ncov-phylodynamics

Phylodynamic estimation of nCoV incidence and prevalence

Source code

github.com/blab/ncov-phylodynamics

Contributors

trvrb

Latest commits

09 Feb 2020 - Include Riou reference
09 Feb 2020 - Update Markdown
09 Feb 2020 - Fix markdown
09 Feb 2020 - Initial commit, containing data, BEAST XML, results and Mathematica notebook

Pages

/

Phylodynamic estimation of incidence and prevalence of novel coronavirus (nCoV) infections through time

Trevor Bedford[1]

[1]Vaccine and Infectious Disease Division, Fred Hutchinson Cancer Research Center, Seattle, WA, USA

Last updated Feb 9, 2020.

Abstract

Here, we use a phylodynamic approach incorporating 53 publicly available novel coronavirus (nCoV) genomes to the estimate underlying incidence and prevalence of the epidemic. This approach uses estimates of the rate of coalescence through time to infer underlying viral population size and then uses assumptions of serial interval and heterogeneity of transmission to provide estimates of incidence and prevalence. We estimate an exponential doubling time of 7.2 (95% CI 5.0-12.9) days. We arrive at a median estimate of the total cumulative number of worldwide infections as of Feb 8, 2020, of 55,800 with a 95\% uncertainty interval of 17,500 to 194,400. Importantly, this approach uses genome data from local and international cases and does not rely on case reporting within China.

Data

Here, we use 53 publicly available nCoV genomes collected between 24 Dec, 2019 and 4 Feb, 2020. These represent cases sequenced from all over the world. There are genomes available from clusters that we did not incorporate here as they are expected to interfere with phylodynamic analysis (coalescent models assume that infections are sampled randomly from the infected population). For this analysis, we removed all but one of each reported cluster. In this case, each international sample can be considered a direct export from the epidemic within China and we can ignore most spatial considerations.

Methods

Here, we followed Andrew Rambaut's work on Virological.org and use the software package BEAST v1.10.4 to infer viral evolutionary dynamics. We began by running the Nextstrain nCov pipeline to align sequences and mask spurious SNPs. We took the output file `masked.fasta` as the starting point for this analysis. We loaded this alignment into BEAST and specified an evolutionary model to estimate:

- strict molecular clock (CTMC rate reference prior)
- exponential growth rate (Laplace prior with scale 100)
- effective population size at time of most recent sampled tip (uniform prior between 0 and 10)

We followed Andrew in using a gamma distributed HKY nucleotide substitution model. MCMC was run for 50M steps, discarding the first 10M as burnin and sampling every 30,000 steps after this to give a dataset of 1335 MCMC samples.

The file `ncov.xml` contains the entire BEAST model specification. To run it will require filling in sequence data; we are not allowed to reshare this data according to GISAID Terms and Conditions. The Mathematica notebook `ncov-phylodynamics.nb` contains code to analyze resulting BEAST output in `ncov.log` and plot figures.

Results

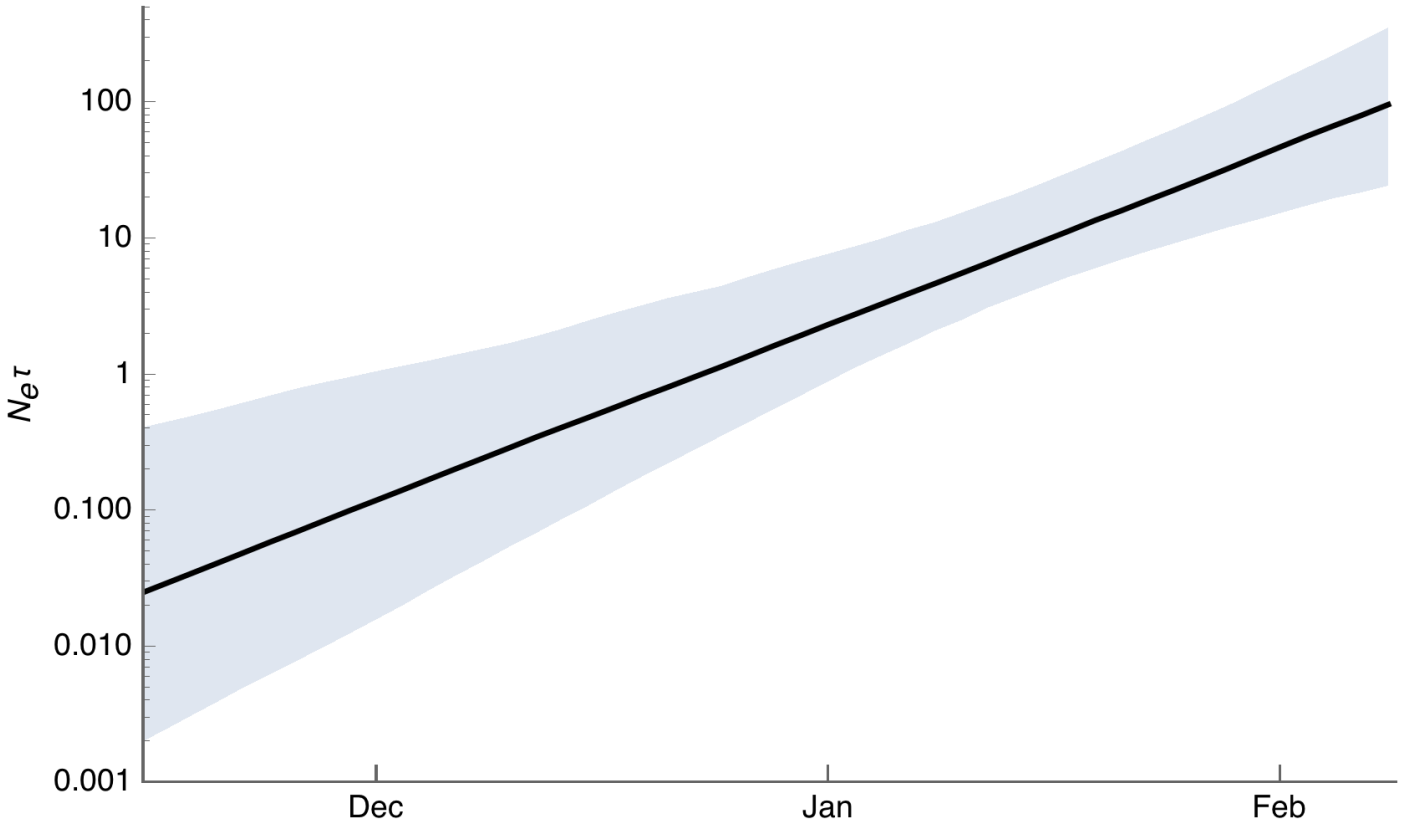
Rate and TMRCA

We find substitution rate consistent with previous work of 0.9×10^{-3} (95% CI $0.5\text{-}1.4 \times 10^{-3}$) substitutions per site per year. We find a median TMRCA of 3 Dec (95% CI 30 Oct to 17 Dec).

Effective population size and exponential growth rate

These phylodynamic approaches can estimate effective size of the virus population by examining rates of coalescence through time. Here, we estimated the exponential growth rate as 35.4 (95% CI 9.6-50.0) per year. This translates to a doubling time of 7.2 (95% CI 5.0-12.9) days. This coincides closely with doubling time reported by modeling groups looking at reported cases in China (Wu et al).

Here, we plot timescale of coalescence $N_e\tau$ through time:



$N_e\tau$ is what is directly measured by phylodynamic methods and is measured in years. Here, it can be seen that coalescence is slowing down, so that pairs of lineages on 1 Feb coalesce at a rate of 1 event 10+ years.

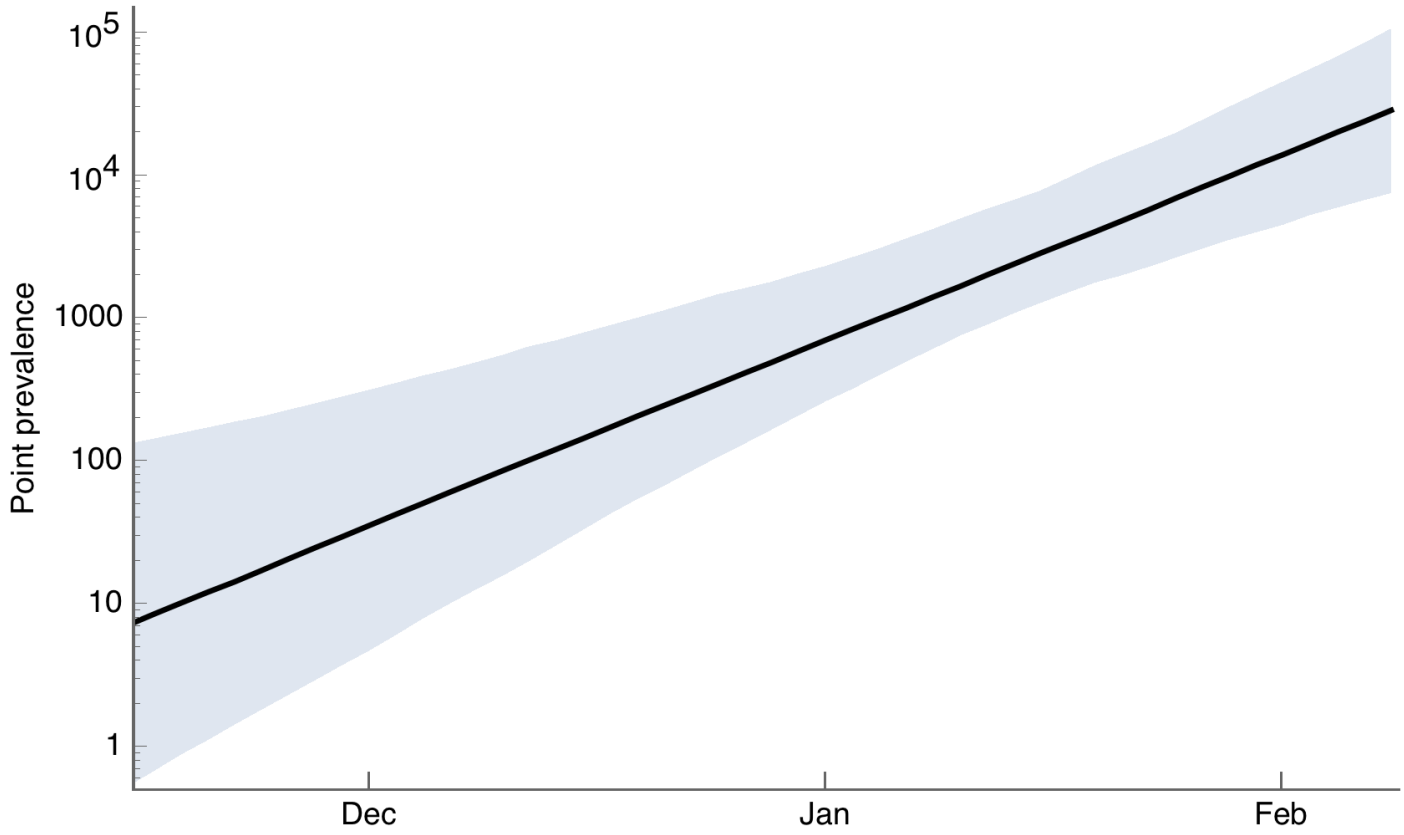
Prevalence

The quantity of $N_e\tau$ can be translated into effective population size N_e by dividing out generation time τ. We assume generation time τ to be 7.5 days following Li et al. Additionally, effective population size N_e can be translated into prevalence with knowledge of the variance in offspring distribution. High variance in distribution of secondary cases reduces prevalence relative to N_e as described by Volz et al. This reduction is equal to

$$\sigma^2 = \frac{1}{E[R_0]} + \frac{1}{k} + 1,$$

where $E[R_0]$ is the mean number of secondary cases and k is the dispersion parameter of secondary cases. We assume $E[R_0]$ to be between 1.8 and 2.8 following Wu et al and others. We assume that variance of secondary cases is at most like SARS with superspreading dynamics with $k = 0.15$, but allow for less variance with $k = 0.30$ (Riou and Althaus). Thus, we convert BEAST estimates of $N_e\tau$ to point prevalence I by following $I = N_e\tau \times \sigma^2/\tau$.

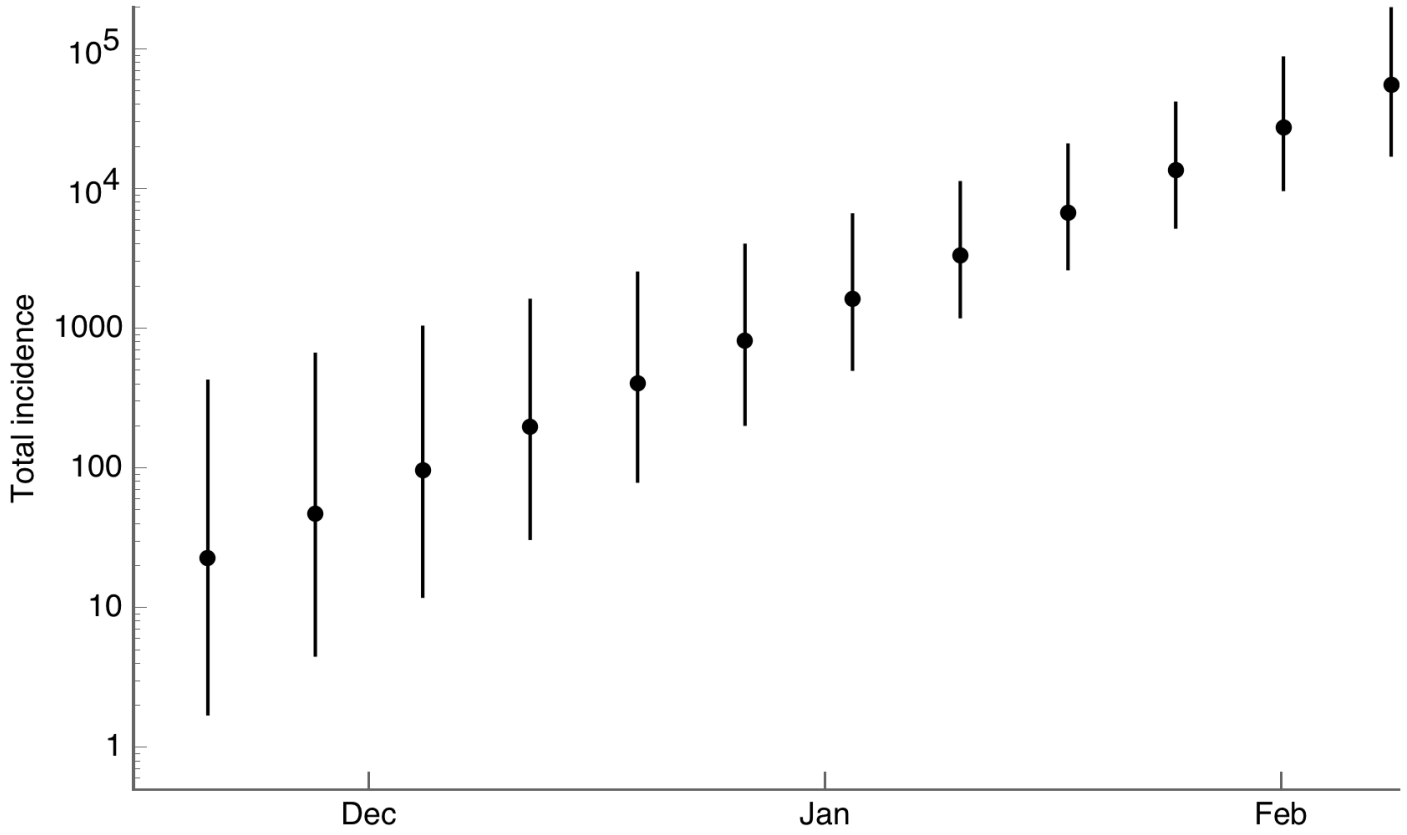
We arrive at the following estimate of prevalence through time:



We estimate a median prevalence on 8 Feb of 28,500 currently infected with a 95% uncertainty interval of between 7500 and 104,300 currently infected.

Total incidence

We estimate incidence in each serial interval and then calculate a cumulative incidence total:



We estimate a median total incidence on 8 Feb of 55,800 total infections since start of epidemic with a 95% uncertainty interval of between 17,500 and 194,400 total infections. On Feb 8, there were 34,886 total cases reported (WHO Sit Rep 19). Importantly, this approach uses genome data from local and international cases and does not rely on case reporting within China.

Our phylodynamic approach estimates an infection-to-case reporting rate of between 18% and 100%. Although, there are obviously wide uncertainty intervals to these estimates, we believe it is safe to conclude that case reporting is largely in line with expectations given spectrum of disease severity. These numbers are consistent with WHO reports of mild cases constituting 82% of the epidemic.

Acknowledgements

The nCoV genomes used in this analysis were generously shared by scientists at the Shanghai Public Health Clinical Center & School of Public Health, Fudan University, Shanghai, China, at the National Institute for Viral Disease Control and Prevention, China CDC, Beijing, China, at the Institute of Pathogen Biology, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing, China, at the Wuhan Institute of Virology, Chinese Academy of Sciences, Wuhan, China, at the Department of Microbiology, Zhejiang Provincial Center for Disease Control and Prevention, Hangzhou, China, at the Guangdong Provincial Center for Diseases Control and Prevention at the Department of Medical Sciences, at the Shenzhen Key Laboratory of Pathogen and Immunity, Shenzhen, China, at the Hangzhou Center for Disease and Control Microbiology Lab, Zhejiang, China, at the National Institute of Health, Nonthaburi, Thailand, at the National Institute of Infectious Diseases, Tokyo, Japan, at the Korea Centers for Disease Control & Prevention, Cheongju, Korea, at the National Public Health Laboratory, Singapore, at the US Centers for Disease Control and Prevention, Atlanta, USA, at the Institut Pasteur, Paris, France, at the Respiratory Virus Unit, Microbiology Services Colindale, Public Health England, and at the Department of Virology, University of Helsinki and Helsinki University Hospital, Helsinki, Finland, and at the University of Melbourne, Peter Doherty Institute for Infection

and Immunity, Melbourne, Australia, at the Victorian Infectious Disease Reference Laboratory, Melbourne, Australia, at the Public Health Virology Laboratory, Brisbane, Australia and at the Institute of Clinical Pathology and Medical Research, University of Sydney, Westmead, Australia via GISAID. We gratefully acknowledge the Authors, Originating and Submitting laboratories of the genetic sequence and metadata made available through GISAID on which this research is based.

Provenance of originating labs, submitting labs and authors is available here.