

# Brief Report: Genomic epidemiology of a densely sampled COVID19 outbreak in China

Erik M Volz<sup>1\*</sup>, Han Fu<sup>1</sup>, Haowei Wang<sup>1</sup>, Xiaoyue Xi<sup>2</sup>, Wei Chen<sup>3</sup>, Dehui Liu<sup>3</sup>,  
Yingying Chen<sup>3</sup>, Mengmeng Tian<sup>3</sup>, Wei Tan<sup>4</sup>, Junjie Zai<sup>5</sup>, Wanying Sun<sup>6</sup>, Jiandong  
Li<sup>6</sup>, Junhua Li<sup>6</sup>, Xingguang Li<sup>7†‡\*</sup>, Qing Nie<sup>3†§\*</sup>

**\*For correspondence:**

[nieqing0454@163.com](mailto:nieqing0454@163.com) (QN);  
[xingguanglee@hotmail.com](mailto:xingguanglee@hotmail.com) (XL)

†These authors contributed equally  
to this work

**Present address:** <sup>‡</sup>Department of  
Microbiology, Weifang Center for  
Disease Control and Prevention,  
Weifang 261061, China. Tel:  
+86-0536-8098503; <sup>§</sup>Hubei  
Engineering Research Center of  
Viral Vector, Wuhan University of  
Bioengineering, Wuhan, 430415,  
China. Tel: +86-027-89648139

<sup>1</sup>Department of Infectious Disease Epidemiology and MRC Centre for Global Infectious  
Disease Analysis, Imperial College London, Norfolk Place, W2 1PG, United Kingdom;  
<sup>2</sup>Department of Mathematics, Imperial College London, London SW7 2AZ, United  
Kingdom; <sup>3</sup>Department of Microbiology, Weifang Center for Disease Control and  
Prevention, Weifang 261061, China.; <sup>4</sup>Department of Respiratory Medicine, Weifang  
People's Hospital, Weifang 261061, China.; <sup>5</sup>Immunology Innovation Team, School of  
Medicine, Ningbo University, Ningbo 315211, China.; <sup>6</sup>Shenzhen Key Laboratory of  
Unknown Pathogen Identification, BGI-Shenzhen, Shenzhen 518083, China.; <sup>7</sup>Hubei  
Engineering Research Center of Viral Vector, Wuhan University of Bioengineering, Wuhan,  
430415, China.

**Abstract** Analysis of genetic sequence data from the pandemic SARS Coronavirus 2 can provide  
insights into epidemic origins, worldwide dispersal, and epidemiological history. With few  
exceptions, genomic epidemiological analysis has focused on geographically distributed data sets  
with few isolates in any given location. Here we report an analysis of 20 whole SARS-CoV 2  
genomes from a single relatively small and geographically constrained outbreak in Weifang,  
People's Republic of China. Using Bayesian model-based phylodynamic methods, we estimate the  
reproduction number for the outbreak to be 2.6 (95% CI:1.5-5). We further estimate the number of  
infections through time and compare these estimates to confirmed diagnoses by the Weifang  
Centers for Disease Control. We find that these estimates are consistent with reported cases and  
there is unlikely to be a large undiagnosed burden of infection over the period we studied.

## Introduction

We report a genomic epidemiological analysis of one of the first geographically concentrated  
community transmission samples of SARS-CoV 2 genetic sequences collected outside of the initial

32 outbreak in Wuhan, China. These data comprise 20 whole genome sequences from confirmed  
33 COVID19 infections in Weifang, Shandong Province, People's Republic of China. The data were  
34 collected over the course of several weeks up to February 10, 2020 and overlap with a period  
35 of intensifying public health and social distancing measures. Phylodynamic analysis allows us to  
36 evaluate epidemiological trends after seeding events which took place in mid to late January, 2020.

37 The objective of our analysis is to evaluate epidemiological trends based on national surveillance  
38 and response efforts by Weifang Centers for Disease Control (CDC). This analysis provides an  
39 estimate of the initial rate of spread and reproduction number in Weifang City. In contrast to the  
40 early spread of COVID19 in Hubei Province of China, most community transmissions within Weifang  
41 took place after public health interventions and social distancing measures were put in place. We  
42 therefore hypothesize that genetic data should reflect a lower growth rate and reproduction number  
43 than was observed in Wuhan. A secondary aim is to estimate the total numbers infected and to  
44 evaluate the possibility that there is a large unmeasured burden of infection due to imperfect case  
45 ascertainment and a large proportion of infections with mild or asymptomatic illness.

46 To analyze the Weifang sequences, we have adapted model-based phylodynamic methods  
47 which were previously used to estimate growth rates and reproduction numbers using sequence  
48 data from Wuhan and exported international cases(Volz *et al.*, 2020). This analysis has several  
49 constraints and requirements:

50

51 *Importation of lineages from Wuhan.* The outbreak in Weifang was seeded by multiple lineages  
52 imported at various times from the rest of China. We use a phylodynamic model that accounts for  
53 location of sampling. Migration is modeled as a bi-directional process with rates proportional to  
54 epidemic size in Weifang. The larger international reservoir of COVID19 cases serves as a source of  
55 new infections and is assumed to be growing exponentially over this period of time.

56 *Nonlinear epidemiological dynamics in Weifang.* The maximum number of daily confirmed COVID19  
57 cases occurred on February 5, but it is unknown when the maximum prevalence of infection oc-  
58 curred. We use a susceptible-exposed-infectious-recovered (SEIR) model(Keeling and Rohani, 2011)  
59 for epidemic dynamics in Weifang. The model accounts for a realistic distribution of generation  
60 times and can potentially capture a nonlinear decrease in cases following epidemic peak.

61 *Variance in transmission rates(Lloyd-Smith *et al.*, 2005).*To estimate total numbers infected, the  
62 phylodynamic model must account for epidemiological variables which are known to significantly  
63 influence genetic diversity. Foremost among these is the variance in offspring distribution (number  
64 of transmissions per primary case). We draw on previous evidence based on the previous SARS  
65 epidemic which indicates that the offspring distribution is highly over-dispersed. High variance of  
66 transmission rates will reduce genetic diversity of a sample and failure to account for this factor  
67 will lead to highly biased estimates of epidemic size(Li *et al.*, 2017). Recent analyses of sequence  
68 data drawn primarily from Wuhan has found that high over-dispersion was required for estimated  
69 cases to be consistent with the epidemiological record(Volz *et al.*, 2020). Models assuming low  
70 variance in transmission rates between people would generate estimates of cases that are lower  
71 than the known number of confirmed cases. Separately, Grantz *et al.*(Grantz *et al.*, ????) have

72 found that high over-dispersion is required to reconcile estimated reproduction numbers with  
 73 the observed frequency of international outbreaks. In this study, we elaborate the SEIR model to  
 74 include a compartment( $J$ ) with higher transmission rates. The variance of the implied offspring  
 75 distribution is calibrated to give similar overdispersion from the SARS epidemic.

## 76 Results

77 Despite an initial rapid increase in confirmed cases in Weifang in late January and early February, the  
 78 number of confirmed cases by Weifang CDC show that outbreak peaked quite early and maximum  
 79 number of cases occurred on February 5. Phylodynamic analysis supports the interpretation that  
 80 control efforts reduced epidemic growth rates and contributed to eventual control. **Figure 1A**  
 81 illustrates the phylodynamic model which was co-estimated with the phylogeny which provides  
 82 estimates of epidemiological parameters summarized in **Table 1**. **Figure 1B** shows the estimated  
 83 time scaled phylogeny (maximum clade credibility) including 20 lineages sampled from distinct  
 84 patients in Weifang and 33 genomes sampled from Wuhan and internationally.

85 The estimated number of infections is shown Figure 1C. The time series of confirmed cases  
 86 should lag the estimated number of infected because of delays from infection to appearance of  
 87 symptoms and delays from symptoms to diagnosis. We also expect that an unknown proportion of  
 88 infections will be missed by the surveillance system due very mild, subclinical, or asymptomatic  
 89 infection. Our estimates do not support the hypothesis that there was a very large hidden burden  
 90 of infection in Weifang over the period that the sequence data were sampled. Indeed, our central  
 91 estimate for the number infected on 10th February is only 142 and the credible intervals cover the  
 92 44 cumulative confirmed cases at the end of February.

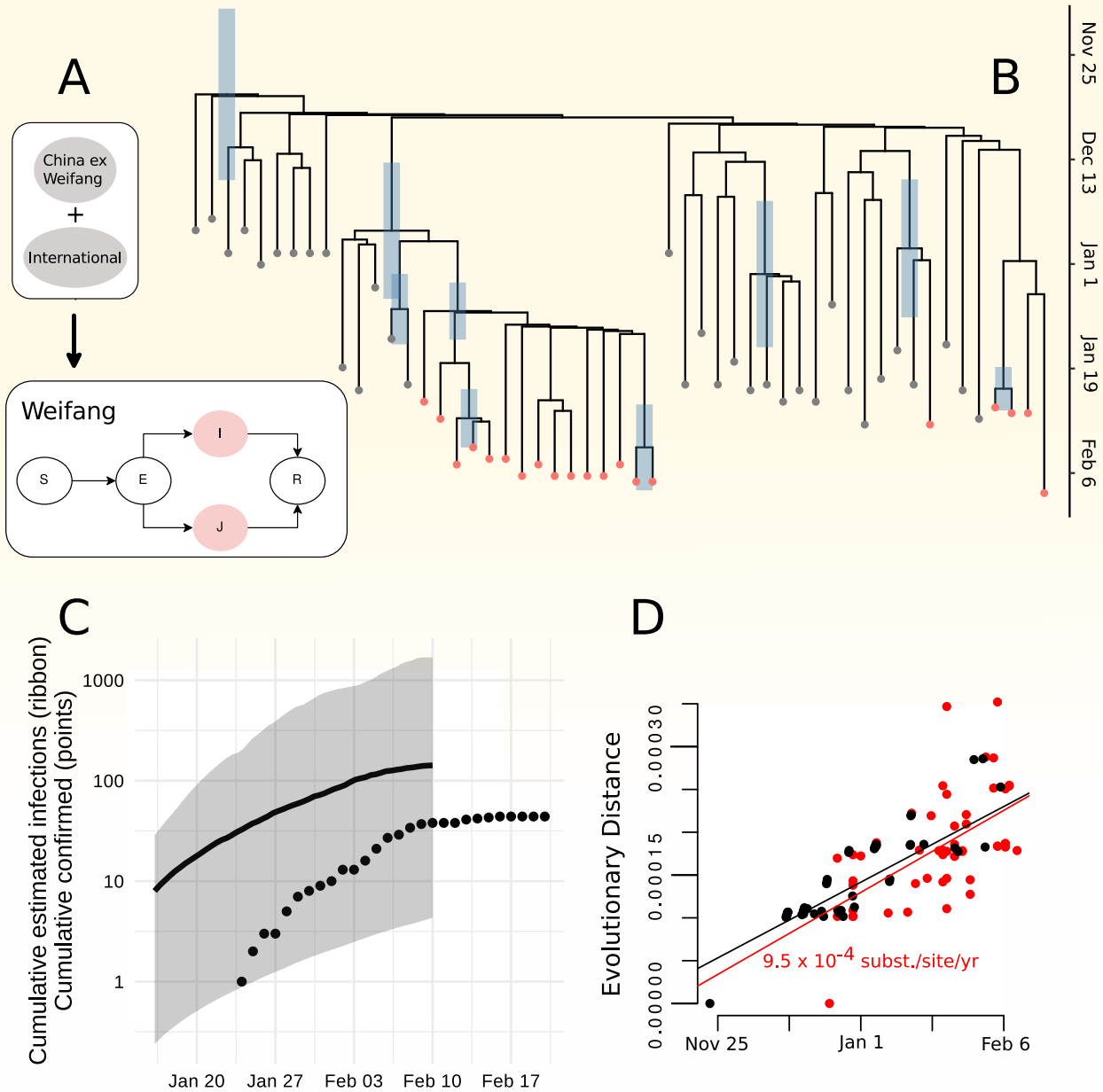
**Table 1.** Summary of primary epidemiological and evolutionary parameters, including Bayesian prior distributions and estimated posteriors. Posterior uncertainty is summarized using a 95% highest posterior density interval.

Parameter	Prior	Posterior mean	95% HPD
Initial infected	Exponential(mean=1)	4.5	0.26-13
Initial susceptible	Log-normal(mean log=6, sd log=1)	660	23-3000
Migration rate <sup>1</sup>	Exponential(mean=10*)	1.6	0.73-2
Reproduction number	Log-normal(mean log=1.03, sd log =0.5)	2.6	1.5-5
Molecular clock rate <sup>2</sup>	Uniform(0.0007,0.003)	0.00098	0.00071-0.0015
Transition/transversion	Log-normal(mean log=1, sd log=1.25)	5.5	3.1-9.4
Gamma shape	Exponential(mean=1)	0.74	0.015-3

<sup>1</sup> Units: Migrations per lineage per year.

<sup>2</sup> Units: Substitutions per site per year.

93 We do not have sufficient data to detect a large decrease in epidemic growth rates as the  
 94 epidemic progressed. We estimate  $R_0 = 2.6$  (95% HPD:1.5-5). The growth rate in estimated  
 95 infections remained positive but decreased substantially over the sampling period. These estimates  
 96 correspond to growth during a period when Weifang was implementing a variety of public health  
 97 interventions and contact tracing to limit epidemic spread. These interventions included public



**Figure 1.** Phylodynamic estimates and epidemiological model. A. Diagram representing the structure of the epidemiological SEIR model which was fitted in tandem with the time scaled phylogeny. Colours correspond to the state of individuals sampled and represented in the tree (B). Note that infected and infectious individuals may occupy a low transmission state (I) or a high transmission rate state (J) to account for high dispersion of the reproduction number. B. A time scaled phylogeny co-estimated with epidemiological parameters. The colour of tips corresponds to location sampling. Red tips were sampled from Weifang, China. The credible interval of TMRCA is shown as a blue bar for all nodes with more than 50% posterior support. C. Cumulative estimated infections through time produced by fitting the SEIR model and the cumulative confirmed cases (points) reported by Weifan CDC. The shaded region shows the 95% HPD and the line shows the posterior median. D. A root to tip regression showing approximately linear increase in diversity with time of sampling.

**Figure 1-Figure supplement 1.** Maximum likelihood time tree.

**Figure 1-Figure supplement 2.** Tree posterior density plot.

**Figure 1-Figure supplement 3.** Tree posterior density plot.

98 health messaging, establishing phone hotlines, encouraging home isolation for recent visitors from  
99 Wuhan (January 23-26), optimizing triage of suspected cases in hospitals (January 24), travel  
100 restrictions (January 26), extending school closures, and establishing 'fever clinics' for consultation  
101 and diagnosis (January 27) (Mao, 2020).

102 As well as providing novel epidemiological estimates, our results point to the significance of  
103 realistic modeling for fidelity of phylogenetic inference. The use of a model-based structured coa-  
104 lescent prior had large influence over estimated molecular clock rates and inferred TMRCA. **Figure**  
105 **Supplement 1** shows that maximum likelihood inference of time-scaled phylogenies produces a  
106 distribution of TMRCA which are substantially different than the Bayesian model-based analysis.  
107 Choice of population genetic prior will have a large influence on phylogenetic inference based on  
108 sparse or poorly informative genetic sequence data. Among the 20 Weifang sequences included in  
109 this analysis, there is mean pairwise difference of only three single nucleotide polymorphisms and  
110 only approximately twice as much diversity observed among the remainder of the sequences we  
111 studied. There is correspondingly low confidence in tree topology (**Figure Supplement 2**), and only  
112 three clades had greater than 50% posterior support including one clade which had a monophyletic  
113 composition of 13 Weifang lineages. The earliest Weifang sequence was sampled on January 25  
114 from a patient who showed first symptoms on January 16. These dates cover a similar range as the  
115 posterior TMRCA of all Weifang sequences (**Figure Supplement 3**).

## 116 Discussion

117 Our analysis of 20 SARS-CoV 2 genomes from Weifang, China has confirmed independent observa-  
118 tions regarding the rate of spread and burden of infection in the city. Surveillance of COVID19 is  
119 rendered difficult by high proportions of illness with mild severity and an unknown proportion of  
120 asymptomatic infection (Guan et al., 2020). The extent of under-reporting and case ascertainment  
121 rates has been widely debated. Analysis of genetic sequence data provides an alternative source of  
122 information about epidemic size which can be more robust to imperfect case ascertainment. We  
123 do not find evidence for a very large hidden burden of infection within Weifang. The relatively low  
124 estimate of  $R_0$  is consistent with a slower rate of spread outside of Wuhan and effective control  
125 strategies implemented in late January.

126 While the value of pathogen genomic analyses is widely recognized for estimating dates of  
127 emergence (Verity Hill, 2020; Gire et al., 2014) and identifying animal reservoirs (Zhou et al., 2020;  
128 Dudas et al., 2018), analysis of pathogen sequences also has potential to inform epidemic surveil-  
129 lance and intervention efforts. With few exceptions (Stadler, 2020; Bedford, 2020), this potential  
130 is currently not being realized for the international response to COVID19. It is worth noting that  
131 the analysis described in this report was accomplished in approximately 48 hours and drew on  
132 previously developed models and packages for BEAST2 (Bouckaert et al., 2019; Volz and Siveroni,  
133 2018). It is therefore feasible for phylodynamic analysis to provide a rapid supplement to epidemio-  
134 logical surveillance, however this requires rapid sequencing and timely sharing of data as well as  
135 randomized concentrated sampling of the epidemic within localities such as individual cities.

## 136 **Methods and Materials**

137 **Epidemiological investigation, sampling, and genetic sequencing.** As of 10 February 2020, 136  
 138 suspected cases, and 214 close contacts were diagnosed by Weifang Center for Disease Control  
 139 and Prevention. 28 cases were detected positive with SARS-CoV-2. Viral RNA was extracted using  
 140 Maxwell 16 Viral Total Nucleic Acid Purification Kit (Promega AS1150) by magnetic bead method  
 141 and RNeasy Mini Kit (QIAGEN 74104) by column method. RT-qPCR was carried out using 2019  
 142 novel coronavirus nucleic acid detection kit (BioGerm, Shanghai, China) to confirm the presence  
 143 of SARS-CoV-2 viral RNA with cycle threshold (Ct) values range from 17 to 37, targeting the high  
 144 conservative region (ORF1ab/N gene) in SARS-CoV-2 genome. Metagenomic sequencing: The  
 145 concentration of RNA samples was measurement by Qubit RNA HS Assay Kit (Thermo Fisher  
 146 Scientific, Waltham, MA, USA). DNase was used to remove host DNA. The remaining RNA was used  
 147 to construct the single-stranded circular DNA library with MGIEasy RNA Library preparation reagent  
 148 set (MGI, Shenzhen, China). Purified RNA was then fragmented. Using these short fragments as  
 149 templates, random hexamers were used to synthesize the first-strand cDNA, followed by the second  
 150 strand synthesis. Using the short double-strand DNA, a DNA library was constructed through end  
 151 repair, adaptor ligation, and PCR amplification. PCR products were transformed into a single strand  
 152 circular DNA library through DNA-denaturation and circularization. DNA nanoballs (DNBs) were  
 153 generated with the single-stranded circular DNA library by rolling circle replication (RCR). The DNBs  
 154 were loaded into the flow cell and pair-end 100bp sequencing on the DNBSEQ-T7 platform 8 (MGI,  
 155 Shenzhen, China). 20 genomes were assembled with length from 26,840 to 29,882 nucleotides.  
 156 The median age of patients was 36 (range:6-75). Two of twenty patients suffered severe or critical  
 157 illness. Weifang sequences were combined with a diverse selection of sequences from China  
 158 outside of Weifang and other countries provided by GISAID *Elbe and Buckland-Merrett (2017)*. The  
 159 new Weifang sequences are deposited in GISAID (EPI\_ISL\_413691, EPI\_ISL\_413692, EPI\_ISL\_413693,  
 160 EPI\_ISL\_413694, EPI\_ISL\_413695, EPI\_ISL\_413696, EPI\_ISL\_413697, EPI\_ISL\_413711, EPI\_ISL\_413729,  
 161 EPI\_ISL\_413746, EPI\_ISL\_413747, EPI\_ISL\_413748, EPI\_ISL\_413749, EPI\_ISL\_413750, EPI\_ISL\_413751,  
 162 EPI\_ISL\_413752, EPI\_ISL\_413753, EPI\_ISL\_413761, EPI\_ISL\_413791, EPI\_ISL\_413809).

**Mathematical model.** The phylodynamic model is designed to account for nonlinear epidemic dynamics in Weifang, a realistic course of infection (incubation and infectious periods), migration of lineages in and out of Weifang, and variance in transmission rates which can influence epidemic size estimates. The model of epidemic dynamics within Weifang is based on a susceptible-exposed-infectious-recovered (SEIR) model. We elaborate the model with with an additional compartment  $J$  which has a higher transmission rate ( $\tau$ -fold higher) than the  $I$  compartment. Upon leaving the incubation period individuals progress to the  $J$  compartment with probability  $p_h$ , or otherwise to  $I$ .

The model is implemented as a system of ordinary differential equations:

$$\dot{S}(t) = -\beta(\beta I(t) + \beta \tau J(t)) \frac{S(t)}{S(t) + I(t) + J(t) + R(t)} \quad (1)$$

$$\dot{E}(t) = \beta(\beta I(t) + \beta \tau J(t)) \frac{S(t)}{S(t) + I(t) + J(t) + R(t)} - \gamma_0 E(t) \quad (2)$$

$$\dot{I}(t) = \gamma_0(1 - p_h)E(t) - \gamma_1 I(t) \quad (3)$$

$$\dot{J}(t) = \gamma_0 p_h E(t) - \gamma_1 J(t) \quad (4)$$

$$\dot{R}(t) = \gamma_1(E(t) + J(t)) \quad (5)$$

We also model an exponentially growing reservoir  $Y(t)$  for imported lineages in to Weifang. The equation governing this population is

$$\dot{Y}(t) = (\rho - \mu)Y(t). \quad (6)$$

163 Migration is modeled as a bidirectional process which only depends on the size of variables in  
164 the Weifang compartment and thus migration does not influence epidemic dynamics; it will only  
165 influence the inferred probability that a lineage resides within Weifang. For a compartment  $X$  ( $E$ ,  
166 or  $J$ ),  $\eta$  is the per lineage rate of migration out of Weifang and the total rate of migration in and out  
167 of Weifang is  $\eta X$ .

168 During phylodynamic model fitting  $\beta$  and  $\rho$  are estimated. Additionally, we estimate initial sizes  
169 of  $Y$ ,  $E$ , and  $S$ . Other parameters are fixed based on prior information. We fix  $1/\gamma_0 = 4.1$  days and  
170  $1/\gamma_1 = 3.8$  days. We set  $p_h = 0.20$  and  $\tau = 74$  which yields a dispersion of the reproduction number  
171 that matches a negative binomial distribution with  $k = 0.22$  if  $R_0 = 2$ , similar to values estimated for  
172 the 2003 SARS epidemic (*Lloyd-Smith et al., 2005*).

173 **Phylogenetic analysis.** We aligned the 20 Weifang sequences using MAFFT(*Katoh and Standley,*  
174 *2013*) with a previous alignment of 35 SARS-CoV 2 sequences from outside of Weifang(*Volz et al.,*  
175 *2020*). Maximum likelihood analysis was carried using IQTree(*Minh et al., 2019*) with a HKY+G4  
176 substitution model and a time-scaled tree was estimated using treedater 0.5.0(*Volz and Frost, 2017*).  
177 Two outliers according to the molecular clock model were identified and removed using 'treedater'  
178 which was also used to compute the root to tip regression.

179 Bayesian phylogenetic analysis was carried out using BEAST 2.6.1(*Bouckaert et al., 2019*) using a  
180 HKY+G4 substitution model and a strict molecular clock. The phylodynamic model was implemented  
181 using the PhyDyn package(*Volz and Siveroni, 2018*) using the QL likelihood approximation and the  
182 RK ODE solver. The model was fitted by running 8 MCMC chains in parallel and combining chains  
183 after removing 50% burn-in.

184 The *ggtree* package was used for all phylogeny visualizations(*Yu et al., 2017*).

185 Code to replicate this analysis and and BEAST XML files can be found at <https://github.com/emvolz/weifang-sarscov2>.  
186

## 187 Funding

188 This work was supported by Centre funding from the UK Medical Research Council under a con-  
189 cordat with the UK Department for International Development. NIHR. J-IDEA. This work was also

190 supported by a grant from the Special Project for Prevention and Control of Pneumonia of New  
 191 Coronavirus Infection in Weifang Science and Technology Development Plan in 2020 (2020YQFK015)  
 192 to Associate Senior Technologist Qing Nie. Role of the Funders: All funders of the study had no role  
 193 in study design, data analysis, data interpretation, or writing of the report.

## 194 Acknowledgements

195 We gratefully acknowledge China National GeneBank at Shenzhen, China for the sequencing strategy  
 196 and capacity support. We also gratefully acknowledge the laboratories that have contributed  
 197 publicly available genomes via GISAID: Shanghai Public Health Clinical Center & School of Public  
 198 Health, Fudan University, Shanghai, China, at the National Institute for Viral Disease Control and  
 199 Prevention, China CDC, Beijing, China, at the Institute of Pathogen Biology, Chinese Academy of  
 200 Medical Sciences & Peking Union Medical College, Beijing, China, at the Wuhan Institute of Virology,  
 201 Chinese Academy of Sciences, Wuhan, China, at the Department of Microbiology, Zhejiang Provincial  
 202 Center for Disease Control and Prevention, Hangzhou, China, at the Guangdong Provincial Center  
 203 for Diseases Control and Prevention at the Department of Medical Sciences, at the Shenzhen Key  
 204 Laboratory of Pathogen and Immunity, Shenzhen, China, at the Hangzhou Center for Disease and  
 205 Control Microbiology Lab, Zhejiang, China, at the National Institute of Health, Nonthaburi, Thailand,  
 206 at the National Institute of Infectious Diseases, Tokyo, Japan, at the Korea Centers for Disease  
 207 Control & Prevention, Cheongju, Korea, at the National Public Health Laboratory, Singapore, at the  
 208 US Centers for Disease Control and Prevention, Atlanta, USA, at the Institut Pasteur, Paris, France,  
 209 at the Respiratory Virus Unit, Microbiology Services Colindale, Public Health England, and at the  
 210 Department of Virology, University of Helsinki and Helsinki University Hospital, Helsinki, Finland,  
 211 and at the University of Melbourne, Peter Doherty Institute for Infection and Immunity, Melbourne,  
 212 Australia, at the Victorian Infectious Disease Reference Laboratory, Melbourne, Australia, at the  
 213 Public Health Virology Laboratory, Brisbane, Australia and at the Institute of Clinical Pathology and  
 214 Medical Research, University of Sydney, Westmead, Australia.

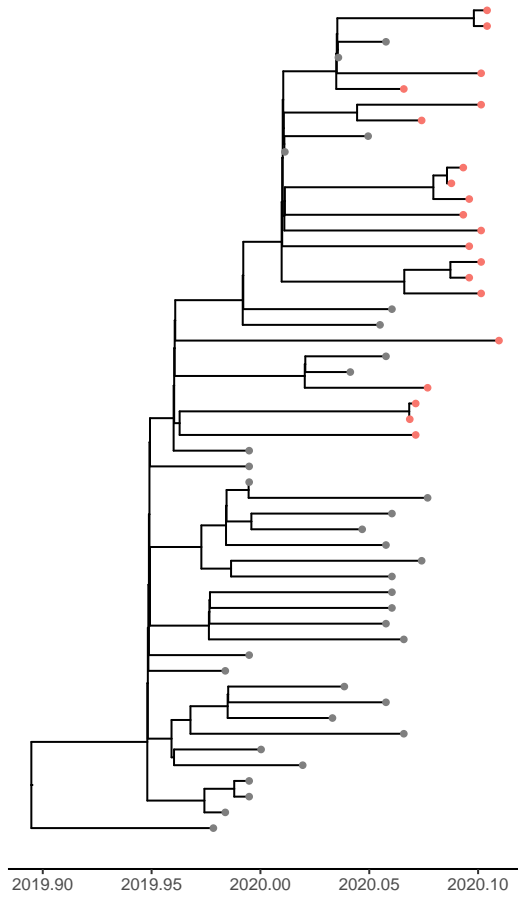
## 215 References

- 216 **Bedford T**, Cryptic transmission of novel coronavirus revealed by genomic epidemiology; 2020. Accessed:  
 217 2020-3-8. <https://bedford.io/blog/ncov-cryptic-transmission/>.
- 218 **Bouckaert R**, Vaughan TG, Barido-Sottani J, Duchêne S, Fourment M, Gavryushkina A, Heled J, Jones G, Kühnert  
 219 D, De Maio N, et al. BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. *PLoS*  
 220 *computational biology*. 2019; 15(4):e1006650.
- 221 **Dudas G**, Carvalho LM, Rambaut A, Bedford T. MERS-CoV spillover at the camel-human interface. *Elife*. 2018  
 222 Apr; 7.
- 223 **Elbe S**, Buckland-Merrett G. Data, disease and diplomacy: GISAID's innovative contribution to global health.  
 224 *Global Challenges*. 2017; 1(1):33–46.
- 225 **Gire SK**, Goba A, Andersen KG, Sealfon RSG, Park DJ, Kanneh L, Jalloh S, Momoh M, Fullah M, Dudas G, Wohl S,  
 226 Moses LM, Yozwiak NL, Winnicki S, Matranga CB, Malboeuf CM, Qu J, Gladden AD, Schaffner SF, Yang X, et al.  
 227 Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak. *Science*. 2014  
 228 Sep; 345(6202):1369–1372.



- 229 **Grantz K**, Metcalf J, Lessler J, Dispersion vs. Control [Internet].[cited 2020 Feb 12];
- 230 **Guan WJ**, Ni ZY, Hu Y, Liang WH, Ou CQ, He JX, Liu L, Shan H, Lei CL, Hui DSC, Du B, Li LJ, Zeng G, Yuen KY, Chen  
231 RC, Tang CL, Wang T, Chen PY, Xiang J, Li SY, et al. Clinical Characteristics of Coronavirus Disease 2019 in China.  
232 *N Engl J Med.* 2020 Feb; .
- 233 **Katoh K**, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance  
234 and usability. *Molecular biology and evolution.* 2013; 30(4):772–780.
- 235 **Keeling MJ**, Rohani P. *Modeling Infectious Diseases in Humans and Animals.* Princeton University Press; 2011.
- 236 **Li LM**, Grassly NC, Fraser C. Quantifying Transmission Heterogeneity Using Both Pathogen Phylogenies and  
237 Incidence Time Series. *Mol Biol Evol.* 2017 Nov; 34(11):2982–2995.
- 238 **Lloyd-Smith JO**, Schreiber SJ, Kopp PE, Getz WM. Superspreading and the effect of individual variation on  
239 disease emergence. *Nature.* 2005 Nov; 438(7066):355–359.
- 240 **Mao H.** Weifang City announces fever clinics. Weifang News Network. 2020 Jan; .
- 241 **Minh BQ**, Schmidt H, Chernomor O, Schrempf D, Woodhams M, von Haeseler A, Lanfear R. IQ-TREE 2: New  
242 models and efficient methods for phylogenetic inference in the genomic era; 2019.
- 243 **Stadler T**, Phylodynamic Analyses based on 11 genomes from the Italian outbreak; 2020. Accessed: 2020-3-8.  
244 <http://virological.org/t/phylogenetic-analyses-based-on-11-genomes-from-the-italian-outbreak/426>.
- 245 **Verity Hill AR**, Phylodynamic analysis of SARS-CoV-2 | Update 2020-03-06; 2020. Accessed: 2020-3-8. <http://virological.org/t/phylogenetic-analysis-of-sars-cov-2-update-2020-03-06/420>.
- 246
- 247 **Volz EM**, Frost SDW. Scalable relaxed clock phylogenetic dating. *Virus Evol.* 2017 Jul; 3(2).
- 248 **Volz E**, Baguelin M, Bhatia S, Boonyasiri A, Cori A, Cucunubá Z, Cuomo-Dannenburg G, Donnelly CA, Dorigatti  
249 I, Fitzjohn R, Fu H, Gaythorpe K, Ghani A, Hamlet A, Hinsley W, Imai N, Laydon D, Nedjati-Gilani L Gemma  
250 abd Okell, Riley S, van Elsland S, et al. Report 5: Phylogenetic analysis of SARS-CoV-2; 2020.
- 251 **Volz EM**, Siveroni I. Bayesian phylodynamic inference with complex models. *PLoS Comput Biol.* 2018 Nov;  
252 14(11):e1006546.
- 253 **Yu G**, Smith DK, Zhu H, Guan Y, Lam TTY. ggtree : an R package for visualization and annotation of phylogenetic  
254 trees with their covariates and other associated data. *Methods Ecol Evol.* 2017 Jan; 8(1):28–36.
- 255 **Zhou P**, Yang XL, Wang XG, Hu B, Zhang L, Zhang W, Si HR, Zhu Y, Li B, Huang CL, Chen HD, Chen J, Luo Y, Guo H,  
256 Jiang RD, Liu MQ, Chen Y, Shen XR, Wang X, Zheng XS, et al. A pneumonia outbreak associated with a new  
257 coronavirus of probable bat origin. *Nature.* 2020 Feb; .

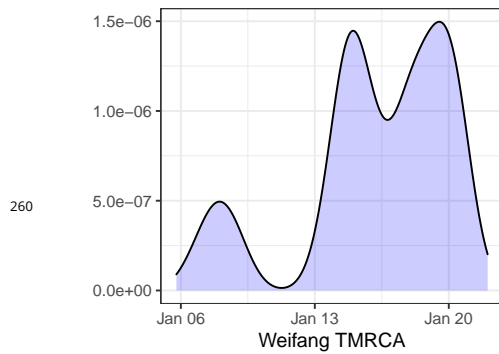
258



**Figure 1-Figure supplement 1.** A time scaled phylogeny estimated using IQTree and treedater and using the same data as used for the Bayesian analysis.



**Figure 1-Figure supplement 2.** A tree density plot based on the posterior distribution of trees computed in BEAST2.



**Figure 1-Figure supplement 3.** The estimated posterior TMRCA among all Weifang lineages.