Global transmission network of SARS-CoV-2: from outbreak to pandemic

Pavel Skums[1,3], Alexander Kirpich[2], Pelin Icer Baykal[1], Alex Zelikovsky[1], and Gerardo Chowell[2]

[1]Department of Computer Science, Georgia State University, Atlanta, GA, USA
[2]School of Public Health, Georgia State University, Atlanta, GA, USA
[3]Corresponding author. *Email: pskums@gsu.edu*

Abstract

Background. The COVID-19 pandemic caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is straining health systems around the world. Although the Chinese government implemented a number of severe restrictions on people's movement in an attempt to contain its local and international spread, the virus had already reached many areas of the world in part due to its potent transmissibility and the fact that a substantial fraction of infected individuals develop little or no symptoms at all. Following its emergence, the virus started to generate sustained transmission in neighboring countries in Asia, Western Europe, Australia, Canada and the United States, and finally in South America and Africa. As the virus continues its global spread, a clear and evidence-based understanding of properties and dynamics of the global transmission network of SARS-CoV-2 is essential to design and put in place efficient and globally coordinated interventions.
Methods. We employ molecular surveillance data of SARS-CoV-2 epidemics for inference and comprehensive analysis of its global transmission network before the pandemic declaration. Our goal was to characterize the spatial-temporal transmission pathways that led to the establishment of the pandemic. We exploited a network-based approach specifically tailored to emerging outbreak settings. Specifically, it traces the accumulation of mutations in viral genomic variants via mutation trees, which are then used to infer transmission networks, revealing an up-to-date picture of the spread of SARS-CoV-2 between and within countries and geographic regions.
Results and Conclusions. The analysis suggest multiple introductions of SARS-CoV-2 into the majority of world regions by means of heterogeneous transmission pathways. The transmission network is scale-free, with a few genomic variants responsible for the majority of possible transmissions. The network structure is in line with the available temporal information represented by sample collection times and suggest the expected sampling time difference of few days between potential transmission pairs. The inferred network structural properties, transmission clusters and pathways and virus introduction routes emphasize the extent of the global epidemiological linkage and demonstrate the importance of internationally coordinated public health measures.
Keywords: SARS-CoV-2; molecular surveillance; transmission network; mutation tree

1 Introduction

The COVID-19 pandemic due to the SARS-CoV-2 virus that emerged out of the city of Wuhan, Hubei Province in China in December, 2019 [15, 31, 71, 46, 41, 66, 53, 16], is now straining or overwhelming health care systems around the world [1]. As of March 2020, hundreds of new confirmed cases have been reported daily in multiple countries of every continent

[54, 51, 59, 35, 30] [13, 17, 39] while the global death toll has passed 13,000. To combat the spread of the virus, in the absence of a vaccine or specific treatments, an increasing number of nations are putting in place social distancing interventions ranging from school closures, quarantine orders on segments of the population, and banning large public gatherings. 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 [31, 12, 49, 46, 37] as well as long-range transmission events (country-to-country) [65]. To characterize the origin, geographic extent and epidemiological parameters of the epidemic, phylogenetics and phylodynamics inference tools have proved useful during this and past epidemic emergencies [68, 69, 38, 61, 10, 63, 11, 10, 5, 48, 70, 42, 60, 67, 19]. However, previous studies indicate that methods based on genetic networks are more powerful and efficient to identify transmission patterns and ascertain transmission links compared to methods based on binary phylogenies [62].

In this study, we used a network-based approach to analyze the global SARS-CoV-2 transmission patterns by constructing the transmission network based on genomic compositions of viral sequences sampled around the world and before the pandemic declaration on March 11, 2020. Here we sought to characterize the transmission pathways that facilitated the virus to establish itself as a pandemic. Such analyses rely on viral genomic data that continues to accumulate in near real time as next-generation sequencing technologies are now more widely used. The richness and accessibility of the genomic data distinguishes this SARS-CoV-2 pandemic from previous large-scale epidemics or pandemics including the 2009/AH1N1 influenza pandemic and the 2003 SARS outbreaks. This provides an unprecedented opportunity to gain insights into the epidemiological and evolutionary dynamics of this emerging virus spreading in an essentially naive population. 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. This indicates that the available sequencing data cover a significant part of the evolutionary space explored from the onset of the epidemics. In this setting, viral evolution could be accurately modeled, reconstructed and visualized by genetic networks models [8, 52], which have been successfully used for the analysis of different viral epidemics [52, 62, 9]. However, in emerging outbreak settings, when all circulating viral genomes are relatively close to each other, methods that allow the analysis of viral population structures at finer resolution than provided by state-of-the-art network models are needed. We achieve such resolution by using mutation trees [34] (related to character-based phylogenies [24]) that keep track of the accumulation of mutations in viral populations. Moreover, as the SARS-CoV-2 population contain relatively few 4-gamete rule violations, it is feasible to efficiently generate and analyze all plausible mutation trees. These trees, in turn, allow for accurate inferences of transmission networks.

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 epidemiological 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.

2 Methods

2.1 Data preprocessing.

We obtained the genomics data and associated metadata for this study from the Global Initiative on Sharing All Influenza Data (GISAID) database [56] that hosts genetic datasets for SARS-CoV-2, which have been self-reported from multiple sources. The sequences identified by GISAID as low-quality have been removed from consideration. The reference genome was identified and taken from the literature [64]. It coincides with the most prevalent sequence from GISAID sampled from Wuhan and a number of other locations. The remaining sequences were aligned to the consensus using MUSCLE [20] and trimmed to the same length, yielding n=319 aligned sequences of length 29772 base pairs (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. Next, 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.

2.2 Dimensionality reduction and clustering.

Viral sequences were embedded into the 2-dimensional space using T-distributed Stochastic Neighbor Embedding (t-SNE) [43] based on pairwise hamming distances between aligned sequences. The embedding points were clustered by k-means clustering, and the optimal number of clusters was estimated using the gap statistics [58].

2.3 Mutation tree reconstruction.

In a mutation tree,

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

Note that this tree does not have to be binary. In the perfect phylogeny model, which is the most widely used character-based phylogenetic model [25], each mutation can be represented by a single internal node implying that each mutation occurs only once. The perfect phylogeny 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. The sequencing data accumulated for SARS-CoV-2 includes several 4-gamete rule violations, thus implying repeated mutations in the same genomic positions. Therefore, instead of using the perfect phylogeny model, we fit the data to Camin-Sokal phylogenetic model, according to which repeated mutations are allowed, mutation losses are not allowed and each mutation could be acquired at most twice [7].

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 in a most parsimonious way using a graph G_{4g}, whose vertices are SNVs, and two vertices are adjacent whenever the corresponding pair of SNVs violates 4-gamete rule. If we remove the SNVs corresponding to vertices of G_{4g}, then the remaining mutation matrix M can be explained without any repeated mutations. Therefore, we are looking for the the minimum vertex cover of G_{4g}, i.e. the minimum number of vertices in G_{4g} whose removal destroys all edges in G_{4g}. The set of genomic positions corresponding to vertices in the minimum vertex cover of G_{4g} forms the most parsimonious set of mutations that should be repeated. We find all minimal vertex covers using Bron-Kerbosch algorithm [6] for maximum independent sets generation; this approach uses the fact that complements of maximum independent sets are exactly minimum vertex covers. Since the number of 4-gamete rule violations for SARS-CoV-2 data is relatively small, this method is fast and allows to significantly simplify the phylogeny reconstruction.

2) *Construct mutation trees.* For each minimum set $R = \{m_1, ..., m_k\}$ of potential repeated mutations found in the previous step, we construct a character-based 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 of Dan Gusfield [26]. 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

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

subject to constraints

$$D_{j,l,1,1} - X_{i,j} - X_{i,l} \geq -1, \tag{2}$$

$$D_{j,l,1,0} - X_{i,j} + X_{i,l} \geq 0, \tag{3}$$

$$D_{j,l,0,1} + X_{i,j} - X_{i,l} \geq 0, \tag{4}$$

$$D_{j,l,0,0} + X_{i,j} + X_{i,l} \geq 1, \tag{5}$$

$$D_{j,l,0,0} + D_{j,l,0,1} + D_{j,l,1,0} + D_{j,l,1,1} \leq 3 \tag{6}$$

$$\sum_{p \in C(j)} X_{i,p} \leq 1 \tag{7}$$

$$i = 1, ..., n; \quad j = 1 : m + k; \quad l = j + 1, ..., m + k \tag{8}$$

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. (Gurobi Optimization, LLC [47]). The trees were constructed for all potential sets of repeated mutations, and the tree with the best objective function was selected.

2.4 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.

To 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 e, its bootstrap probability p_e was calculated. The final consensus transmission tree was estimated as the maximum-weight spanning arborescence [18, 21] with respect to the weights p_e.

3 Results

3.1 Transmission clusters.

t-SNE plots combined with k-means clustering suggest that as of March 11, 2020 five distinct viral subpopulations have been circulating globally (Fig. 1). These subpopulations could be roughly classified as follows:

1) The original cluster(shown in 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 (shown in red) that includes sequences almost exclusively from European countries, 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 (shown in blue) that includes sequences from mainland China, South Korea, Hong Kong, Singapore, Vietnam, Australia, USA and

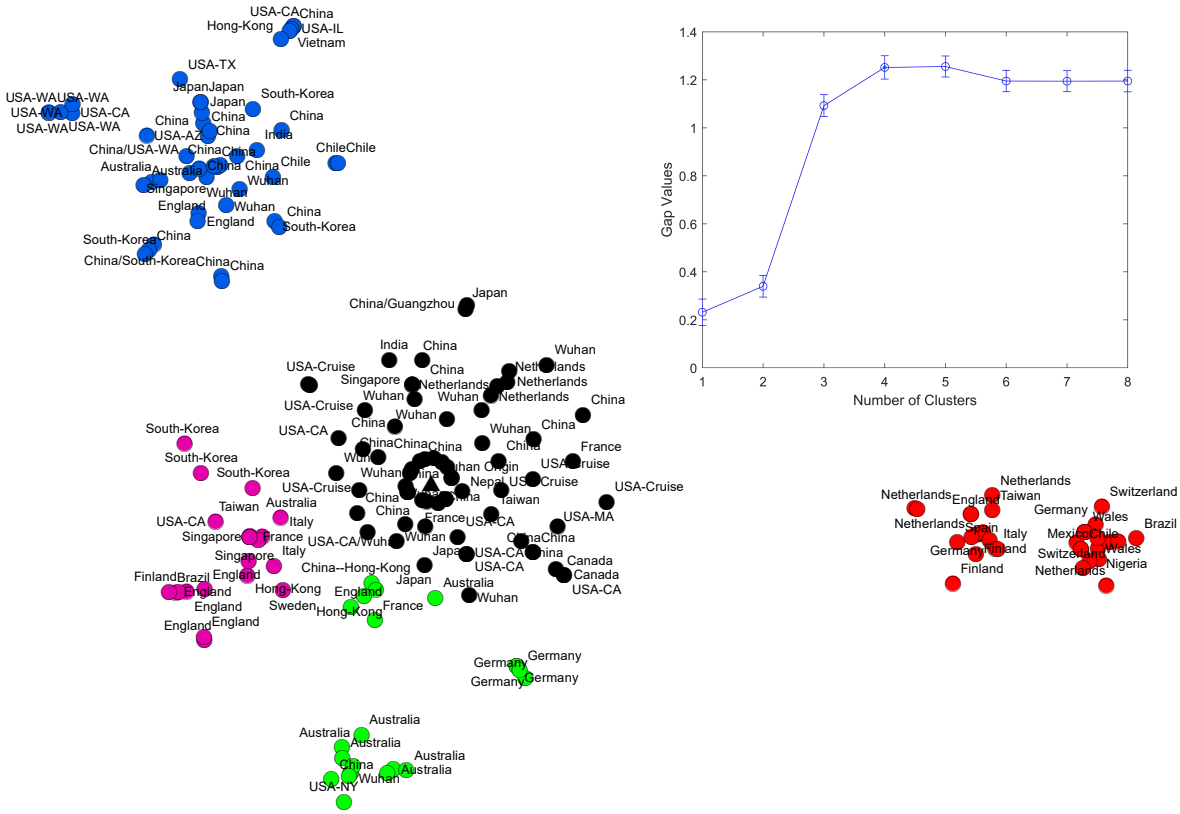


Figure 1: t-SNE plot of observed SARS-CoV-2 genomes. Wuhan-1 haplotype is depicted as a triangle. Identified clusters are highlighted in different colors. In the upper right corner, gap statistic values are shown.

Chile, including the subcluster of sequences from the US state of Washington corresponding to the ongoing outbreak there.

4) The cluster that include significant portion of genomes sampled in Australia and New Zealand (green). It should be noted that 2 infected hosts from Australia and 1 infected host from New Zealand are 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, and two sequences from China that belong to this cluster are related to their common ancestor. However, full-length sequences from Iran are currently unavailable, and therefore this scenario is currently hypothetical.

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).

It should be noted that these clusters do not consist exclusively of viral variants from the aforementioned regions, as air traffic probably drives the spread of SARS-CoV-2's viral genomic variants from each cluster around the globe.

We observed the negative log-linear relationship between cluster sizes and maximum pairwise Hamming distances between the genomes inside the clusters (Fig. 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 the older genomes. This mechanism is further confirmed in the following analysis.

3.2 Transmission network.

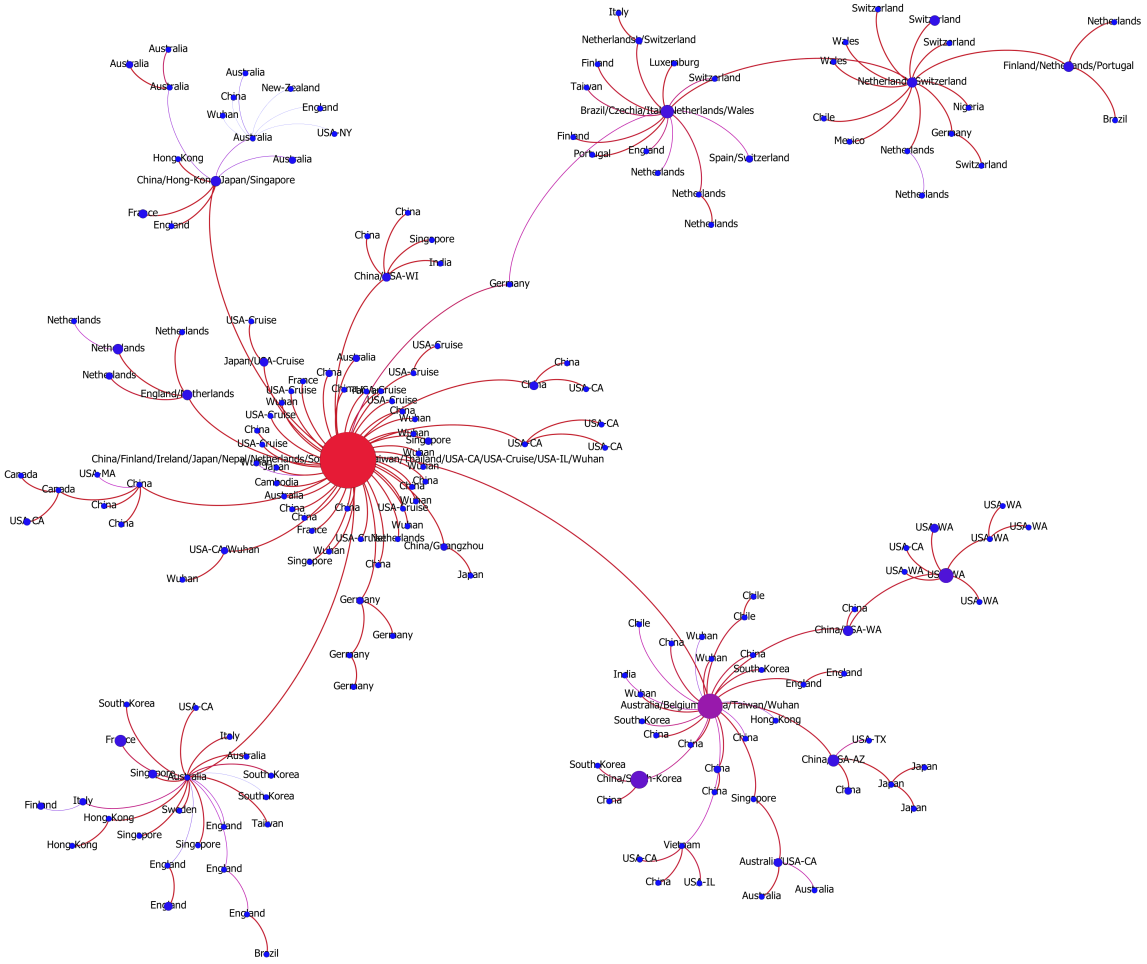
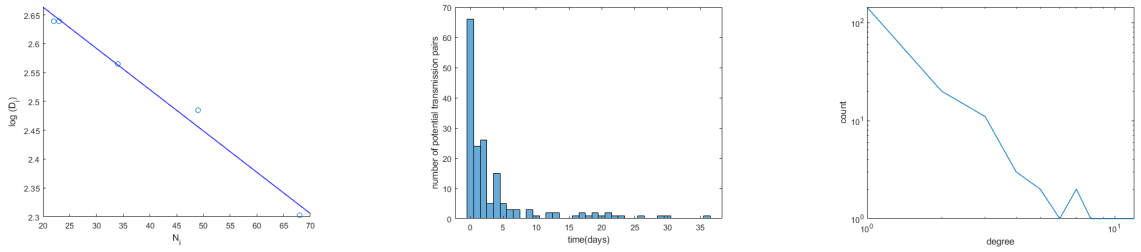


Figure 2: The transmission network constructed using SARS-CoV-2 genomes available as of March 10, 2020 and vizualized using Gephi [3]. Vertices represent viral genomes, and two vertices are connected by an arc if their mutational composition suggests potential direct or indirect transmission linkage 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.

The transmission network produced by the algorithms described above is visualized in Fig. 2. Although the majority of viral haplotypes (80%) were sampled in a single geographical



(a) (b) (c)

Figure 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).

location, many of them were sampled in multiple locations. Such haplotypes are annotated on Fig. 2 by the full list of their sampling locations. In what follows, the most prevalent haplotype sampled in Wuhan and associated with the initial phase of the epidemic (highlighted in red on Fig. 2) will be referred to as Wuhan-1 haplotype. For the potential transmission links e discussed below, their bootstrapping-based probabilities p_e are reported.

3.2.1 Network and temporal information

For the majority of genomes, their sampling times are known. The network structure was found to be in line with this temporal information, even though the network was constructed using the genomic data alone. Indeed, the correlation between network distances and differences in first sampling times between ancestor-descendant pairs of network nodes was 0.78 ($p < 10^{-116}$). The fact that this correlation is not absolutely perfect is not surprising, as sampling times are much more prone to different 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 potential source was sampled earlier than the potential recipient; and for 94.25% of these pairs their sampling times either agree or differ by up to 7 days. Finally, the mean minimum time difference between sampling times of potential transmission pairs was 3.74 days (95% $CI = [2.75, 4.73]$), while for the random pair of haplotypes the expected time difference was 20.48 days (95% $CI = [20.21, 20.74]$). This difference is statistically significant ($p < 10^{-45}$, Kolmogorov-Smirnov test). The distribution of sampling time differences for linked pairs is shown in Fig. 3b.

3.2.2 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. 3c). Degree distributions of such networks follow power law (i.e. the probability of having a particular degree is

Table 1: Comaprison 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

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 [62, 29], we fitted the following distributions to the observed degree distribution of the transmission network: negative binomial, Yule, Pareto and Waring. To compare the goodness of fit yielded by different models, we used the Akaike information criterion (AIC) and Bayesian Information Criterion (BIC) (Table 1). 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).

Furthermore, 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.

3.2.3 Transmission history

In general, the structure of the potential transmission network agrees with the distribution of t-SNE clusters (Fig. 1) and allows to hypothesize multiple transmission routes. In particular, the virus spread is characterized by multiple introductions of SARS-CoV2 into regions and countries: at this point for 14 out of 34 countries with reported sequences multiple introductions could be claimed, although additional data could adjust these estimations. Below we summarize the information about the transmission pathways in different regions outside the mainland China (whose subnetwork is depicted on Fig. A1) that could be deduced from the inferred network. It should be noted that we are currently lacking whole-genome sequencing data from Iran and Eastern Europe, therefore we concentrate on the analysis of regions with a sufficiently large number of available samples.

USA (Fig. A2) At this point, there are indications of multiple introductions of SARS-CoV2 into the country (not counting the cases from the Grand Princess cruise ship), as well as of sustained human-to-human transmissions inside the country. 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 $d = 1.79$. 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 form a connected subtree and thus suggest a single introduction from mainland China (since the root of the Washington subtree was also sampled in China) followed by the sustainable 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. So far, two possible cases of virus transmission between US states could be identified: from Arizona to Texas ($p_e = 0.83$) and from Washington to California ($p_e = 1$). Finally, as expected, 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 is linked to the Wuhan-1 haplotype through the haplotype sampled in Germany on January, 28, 2020 ($p_e = 0.79$) (Fi. A3). The parent of this haplotype is the Wuhan-1 haplotype ($p_e = 0.91$) and its only child is the haplo-

type later sampled in Italy. This potential transmission route is in agreement with epidemiological and molecular evidence reported by other sources [4, 22, 23]. For Italy, the analysis suggests that there was another independent SARS-CoV-2 introduction with no genomic evidence that it led to further spread. Similarly, at least two introductions are hypothesized for Netherlands (Fig. A4); however in that case both resulted in sustained host-to-host transmissions inside the country. The first Netherlands cluster is part of the major European cluster, while the second one could be linked to the Wuhan-1 haplotype ($p_e = 1$). Interestingly, this cluster has the genetic signature in the form of codon deletion in nsp2 genomic region that was observed only there; furthermore, both viral substrains are not geographically separated and co-exist in the same cities. A similar situation is observed in Germany, connected to the major European cluster and a separate branch sampled at North Rhine-Westphalia and 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 Italy.

The epidemiological history in the United Kingdom seems to be quite different from that of continental Europe (Fig. A5). There are multiple separate clusters detected there, three of which cannot be associated with other European cases, but are directly linked to the sequences sampled in China and Australia in January, 2020 (mean $p_e = 0.82$). Regarding the remaining haplotypes, one belongs to the second Netherlands cluster, while the remaining belong to the major European cluster (the majority of them being sampled in Wales). Two clusters show indications of intra-country transmissions (mean $p_e = 0.93$).

East Asia. All introductions in Singapore, Japan and South Korea (Figs. A6, A7 and A8) 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 could be 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. These regions currently reported few SARS-CoV-2 haplotypes, and the majority of cases most probably represent the third wave of the epidemics. Indeed, viral variants sampled in Nigeria, Mexico, 2 variants from Brazil and 1 variant from Chile are 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 supporting the hypothesis 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, Australia and Belgium (mean $p_e = 0.92$).

4 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 [14, 32, 33]. To achieve this aim, we proposed, implemented and used a computational framework to recover the network of potential SARS-CoV-2 transmissions using the aggre-

gated genomic data retrieved from GISAID repository [56]. The analysis allowed to identify the potential transmission links and routes of disease introduction in different regions of the planet, and confirm the hypothesis that there were multiple sources of introduction to the majority of countries. This conclusion supports the implementation of travel restrictions and border screening which have already been implemented in different locations. It should be also emphasized that the fact that a number of countries exhibited multiple introductions demonstrates how different susceptible human subpopulations are interconnected nowadays. This allowed the virus to exploit multiple transmission pathways and spread so rapidly. At the same time, our work 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 a 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 remains open and will require 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 [44, 50].

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 in China. Since the pandemic is currently only a few months old, no epidemiologically relevant lineages from the beginning of the epidemics are expected to die out anytime soon. Indeed, although the virus genetic diversity is gradually increasing, particular genomic variants continue to be repeatedly sequenced at different time points and geographical locations. It provides the means of tracking virus spread and evolution across time and space from the beginning of the epidemics using the methods of computational genomics and molecular epidemiology, and make conclusions about the potential routes of transmission.

We should warn that the results of such molecular surveillance analyses should be interpreted with caution. First of all, the genomic analyses do not necessarily replace traditional epidemiology methods that aim to investigate sources of transmission based on traditional surveillance systems that keeps track of the trajectory of the epidemic. Rather, genomic analyses complement and confirm other epidemiological findings using the sequencing data as an independent source of information that is not subject to the biases associated with the traditional epidemiological data [27, 2]. 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, the analysis is subject to the limitations associated with the nature of genomic data that include a small amount of cases in the beginning of the epidemics, underreporting and potential multiple sources of epidemic introductions. Further, 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 [57, 55, 45], and the decision to sequence each particular case is largely done subjectively in each specific country and lab.

In response to the rapidly growing number of cases of COVID-19 the authorities in different countries around the world have implemented unprecedentedly stringent travel and movement restrictions. Those measures were taken separately and independently country by country with different levels of escalation and at different times. In this context it is important to emphasize

the importance of globally coordinated measures and collaborations that should be supported by timely and evidence-supported analysis of reliable epidemiological data of diverse nature. Automatic high-performance computing-based molecular near real-time surveillance systems such as Nextstrain [28], HIV-Trace [36] and GHOST [40] could be instrumental in such public health global surveillance and decision making.

5 Footnotes

Competing interests. All the authors declare that they have no competing interests.
Funding. PS and AZ were supported by National Institutes of Health, [grant number 1R01EB025022]. GC was supported by National Science Foundation, [grant number 1414374]. PIB was supported by GSU Molecular Basis of Disease fellowship. The funding bodies have not played any roles in the design of the study and collection, analysis and interpretation of data in writing the manuscript.
Acknowledgements. We acknowledge all reserachers and laboratories who contributed their data to GISAID database. A full listing of all originating laboratories and authors is available at https://www.dropbox.com/s/4wjpi9lv5tlnsxh/gisaid.xlsx?dl=0
Software and data availability. The scripts for network reconstruction are available at https://github.com/compbel/SARS-CoV-2. The results of our analysis are posted at https://publichealth.gsu.edu/research/genomic-coronavirus/
Authors' Contributions. PS designed and implemented algorithms, processed and analyzed the data and wrote the paper; AK analyzed the data and wrote the paper; PIB designed algorithms and processed the data; AZ designed the algorithms and wrote the paper; GC analyzed the data and wrote the paper. All authors read and approved the final manuscript.

References

[1] WHO. `https://www.who.int/dg/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020`. Accessed: 2020-3-19.

[2] Armstrong, G. L., MacCannell, D. R., Taylor, J., Carleton, H. A., Neuhaus, E. B., Bradbury, R. S., Posey, J. E., and Gwinn, M. Pathogen genomics in public health. *New England Journal of Medicine 381*, 26 (2019), 2569–2580.

[3] Bastian, M., Heymann, S., and Jacomy, M. Gephi: an open source software for exploring and manipulating networks. In *Third international AAAI conference on weblogs and social media* (2009).

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

[5] Benvenuto, D., Giovanetti, M., Ciccozzi, A., Spoto, S., Angeletti, S., and Ciccozzi, M. The 2019-new coronavirus epidemic: Evidence for virus evolution, 2020.

[6] Bron, C., and Kerbosch, J. Algorithm 457: finding all cliques of an undirected graph. *Commun. ACM 16*, 9 (Sept. 1973), 575–577.

[7] Camin, J. H., and Sokal, R. R. A method for deducing branching sequences in phylogeny. *Evolution 19*, 3 (Sept. 1965), 311–326.

[8] Campo, D. S., Dimitrova, Z., Yamasaki, L., Skums, P., Lau, D. T., Vaughan, G., Forbi, J. C., Teo, C.-G., and Khudyakov, Y. Next-generation sequencing reveals large connected networks of intra-host hcv variants. *BMC genomics 15*, 5 (2014), S4.

[9] Campo, D. S., Xia, G.-L., Dimitrova, Z., Lin, Y., Forbi, J. C., Ganova-Raeva, L., Punkova, L., Ramachandran, S., Thai, H., Skums, P., et al. Accurate genetic detection of hepatitis c virus transmissions in outbreak settings. *The Journal of infectious diseases 213*, 6 (2016), 957–965.

[10] Ceraolo, C., and Giorgi, F. M. Genomic variance of the 2019-ncov coronavirus. *Journal of Medical Virology* (2020).

[11] Ceraolo, C., and Giorgi, F. M. Phylogenomic analysis of the 2019-ncov coronavirus. *bioRxiv* (2020).

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

[13] Chantal B E, Broberg, E. K., Haagmans, B., Meijer, A., Corman, V. M., Papa, A., Charrel, R., Drosten, C., Koopmans, M., Leitmeyer, K., 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 25*, 6 (Feb. 2020), 2000082.

[14] Chen, N., Zhou, M., Dong, X., Qu, J., Gong, F., Han, Y., Qiu, Y., Wang, J., Liu, Y., Wei, Y., et al. Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in wuhan, china: a descriptive study. *The Lancet 395*, 10223 (2020), 507–513.

[15] Chen, N., Zhou, M., Dong, X., Qu, J., Gong, F., Han, Y., Qiu, Y., Wang, J., Liu, Y., Wei, Y., Xia, J., Yu, T., Zhang, X., and Zhang, L. Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in wuhan, china: a descriptive study, 2020.

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

[17] Cheng, Z. J., and Shan, J. 2019 novel coronavirus: where we are and what we know. *Infection* (Feb. 2020).

[18] Chu, Y.-J. On the shortest arborescence of a directed graph. *Scientia Sinica 14* (1965), 1396–1400.

[19] Cleemput, S., Dumon, W., Fonseca, V., Karim, W. A., Giovanetti, M., Alcantara, L. C., Deforche, K., and de Oliveira, T. Genome detective coronavirus typing tool for rapid identification and characterization of novel coronavirus genomes. *Bioinformatics* (Feb. 2020).

[20] Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res. 32*, 5 (Mar. 2004), 1792–1797.

[21] Edmonds, J. Optimum branchings. *Journal of Research of the national Bureau of Standards B 71*, 4 (1967), 233–240.

[22] Giovanetti, M., Angeletti, S., Benvenuto, D., and Ciccozzi, M. A doubt of multiple introduction of sars-cov-2 in italy: a preliminary overview. *Journal of Medical Virology*.

[23] Giovanetti, M., Benvenuto, D., Angeletti, S., and Ciccozzi, M. The first two cases of 2019-ncov in italy: where they come from? *Journal of Medical Virology* (2020).

[24] Gusfield, D. Algorithms on stings, trees, and sequences: Computer science and computational biology. *Acm Sigact News 28*, 4 (1997), 41–60.

[25] Gusfield, D. Algorithms on strings, trees, and sequences. 1997. *Computer Science and Computational Biology. New York: Cambridge University Press* (1997).

[26] Gusfield, D., Frid, Y., and Brown, D. Integer programming formulations and computations solving phylogenetic and population genetic problems with missing or genotypic data. In *Computing and Combinatorics* (2007), Springer Berlin Heidelberg, pp. 51–64.

[27] Gwinn, M., MacCannell, D., and Armstrong, G. L. Next-generation sequencing of infectious pathogens. *Jama 321*, 9 (2019), 893–894.

[28] Hadfield, J., Megill, C., Bell, S. M., Huddleston, J., Potter, B., Callender, C., Sagulenko, P., Bedford, T., and Neher, R. A. Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics 34*, 23 (Dec. 2018), 4121–4123.

[29] Hamilton, D. T., Handcock, M. S., and Morris, M. Degree distributions in sexual networks: a framework for evaluating evidence. *Sexually transmitted diseases 35*, 1 (2008), 30.

[30] Holshue, M. L., DeBolt, C., Lindquist, S., Lofy, K. H., Wiesman, J., Bruce, H., Spitters, C., Ericson, K., Wilkerson, S., Tural, A., Diaz, G., Cohn, A., Fox, L., Patel, A., Gerber, S. I., Kim, L., Tong, S., Lu, X., Lindstrom, S., Pallansch, M. A., Weldon, W. C., Biggs, H. M., Uyeki, T. M., Pillai, S. K., and Washington State 2019-nCoV Case Investigation Team. First case of 2019 novel coronavirus in the united states. *N. Engl. J. Med.* (Jan. 2020).

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

[32] Huang, C., Wang, Y., Li, X., Ren, L., Zhao, J., Hu, Y., Zhang, L., Fan, G., Xu, J., Gu, X., et al. Clinical features of patients infected with 2019 novel coronavirus in wuhan, china. *The Lancet 395*, 10223 (2020), 497–506.

[33] Hui, D. S., I Azhar, E., Madani, T. A., Ntoumi, F., Kock, R., Dar, O., Ippolito, G., Mchugh, T. D., Memish, Z. A., Drosten, C., 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 91* (2020), 264–266.

[34] Jahn, K., Kuipers, J., and Beerenwinkel, N. Tree inference for single-cell data. *Genome biology 17*, 1 (2016), 86.

[35] Kim, J. Y., Choe, P. G., Oh, Y., Oh, K. J., Kim, J., Park, S. J., Park, J. H., Na, H. K., and Oh, M. D. The first case of 2019 novel coronavirus pneumonia imported into korea from wuhan, china: Implication for infection prevention and control measures. *J. Korean Med. Sci. 35*, 5 (Feb. 2020), e61.

[36] Kosakovsky Pond, S. L., Weaver, S., Leigh Brown, A. J., and Wertheim, J. O. Hiv-trace (transmission cluster engine): a tool for large scale molecular epidemiology of hiv-1 and other rapidly evolving pathogens. *Molecular biology and evolution 35*, 7 (2018), 1812–1819.

[37] Li, W. Bats are natural reservoirs of SARS-Like coronaviruses, 2005.

[38] Li, X., Wang, W., Zhao, X., Zai, J., Zhao, Q., Li, Y., and Chaillon, A. Transmission dynamics and evolutionary history of 2019-ncov. *Journal of Medical Virology* (2020).

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

[40] Longmire, A. G., Sims, S., Rytsareva, I., Campo, D. S., Skums, P., Dimitrova, Z., Ramachandran, S., Medrzycki, M., Thai, H., Ganova-Raeva, L., et al. Ghost: global hepatitis outbreak and surveillance technology. *BMC genomics 18*, 10 (2017), 916.

[41] Lu, H., Stratton, C. W., and Tang, Y. Outbreak of pneumonia of unknown etiology in wuhan china: the mystery and the miracle, 2020.

[42] Lu, R., Zhao, X., Li, J., Niu, P., Yang, B., Wu, H., Wang, W., Song, H., Huang, B., Zhu, N., et al. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. *The Lancet 395*, 10224 (2020), 565–574.

[43] Maaten, L. v. d., and Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res. 9*, Nov (2008), 2579–2605.

[44] May, R. M., and Lloyd, A. L. Infection dynamics on scale-free networks. *Physical Review E 64*, 6 (2001), 066112.

[45] Muir, P., Li, S., Lou, S., Wang, D., Spakowicz, D. J., Salichos, L., Zhang, J., Weinstock, G. M., Isaacs, F., Rozowsky, J., et al. The real cost of sequencing: scaling computation to keep pace with data generation. *Genome biology 17*, 1 (2016), 53.

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

[47] Optimization, G. Inc.,"gurobi optimizer reference manual," 2015, 2014.

[48] Paraskevis, D., Kostaki, E. G., Magiorkinis, G., Panayiotakopoulos, G., Sourvinos, G., and Tsiodras, S. Full-genome evolutionary analysis of the novel corona virus (2019-ncov) rejects the hypothesis of emergence as a result of a recent recombination event. *Infection, Genetics and Evolution 79* (2020), 104212.

[49] Parry, J. China coronavirus: cases surge as official admits human to human transmission. *BMJ 368* (Jan. 2020), m236.

[50] Pastor-Satorras, R., and Vespignani, A. Epidemic spreading in scale-free networks. *Physical review letters 86*, 14 (2001), 3200.

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

[52] Ragonnet-Cronin, M., Hu, Y. W., Morris, S. R., Sheng, Z., Poortinga, K., and Wertheim, J. O. Hiv transmission networks among transgender women in los angeles county, ca, usa: a phylogenetic analysis of surveillance data. *The Lancet HIV 6*, 3 (2019), e164–e172.

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

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

[55] Schwarze, K., Buchanan, J., Fermont, J. M., Dreau, H., Tilley, M. W., Taylor, J. M., Antoniou, P., Knight, S. J., Camps, C., Pentony, M. M., 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 22*, 1 (2020), 85–94.

[56] Shu, Y., and McCauley, J. Gisaid: Global initiative on sharing all influenza data–from vision to reality. *Eurosurveillance 22*, 13 (2017).

[57] Taber, K. A. J., Dickinson, B. D., and Wilson, M. The promise and challenges of next-generation genome sequencing for clinical care. *JAMA internal medicine 174*, 2 (2014), 275–280.

[58] Tibshirani, R., Walther, G., and Hastie, T. Estimating the number of clusters in a data set via the gap statistic. *J. R. Stat. Soc. Series B Stat. Methodol. 63*, 2 (May 2001), 411–423.

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

[60] Wan, Y., Shang, J., Graham, R., Baric, R. S., and Li, F. Receptor recognition by novel coronavirus from wuhan: An analysis based on decade-long structural studies of SARS. *J. Virol.* (Jan. 2020).

[61] Wang, H., Wang, Z., Dong, Y., Chang, R., Xu, C., Yu, X., Zhang, S., Tsamlag, L., Shang, M., Huang, J., Wang, Y., Xu, G., Shen, T., Zhang, X., and Cai, Y. Phase-adjusted estimation of the number of coronavirus disease 2019 cases in wuhan, china. *Cell Discovery 6*, 1 (Feb. 2020), 1–8.

[62] Wertheim, J. O., Leigh Brown, A. J., Hepler, N. L., Mehta, S. R., Richman, D. D., Smith, D. M., and Kosakovsky Pond, S. L. The global transmission network of hiv-1. *The Journal of infectious diseases 209*, 2 (2014), 304–313.

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

[64] Wu, F., Zhao, S., Yu, B., Chen, Y.-M., Wang, W., Hu, Y., Song, Z.-G., Tao, Z.-W., Tian, J.-H., Pei, Y.-Y., Yuan, M.-L., Zhang, Y.-L., Dai, F.-H., Liu, Y., Wang, Q.-M., Zheng, J.-J., Xu, L., Holmes, E. C., and Zhang, Y.-Z. Complete genome characterisation of a novel coronavirus associated with severe human respiratory disease in wuhan, china. Feb. 2020.

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

[66] Xu, X., Chen, P., Wang, J., Feng, J., Zhou, H., Li, X., Zhong, W., and Hao, P. Evolution of the novel coronavirus from the ongoing wuhan outbreak and modeling of its spike protein for risk of human transmission. *Sci. China Life Sci.* (Jan. 2020).

[67] Zhang, L., Shen, F.-M., Chen, F., and Lin, Z. Origin and evolution of the 2019 novel coronavirus. *Clin. Infect. Dis.* (Feb. 2020).

[68] Zhao, S., Lin, Q., Ran, J., Musa, S. S., Yang, G., Wang, W., Lou, Y., Gao, D., Yang, L., He, D., et al. Preliminary estimation of the basic reproduction number of novel coronavirus (2019-ncov) in china, from 2019 to 2020: A data-driven analysis in the early phase of the outbreak. *International Journal of Infectious Diseases 92* (2020), 214–217.

[69] Zhao, S., Musa, S. S., Lin, Q., Ran, J., Yang, G., Wang, W., Lou, Y., Yang, L., Gao, D., He, D., et al. Estimating the unreported number of novel coronavirus (2019-ncov) cases in china in the first half of january 2020: a data-driven modelling analysis of the early outbreak. *Journal of clinical medicine 9*, 2 (2020), 388.

[70] Zhou, P., Yang, X.-L., Wang, X.-G., Hu, B., Zhang, L., Zhang, W., Si, H.-R., Zhu, Y., Li, B., Huang, C.-L., Chen, H.-D., Chen, J., Luo, Y., Guo, H., Jiang, R.-D., Liu, M.-Q., Chen, Y., Shen, X.-R., Wang, X., Zheng, X.-S., Zhao, K., Chen, Q.-J., Deng, F., Liu, L.-L., Yan, B., Zhan, F.-X., Wang, Y.-Y., Xiao, G.-F., and Shi, Z.-L. A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* (Feb. 2020), 1–4.

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

Appendix

Below are transmission subnetworks for the regions described in the paper.

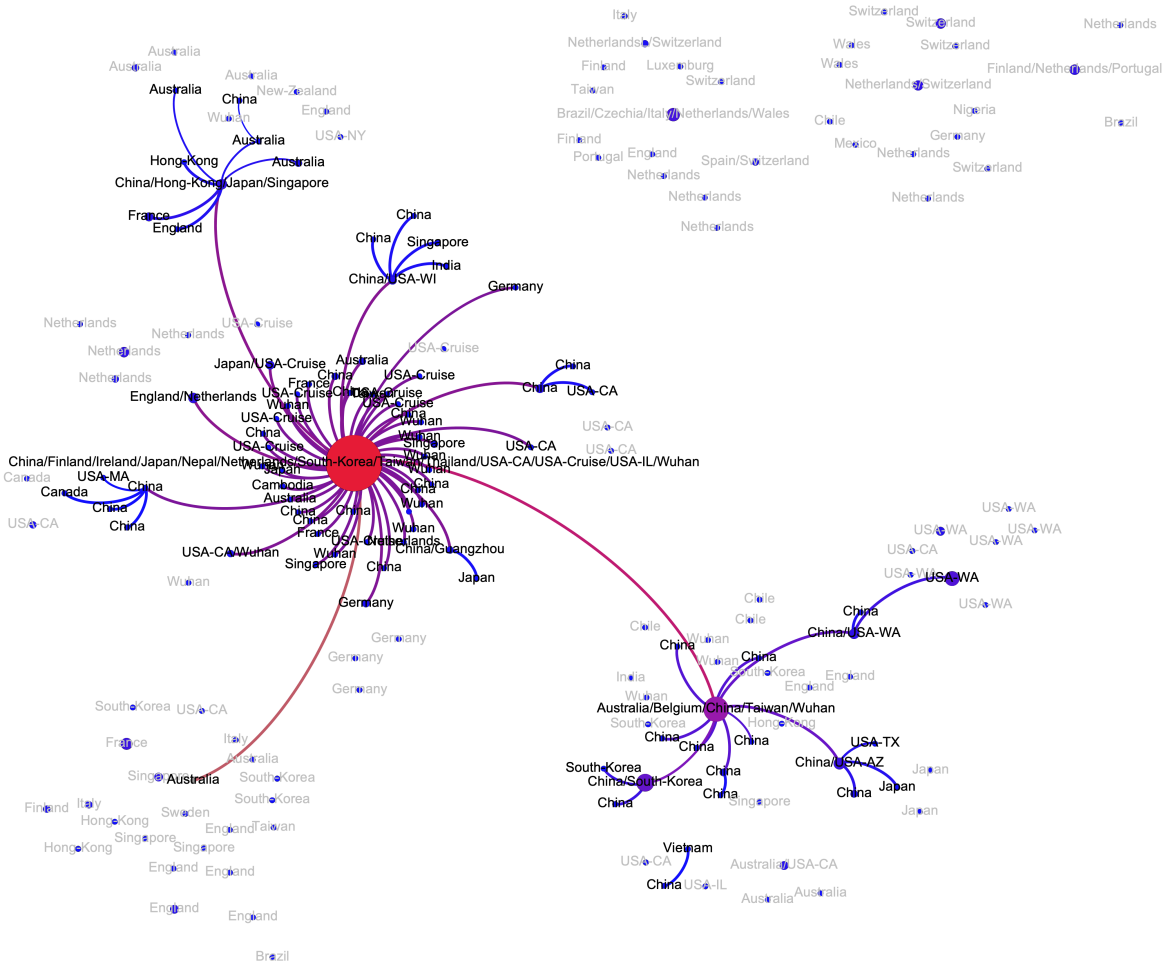


Figure A1: Transmission subnetwork of genomes sampled in China

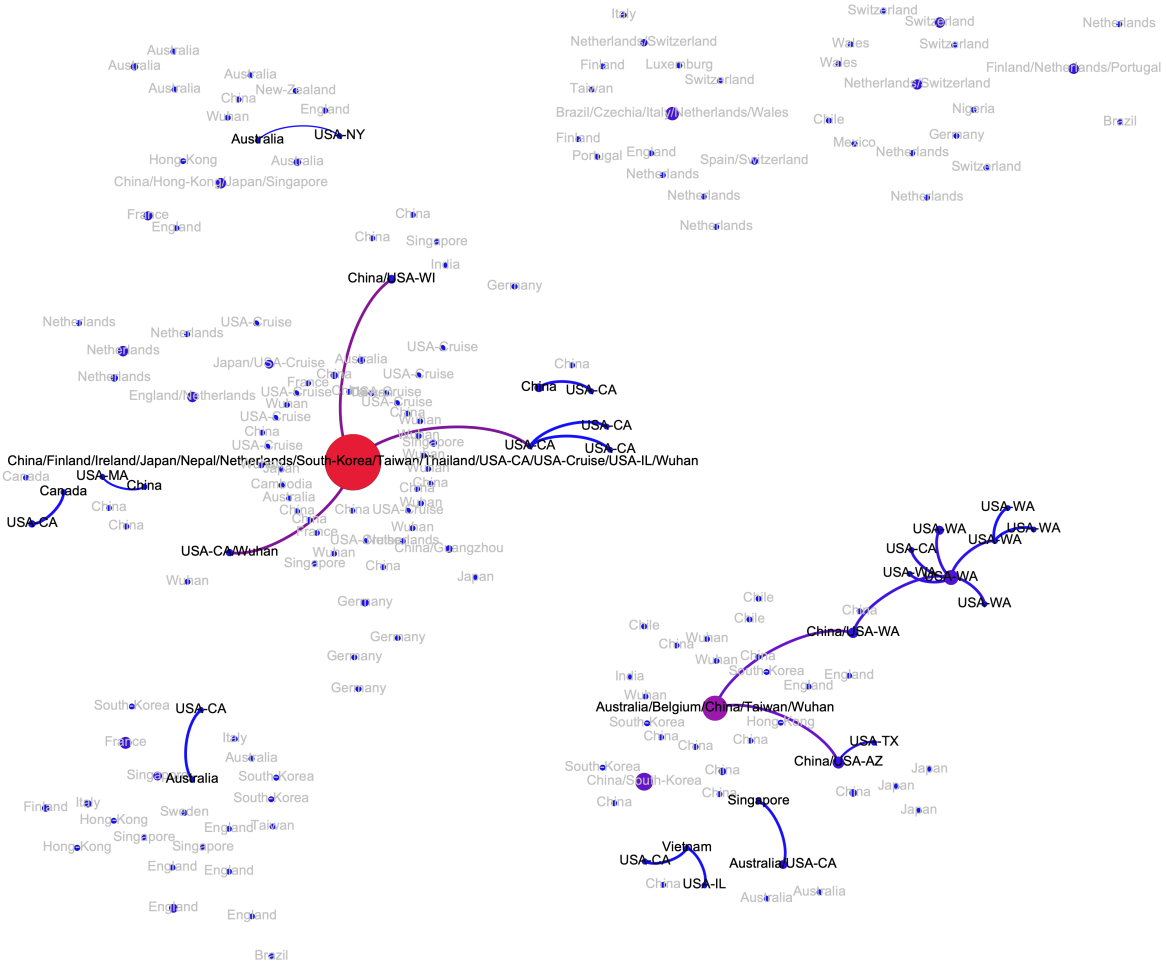


Figure A2: Transmission subnetwork of genomes sampled in USA

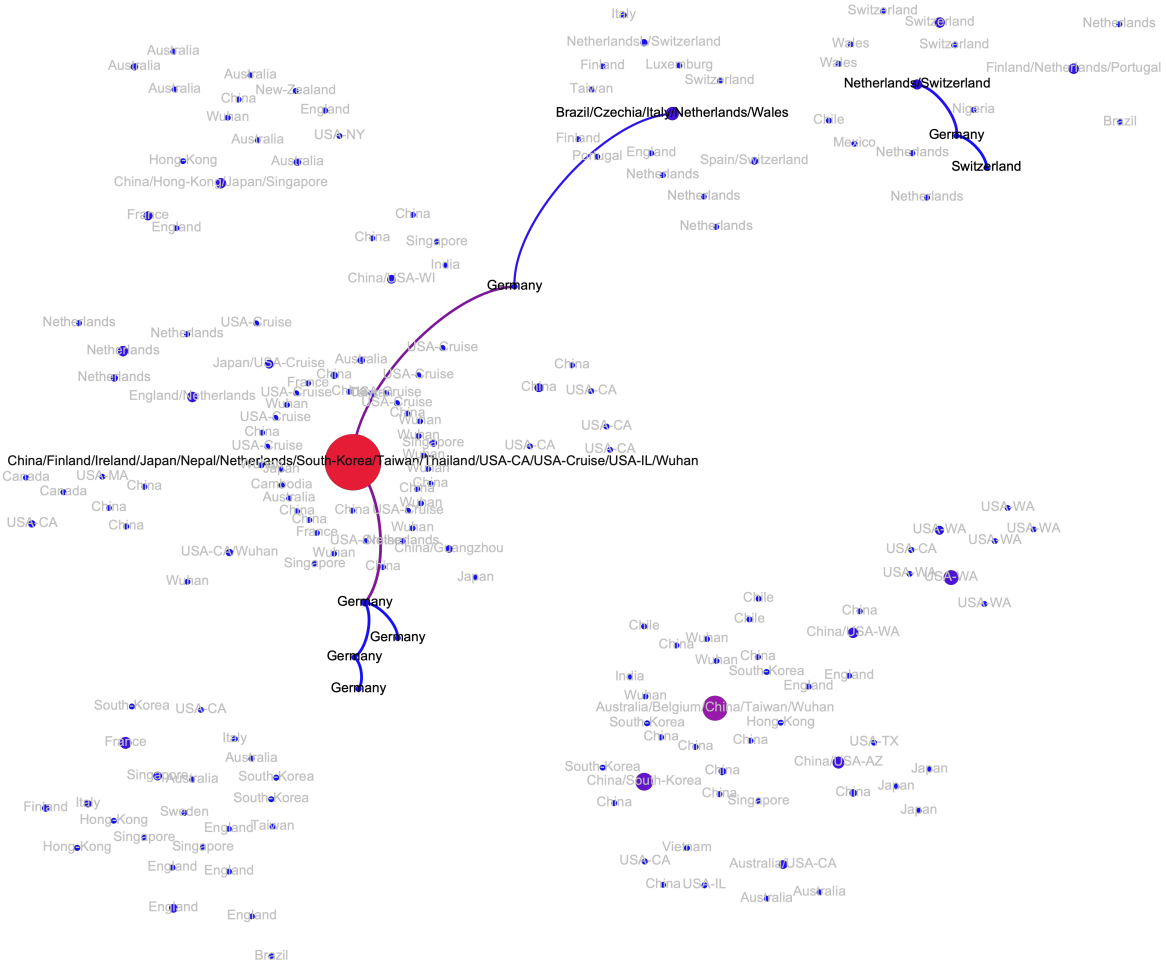


Figure A3: Transmission subnetwork of genomes sampled in Germany

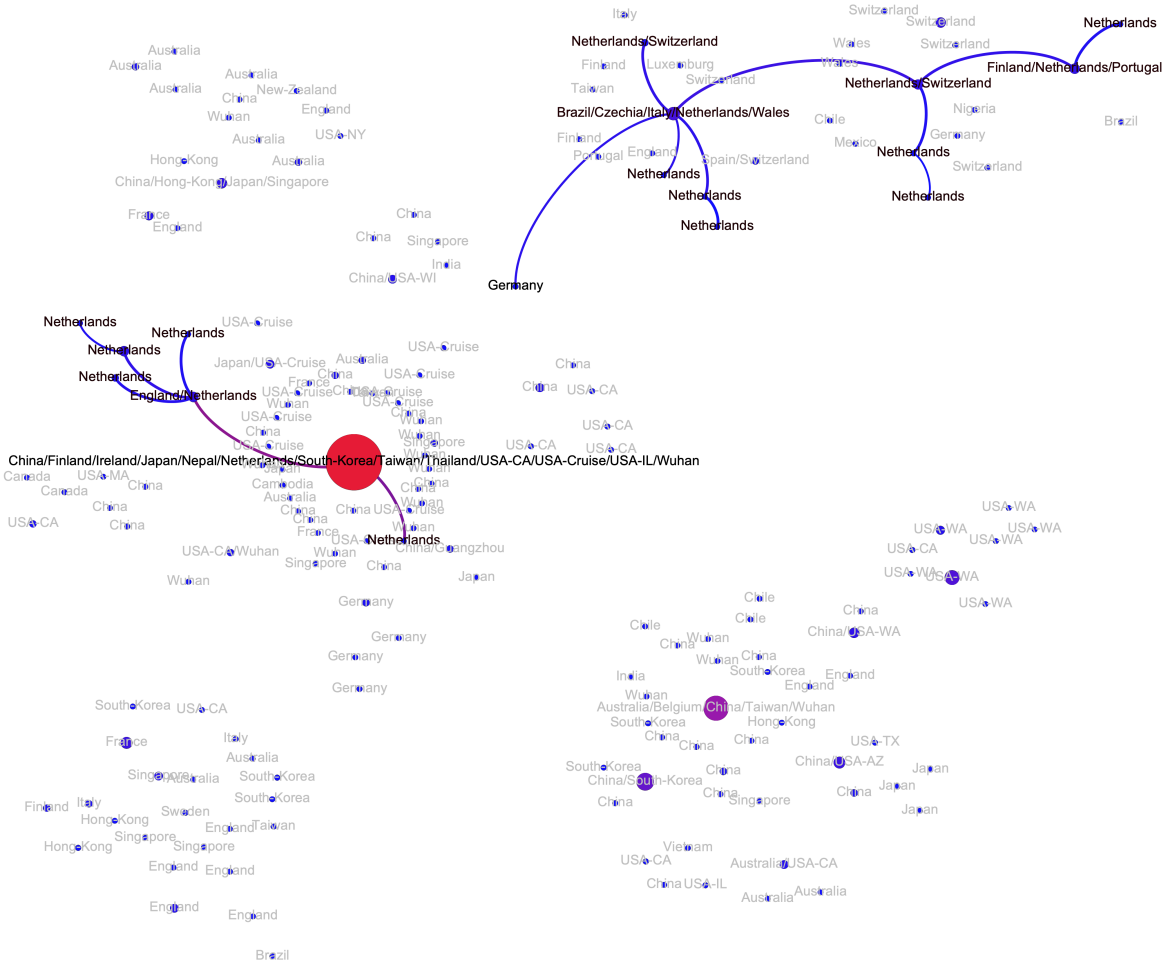


Figure A4: Transmission subnetwork of genomes sampled in Netherlands

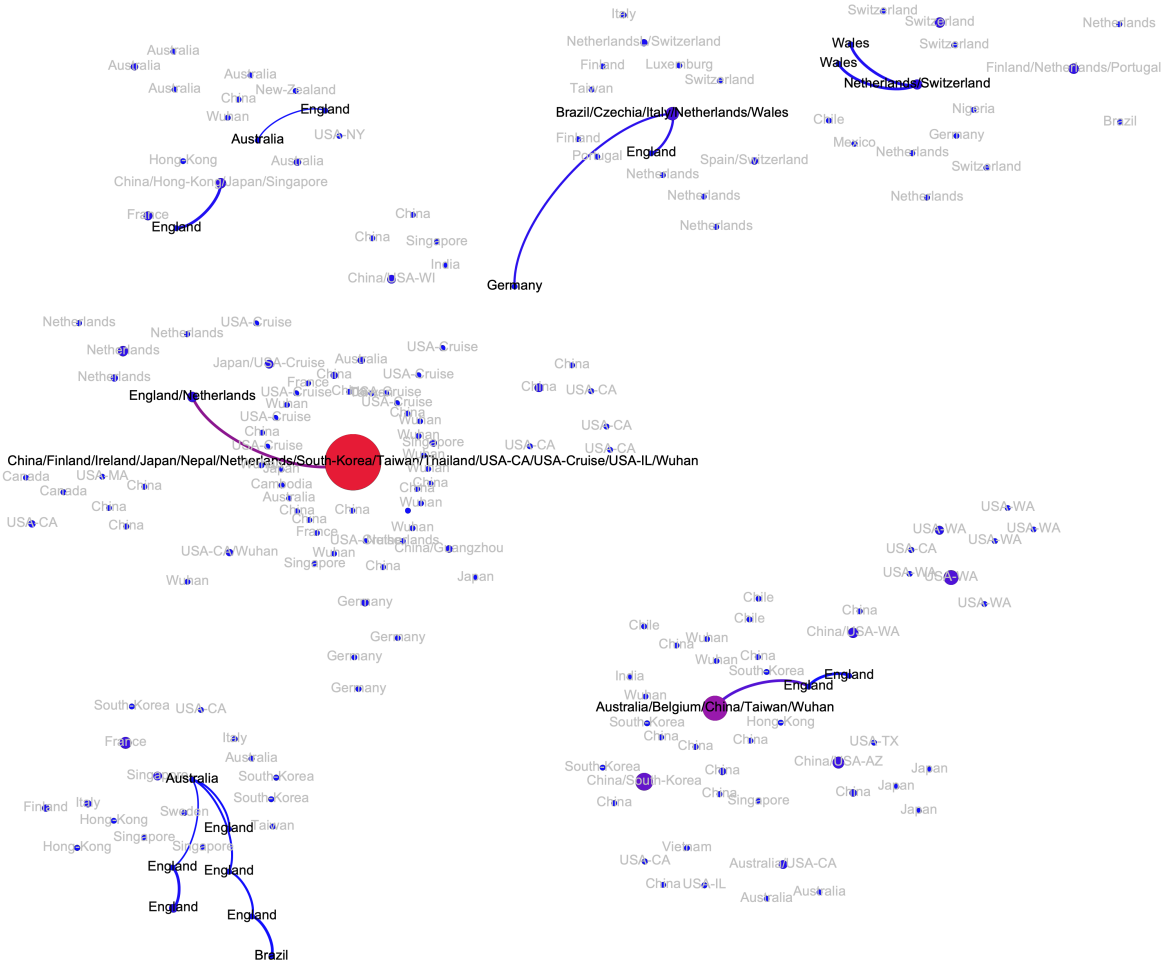


Figure A5: Transmission subnetwork of genomes sampled in UK

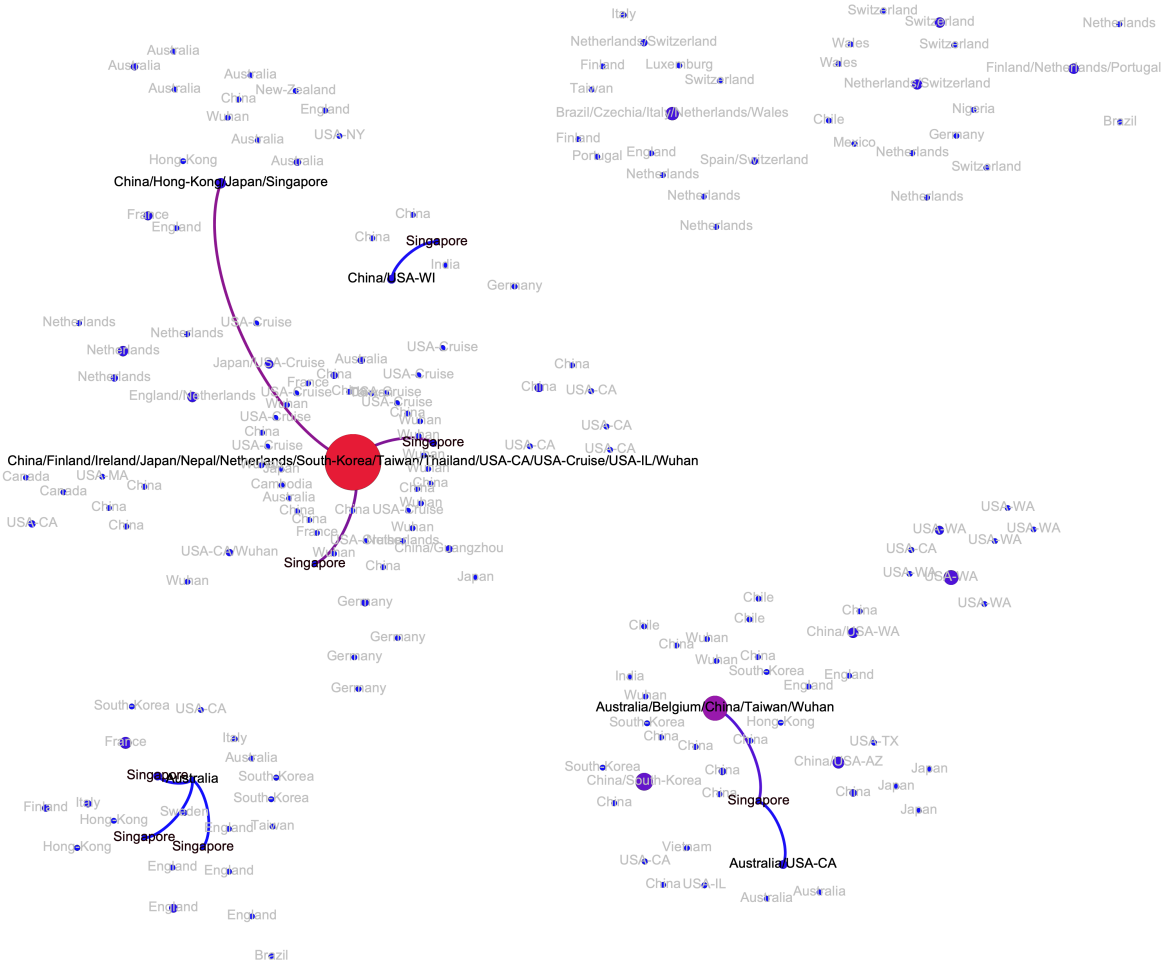


Figure A6: Transmission subnetwork of genomes sampled in Singapore

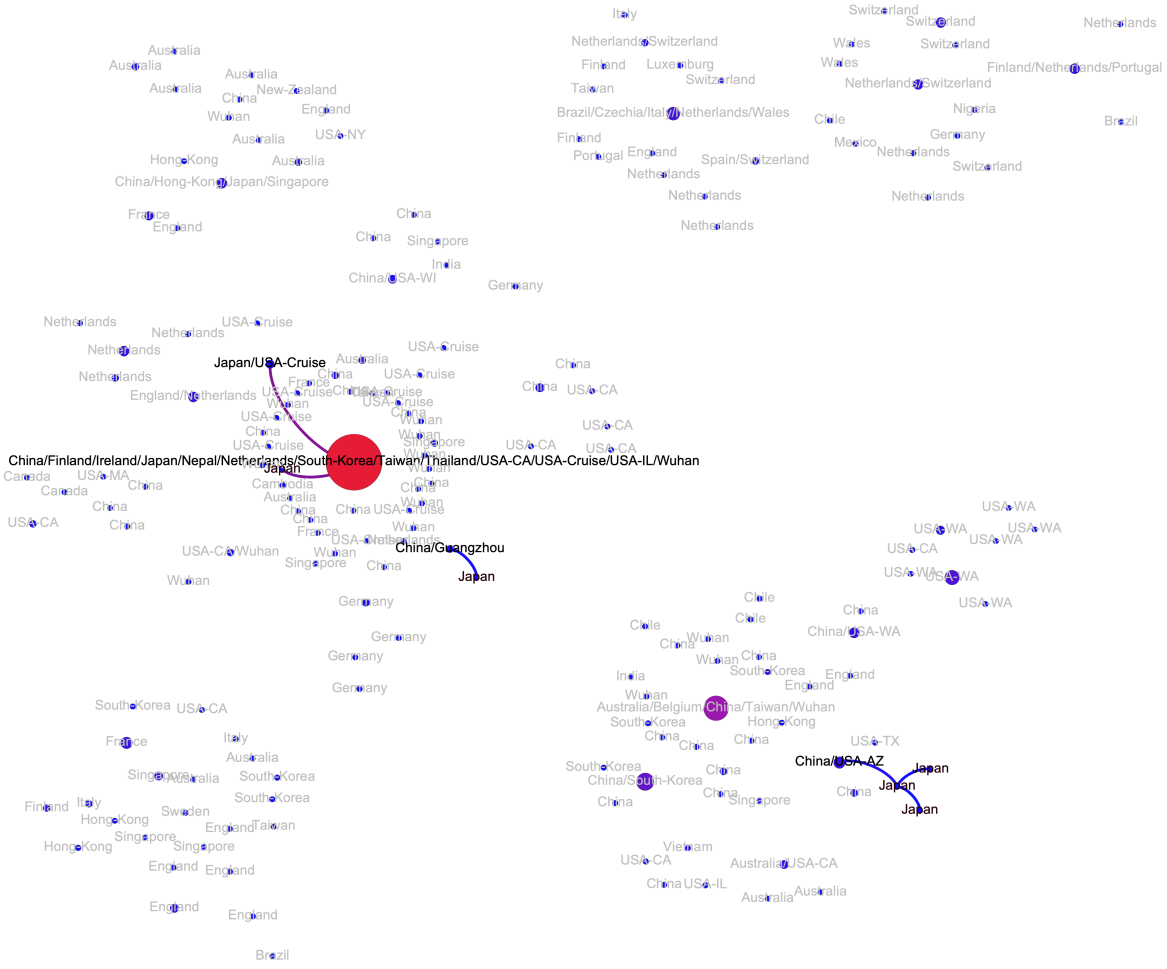


Figure A7: Transmission subnetwork of genomes sampled in Japan

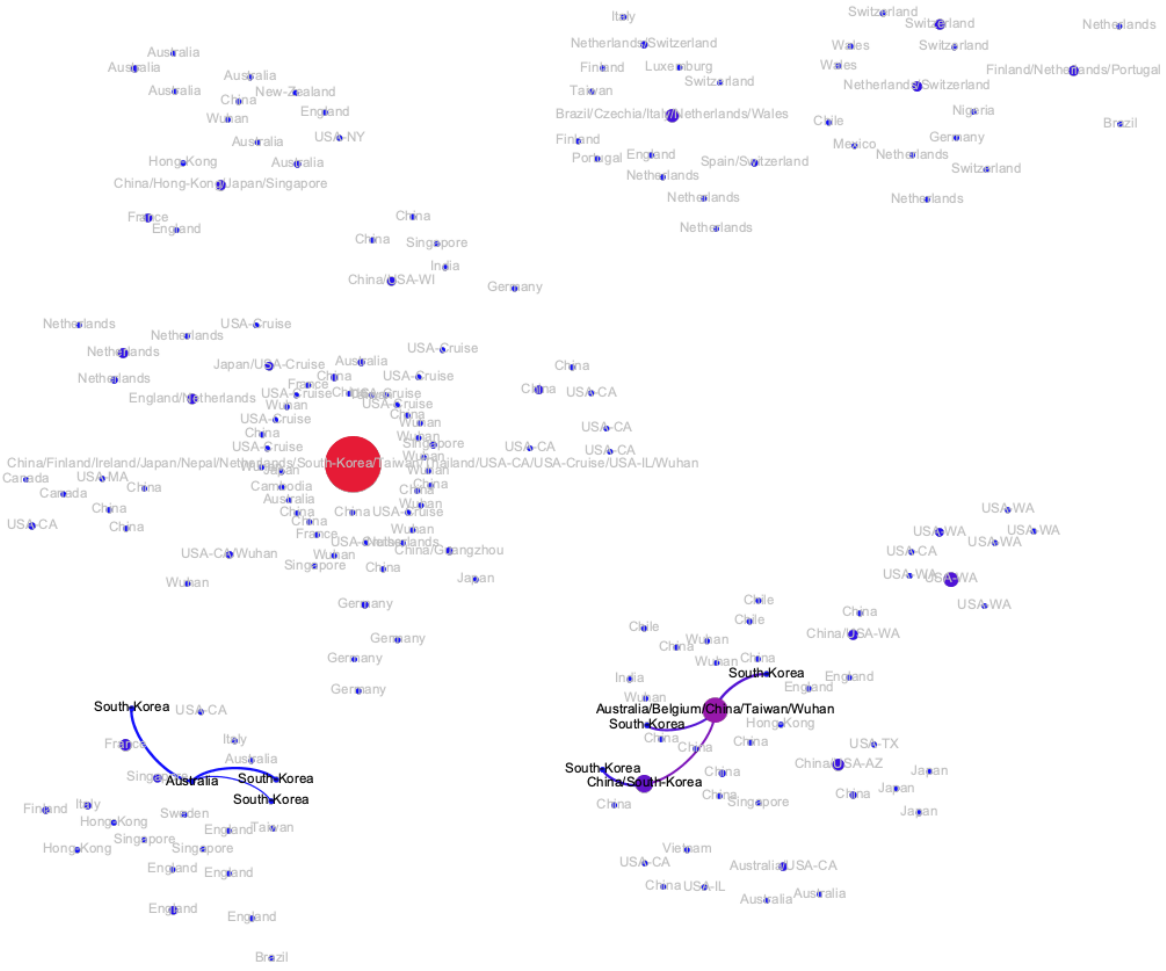


Figure A8: Transmission subnetwork of genomes sampled in South Korea

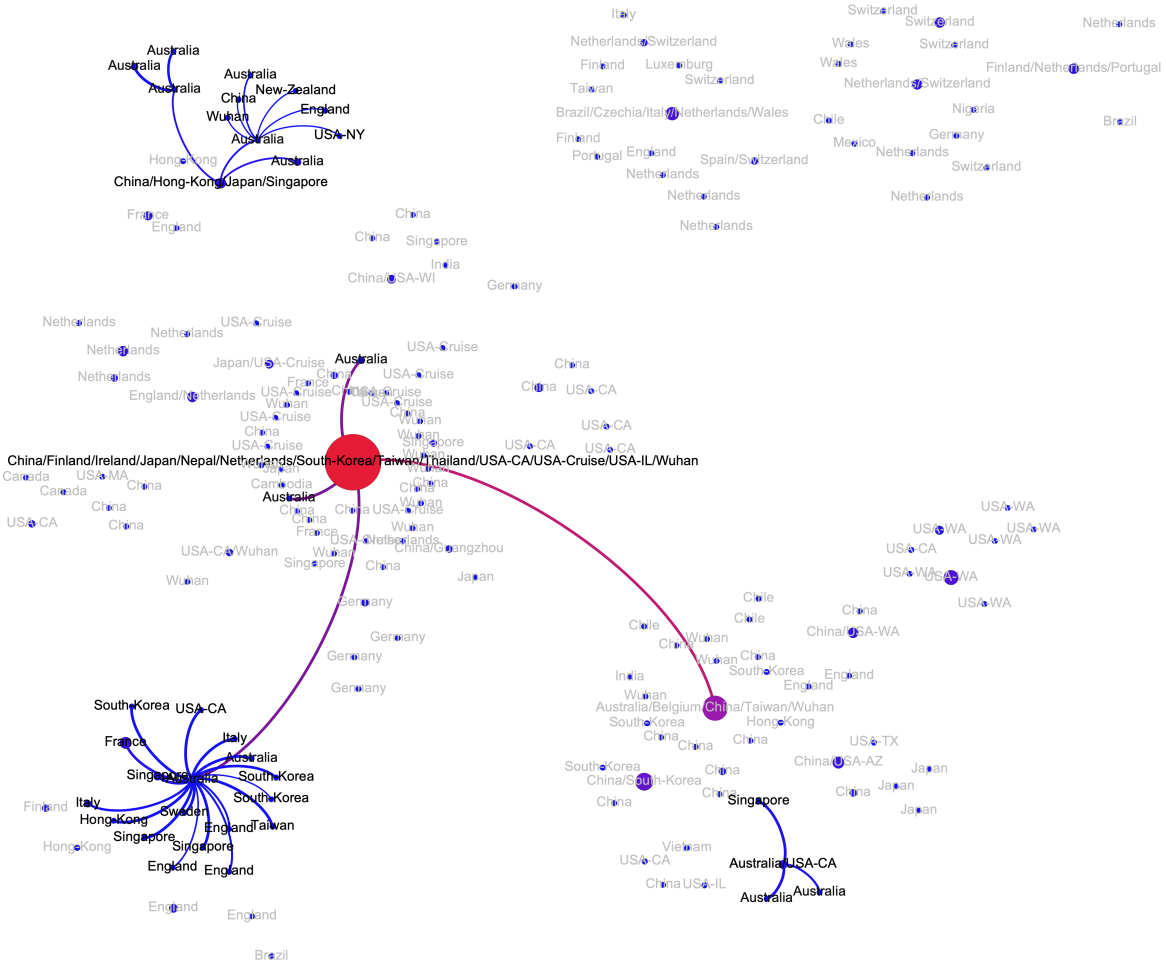


Figure A9: Transmission subnetwork of genomes sampled in Australia

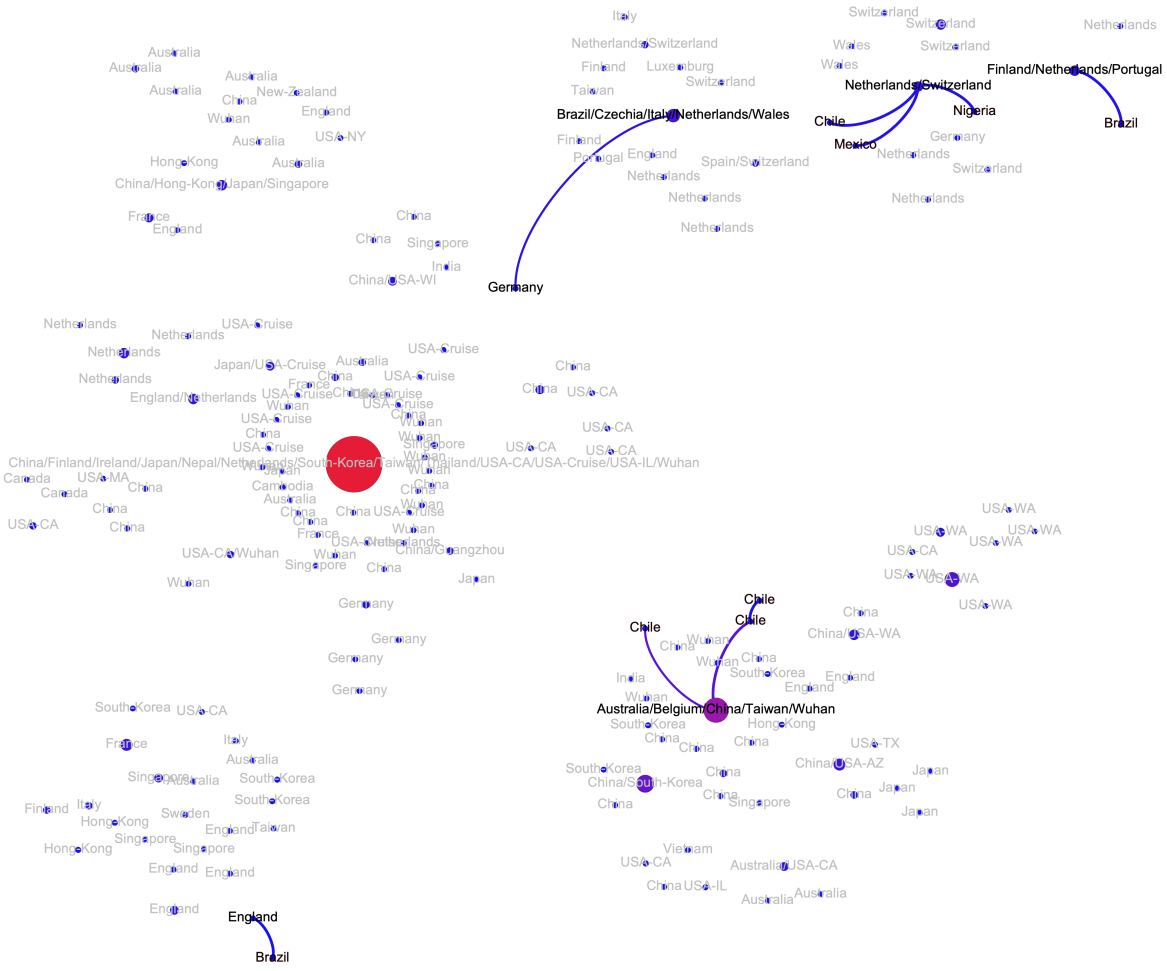


Figure A10: Transmission subnetwork of genomes sampled in Africa, Central and South America