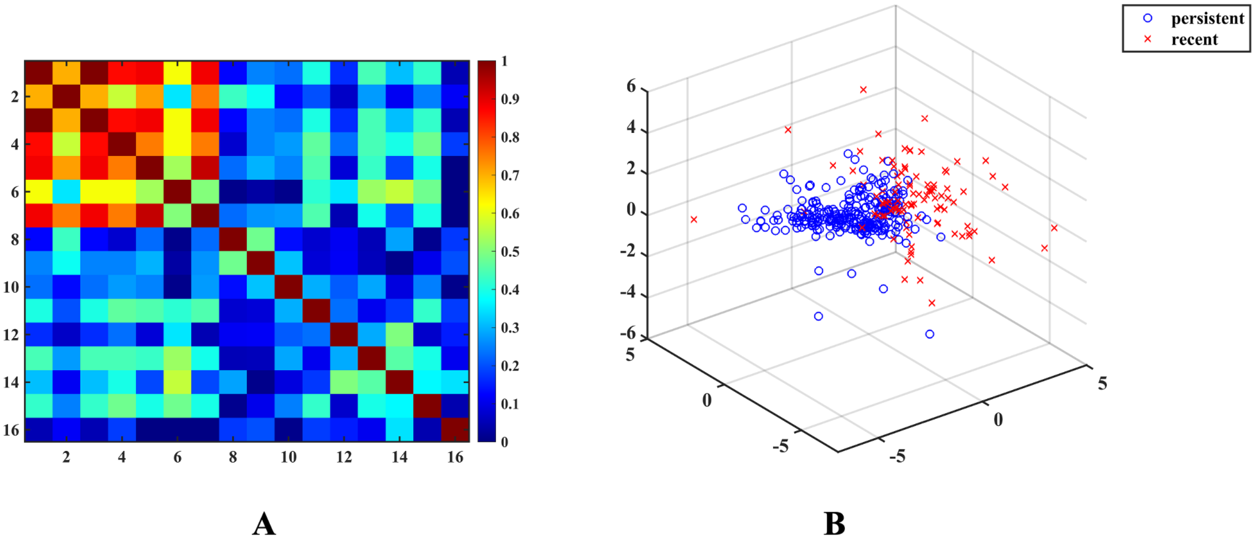


Figure (2.2) **(A)** Heatmap of absolute values of pairwise correlations between features. **(B)** 3-Dimensional projection of recently and persistently infected hosts (with highly correlated features removed).

As expected, diversities are on average higher for persistent than recent populations (Fig.

2.3 (1-7)). Higher genetic diversity of persistent populations is accompanied by significantly lower PCA ($p < 0.01$) and Kolmogorov complexities (feature 10 and 11) (Fig. 2.3 (10, 11)). The decrease of complexity points to a higher level of adaptation and organization of intra-host populations (see Discussion). The emergence of intra-host adaptation is further supported by the increase in negative selection (feature 9) (Fig. 2.3 (9)) at later stages of HCV infection. It can be claimed that this trend is not due to spurious correlations associated with the difference in numbers of sequences for recent and persistent populations. When the patients with less than q sampled sequences are discarded, then the numbers of sequences for recent and chronic patients are more balanced (1139 and 1340 in average, respectively, for $q = 200$ and 1350 and 1594, respectively, for $q = 300$), and the trend to have lower PCA and Kolmogorov complexities (feature 10 and 11) for persistent populations remain to be statistically significant ($P < 0.01$). Furthermore, the significant difference between PCA complexities (feature 10) of recent and persistent populations is observed for different values of the variance threshold. Indeed, the areas under the curves $f_p(v)$, $(0 \leq v \leq 0.9)$, for persistent populations P remain significantly lower than for recent populations ($P < 0.02$).

Recent and persistent HCV populations are also separable by an ODE feature c_{ODE} (feature 15) (Fig. 2.3 (15)). However, it should be noted that the minimal Jensen-Shannon divergence between model-based and observed frequencies is significantly lower for recent than for chronic patients (0.07 vs 0.12, respectively, $P < 0.001$). Thus, the immune escape-based model (2.2)-(2.5) describes recent populations more accurately than persistent populations. This observation could point to a declining role of immune escape.

Transition mutations were overwhelmingly more frequent than transversion mutations (feature 8) for both classes of samples. This fact agrees with the previously published results [39], 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 NS5B [39]. However, prevalence of transversions was ~ 2 times higher in persistent populations (Fig. 2.3 (8)). This property of intra-host viral evolution also agrees with previous studies [52] on the between-host level that suggests that the propensity of transition/transversion ratio to decline could be associated with the growth of genetic saturation.

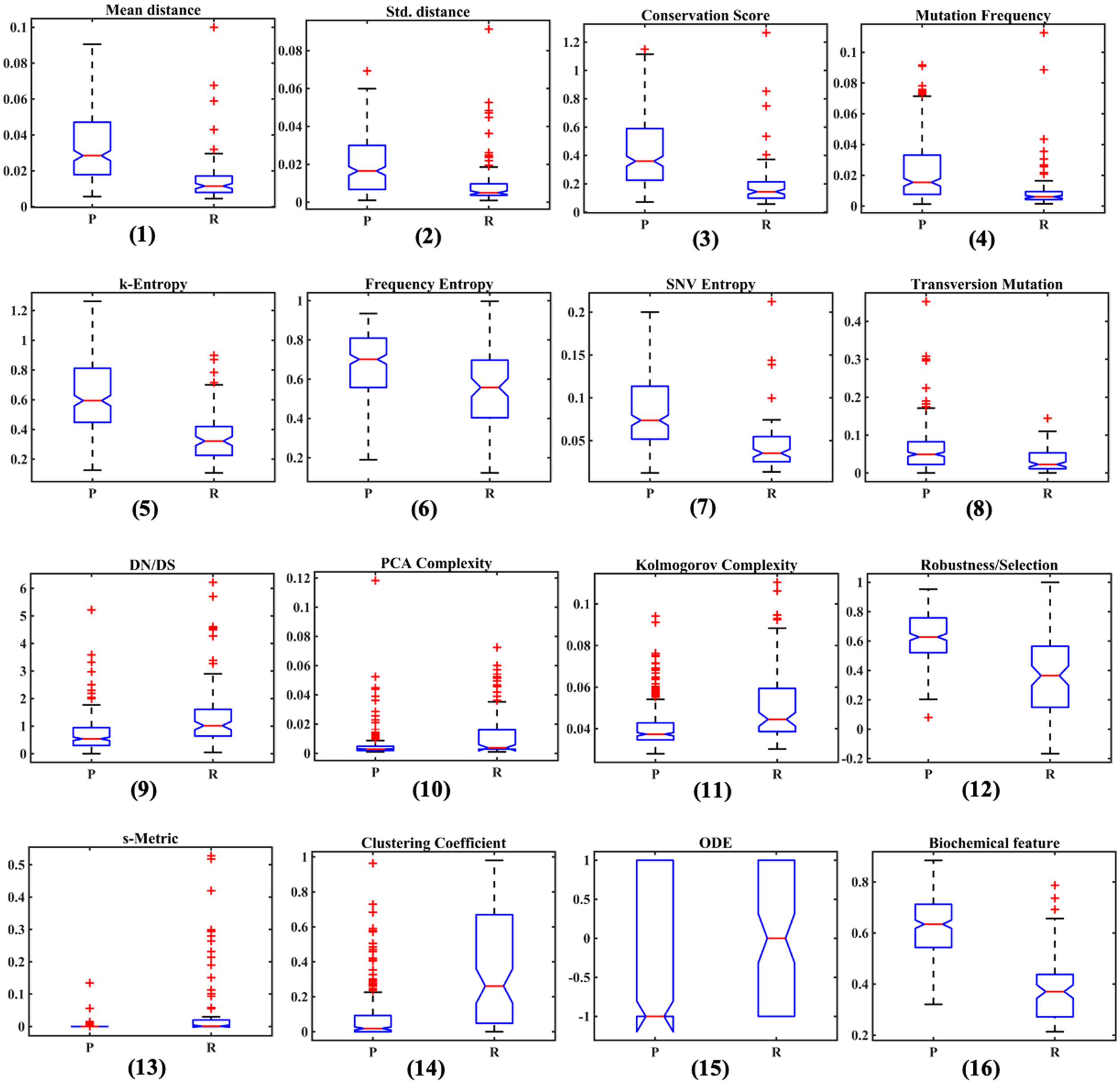


Figure (2.3) Boxplots of feature distributions for persistent (left box plot on each graph) and recent (right box plot on each graph) intra-host HCV populations. The plots are in the same order as in Table B.1.

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 (feature 13 and 14) (2.3 (13, 14)). It indicates that, in contrast to the persistent

Table (2.1) Prediction accuracies of machine learning methods

Method	**Prediction Accuracy**	**95 % CI**
SVM - linear kernel	95.17%	$(94.009, 95.233)$
SVM - quadratic kernel	95.22%	$(95.004, 95.356)$
Logistic regression	93.17%	$(92.908, 93.051)$

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. 2.1). This observation reflects the role of founder viral variants at the earlier stage of infection (see Discussion). A significantly higher correlation between frequencies and network centralities of variants in persistent populations (feature 12) (2.3 (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.

Finally, individual sequences of recent and persistent populations have distinct physico-chemical properties (feature 16) (2.3 (16)).

2.4.2 Machine learning classification

Mutation frequency, k-entropy and frequency entropy (feature 4, 5 and 6) have been excluded from the prediction model as they are highly correlated with feature 1 and with each other. The remaining 13 features 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”. Although these classifiers do not necessarily require removal of highly correlated features, they were dropped to reduce a likelihood of overfitting. 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 1. Classification performance evaluation of all methods indicates a high accuracy of infection stage inference, with SVM with quadratic kernel demonstrating the highest accuracy of 95.18%.

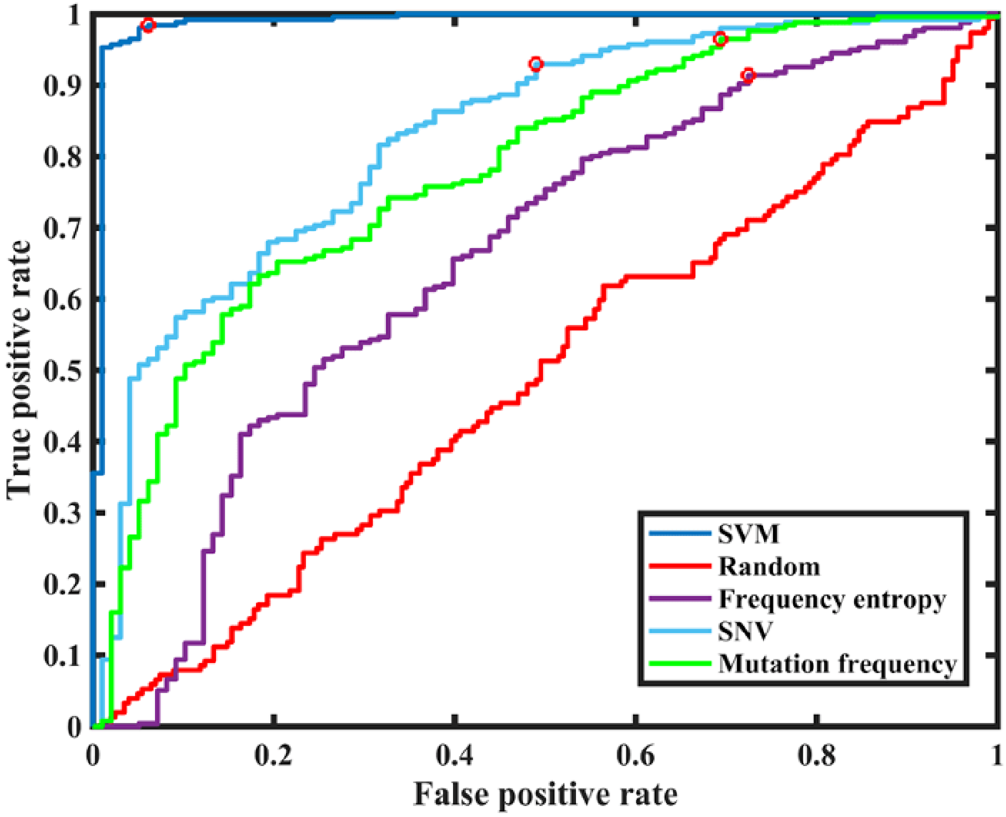


Figure (2.4) ROC curves of classification models

SVM classifier with quadratic kernel has been compared to the previously published HCV infection staging models [31] which classify intra-host viral populations as recent or persistent using frequency entropy (feature 6), SNV entropy (feature 7) or mutation frequency (feature 4). The ROC curves of the classifiers are shown in Fig. 2.4. 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 features 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 = 0.5), thus indicating that the associations between feature distributions and infection stages are likely due to the structural and evolutionary factors rather than to random statistical correlations in the data.

2.5 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 studies [31]. 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 polyprotein [53], and is associated with immune escape [54], drug resistance [55, 53] and virulence [56]. 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.

Our results indicate significant differences in the structure of HCV populations sampled from recently and persistently infected hosts and suggest that intra-host HCV populations develop in a complex but ordered and predictable manner during the course of infection. This form of evolutionary regularity manifests itself in the presence of viral genetic features strongly associated with stages of infection. Utilization of these features for machine learning allowed us to train 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 [28, 57, 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 [22, 58], simple metrics of genetic heterogeneity are insufficient for the accurate staging of HCV infections. High accuracy could be achieved by using a combination of features measuring different structural and evolutionary properties of viral populations. Furthermore, most of the analyzed features are easily computable and do not require computationally intensive phylogenetic and phylodynamic inference. Thus, the proposed prediction models may serve as accurate and scalable 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) [59] — a web-based molecular surveillance system developed and maintained by CDC. Moreover, the overall strategy described in this study may serve as a foundation for cyber-molecular diagnostics [60].

All feature changes described in this study reflect major trends of intra-host viral evolution that have been previously explored and described [21, 52, 22]. Some changes such as the increase of transversions and genetic heterogeneity of viral populations seem to be more directly than others associated with duration of infection and linked to high mutability and genetic saturation [52, 31]. Changes of other features are likely associated with variation in selection pressures operating at different stages of infection.

The dynamics of analyzed features suggests that intra-host HCV evolution at the initial stage of infection is largely different from evolution at later stages of infection. In particular, the physico-chemical properties of HVR1 variants appear to be influenced by and responsive to intra-host environmental factors specific to the recent and persistent stages of HCV infection. The change of properties is likely to be associated with different evolutionary mechanisms operating at different infection stages. Early evolution is likely defined by a founder-flush process [61, 21], which rapidly generates massive selectable genetic heterogeneity. This is reflected, in particular, by pronounced scale-free properties of recent genetic networks. Indeed, star-like networks (the simplest scale-free networks), just like star-like phylogenies, typically represent populations that recently underwent a population expansion from a single founder. For HCV, transmissions of higher multiplicity have been observed [62]; furthermore, the population could expand by the time of sampling. In that case, the observed genetic network may not be exactly star-like, but multiple founder variants (median $=4$ [63]) should still serve as most central vertices, resulting in more general scale-free structure. Evolution of recent populations seems to be driven by positive selection ($DN/DS > 1$). In such settings, the contribution of mutational robustness is less pronounced, resulting in the lower values of the robustness/selection balance feature (feature 12) for recent populations. In particular, the network centers are likely founders transmitted from their previous hosts, some of which may be less fit than newly generated variants in the new host, resulting in the eventual decrease of their observed frequencies over time.

In contrast, later stages are likely to be defined by the virus adaptation to the host environment and varying immune selection pressures. The process of adaptation results in an orderly development of intra-host viral populations that is reflected by the increase of negative selection and decrease of PCA and Kolmogorov complexities (feature 10 and 11). The major role here is played by epistasis. For HCV, it is frequently detected in the form of coordinated substitutions, which are organized into a complex network of epistatic connectivity [64]. Coordinated substitutions reflect selection pressures acting on intra-host viral populations and represent dependence of phenotypic effects of mutations on other mutations or on genetic background to which these mutations occur [64]. HCV epistatic interactions have been shown to be associated with host factors [53, 65], drug resistance [55, 56, 66], disease severity [65], and coinfections [50]. Many features analyzed here could be essentially linked with the underlying epistatic networks. For example, high complexity indicates a high level of randomness of a sequence, while low complexity implies the presence of specific structural patterns inside a sequence [42]. The reduction in complexity of the late-stage intra-host HCV populations indicates increase in epistatic connectivity among polymorphic sites resulting from strong functional constraints experienced by these populations. Such constraints shape adaptation of viral populations to specific intra-host environments. At the earlier stages of infection, nucleotide changes are seemingly more random, resulting in populations with higher dimensionality. Changes in physico-chemical properties based on auto-correlation also likely reflect variations in the structure of epistatic networks established during early and late stages of infection in addition to variation in the strength of connectivity in these networks [35].

Given the above observations, it is unlikely that the entire intra-host HCV evolution is driven by a single evolutionary mechanism. The changes of evolutionary parameters are consistent with the hypothesis that intra-host HCV evolution is not a simple accrual of genetic heterogeneity resulting from the "arms race" between the virus and the host's immune system or a random genetic drift within the space of effectively neutral genomic variants. It is rather a complex process defined by the recurring presentation of a succession of selection challenges specific to each stage of infection [21, 22]. In this process, different modes of evolution can be dominant at different stages. The "arms race" seems to drive intra-host HCV evolution at its early stages, as indicated by the positive selection, smaller impact of mutational robustness and the fact that ODE model (feature 15) (2)-(5) describes recent populations quite accurately. However, the "perpetual arms race" model is inconsistent with the observed increase of negative selection, long-term persistence of particular genomic variants [21, 67] and antigenic convergence [68]. Similarly, strong coordination of variability and functionality among genomic sites is unlikely to be established as a result of a neutral evolution and should be the result of selection. However, it is quite possible that the stable "end-game" or equilibrium population observed at its later stages will be internally neutral (under the network neutrality definition from [46]). Such a population can be located at a local fitness landscape plateau and form a neutral genetic network, where fitnesses of its nodes are equal and greater than for variants outside the network. This can lead to the higher impact of mutational robustness on the relative variant frequencies inside this network (van Nimwegen, Crutchfield and Huynen 1999). However, it should be remembered that the viral fitness landscape is essentially dynamic and is defined by emerging immune responses and cross-immunoreactivity between current and past viral genotypes [22]. Thus, the currently observed internally neutral network should have been selected over the course of evolution.

The most intriguing stage of intra-host HCV evolution is the transition between two aforementioned stages, i.e., between the immune escape under positive selection and a conditionally stable state under the negative selection. The reported results are consistent with the hypothesis that this transition can be caused by the development of specific cooperative interactions among intra-host viral variants [22, 69]. Under this model, HCV immune adaptation is associated with antigenic cooperation among intra-host HCV variants [21, 22], owing to complementary roles played by viral variants in mitigation of neutralizing immune responses defined by their topological location in cross-immunoreactivity networks [68]. This model posits that intra-host viral populations evolve as quasi-social systems of functionally complementary variants [22, 69]. Such functional differentiation enables HCV adaptation to the changing intra-host environment as a group of cooperators rather than independent variants.

PART 3

RECONSTRUCTION OF VIRAL TRANSMISSIONS IN CONTACT NETWORKS WITH EXPECTED DEGREE DISTRIBUTIONS

3.1 Abstract

Reconstruction of transmission networks is a major problem of genomic epidemiology. Many methods have been developed in the past decade, which methods have showed that genomic data alone is not enough to reconstruct the transmission network of an outbreak. In this research we proposed to combine genomic data with information concerning the social component of epidemics. We considered that the contact network (outbreak network) is a random graph with expected degree distribution (for example, it can be scale-free). We proposed a Maximum Likelihood approach to find the transmission network as the tree with the highest embedding likelihood to a random contact network. In order to calculate the likelihood of embedding of a candidate tree within a contact network, we used a *uncapacitated facility location* problem that, in turn, can be reduced to a *maximum perfect matching* problem for bipartite graphs. In our approach, candidate transmission trees are produced using *simulated annealing* and *local search* algorithms. When applied to synthetic data, the proposed method in average correctly predicted $\sim 82\%$ of transmission links and $\sim 83\%$ of transmission ancestries. Thus, the proposed method is capable of inferring the majority of transmission links of an outbreak, and with more information specific to the outbreak better predictions can be achieved.

3.2 Introduction

Outbreaks of RNA based viral pathogens such as Hepatitis C virus (HCV) and Human Immunodeficiency Virus (HIV) are major concerns in public health. Reconstruction of viral transmissions of an outbreak is crucial for understanding the dynamics of viral spread. The history of viral transmissions can be formed based on the following information:

1. Transmission clusters with host involved in outbreaks.

2. Transmission history of each outbreak

Phylogenetic tree reconstruction from genomic data helps to detect transmission clusters of an outbreak. However, the question of 'who infected whom' within each cluster remains a complex problem to solve. Extraction of such information is a challenging task due to factors caused by the RNA virus nature. In particular, topologies of phylogenetic trees alone may not reflect the true transmission history [70]. Therefore, additional epidemiological and biological related information should be used along with sequencing data. In this research we propose a Maximum Likelihood approach that uses expected properties of contact networks of susceptible individuals for the accurate transmission network inference.

3.3 Materials and Methods

3.3.1 Method Description

A set of heterogenous population $P = P_i, ..., P_N$ is given where each population P_i consists of n haplotypes of length L sampled from a particular infected host. Based on this population, a Maximum Likelihood approach is used to reconstruct the transmission network. This approach requires the information mentioned below.

- A relatedness matrix W consisting evolutionary distances.

- A prior information I_s which is the expected degree distribution (EDD) of a contact network G_c. EDD reflects properties from spectral and structural properties of contact networks [71, 72]. In particular, here it is adjusted to account for some epidemiological settings such as needle-sharing.

Based on the relatedness matrix W of a genetic network (Fig. 3.1 (A)), a relatedness graph G_R (Fig. 3.1 (B)) is constructed. We assume that tree T is a subgraph of G_c and G_R at the same time. Tree T is one of the candidate spanning trees with the highest embedding likelihood inferred from the Maximum Likelihood approach. In this example, if the contact network is assumed to be scale-free, then the left tree on the Fig. 3.1 (C)) would be the most likely transmission tree (tree T).

In our approach, the relatedness network G_R is constructed as follows: given a threshold T, estimated as the minimal integer such that G_R is connected, where a pair of vertices i, j in G_R is adjacent if the distance $W_{i,j}$ between the corresponding populations does not exceed T. The obtained graph is then further sparsified out by applying the aforementioned procedure to each of its biconnected components.

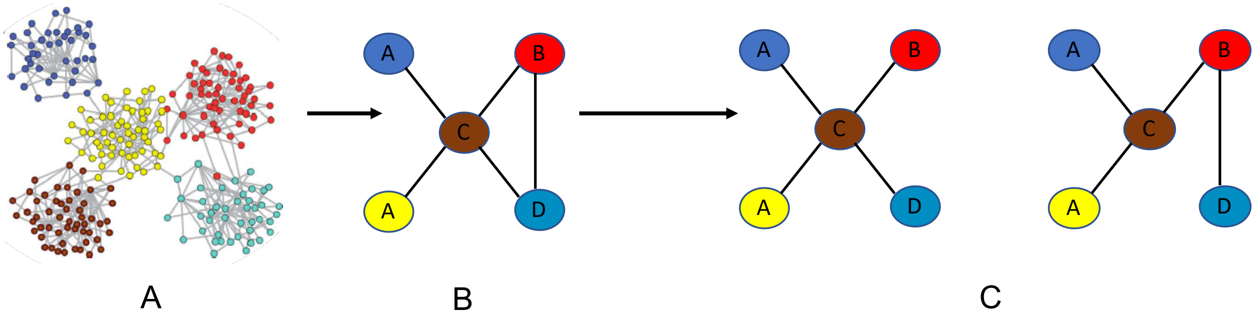


Figure (3.1) **(A)** Network of viral haplotypes. Different colors represent different hosts [73]. **(B)** Relatedness graph G_R. **(C)** Spanning trees derived from G_R.

3.3.2 Estimation of tree likelihood

We assume that the transmission tree T is a subgraph of G_C and one of the spanning trees of G_R. Then, every vertex $v \in V(T) = V(G_R)$ has a degree d_v in T.

Since the contact network is not observed we use only its expected degree counts $C = (C_1, ..., C_{N-1})$, where C_j is an expected number of vertices of degree j in G_C. An embedding of T in G_C, where I_s is the prior information EDD, is the tree likelihood calculated as follows:

$$p(T|I_s) = \max_D p(G_T|D)p(D) \tag{3.1}$$

In this study a Maximum Likelihood approach is used to solve this objective function (Equation 3.1).

3.3.3 Estimation likelihood of an embedding

Let C_j and σ_j be the numbers of vertices with expected degree j in G_C and G_T, respectively. The conditional probability $p(G_T|D)$, which is the likelihood of an embedding, is calculated as here [71], in which the probability of an edge between vertices with expected degrees D_i and D_j is equal to $\frac{D_i D_j}{2M}$. Based on that the following equation is deducted:

$$p(G_T|D) = \prod_{i,j \in E(G_T)} \frac{D_i D_j}{2M} = \frac{1}{(2M)^m} \prod_{i=1}^{n} D_i^{d_i}. \tag{3.2}$$

The counts σ_k follow a multivariate hypergeometric distribution such as $p(D) = \frac{1}{\binom{N}{n}} \prod_{j=1}^{N-1} \binom{C_j}{\sigma_j}$. After transition to log-probability, the problem in Equation. 3.2 can be represented as a *generalized uncapacitated facility location problem* with convex costs. This problem can be solved in polynomial time as shown by Hajiaghayi et al. [74].

3.3.4 Uncapacitated facility location algorithm

Let $C = \{u_1, ..., u_n\}$ represents the set of clients and $F = \{1, ..., N-1\}$ represents the set of facilities, and $d = (d_1, ..., d_n)$ is the degree sequence vector of G_T and $C = (C_1, ..., C_N - 1)$ is the expected degree count vector of G_C.

If client i is assigned to facility j, then (i.e. the vertex v_i has a degree j in G_C):

$$b_i j = \begin{cases} d_i \log j \text{ if } d_i \geq j \\ -\inf \text{ otherwise} \end{cases} \tag{3.3}$$

Assignment of σ_j number of clients to a facility j generates a profit of:

$$f_j = f_j \sigma_j = \log \binom{C_j}{\sigma_j} \tag{3.4}$$

Here the functions $-f_j(\sigma_j)$ are convex, so this problem can be reduced to the *maximum-weight perfect matching* problem for bipartite graphs [74]. Based on this algorithm, for each facility $j \in F$, one copy for each facility is created for each client i. Considering, the general convex function the opening costs of the j^{th} copy of a facility F will be $f_j(i+1) - f_j(i)$ for $0 \leq i \leq N-1$.

Since $-f_j(\sigma_j)$ is convex, the facility profit becomes $f_j = -\log\binom{C_j}{\sigma_{j+1}} + \log\binom{C_j}{\sigma_j}$. From the reduction if a j^{th} copy of a facility is used k^{th} copy must also be used for $k \leq j$, because f_j is convex. Finally, a bipartite graph $G = (X \cup Y, E)$ is constructed, where for each client i a vertex in the set of X is placed and for each facility j a vertex in the set of Y is placed. Finally, an edge i, j, is placed in E, where this edge is between a client i and facility j and its weight is:

$$-b_ij - f_j(\sigma_{k+1}) + f_j(\sigma_k) \tag{3.5}$$

Based on this bipartite graph G, a *maximum-weight perfect matching* is performed to assign clients to facilities in the most profitable way.

3.3.5 Simulated annealing and local search

Given G_R and the EDD of the contact network (G_C) our goal is to find the tree T with the highest likelihood of embedding within G_C. To solve this problem we used simulated annealing (SA) algorithm as described in Fig. 3.2, followed by a local search (LS) algorithm as described in Fig. 3.3.

Here $s(T)$ is the minimum spanning tree of G_R. T_{act} and P_{act} represent the actual (current) tree and its likelihood respectively. T_{new} is the new tree generated from T_{act} and P_{act} is again the likelihood of this new tree. The probability of accepting the new tree is calculated as follows: $p = \frac{\exp(P_{new} - P_{act})}{t}$, where t is the temperature used in SA algorithm. Temperature t is reduced by the following formula: $t \times (1 - \alpha)$, where α is the reduced factor.

In the LS diagram (Fig. 3.3), n and $iteration$ represent the total number of steps required for LS algorithm and the actual iteration number respectively. Finally T_{final} represents the tree with the highest embedding likelihood obtained after applying SA and LS algorithms.

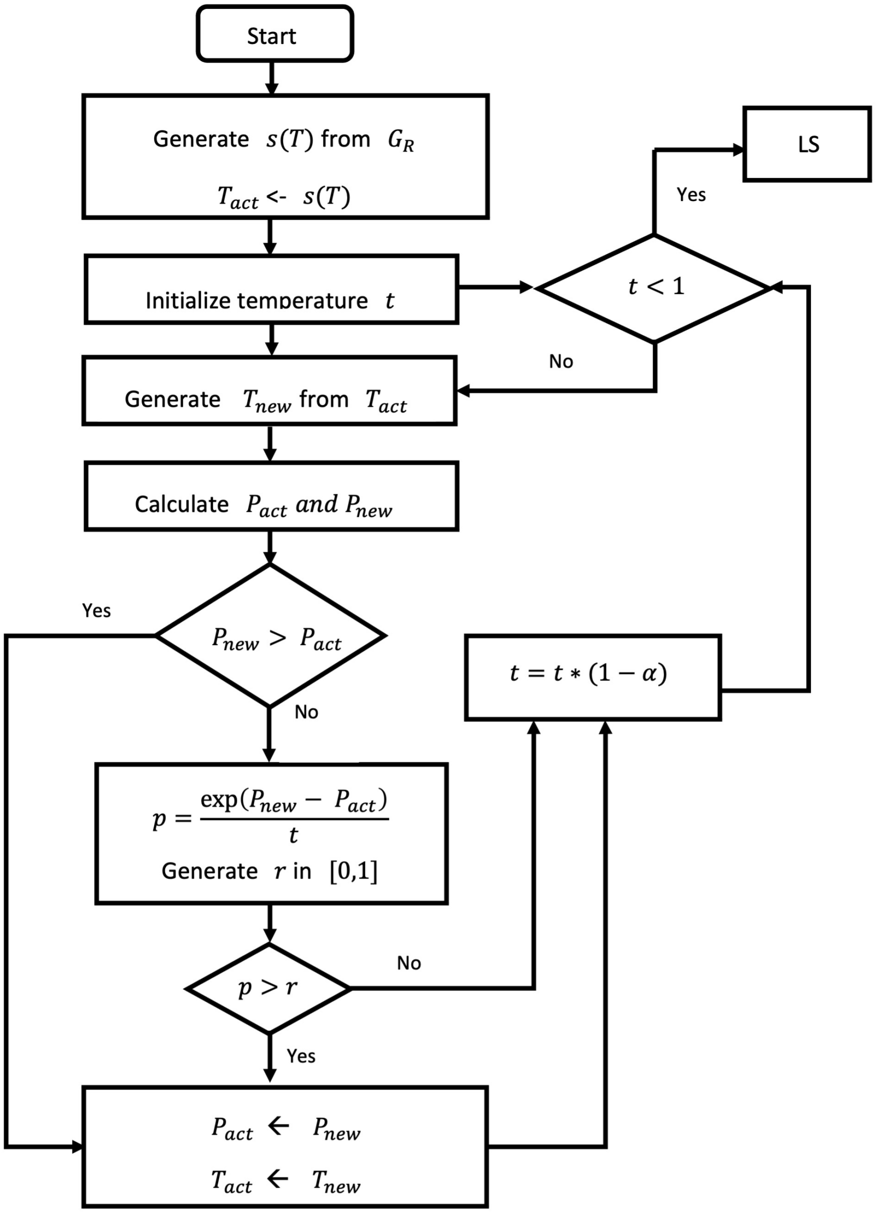

Start
Generate $s(T)$ from G_R
T_{act} <- $s(T)$
Initialize temperature t
$t < 1$
Yes
LS
No
Generate T_{new} from T_{act}
Calculate P_{act} and P_{new}
Yes
$P_{new} > P_{act}$
No
$t = t * (1 - \alpha)$
$p = \frac{\exp(P_{new} - P_{act})}{t}$
Generate r in [0,1]
No
$p > r$
Yes
$P_{act} \leftarrow P_{new}$
$T_{act} \leftarrow T_{new}$

Figure (3.2) Diagram of simulated annealing (SA).

3.3.6 Tree rearrangement

As shown in the SA diagram (Fig. 3.2) and the LS diagram (Fig. 3.3) T_{new} is generated from T_{act}. The tree modification here is performed by deleting a random edge from T_{act} which will result in two sub-trees. Then, two random vertices are chosen from these two sub-trees and connected to each other which creates our new spanning tree T_{new}.

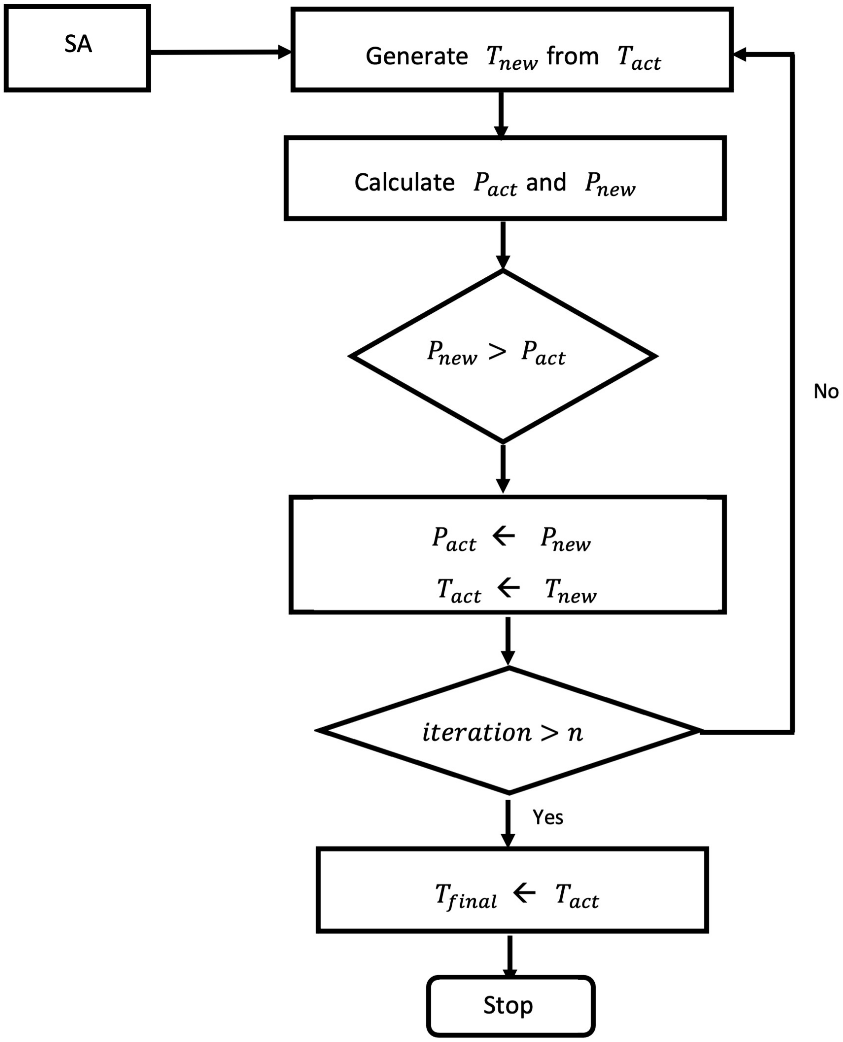

SA
Generate T_{new} from T_{act}
Calculate P_{act} and P_{new}
$P_{new} > P_{act}$
No
$P_{act} \leftarrow P_{new}$
$T_{act} \leftarrow T_{new}$
$iteration > n$
Yes
$T_{final} \leftarrow T_{act}$
Stop

Figure (3.3) Diagram of local search (LS).

Table (3.1) Common pairwise link and parent-ancestor link accuracies.

Number of nodes	**Mean link accuracy**	**Mean ancestry accuracy**
10	84.5%	85.3%
20	83.2%	85.6%
30	78.3%	77.6%

3.4 Results

We tested our method with simulated data with three different tree sizes ($n = 10, 20, 30$) where n is the number of vertices. We simulated the infection spreads over contact network according to the SI model with the transmission rate $\rho = 10^{-2}$. Each infected individual is assumed to carry a viral sequence, at each transmission events the source's sequence is transmitted to the recipient. The sequences mutate at the basic rate of $\mu = 10^{-6}$. The transmission tree generated from this simulator is used as the true tree which is compared to the tree inferred by our proposed method. For each different size of tree we used 10 different simulated data. As ground truth we used the tree generated from the simulator. As for the contact network (G_C) we set the maximum degree of vertices to a large enough number to cover all possible transmission networks. We used the EDD of the assumed network G_C calculated following a power-law degree distribution.

We applied our method as described in Fig. 3.2 and Fig. 3.3 starting with the G_R obtained from the simulator. As default values we set the parameters as initial temperature $t = 30$ with a reduced factor $\alpha = 0.003$ and following with an iteration number $n = 30$ for local search.

We applied our method 10 times for each tree with different nodes. We recorded the average accuracies of matching direct links and matching parent-ancestry links between the true tree and the inferred tree T_{final} with the maximum embedding likelihood (see Table 3.1).

3.5 Discussion

In this research we studied the problem of "who infected whom" within an outbreak. In addition to genomic data we incorporated the structure of an outbreak contact network. We suggested to solve this problem by finding the most "scale-free-like" spanning tree given a relatedness network and the EDD of a random scale-free contact network.

It is known that infection occurs through contact between susceptible individuals within a social network, even though the contact networks of outbreaks are not exactly known. Therefore, the transmission networks reflect the characteristics of these social networks [75, 76, 77, 78] Our model shows that the introduction of such prior information specific to the outbreak is useful to detect the source of the infection transmission and the structure of the transmission tree. However, this information alone is not always sufficient to find the most likely transmssion tree. The more prior information is introduced, the better the results would be.

Here in this study we tested our method with simulated HCV data and our preliminary results suggest that the concept of scale-free contact network is an important information for HCV outbreak transmission reconstruction. The general method composed of *uncapacitated facility location* algorithm reduced to *maximum-weight perfect matching* can be applied to different types of outbreak such as HIV and SARS-CoV-2 outbreaks. Especially during an emerging outbreak, this method could potentially complement genomic data analysis with adequate prior information.

PART 4

GLOBAL TRANSMISSION NETWORK OF SARS-COV-2: FROM OUTBREAK TO PANDEMIC

4.1 Abstract

The pandemic caused by the SARS-CoV-2 virus is straining health systems around the world. Although the Chinese government implemented severe restrictions on people's movement in an attempt to contain its initial spread, the virus had reached the majority of countries, partially due to its potent transmissibility and frequency of asymptomatic cases. As the pandemic continues, understanding its global transmission network properties is essential. The goal of this study is to characterize the network associated with the establishment of the pandemic.

We employ molecular surveillance data for inference and analysis of SARS-CoV-2 global transmission network, and exploit an algorithmic approach specifically tailored to emerging outbreak settings. It traces accumulation of viral genomic heterogeneity via mutation trees, that are then transformed into transmission networks.

SARS-CoV-2 into the majority of regions via heterogeneous transmission pathways. The transmission network is scale-free, with few genomic variants responsible for the majority of transmissions. The network structure is in line with the temporal data and suggest the expected sampling time difference of few days between potential transmission pairs. The findings emphasize the extent of the global epidemiological linkage and demonstrate importance of internationally coordinated containment measures.

4.2 Introduction

The COVID-19 pandemic due to the SARS-CoV-2 virus that emerged out of the city of Wuhan, China in December, 2019 [79, 80, 81, 82, 83], is now straining or overwhelming health care systems around the world. Hundreds of new confirmed cases have been reported daily in countries of every continent [84, 85, 86, 87, 88, 89]. To combat the spread of the virus, in the absence of a vaccine or specific treatments, travel restrictions and social distancing interventions have been put in place. In order to devise effective control strategies at different spatial scales, it is critical to have a clear understanding of transmission pathways of SARS-CoV-2, including local human-to-human transmission dynamics [90, 91, 81] as well as long-range transmission events (country-to-country) [92].

In this study, we sought to characterize the transmission pathways that facilitated the virus to establish itself as a pandemic. We utilized a network-based approach to infer and analyze the global SARS-CoV-2 transmission network using viral genomes sampled around the world before the pandemic declaration by WHO on March 11, 2020. The richness and accessibility of genomic data accumulated in almost real time owing to wide utilization of next-generation sequencing technologies distinguishes this SARS-CoV-2 pandemic from previous large-scale epidemics including the 2009/AH1N1 influenza pandemic and the 2003 SARS outbreaks. Another feature of the scope of the genomic data for COVID-19 is the high sampling density available to investigate the early transmission stages. Indeed, although the virus genetic diversity gradually increases as the virus spreads, particular genomic variants have been repeatedly sequenced at different time points and geographical locations. These observations indicate that the available sequencing data cover a significant part of the evolutionary space explored from the onset of the epidemics and is less susceptible to the effects of sampling and reporting biases.

For such molecular surveillance data, viral evolution could be accurately modeled, reconstructed and visualized by genetic networks models [34, 93, 76], that have been successfully used for the analysis of different viral epidemics [93, 76, 23] and shown to be particularly efficient to ascertain transmission links compared to methods based on binary phylogenies [76]. However, in emerging outbreak settings, when all circulating viral genomes are relatively close to each other, it is necessary to employ methods that allow the analysis of viral population structures at finer resolution than provided by state-of-the-art network models. We achieve such resolution by using mutation trees [94] associated with character-based phylogenies that keep track of the accumulation of mutations in viral populations. These trees, in turn, allow for accurate inferences of transmission networks. Note that besides us, other researchers simultaneously and independently suggested the usefulness of character-based phylogenies for SARS-CoV-2 evolutionary analysis [95].

Our results presented here summarize the up-to-date picture of the spread of SARS-CoV-2 between and within countries and geographic regions. The structural properties of the inferred network, transmission clusters and pathways as well as virus introduction routes emphasize the extent of the global transmission network. It also demonstrates the importance of internationally coordinated public health measures and highlights how epidemiological and molecular surveillance analyses complement each other to characterize the spatial-temporal spread of epidemics.

4.3 Materials and Methods

4.3.1 Data preprocessing.

We obtained the genomics data and associated metadata from the first stage of COVID-19 epidemics (before March 11, 2020) from the GISAID database [96]. The sequences identified by GISAID as low-quality have been removed. The reference genome was taken from the literature [97]. It coincides with the most prevalent sequence sampled from Wuhan at the outbreak onset and from a number of other locations. Given this fact, we assume that the reference represent the original states (major alleles) of analyzed genomic positions. The sequences were aligned to the reference using MUSCLE [37] and trimmed to the same length of $m = 29772$ bp. In order to be as conservative as possible in the mutation calling, gaps and non-identifiable positions have been assumed to have major alleles. As a result, the set of $n = 319$ sequences $\mathcal{S} = \{s_1, ..., s_n\}$ have been selected for the further analysis.

4.3.2 Dimensionality reduction and clustering.

Viral sequences were embedded into the 2-dimensional space using T-distributed Stochastic Neighbor Embedding (t-SNE) [98] based on the Hamming distance. The embedding points were clustered by k-means clustering, and the optimal number of clusters was estimated using the gap statistics [99].

4.3.3 Mutation tree reconstruction.

First, genomic positions without variation have been removed, leaving $m = 274$ single nucleotide variants (SNVs) for further analysis. Finally, the alignment was represented by the $n \times m$ $(0, 1)$-mutation matrix M with rows corresponding to sequences and columns corresponding to SNVs, where $M_{i,j} = 1$ whenever the i-th genome has a minor allele at the position j with respect to the reference.

The major structure used for such inference is a *mutation tree,* where

- internal nodes correspond to mutations, with the root representing the zero mutation (the absence of mutations);
- leafs represent sampled genomes;
- mutational profile of each sequence consists of mutations on the path from the corresponding leaf to the root.

This tree does not have to be binary. In a perfect phylogeny, which is the simplest character-based phylogenetic model [100], each mutation occurs only once and can be represented by a single internal node. This model can only explain the data without 4-gamete rule violation, i.e. whenever for each pair of columns in the mutation matrix M, there are no 4 sequences that have all possible combinations of alleles $(0, 0), (0, 1), (1, 0), (1, 1)$ at that positions. SARS-CoV-2 sequences contain several 4-gamete rule violations, thus implying repeated mutations in the same genomic positions. Therefore, we fit the data to more general Camin-Sokal phylogenetic model, that allows homoplasy in the form of repeated mutations, while mutation losses are not allowed and each mutation could be acquired at most twice [101].

We first identify potential repeated mutations and then construct plausible mutations trees taking into account for the possibility of false SNV calls resulting from the sequencing noise as follows:

1) *Identify potential repeated mutations.* This is achieved using a graph G_{4g}, whose vertices are SNVs, and two vertices are adjacent whenever the corresponding pair of SNVs violates 4-gamete

rule [102].

In this graph, we are looking for the minimum vertex cover, i.e. the minimum set of vertices whose removal destroys all edges. The set of genomic positions corresponding to vertices in a minimum vertex cover forms the most parsimonious set of mutations that should be repeated.

We find all minimal vertex covers using Bron-Kerbosch algorithm [103] for maximum independent sets generation; the complements of maximum independent sets are exactly minimum vertex covers. Since the number of 4-gamete rule violations for SARS-CoV-2 data was relatively small, this method was fast and allowed to significantly simplify the phylogeny reconstruction.

2) *Construct mutation trees.*

For each minimum set $R = \{m_1, ..., m_k\}$ of potential repeated mutations from the previous step, we construct a Camin-Sokal phylogeny which minimizes the number of mismatches with the original mutation matrix M as follows. We generate an extended set of mutations $P = \{1, ..., m, m+1, ..., m+k\}$, where each original mutation $j \in \{1, ..., m\} \setminus R$ is represented by a single copy and each mutation $j \in R$ is represented by a pair of copies $C(j)$. The sought-for Camin-Sokal phylogeny T would be a perfect phylogeny with respect to the extended set of mutations P. To construct this phylogeny and the corresponding extended mutation matrix X, we utilize an integer linear programming (ILP) approach [102]. The following binary variables are used:

(a) $X_{i,j} = 1$ whenever the genomic variant i has a mutation j from the extended set P, $i = 1, .., n$; $j = 1, ..., m+k$.

(b) $D_{j,l,a,b} = 1$ whenever there is a sequence that have an allele combination (a, b) at the positions j and l, $j = 1, ..., m+k$; $l = j+1, ..., m+k$; $(a, b) \in \{(0,0), (0,1), (1,0), (1,1)\}$. Then we seek to minimize the total number of mismatches between the observed mutation matrix M and the mutation profiles defined by the tree T by minimizing the objective function

$$\sum_{(i,j):M_{i,j}=0} \sum_{p \in C(j)} X_{i,p} + \sum_{(i,j):M_{i,j}=1} \sum_{p \in C(j)} (1 - X_{i,p}) \tag{4.1}$$

subject to constraints

$$D_{j,l,1,1} - X_{i,j} - X_{i,l} \geq -1, \tag{4.2}$$

$$D_{j,l,1,0} - X_{i,j} + X_{i,l} \geq 0, \tag{4.3}$$

$$D_{j,l,0,1} + X_{i,j} - X_{i,l} \geq 0, \tag{4.4}$$

$$D_{j,l,0,0} + X_{i,j} + X_{i,l} \geq 1, \tag{4.5}$$

$$D_{j,l,0,0} + D_{j,l,0,1} + D_{j,l,1,0} + D_{j,l,1,1} \leq 3 \tag{4.6}$$

$$\sum_{p \in C(j)} X_{i,p} \leq 1 \tag{4.7}$$

$$i = 1, ..., n; \;\; j = 1 : m + k; \;\; l = j + 1, ..., m + k \tag{4.8}$$

The first 4 sets of constraints enforces the relations between the variables $X_{i,j}$ and $D_{i,j,a,b}$ specified by (a) and (b), the fifth set of constraints guarantees that T is the perfect phylogeny with respect to the extended set of mutations P, and the last set of constraints ensures that two gains of the same mutation appear only in parallel lineages. The instances of the ILP problem were solved to optimality using Gurobi 8.1. The trees were constructed for all potential sets of repeated mutations, and the tree with the best objective function was selected.

4.3.4 Transmission network construction and bootstrapping.

The transmission network defined by the mutation tree T is a directed graph, whose vertices represent viral genomes, and two genomes are connected by an arc if their mutational composition

suggests potential direct or indirect transmission linkage between their hosts. For a given mutation tree T, the corresponding transmission network G_T is constructed as follows:
1) Collapse sequences that share the same parent in T into a single *haplotype*. The set of haplotypes forms the vertex set of G_T.
2) A pair of haplotypes h_i and h_j are connected by a directed arc whenever (a) the parent of h_i is an ancestor of the parent of h_j and (b) there is no haplotype h_k whose parent belongs to the path between the parents of h_i and h_j.
In graph-theoretical terms, G_T is the transitive reduction of the reachability graph of the parents of observed haplotypes.

Some SNV calls could be possibly misplaced due to the sequencing noise and underlying complexity of the SARS-Cov-2 genome. To account and quantify the uncertainty for the hypothesized transmission links, bootstrapping of the transmission networks was performed. At each bootstrap, m mutations were sampled with replacement from the the original set of mutations, and the mutation tree and the transmission network were constructed using the obtained mutation matrix as discussed above. For each potential transmission link, its bootstrap probability was calculated. The final consensus transmission tree was estimated as the maximum-weight spanning arborescence [104] with respect to the edge probabilities.

4.4 Results

4.4.1 Transmission clusters.

t-SNE/clustering analysis suggest that as of March 11, 2020 five distinct viral subpopulations have been circulating globally (Fig. 4.1):

1) The original cluster(black) that includes sequences sampled from Wuhan and other mainland China provinces during the early transmission phase, as well as from USA, Singapore, Taiwan, Thailand, South Korea, Nepal and several European countries are most probably epidemiologically linked with China at earlier outbreak stages.

2) European cluster (red) that includes sequences almost exclusively from European countries,

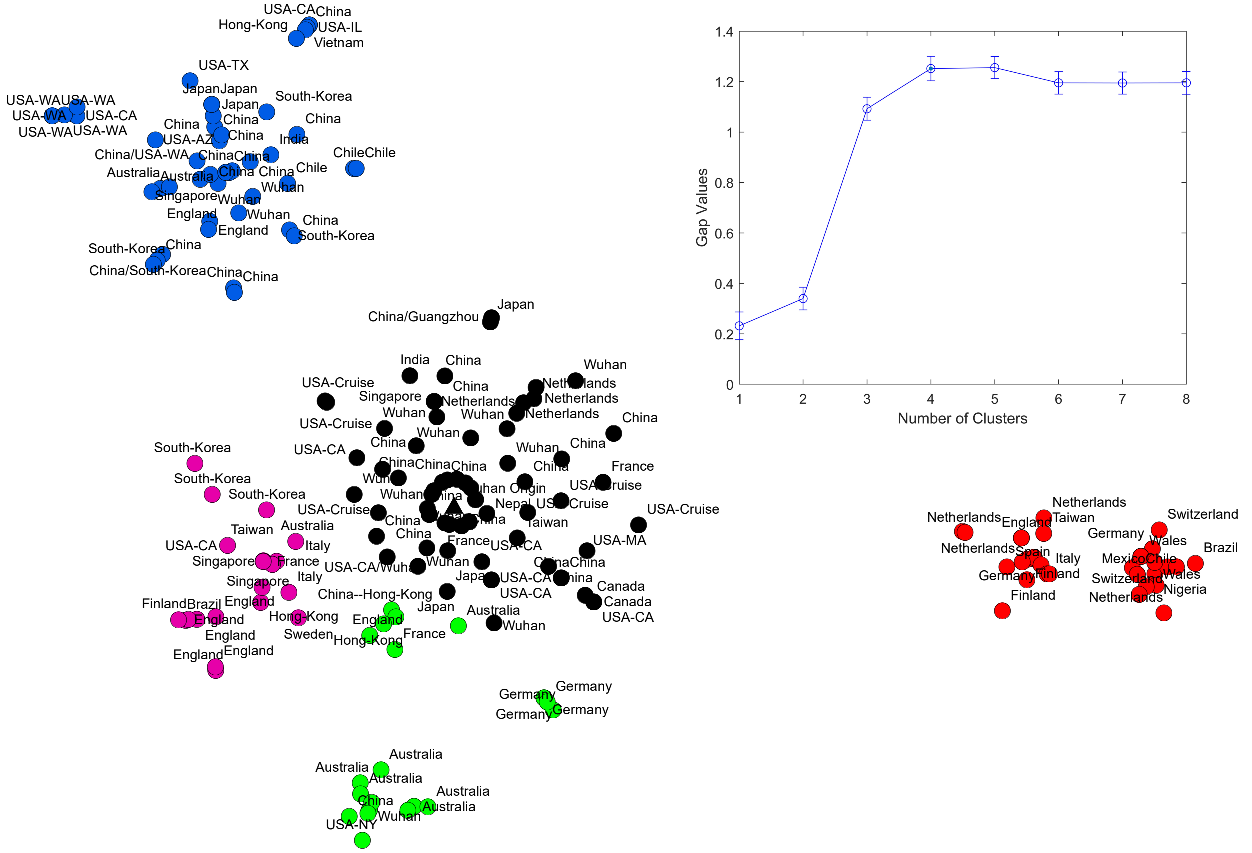


Figure (4.1) t-SNE plot of observed SARS-CoV-2 genomes. Wuhan-1 haplotype is depicted as a triangle. Identified clusters are highlighted in different colors. In the upper right corner, gap statistic values are shown.

as well as from a few non-European countries (Nigeria, Mexico, Brazil) most of whom are documented to be epidemiologically linked to Europe.

3) The cluster from mostly Pacific countries (blue) that includes sequences from mainland China, South Korea, Hong Kong, Singapore, Vietnam, Australia, USA and Chile, including the subcluster of sequences from the US Washington State.

4) The mixed cluster that include significant portion of genomes sampled in Australia and New Zealand (green). Among their hosts, three were reported to have travel history to Iran. It may mean that this strain is a branch of an Iranian strain tied to the epidemics at a much larger scale in that country.

5) The cluster that includes sequences sampled across several regions of Southeastern Asia (Hong Kong, South Korea, Taiwan, Singapore), as well as the major United Kingdom cluster (violet).

We observed the negative log-linear relationship between cluster sizes and maximum pairwise Hamming distances between the genomes inside the clusters (Fig. 4.3a, $R^2 = 0.984$, $p < 0.001$). It suggest a preferential attachment-like mechanism of cluster formation, where the majority of newly appearing genomes tend to concentrate around few older genomes. This mechanism is further confirmed by the following analysis.

4.4.2 Transmission network.

The inferred transmission network is visualized in Fig. 4.2, haplotypes are annotated by the full list of their sampling locations. Although the majority of haplotypes (80%) were sampled in a single geographical location, a number of them were sampled in multiple locations, including the most prevalent haplotype sampled in Wuhan and associated with the initial phase of the epidemic (highlighted in red) that is further referred to as Wuhan-1 haplotype. For the potential transmission links e, their bootstrapping-based probabilities p_e are reported.

4.4.3 Comparison of different transmission detection tools.

We compared the performance of our tool (CS-phylogeny) to others that are currently available. For this purpose we compiled epidemiological links prior to March 11th, 2020 from different sources, such as news and articles. These compiled epidemiological links were marked as true transmission links. Table shows sensitivity and specificity of the transmission detection tools based on the detected transmission links. With a sensitivity of 80% CS-phylogeny outperformed other tools (Table 4.1).

Network and temporal information

The network structure was found to be in line with the reported genome sampling times, even though the network was constructed using the genomic data alone. Indeed, the correlation

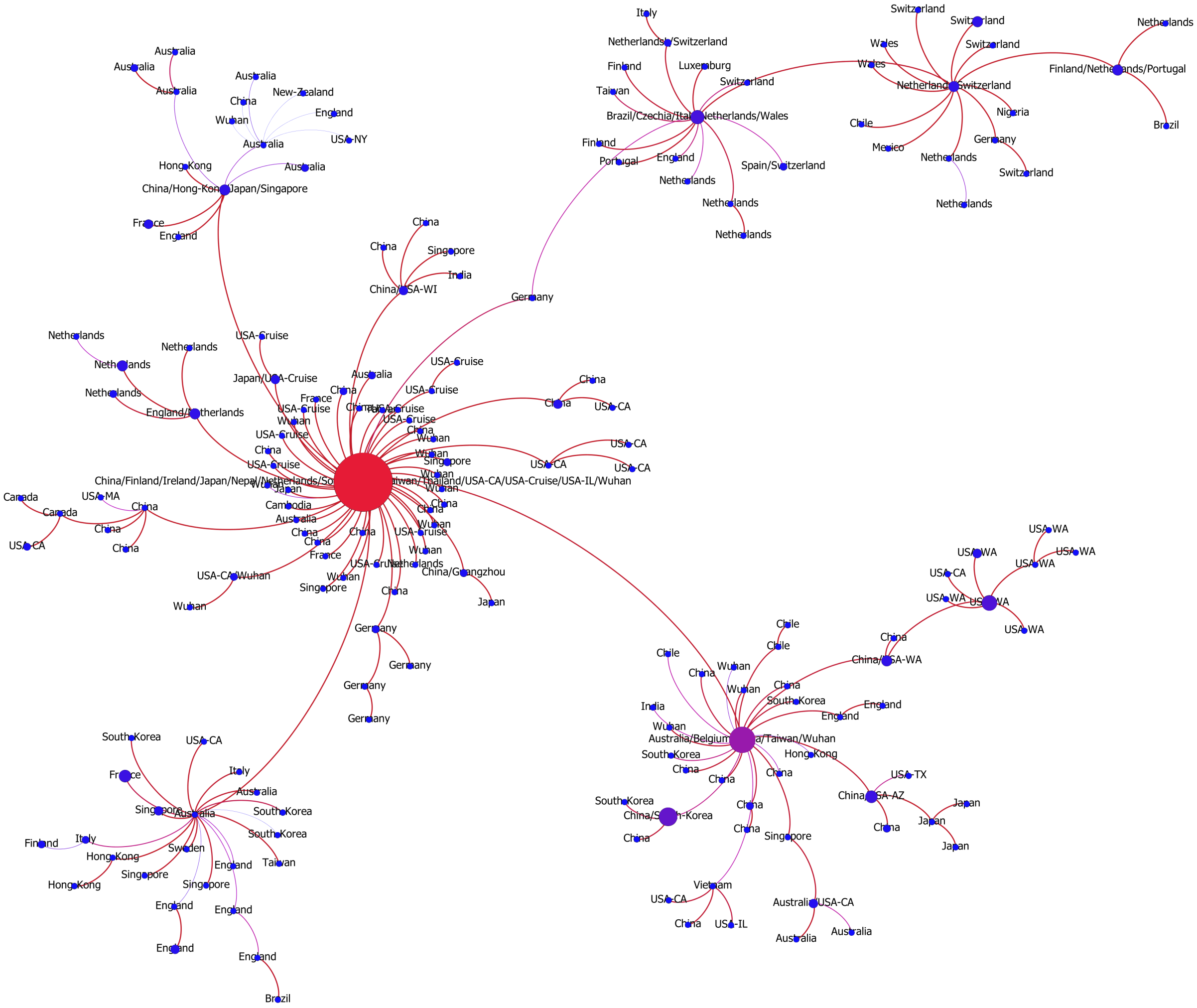
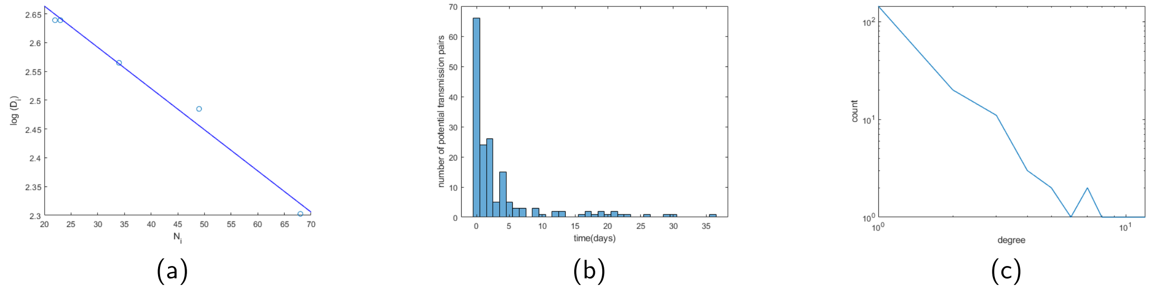


Figure (4.2) The transmission network constructed using SARS-CoV-2 genomes available as of March 10, 2020 and vizualized using Gephi [105]. Vertices represent viral genomes, and two vertices are connected by an arc if their mutational composition suggests potential direct or indirect transmission linkage between their hosts. Each vertex is annotated by the list of geographical locations were it was sampled. The thickness and color (from blue to red) of the edges are proportional to their bootstrapping probabilities.

between network distances and differences in first sampling times between ancestor-descendant pairs of network nodes is 0.78 ($p < 10^{-116}$). The fact that this correlation is not absolutely

Table (4.1) Sensitivity is defined as the ratio of true transmission links formed by the tool to the total number of true links. Specificity is defined as the ratio of true links formed by the tool to the total number of links formed by the tool.

Tool	**Sensitivity**	**Specificity**
CS-phylogeny	80%	4.76%
NETWORK5011CS	72%	4.99%
RAxML	64%	4.26%
bitrugs	52%	3.38%
hivtrace (t=0)	40%	7.81%
outbreaker	28%	5.83%
phybreak	4%	0.83%



(a) (b) (c)

Figure (4.3) (a) Relationship between cluster sizes and the logarithms of maximum pairwise distances. (b) The distribution of sampling time differences for linked pairs. (c) Degree distribution of the transmission network (in log-log scale).

perfect is not surprising, as sampling times are prone to sampling and reporting biases and may not accurately reflect actual transmission times. Even in such settings, for 86.10% of potential transmission pairs, their sampling times agree with each other, i.e. the source was sampled earlier than the recipient; and for 94.25% of these pairs their sampling times either agree or differ by at most 7 days. Finally, the mean minimum time difference between sampling times of potential transmission pairs was 3.74 days (95% $CI = [2.75, 4.73]$, Fig. 4.3b), while for the random pair of haplotypes the expected time difference was 20.48 days (95% $CI = [20.21, 20.74]$), making this difference significant ($p < 10^{-45}$, Kolmogorov-Smirnov test).

Network structure

The transmission network is robust to an input data variation, with 97.34% of its edges being supported by the majority of bootstrap experiments, and 78.19% of edges having bootstrapping probabilities above 95%.

SARS-CoV-2 transmission network appears to be scale-free, with the the right-skewed degree distribution (Fig. 4.3c). Degree distributions of such networks follow power law (i.e. the probability of having a particular degree is proportional to the power of that degree), and they are often the result of a preferential attachment process, where a vertex joining a network gets connected to an existing vertex with the probability proportional to the degree of that vertex - the model is often described by the metaphor "the rich get richer". Following [76, 106], we fitted negative binomial, Yule, Pareto and Waring distributions to the observed degree distribution of the transmission network. To compare the goodness of fit yielded by different models, we used the Akaike (AIC) and Bayesian (BIC) Information Criteria (Table 4.2). The Pareto distribution, that represent the classical power-law, demonstrated the best fit. The exponent of the Pareto distribution was estimated to be 1.20 ($95\%CI = [1.12, 1.34]$), indicating the higher tendency of vertices to be connected to hubs (high-degree vertices). The correlation between vertex degrees and sampling frequencies of the corresponding genomes was high: $\rho = 0.8932$, $p < 10^{-65}$. All these observations suggest that few genomes were responsible for the majority of possible transmissions.

Transmission history

The structure of the potential transmission network mostly agrees with the distribution of t-SNE clusters (Fig. 4.1) and allows to hypothesize multiple transmission routes. The virus spread is characterized by multiple introductions of SARS-CoV2 into regions and countries: for 14 out of 34 countries with reported sequences multiple introductions could be claimed. Below we summarize the information about the transmission pathways in different regions outside the mainland China (whose subnetwork is depicted on Fig. C.1) that could be deduced from the network.

USA (Fig. C.2). There are indications of multiple introductions of SARS-CoV2 into the country, as well as of sustained human-to-human transmissions inside the country prior to March

Table (4.2) Comparison of different models for the transmission network degree distribution.

Distribution	AIC	BIC
Negative Binomial	708.49	714.96
Pareto	603.48	609.95
Waring	704.59	711.06
Yule	804.91	808.15

11. Most of introduced haplotypes could be directly linked to the first epidemic wave in mainland China, with the average graph distance between US haplotypes (Washington cases excluded) and Wuhan-1 haplotype being equal to $d = 1.79$ links. In particular, the state of California alone could have exhibited multiple introductions with no identified significant clustering of cases, as its observed viral haplotypes were either also sampled in China (2 cases) or linked directly to the haplotypes in mainland China (2 cases, $p_e = 1$), or to haplotypes from Singapore, Vietnam, Australia and Canada (4 cases, $p_e = 0.98, 0.97, 0.99$ and 1, respectively) which are, in turn, linked to haplotypes sampled in mainland China ($p_e = 0.98, 0.77, 1$ and 1). In contrast, the haplotypes from the state of Washington formed a connected subtree with the root sampled in China, thus suggesting a single introduction followed by the sustained human-to-human transmissions inside the state (mean $p_e = 0.99$). In addition, the network suggests independent introductions to Massachusetts, Wisconsin, New York, Illinois and Arizona. Two possible cases of virus transmission between US states were identified: from Arizona to Texas ($p_e = 0.83$) and from Washington to California ($p_e = 1$). The sequences from the Grand Princess cruise ship could be linked to a single case identical to the Wuhan-1 haplotype.

Western Europe. The major European cluster was linked to the Wuhan-1 haplotype through the haplotype sampled in Germany on January, 28, 2020 ($p_e = 0.79$) (Fig. C.3). The parent of this haplotype is the Wuhan-1 haplotype ($p_e = 0.91$) and its only child is the haplotype later sampled in Italy. This potential transmission route is in agreement with epidemiological and molecular evidence reported by other sources [107, 108, 109, 95]. For Italy, the analysis suggests that there was another independent SARS-CoV-2 introduction with no genomic evidence of the further spread. Similarly, at least two introductions are hypothesized for Netherlands (Fig. C.4);

however in that case both resulted in sustained host-to-host transmissions inside the country. The first Netherlands cluster was the part of the major European cluster, while the second one could be linked to the Wuhan-1 haplotype ($p_e = 1$). This cluster has the genetic signature in the form of codon deletion in nsp2 genomic region that was observed only there; furthermore, both viral subpopulations are not geographically separated and co-exist in the same cities. A similar situation was observed in Germany, where two separate subpopulations coexisted: first of them is linked to the major European cluster, while the second one was directly linked to the Wuhan-1 haplotype ($p_e = 1$). Multiple introductions have also been observed in Finland. Haplotypes from Switzerland, Spain and Czech Republic are only observed in the main European cluster and most probably were introduced from Italy; in the latter case, this claim has an epidemiological support as the infected person reportedly had traveled to Italy.

The epidemiological history in the United Kingdom seems to be quite different from that of continental Europe (Fig. C.5). There were multiple separate clusters detected there, three of which were not associated with other European cases and directly linked to the sequences sampled in China and Australia in January, 2020 (mean $p_e = 0.82$). Of the remaining haplotypes, one is linked to the second Netherlands cluster, while others belong to the major European cluster (the majority of them being sampled in Wales). Two clusters showed indications of intra-country transmissions (mean $p_e = 0.93$).

East Asia. All introductions in Singapore, Japan and South Korea (Figs. C.6, C.7 and C.8) were linked to the haplotypes observed in mainland China. Singapore possibly experienced four such introductions (mean $p_e = 0.99$). In both Japan and South Korea, two potential introductions were predicted, one linked to Wuhan-1 haplotype, and the other linked to intra-country transmissions (mean $p_e = 0.99$ in Japan and $p_e = 0.86$ for South Korea).

Australia. There are indications of at least three potential virus introductions to Australia either from mainland China or, as discussed in the previous subsection, from Iran, although the latter claim is currently based only on epidemiological evidence. Three of the corresponding clusters have evidences of intra-country transmissions (mean $p_e = 0.88$).

Central and South America and Africa. All viral variants sampled in Nigeria and Mexico, 2 variants from Brazil and 1 variant from Chile were linked to haplotypes from the major European cluster (mean $p_e = 1$) and have reported travel history to Italy, while one variant from Brazil is linked to the genome from the United Kingdom ($p_e = 0.99$), thus confirming that the virus was imported from continental Europe. On the other hand, three other Chilean haplotypes belong to the Pacific cluster and could be linked to the haplotype sampled in mainland China, Taiwan and Australia (mean $p_e = 0.92$).

4.5 Discussion

In this work, we report the results of molecular surveillance analyses of SARS-CoV-2 prior to the transition from epidemic to pandemic state. Our aim was to identify and analyze the transmission pathways that allowed the epidemic to rapidly progress from the initial outbreak in Wuhan to the pandemic that is now affecting almost every geographic region of the globe [110, 79, 111]. To achieve this aim, we implemented a computational framework to recover the network of potential SARS-CoV-2 transmissions from the aggregated genomic data. The analysis allowed to identify potential transmission links and routes of disease introduction in different geographic regions, and confirm the presence of multiple sources of introduction to the majority of countries. This conclusion supports the implementation of travel restrictions and border screening. The fact that a number of countries exhibited multiple introductions demonstrates how the global transportation network allowed the virus to exploit multiple transmission pathways and spread so rapidly. It underscores the need to put in place coordinated efforts involving multiple countries in order to achieve epidemic control.

The scale-free structure of the global SARS-CoV-2 transmission network suggests that few genomes were responsible for the spread of the virus. The question whether this pattern is due to founder effects, epidemiological settings or differences in phenotypic features across SARS-CoV-2 genomic variants requires further investigation. Furthermore, scale-freeness of the network should be taken into account in epidemiological modelling and planning of public health intervention measure, since epidemic processes on such networks exhibit a specific behaviour [112, 113]. Although additional data that appeared since the declaration of the pandemic will add vertices and links to the network, its general structural properties will likely be preserved.

The ongoing pandemic of SARS-CoV-2 is the first global public health emergency for which next-generation sequencing technologies have been employed at such scale. This has led to a high density sampling that is unprecedented for both the geographic extent of virus spread and the evolutionary space explored by the virus since its emergence. It provides the means of tracking virus spread and evolution across time and space using the methods of computational genomics and molecular epidemiology. Automatic high-performance computing-based molecular surveillance systems such as Nextstrain [114], HIV-Trace [115] and GHOST [59] could be instrumental in such global surveillance and decision making. However, the results of such molecular surveillance analyses should be interpreted with caution. First of all, the genomic analysis do not necessarily replace traditional epidemiology methods. It rather complements other epidemiological investigations using the sequencing data as an independent source of information that is not subject to the biases associated with the traditional epidemiological data [116, 117]. Second, it is important to understand that the edges in the estimated global transmission network may not be synonymous with actual transmission events, but rather link infected hosts from the same epidemiological transmission clusters. Furthermore, underreporting and undersampling effects still could be sensable. The dataset available for this analysis is a convenience sample rather than a random sample within infected individuals, which results from the aggregation of data from different countries and sequencing labs and instruments. This is an inevitable consequence of sequencing data analysis since the procedure itself can be relatively expensive when implemented on a large scale [118, 119], and the decision to sequence each particular case is largely done subjectively in each specific country and lab.

REFERENCES

[1] D. B. Burkhardt, J. S. Stanley, A. Tong, A. L. Perdigoto, S. A. Gigante, K. C. Herold, G. Wolf, A. J. Giraldez, D. van Dijk, and S. Krishnaswamy, "Quantifying the effect of experimental perturbations at single-cell resolution," *Nature Biotechnology*, pp. 1–11, 2021.

[2] S. A. Byron, K. R. Van Keuren-Jensen, D. M. Engelthaler, J. D. Carpten, and D. W. Craig, "Translating rna sequencing into clinical diagnostics: opportunities and challenges," *Nature Reviews Genetics*, vol. 17, no. 5, pp. 257–271, 2016.

[3] S. Consortium *et al.*, "A comprehensive assessment of rna-seq accuracy, reproducibility and information content by the sequencing quality control consortium," *Nature biotechnology*, vol. 32, no. 9, p. 903, 2014.

[4] B. Langmead and S. L. Salzberg, "Fast gapped-read alignment with bowtie 2," *Nature methods*, vol. 9, no. 4, p. 357, 2012.

[5] H. Li, "Aligning sequence reads, clone sequences and assembly contigs with bwa-mem," *arXiv preprint arXiv:1303.3997*, 2013.

[6] H. Xin, D. Lee, F. Hormozdiari, S. Yedkar, O. Mutlu, and C. Alkan, "Accelerating read mapping with fasthash," in *BMC genomics*, vol. 14, pp. 1–13, Springer, 2013.

[7] J. Sirén, N. Välimäki, and V. Mäkinen, "Indexing graphs for path queries with applications in genome research," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 11, no. 2, pp. 375–388, 2014.

[8] H. Li, "Minimap2: pairwise alignment for nucleotide sequences," *Bioinformatics*, vol. 34, no. 18, pp. 3094–3100, 2018.

[9] F. J. Sedlazeck, P. Rescheneder, and A. Von Haeseler, "Nextgenmap: fast and accurate read mapping in highly polymorphic genomes," *Bioinformatics*, vol. 29, no. 21, pp. 2790–2791, 2013.

[10] H. Ponstingl and Z. Ning, "Smalt-a new mapper for dna sequencing reads. 2010," *F1000 Posters*, vol. 1.

[11] M. Zaharia, W. J. Bolosky, K. Curtis, A. Fox, D. Patterson, S. Shenker, I. Stoica, R. M. Karp, and T. Sittler, "Faster and more accurate sequence alignment with snap," *arXiv preprint arXiv:1111.5572*, 2011.

[12] Y. Liao, G. K. Smyth, and W. Shi, "The subread aligner: fast, accurate and scalable read mapping by seed-and-vote," *Nucleic acids research*, vol. 41, no. 10, pp. e108–e108, 2013.

[13] K. Chen, J. W. Wallis, M. D. McLellan, D. E. Larson, J. M. Kalicki, C. S. Pohl, S. D. McGrath, M. C. Wendl, Q. Zhang, D. P. Locke, *et al.*, "Breakdancer: an algorithm for high-resolution mapping of genomic structural variation," *Nature methods*, vol. 6, no. 9, pp. 677–681, 2009.

[14] T. Rausch, T. Zichner, A. Schlattl, A. M. Stütz, V. Benes, and J. O. Korbel, "Delly: structural variant discovery by integrated paired-end and split-read analysis," *Bioinformatics*, vol. 28, no. 18, pp. i333–i339, 2012.

[15] S. Niehus, H. Jónsson, J. Schönberger, E. Björnsson, D. Beyter, H. P. Eggertsson, P. Sulem, K. Stefánsson, B. V. Halldórsson, and B. Kehr, "Popdel identifies medium-size deletions simultaneously in tens of thousands of genomes," *Nature communications*, vol. 12, no. 1, pp. 1–10, 2021.

[16] N. L. Bray, H. Pimentel, P. Melsted, and L. Pachter, "Near-optimal probabilistic rna-seq quantification," *Nature biotechnology*, vol. 34, no. 5, pp. 525–527, 2016.

[17] R. Patro, G. Duggal, M. I. Love, R. A. Irizarry, and C. Kingsford, "Salmon provides fast and bias-aware quantification of transcript expression," *Nature methods*, vol. 14, no. 4, pp. 417–419, 2017.

[18] S. Blach, S. Zeuzem, M. Manns, I. Altraif, A.-S. Duberg, D. H. Muljono, I. Waked, S. M. Alavian, M.-H. Lee, F. Negro, *et al.*, "Global prevalence and genotype distribution of

hepatitis c virus infection in 2015: a modelling study," *The Lancet Gastroenterology & Hepatology*, vol. 2, no. 3, pp. 161–176, 2017.

[19] J. E. Zibbell, K. Iqbal, R. Patel, A. Suryaprasad, K. Sanders, L. Moore-Moravian, J. Serrecchia, S. Blankenship, J. Ward, and D. Holtzman, "Increases in hepatitis c virus infection related to injection drug use among persons aged$\leq$ 30 years-kentucky, tennessee, virginia, and west virginia, 2006-2012.," *MMWR. Morbidity and mortality weekly report*, vol. 64, no. 17, pp. 453–458, 2015.

[20] E. Domingo, J. Sheldon, and C. Perales, "Viral quasispecies evolution," *Microbiology and Molecular Biology Reviews*, vol. 76, no. 2, pp. 159–216, 2012.

[21] S. Ramachandran, D. S. Campo, Z. E. Dimitrova, G.-l. Xia, M. A. Purdy, and Y. E. Khudyakov, "Temporal variations in the hepatitis c virus intrahost population during chronic infection," *Journal of virology*, vol. 85, no. 13, pp. 6369–6380, 2011.

[22] P. Skums, L. Bunimovich, and Y. Khudyakov, "Antigenic cooperation among intrahost hcv variants organized into a complex network of cross-immunoreactivity," *Proceedings of the National Academy of Sciences*, vol. 112, no. 21, pp. 6653–6658, 2015.

[23] D. S. Campo, G.-L. Xia, Z. Dimitrova, Y. Lin, J. C. Forbi, L. Ganova-Raeva, L. Punkova, S. Ramachandran, H. Thai, P. Skums, *et al.*, "Accurate genetic detection of hepatitis c virus transmissions in outbreak settings," *The Journal of infectious diseases*, vol. 213, no. 6, pp. 957–965, 2016.

[24] S. J. Gould, *Wonderful life: the Burgess Shale and the nature of history*. WW Norton & Company, 1990.

[25] M. Lässig, V. Mustonen, and A. M. Walczak, "Predicting evolution," *Nature ecology & evolution*, vol. 1, no. 3, pp. 1–9, 2017.

[26] S. Seo, M. J. Silverberg, L. B. Hurley, J. Ready, V. Saxena, D. Witt, C. B. Hare, J. H. Champsi, D. G. Korn, M. P. Pauly, *et al.*, "Prevalence of spontaneous clearance of hepatitis

c virus infection doubled from 1998 to 2017," *Clinical Gastroenterology and Hepatology*, vol. 18, no. 2, pp. 511–513, 2020.

[27] D. G. Bowen and C. M. Walker, "Adaptive immune responses in acute and chronic hepatitis c virus infection," *Nature*, vol. 436, no. 7053, pp. 946–952, 2005.

[28] A. C. Araujo, I. V. Astrakhantseva, H. A. Fields, and S. Kamili, "Distinguishing acute from chronic hepatitis c virus (hcv) infection based on antibody reactivities to specific hcv structural and nonstructural proteins," *Journal of clinical microbiology*, vol. 49, no. 1, pp. 54–57, 2011.

[29] T. Tsertsvadze, L. Sharvadze, N. Chkhartishvili, L. Dzigua, M. Karchava, L. Gatserelia, A. Abutidze, and K. E. Nelson, "The natural history of recent hepatitis c virus infection among blood donors and injection drug users in the country of georgia," *Virology journal*, vol. 13, no. 1, p. 22, 2016.

[30] I. V. Astrakhantseva, D. S. Campo, A. Araujo, C.-G. Teo, Y. Khudyakov, and S. Kamili, "Differences in variability of hypervariable region 1 of hepatitis c virus (hcv) between acute and chronic stages of hcv infection," *In silico biology*, vol. 11, no. 5, pp. 163–173, 2011.

[31] V. Montoya, A. D. Olmstead, N. Z. Janjua, P. Tang, J. Grebely, D. Cook, P. Richard Harrigan, and M. Krajden, "Differentiation of acute from chronic hepatitis c virus infection by nonstructural 5b deep sequencing: A population-level tool for incidence estimation," *Hepatology*, vol. 61, no. 6, pp. 1842–1850, 2015.

[32] M. I. Gismondi, J. M. D. Carrasco, P. Valva, P. D. Becker, C. A. Guzmán, R. H. Campos, and M. V. Preciado, "Dynamic changes in viral population structure and compartmentalization during chronic hepatitis c virus infection in children," *Virology*, vol. 447, no. 1, pp. 187–196, 2013.

[33] L. Lu, N. Tatsunori, C. Li, S. Waheed, F. Gao, and B. H. Robertson, "Hcv selection and hvr1 evolution in a chimpanzee chronically infected with hcv-1 over 12 years," *Hepatology Research*, vol. 38, no. 7, pp. 704–716, 2008.

[34] D. S. Campo, Z. Dimitrova, L. Yamasaki, P. Skums, D. T. Lau, G. Vaughan, J. C. Forbi, C.-G. Teo, and Y. Khudyakov, "Next-generation sequencing reveals large connected networks of intra-host hcv variants," *BMC genomics*, vol. 15, no. Suppl 5, p. S4, 2014.

[35] J. Lara, M. Teka, and Y. Khudyakov, "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*, vol. 18, no. 10, p. 880, 2017.

[36] P. Skums, Z. Dimitrova, D. S. Campo, G. Vaughan, L. Rossi, J. C. Forbi, J. Yokosawa, A. Zelikovsky, and Y. Khudyakov, "Efficient error correction for next-generation sequencing of viral amplicons," in *BMC bioinformatics*, vol. 13, p. S6, BioMed Central, 2012.

[37] R. C. Edgar, "Muscle: multiple sequence alignment with high accuracy and high throughput," *Nucleic acids research*, vol. 32, no. 5, pp. 1792–1797, 2004.

[38] K. Nguyen, X. Guo, and Y. Pan, *Multiple biological sequence alignment: scoring functions, algorithms and evaluation*. John Wiley & Sons, 2016.

[39] M. H. Powdrill, E. P. Tchesnokov, R. A. Kozak, R. S. Russell, R. Martin, E. S. Svarovskaia, H. Mo, R. D. Kouyos, and M. Götte, "Contribution of a mutational bias in hepatitis c virus replication to the genetic barrier in the development of drug resistance," *Proceedings of the National Academy of Sciences*, vol. 108, no. 51, pp. 20509–20513, 2011.

[40] N. Patterson, A. L. Price, and D. Reich, "Population structure and eigenanalysis," *PLoS genet*, vol. 2, no. 12, p. e190, 2006.

[41] L. L. Cavalli-Sforza, L. Cavalli-Sforza, P. Menozzi, and A. Piazza, *The history and geography of human genes*. Princeton university press, 1994.

[42] M. Li, P. Vitányi, *et al.*, *An introduction to Kolmogorov complexity and its applications*, vol. 3. Springer, 2008.

[43] A. Lempel and J. Ziv, "On the complexity of finite sequences," *IEEE Transactions on information theory*, vol. 22, no. 1, pp. 75–81, 1976.

[44] F. Kaspar and H. Schuster, "Easily calculable measure for the complexity of spatiotemporal patterns," *Physical Review A*, vol. 36, no. 2, p. 842, 1987.

[45] S. Schaper, I. G. Johnston, and A. A. Louis, "Epistasis can lead to fragmented neutral spaces and contingency in evolution," *Proceedings of the Royal Society of London B: Biological Sciences*, p. rspb20112183, 2011.

[46] E. Van Nimwegen, J. P. Crutchfield, and M. Huynen, "Neutral evolution of mutational robustness," *Proceedings of the National Academy of Sciences*, vol. 96, no. 17, pp. 9716–9720, 1999.

[47] L. Li, D. Alderson, J. C. Doyle, and W. Willinger, "Towards a theory of scale-free graphs: Definition, properties, and implications," *Internet Mathematics*, vol. 2, no. 4, pp. 431–523, 2005.

[48] D. Wodarz, "Hepatitis c virus dynamics and pathology: the role of ctl and antibody responses," *Journal of General Virology*, vol. 84, no. 7, pp. 1743–1750, 2003.

[49] L. Rong, H. Dahari, R. M. Ribeiro, and A. S. Perelson, "Rapid emergence of protease inhibitor resistance in hepatitis c virus," *Science translational medicine*, vol. 2, no. 30, pp. 30ra32–30ra32, 2010.

[50] H. Dahari, J. E. Layden-Almer, E. Kallwitz, R. M. Ribeiro, S. J. Cotler, T. J. Layden, and A. S. Perelson, "A mathematical model of hepatitis c virus dynamics in patients with high baseline viral loads or advanced liver disease," *Gastroenterology*, vol. 136, no. 4, pp. 1402–1409, 2009.

[51] P. B. I. Baykal, J. Lara, Y. Khudyakov, A. Zelikovsky, and P. Skums, "Quantitative differences between intra-host hcv populations from persons with recently established and persistent infections," *bioRxiv*, 2020.

[52] S. Duchêne, S. Y. Ho, and E. C. Holmes, "Declining transition/transversion ratios through

time reveal limitations to the accuracy of nucleotide substitution models," *BMC evolutionary biology*, vol. 15, no. 1, pp. 1–10, 2015.

[53] J. Lara, J. E. Tavis, M. J. Donlin, W. M. Lee, H.-J. Yuan, B. L. Pearlman, G. Vaughan, J. C. Forbi, G.-L. Xia, and Y. E. Khudyakov, "Coordinated evolution among hepatitis c virus genomic sites is coupled to host factors and resistance to interferon," *In silico biology*, vol. 11, no. 5, 6, pp. 213–224, 2011.

[54] J. L. Law, M. Logan, J. Wong, J. Kundu, D. Hockman, A. Landi, C. Chen, K. Crawford, M. Wininger, J. Johnson, *et al.*, "Role of the e2 hypervariable region (hvr1) in the immunogenicity of a recombinant hepatitis c virus vaccine," *Journal of virology*, vol. 92, no. 11, pp. e02141–17, 2018.

[55] R. Aurora, M. J. Donlin, N. A. Cannon, and J. E. Tavis, "Genome-wide hepatitis c virus amino acid covariance networks can predict response to antiviral therapy in humans," *The Journal of clinical investigation*, vol. 119, no. 1, pp. 225–236, 2009.

[56] J. Lara and Y. Khudyakov, "Epistatic connectivity among hcv genomic sites as a genetic marker of interferon resistance," *Antiviral therapy*, vol. 17, no. 7 Pt B, pp. 1471–5, 2012.

[57] C. Shen, P. Gupta, X. Xu, A. Sanyal, C. Rinaldo, E. Seaberg, J. B. Margolick, O. Martinez-Maza, and Y. Chen, "Transmission and evolution of hepatitis c virus in hcv seroconverters in hiv infected subjects," *Virology*, vol. 449, pp. 339–349, 2014.

[58] J. Raghwani, R. Rose, I. Sheridan, P. Lemey, M. A. Suchard, T. Santantonio, P. Farci, P. Klenerman, and O. G. Pybus, "Exceptional heterogeneity in viral evolutionary dynamics characterises chronic hepatitis c virus infection," *PLoS pathogens*, vol. 12, no. 9, p. e1005894, 2016.

[59] A. G. Longmire, S. Sims, I. Rytsareva, D. S. Campo, P. Skums, Z. Dimitrova, S. Ramachandran, M. Medrzycki, H. Thai, L. Ganova-Raeva, *et al.*, "Ghost: global hepatitis outbreak and surveillance technology," *BMC genomics*, vol. 18, no. 10, p. 916, 2017.

[60] D. S. Campo and Y. Khudyakov, "Machine learning can accelerate discovery and application of cyber-molecular cancer diagnostics," *Journal of medical artificial intelligence*, vol. 3, no. 7, 2020.

[61] A. R. Templeton, "The reality and importance of founder speciation in evolution," *BioEssays*, vol. 30, no. 5, pp. 470–479, 2008.

[62] H. Li, M. B. Stoddard, S. Wang, L. M. Blair, E. E. Giorgi, E. H. Parrish, G. H. Learn, P. Hraber, P. A. Goepfert, M. S. Saag, *et al.*, "Elucidation of hepatitis c virus transmission and early diversification by single genome sequencing," *PLoS Pathog*, vol. 8, no. 8, p. e1002880, 2012.

[63] M. Friedel, S. Nikolajewa, J. Sühnel, and T. Wilhelm, "Diprodb: a database for dinucleotide properties," *Nucleic acids research*, vol. 37, no. suppl_1, pp. D37–D40, 2009.

[64] D. Campo, Z. Dimitrova, R. J. Mitchell, J. Lara, and Y. Khudyakov, "Coordinated evolution of the hepatitis c virus," *Proceedings of the National Academy of Sciences*, vol. 105, no. 28, pp. 9685–9690, 2008.

[65] J. Lara, M. A. Purdy, and Y. E. Khudyakov, "Genetic host specificity of hepatitis e virus," *Infection, Genetics and Evolution*, vol. 24, pp. 127–139, 2014.

[66] H. Thai, D. S. Campo, J. Lara, Z. Dimitrova, S. Ramachandran, G. Xia, L. Ganova-Raeva, C.-G. Teo, A. Lok, and Y. Khudyakov, "Convergence and coevolution of hepatitis b virus drug resistance," *Nature communications*, vol. 3, p. 789, 2012.

[67] B. A. Palmer, Z. Dimitrova, P. Skums, O. Crosbie, E. Kenny-Walsh, and L. J. Fanning, "Analysis of the evolution and structure of a complex intrahost viral population in chronic hepatitis c virus mapped by ultradeep pyrosequencing," *Journal of virology*, vol. 88, no. 23, pp. 13709–13721, 2014.

[68] D. S. Campo, Z. Dimitrova, J. Yokosawa, D. Hoang, N. O. Perez, S. Ramachandran, and

Y. Khudyakov, “Hepatitis c virus antigenic convergence,” *Scientific reports*, vol. 2, p. 267, 2012.

[69] P. Domingo-Calap, E. Segredo-Otero, M. Durán-Moreno, and R. Sanjuán, “Social evolution of innate immunity evasion in a virus,” *Nature microbiology*, vol. 4, no. 6, pp. 1006–1013, 2019.

[70] L. Villandre, D. A. Stephens, A. Labbe, H. F. Günthard, R. Kouyos, T. Stadler, and S. H. C. Study, “Assessment of overlap of phylogenetic transmission clusters and communities in simple sexual contact networks: applications to hiv-1,” *PloS one*, vol. 11, no. 2, p. e0148459, 2016.

[71] F. Chung and L. Lu, “The average distance in a random graph with given expected degrees,” *Internet Mathematics*, vol. 1, no. 1, pp. 91–113, 2004.

[72] F. Chung, L. Lu, and V. Vu, “The spectra of random graphs with given expected degrees,” *Internet Mathematics*, vol. 1, no. 3, pp. 257–275, 2004.

[73] M. Müller-Linow, C. C. Hilgetag, and M.-T. Hütt, “Organization of excitable dynamics in hierarchical biological networks,” *PLoS Comput Biol*, vol. 4, no. 9, p. e1000190, 2008.

[74] M. T. Hajiaghayi, M. Mahdian, and V. S. Mirrokni, “The facility location problem with general cost functions,” *Networks: An International Journal*, vol. 42, no. 1, pp. 42–47, 2003.

[75] F. Liljeros, C. R. Edling, L. A. N. Amaral, H. E. Stanley, and Y. Åberg, “The web of human sexual contacts,” *Nature*, vol. 411, no. 6840, pp. 907–908, 2001.

[76] J. O. Wertheim, A. J. L. Brown, N. L. Hepler, S. R. Mehta, D. D. Richman, D. M. Smith, and S. L. K. Pond, “The global transmission network of hiv-1,” *Journal of Infectious Diseases*, vol. 209, no. 2, pp. 304–313, 2014.

[77] G. J. Hughes, E. Fearnhill, D. Dunn, S. J. Lycett, A. Rambaut, A. J. L. Brown, *et al.*, "Molecular phylodynamics of the heterosexual hiv epidemic in the united kingdom," *PLoS pathogens*, vol. 5, no. 9, p. e1000590, 2009.

[78] C. M. Romano, I. M. G. de Carvalho-Mello, L. F. Jamal, F. L. de Melo, A. Iamarino, M. Motoki, J. R. R. Pinho, E. C. Holmes, P. M. de Andrade Zanotto, V. Consortium, *et al.*, "Social networks shape the transmission dynamics of hepatitis c virus," *PLoS One*, vol. 5, no. 6, p. e11170, 2010.

[79] C. Huang, Y. Wang, X. Li, L. Ren, J. Zhao, Y. Hu, L. Zhang, G. Fan, J. Xu, X. Gu, *et al.*, "Clinical features of patients infected with 2019 novel coronavirus in wuhan, china," *The Lancet*, vol. 395, no. 10223, pp. 497–506, 2020.

[80] N. Zhu, D. Zhang, W. Wang, X. Li, B. Yang, J. Song, X. Zhao, B. Huang, W. Shi, R. Lu, P. Niu, F. Zhan, X. Ma, D. Wang, W. Xu, G. Wu, G. F. Gao, W. Tan, and China Novel Coronavirus Investigating and Research Team, "A novel coronavirus from patients with pneumonia in china, 2019," *N. Engl. J. Med.*, Jan. 2020.

[81] V. J. Munster, M. Koopmans, N. van Doremalen, D. van Riel, and E. de Wit, "A novel coronavirus emerging in china - key questions for impact assessment," *N. Engl. J. Med.*, Jan. 2020.

[82] L.-L. Ren, Y.-M. Wang, Z.-Q. Wu, Z.-C. Xiang, L. Guo, T. Xu, Y.-Z. Jiang, Y. Xiong, Y.-J. Li, H. Li, G.-H. Fan, X.-Y. Gu, Y. Xiao, H. Gao, J.-Y. Xu, F. Yang, X.-M. Wang, C. Wu, L. Chen, Y.-W. Liu, B. Liu, J. Yang, X.-R. Wang, J. Dong, L. Li, C.-L. Huang, J.-P. Zhao, Y. Hu, Z.-S. Cheng, L.-L. Liu, Z.-H. Qian, C. Qin, Q. Jin, B. Cao, and J.-W. Wang, "Identification of a novel coronavirus causing severe pneumonia in human: a descriptive study," *Chin. Med. J.*, Jan. 2020.

[83] Y. Chen, Q. Liu, and D. Guo, "Emerging coronaviruses: genome structure, replication, and pathogenesis," *Journal of medical virology*, vol. 92, no. 4, pp. 418–423, 2020.

[84] C. Rothe, M. Schunk, P. Sothmann, G. Bretzel, G. Froeschl, C. Wallrauch, T. Zimmer, V. Thiel, C. Janke, W. Guggemos, M. Seilmaier, C. Drosten, P. Vollmar, K. Zwirglmaier, S. Zange, R. Wölfel, and M. Hoelscher, "Transmission of 2019-nCoV infection from an asymptomatic contact in germany," *N. Engl. J. Med.*, Jan. 2020.

[85] L. T. Phan, T. V. Nguyen, Q. C. Luong, T. V. Nguyen, H. T. Nguyen, H. Q. Le, T. T. Nguyen, T. M. Cao, and Q. D. Pham, "Importation and Human-to-Human transmission of a novel coronavirus in vietnam," *N. Engl. J. Med.*, Jan. 2020.

[86] L. J. Walker and COVID-19 National Incident Room Surveillance Team, "COVID-19, australia: Epidemiology report 2: Reporting week ending 19:00 AEDT 8 february 2020," 2020.

[87] Chantal B E, E. K. Broberg, B. Haagmans, A. Meijer, V. M. Corman, A. Papa, R. Charrel, C. Drosten, M. Koopmans, K. Leitmeyer, and on behalf of EVD-LabNet and ERLI-Net, "Laboratory readiness and response for novel coronavirus (2019-nCoV) in expert laboratories in 30 EU/EEA countries, january 2020," *Eurosurveillance*, vol. 25, p. 2000082, Feb. 2020.

[88] Y. Chen, Q. Liu, and D. Guo, "Emerging coronaviruses: genome structure, replication, and pathogenesis," *J. Med. Virol.*, Jan. 2020.

[89] X. Liao, B. Wang, and Y. Kang, "Novel coronavirus infection during the 2019–2020 epidemic: preparing intensive care units—the experience in sichuan province, china," *Intensive Care Med.*, pp. 1–4, Feb. 2020.

[90] C. Huang, Y. Wang, X. Li, L. Ren, J. Zhao, Y. Hu, L. Zhang, G. Fan, J. Xu, X. Gu, Z. Cheng, T. Yu, J. Xia, Y. Wei, W. Wu, X. Xie, W. Yin, H. Li, M. Liu, Y. Xiao, H. Gao, L. Guo, J. Xie, G. Wang, R. Jiang, Z. Gao, Q. Jin, J. Wang, and B. Cao, "Clinical features of patients infected with 2019 novel coronavirus in wuhan, china," *Lancet*, Jan. 2020.

[91] J. F.-W. Chan, S. Yuan, K.-H. Kok, K. K.-W. To, H. Chu, J. Yang, F. Xing, J. Liu, C. C.-Y. Yip, R. W.-S. Poon, H.-W. Tsoi, S. K.-F. Lo, K.-H. Chan, V. K.-M. Poon, W.-M. Chan, J. D. Ip, J.-P. Cai, V. C.-C. Cheng, H. Chen, C. K.-M. Hui, and K.-Y. Yuen,

"A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: a study of a family cluster," *Lancet*, Jan. 2020.

[92] J. T. Wu, K. Leung, and G. M. Leung, "Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in wuhan, china: a modelling study," 2020.

[93] M. Ragonnet-Cronin, Y. W. Hu, S. R. Morris, Z. Sheng, K. Poortinga, and J. O. Wertheim, "Hiv transmission networks among transgender women in los angeles county, ca, usa: a phylogenetic analysis of surveillance data," *The Lancet HIV*, vol. 6, no. 3, pp. e164–e172, 2019.

[94] K. Jahn, J. Kuipers, and N. Beerenwinkel, "Tree inference for single-cell data," *Genome biology*, vol. 17, no. 1, p. 86, 2016.

[95] P. Forster, L. Forster, C. Renfrew, and M. Forster, "Phylogenetic network analysis of sars-cov-2 genomes," *Proceedings of the National Academy of Sciences*, 2020.

[96] Y. Shu and J. McCauley, "Gisaid: Global initiative on sharing all influenza data–from vision to reality," *Eurosurveillance*, vol. 22, no. 13, 2017.

[97] F. Wu, S. Zhao, B. Yu, Y.-M. Chen, W. Wang, Y. Hu, Z.-G. Song, Z.-W. Tao, J.-H. Tian, Y.-Y. Pei, *et al.*, "Complete genome characterisation of a novel coronavirus associated with severe human respiratory disease in wuhan, china," *bioRxiv*, 2020.

[98] L. v. d. Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. Nov, pp. 2579–2605, 2008.

[99] R. Tibshirani, G. Walther, and T. Hastie, "Estimating the number of clusters in a data set via the gap statistic," *J. R. Stat. Soc. Series B Stat. Methodol.*, vol. 63, pp. 411–423, May 2001.

[100] D. Gusfield, "Algorithms on stings, trees, and sequences: Computer science and computational biology," *Acm Sigact News*, vol. 28, no. 4, pp. 41–60, 1997.

[101] J. H. Camin and R. R. Sokal, "A method for deducing branching sequences in phylogeny," *Evolution*, pp. 311–326, 1965.

[102] D. Gusfield, *Integer linear programming in computational and systems biology: an entry-level text and course*. Cambridge University Press, 2019.

[103] C. Bron and J. Kerbosch, "Algorithm 457: finding all cliques of an undirected graph," *Communications of the ACM*, vol. 16, no. 9, pp. 575–577, 1973.

[104] J. Edmonds, "Optimum branchings," *Journal of Research of the national Bureau of Standards B*, vol. 71, no. 4, pp. 233–240, 1967.

[105] M. Bastian, S. Heymann, and M. Jacomy, "Gephi: an open source software for exploring and manipulating networks," in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 3, 2009.

[106] D. T. Hamilton, M. S. Handcock, and M. Morris, "Degree distributions in sexual networks: a framework for evaluating evidence," *Sexually transmitted diseases*, vol. 35, no. 1, p. 30, 2008.

[107] T. Bedford, R. Neher, J. Hadfield, E. Hodcroft, M. Ilcisin, and N. Muller, "Genomic analysis of ncov spread. situation report 2020-01-23.," tech. rep., 2020.

[108] M. Giovanetti, S. Angeletti, D. Benvenuto, and M. Ciccozzi, "A doubt of multiple introduction of sars-cov-2 in italy: a preliminary overview," *Journal of medical virology*, vol. 92, no. 9, pp. 1634–1636, 2020.

[109] M. Giovanetti, D. Benvenuto, S. Angeletti, and M. Ciccozzi, "The first two cases of 2019-ncov in italy: Where they come from?," *Journal of medical virology*, vol. 92, no. 5, pp. 518–521, 2020.

[110] N. Chen, M. Zhou, X. Dong, J. Qu, F. Gong, Y. Han, Y. Qiu, J. Wang, Y. Liu, Y. Wei, *et al.*, "Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus

pneumonia in wuhan, china: a descriptive study," *The lancet*, vol. 395, no. 10223, pp. 507–513, 2020.

[111] D. S. Hui, E. I Azhar, T. A. Madani, F. Ntoumi, R. Kock, O. Dar, G. Ippolito, T. D. Mchugh, Z. A. Memish, C. Drosten, *et al.*, "The continuing 2019-ncov epidemic threat of novel coronaviruses to global health—the latest 2019 novel coronavirus outbreak in wuhan, china," *International Journal of Infectious Diseases*, vol. 91, pp. 264–266, 2020.

[112] R. M. May and A. L. Lloyd, "Infection dynamics on scale-free networks," *Physical Review E*, vol. 64, no. 6, p. 066112, 2001.

[113] R. Pastor-Satorras and A. Vespignani, "Epidemic spreading in scale-free networks," *Physical review letters*, vol. 86, no. 14, p. 3200, 2001.

[114] J. Hadfield, C. Megill, S. M. Bell, J. Huddleston, B. Potter, C. Callender, P. Sagulenko, T. Bedford, and R. A. Neher, "Nextstrain: real-time tracking of pathogen evolution," *Bioinformatics*, vol. 34, no. 23, pp. 4121–4123, 2018.

[115] S. L. Kosakovsky Pond, S. Weaver, A. J. Leigh Brown, and J. O. Wertheim, "Hiv-trace (transmission cluster engine): a tool for large scale molecular epidemiology of hiv-1 and other rapidly evolving pathogens," *Molecular biology and evolution*, vol. 35, no. 7, pp. 1812–1819, 2018.

[116] M. Gwinn, D. MacCannell, and G. L. Armstrong, "Next-generation sequencing of infectious pathogens," *Jama*, vol. 321, no. 9, pp. 893–894, 2019.

[117] G. L. Armstrong, D. R. MacCannell, J. Taylor, H. A. Carleton, E. B. Neuhaus, R. S. Bradbury, J. E. Posey, and M. Gwinn, "Pathogen genomics in public health," *New England Journal of Medicine*, vol. 381, no. 26, pp. 2569–2580, 2019.

[118] K. Schwarze, J. Buchanan, J. M. Fermont, H. Dreau, M. W. Tilley, J. M. Taylor, P. Antoniou, S. J. Knight, C. Camps, M. M. Pentony, *et al.*, "The complete costs of genome

sequencing: a microcosting study in cancer and rare diseases from a single center in the united kingdom," *Genetics in Medicine*, vol. 22, no. 1, pp. 85–94, 2020.

[119] P. Muir, S. Li, S. Lou, D. Wang, D. J. Spakowicz, L. Salichos, J. Zhang, G. M. Weinstock, F. Isaacs, J. Rozowsky, *et al.*, "The real cost of sequencing: scaling computation to keep pace with data generation," *Genome biology*, vol. 17, no. 1, p. 53, 2016.

APPENDICES

Appendix A

Table A.1 contains data used to generate Figure. 1.2 for clarity due to the presence of small fractions that are difficult to see on the full size image.

Table (A.1) Percentage of reads for different categories (identical, CG - IL: constant global and inconsistent local, IG: inconsistent global, MM: multi-mapped, IT2: inconsistent type 1, inconsistent type 2) between original data and replicate data (reverse complement and shuffled).

Data type	Tool	Identical	CG - IL	IG	MM	IT1	IT2
rev. comp.	Bowtie2	**78.706**	0.397	0.009	0.425	**19.746**	0.717
	BWA-MEM	95.470	0.069	2.352	0.915	0.641	0.553
	BWA-MEM2	95.470	0.069	2.352	0.915	0.641	0.553
	HISAT2	94.162	0.048	0.232	0.642	4.541	0.375
	Minimap2	82.958	0.003	**5.681**	**7.036**	4.252	0.070
	NextGenMap	99.643	0.000	0.000	0.000	0.178	0.179
	SMALT	92.164	0.002	0.548	0.962	5.178	1.146
	SNAP	93.167	0.000	1.990	1.421	2.905	0.517
	Subread	97.544	0.002	0.083	0.194	1.058	1.118
shuffled	Bowtie2	100.000	0.000	0.000	0.000	0.000	0.000
	BWA-MEM	96.498	0.000	**3.096**	0.379	0.014	0.013
	BWA-MEM2	96.499	0.000	**3.095**	0.378	0.014	0.013
	HISAT2	100.000	0.000	0.000	0.000	0.000	0.000
	Minimap2	100.000	0.000	0.000	0.000	0.000	0.000
	NextGenMap	99.961	0.000	0.002	0.007	0.016	0.013
	SMALT	100.000	0.000	0.000	0.000	0.000	0.000
	SNAP	100.000	0.000	0.000	0.000	0.000	0.000
	Subread	99.714	0.064	0.013	0.098	0.055	0.056

Appendix B

Table B.1 describes all statistical results of analyzed features, their dynamics in time and interpretation in part 2: Quantitative difference between intra-host HCV populations from persons with recently established and persistent infection.

Table (B.1) Summary of features (P. mean refers to persistent mean and R. mean refers to recent mean).

Features	**p-value**	**P.mean**	**R.mean**	**Dynamics in time**
1. Mean distance	$1.41E-24$	0.034	0.015	↑
2. Std distance	$4.10E-14$	0.019	0.010	↑
3. Conservation score	$5.11E-25$	0.422	0.188	↑
4. Mutation frequency	$3.07E-16$	0.023	0.010	↑
5. k-entropy $(k=10)$	$7.00E-23$	0.630	0.357	↑
6. Frequency entropy	$4.56E-06$	0.668	0.567	↑
7. SNV entropy	$1.14E-21$	0.084	0.043	↑
8. Transversion mutation	$1.06E-07$	0.061	0.032	↑
9. DN/DS	$5.39E-10$	0.713	1.330	↓
10. PCA complexity	$1.4E-03$	0.0053	0.0125	↓
11. Kolmogorov complexity	$1.55E-11$	0.041	0.052	↓
12. Robustnesss/Selection	$3.66E-15$	0.628	0.386	↑
13. s-metric	$1.93E-20$	0.001	0.044	↓
14. Clustering coefficient	$2.08E-13$	0.082	0.356	↓
15. ODE feature	$2.42E-06$	-0.270	0.224	N/A
16. Biochemical feature	$2.92E-35$	0.628	0.379	↑

Appendix C

Figures C.1 - C.10 represent transmission networks for the regions described in part6: Global transmission network of SARS-CoV-2 from outbreak to pandemic.

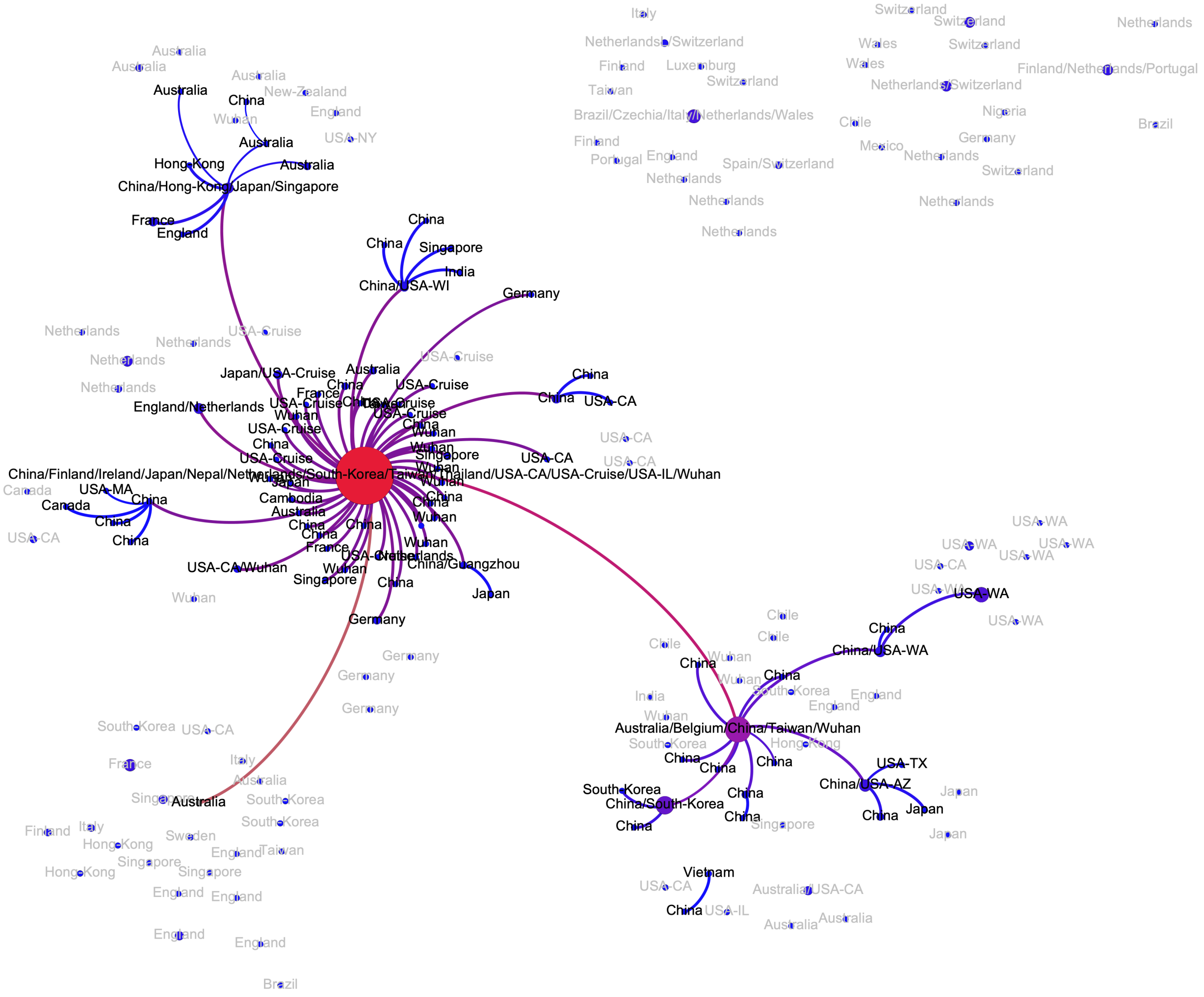


Figure (C.1) Transmission subnetwork of genomes sampled in China

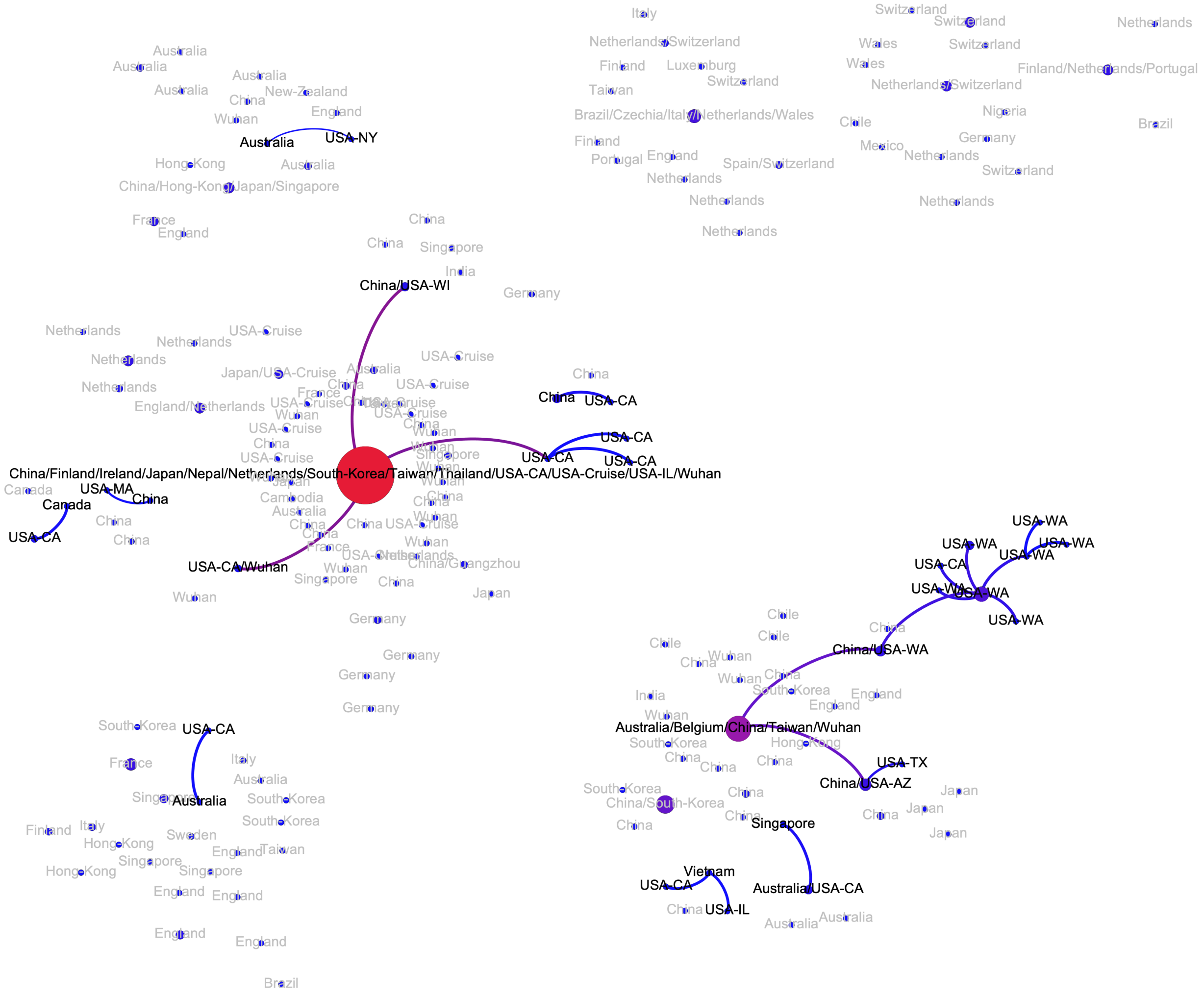


Figure (C.2) Transmission subnetwork of genomes sampled in USA

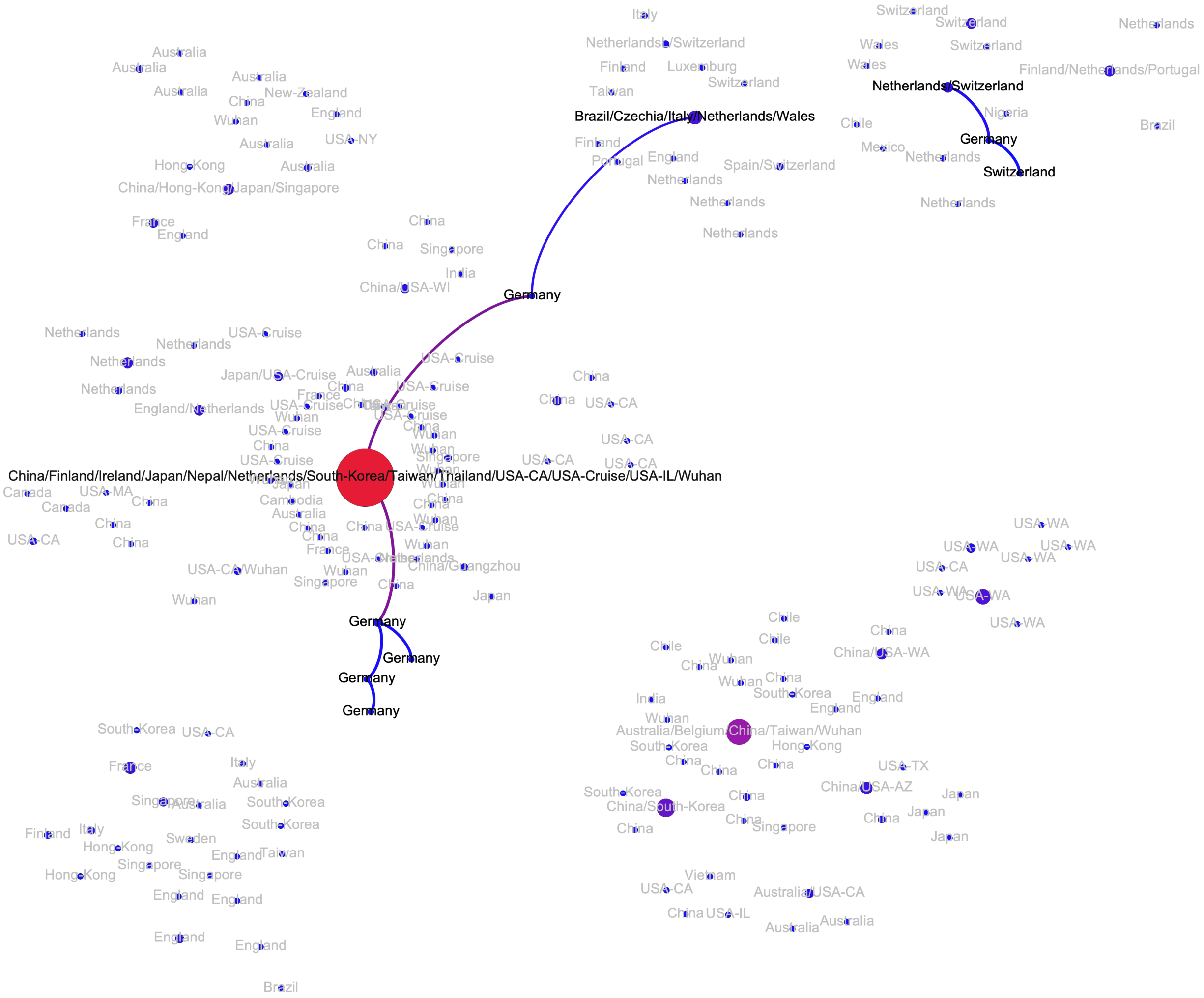


Figure (C.3) Transmission subnetwork of genomes sampled in Germany

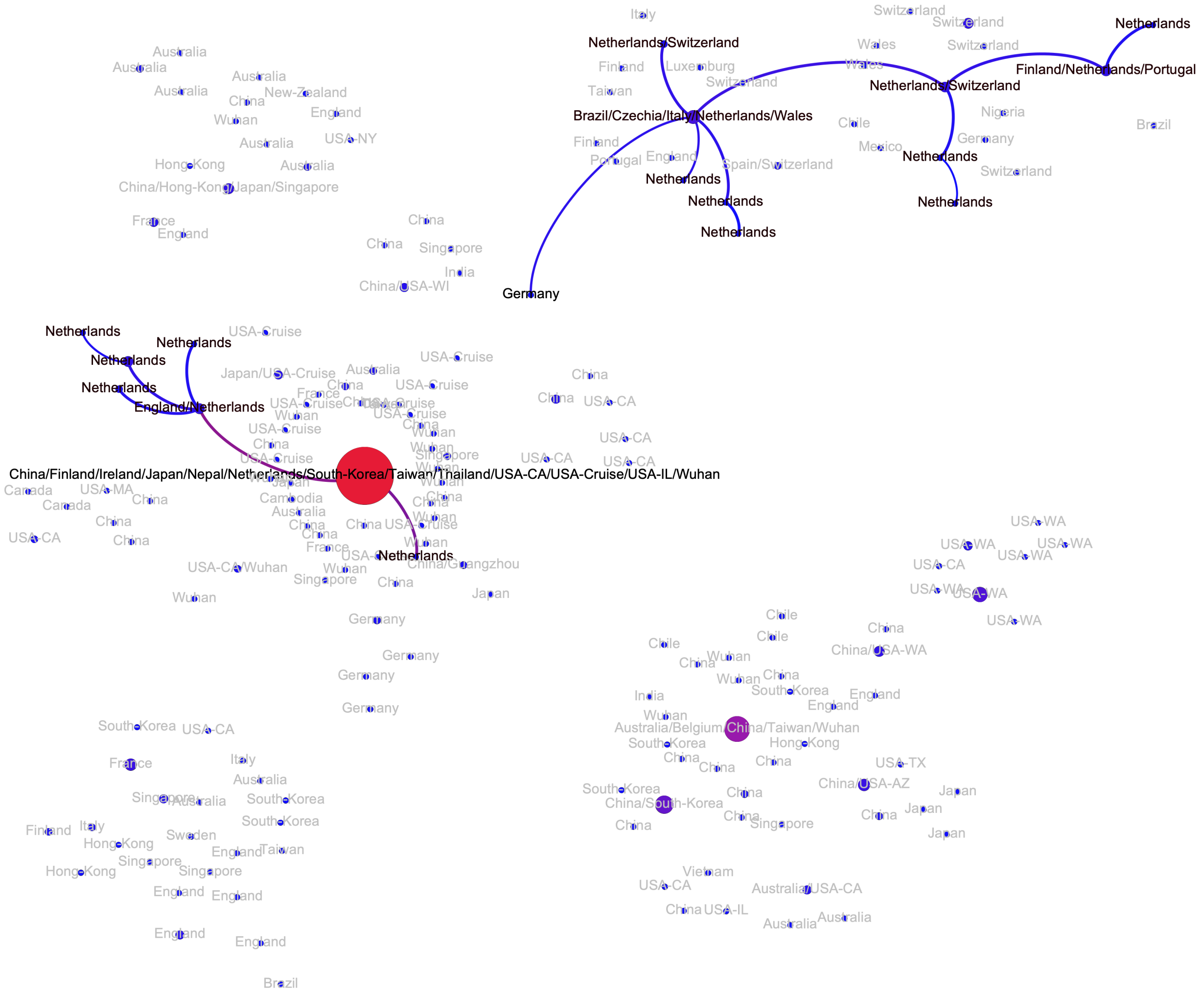


Figure (C.4) Transmission subnetwork of genomes sampled in Netherlands

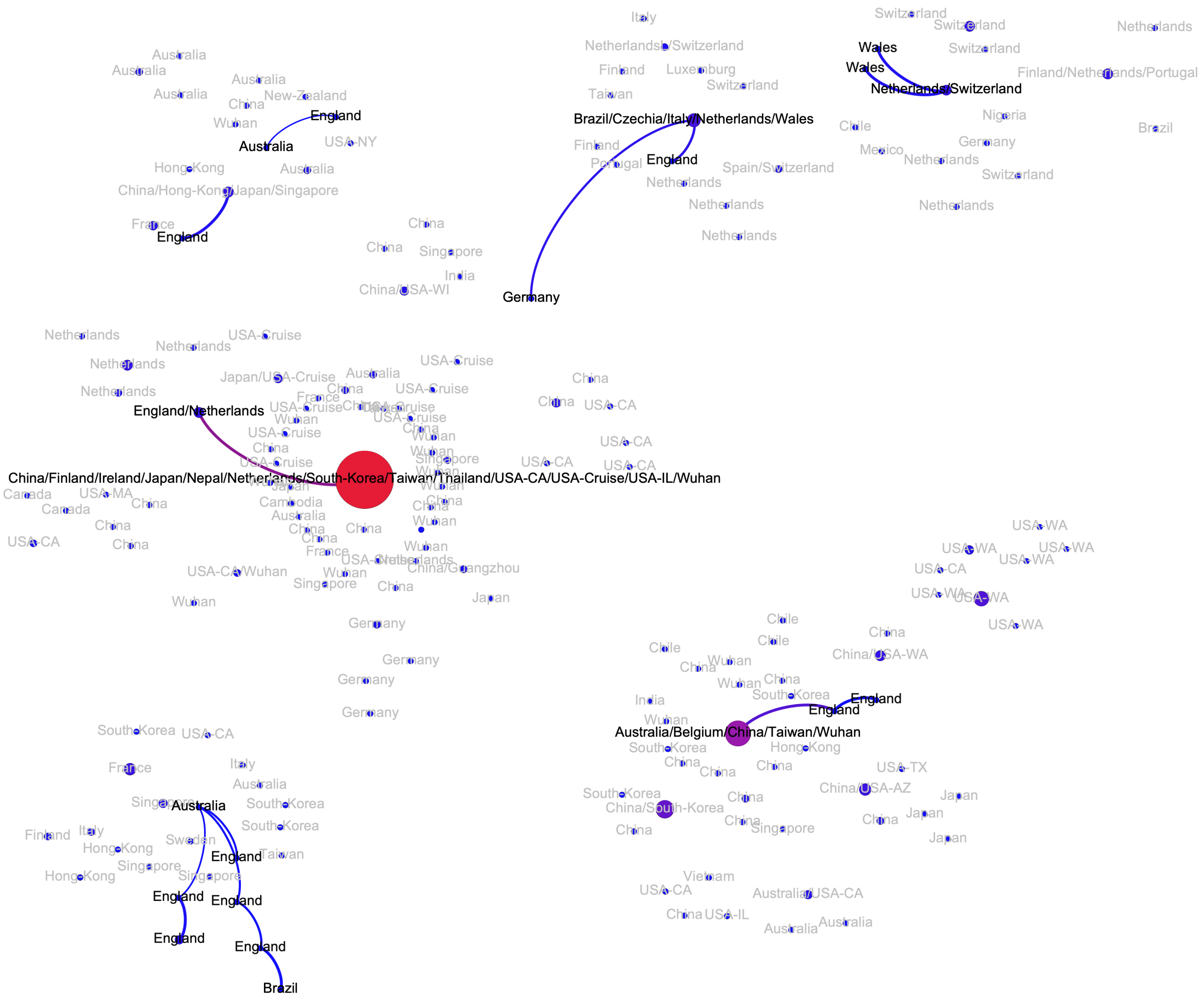


Figure (C.5) Transmission subnetwork of genomes sampled in UK

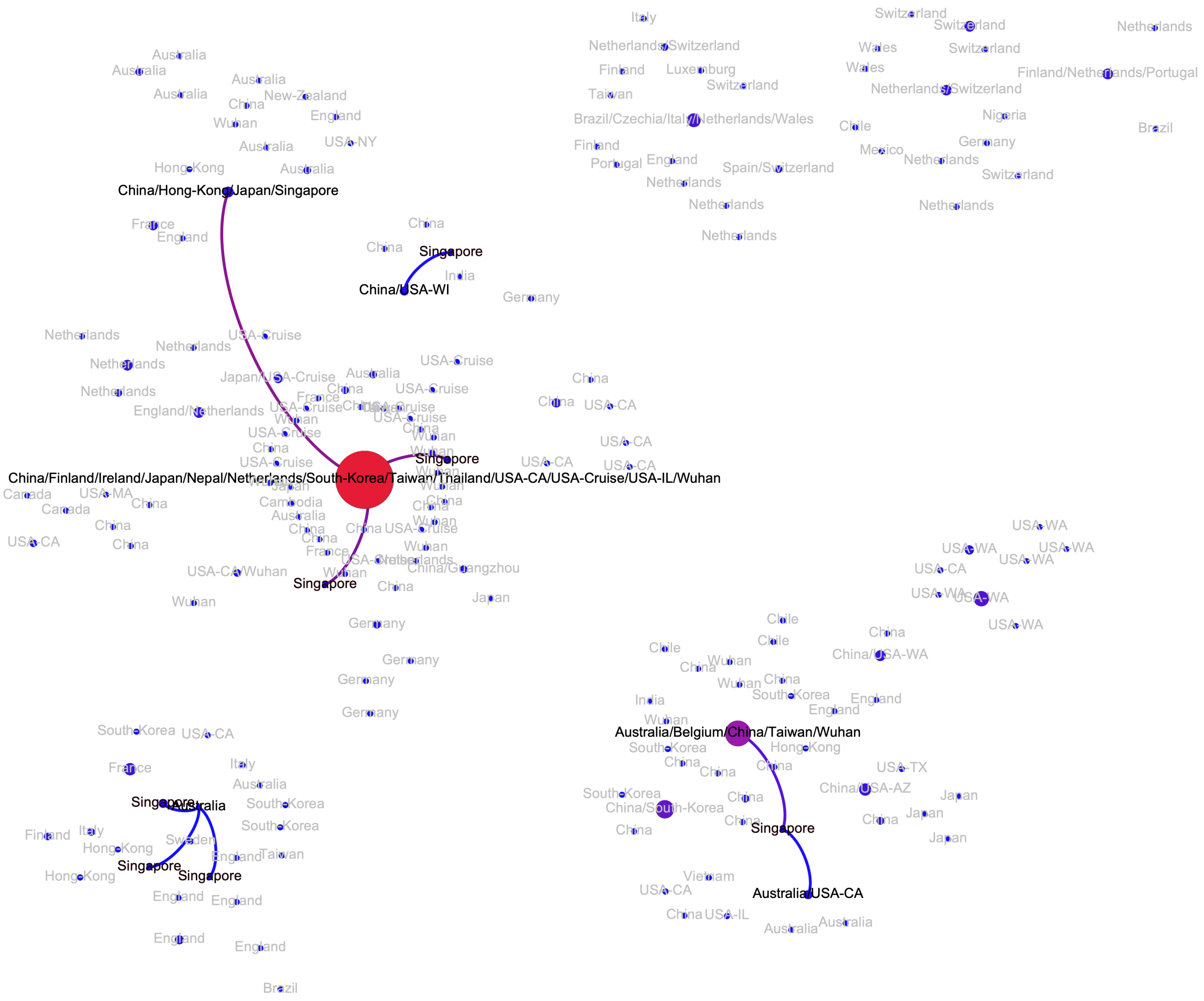


Figure (C.6) Transmission subnetwork of genomes sampled in Singapore

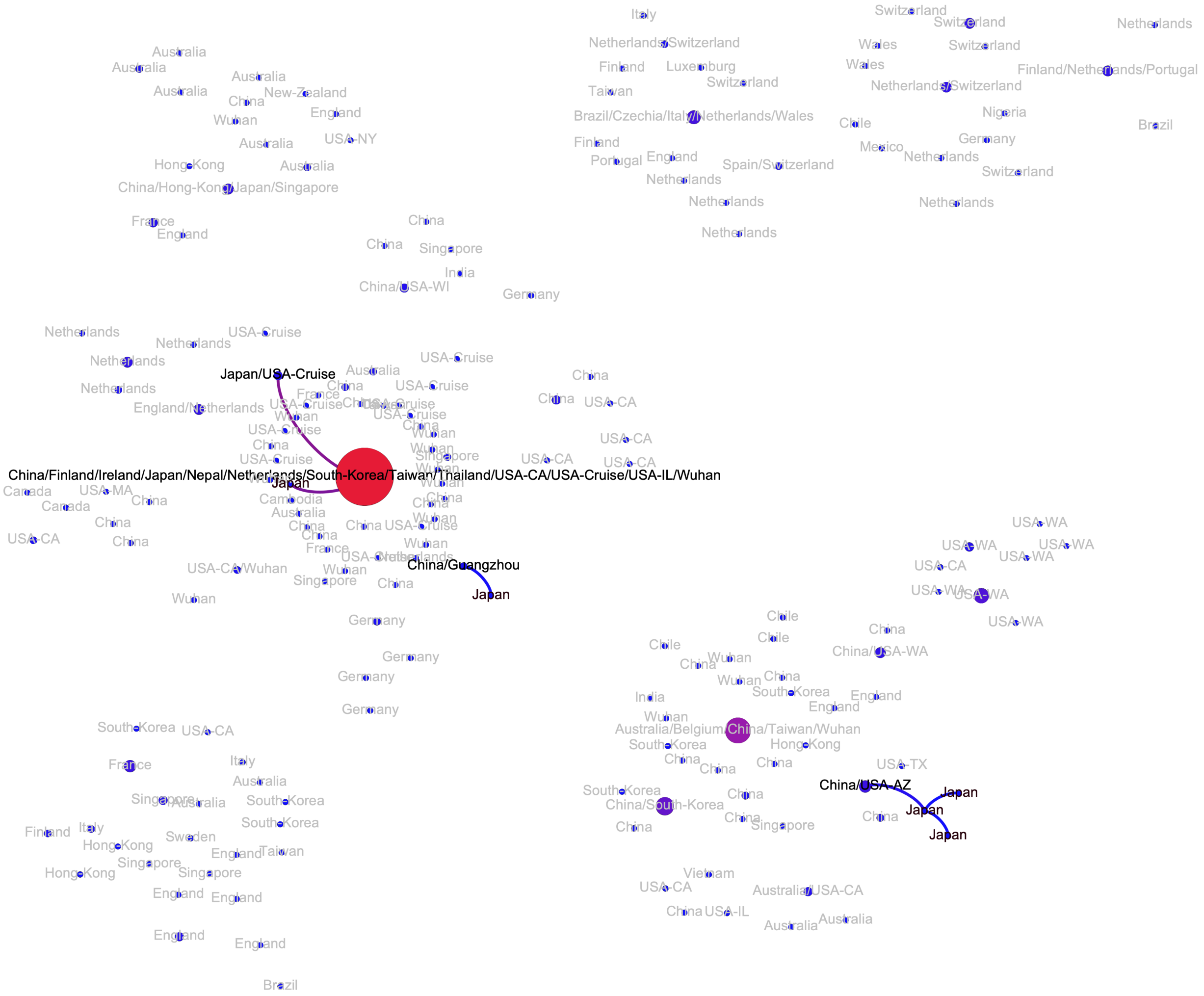


Figure (C.7) Transmission subnetwork of genomes sampled in Japan

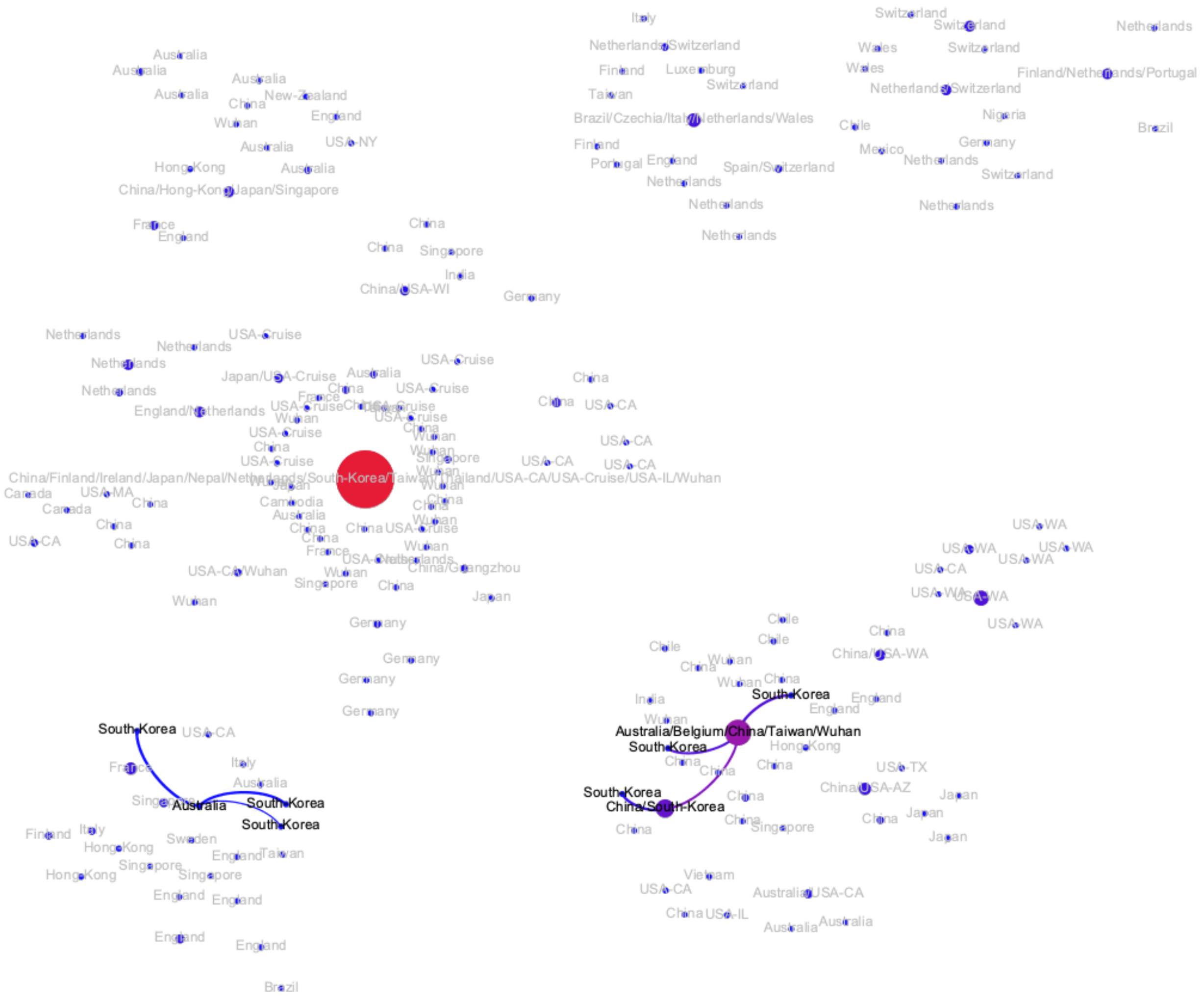


Figure (C.8) Transmission subnetwork of genomes sampled in South Korea

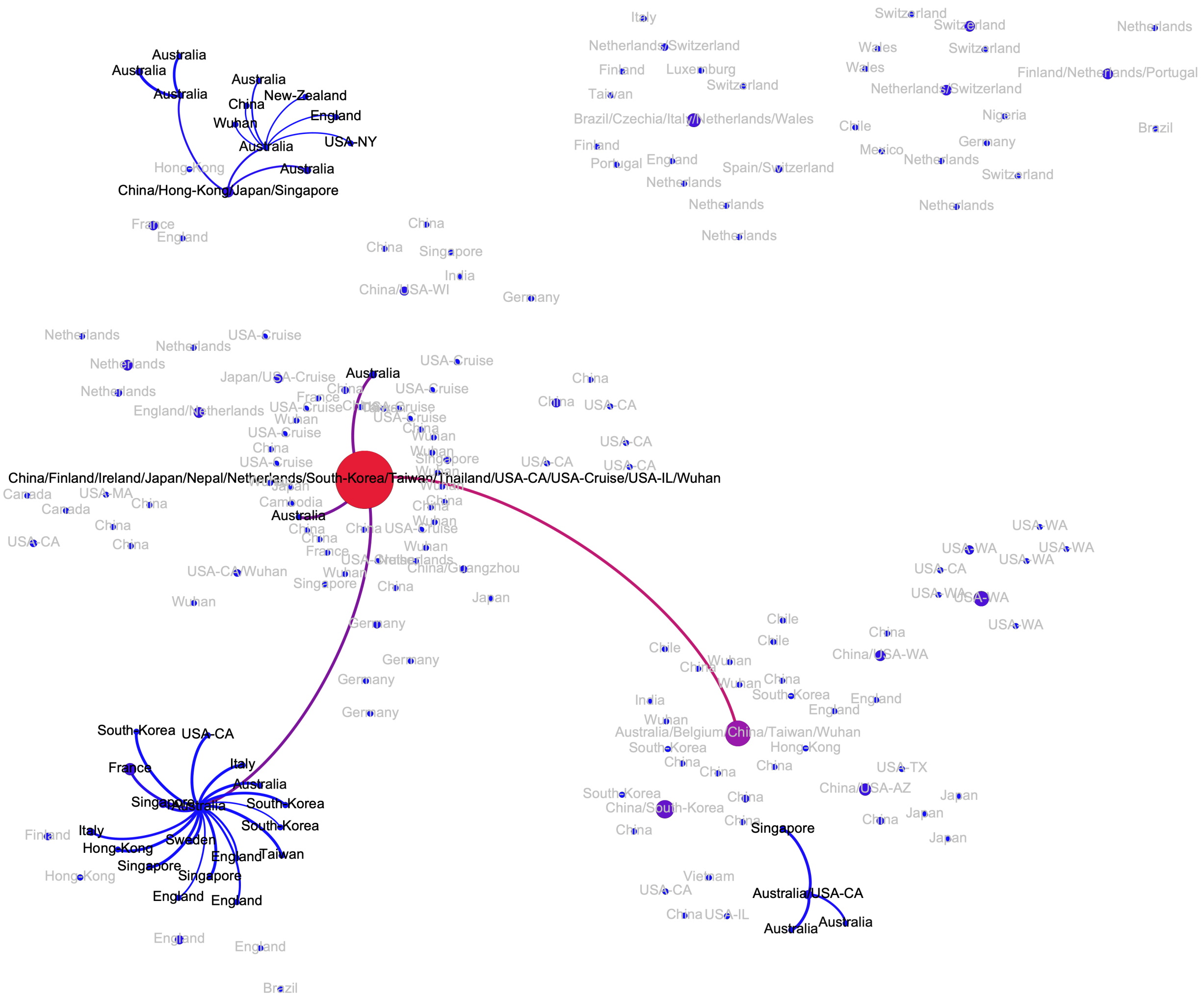


Figure (C.9) Transmission subnetwork of genomes sampled in Australia

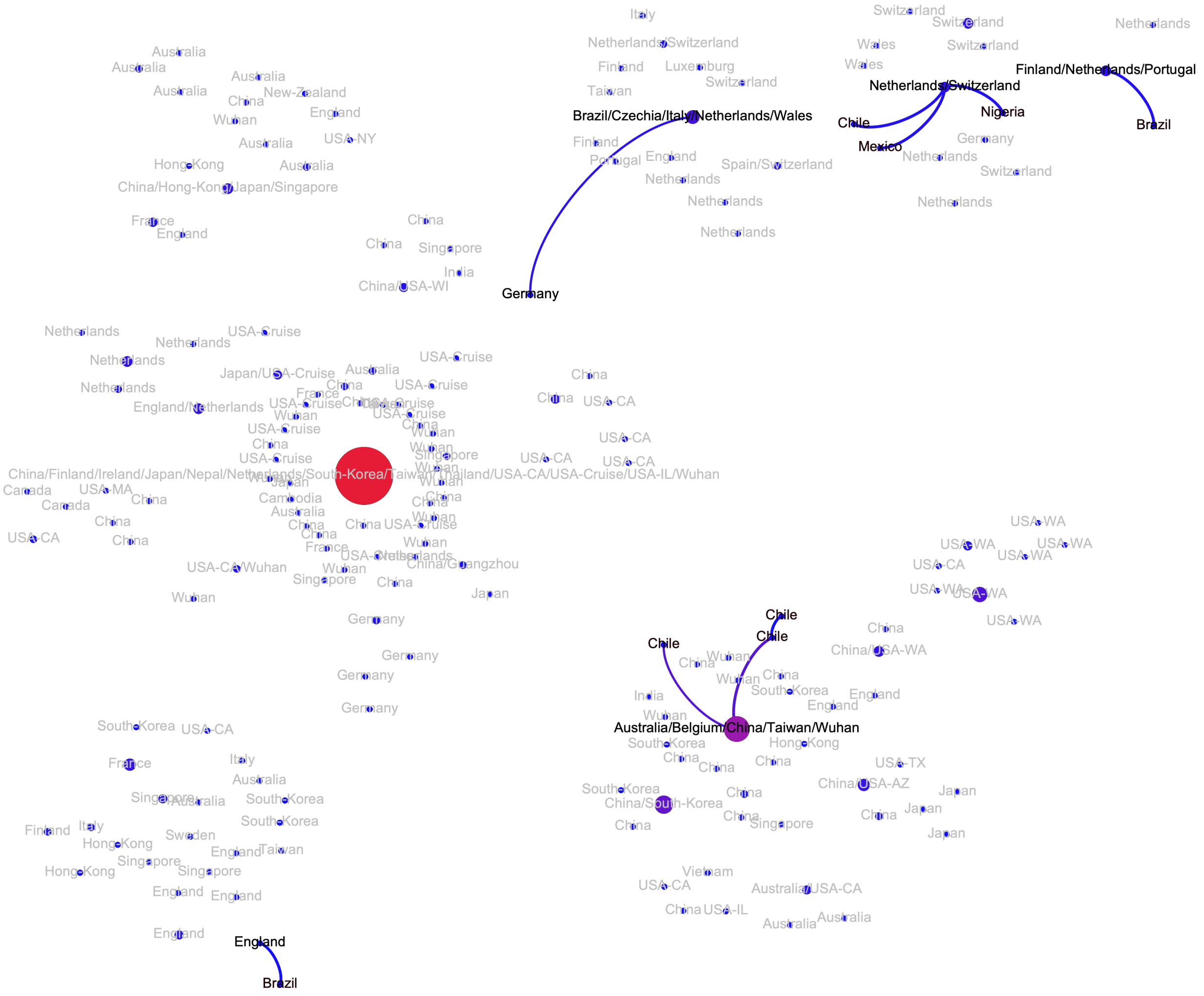


Figure (C.10) Transmission subnetwork of genomes sampled in Africa, Central and South America