Explore the content by scrolling the left hand side (or click on the arrows), and the data visualizations will change accordingly. Clicking "explore the data yourself" above will display sidebar controls.

Genomic analysis of nCoV spread. Situation report 2020-01-23.

Author: Trevor Bedford, Richard Neher, James Hadfield, Emma Hodcroft, Misja Ilcisin, Nicola Müller [1]

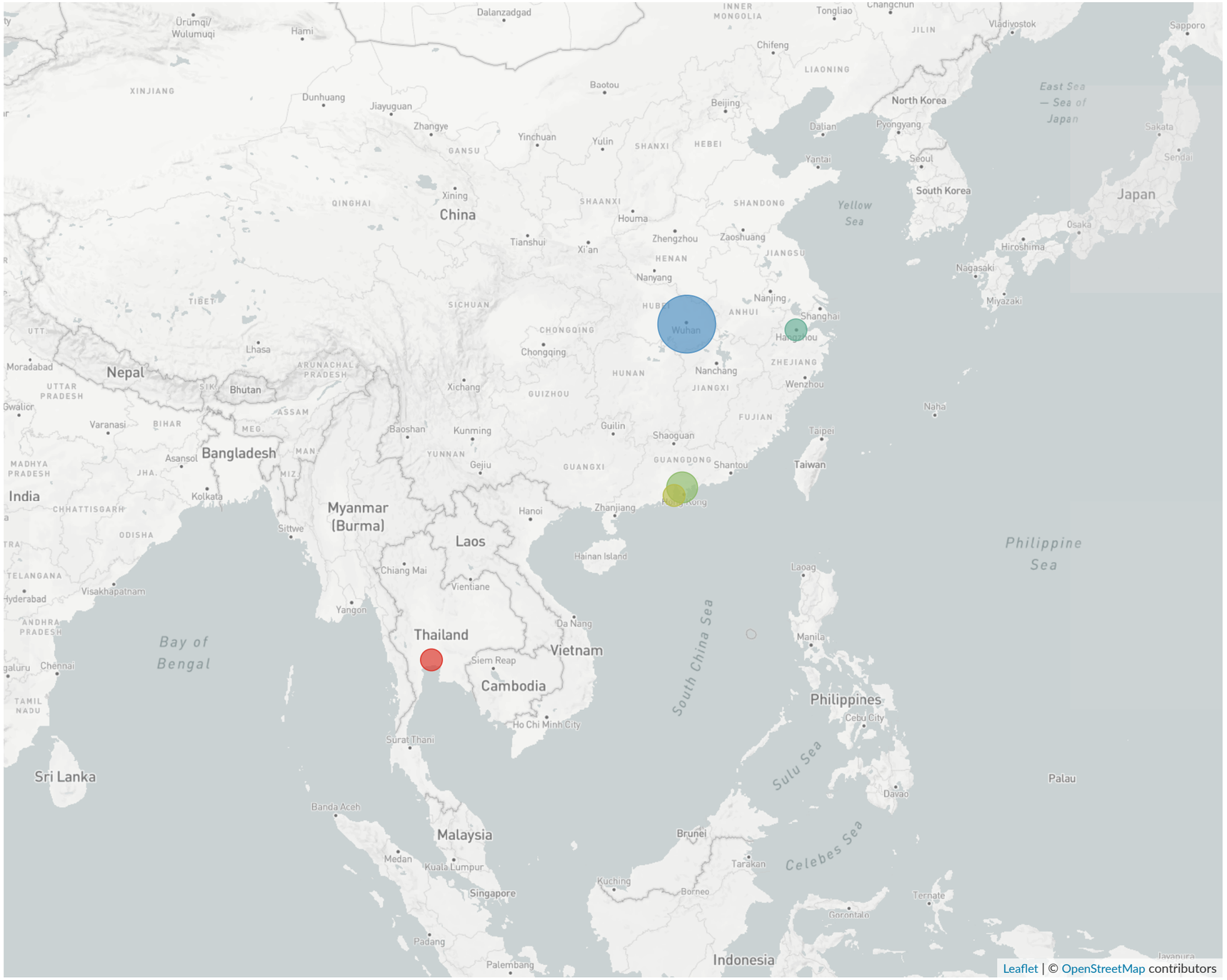
[1] Fred Hutch, Seattle, USA and Biozentrum, Basel, Switzerland

Created: 2020 Jan 23

This report uses publicly shared novel coronavirus (nCoV) genomic data from GISAID to estimate rates and patterns of viral epidemic spread. We plan to issue updated situation reports as new data is produced and shared. This website is optimized for display on desktop browsers.

Geography

EXPLORE THE DATA YOURSELF



Executive summary

Executive summary

Using 24 public shared novel coronavirus (nCoV) genomes, we examined genetic diversity to infer date of common ancestor and rate of spread. We find:

- 24 sampled genomes are nearly identical, differing by 0-3 mutations
- This lack of genetic diversity has a parsimonious explanation that the outbreak descends either from a single introduction into the human population or a small number of animal to human transmissions of very similar viruses.
- This event most likely occurred in November or early December 2019.
- There has been ongoing human-to-human spread since this point resulting in observed cases.
- Using estimates of total case count from Imperial College London of several thousand cases, we infer a reproductive number between 1.5 and 3.5 indicating rapid growth in the Nov-Jan period.

Coronaviruses

Further Reading:

- General information on coronaviruses on Wikipedia *2020-01-23*
- Material provided by the US CDC *2020-01-23*
- Organization and genome on ViralZone *2020-01-23*

Different human coronaviruses

Coronaviruses (CoV) are members of a diverse species of positive-sense single-stranded RNA ((+)ssRNA) viruses which have a history of causing respiratory infections in humans. Some variants of coronaviruses are associated with outbreaks, others are continuously circulating and cause mostly mild respiratory infections (e.g. the common cold).

SARS-CoV & MERS-CoV

The most well known of these coronaviruses is SARS-CoV ("severe acute respiratory syndrome"), which in a Nov 2002 to Jul 2003 outbreak spread around the world and resulted in over 8000 cases and 774 deaths, with a case fatality rate of around 9–11%.

In 2012, a novel coronavirus, MERS-CoV ("Middle East respiratory syndrome"), causing severe respiratory symptoms was identified. MERS has resulted in fatalities comparable to SARS, however the transmission route of MERS is very different. Whereas SARS was efficiently spread from one human to another, human MERS infections were generally a result of independent zoonoses (animal to human transmissions) from camels (see Dudas *et al.* for more information). This has lead to a self-limiting outbreak largely restricted to the Arabian Peninsula.

Seasonal CoV

However, not all coronaviruses are as deadly as SARS-CoV and MERS-CoV. There are four "seasonal" coronaviruses that commonly infect humans each year. Compared with SARS, these seasonal coronavirus strains are "much more prevalent, much less severe, and common causes of influenza-like illness (ILI)". In fact, 5–12% of all ILI cases test positive for coronaviruses, so they are rather common, resulting in millions of infections every year with low severity. These seasonal coronaviruses are the results of separate spillovers from the bat animal reservoir into humans in the past ~100 years, in which after spillover, each seasonal virus established itself and spread widely in the human population.

Animal reservoirs

Coronaviruses infect a wide range of animals, and the human outbreaks described above are a result of one or more "jumps" from these animal reservoirs into the human population. SARS is believed to have arrived in the human population from horseshoe bats via a masked palm civet intermediary.

Human-to-human transmission

The ability for different lineages to be transmitted between humans is extremely important to understand the potential development of an outbreak. Due to the ability of SARS to spread between humans and the high case fatality rate, SARS (or a SARS-like virus) is considered a global public health threat by the WHO.

Novel coronavirus (nCoV) 2019-2020

Further Reading:

- New China virus: Five questions scientists are asking Nature news *2020-01-22*
- China virus latest: first US case confirmed Nature news *2020-01-21*
- New virus surging in Asia rattles scientists Nature news *2020-01-20*
- New virus identified as likely cause of mystery illness in China Nature news *2020-01-08*

Recent outbreak of a novel coronavirus

In December 2019, a new illness was first detected in Wuhan, China. We now know this to be another outbreak of coronavirus in humans (the 7th), and it is provisionally being called nCoV (novel coronavirus).

As of January 23rd, 2020 over 624 cases and 17 deaths have been reported. It's still too early to know the case fatality rate, but early indications are that it is significantly less than SARS-CoV. The case counts are dramatically rising in part due to increased surveillance and testing, with 131 new cases reported on Jan 22nd, 2020.

While the outbreak seems to be centered in Wuhan, which is now under quarantine, the virus has spread throughout China and abroad, including Hong Kong, Macau, Thailand, USA, Japan and South Korea, but local transmission outside of China has yet to be reported.

The origin of the virus is still unclear, however genomic analyis suggests nCoV is most closely related to viruses previously identified in bats. It is plausible that there were other intermediate animal transmissions before the introduction into humans. There is no evidence of snakes as an intermediary.

Nextstrain narratives

The following pages contain analysis performed using Nextstrain. Scrolling through the left hand sidebar will reveal paragraphs of text with a corresponding visualization of the genomic data on the right hand side.

To have full genomes of a novel and large RNA virus this quickly is a remarkable achievement. These analyses have been made possible by the rapid and open sharing of genomic data and interpretations by scientists all around the world (see the final slide for a visualization of sequencing authorship).

How to interpret the phylogenetic trees

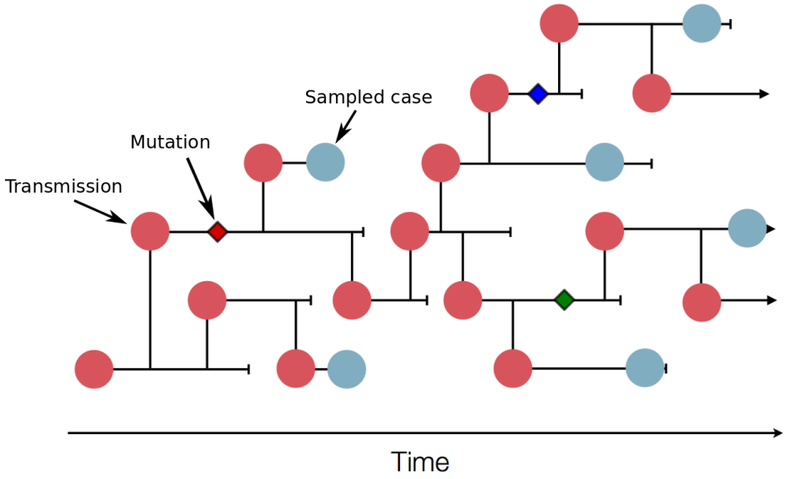
Further Reading:

- Exploring interactive phylogenies with Auspice *2019-01-24*

Transmission trees vs phylogenetic trees

Pathogens spread through rapid replication in one host followed by transmission to another host. An epidemic can only take off when one infection results in more than one subsequent infections.

As the pathogen replicates and spreads, its genome needs to be replicated many times and random mutations (copying mistakes) will accumulate in the genome. Such random mutations can help to track the spread of the pathogen and learn about its transmission routes and dynamics.

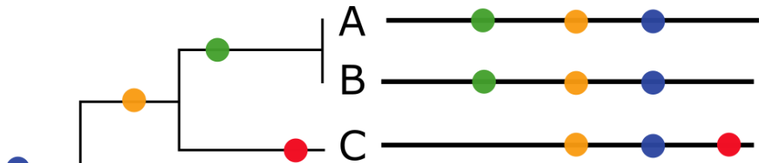


The illustration above shows a sketch of a transmission tree with a subset of cases that were sampled (blue). In practice, the transmission tree is unknown and typically only rough estimates of case counts are available. Genome sequences allow us to infer parts of the transmission tree. In this example, three mutations (little diamonds) are indicated on the tree. Sequences that have the same mutations are more closely related, so these mutations allow us to group samples into clusters of closely related viruses that belong to the same transmission chains.

Reading a Phylogenetic Tree

Below, we see an illustration with a phylogenetic tree on the left, where mutations are shown as colored circles. On the right are the corresponding sequences, also with mutations shown as colored circles. We can see that sequences that share the same mutations group together. When sequences appear linked by a flat vertical line, like A and B, this means there are no differences between them – their sequences are identical.

When a sequence sits on a long line on its own, like C or E, this means it has unique mutations not found in other sequences. The longer the line, the more mutations. A and B also have unique mutations (the green circle) not shared by the other sequences, but they are identical to each other.



Phylogenetic analysis

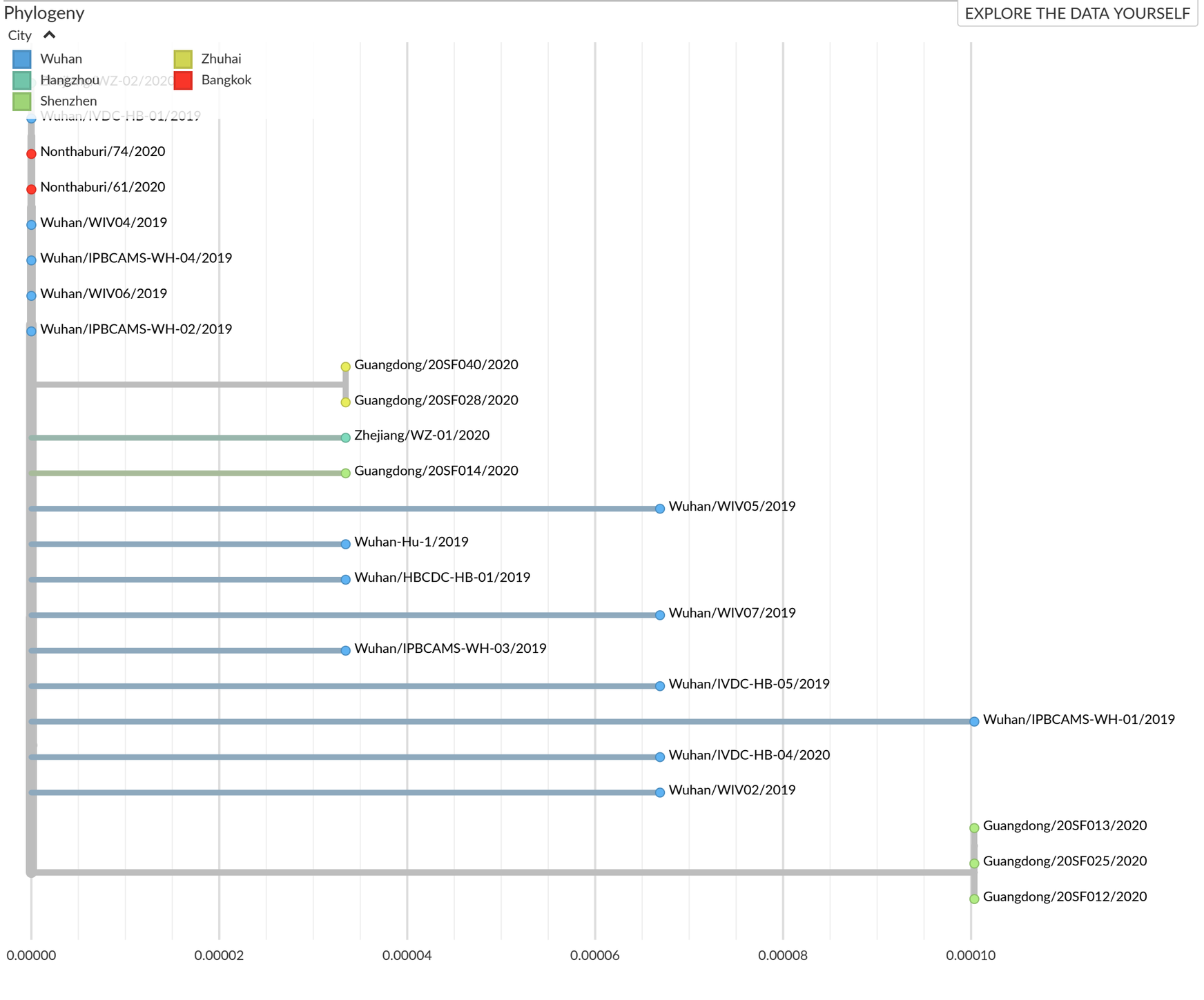
Here we present a phylogeny of 24 strains of nCoV that have been publicly shared. Information on how the analysis was performed is available in this GitHub repository.

The colours represent the city of isolation, with the x-axis representing nucleotide divergence.

Divergence is measured as the number of changes (mutations) per base. Since the nCoV genome is 29,000 bases long, one mutation corresponds to a divergence of 1/29,000 = 0.0000335.

Sequences that have just one mutation sit just to the left of the 0.00004 line on the x-axis.

Sequencing the genome of a large novel RNA virus in an evolving outbreak situation is challenging. Some of the differences observed in these sequences may be sequencing errors rather than actual mutations. Insertions, deletions, and differences at the ends of the genome are more likely to be errors and so we masked these for the purposes of this analysis.



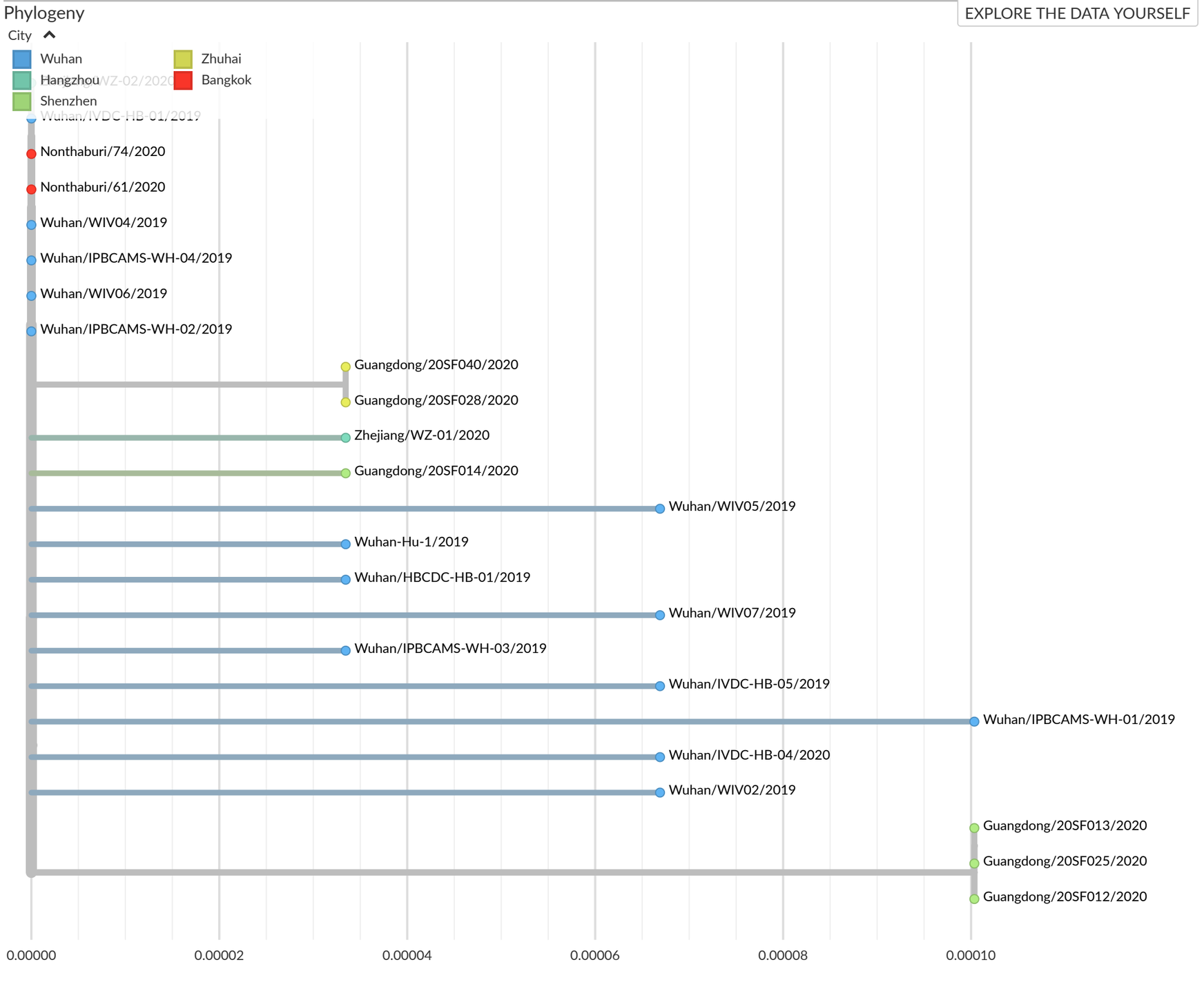
Phylogenetic Interpretation

We currently see little genetic diversity across the nCoV sequences, with 8 out of 24 sequences having no unique mutations.

Low genetic diversity across these sequences suggests that the most recent common ancestor of all nCoV sequences was fairly recent, since mutations generally accumulate slowly, around 1-2 mutations per month for coronaviruses. Generally, repeated introductions from an animal reservoir will show significant diversity (this has been true for Lassa, Ebola, MERS-CoV and avian flu). The observation of such strong clustering of human infections can be explained by an outbreak that descends from a single zoonotic introduction event into the human population followed by human-to-human epidemic spread.

At the moment, most mutations that can be observed are singletons – they are unique to individual genomes. Only the sequences that form the two

clusters from Guangdong share mutations – we will explore these in later slides.

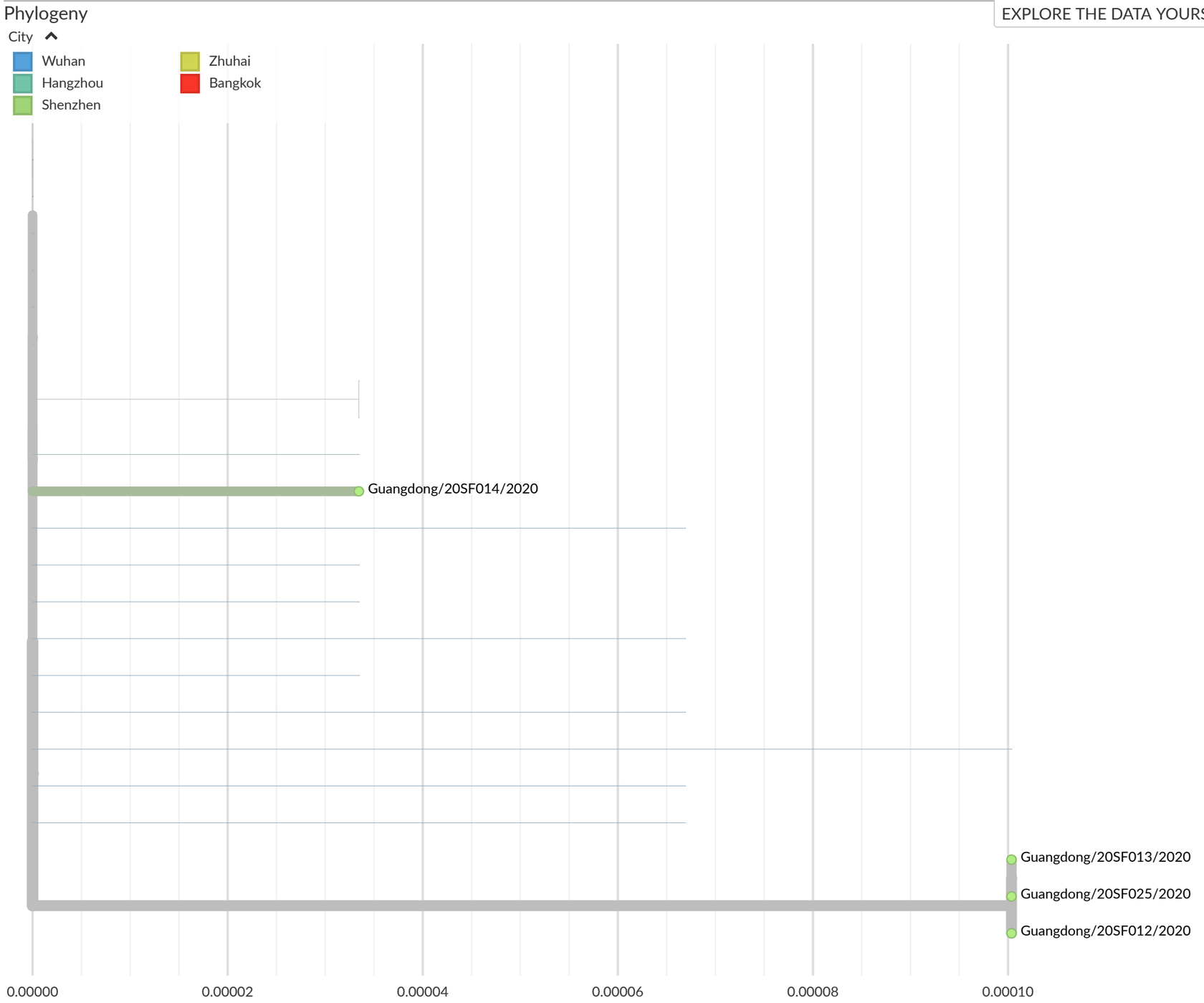


Potential within-family transmission 1

Of the four isolates from Shenzhen (Southeastern China, Guangdong Province) we see three isolates which are genetically identical and share three mutations unique to those three samples (you can hover your mouse over the branches to see which mutations are present).

These three samples are known to come from a single family, and almost certainly represent human-to-human transmission.

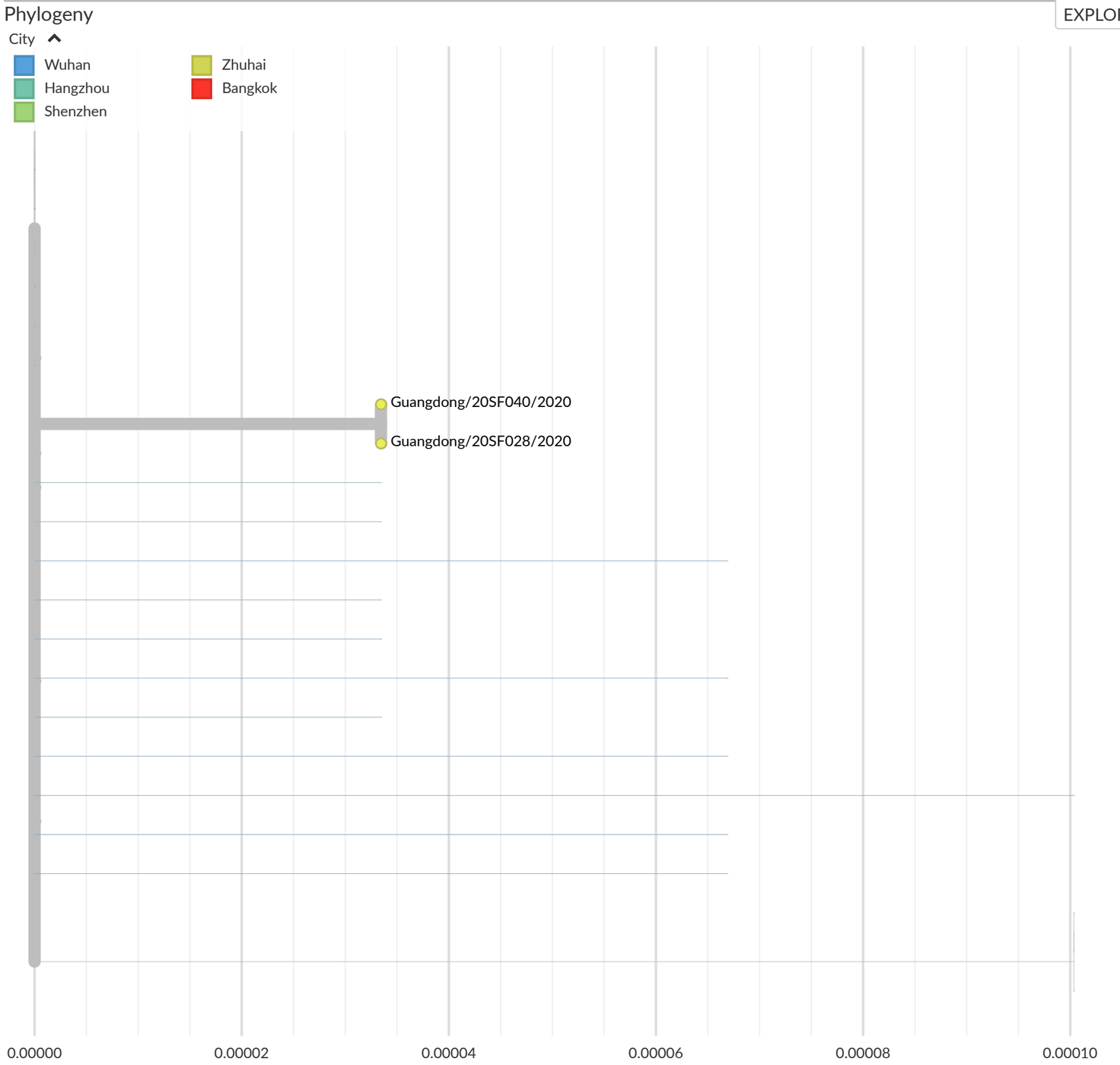
The fourth sample does not seem to be related to the other three, or to any other of the available sequences. Its genome has one mutation not seen in any other genome.



Potential within-family transmission 2

Similarly, there are two genetically-identical isolates from Zhuhai (Southeastern China, Guangdong Province) which form a cluster, sharing one unique mutation seen in no other isolate.

These two cases are also known to come from a single family, again indicating human-to-human transmission.

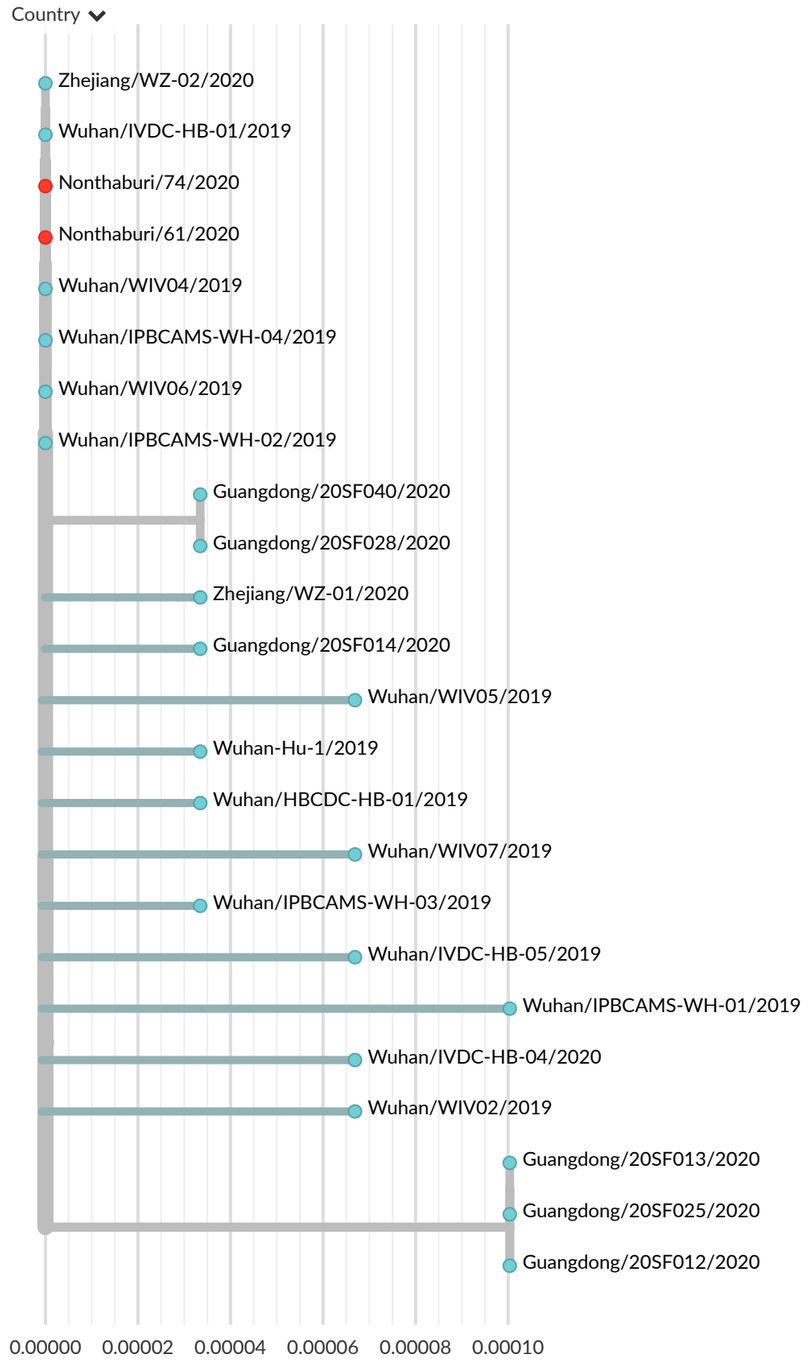


Cases outside China

There are reported diagnostically confirmed nCoV cases in Thailand, USA, Japan and South Korea. These cases are all linked to Wuhan, and we are not aware of evidence for local nCoV spread in these countries.

The only currently available sequence data for cases outside of China are the two cases from Thailand, which are coloured here in red. These samples are genetically identical to six Chinese sequences, including five isolated in Wuhan.

Phylogeny



Geography

EXPLORE THE DATA YOURSELF

Dating the time of the most recent common ancestor

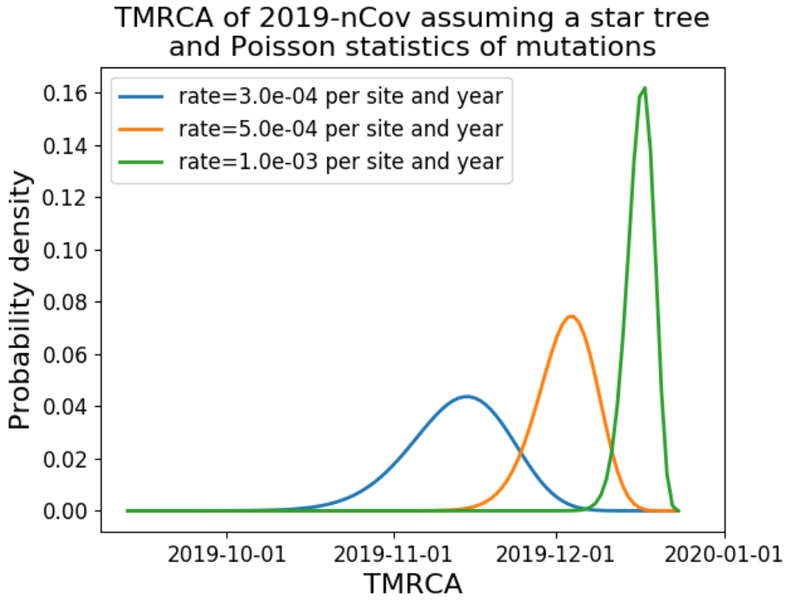
The high similarity of the genomes suggests they share a recent common ancestor (i.e. that they have descended from the same ancestral virus recently). Otherwise, we would expect a higher number of differences between the samples.

Previous research on related coronavirus suggests that these viruses accumulate between 1 and 3 changes in their genome per month (rates of 3×10^{-4} to 1×10^{-3} per site per year).

On the right, we explore how different assumptions about the rate of change, and the observed genetic diversity, give us estimates for the timing of the outbreak.

Date of the common ancestor of outbreak viruses

Here, we assume a star-like phylogeny structure along with a Poisson distribution of mutations through time to estimate the time of the most recent common ancestor ('TMRCA') of sequenced viruses. We find that the common ancestor most likely existed between mid-Nov and early-Dec 2019.



As the more samples are sequenced, we expect the tree to show more structure, such that the star-like phylogeny topology is no longer a good assumption. At this point, phylodynamic estimates of the age of the epidemic will become feasible.

Estimating the growth rate

An important quantity in the spread of a pathogen is the average number of secondary cases each infection produces.

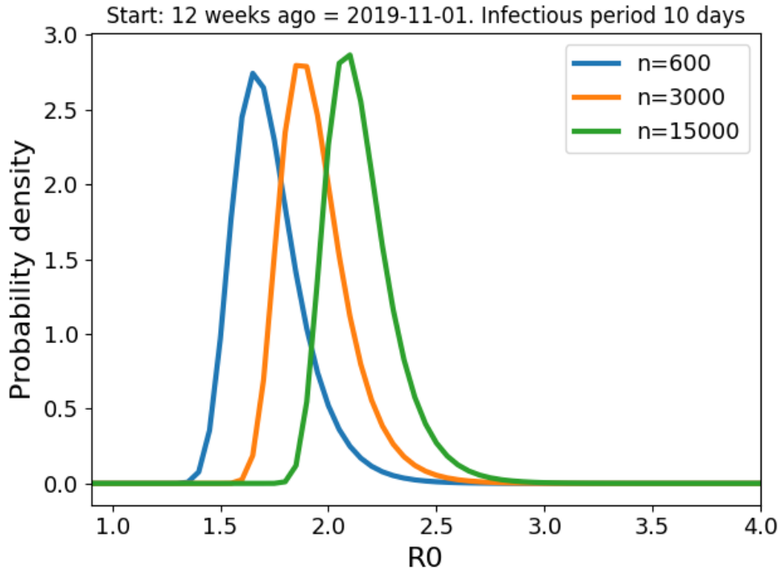
This number is known as R0 ("R-zero" or "R-nought"). One the right, we present simple estimates of R0.

Estimates of epidemic growth rate

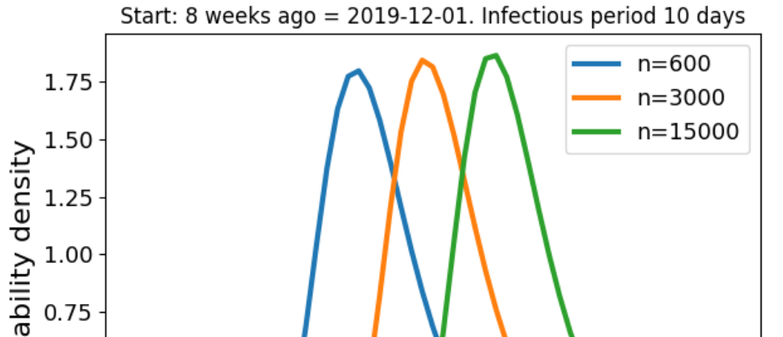
Scientists at Imperial College London have used the number of cases observed outside of China to estimate the total number of cases and suggested that there have been at least several thousand cases. Together with our previous estimates of the age of the outbreak and information on the infectious period, we can estimate plausible ranges of R0 using a branching process model.

We find plausible estimates of R0 between 1.5 and 3.5.

If we assume the outbreak started at the beginning of November 2019 (12 weeks ago), we find that R0 should range between 1.5 and 2.5, depending on how large ('n') the outbreak is now.



If we assume a more recent start, at the beginning of December 2019 (8 weeks ago), the estimates for R0 range between 1.8 and 3.5:



Scientific credit

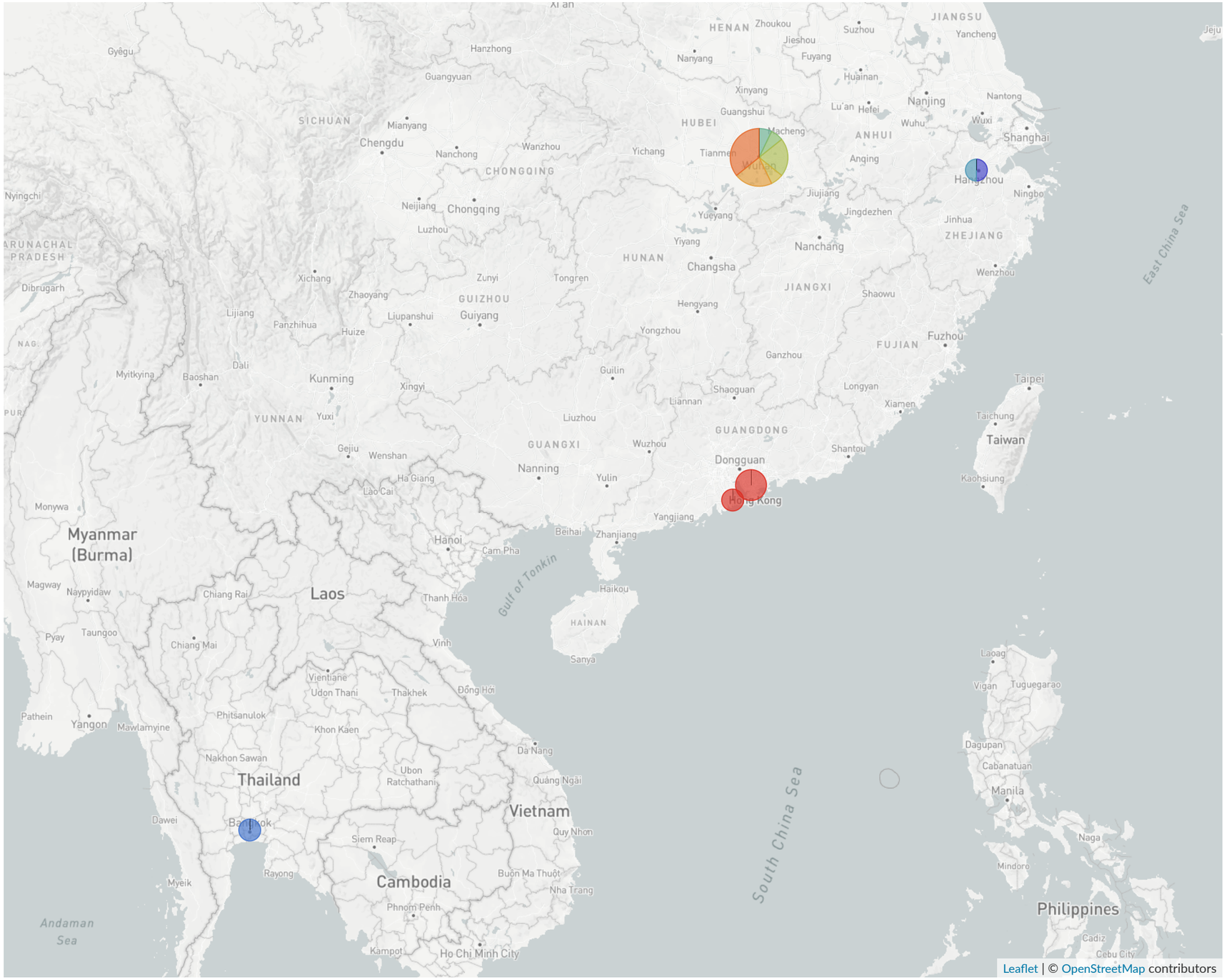
We would like to acknowledge the amazing and timely work done by all scientists involved in this outbreak, but particularly those working in China. Only through the rapid sharing of genomic data and metadata are analyses such as these possible.

The nCoV genomes were generously shared by scientists at the:

- Shanghai Public Health Clinical Center & School of Public Health, Fudan University, Shanghai, China
- National Institute for Viral Disease Control and Prevention, China CDC, Beijing, China
- Institute of Pathogen Biology, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing, China
- Wuhan Institute of Virology, Chinese Academy of Sciences, Wuhan, China
- Department of Microbiology, Zhejiang Provincial Center for Disease Control and Prevention, Hangzhou, China
- Guangdong Provincial Center for Diseases Control and Prevention
- Department of Medical Sciences, National Institute of Health, Nonthaburi, Thailand

Geography

EXPLORE THE DATA YOURSELF



Detailed scientific credit

These data were shared via GISAID. We gratefully acknowledge their contributions.

To the right we give specific sequences shared by each lab.

The nCoV genomes were generously shared by scientists at the

- Shanghai Public Health Clinical Center & School of Public Health, Fudan University, Shanghai, China
  - Wuhan-Hu-1/2019
- National Institute for Viral Disease Control and Prevention, China CDC, Beijing, China
  - Wuhan/IVDC-HB-01/2019
  - Wuhan/IVDC-HB-04/2020
  - Wuhan/IVDC-HB-05/2019)
- Institute of Pathogen Biology, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing, China
  - Wuhan/IPBCAMS-WH-01/2019
  - Wuhan/IPBCAMS-WH-02/2019
  - Wuhan/IPBCAMS-WH-03/2019
  - Wuhan/IPBCAMS-WH-04/2019
- Wuhan Institute of Virology, Chinese Academy of Sciences, Wuhan, China
  - Wuhan/WIV02/2019
  - Wuhan/WIV04/2019
  - Wuhan/WIV05/2019
  - Wuhan/WIV06/2019
  - Wuhan/WIV07/2019
- Department of Microbiology, Zhejiang Provincial Center for Disease Control and Prevention, Hangzhou, China
  - Zhejiang/WZ-01/2020
  - Zhejiang/WZ-02/2020
- Guangdong Provincial Center for Diseases Control and Prevention
  - Guangdong/20SF012/2020
  - Guangdong/20SF013/2020
  - Guangdong/20SF014/2020
  - Guangdong/20SF025/2020
  - Guangdong/20SF028/2020
  - Guangdong/20SF040/2020
- Department of Medical Sciences, National Institute of Health, Nonthaburi, Thailand
  - Nonthaburi/61/2020
  - Nonthaburi/74/2020

END OF NARRATIVE

Scroll back to the beginning

Leave the narrative & explore the data yourself

The nCoV genomes were generously shared by scientists at the

- Shanghai Public Health Clinical Center & School of Public Health, Fudan University, Shanghai, China
  - Wuhan-Hu-1/2019
- National Institute for Viral Disease Control and Prevention, China CDC, Beijing, China
  - Wuhan/IVDC-HB-01/2019
  - Wuhan/IVDC-HB-04/2020
  - Wuhan/IVDC-HB-05/2019)
- Institute of Pathogen Biology, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing, China
  - Wuhan/IPBCAMS-WH-01/2019
  - Wuhan/IPBCAMS-WH-02/2019
  - Wuhan/IPBCAMS-WH-03/2019
  - Wuhan/IPBCAMS-WH-04/2019
- Wuhan Institute of Virology, Chinese Academy of Sciences, Wuhan, China
  - Wuhan/WIV02/2019
  - Wuhan/WIV04/2019
  - Wuhan/WIV05/2019
  - Wuhan/WIV06/2019
  - Wuhan/WIV07/2019
- Department of Microbiology, Zhejiang Provincial Center for Disease Control and Prevention, Hangzhou, China
  - Zhejiang/WZ-01/2020
  - Zhejiang/WZ-02/2020
- Guangdong Provincial Center for Diseases Control and Prevention
  - Guangdong/20SF012/2020
  - Guangdong/20SF013/2020
  - Guangdong/20SF014/2020
  - Guangdong/20SF025/2020
  - Guangdong/20SF028/2020
  - Guangdong/20SF040/2020
- Department of Medical Sciences, National Institute of Health, Nonthaburi, Thailand
  - Nonthaburi/61/2020
  - Nonthaburi/74/2020