PRE-PRINT – Not Peer Reviewed

22-02-2020

The effect of cordon sanitaire on the COVID-19 epidemic in China

One sentence summary: Human mobility from Wuhan predicts epidemic establishment and size across China.

Moritz U.G. Kraemer[1,2,3,$], Chia-Hung Yang[4], Bernardo Gutierrez[1,5], Chieh-Hsi Wu[6], Brennan Klein[4], David M. Pigott[7], open COVID-19 data working group*, Louis du Plessis[1], Nuno R. Faria[1], Ruoran Li[8], William P. Hanage[8], John S. Brownstein[2,3], Maylis Layan[9], Alessandro Vespignani[4], Huaiyu Tian[10], Christopher Dye[1], Simon Cauchemez[9], Oliver G. Pybus[1,$], Samuel V. Scarpino[4,$]

1. Department of Zoology, University of Oxford, United Kingdom
2. Harvard Medical School, Harvard University, Boston, United States
3. Boston Children's Hospital, Boston, United States
4. Network Science Institute, Northeastern University, Boston, United States
5. School of Biological and Environmental Sciences, Universidad San Francisco de Quito USFQ, Quito, Ecuador
6. Mathematical Sciences, University of Southampton, Southampton, United Kingdom
7. Institute for Health Metrics and Evaluation, Department of Health Metrics, University of Washington, Seattle, United States
8. Harvard T.H. Chan School of Public Health, Boston, United States
9. Mathematical Modelling of Infectious Diseases Unit, Institut Pasteur, UMR2000, CNRS, Paris, France
10. State Key Laboratory of Remote Sensing Science, College of Global Change and Earth System Science, Beijing Normal University, Beijing, China

[$]Correspondence should be addressed to: s.scarpino@northeastern.edu; oliver.pybus@zoo.ox.ac.uk; moritz.kraemer@zoo.ox.ac.uk

*members of the open COVID-19 data working group are listed at the end of the manuscript

Abstract

The ongoing COVID-19 outbreak has rapidly expanded throughout China. Major behavioral, clinical, and state interventions are currently underway to mitigate the epidemic and prevent the persistence of the virus in human populations in China and globally. It remains unclear how these unprecedented interventions, including travel restrictions, have affected COVID-19 spread in China. We use real-time mobility data from Wuhan and detailed case data including travel history to elucidate the role of case importation on transmission in cities across China. We find that travelers spread the virus across China before the cordon sanitaire was implemented and that human mobility out of Wuhan predicts the subsequent case distribution across China. In light of these findings, local public health responses focusing on identifying new importations will be crucial for COVID-19 control and elimination.

Main text

The outbreak of COVID-19 has spread rapidly from its origin in Wuhan, Hubei Province, China (*1*). A range of interventions have been implemented following the detection in late December 2019 of a cluster of pneumonia cases of unknown etiology, and identification of the causative virus SARS-CoV-2 in early January 2020 (*2*). Interventions include improved rates of diagnostic testing, clinical management, rapid isolation of suspected and confirmed cases and, most notably, restrictions on mobility (hereafter called cordon sanitaire) imposed on Wuhan city on 23rd January. Travel restrictions were subsequently imposed on 14 other cities across Hubei Province and partial movement restrictions have been enacted in many cities across China. Initial analysis suggests that the Wuhan cordon sanitaire resulted in an average delay of COVID-19 spread to other cities of 3 days (*3*), but the true extent of the effect of the mobility restrictions on transmission has not been examined in detail (*4*, *5*). Questions remain over how the timing of these interventions affected the spread of SARS-CoV-2 to locations outside of Wuhan. We here use real-time mobility data, crowdsourced line-list data of cases with reported travel history, and timelines of reporting changes to identify early shifts in the epidemiological dynamics of the COVID-19 epidemic in

China, from an outbreak driven by importations from Wuhan to a locally-sustained epidemic across provinces in China.

Transmission from Wuhan drove geographic expansion before the cordon sanitaire

To identify accurately a timeframe for evaluating early changes in SARS-CoV-2 transmission in China, we first estimated from case data the average incubation period of COVID-19 infection (i.e. the duration between time of infection and symptom onset (*6*, *7*)). Since infection events are typically not directly observed, we estimate incubation period from the span of exposure during which infection likely occurred. Using detailed information on 38 cases for whom both the dates of entry to and exit from Wuhan are known, we estimate the mean incubation period to be 5.1 days (std. dev. = 3.0 days; **Fig. S1**), similar to previous estimates from other data (*8*, *9*). In subsequent analyses we add an upper estimate of one incubation period (mean + 1 standard deviation = 8 days) to the date of Wuhan shutdown, in order to delineate the date before which cases recorded in other provinces might represent infections acquired in Hubei (i.e. 31st January 2020; **Fig. 1a**). After that, only a small number of travel-related infections are expected, due to limited mobility from Wuhan (**Fig. 3a**).

As of 16th February 2020, 68584 cases of COVID-19 have been reported worldwide (**Fig. 1a**) (*10*). Case reports were mostly restricted to Hubei until 23rd January 2020 (81% of all cases), after which most provinces reported rapid increases in cases (**Fig. 1a**). Many early cases (before 23rd January 2020) reported outside of Wuhan had known travel history to Wuhan (57%; **Fig. 1b**) and were distributed across China, confirming the importance of Wuhan as the major source. Among cases known to have traveled from Wuhan before 23rd Jan, the time from symptom onset to confirmation was 6.5 days (SD: 4.2, **Fig. S2**), providing opportunity for onward transmission at the destination. More active surveillance brought this interval down to 4.8 days (SD: 3.03, **Fig. S2**) for those who travelled after 23rd Jan.

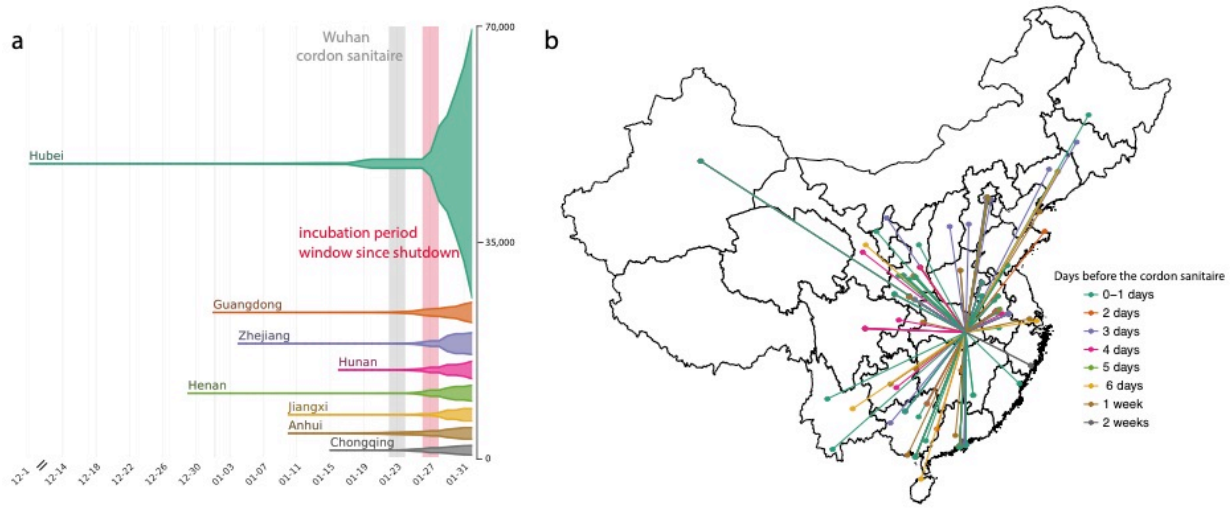

Figure 1: Number of cases and key dates during the epidemic. *(a) The epidemic curve of the COVID-19 outbreak in provinces in China. Vertical lines and boxes indicate key dates such as implementation of cordon sanitaire of Wuhan (grey) and the end of the first incubation period after the travel restrictions (red). The thin grey line represents the closure of Wuhan seafood market on 1st January 2020. The width of each horizontal tube represents the number of reported cases in that province. (b) Map of COVID-19 confirmed cases (n = 554) that had reported travel history from Wuhan before travel restrictions were implemented on January 23, 2020. Colors of the arrows indicate date of travel relative to the date of travel restrictions.*

Age and sex distributions can reflect heterogeneities in the risk of infection within affected populations. To investigate meaningful shifts in the epidemiology of the COVID-19 outbreak through time, we examined age and sex data for cases from different periods of the outbreak, and from individuals with and without travel from Wuhan. However, details of travel history exist for only a fraction of confirmed cases and this information is particularly scant for some provinces (e.g. Zhejiang and Guangdong). Consequently, we grouped confirmed cases into four categories: (*I*) early cases with travel history (early = reported before 1st Feb), (*II*) early cases without travel history, (*III*) later cases with travel history (later = reported between 1st – 10th Feb), (*IV*) later cases without travel history.

Using crowdsourced case data, we found that cases with travel history (categories *I* and *III*) had similar median ages and sex ratios in both the early and later phases of the outbreak (41 vs 42 years old, respectively; 1.47 vs. 1.45 males per female, respectively; **Fig. 2d**). After January 31st, however, the sex ratio of cases *without* reported travel history (category *IV*) shifted dramatically to approximately 1:1 (57 male vs. 62 female) as expected under a null hypothesis of equal transmission risk **(Fig. 2a,b,d,** see also reference (*11*, *12*), **Materials and Methods)** and the median age increased to 46 (**Figs. 2a,b,c**). Early cases with no information on travel history (category *II)* had a similar median age and sex ratio to those with known travel history (42 years old and 1.80 males per female, **Fig. 2d**). We hypothesize that many of the cases with no known travel history in the early period were indeed travelers that contributed to disseminating SARS-CoV-2 outside of Wuhan. The shift towards more equal sex ratios and older ages after Jan 31st can be explained by an early phase dominated by infections (typically younger, biased towards males) related to travel out of Wuhan, followed by a later phase dominated by local transmission in the general population.

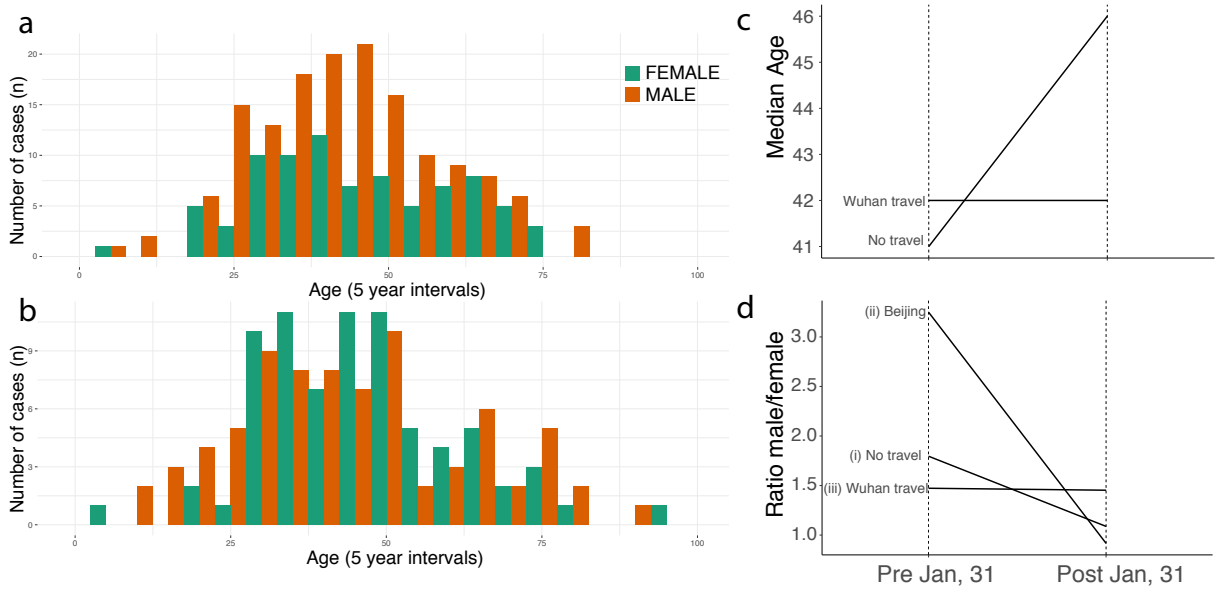

Figure 2: Shifting age and sex distributions through time. *(a) Age and sex distributions of confirmed cases with known travel history to Wuhan, (b) Age and sex distributions of confirmed cases that had no*

travel history. (c) Difference in mean age between cases reported early (before 1st Feb) and those reported later (between 1st – 10th February. (d) Change through time in the sex ratio of (i) all reported cases in China with no reported travel history, (ii) cases reported in Beijing without travel history, and (iii) cases known to have travelled from Wuhan.

Given the increasing evidence that most COVID-19 infections may be characterized by relatively mild symptoms it is plausible to assume that, in addition to nosocomial transmission (*13*), transmission occurs in the community before hospitalization. Using data on travel history and reported date of onset of symptoms we find that 22% of domestic travelers in China had onset of symptoms before they travelled (for international travelers, including travel to Hong Kong, Macao and Taiwan, the value was 20%; **Fig. S1)**. These data are consistent with the hypothesis that a substantial fraction of infectious individuals have mild disease.

Evidence for establishment of local transmission across China

In order to understand whether the volume of travel within China can predict the epidemic outside of Wuhan, we analyzed real-time human mobility data from Baidu, together with epidemiological data from each province. Ongoing local transmission of COVID-19 was first reported in Guangzhou, Guangdong province (*14*), in a family cluster. We investigated spatio-temporal disease spread to elucidate the relative contribution of Wuhan to transmission elsewhere and evaluate how the cordon sanitaire may have impacted it.

Among all cases reported outside Hubei province, we observe 515 cases with known travel history to Wuhan and a confirmation date before January 31st, compared with only 39 after February 1st, 2020, illustrating the effect of travel restrictions **(Figs. 1b, 3a, Fig. S3)**. We confirm the expected decline of importation with real-time human mobility data from Baidu Inc.. Travel out of Wuhan was significantly reduced compared to the time period 2019 (**Figs. 3a, b**). Movements of individuals out of Wuhan

increased in the days before the Lunar New Year and the establishment of the cordon sanitaire, before rapidly decreasing to effectively no movement **(Figs. 3a,b)**. The travel ban appears to have prevented travel in and out of Wuhan around the time of the Lunar New Year celebration (**Fig. 3a**) and likely prevented further dissemination of SARS-CoV-2 from Wuhan.

To test the contribution of the epidemic in Wuhan to seeding epidemics elsewhere we build a naïve COVID-19 GLM (*15*) model of exponential growth in daily case counts, assuming previously reported doubling times of cases of 3 - 7 days (*5*, *9*) which correspond to R0 values of 1.4 - 3.8, and a mean serial interval of 6 – 16 days (*5*, *9*, *16*, *17*). Such a model can predict daily case counts across all provinces with relatively high accuracy (as measured with R^2) throughout the epidemic, and when accounting for human mobility. We also find that the correlation increases one incubation period after the shutdown (**Figs. 3c,d**) indicating that human mobility contributed to the successful establishment of COVID-19.

We find the magnitude of the epidemic (total number of cases until February 10, 2020) outside of Wuhan is remarkably well predicted by the volume of human movement out of Wuhan ($R^2 = 0.87$; Spearman's rank correlation rho = 0.94, 95% CI: 0.82 – 0.98, p-value < 0.001, **Fig. 4a**). Therefore cases exported from Wuhan prior to the cordon sanitaire appear to have not only contributed to initiating local chains of transmission, both in neighboring provinces such as Henan, and in comparatively more distant provinces, e.g., Guangdong and Zhejiang **(Figs. 1a, 3b)** but that the frequency of introductions from Wuhan are also predictive of the size of the epidemic in other provinces (controlling for population size).

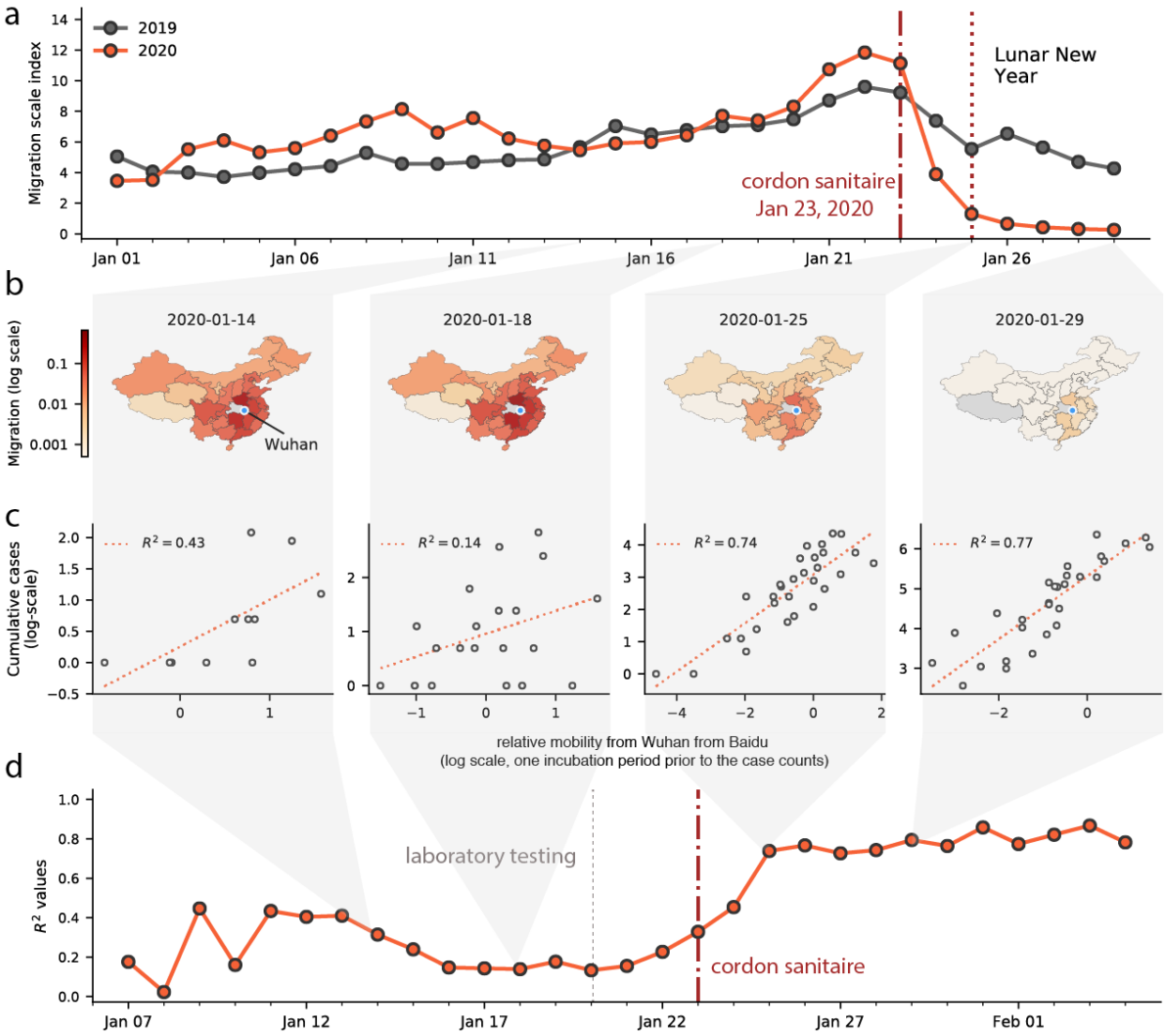

Figure 3: Human mobility, spread and synchrony of COVID-19 outbreak in China. *a) Human mobility data extracted in real time from Baidu. Date of start of travel ban of movements out of Wuhan on January 23,2020. Dark and red lines represent migration scale indexes for 2019 and 2020, respectively. b) Geographic locations of movements across Chinese provinces. c) Correlation between epidemic curve in Wuhan and epidemic curves in all other provinces weighted by human mobility from Wuhan. Analyses were also performed on daily case counts (see Table S1 and Table S2). d) Timeline of daily correlation*

between incidence in Wuhan and incidence in all other provinces weighted by human mobility. Dashed grey line indicates the date when large scale laboratory testing started in provinces across China.

To evaluate the relative contribution of introductions from Wuhan we estimate the relative growth rates of the epidemic in all other provinces (**Materials and Methods**). Interestingly, we find that all provinces outside Wuhan experienced high epidemic growth rates between January 24 – February 3rd, 2020 relative to Wuhan (**Fig. 4c, Fig. S4b**). The epidemic growth rates (relative to the estimated epidemic growth in Wuhan (*5*, *18*)) in each province spikes between January 27-29th, consistent with one incubation period of the virus since the shutdown of Wuhan (**Fig. 4b, Fig. S6**); this delay is also evident from the cumulative case counts by province (**Fig. S6**). In the same period, we show that the variation in the growth rate is almost entirely explained by human movements from Wuhan **(Fig. S4**), and consistent with theory of infectious disease spread in highly coupled metapopulations (*19*, *20*). We thereby identify the mechanism of spread of COVID-19 where each successful introduction sparks independent transmission chains and that the frequency of introductions from Wuhan dictate the probability of establishment and size of the epidemics across provinces in China. After February 2nd, and the effective stop of seeding from Wuhan, epidemics follow doubling times in Wuhan and are increasingly synchronized in their behaviour (**Supplementary Fig. 6**, (*21*, *22*)) and counterintuitively, due to the effective decoupling of metapopulations, global elimination is less likely (*19*, *23*). Early in an epidemic, observations may be biased towards identifying large clusters which can inflate the growth rate estimation.

Laboratory testing and human mobility interact in predicting the epidemic

Early during an emerging epidemic, laboratory testing capacity can vary significantly. To understand how testing differentially impacts our model performance, we test the hypothesis that changes in human mobility interact with rapid shifts in testing capacity, which started around January 20th, 2020 (**Materials and Methods**). The inclusion of mobility data from Wuhan and an indicator for the incubation period of cases leaving Wuhan prior to the shutdown produces a significant improvement (delta-BIC > 90, (*24*))

over the naive model that considers only a doubling in cases every 3-7 days **(Fig. 4b)**. However, prior to 20th January, 2020, we note that a sizeable overprediction in the number of cases can be explained by testing capacity which is consistent with reports that cases were underreported in that period. This signal becomes even more clear when considering the heterogeneity among provinces. We plot the relative improvement in the model fit (based on normalized residual error) for a model with lagged-cases and mobility from Wuhan and a model with lagged cases and testing availability (see more details in **Materials and Methods**). Of the 27 provinces in China reporting cases through February 6th, 2020, we find that in 12 provinces the largest improvements in prediction can be achieved using mobility only (**Supplementary Fig. 5**). In 10 provinces, both testing and mobility improve the model and in only one province testing is the most important factor improving the model (**Fig. S5**). We conclude that accounting for testing during the early phase of the epidemic is critical, however, mobility out of Wuhan remains the main driver of spread prior to the shutdown.

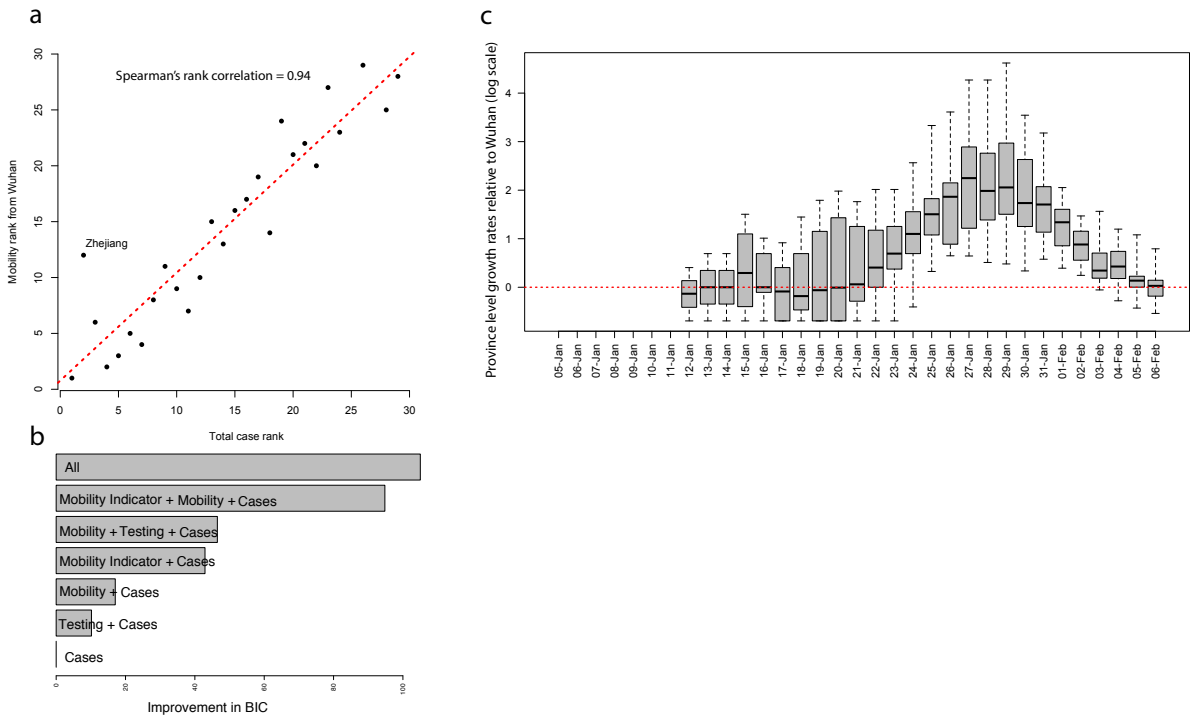

Figure 4: Total number of cases are well predicted by human mobility. (*a) Ranked total number of cases against rank of mobility from Wuhan. (b) Total improvement in model fit as measured by Bayesian Information Criteria (BIC), where BIC >4 was considered the cutoff for substantial model improvement, for models that included lagged cases (one incubation period), testing capacity as reported by the Chinese CDC, and human mobility out of Wuhan using Baidu data. Detailed sensitivity analyses and model selection using regularized regression confirm these results (see Supplementary Information). c) Time series of growth rates of the COVID-19 epidemic across provinces in China relative to the epidemic in Wuhan.*

Discussion

Containment of respiratory infections is particularly difficult if they are characterized by relatively mild symptoms or transmission before the onset of disease (*25*). Intensive restrictions on movement have been suggested as a method to limit the geographic spread of COVID-19 from regions where transmission is established. Using multiple data sources, we conclude that the period before hospitalization, particularly among travelers, was a crucial factor in the geographic expansion of the outbreak. Further, the volume of infected travelers appears to explain the size of the epidemic in secondary locations. In light of these findings, local public health responses focusing on identifying new importations will be crucial for COVID-19 control and elimination.

References:

1. S. Chen, J. Yang, W. Yang, C. Wang, T. Baerninghausen, COVID-19 control in China during mass population movements at New Year. *Lancet* (2020), doi:10.1016/ S0140-6736(20)30421-9.
2. N. Zhu *et al.*, A Novel Coronavirus from Patients with Pneumonia in China, 2019. *N. Engl. J. Med.*, NEJMoa2001017 (2020).
3. H. Tian *et al.*, *medRxiv*, in press, doi:10.1101/2020.01.30.20019844.
4. Z. Du *et al.*, Risk for Transportation of 2019 Novel Coronavirus Disease from Wuhan to Other Cities in China. *Emerg. Infect. Dis.* **26** (2020), doi:10.3201/eid2605.200146.
5. J. T. Wu, K. Leung, G. M. Leung, Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: a modelling study. *Lancet*. **6736** (2020), doi:10.1016/S0140-6736(20)30260-9.
6. S. Cauchemez *et al.*, Middle East respiratory syndrome coronavirus: Quantification of the extent of the epidemic, surveillance biases, and transmissibility. *Lancet Infect. Dis.* **14**, 50–56 (2014).

7. J. Lessler *et al.*, Incubation periods of acute respiratory viral infections: a systematic review. *Lancet Infect. Dis.* **9**, 291–300 (2009).
8. J. A. Backer, D. Klinkenberg, J. Wallinga, Incubation period of 2019 novel coronavirus (2019-nCoV) infections among travellers from Wuhan, China, 20–28 January 2020. *Eurosurveillance*. **25**, 20–28 (2020).
9. Q. Li *et al.*, Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus–Infected Pneumonia. *N. Engl. J. Med.*, 1–9 (2020).
10. World Health Organization (WHO), Coronavirus disease 2019 (COVID-19) Situation Report – 27 (2020) (available at https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200216-sitrep-27-covid-19.pdf?sfvrsn=78c0eb78_2).
11. Novel Coronavirus Pneumonia Emergency Response Epidemiology Team, The Epidemiological Characteristics of an Outbreak of 2019 Novel Coronavirus Diseases (COVID-19) — China, 2020. *China CDC Wkly.* **41**, 145–151 (2020).
12. E. Goldstein, V. E. Pitzer, J. J. O'Hagan, M. Lipsitch, Temporally Varying Relative Risks for Infectious Diseases. *Epidemiology*. **28**, 136–144 (2017).
13. World Health Organization (WHO), Novel Coronavirus (2019-nCoV) SITUATION REPORT - 3 (2020) (available at https://apps.who.int/iris/bitstream/handle/10665/330762/nCoVsitrep23Jan2020-eng.pdf).
14. J. F.-W. Chan *et al.*, A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: a study of a family cluster. *Lancet*. **395**, 514–523 (2020).
15. T. J. Hastie, D. Pregibon, in *Statistical Models in S*, J. M. Hastie, C. T. J., Eds. (Wadsworth & Brooks/Cole, 1992).
16. J. Riou, C. L. Althaus, Pattern of early human-to-human transmission of Wuhan 2019 novel coronavirus (2019-nCoV), December 2019 to January 2020. *Eurosurveillance*. **25**, 1–5 (2020).
17. A. R. Tuite, D. N. Fisman, Reporting, Epidemic Growth, and Reproduction Numbers for the 2019 Novel Coronavirus (2019-nCoV) Epidemic. *Ann. Intern. Med.*, 2019–2020 (2020).
18. Q. Li *et al.*, Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus–Infected Pneumonia. *N. Engl. J. Med.*, NEJMoa2001316 (2020).
19. M. J. Keeling, O. N. Bjørnstad, B. T. Grenfell, in *Ecology, Genetics and Evolution of Metapopulations* (Elsevier, 2004; https://linkinghub.elsevier.com/retrieve/pii/B9780123234483500192), pp. 415–445.
20. D. J. Watts, R. Muhamad, D. C. Medina, P. S. Dodds, Multiscale, resurgent epidemics in a hierarchical metapopulation model. *Proc. Natl. Acad. Sci.* **102**, 11157–11162 (2005).
21. G. Chowell, L. Sattenspiel, S. Bansal, C. Viboud, Mathematical models to characterize early epidemic growth: A review. *Phys. Life Rev.* **18**, 66–97 (2016).
22. C. Viboud *et al.*, Synchrony, waves, and spatial hierarchies in the spread of influenza. *Science*. **312**, 447–51 (2006).
23. M. J. Keeling, B. T. Grenfell, Individual-based Perspectives on R0. *J. Theor. Biol.* **203**, 51–61 (2000).
24. K. P. Burnham, D. R. Anderson, Multimodel Inference. *Sociol. Methods Res.* **33**, 261–304 (2004).
25. C. Fraser, S. Riley, R. M. Anderson, N. M. Ferguson, Factors that make an infectious disease outbreak controllable. *Proc. Natl. Acad. Sci.* **101**, 6146–6151 (2004).

Acknowledgements: We would like to thank all individuals who are collecting epidemiological data of the COVID-19 outbreak around the world. **Funding:** HT, OGP and MUGK acknowledge support from the Oxford Martin School. MUGK is supported by a Branco Weiss Fellowship. NRF is supported by a Sir Henry Dale Fellowship. The funders had no role in study design, data collection and analysis, decision to

publish or preparation of the manuscript. WPH was supported by the National Institute of General Medical Sciences (#U54GM088558). **Authors contributions:** MUGK, OGP, SVS developed the idea and research. MUGK and SVS wrote the first draft of the manuscript and all other authors discussed results and edited the manuscript. MUGK, BG, SVS, DMP and the open COVID-19 data working group collected and validated epidemiological data. RL collected intervention data. C-HY, BK, and SVS collected and processed human mobility data. **Competing interests:** The authors declare no competing interests. **Data and materials availability:** Code and data are available on the following GitHub repository: https://github.com/Emergent-Epidemics/covid19_cordon

Supplementary Materials

Materials and Methods

Supplementary Text

Table S1-S2

Fig S1 – S6

References 26 – 38

Materials and Methods

Epidemiological data: No officially reported line list was available for cases in China (*26*). We use a standard protocol (*27*, *28*) to extract individual level data from December 1st, 2019 - February 10th, 2020. Sources are mainly official reports from provincial, municipal or national health governments. Data was entered by a team of data curators on a rolling basis and technical validation and geo-positioning protocols were applied continuously to ensure validity. A detailed description of the methodology is available (*27*) (attached in Supplementary information). Lastly, total numbers were matched with officially reported data from China and other government reports. For sensitivity, GLM analyses (see below) were performed with case counts from the World Health Organization.

Proportions of symptomatic travelers: The proportion of cases who travelled while symptomatic was assessed from a subset of 236 cases for whom the dates of symptom onset and departure from Wuhan were available. Residency was split into three categories: Wuhan, China and International. Foreigners living in Wuhan were categorized as Wuhan and patients with missing Wuhan residency were either kept as missing values or categorized according to their country of origin. Both parametric (χ^2 test, (*29*)) and non-parametric (exact Fisher test, (*30*)) tests were performed and the uncertainty in proportions was assessed by the standard deviation of sample proportions.

Statistical inference of the incubation period: The incubation period is the time interval between infection and symptom onset. We assumed that cases travelling from Wuhan were exposed during their stay in Wuhan. We estimated the incubation period from 38 travelling cases returning from Wuhan with known dates of symptom onset, entry and exit. The end of the exposure period was assumed to be the exit travel date except if symptom onset occurred prior to the exit date (in which case exposure was assumed to have occurred prior to symptom onset). The start of the exposure period corresponded to the entry date. We assumed that the incubation period could not exceed 30 days.

For each case, the minimum and maximum incubation period was derived from the dates of entry, exit and symptom onset

$$IC_{max} = onset - entry$$

$$IC_{min} = onset - exit$$

We fitted a truncated gamma distribution (0 to 30 days) and estimated the mean and variance of the incubation period using Markov Chain Monte Carlo (MCMC) in a Bayesian framework using an uninformative prior distribution. We derived the likelihood as follows:

$$L = \frac{P_\Gamma(IC \leq IC_{max} + 1) - P_\Gamma(IC \leq IC_{min})}{P_\Gamma(IC \leq 30)}$$

A Metropolis-Hastings algorithm was implemented in R. Marginal posteriors were sampled from a chain of 5,000 steps after discarding a burn-in of 50 steps. Convergence was inspected visually.

Models of shifting age and sex distributions: Age and sex distributions are important in understanding risk of infection across populations. Assuming risk to be distributed relatively equally across a population, as an outbreak evolves age and sex distributions should follow the underlying population structure. Varying degrees of immunity and exposure may shift these distributions (*31*). To examine whether the ongoing outbreak shifted from an epidemic concentrated in Wuhan and among travelers from Wuhan to an epidemic that was self-sustained in provinces across China we use age and sex data from different periods of the outbreak for individuals with reported travel history and no known travel history. We define two periods of the outbreak, an "early" phase, starting with the first reports in early December and ending a set number of days after the Wuhan shutdown. This was selected to be 8 days after the Wuhan shutdown, which conservatively corresponds to one incubation period (see above) after the shutdown. After that date (i.e. 1st Feb 2020) we assume that most reported transmissions in provinces outside of Wuhan are the result of local transmission. We further divided our data in those that had cases

with known travel history to Wuhan and those who did not. Then we produce the following summary statistics:

1. Average age stratified by sex for all cases with reported travel history to Wuhan.
2. Average age stratified by sex for all cases with no reported travel history to Wuhan in the period between December 1, 2019 - January 31, 2020.

We then compare these with:

3. Average age stratified by sex for all cases with no reported travel history to Wuhan in the period between January 31, 2020 - February 10, 2020.

Model M1 compares the distribution of age and sex among travelers to the reported infections outside Wuhan with no known travel history. In case these two distributions are similar, import driven epidemic can be concluded. Under our *model assumptions M2*, if the epidemic was driven largely by importations across the two time periods, all age and sex distributions should mirror those of the reported traveler infections. Under our *model* assumption *M3*, if the epidemic was driven by other factors (i.e., local transmission), the two distributions should vary across the two time periods.

We cannot exclude the possibility that shifts in distributions may be due to heightened awareness among the general population which may have increased reporting in female cases later in the epidemic. Further, more work will be necessary to understand the differential risk of severe or symptomatic disease to fully understand the age and sex distributions in this outbreak. For example, why there are relatively few reports of cases <18y old. However, as for other respiratory pathogens symptomatic and severe infection were more concentrated in older populations. We do not intend to make any general statements about differential risk but were more interested in shifts in reported cases across multiple geographies in China.

Real time human mobility data: We extract human mobility data from the Baidu Qianxi web platform, which presents daily population travels between cities or provinces tracked through the Baidu Huiyan system. The data do not represent numbers of individual travelers but rather an index of relative movements constructed by Baidu's proprietary methods which are correlated with human mobility (*32*) (http://qianxi.baidu.com/). In particular, two pieces of information are collected. First, we extract a series of migration scale indices for traveling out of Wuhan, from January 1st to February 10, both in 2019 and 2020. Second, we obtain the proportion of human movement from Wuhan were bound for each of 31 provinces in China. These proportions are available for January 1st - February 10, 2020. Based on this data we had access to both changes in mobility volume and changes in mobility direction. See more detailed descriptions of the human movement data here: (*33*, *34*)

Review of interventions and reporting shifts: We reviewed the literature and online social media to understand the key timings of interventions and announcements that are relevant for disease transmission across China. We collated information about the type (e.g., announcement of outbreak, travel restrictions, isolation of patients, etc.), geographic location (e.g., city where available, province), and timing (specific date or date range).

COVID-19 case definitions:

Definitions of probable and confirmed COVID-19 cases have changed throughout the epidemic. We collected data from official sources describing the timing and specifics of the case definitions.

From January 18-22:

Probable: Need to satisfy (i) and (ii):

i. Clinical symptoms: (1) fever; (2) imaging showing pneumonia typical of the disease; (3) during early disease, total white cells normal or reduced, or lymph cell count reduced.

ii. Epidemiologic history: (1) within 2 weeks of symptom onset, Wuhan travel or resident history; or within 2 weeks of symptom onset, contact with persons from Wuhan who had fever with respiratory symptoms; or belong to a cluster.

Confirmed: Need to satisfy criteria for probable case and have a real-time quantitative polymerase chain reaction (RT-qPCR) positive result from sputum, nasopharyngeal swabs, lower respiratory tract secretions or other sample tissue, or genome sequencing highly similar with known SARS-CoV-2. available strains.

From January 22-23:

Probable: Need to satisfy (i) and any one epidemiologic history described in (ii):

i. Clinical symptoms: (1) fever; (2) imaging showing pneumonia typical of the disease; (3) during early disease, total white cells normal or reduced, or lymph cell count reduced

ii. Epidemiologic history: (1) within 2 weeks of symptom onset, Wuhan travel or resident history; (2) within 2 weeks of symptom onset, contact with persons from Wuhan who had fever with respiratory symptoms; (3) belong to a cluster or had epidemiologic link with confirmed cases.

Confirmed: Need to satisfy criteria for probable case and have a RT-qPCR positive result from respiratory or blood samples, or genome sequencing highly similar with known SARS-CoV-2. available strains.

From January 23-27:

Probable: Need to satisfy (i) and (ii):

i. Clinical symptoms: (1) fever; (2) imaging showing pneumonia typical of the disease; (3) during early disease, total white cells normal or reduced, or lymph cell count reduced

ii. Epidemiologic history: within 2 weeks of symptom onset, Wuhan travel or resident history; or within 2 weeks of symptom onset, contact with persons from Wuhan who had fever with respiratory symptoms, or belong to a cluster.

Confirmed: Need to satisfy criteria for probable case and have a RT-qPCR positive result from sputum, nasopharyngeal swabs, lower respiratory tract secretions, or other samples, or genome sequencing highly similar with known SARS-CoV-2. available strains.

From January 27-February 5:

Probable: Need to satisfy any two of the symptoms described in (i) and any of the epidemiological history described in (ii):

i. Clinical symptoms: (1) fever; (2) imaging showing pneumonia typical of the disease; (3) during early disease, total white cells normal or reduced, or lymph cell count reduced

ii. Epidemiologic history: (1) within 2 weeks of symptom onset, travel or resident history in Wuhan region or other places with sustained local transmission; (2) within 2 weeks of symptom onset, contact with persons from Wuhan city or other places with sustained local transmission who had fever with respiratory symptoms, (3) belong to a cluster or epidemiologic connection with COVID-19 infected persons.

Confirmed: Need to satisfy criteria for probable case and have a RT-qPCR positive result from respiratory or blood samples, or genome sequencing highly similar with known SARS-CoV-2. available strains from lab test of respiratory or blood samples.

Comparing predictive models of epidemic trajectories: BIC scores shown in **Fig. 4b** are calculated on Generalized Linear Models (GLM) of the form $Y(t) = Y(t-7) + IT(t) + M(t-5) + IM(t)$ where $Y(t)$ is either the cumulative number of cases observed through day t or the number of new cases observed on day t, $Y(t-7)$ represents the cumulative (or new) number of cases seven days prior (median doubling time), $IT(t)$ is an indicator function for PCR test availability that is 1 after Jan. 19th 2020 and 0 otherwise, $M(t-5)$ is the Baidu-estimated mobility between Wuhan and each province 5 days prior (median incubation period), and $IM(t)$ is an indicator function which is set to 1 after Jan. 26th and 0 otherwise (one incubation period after Jan. 22nd). Models were fit to province-level data. The GLM were compared using differences in

Bayesian Information Criteria (BIC), where larger values indicate models with lower relative support, and BIC>4 considered the cutoff for substantial model improvement. We performed a detailed sensitivity analysis on the availability of PCR tests, doubling time, and incubation periods. We confirmed obtained qualitatively similar results for log-linear regressions fit to cumulative cases, Poisson GLM fit to daily case counts, and Negative Binomial GLM fit to daily case counts (**Table S2**). In addition, we provide a full time series analysis of the optimal lag structure for cases and mobility for each province. Lastly, although BIC is considered more conservative, model selection results were confirmed using AIC for model selection (see **Fig. 4** and **Table S2**). Lastly, we validated our model selection results using elastic-net regression and n-fold cross validation as implemented in the R package GLMNET (*35*, *36*).

Supplementary text:

To ascertain whether earlier travel restrictions could have prevented the wide-spread increase in cases witnessed in late-January we constructed a simple forecasting model for COVID-19. Briefly, we forecast the cumulative number of cases in each Chinese province by simply doubling the number of cumulative cases reported six days prior. For dates prior to Jan. 28th and after Feb 3rd, this naive forecast produces an accurate estimate of the cumulative number of cases in each province **(Fig. S4)**. However, the cumulative number of cases reported on Jan 28th is poorly estimated using this model (**Fig. S4).** In order to accurately forecast the number of cases on Jan 28th, we must also include the relative amount of mobility out of Wuhan into various provinces in the regression model. In **Fig. S4**, we show how a model including only movement from Wuhan on January 22nd fit to the residuals from **Fig. S4** is once again able to accurately forecast cumulative cases. This indicates that for any hope of success, movement restrictions must be prompt.

Supplementary Tables:

Table S1: Table shows the pseudo- R^2 values for Poisson and Negative Binomial GLM of daily case counts and 5-day lagged log mobility from Wuhan, where pseudo- R^2 were calculated using model deviances as described (*37*, *38*).

Date	Poisson (pseudo R^2)	Negative Binomial (pseudo R^2)
01-14-2020	0.03	0.03
01-18-2020	0.09	0.30
01-25-2020	0.58	0.42
01-29-2020	0.99	0.70

Table S2: Table shows the AIC and BIC values for a log-linear regression based on cumulative cases, a Poisson GLM of daily case counts and a Negative Binomial GLM of daily case counts using seven combinations of predictors.

Model	AIC (log linear)	BIC (log linear)	AIC (Poisson)	BIC (Poisson)	AIC (Neg. Binom.)	BIC (Neg. Binom.)
All	1429.898	1455.076	22262.80	22290.92	5473.462	5507.198
Cases + Mobility Ind. + Mobility	1471.191	1492.173	25392.58	25415.07	6038.094	6066.207
Cases + Testing + Mobility	1698.308	1719.290	23155.21	23177.70	5498.274	5526.387
Cases + Mobility Ind.	1578.920	1595.705	27008.11	27024.97	6125.860	6148.351
Cases + Mobility	1839.027	1855.813	36893.33	36910.20	6368.911	6391.401
Cases + Testing	1753.424	1770.210	24565.45	24582.32	5580.771	5603.261
Cases	1858.853	1871.442	38118.68	38129.93	6403.403	6420.271

Supplementary Figures:

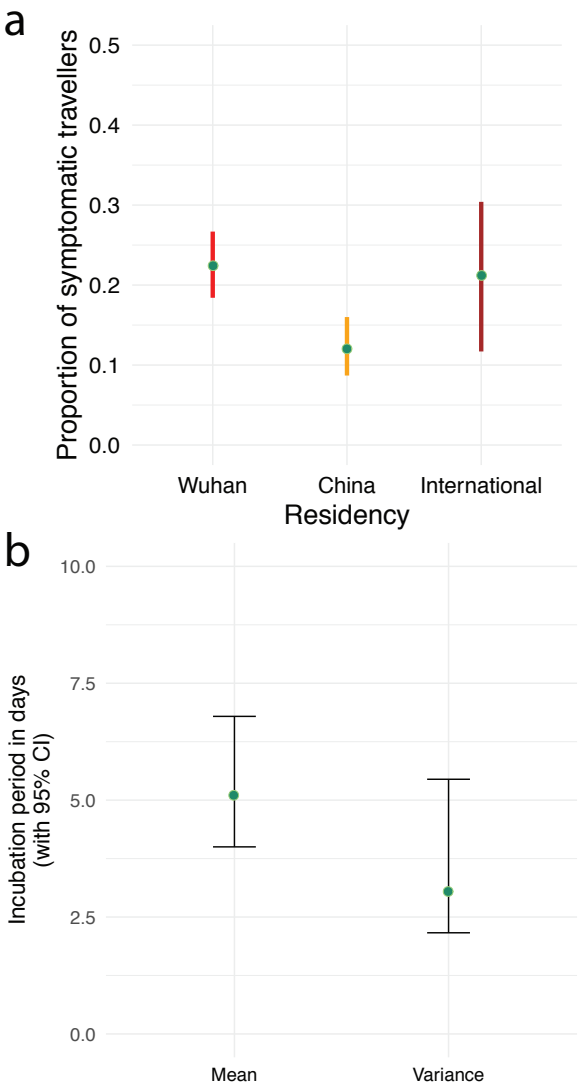

Figure S1: a) Dates of symptom onset before date of travel from Wuhan. b) Incubation period estimates and standard deviation.

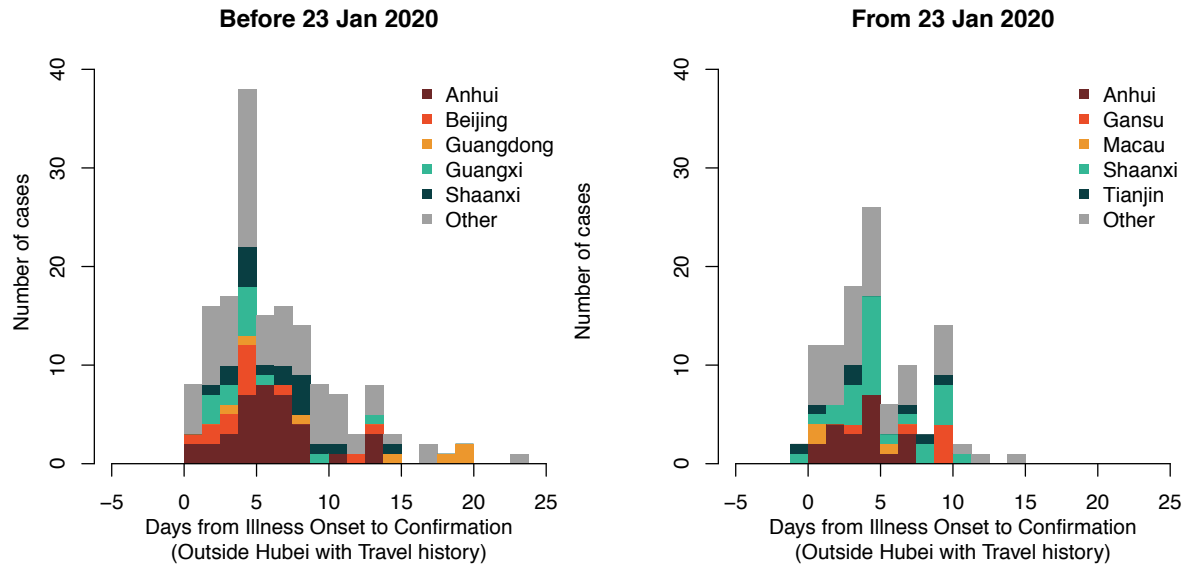

Figure S2: Interval between symptom onset and date of confirmation in confirmed cases with reported travel history in two key periods, before and after January 23, 2020.

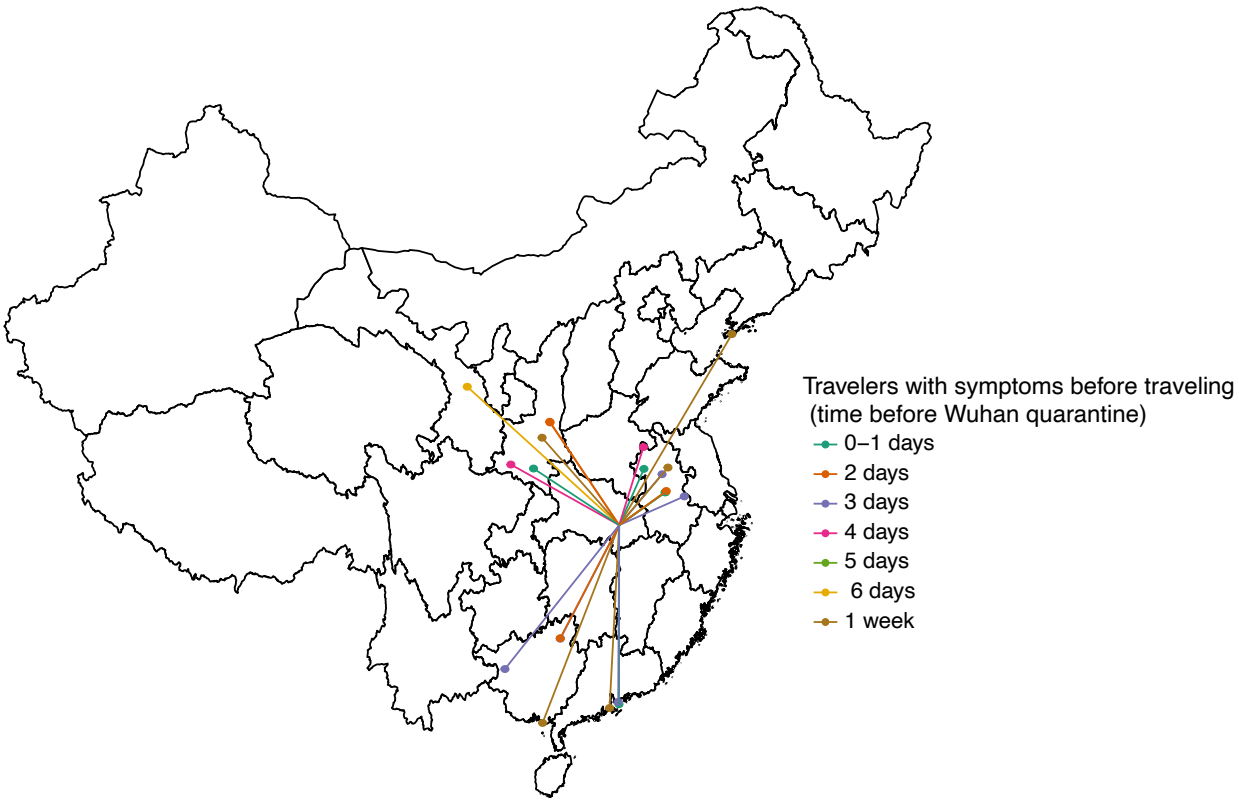

Figure S3: Map of confirmed cases of COVID-19 with known travel history and date of onset date before date of travel.

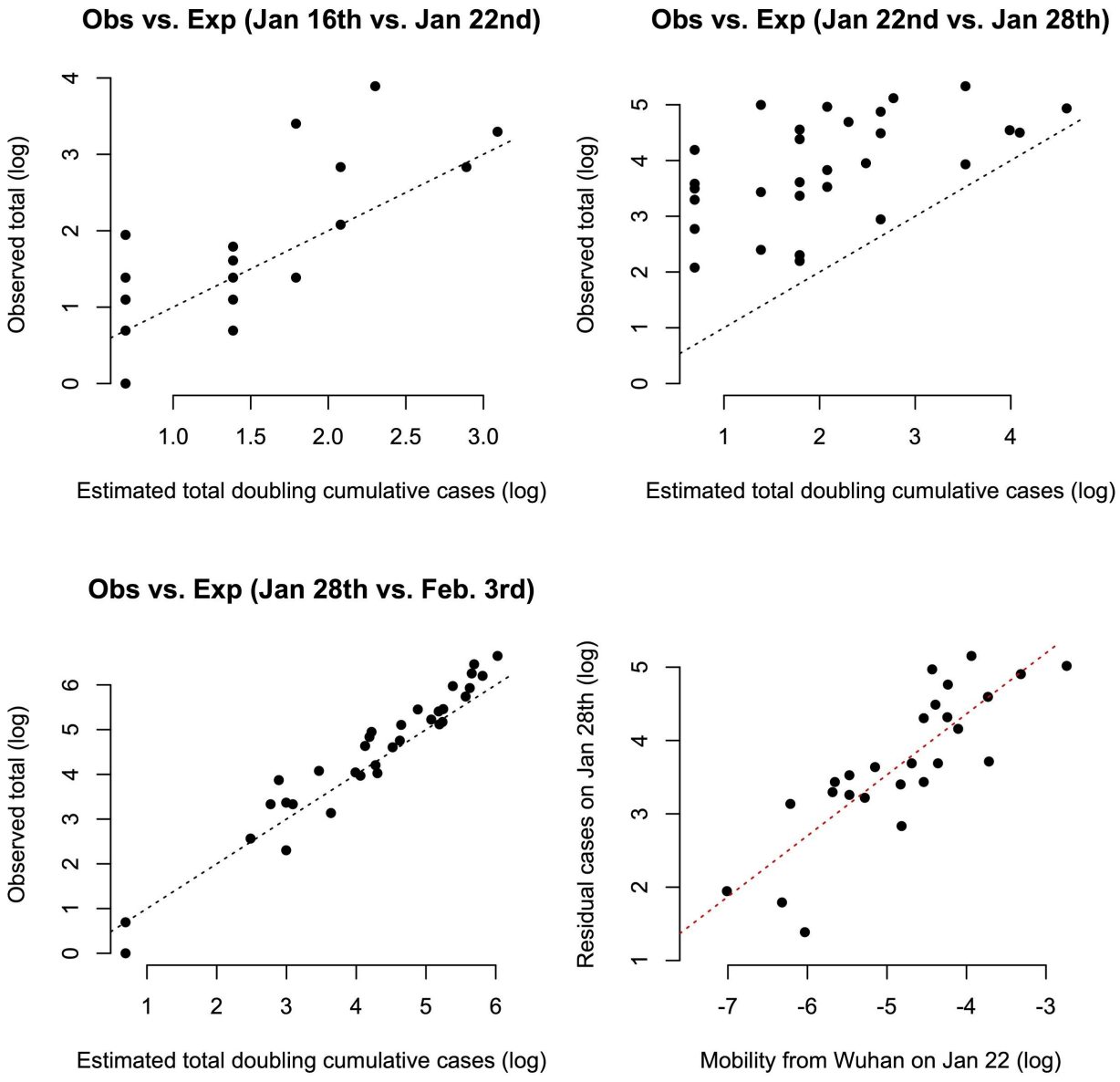

Figure S4: Predicting COVID-19 cases using mobility data. a) Province-level cumulative cases on January 22nd can be accurately predicted based on simply doubling the number of cumulative number of cases occurring on January 16th. b) However, by Jan. 28th, the expected number of cases has significantly increased with respect to predictions based on cases through January 22nd. c) By Feb. 34rd, cumulative cases are once again well estimated based on the cumulative number of cases in each province six days earlier, i.e., on Jan. 28th. d) The deviation in cases on January 28th is well explained by the relative amount of migration out of Wuhan on January 22nd.

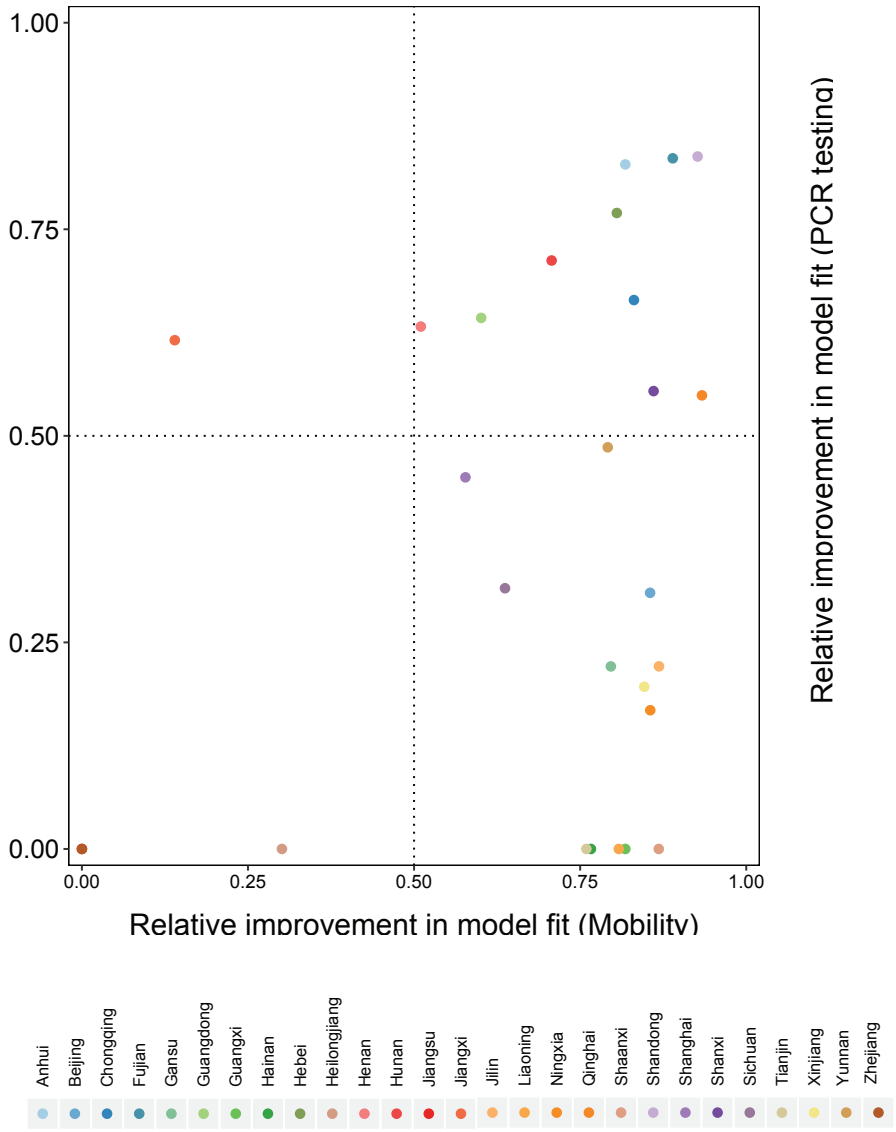

Figure S5: Relative importance of PCR testing vs. human mobility to improve a simple GLM of COVID-19 when estimating exponential growth in province-level cases. Relative improvement is measured as one minus the residuals of a GLM with lagged cases + PCR testing availability (y-axis) and a GLM with lagged cases + mobility from Wuhan. Values were normalized by the observed number of cases such that they ranged between 0 and 1. The resulting metric has a value of 0 for a model where the residual error vastly eclipses the observed data and a value of 1 when residual error is 0, i.e., a perfect model fit.

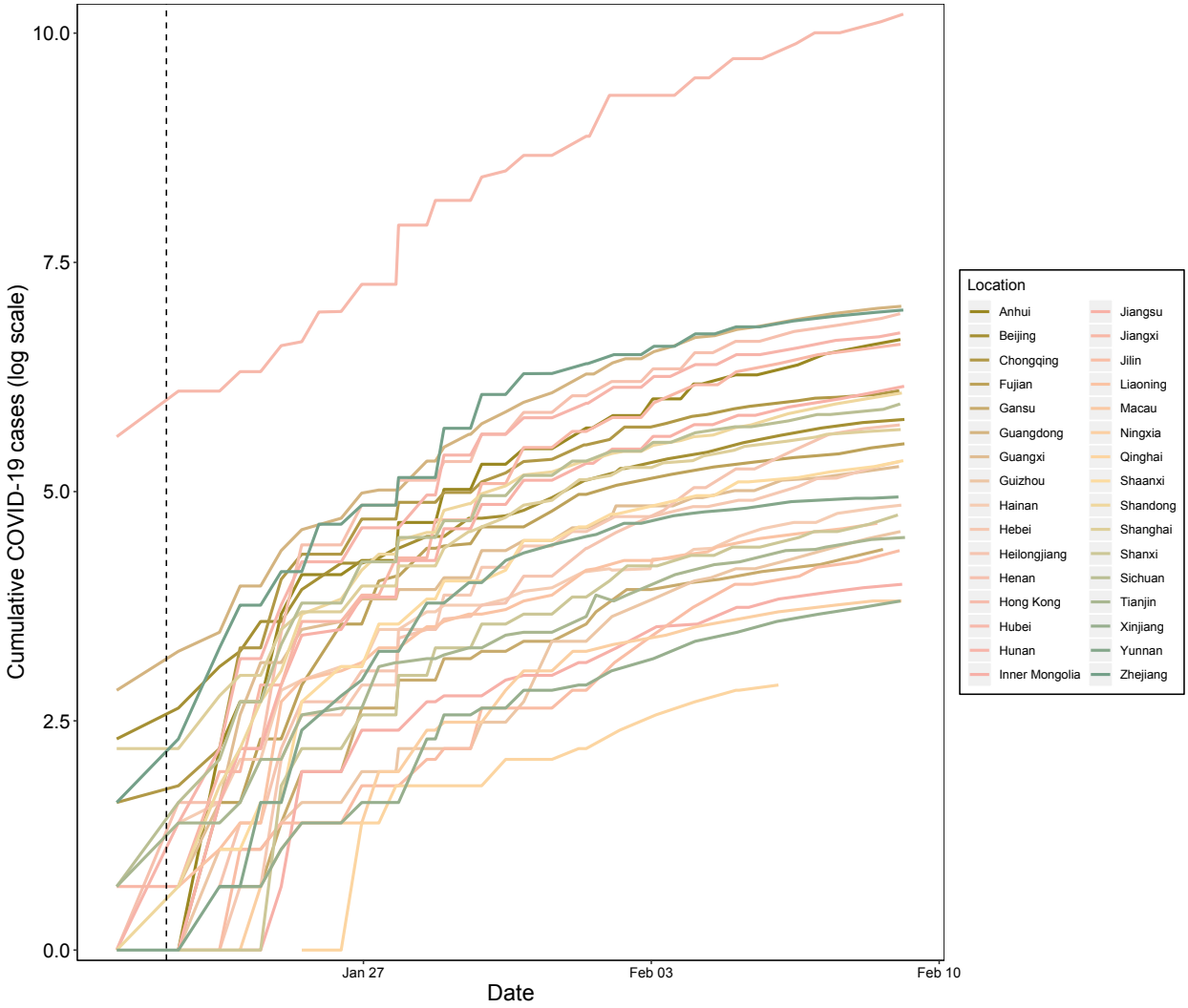

Figure S6: Cumulative case counts of COVID-19 in China between January 21 and February 10, 2020 (log scale).

Members of the COVID-19 working group: Bo Xu, Sumiko Mekaru, Kara Sewalk, Lauren Goodwin, Alyssa Loskill, Emily Cohn, Yulin Hswen, Sarah C. Hill, Maria M. Cobo, Alexander Zarebski, Sabrina Li, Erin Hulland, Julia Morgan, Lin Wang, Katelynn O'Brien.

Additional references

26. B. Xu, M. U. G. Kraemer, Open access epidemiological data from the COVID-19. *Lancet Infect. Dis.* **3099**, 30119 (2020).
27. B. Xu *et al.*, *Sci. Data*, in press.
28. R. E. Ramshaw *et al.*, A database of geopositioned Middle East Respiratory Syndrome Coronavirus occurrences. *Sci. Data*. **6**, 318 (2019).
29. M. L. McHugh, The Chi-square test of independence. *Biochem. Medica*. **23**, 143–149 (2013).

30. J. H. McDonald, *Handbook of Biological Statistics* (Sparky House Publishing, Baltimore, Maryland, ed. 3, 2014).
31. N. R. Faria *et al.*, Genomic and epidemiological monitoring of yellow fever virus transmission potential. *Science*. **361**, 894–899 (2018).
32. M. U. G. Kraemer *et al.*, Past and future spread of the arbovirus vectors Aedes aegypti and Aedes albopictus. *Nat. Microbiol.* **4**, 854–863 (2019).
33. G. Jin, J. Yu, L. Han, S. Duan, The impact of traffic isolation in Wuhan on the spread of 2019-nCov. *medRxiv* (2020), doi:https://doi.org/10.1101/2020.02.04.20020438.
34. A. S. Ai *et al.*, Population movement, city closure and spatial transmission of the 2019-nCoV infection in China. *medRxiv* (2020), doi:doi.org/10.1101/2020.02.04.20020339.
35. J. Friedman, T. Hastie, R. Tibshirani, Regularization Paths for Generalized Linear Models via Coordinate Descent. *J. Stat. Softw.* **33** (2010), doi:10.18637/jss.v033.i01.
36. R Core Team, R: A language and environment for computing. Vienna, Austria. *R Found. Stat. Comput.* (2019).
37. H. Heinzl, M. Mittlböck, Pseudo R-squared measures for Poisson regression models with over- or underdispersion. *Comput. Stat. Data Anal.* **44**, 253–271 (2003).
38. A. Zeileis, C. Kleiber, S. Jackman, Regression models for count data in R. *J. Stat. Softw.* **27**, 1–25 (2008).