

1 **An ancient coronavirus-like epidemic drove adaptation in East Asians from**  
2 **25,000 to 5,000 years ago**

3

4 Yassine Souilmi<sup>1,2</sup>, M. Elise Lauterbur<sup>3</sup>, Ray Tobler<sup>1</sup>, Christian D. Huber<sup>1</sup>, Angad S. Johar<sup>1</sup>,  
5 David Enard<sup>3</sup>

6

7 <sup>1</sup> Australian Centre for Ancient DNA, School of Biological Sciences, University of Adelaide,  
8 Adelaide, SA 5005, Australia.

9 <sup>2</sup> National Centre for Indigenous Genomics, Australian National University, Canberra, ACT  
10 0200, Australia.

11 <sup>3</sup> University of Arizona Department of Ecology and Evolutionary Biology, Tucson, Arizona, USA.

12

## 13 Summary

14 The current SARS-CoV-2 pandemic has emphasized the vulnerability of human  
15 populations to novel viral pressures, despite the vast array of epidemiological and  
16 biomedical tools now available. Notably, modern human genomes contain evolutionary  
17 information tracing back tens of thousands of years, which may help identify the viruses  
18 that have impacted our ancestors – pointing to which viruses have future pandemic  
19 potential. Here, we apply evolutionary analyses to human genomic datasets to recover  
20 selection events involving tens of human genes that interact with coronaviruses,  
21 including SARS-CoV-2, that started 25,000 years ago. These adaptive events were  
22 limited to ancestral East Asian populations, the geographical origin of several modern  
23 coronavirus epidemics. An arms race with an ancient corona-like virus may thus have  
24 taken place in ancestral East Asian populations. By learning more about our ancient  
25 viral foes, our study highlights the promise of evolutionary information to combat the  
26 pandemics of the future.

27

## 28 Introduction

29 In the past 20 years, strains of the beta coronavirus genus (family Coronaviridae; Richman et  
30 al., 2020) have been behind three major zoonotic outbreaks with grave impacts for human  
31 populations (Ou et al., 2020). The first outbreak, commonly known as SARS-CoV (Severe Acute  
32 Respiratory Syndrome), originated in China in late 2002 and eventually spread to 30 additional  
33 counties where it infected more than 8,000 people and claimed nearly 800 lives (Hoffmann and  
34 Kamps, 2003). Four years later, MERS-CoV (Middle East respiratory syndrome coronavirus)  
35 affected >2,400 people and caused over 850 deaths, mostly in Saudi Arabia (World Health  
36 Organization, 2019). The most recent outbreak began in late 2019 when SARS-CoV-2 – a less  
37 virulent but far more contagious strain than those behind the two previous epidemics – emerged  
38 in mainland China before spreading rapidly across the rest of the world, triggering an ongoing  
39 pandemic (COVID-19) that so far has infected 45 million people and resulted in over one million  
40 deaths worldwide (Dong et al., 2020).

41 The devastation caused by SARS-CoV-2 has inspired a worldwide research effort to develop  
42 new vaccines and strategies that aim to curb its impact by determining the factors that underlie  
43 its epidemiology. The resulting research has revealed that socioeconomic (e.g. access to  
44 healthcare and testing facilities), demographic (e.g. population density and age structure), and  
45 personal health factors all play a major role in SARS-CoV-2 epidemiology (Balogun et al., 2020;  
46 Sattar Naveed et al., 2020; Scarpone et al., 2020). Additionally, several genetic loci that mediate  
47 SARS-CoV-2 susceptibility and severity have been found in contemporary European  
48 populations (Ellinghaus et al., 2020; Roberts et al., 2020), one of which contains a genetic  
49 variant that increases SARS-CoV-2 susceptibility that was likely introduced into the ancestors of  
50 modern Europeans after they interbred with Neanderthals ~40,000 years ago (Zeberg and  
51 Pääbo, 2020). This historical admixture event has led to genetic differences within and between  
52 contemporary human populations that directly impact COVID-19 epidemiology – the  
53 Neanderthal-derived variant haplotype is now carried by 8% of modern Europeans, but at lower  
54 frequencies in African populations whose ancestors did not experience this admixture event –  
55 and suggests that evolutionary analyses of human populations may help reveal these genetic  
56 differences and ultimately assist in the development of novel drugs and therapies to combat the  
57 negative impacts of SARS-CoV-2.

58 Throughout the evolutionary history of our species, positive natural selection has frequently  
59 targeted proteins that physically interact with viruses – e.g. those involved in immunity, or used  
60 by viruses to hijack the host cellular machinery (Barreiro et al., 2009; Enard et al., 2016; Sawyer  
61 et al., 2005). In the ~6 million years since the ancestors of humans and chimpanzees  
62 separated, selection has led to the fixation of gene variants encoding virus-interacting proteins  
63 (VIPs) at three times the rate observed for other classes of genes (Enard et al., 2016; Uricchio  
64 et al., 2019). Moreover, strong selection on VIPs has continued in human populations during the  
65 past 50,000 years, as evidenced by VIP genes being enriched for adaptive introgressed  
66 Neanderthal variants and also selective sweep signals (i.e. selection that drives a beneficial  
67 variant to substantial frequencies in a population), particularly around VIPs that interact with  
68 RNA viruses, a viral class that includes the coronaviruses (Enard and Petrov, 2018, 2020).

69 The accumulated evidence suggests that ancient viral epidemics have occurred frequently  
70 during the history of our species; however, we currently do not know if selection has made a  
71 substantial contribution to individual- and population-level differences in the SARS-CoV-2  
72 response in modern humans. Indeed, it is possible that many ancient epidemics were restricted

73 to specific ancestral human populations that will have resulted in differences in the distribution  
74 of genetic variants that mediate viral epidemiology in modern human populations.

75 Accordingly, here we investigate whether ancient coronavirus epidemics have created genetic  
76 differences within and across modern human populations, by examining if selection signals are  
77 enriched within a set of 420 VIPs specific to the coronavirus (denoted CoV-VIPs; Table S1)  
78 across 26 worldwide human populations from the 1000 Genomes Project (1000 Genomes  
79 Project Consortium, 2015). These CoV-VIPs comprise 332 SARS-CoV-2 VIPs that were  
80 recently identified by high-throughput mass spectrometry (Gordon et al., 2020) and an additional  
81 88 proteins that were manually curated from the coronavirus literature (e.g. SARS-CoV-1,  
82 MERS, HCoV-NL63, etc; Table S1; Enard and Petrov, 2018), and form part of a larger set of  
83 5,291 previously published VIPs (SI; Table S1) from multiple viruses known to infect humans  
84 (Enard and Petrov, 2018). Our focus upon host adaptation at VIPs is motivated by evidence  
85 indicating that these protein interactions are the central mechanism that viruses use to hijack  
86 the host cellular machinery (Enard and Petrov, 2018). Accordingly, VIPs are much more likely to  
87 have functional impacts on viruses than proteins not known to interact with viruses (see SI: *Host*  
88 *adaptation is expected at VIPs*). Our enrichment-based approach is expected to be particularly  
89 powerful if the ancestors of one or more of the 26 modern human populations were exposed to  
90 epidemics driven by coronavirus-like viruses that resulted in selection upon multiple CoV-VIPs  
91 (see Discussion).

92 Our analyses of CoV-VIPs find a strong enrichment in sweep signals in these proteins across  
93 multiple East Asian populations, which is absent from other human populations. This suggests  
94 that an ancient coronavirus epidemic drove an adaptive response in the ancestors of East  
95 Asians, which is in strong agreement with the current geographic range of the major known  
96 animal reservoirs of coronaviruses (Wong et al., 2019). Further, by leveraging ancestral  
97 recombination graph approaches (Speidel et al., 2019; Stern et al., 2019) we find that amongst  
98 the putatively selected CoV-VIPs, 42 first came under selection around 900 generations  
99 (~25,000 years) ago and exhibit a coordinated adaptive response that lasted until around 200  
100 generations (~5,000 years) ago. By drawing upon other publicly available datasets, we show  
101 that the CoV-VIP genes are enriched for anti- and proviral effects and variants that affect  
102 COVID-19 etiology in modern European populations (<https://grasp.nhlbi.nih.gov/Covid19GWAS>  
103 [Results.aspx](#)), and that the inferred underlying causal mutations are situated near to regulatory  
104 variants active in lungs and other tissues negatively impacted by COVID-19. Taken together,

105 these independent lines of evidence provide strong support for an ancient coronavirus (or  
106 related virus) epidemic that emerged 25,000 years ago in the ancestors of contemporary East  
107 Asian populations, whose genetic signature remains apparent in the genomes of the present-  
108 day populations now living in this region.

109 The SARS-CoV-2 pandemic represents a great threat for humanity, a point clearly emphasized  
110 by the COVID-19 Host Genetics Initiative, who stated that “insights into how to better  
111 understand and treat COVID-19 are desperately needed” and that “it is critical for the scientific  
112 community to come together around this shared purpose” (<https://www.covid19hg.org>; The  
113 COVID-19 Host Genetics Initiative, 2020). In this study, we show how by extracting evolutionary  
114 information from publicly available genomic datasets we can achieve a small but significant step  
115 towards this important goal, and highlight the potential for similar analyses to play an important  
116 role in the future of medical research.

## 117 **Results**

### 118 **Signatures of adaptation to an ancient epidemic**

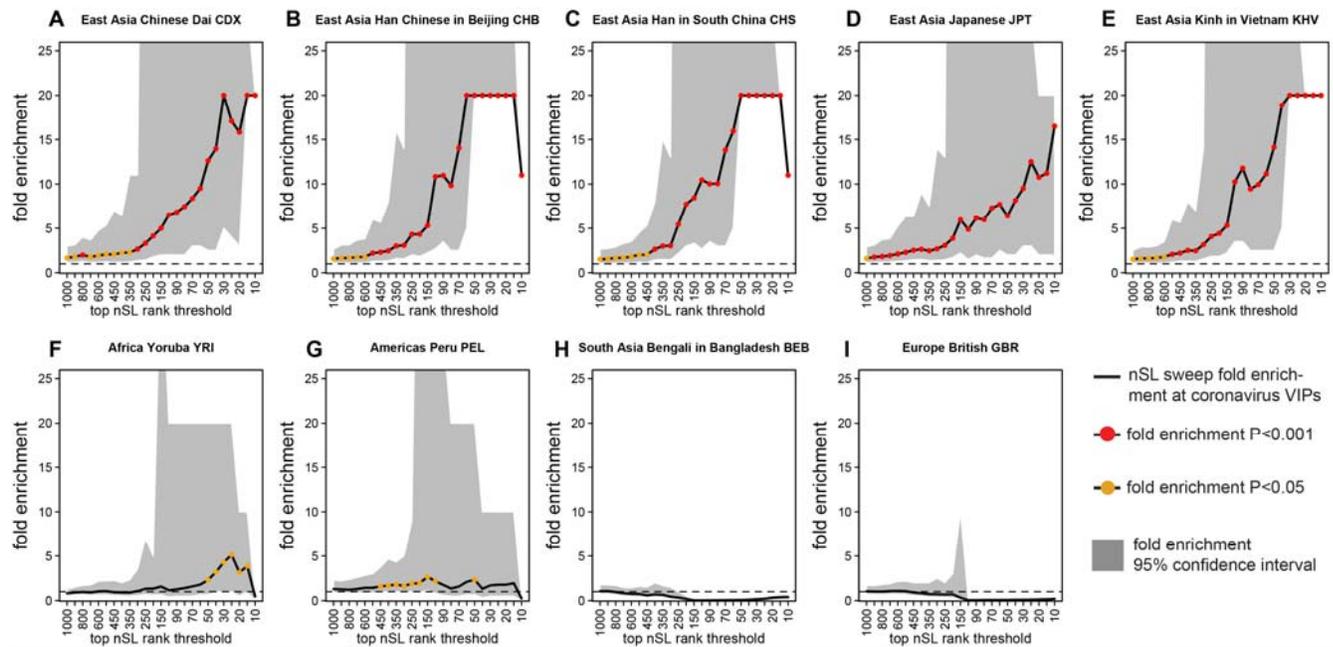
119 Viruses have exerted strong selective pressures on the ancestors of modern humans (Enard  
120 and Petrov, 2020; Uricchio et al., 2019). Accordingly, we use two population genetic statistical  
121 tests that are sensitive to such genetic signatures (i.e. selective sweeps) – nSL (Ferrer-Admetlla  
122 et al., 2014) and iHS (Voight et al., 2006) – and which are able to detect genomic regions  
123 impacted by strong selection across a wide range of parameters (e.g. different starting and end  
124 frequencies of the selected allele). Both statistics also have the advantage of being insensitive  
125 to background selection (Enard et al., 2014; Schrider, 2020), thereby reducing the potential  
126 impact of false positives in our analyses.

127 After scanning each of the 26 populations for signals of selection, we apply an enrichment test  
128 that was previously used to detect enriched selection signals in RNA VIPs in human populations  
129 (Enard and Petrov, 2020). Briefly, for each population and selection statistic, we rank all genes  
130 based on the average selection statistic score observed in genomic windows ranging from 50kb  
131 to 2Mb. Different windows sizes are used because smaller windows tend to be more sensitive to  
132 weaker sweeps, whereas larger windows tend to be more sensitive to stronger sweeps (Enard  
133 and Petrov, 2020; STAR Methods). After ranking the gene scores, we estimate an enrichment  
134 curve (Figure 1) for gene sets ranging from the top 10 to 10,000 ranked loci (STAR Methods).

135 Enrichment is then calculated for each set of top ranked genes using a block-randomization  
136 approach that accounts for the genomic clustering of neighboring CoV-VIPs (Enard and Petrov,  
137 2020; STAR Methods). For our control gene set, we use protein-coding genes situated at least  
138 500kb from CoV-VIPs to avoid overlapping the same sweep signals. Additionally, genes in the  
139 control sets are chosen to have similar characteristics as the CoV-VIPs (e.g. similar  
140 recombination rates, density of coding and regulatory sequences, percentage of immune genes,  
141 percentage of genes that interact with bacteria; see STAR Methods for the complete list of  
142 factors) to ensure that any detected enrichment is virus-specific rather than due to a  
143 confounding factor (Enard and Petrov, 2020). Finally, we also exclude the possibility that  
144 functions other than viral interactions might explain our results by running a Gene Ontology  
145 analysis (Gene Ontology Consortium, 2015; SI; Tables S2, S3 and Figure S1).

146 Applying this approach to each of the 26 human populations from the 1,000 genomes dataset,  
147 we find a very strong enrichment of sweep signals in CoV-VIPs across all top-ranked gene set  
148 sizes that is specific to the five East Asian populations (whole enrichment curve for nSL and iHS  
149 combined  $P=2.10^{-4}$ ; Figures 1 & S2; STAR Methods). No enrichment is observed for populations  
150 from other continental regions, including in neighboring South Asia (whole enrichment curve for  
151 nSL and iHS combined  $P>0.05$  in all cases; Figures 1 & S2). Further, no enrichment is detected  
152 for VIP sets for 17 other viruses in East Asian populations (whole enrichment curve for nSL and  
153 iHS separately or combined,  $P>0.05$  in all cases; Figures S3 & S4). Taken together, these  
154 results suggest that coronaviruses, or a phylogenetically related viral taxon, have driven ancient  
155 epidemics in ancient human populations that are ancestral to modern East Asians, and this  
156 enrichment is unlikely to have been caused by any other virus represented in our set of 5,291  
157 VIPs (Table S1). The enrichment is most substantial for the top-ranked gene sets ranging  
158 between the top 10 and top 1,000 loci (Figure 1; whole enrichment curve  $P=3.10^{-6}$  for nSL,  
159  $P=4.10^{-3}$  for iHS,  $P=6.10^{-5}$  for iHS and nSL combined), and is particularly strong for the top 200  
160 loci in large windows (1 Mb) where a four-fold enrichment is observed for both nSL and iHS  
161 statistics (pertaining to between 10 to 13 selected CoV-VIPs amongst the top 200 ranked  
162 genes; Table S4). This suggests that strong selection targeted multiple CoV-VIPs in the  
163 common ancestors of modern East Asian populations. That the selected haplotype structures  
164 are detected by both the iHS and nSL methods suggests that they are unlikely to have occurred  
165 prior to 30,000 years ago, as both nSL and iHS have little power to detect adaptive events  
166 arising before this time point in human evolution (Sabeti et al., 2006).

167



168  
 169 **Figure 1. Coronavirus VIPs nSL ranks enrichment** A,B,C,D,E are East Asian populations,  
 170 F,G,H,I are populations from other continents. The y axis represents the bootstrap test (STAR  
 171 Methods) relative fold enrichment of the number of genes in putative sweeps at CoV-VIPs,  
 172 divided by the number of genes in putative sweeps at control genes matched for multiple  
 173 confounding factors. The x axis represents the top rank threshold to designate putative sweeps.  
 174 Black full line: average fold enrichment over 5,000 bootstrap test control sets. Fold enrichments  
 175 greater than 20 are represented at 20. Grey area: 95% confidence interval of the fold  
 176 enrichment over 5,000 bootstrap test control sets. The rank thresholds where the confidence  
 177 interval lower or higher fold enrichment has a denominator of zero are not represented (For  
 178 example, graph B, top 10 rank threshold). Lower confidence interval fold enrichments higher  
 179 than 20 are represented at 20 (for example, graph B, top 30 rank threshold). Red dots:  
 180 bootstrap test fold enrichment P < 0.001. Orange dots: bootstrap test fold enrichment P < 0.05.  
 181 Note that the bootstrap test p-values are not the same as the whole curve enrichment p-value  
 182 estimated using block-randomized genomes on top of the bootstrap test (STAR Methods).

183 **An ancient epidemic in the ancestors of East Asians starting around 25,000 years ago**

184 To further unravel the history of the coronavirus-like epidemics in the ancestors of East Asians,  
 185 we use a recent ancestral recombination graph (ARG)-based method, Relate (Speidel et al.,  
 186 2019), to infer the timing and trajectories of selected loci for the CoV-VIPs. By estimating ARGs  
 187 at variants distributed across the entire genome, Relate can reconstruct coalescent events  
 188 across time and detect genomic regions impacted by positive selection while explicitly  
 189 controlling for historical variation in population demography. To approximate the start time of  
 190 selection, Relate estimates the first historical time point that a putatively selected variant had an  
 191 observable frequency greater than zero (STAR Methods). We use this approximation as the  
 192 likely starting time of selection, although we note that this method does not account for selection

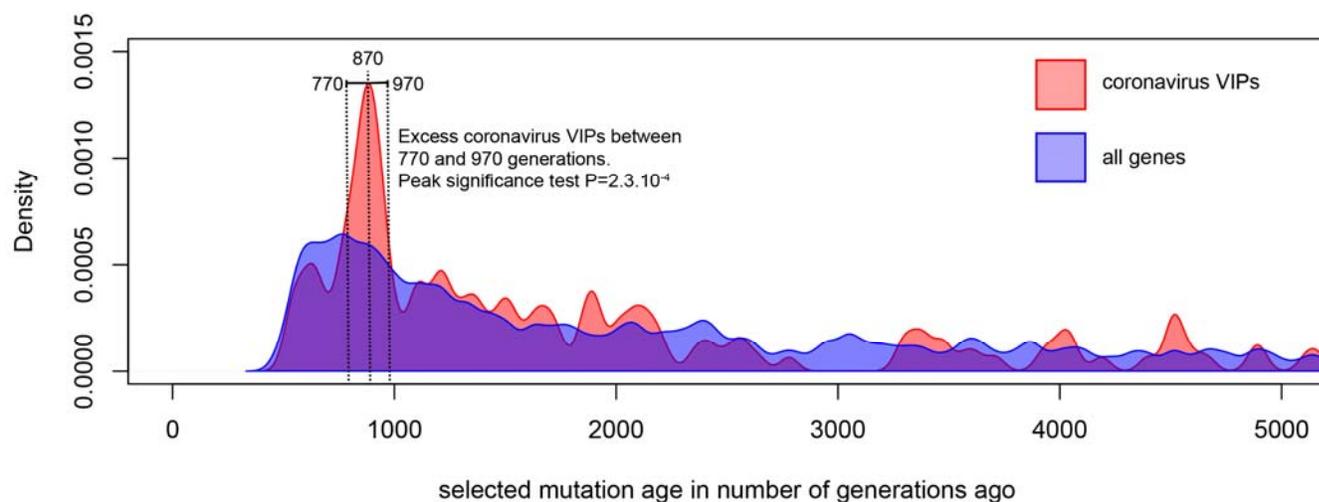
193 on standing variants that had non-zero frequencies at the onset of selection (STAR Methods).  
194 Additionally, we use the iSAFE software – which enables the localization of selected mutations  
195 (Akbari et al., 2018) – along with a curated set of regulatory variants (expression QTLs; eQTLs)  
196 from the eGTEX Project (2017) to help identify the likely causal mutations in the selected CoV-  
197 VIP genes. There is good evidence that the majority of adaptive mutations in the human  
198 genome are regulatory mutations (Enard et al., 2014; Kudaravalli et al., 2009; Nédélec et al.,  
199 2016; Quach et al., 2016) and, accordingly, we find that iSAFE peaks are significantly closer to  
200 GTEx eQTLs proximal to CoV-VIP genes than expected by chance (iSAFE peak proximity test,  
201  $P < 10^{-9}$ ; STAR Methods). Therefore, for each CoV-VIP gene, we chose a variant with a low  
202 Relate p-value ( $< 10^{-3}$ ; STAR Methods) that is situated at or close to a GTEx eQTL associated  
203 with the focal gene to estimate the likely starting time of selection for that gene (STAR Methods;  
204 Figure S5).

205 Using this approach, we observe 42 CoV-VIPs (Table S5 and Figure S5) with selection starting  
206 times clustered around a peak 870 generations ago (~200 generations wide, potentially due to  
207 noise in our estimates; Figure 2). While this amounts to about four times more selected CoV-  
208 VIP genes than were detected using either nSL or iHS (both detected around ten CoV-VIPs  
209 amongst the top 200 ranked genes; Table S4) this is not unexpected as Relate has more power  
210 to detect selection events than nSL and iHS when the beneficial allele is at intermediate  
211 frequencies at the point of measurement (typically  $< 60\%$ ; Figure 3; Enard and Petrov, 2020;  
212 Ferrer-Admetlla et al., 2014; Voight et al., 2006). The relatively tight temporal clustering of  
213 starting times forms a highly significant peak (peak significance test  $P = 2.3 \cdot 10^{-4}$ ; Figure 2) when  
214 comparing the observed clustering of CoV-VIPs start times with the distribution of inferred start  
215 times for randomly sampled sets of genes (STAR Methods). Note that this peak significance test  
216 is gene clustering-aware (STAR Methods). Further, this significance test is not biased by the  
217 fact that CoV-VIPs are enriched for sweep signals, as the test remains highly significant  
218 ( $P = 1 \cdot 10^{-4}$ ) when using random control sets with comparable high-scoring nSL statistics (STAR  
219 Methods). This suggests that the tight temporal clustering of selection events is a specific  
220 feature of the CoV-VIPs, rather than a confounding aspect of any gene set similarly enriched for  
221 sweeps.

222

223

224



225

## 226 **Figure 2. Timing of selection at CoV-VIPs**

227 The figure shows the distribution of selection start times at CoV-VIPs (pink distribution)  
228 compared to the distribution of selection start times at all loci in the genome (blue distribution).  
229 Details on how the two distributions are compared by the peak significance test, and how the  
230 selection start times are estimated with Relate, are provided in STAR Methods.

231 The genes with clustered selection starting times around 900 generations ago are enriched in  
232 strong nSL signals, as shown by running the peak significance test using only CoV-VIPs and  
233 controls with strong nSL signals (Figure S6). Conversely, the peak disappears when restricting  
234 this test to weaker nSL signals ( $P=0.53$  when using the lowest 50% of nSL statistics; STAR  
235 Methods). Importantly, our estimates of the timing of selection are not biased by our use of  
236 methods that rely on selected variants not being fixed in the population at the time of genome  
237 sampling (i.e. nSL and iHS). When rerunning our analytical pipeline focusing only on strong  
238 candidate loci according to Tajima's D (Tajima, 1989), a statistic developed to detect complete  
239 sweeps (i.e. fixed mutations), we observe the same clustering of selection events starting  
240 around 900 generations ago (Figure S7). Further, the remaining 382 CoV-VIPs that are not part  
241 of this temporal cluster around 900 generations ago are not more likely to have significant  
242 Tajima's D values than controls (whole enrichment curve  $P=0.07$ ). Consequently, our results are  
243 consistent with the emergence of a coronavirus epidemic ~900 generations, or ~25,000 years  
244 (900 generations \* 28 years per generation; Moorjani et al., 2016), ago that drove a burst of  
245 strong positive selection in the ancestors of East Asians, which may represent a genetic record  
246 of a multi-generational coronavirus (or related virus) epidemic amongst the 26 human  
247 populations tested here.

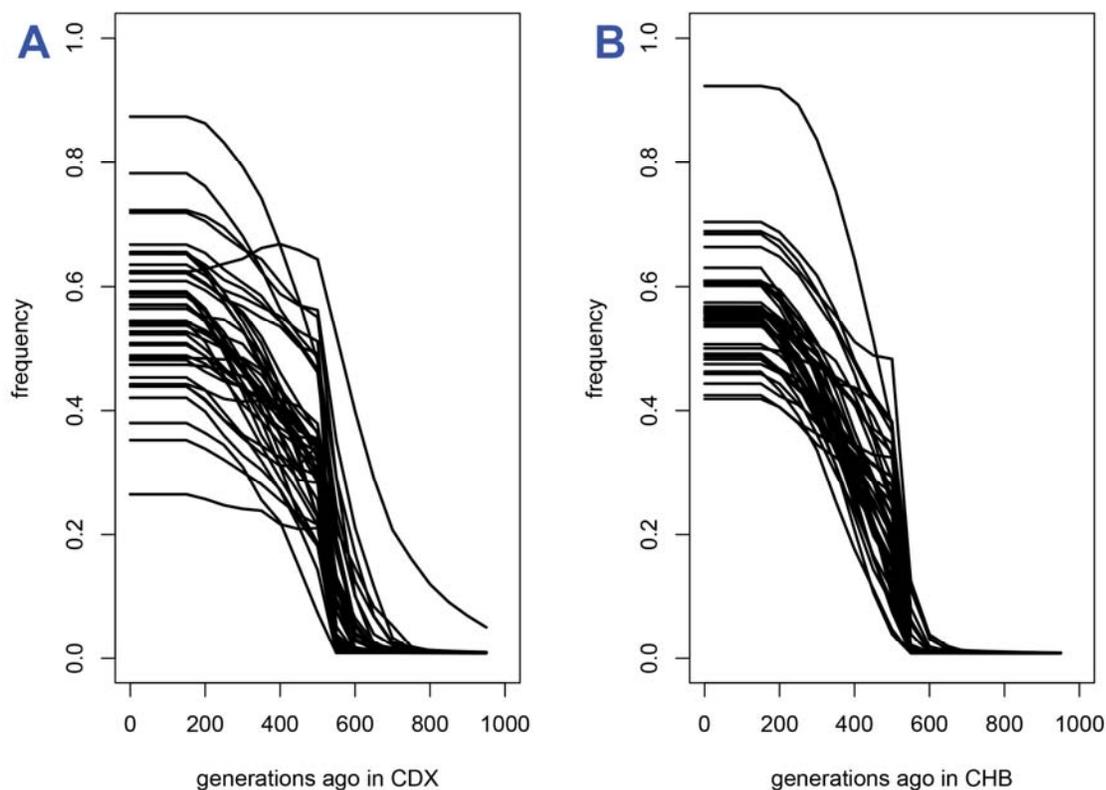
## 248 **Strong selection drove coordinated changes in multiple CoV-VIP genes over 20,000 years**

249 To learn more about the likely duration of the selection pressure acting on the ancestors of East  
250 Asians, we use CLUES (Stern et al., 2019) to infer allele frequency trajectories and selection  
251 coefficients for the inferred beneficial mutations proximal to the 42 CoV-VIP genes with  
252 selection starting 900 generations ago (Figure 3). CLUES uses the temporal variation in  
253 population size and coalescence rates inferred by Relate to reconstruct frequency trajectories  
254 while taking demographic fluctuations into account. Our observation of sweep signals at 42  
255 CoV-VIP genes in the ancestors of East Asians suggests that the putative underlying viral  
256 epidemic likely spanned multiple generations (i.e. the time needed for selection to drive initially  
257 rare alleles to intermediate/high frequencies). Accordingly, we anticipate that selection was  
258 probably strongest when the naive host population was first infected by the virus, before  
259 gradually waning as the host population adapted to the viral pressure (Hayward and Sella,  
260 2019). Similarly, a decrease in the virulence of the virus over time, a phenomenon that has been  
261 reported during the long term bouts of host-virus coevolution (Best and Kerr, 2000), would also  
262 result in the gradual decrement of selection coefficients across time. Hence, for each of the 42  
263 CoV-VIPs predicted to have started coming under selection ~900 generations ago, we use  
264 CLUES to estimate the selection coefficient in two successive time-intervals (between 1,000  
265 and 500 generations ago, and from 500 generations ago to the present), predicting that  
266 selection would be stronger in the oldest interval. We note that a 500 generation interval was  
267 reported as the approximate timespan that CLUES provides reliable estimates for humans  
268 (Stern et al., 2019); however, intervals of this length are not adequate to obtain reasonable  
269 estimates of the precise duration of the selective pressure (Stern et al., 2019), so we do not  
270 attempt to estimate this parameter here. Also, because CLUES uses a computationally  
271 intensive algorithm, following the recommendations of Stern et al. (2020) we base our estimates  
272 on only two of the five East Asian populations (i.e. Dai and Beijing Han Chinese; Figure 3A, B  
273 respectively).

274 Using this two-interval configuration, CLUES infers that beneficial mutations experienced an  
275 exponential increase in frequency between 800 and 500 generations ago for most of the CoV-  
276 VIPs, which was preceded by a more moderate establishment phase (Figure 3A, B). The  
277 selected mutations are estimated to have continually increased in frequency until ~200  
278 generations (approximately 5,000 years) ago, after which they remained relatively stable (Figure  
279 3A, B). Accordingly, CLUES estimates very high selection coefficients in the interval between  
280 1,000 and 500 generations ago (Dai average  $s = 0.05$ , Beijing Han  $s = 0.11$ ), but much weaker

281 selection coefficients from 500 generations ago up to the present (Dai average  $s = 0.003$ ,  
282 Beijing Chinese  $s = 0.004$ ; Figure 4A, B). These patterns are consistent with the appearance of  
283 a strong selective pressure that triggered a coordinated adaptive response across multiple  
284 independent loci, which waned through time as the host population adapted to the viral pressure  
285 and/or as the virus became less virulent. Although selective pressures other than a coronavirus  
286 or related virus might also contribute to these patterns, we note that the signal is restricted  
287 specifically at CoV-VIPs and none of 17 other viruses that we tested exhibit the same temporal  
288 clustering ~900 generations ago in East Asia (peak significance test  $P > 0.05$  in all cases; STAR  
289 Methods). Further, this test remained highly significant when retesting the temporal clustering of  
290 CoV-VIPs using only other RNA VIPs as the control set ( $P = 4.10 \cdot 10^{-4}$ ; Table S1), consistent with  
291 the clustered selection signals being a coordinated adaptive response to coronaviruses or a  
292 phylogenetically related viral taxon.

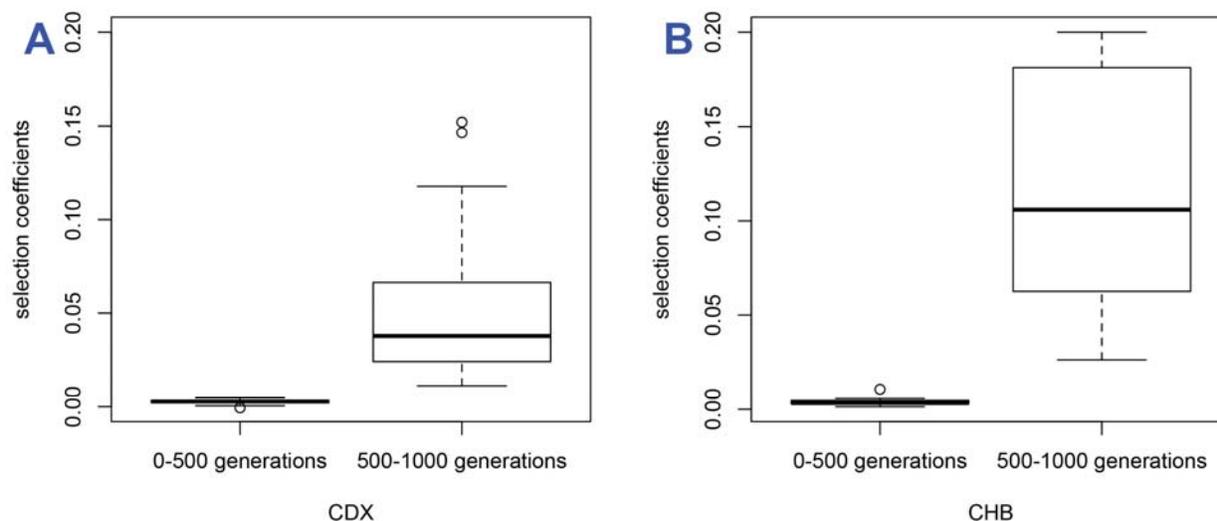
293



294

295 **Figure 3. Coronavirus selected VIPs allele frequency trajectories over time estimated by**  
296 **CLUES**

297 Each frequency trajectory is for one of the 42 Relate selected mutations at CoV-VIPs within the  
298 peak around 900 generations ago (STAR Methods). A) Frequency trajectories in the Chinese  
299 Dai CDX 1,000 Genomes population. B) Frequency trajectories in the Han Chinese from Beijing  
300 CHB 1,000 Genomes population.



301

#### 302 **Figure 4. Coronavirus selected VIPs selection coefficients estimated by CLUES**

303 This figure shows classic R boxplots of selected coefficients at the 42 Relate selected mutations  
304 within the peak around 900 generations ago (STAR Methods). A) Selection coefficients in the  
305 Chinese Dai CDX 1,000 Genomes population. B) Selection coefficients in the Han Chinese from  
306 Beijing CHB 1,000 Genomes population. Left: average selection coefficients between 0 and 500  
307 generations ago. Right: average selection coefficients between 500 and 1,000 generations ago.

#### 308 **Selected CoV-VIPs are enriched for antiviral and proviral factors**

309 To further clarify that an ancient viral epidemic caused the strong burst of selection we observe  
310 ~900 generations ago in the ancestors of East Asians, and not another ecological pressure  
311 acting on the same set of genes, we test if the 42 selected CoV-VIPs are enriched for genes  
312 with anti- or proviral effects relative to other CoV-VIPs (i.e. loci that are known to have a  
313 detrimental or beneficial effect on the virus, respectively). Because the relevant literature for  
314 coronaviruses is currently limited – which also applies to the relatively recent SARS-CoV-2 virus  
315 – we extend our set of anti- and proviral loci beyond those associated with coronaviruses to  
316 include loci reported for diverse viruses with high confidence from the general virology literature  
317 (see SI: *Host adaptation is expected at VIPs*; Table S1). We find that 21 (50%) of the 42 CoV-  
318 VIPs that came under selection ~900 generations ago have high-confidence anti- or proviral  
319 effects (vs. 29% for all 420 CoV-VIPs), a significant inflation in anti- and proviral effects  
320 (hypergeometric test  $P=6.10^{-4}$ ) that further supports our claim that the underlying selective  
321 pressure was most likely a viral epidemic.

322 **Selected mutations lie near regulatory variants active in SARS-CoV-2 affected tissues**

323 Coronavirus infections in humans are known to have pathological consequences for specific  
324 bodily tissues, whereby we investigate if the genes targeted by selection in the ancestors of  
325 East Asians are also enriched for regulatory functions in similar tissues. In light of our finding  
326 that many putative causal mutations in CoV-VIPs were proximal to eQTLs, we investigate  
327 whether selected mutations are situated closer to eQTLs for a specific tissue than expected by  
328 chance, as this would indicate that the tissue was negatively impacted by the virus (prompting  
329 the adaptive response). Briefly, we estimate a proximity-based metric that quantifies the  
330 distance between the location of the causal mutation estimated by iSAFE and the tissue-specific  
331 eQTLs for the 42 loci that likely started coming under selection ~900 generations ago, and  
332 compare this to the same distances observed amongst randomly sampled sets of CoV-VIPs  
333 (Figure 5; STAR Methods).

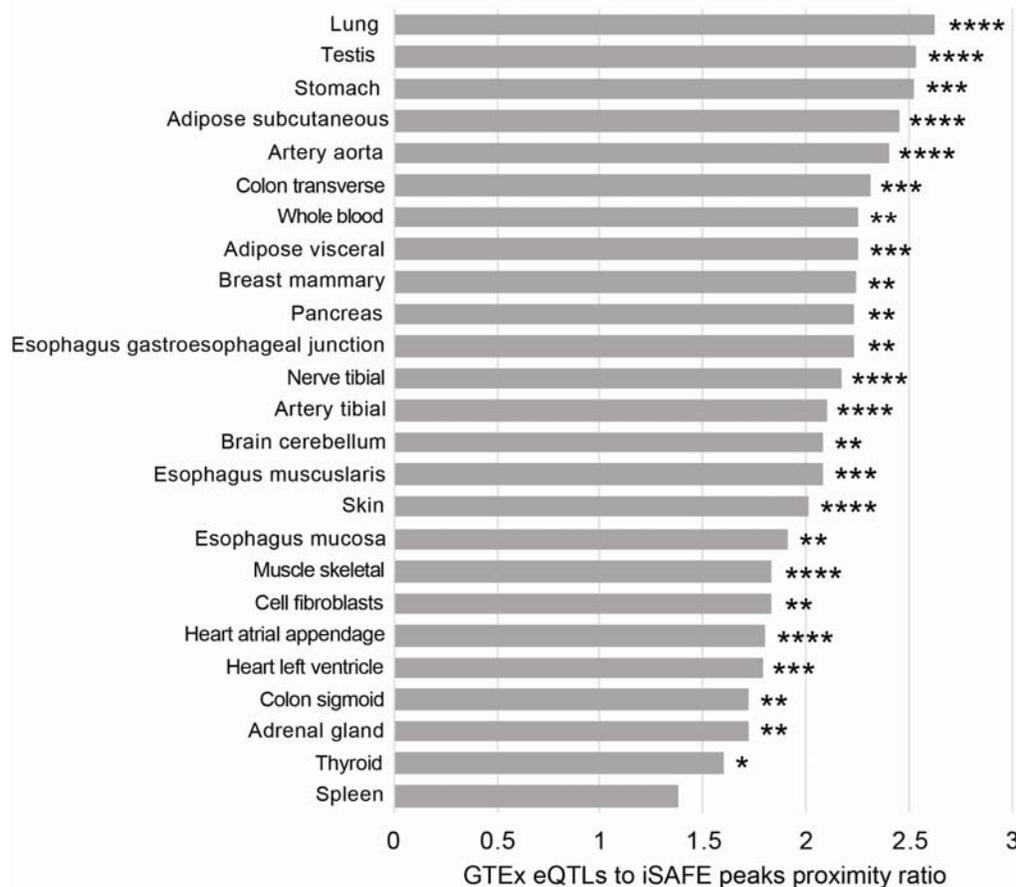
334 Using this approach, we find that GTEx lung eQTLs lie closer to predicted causal mutations  
335 amongst the 42 putative selected loci than for any other tissue ( $P=3.10^{-5}$ ; Figure 5). Several  
336 additional tissues known to be negatively affected by coronavirus – blood and arteries (Bao et  
337 al., 2020; Grosse et al., 2020), adipose tissue (Michalakis and Ilias, 2020) and the digestive  
338 tract (Elmunzer et al., 2020) – also exhibit closer proximities between putative causal loci and  
339 tissue-specific eQTLs than expected by chance (Figure 5). Interestingly, the spleen shows no  
340 tendency for eQTLs to lie closer to selected loci than expected around 900 generations ago  
341 compared to other evolutionary times, perhaps because the spleen is replete with multiple types  
342 of immune cells that might be more prone to more regular adaptation in response to diverse  
343 pathogens over time, and less prone to adaptive bursts restricted over time in response to a  
344 specific pathogen (Quintana-Murci, 2019). Our results indicate that the tissues impacted in the  
345 inferred multi-generational viral epidemic in ancestors of East Asians match those pathologically  
346 affected by the SARS-CoV-2 infection in contemporary populations, providing further evidence  
347 that this ancient infection might have been a coronavirus or related virus.

348

349

350

351



352

353 **Figure 5. Proximity of selection signals to GTEx eQTLs at the 42 selected CoV-VIPs**  
 354 **compared to random CoV-VIPs**

355 The histogram shows how close selection signals localized by iSAFE peaks are to the GTEx  
 356 eQTLs from 25 different tissues, at peak-VIPs compared to randomly chosen CoV-VIPs (STAR  
 357 Methods). How close iSAFE peaks are to GTEx eQTLs compared to random CoV-VIPs is  
 358 estimated through a proximity ratio. The proximity ratio is described in the STAR Methods. It  
 359 quantifies how much closer iSAFE peaks are to eQTLs of a specific GTEx tissue, compared to  
 360 random expectations that take the number and structure of iSAFE peaks, as well as the number  
 361 and structure of GTEx eQTLs into account (STAR Methods). Four stars: proximity ratio test  
 362  $P < 0.0001$ . Three stars: proximity ratio test  $P < 0.001$ . Two stars:  $P < 0.01$ . One star:  $P < 0.05$ . Note  
 363 that lower proximity ratios can be associated with smaller p-values for tissues with more eQTLs  
 364 (due to decreased null variance; for example, skeletal muscle vs. pancreas).

365 **Coronavirus VIPs are enriched for SARS-CoV-2 susceptibility and COVID-19 severity loci**

366 Our results indicate that many of the selected CoV-VIPs now sit at intermediate to high  
 367 frequencies in modern East Asian populations. Accordingly, we anticipate that these  
 368 segregating loci should make a measurable contribution to the inter-individual variation in

369 SARS-CoV-2 susceptibility and (COVID-19) severity amongst contemporary populations in East  
370 Asia, and predict that such loci would be readily detectable in a reasonably-powered genome  
371 wide association study (GWAS) investigating these traits in East Asian populations. While such  
372 a scan has yet to be reported for a large East Asian cohort, two GWASs were recently released  
373 that used sizable British cohorts to investigate SARS-CoV-2 susceptibility (1,454 cases and  
374 7,032 controls; henceforth called the susceptibility GWAS) and severity (325 cases [deaths]  
375 versus 1,129 positive controls; henceforth called the severity GWAS) (data from the UK  
376 Biobank; Sudlow et al., 2015; <https://grasp.nhlbi.nih.gov/Covid19GWASResults.aspx>). While we  
377 are unable to precisely identify the causal variants for the selected CoV-VIP genes observed in  
378 the ancestors of East Asians – nor would these variants necessarily occur as outliers in a  
379 GWAS conducted on the British population – we note that it is possible that other variants in the  
380 same CoV-VIP genes may also produce variation in SARS-CoV-2 susceptibility and severity  
381 amongst modern British individuals.

382 By contrasting variants in CoV-VIPs against those in random sets of genes, we find that variants  
383 in CoV-VIPs have significantly lower p-values for both the susceptibility GWAS and severity  
384 GWAS than expected (simple permutation test  $P < 10^{-9}$  for both GWAS tests; STAR Methods).  
385 More importantly, the 42 CoV-VIPs from the selection event starting ~900 generations ago have  
386 even lower GWAS p-values compared to other CoV-VIPs ( $P = 0.0015$  for susceptibility GWAS  
387 and  $P = 0.023$  for severity GWAS; STAR Methods). This result indicates that the selected genes  
388 inferred in our study might contribute to individual variation in COVID-19 etiology in modern  
389 human populations in the UK, providing further evidence that a coronavirus or related virus may  
390 have been the selection pressure behind the adaptive response we observe in the ancestors of  
391 East Asians.

### 392 **Selected CoV-VIP genes include multiple known drug targets**

393 Our analyses suggest that the 42 CoV-VIPs identified as putative targets of an ancient  
394 coronavirus (or related virus) epidemic might play a functional role in SARS-CoV-2 etiology in  
395 modern human populations. We find that four of these genes (*SMAD3*, *IMPDH2*, *PPIB*, *GPX1*)  
396 are targets of eleven drugs being currently used or investigated in clinical trials to mitigate  
397 COVID-19 symptoms (STAR methods). While this number is not higher than expected when  
398 compared to other CoV-VIPs (hypergeometric test  $P > 0.05$ ), we note that most of the 42 genes  
399 identified here have yet to be the focus of clinical trials for SARS-CoV-2-related drugs. In  
400 addition to the four selected CoV-VIP genes targeted by coronavirus-specific drugs, five

401 additional selected CoV-VIP genes are targeted by multiple drugs to treat a variety of non-  
402 coronavirus pathologies (Table S6). This raises the possibility that such drugs could be  
403 repurposed for therapeutic use in the current SARS-CoV-2 pandemic. Indeed, an additional six  
404 of the 42 selected CoV-VIPs have been identified by (Finan et al., 2017) as part of the  
405 “druggable genome” (Table S6).

## 406 **Discussion**

407 By scanning 26 diverse human populations from five continental regions for evidence of strong  
408 selection acting on proteins that interact with coronavirus strains (CoV-VIPs), we identified a set  
409 of 42 CoV-VIPs exhibiting a coordinated adaptive response that emerged around ~25,000 years  
410 ago (900 generations; Figure 2). This pattern was unique to the ancestors of East Asian  
411 populations, being absent from any of the 21 non East-Asian human populations tested here. By  
412 using ARG methods to reconstruct the trajectories of selected alleles, we show that this  
413 selection pressure produced a strong response across the 42 CoV-VIP genes that gradually  
414 waned and resulted in the selected loci plateauing at intermediate frequencies. Further, we  
415 demonstrate that this adaptive response is likely the outcome of a multigenerational coronavirus  
416 epidemic, as attested by the clustering of putatively selected loci around variants that regulate  
417 tissues known to exhibit COVID-19-related pathologies, and the enrichment of variants  
418 associated with SARS-CoV-2 susceptibility and severity, as well as anti- and proviral functions,  
419 amongst the 42 CoV-VIP genes selected starting 900 generations ago.

420 An important limitation of our study is that some of our analyses rely upon comparative datasets  
421 that were generated in contemporary human populations that have different ancestries than the  
422 East Asian populations where the selected CoV-VIP genes were detected. In particular, both of  
423 the eQTL and GWAS datasets come from large studies that are primarily focused on  
424 contemporary populations from Europe, and none of the five European populations in our study  
425 exhibit the selection signals observed in the genomes of East Asians. Accordingly, more direct  
426 confirmation of the causal role of 42 CoV-VIP genes in COVID-19 etiology will require the  
427 appropriate GWAS to be conducted in East Asian populations. The detection of genetic  
428 associations amongst the 42 CoV-VIPs in a GWAS on contemporary East Asians would provide  
429 further evidence that coronaviruses comprised the selection pressure that drove the observed  
430 adaptive response. Moreover, a high-powered GWAS in East Asian populations could also  
431 identify the putative causal loci that currently impact individual variation in COVID-19 etiology in  
432 East Asian individuals.

### 433 **Insights into ancient viral epidemics from modern human genomes**

434 A particularly salient feature of the adaptive response observed for the 42 CoV-VIPs is that  
435 selection appears to be acting continuously over a 20,000 year period. The activity of a viral  
436 pressure over such an extensive time period is not consistent with epidemics that started in  
437 recorded human history, which tend to be circumscribed to a few generations. This apparent  
438 disparity is reconcilable if coronavirus epidemics had occurred at regular intervals across the  
439 history of East Asian populations. In recent human history, coronavirus epidemics have  
440 occurred on a decadal scale – MERS from 2002 to 2004, SARS in 2007, and SARS-CoV-2 from  
441 2019 until the present – which suggests that a similar temporal distribution of coronavirus  
442 outbreaks might have also occurred across the history of contemporary East Asians. An  
443 alternate hypothesis is that the viral pressure remained present throughout the 20,000 year  
444 period, and initially exerted a strong selective pressure before becoming less severe over time  
445 as a consequence of host adaptation and/or a reduction in virulence. This scenario might be  
446 similar to coevolutionary history of seasonal influenza and the Variola virus in human  
447 populations. As this manuscript was in the final stages of preparation, the first host-virus  
448 interactomes were published for SARS-CoV-1 and MERS-CoV, which exhibit an extensive  
449 overlap with the SARS-CoV-2 interactome used in the present study (Gordon et al., 2020). This  
450 suggests that coronaviruses share a broad set of host proteins that they interact with, which  
451 should also apply to ancient coronaviruses. These patterns are consistent with one or more  
452 coronaviruses driving episodic selection events in East Asian prehistory that produced the  
453 signals that we report here.

454 Further validation of the historical trajectories of the causal mutations at selected genes is still  
455 needed, including more finely resolved temporal and geographic patterns that could be derived  
456 from ancient DNA sampled from across East Asia that span the human occupation of this  
457 region; however, the requisite ancient samples are lacking at the moment. Nonetheless, we  
458 note the geographic origin of several modern outbreaks of coronaviruses in East Asia, and their  
459 exclusion from other parts of the world, point to East Asia being the most likely location where  
460 these ancient populations came into contact with the virus. Given that most recently recorded  
461 coronavirus outbreaks have been traced to zoonoses (direct or indirect with other animal  
462 intermediates) from East Asian bats (Wong et al., 2019), our results suggest that East Asia was  
463 also a natural range for coronavirus reservoir species during the last 25,000 years.

464 **Applied evolutionary medicine: using evolutionary information to combat COVID-19**

465 The net result of the ancient selection patterns on the CoV-VIPs in ancient human populations is  
466 the creation of genetic differences amongst individuals now living in East Asia, and between  
467 East Asians and populations distributed across the rest of the world. As we demonstrate in this  
468 study, this evolutionary genetic information can be exploited by statistical analyses to identify  
469 loci that are potentially involved in the epidemiology of modern diseases – COVID-19 in the  
470 present case. Such evolutionary information may ultimately assist in the development of future  
471 drugs and therapies, by complementing information obtained from more traditional  
472 epidemiological and biomedical research. For example, a recent study focusing on *TMPRSS2* –  
473 a gene encoding for a transmembrane protein that plays a key role in SARS-CoV-2 infection –  
474 found that East Asian populations carry two protein coding variant that are correlated with low  
475 fatality rate for COVID-19 cases (Jeon et al., 2020). While such studies provide high quality  
476 information on a specific gene, the evolutionary approach adopted here is able to leverage  
477 evolutionary information embedded in modern genomes to identify candidate genomic regions  
478 of interest. This is similar to the information provided by GWAS – i.e. lists of variants or genes  
479 that are potentially associated with a particular trait or disease – though we note that the  
480 information provided by evolutionary analyses comes with an added understanding about the  
481 historical processes that created the underlying population genetic patterns.

482 The current limitation shared by population genomic approaches such as GWAS and the  
483 evolutionary analyses presented here, is that they identify statistical associations, rather than  
484 causal links, between genomic regions and traits, thereby necessitating additional research to  
485 confirm causality. In addition to the various forms of empirical information that we provide here,  
486 further evidence of causal relationships between the CoV-VIPs and COVID-19 etiology could be  
487 obtained by examining which viral proteins the selected CoV-VIPs interact with, thereby  
488 establishing the specific viral functions that are affected. As a preliminary observation, we find  
489 that the 35 of the 42 selected SARS-CoV-2 VIPs tend to interact with more viral proteins than  
490 expected by chance (SI). Such information will help establish genetic causality and will also  
491 improve our understanding how hosts adapt in response to viruses.

492 The ultimate confirmation of causality requires functional validation that the genes interact with  
493 the virus, or that drugs targeting these genes have a knock-on impact for the virus. Notably,  
494 several CoV-VIP genes are existing drug targets showing the functional importance of these  
495 particular loci (Table S6), several of which are currently being investigated or used to treat

496 severe cases in the current COVID-19 pandemic. It remains to be established if the other genes  
497 we have identified in this study might also help guide drug repurposing efforts and provide a  
498 basis for future drug and therapeutic development to combat COVID-19 and related  
499 pathologies.

## 500 Conclusion

501 By leveraging the evolutionary information contained in publicly available human genomic  
502 datasets, we were able to infer ancient viral epidemics impacting the ancestors of contemporary  
503 East Asian populations, which initially arose around 25,000 years ago, resulting in coordinated  
504 adaptive changes across at least 42 genes. Importantly, our evolutionary genomic analyses  
505 have identified several new candidate genes that might benefit current efforts to combat COVID-  
506 19, either by providing novel drug targets or by repurposing currently available drugs that target  
507 these candidate genes (Tables S4 & S6). More broadly, our findings highlight the utility of  
508 incorporating evolutionary genomic approaches into standard medical research protocols.  
509 Indeed, by revealing the identity of our ancient pathogenic foes, evolutionary genomic methods  
510 may ultimately improve our ability to predict – and thus prevent – the epidemics of the future.

## 511 Star Methods

512 **Important note: for convenience, the 42 CoV-VIPs that we infer to have started coming**  
513 **under selection around 900 generations ago are called peak-VIPs in the Methods.**

## 514 Key resources table

515

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited Data		
1000 Genome Project - Phase 3	(1000 Genomes Project Consortium, 2015)	<a href="ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/">ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/</a>
VIPs	–this manuscript	–Table S1
Relate-estimated coalescence rates, allele ages and selection P-values for the 1000GP	(Speidel et al., 2019)	<a href="https://zenodo.org/record/3234689">https://zenodo.org/record/3234689</a>

GTEx expression	(GTEx Project, 2017)	<a href="https://gtexportal.org/home/datasets">https://gtexportal.org/home/datasets</a>
Protein-protein interactions (IntAct)	(Luisi et al., 2015)	<a href="https://www.ebi.ac.uk/intact">https://www.ebi.ac.uk/intact</a>
The density of conserved segments (PhastCons)	(Siepel et al., 2005)	<a href="http://hgdownload.cse.ucsc.edu/goldenPath/hg19/phastCons46way/">http://hgdownload.cse.ucsc.edu/goldenPath/hg19/phastCons46way/</a>
The density of regulatory elements	–	<a href="http://hgdownload.soe.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeRegDnaseClustered">http://hgdownload.soe.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeRegDnaseClustered</a>
The recombination rate	(Hinch et al., 2011)	<a href="https://www.well.ox.ac.uk/~anjali/AAMap/">https://www.well.ox.ac.uk/~anjali/AAMap/</a>
Software and Algorithms		
selscan (compute nSL).	(Szpiech and Hernandez, 2014)	<a href="https://github.com/szpiech/selscan">https://github.com/szpiech/selscan</a>
hapbin (compute iHS)	(Maclean et al., 2015)	<a href="https://github.com/evotools/hapbin">https://github.com/evotools/hapbin</a>
Gene Set Enrichment Pipeline	(Enard and Petrov, 2020)	<a href="https://github.com/DavidPierreEnard/Gene_Set_Enrichment_Pipeline">https://github.com/DavidPierreEnard/Gene_Set_Enrichment_Pipeline</a>
Relate	(Speidel et al., 2019)	<a href="https://myersgroup.github.io/relate/">https://myersgroup.github.io/relate/</a>
CLUES	(Stern et al., 2019)	<a href="https://github.com/35ajstern/clues">https://github.com/35ajstern/clues</a>
iSAFE	(Akbari et al., 2018)	<a href="https://github.com/alek0991/iSAFE">https://github.com/alek0991/iSAFE</a>

## 517 **Coronavirus VIPs**

518 We used a dataset of 5,291 VIPs (Table S1). Of these, 1,920 of these VIPs are high confidence  
519 VIPs identified by low-throughput molecular methods, while the remaining VIPs were identified  
520 by diverse high-throughput mass-spectrometry studies. For a more detailed description of the  
521 VIPs dataset, please refer to SI: Host adaptation is expected at VIPs.

## 522 **Genomes and sweeps summary statistics**

523 To detect signatures of adaptation in various human populations, we used the 1,000 Genome  
524 Project phase 3 dataset which provides chromosome level phased data for 26 distinct human  
525 populations representing all major continental groups (1000 Genomes Project Consortium,  
526 2015). To measure nSL separately in each of the 26 populations, we used the selscan software  
527 available at <https://github.com/szpiech/selscan> (Szpiech and Hernandez, 2014). To measure  
528 iHS, we used the hapbin software available at <https://github.com/evotools/hapbin> (Maclean et  
529 al., 2015).

## 530 **Ranking of sweep signals at protein-coding genes and varying window sizes**

531 To detect sweep enrichments at CoV-VIPs, we first order, separately in each of the 26 1,000  
532 Genomes populations, human Ensembl (Cunningham et al., 2019) (version 83) protein-coding  
533 genes according to the intensity of the sweep signals at each gene. As a proxy for the intensity  
534 of these signals, we use the average of either iHS or nSL across all the SNPs with iHS or nSL  
535 values within a window of fixed size, centered at the genomic center of genes, halfway between  
536 the most upstream transcription start site and the most downstream transcription end site. We  
537 then rank the genes according to the average iHS or nSL (more precisely their absolute values)  
538 in these windows. We get six rankings for six different fixed window sizes: 50kb, 100kb, 200kb,  
539 500kb, 1,000kb and 2,000kb. We do this to account for the variable size of sweeps of different  
540 strengths. We then estimate the sweep enrichment at CoV-VIPs compared to controls over all  
541 these different window sizes considered together, or at specific sizes, as described below and in  
542 Enard & Petrov (Enard and Petrov, 2020).

## 543 **Estimating the whole ranking curve enrichment at CoV-VIPs and its statistical** 544 **significance**

545 To estimate a sweep enrichment in a set of genes, a typical approach is to use the outlier  
546 approach to select, for example, the top 1% of genes with the most extreme signals. Here we  
547 use a previously described approach to estimate a sweep enrichment while relaxing the

548 requirement to identify a single top set of genes. Instead of, for example, only estimating an  
549 enrichment in the top 100 genes with the strongest sweep signals, we estimate the enrichment  
550 over a wide range of top X genes, where X is allowed to vary from the top 10,000 to the top 10  
551 with many intermediate values. This creates an enrichment curve as in Figure 1. Figure 1 shows  
552 the estimated relative fold enrichments at CoV-VIPs compared to controls, from the top 1,000 to  
553 the top 10 nSL. The statistical significance of the whole enrichment curve can then be estimated  
554 by using block-randomized genomes, as described in Enard & Petrov (Enard and Petrov, 2020).  
555 In brief, block-randomized genomes make it possible to generate a large number of random  
556 whole enrichment curves while maintaining the same level of clustering of genes in the same  
557 candidate sweeps as in the real genome, which effectively controls for gene clustering.  
558 Comparing the real whole enrichment curve to the random ones then makes it possible to  
559 estimate an unbiased false-positive risk (also known as False Discovery Rate in the context of  
560 multiple testing) for the observed whole enrichment curve at CoV-VIPs. A single false positive  
561 risk can be estimated for not just one curve but by summing over multiple curves combined,  
562 thus making it possible to estimate a single false positive risk over any arbitrary numbers of rank  
563 thresholds, window sizes, summary statistics, and populations. For instance, we estimate the  
564 false-positive enrichment risk of  $P=2.10^{-4}$  at CoV-VIPs for rank threshold from the top 10,000 to  
565 top 10, over six window sizes, for the five East Asian populations in the 1,000 Genomes data,  
566 and for both nSL and iHS, all considered together at once. This makes our approach more  
567 versatile and sensitive to selection signals ranging from a few very strong sweeps, to many,  
568 more moderately polygenic hitchhiking signals. The entire pipeline to estimate false-positive  
569 risks with block-randomized genomes is available at  
570 [https://github.com/DavidPierreEnard/Gene\\_Set\\_Enrichment\\_Pipeline](https://github.com/DavidPierreEnard/Gene_Set_Enrichment_Pipeline) (Enard and Petrov, 2020).

### 571 **Building sets of controls matching for confounding factors**

572 To estimate a sweep enrichment at CoV-VIPs, we compare CoV-VIPs with random control sets  
573 of genes selected far enough (>500kb) from CoV-VIPs that they are unlikely to overlap the  
574 same large sweeps. We do not compare CoV-VIPs with completely random sets of control  
575 genes. Instead, we use a previously described bootstrap test to build random control sets of  
576 genes that match CoV-VIPs for a number of potential confounding factors that might explain a  
577 sweep enrichment, rather than interactions with viruses. The bootstrap test has been described  
578 in detail (Enard and Petrov, 2020), and is available at  
579 [https://github.com/DavidPierreEnard/Gene\\_Set\\_Enrichment\\_Pipeline](https://github.com/DavidPierreEnard/Gene_Set_Enrichment_Pipeline).

580 We include 11 different potential confounding factors in the bootstrap test:

581 - average GTEx expression in 53 GTEx V6 tissues.

582 - GTEx expression in lymphocytes.

583 - GTEx expression in testis.

584 - the number of protein-protein interactions from the Intact database, curated by Luisi et al.

585 (Luisi et al., 2015).

586 - the Ensembl (v83) coding sequence density in a 50kb window centered on each gene.

587 - the density of conserved segments identified by PhastCons (Siepel et al., 2005)

588 (<http://hgdownload.cse.ucsc.edu/goldenPath/hg19/phastCons46way/>).

589 - the density of regulatory elements, estimated by the density of Encode DNase I V3 Clusters

590 (<http://hgdownload.soe.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeRegDnaseClustered/>)

591 in a 50kb window centered on each gene.

592 - the recombination rate in a 200kb window centered on each gene (Hinch et al., 2011).

593 - the GC content in a 50kb window centered on each gene.

594 - the number of bacteria each gene interacts with, according to the Intact database (as of June

595 2019; <https://www.ebi.ac.uk/intact/>).

596 - the proportion of genes that are immune genes according to Gene Ontology annotations

597 GO:0006952 (defense response), GO:0006955 (immune response), and GO:0002376 (immune

598 system process) as of May 2020.

### 599 **Estimating adaptation start times at specific genes with Relate**

600 As times of emergence of adaptive mutations, we use the publicly available estimates from

601 Relate (<https://myersgroup.github.io/relate/>). Relate estimates mutation emergence times while

602 controlling for fluctuations of population size over time, based on the coalescence rates it

603 reconstructs after inferring ancestral recombination graphs at the scale of the whole genome

604 (Speidel et al., 2019). Relate provides two times of emergence of mutations, one low estimate

605 (less generations ago), and one high estimate (more generations ago). The low time estimate

606 corresponds to the time when Relate estimates an elevated probability that the frequency of the

607 mutation is different from zero. The high time estimate corresponds to the time when Relate

608 estimates that the probability is not too small that the frequency of the mutation is different from

609 zero. For our purpose of estimating when selection started, the low time estimate is the best

610 suited, because it provides an estimate of when the frequency of a selected mutation was

611 already high enough to distinguish from zero, for those mutations where selection started from a

612 very low frequency. For cases where selection started with standing genetic variants that were  
613 already distinguishable from zero, the Relate low time estimates for the emergence of mutations  
614 do not provide a good proxy for when selection actually started. Thus, if we were able to  
615 estimate when selection started for standing genetic variants, we might be able to observe an  
616 even stronger peak than the one we see when just relying on those variants where selection  
617 started from low frequencies.

618 Using the low Relate time estimates is also justified due to the fact that the sweep establishment  
619 phase can take very variable amounts of time before the start of the sweep exponential phase.  
620 During the establishment phase, selected alleles are still mostly governed by drift which makes  
621 pinpointing the actual starting time of selection difficult. In this context, the low Relate time  
622 estimates provide an estimate of the time when the selected alleles were no longer at very low  
623 frequencies not statistically different from zero, and closer to entering the exponential phase,  
624 which provides a more certain time estimate for when selection started for certain.

625 An important step is then to choose at each CoV-VIP locus, and all the other control loci, which  
626 Relate mutation to use to get a single time estimate for each locus. Note that here we make an  
627 assumption that each locus has experienced only one single adaptive event. Given our finding  
628 that iSAFE peaks at CoV-VIPs are much closer to GTEx V8 eQTLs than expected by chance, it  
629 is likely that the selected adaptive mutations are regulatory mutations at, or close to annotated  
630 eQTLs for a specific gene. They are not necessarily exactly located at eQTLs, because current  
631 eQTLs annotations may still be incomplete, and in our case we use eQTLs identified in GTEx  
632 V8 using mostly European individuals, even though we analyse selection signals in East Asian  
633 populations. Because of these limitations, we use the Relate estimated time at the mutation  
634 where Relate estimates the lowest positive selection p-value within 50kb windows centered on  
635 eQTLs. We also only consider variants with a minor allele frequency greater than 20%, given  
636 the signals detected by iHS and nSL that only have some power to detect incomplete sweeps  
637 above 20% frequencies (Ferrer-Admetlla et al., 2014; Voight et al., 2006). This also excludes a  
638 potential risk of confounding by low frequency neutral or weakly deleterious variants, that can  
639 show selection-like patterns when their only way to escape removal early on is through a  
640 chance, rapid frequency increase that can look like selection. The Relate selection test is based  
641 on faster than expected coalescence rates given the population size at any given time, and its  
642 results are publicly available at <https://myersgroup.github.io/relate/>. Note that the mutation with  
643 the lowest Relate p-value does not always overlap with an iSAFE peak (Figure S5), which is not  
644 entirely surprising if the haplotype signals exploited by both Relate and iSAFE partly

645 deteriorated due to recombination since the time selection at CoV-VIPs was strong (Figures 3  
646 and 4). Both of these methods are indeed designed to locate the selected variant right after, or  
647 during, active selection.

648 Because we work with five different East Asian populations, we more specifically select the  
649 variant with the lowest Relate selection test p-value on average across all the five East Asian  
650 populations. Then, we also use the corresponding average low Relate mutation time estimate  
651 across the five East Asian populations. We do not attempt to estimate the selection time and p-  
652 value by considering all 1,000 Genomes East Asian individuals tested together by Relate,  
653 because then the Relate selection test is at a greater risk of being confounded by population  
654 structure. Finally, we only consider CoV-VIPs and other control genes with an average Relate  
655 selection test p-value lower than  $10^{-3}$ , to make sure that we indeed use estimated times at  
656 selected variants.

### 657 **The peak significance test**

658 To test if the peak of Relate time estimates around 900 generations ago at CoV-VIPs (Figure 2)  
659 is expected simply by chance or not, we designed a peak significance test. The test compares  
660 the peak at CoV-VIPs, with the top peaks obtained when repeatedly randomly sampling sets of  
661 genes. We first identify the most prominent peak at CoV-VIPs by visual inspection of the pink  
662 distribution of Relate times for CoV-VIPs compared to the blue distribution of Relate times for all  
663 protein-coding genes with an estimated Relate time (Figure 2). To build these distributions, top  
664 Relate selected mutations shared between multiple neighboring genes (CoV-VIPs or controls)  
665 are counted only once, to avoid a confounding effect of gene clustering (152 selected variants at  
666 CoV-VIPs, 1771 selected variants for all protein coding genes). The peak around 900  
667 generations ago (870 generations more exactly) spans approximately 200 generations, where  
668 the pink distribution is clearly above the blue one. We then use a 200 generations-wide window,  
669 sliding every generation from 0 to 6,000 generations to verify the peak more rigorously. Sliding  
670 one generation after another, each time we count the difference between the number of Relate  
671 selected variants at CoV-VIPs that fall in the sliding 200 generations window, and the number of  
672 Relate selected variants at all other genes that are not CoV-VIPs, weighted by the percentage  
673 of variants found at CoV-VIPs, to correct for the different size of the two sets of variants. Using  
674 this sliding window approach, the top of the peak is found at 870 generations, with a difference  
675 of 19.5 additional Relate selected variants between 770 and 970 (870 plus or minus 100) at  
676 CoV-VIPs compared to the null expectation.

677 We then repeat the sliding of a 200 generations window to identify the maximum peak and  
678 measure the same difference, but this time for random sets of Relate selected variants of the  
679 same size (152 selected variants out of the 1,771 selected variants). To estimate p-values, we  
680 then compare the actual observed difference with the distribution of differences generated with  
681 one million random samples.

682 As mentioned in the Results, one potential issue is that we run the peak significance test after  
683 we already know that CoV-VIPs are enriched for iHS and nSL top sweeps, and especially  
684 enriched for nSL top sweeps. This enrichment may skew the null expectation for the distribution  
685 of Relate times at CoV-VIPs. In other words, there is a risk that any set of genes with the same  
686 sweep enrichment might exhibit the same peak as CoV-VIP. As a result, comparing CoV-VIPs  
687 with randomly chosen non-CoV-VIPs may not be appropriate. To test this, we repeat the peak  
688 significance test, but this time comparing the peak at CoV-VIPs with the peaks at random sets  
689 of non-CoV-VIPs that we build to have the same distribution of nSL ranks as CoV-VIPs. To do  
690 this, we define nSL bins between ranks 1 and the highest rank with a rank step of 100 between  
691 each bin, and we count how many Relate selected variants fall in each bin (each gene has one  
692 nSL rank and one Relate selected variant). To build the random set, we then fill each of the 100  
693 bins with the same number of random non-CoV-VIPs, as long as their nSL rank falls within that  
694 bin. We use the average nSL rank over the five East Asian populations, and the lower  
695 population-averaged rank of either 1 Mb or 2Mb window sizes (where we observe the strongest  
696 enrichment at CoV-VIPs, see Results). The results of the peak significance test are unchanged  
697 when using the matching nSL distribution (peak significance test  $P=1.10^{-4}$  vs.  $P=2.3 \cdot 10^{-4}$   
698 without matching nSL distribution).

699 In further agreement with the fact that the sweep enrichment does not confound the peak  
700 significance test, the peak at CoV-VIPs stands out more when repeating the peak significance  
701 test using a smaller nSL top rank limit (Figure S6). In this case, we compare sets of CoV-VIPs  
702 and sets of controls both enriched in stronger sweep signals. Thus, if stronger sweep signals at  
703 CoV-VIPs biased the peak significance test, we would expect the peak to fade away when  
704 comparing only CoV-VIPs and controls both with stronger nSL signals. Conversely, we observe  
705 that half of the CoV-VIPs with the weaker nSL signals (population-averaged nSL rank higher  
706 than 7,200 for both 1Mb and 2Mb windows) do not show a significant peak (peak significance  
707 test  $P=0.53$ ).

## 708 **The iSAFE peaks/eQTL proximity test**

709 Adaptation in the human genome was likely mostly regulatory adaptation through gene  
710 expression changes (Enard et al., 2014; Kudaravalli et al., 2009; Nédélec et al., 2016; Quach et  
711 al., 2016). To test if positive selection at CoV-VIPs likely involved regulatory changes, we ask  
712 whether the signals of adaptation around CoV-VIPs are localized closer than expected by  
713 chance to GTEx eQTLs that affect the expression of CoV-VIPs in present human populations.  
714 Indeed, the genomic regions at or close to CoV-VIP GTEx eQTLs are likely enriched for CoV-  
715 VIP regulatory elements, and therefore the most likely place to find CoV-VIP-related adaptations  
716 in the genome. To localize where adaptation occurred, we use the iSAFE method that was  
717 specifically designed for this purpose (Akbari et al., 2018). iSAFE scans the genome and  
718 estimates a score that increases together with proximity to the actual selected mutation. The  
719 higher the score, the higher the odds that the scored variant is itself the selected one, or close  
720 to the selected one. An important caveat is that iSAFE is designed to localize where selection  
721 happened right after it happened, or as selection is still ongoing. In our case, we have evidence  
722 that selection was strong at CoV-VIPs only more than 500 generations (~14,000 years) ago,  
723 and then much weaker more recently (Figure 4). This could be an issue, because we expect  
724 that recombination events that occurred after the strong selection might have deteriorated the  
725 iSAFE signal that relies on haplotype structure. This is because recombination mixes together  
726 the haplotypes that hitchhiked with the selected mutation, with those that did not. In line with  
727 this, we often do not observe simple, clean iSAFE score peaks, but instead, iSAFE score  
728 plateaus and more rugged peaks (Figure S5). For this reason, we designed an approach to not  
729 only identify the top of simple iSAFE peaks, but also more rugged peaks or plateaus. First, to  
730 measure iSAFE scores, we combine all the haplotypes from the five East Asian populations  
731 together as input, since we found that the selection signal at CoV-VIPs is common to all these  
732 populations (iSAFE parameters: --IgnoreGaps --MaxRegionSize 250000000 --window 300 --  
733 step 100 --MaxFreq .95 --MaxRank 15). We then use a 500kb window sliding every 10kb to  
734 identify the highest local iSAFE value in the 500kb window (Figure S8). Once we have the  
735 highest local iSAFE value and coordinate, we define a broader iSAFE peak as the region both  
736 upstream and downstream where the iSAFE values are still within 80% of the maximum value  
737 (Figure S8). This way, we can better annotate iSAFE plateaus and rugged peaks, and take into  
738 account the fact that they can span more than just a narrow local maximum (Figure S5).

739 Once the local iSAFE peaks are identified, we can ask how close GTEx eQTLs are to these  
740 peaks compared to random expectations. We first measure the distance of each CoV-VIP GTEx

741 eQTL to the closest iSAFE peak. To avoid redundancy, we merge eQTLs closer than 1kb to  
742 each other into one test eQTL at the closest, lower multiple of 1,000 genomic coordinates (for  
743 example 3,230 and 3,950 would both become 3,000). We then measure the average of the log  
744 of the distance between all CoV-VIPs and their closest iSAFE peak. We use the log (base 10) of  
745 the distance, because it matters if the eQTL/iSAFE peak distance is 100 bases instead of  
746 200kb, but it does not really matter if the distance is 200kb or 600kb, because the iSAFE peak  
747 at 300kb is likely not related to the eQTL more than the peak at 600kb. Once we have the  
748 average of log-distances, we compare it to its random expected distribution. To get this random  
749 distribution, we measure the log-distance between each CoV-VIP eQTL and the iSAFE peaks,  
750 but after shifting the iSAFE scores left or right by a random value between 1Mb and 2Mb (Figure  
751 S8; less, or no shift at all if this falls within telomeres or centromeres). We shift by at least 1Mb  
752 to make sure that we do not rebuild the original overlap of iSAFE peaks with eQTLs again and  
753 again (some iSAFE peaks, or more precisely rugged peaks and plateaus can be wide and  
754 include several hundred kilobases; see Figure S5). The random shifting effectively breaks the  
755 relationship between eQTLs and iSAFE peaks, while maintaining the same overall eQTL and  
756 peak structure (and thus variance for the test). The random log-distance distribution then  
757 provides an overall random average log-distance to compare the observed average long-  
758 distance with, as well as estimate a p-value.

759 Then, to more specifically ask if lung eQTLs at CoV-VIPs or the eQTLs of other specific tissues  
760 are closer to iSAFE peaks than expected by chance, we can do the same but only using the  
761 eQTLs of that specific tissue. The analysis represented in figure 5 is however more complicated  
762 than just testing if CoV-VIP eQTLs for a specific tissue are closer to iSAFE peaks than expected  
763 by chance by randomly sliding iSAFE values. Instead, what we ask is whether the 42 peak-VIPs  
764 have eQTLs for a given tissue that are even closer to iSAFE peaks than the eQTLs of all CoV-  
765 VIPs in general. To test this, for example with lung eQTLs, we first estimate how close lung  
766 eQTLs are to iSAFE peaks at peak-VIPs, compared to random expectations, by measuring the  
767 difference between the observed and the average random log-distance, just as described  
768 before. We then count the number of peak-VIPs with lung eQTLs (19 out of 25 peak-VIPs with  
769 GTEx eQTLs), and we randomly select the same number of any CoV-VIP (which may randomly  
770 include peak-VIPs) as long as the random set of CoV-VIPs has the same number of lung eQTLs  
771 (plus or minus 10%) as the set of peak VIPs with lung eQTLs (the same gene can have multiple  
772 eQTLs for one tissue). We make sure that the tested and the random sets have similar numbers  
773 of genes and eQTLs so that the test has the appropriate null variance. We then measure the

774 difference between the observed log-distance, and the randomly expected average log-distance  
775 for the random set of CoV-VIPs, exactly the same way we did before for the actual set of peak-  
776 VIPs. We then measure the ratio of the observed difference in log-distance between peak-VIPs  
777 and the random expectation after many random shiftings (1,000), divided by the average of the  
778 same difference measured over many random sets of CoV-VIPs. The final ratio tells us how  
779 much closer lung eQTLs are to iSAFE peaks at peak-VIPs compared to CoV-VIPs in general,  
780 and still takes the specific eQTLs and iSAFE peak structures at each locus into account, since  
781 we compare differences in log-distances expected while preserving the same eQTL and iSAFE  
782 peak structure (see above the description of the random coordinate shifting). One important last  
783 detail about the test is that because we already found that the 50% of loci with the lowest nSL  
784 signals do not show a peak of selection at CoV-VIPs around 900 generations ago (see Results),  
785 we do not use these loci in this test since any iSAFE peak there is much more likely to represent  
786 random noise, not actual selection locations, and thus likely to dilute genuine signals. Using this  
787 test, we find that lung and other tissues' eQTLs at peak-VIPs are much closer to iSAFE peaks  
788 than they are at CoV-VIPs in general. This test thus specifically tells that adaptation happened  
789 closer to lung eQTLs, specifically around 900 generations ago compared to other evolutionary  
790 times. By estimating the same ratio for 24 other tissues with at least 10 peak-VIPs with the  
791 specific tested tissue eQTLs, we can finally rank each tissue for its more pronounced  
792 involvement in adaptation ~900 generations ago, as done in figure 5. It is particularly interesting  
793 in this respect that the tissue with least evidence for being more involved in adaptation at that  
794 time more than other evolutionary times is spleen. Spleen indeed likely represents a good  
795 negative control as a tissue strongly enriched in immune cell types and likely to have evolved  
796 adaptively for most of evolution.

### 797 **UK Biobank GWAS analysis**

798 To compare the UK Biobank GWAS p-values at different loci, we assigned one p-value for each  
799 gene, either CoV-VIPs, peak-VIPs or other genes, even though each gene locus can have many  
800 variants with associated GWAS p-values. To assign just one single GWAS p-value to each  
801 gene, we selected the variant with the lowest p-value at or very close (<1kb) to GTEx eQTLs for  
802 a specific gene, in line with the fact that GWAS hits tend to overlap eQTLs (Hormozdiari et al.,  
803 2016), and to remain consistent with the rest of our manuscript. We then compared the average  
804 p-value between different sets of genes using classic permutations (one billion iterations).

## 805 **Drug targets identification**

806 We queried the databases DGldb (Cotto et al., 2017), and PanDrugs (Piñeiro-Yáñez et al.,  
807 2018) for drugs targeting CoV-VIPs and peak-VIPs. For hits from PanDrugs we limited the  
808 results to only genes that are in direct interaction with the designated drug. Drugs targeting  
809 peak-VIPs are presented in Table S6. In addition, we present a list of peak-VIPs that are not  
810 currently drug targets, but have been previously identified in (Finan et al., 2017) as viable drug  
811 targets (druggable genome).

## 812 **Acknowledgments**

813 We wish to thank Leo Speidel and Aaron Stern for their valuable help using Relate and CLUES,  
814 respectively. Y.S. is supported by the Australian Research Council (ARC DP190103705). R.T. is  
815 an ARC DECRA fellow (DE190101069).

816

## 817 **Authors Contributions**

818 Conceived and designed the experiments: YS, RT, DE. Performed the experiments: YS, MEL,  
819 RT, DE. Interpreted the results: YS, MEL, RT,CDH, ASJ, DE. Wrote the manuscript: YS, RT, DE

820

## 821 **References**

822 1000 Genomes Project Consortium (2015). A global reference for human genetic variation.  
823 *Nature* 526, 68–74.

824 Akbari, A., Vitti, J.J., Iranmehr, A., Bakhtiari, M., Sabeti, P.C., Mirarab, S., and Bafna, V. (2018).  
825 Identifying the favored mutation in a positive selective sweep. *Nat. Methods* 15, 279–282.

826 Balogun, O.D., Bea, V.J., and Phillips, E. (2020). Disparities in Cancer Outcomes Due to  
827 COVID-19—A Tale of 2 Cities. *JAMA Oncol.*

828 Bao, C., Tao, X., Cui, W., Yi, B., Pan, T., Young, K.H., and Qian, W. (2020). SARS-CoV-2  
829 induced thrombocytopenia as an important biomarker significantly correlated with abnormal  
830 coagulation function, increased intravascular blood clot risk and mortality in COVID-19 patients.  
831 *Exp. Hematol. Oncol.* 9, 16.

832 Barreiro, L.B., Ben-Ali, M., Quach, H., Laval, G., Patin, E., Pickrell, J.K., Bouchier, C., Tichit, M.,  
833 Neyrolles, O., Gicquel, B., et al. (2009). Evolutionary dynamics of human Toll-like receptors and  
834 their different contributions to host defense. *PLoS Genet.* 5, e1000562.

835 Best, S.M., and Kerr, P.J. (2000). Coevolution of host and virus: the pathogenesis of virulent  
836 and attenuated strains of myxoma virus in resistant and susceptible European rabbits. *Virology*

- 837 267, 36–48.
- 838 Cotto, K.C., Wagner, A.H., Feng, Y.-Y., Kiwala, S., Coffman, A.C., Spies, G., Wollam, A., Spies,  
839 N.C., Griffith, O.L., and Griffith, M. (2017). DGIdb 3.0: a redesign and expansion of the drug–  
840 gene interaction database. *Nucleic Acids Res.* *46*, D1068–D1073.
- 841 Cunningham, F., Achuthan, P., Akanni, W., Allen, J., Amode, M.R., Armean, I.M., Bennett, R.,  
842 Bhai, J., Billis, K., Boddu, S., et al. (2019). Ensembl 2019. *Nucleic Acids Res.* *47*, D745–D751.
- 843 Dong, E., Du, H., and Gardner, L. (2020). An interactive web-based dashboard to track COVID-  
844 19 in real time. *Lancet Infect. Dis.* *20*, 533–534.
- 845 eGTEx Project (2017). Enhancing GTEx by bridging the gaps between genotype, gene  
846 expression, and disease. *Nat. Genet.* *49*, 1664–1670.
- 847 Ellinghaus, D., Degenhardt, F., Bujanda, L., Buti, M., Albillos, A., Invernizzi, P., Fernández, J.,  
848 Prati, D., Baselli, G., Asselta, R., et al. (2020). Genomewide Association Study of Severe Covid-  
849 19 with Respiratory Failure. *N. Engl. J. Med.*
- 850 Elmunzer, B.J., Spitzer, R.L., Foster, L.D., Merchant, A.A., Howard, E.F., Patel, V.A., West,  
851 M.K., Qayed, E., Nustas, R., Zakaria, A., et al. (2020). Digestive Manifestations in Patients  
852 Hospitalized with COVID-19. *Clin. Gastroenterol. Hepatol.*
- 853 Enard, D., and Petrov, D.A. (2018). Evidence that RNA Viruses Drove Adaptive Introgression  
854 between Neanderthals and Modern Humans. *Cell* *175*, 360–371.e13.
- 855 Enard, D., and Petrov, D.A. (2020). Ancient RNA virus epidemics through the lens of recent  
856 adaptation in human genomes. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* *375*, 20190575.
- 857 Enard, D., Messer, P.W., and Petrov, D.A. (2014). Genome-wide signals of positive selection in  
858 human evolution. *Genome Res.* *24*, 885–895.
- 859 Enard, D., Cai, L., Gwennap, C., and Petrov, D.A. (2016). Viruses are a dominant driver of  
860 protein adaptation in mammals. *Elife* *5*, 56.
- 861 Ferrer-Admetlla, A., Liang, M., Korneliussen, T., and Nielsen, R. (2014). On detecting  
862 incomplete soft or hard selective sweeps using haplotype structure. *Mol. Biol. Evol.* *31*, 1275–  
863 1291.
- 864 Finan, C., Gaulton, A., Kruger, F.A., Lumbers, R.T., Shah, T., Engmann, J., Galver, L., Kelley,  
865 R., Karlsson, A., Santos, R., et al. (2017). The druggable genome and support for target  
866 identification and validation in drug development. *Sci. Transl. Med.* *9*.
- 867 Gene Ontology Consortium (2015). Gene Ontology Consortium: going forward. *Nucleic Acids*  
868 *Res.* *43*, D1049–D1056.
- 869 Gordon, D.E., Jang, G.M., Bouhaddou, M., Xu, J., Obernier, K., White, K.M., O’Meara, M.J.,  
870 Rezelj, V.V., Guo, J.Z., Swaney, D.L., et al. (2020). A SARS-CoV-2 protein interaction map  
871 reveals targets for drug repurposing. *Nature* *583*, 459–468.
- 872 Grosse, C., Grosse, A., Salzer, H.J.F., Dünser, M.W., Motz, R., and Langer, R. (2020). Analysis  
873 of cardiopulmonary findings in COVID-19 fatalities: High incidence of pulmonary artery thrombi  
874 and acute suppurative bronchopneumonia. *Cardiovasc. Pathol.* *49*, 107263.

- 875 Hayward, L.K., and Sella, G. (2019). Polygenic adaptation after a sudden change in  
876 environment.
- 877 Hinch, A.G., Tandon, A., Patterson, N., Song, Y., Rohland, N., Palmer, C.D., Chen, G.K., Wang,  
878 K., Buxbaum, S.G., Akylbekova, E.L., et al. (2011). The landscape of recombination in African  
879 Americans. *Nature* 476, 170–175.
- 880 Hoffmann, C., and Kamps, B.S. (2003). SARS Reference (Flying Publisher).
- 881 Hormozdiari, F., van de Bunt, M., Segrè, A.V., Li, X., Joo, J.W.J., Bilow, M., Sul, J.H.,  
882 Sankararaman, S., Pasaniuc, B., and Eskin, E. (2016). Colocalization of GWAS and eQTL  
883 Signals Detects Target Genes. *Am. J. Hum. Genet.* 99, 1245–1260.
- 884 Jeon, S., Blazyte, A., Yoon, C., Ryu, H., Jeon, Y., Bhak, Y., Bolser, D., Manica, A., Shin, E.-S.,  
885 Cho, Y.S., et al. (2020). Ethnicity-dependent allele frequencies are correlated with COVID-19  
886 case fatality rate (Authorea, Inc.).
- 887 Kudaravalli, S., Veyrieras, J.-B., Stranger, B.E., Dermitzakis, E.T., and Pritchard, J.K. (2009).  
888 Gene expression levels are a target of recent natural selection in the human genome. *Mol. Biol.*  
889 *Evol.* 26, 649–658.
- 890 Luisi, P., Alvarez-Ponce, D., Pybus, M., Fares, M.A., Bertranpetit, J., and Laayouni, H. (2015).  
891 Recent positive selection has acted on genes encoding proteins with more interactions within  
892 the whole human interactome. *Genome Biol. Evol.* 7, 1141–1154.
- 893 Maclean, C.A., Chue Hong, N.P., and Prendergast, J.G.D. (2015). hapbin: An Efficient Program  
894 for Performing Haplotype-Based Scans for Positive Selection in Large Genomic Datasets. *Mol.*  
895 *Biol. Evol.* 32, 3027–3029.
- 896 Michalakis, K., and Ilias, I. (2020). SARS-CoV-2 infection and obesity: Common inflammatory  
897 and metabolic aspects. *Diabetes Metab. Syndr.* 14, 469–471.
- 898 Moorjani, P., Sankararaman, S., Fu, Q., Przeworski, M., Patterson, N., and Reich, D. (2016). A  
899 genetic method for dating ancient genomes provides a direct estimate of human generation  
900 interval in the last 45,000 years. *Proc. Natl. Acad. Sci. U. S. A.* 113, 5652–5657.
- 901 Nédélec, Y., Sanz, J., Baharian, G., Szpiech, Z.A., Pacis, A., Dumaine, A., Grenier, J.-C.,  
902 Freiman, A., Sams, A.J., Hebert, S., et al. (2016). Genetic Ancestry and Natural Selection Drive  
903 Population Differences in Immune Responses to Pathogens. *Cell* 167, 657–669.e21.
- 904 Ou, X., Liu, Y., Lei, X., Li, P., Mi, D., Ren, L., Guo, L., Guo, R., Chen, T., Hu, J., et al. (2020).  
905 Characterization of spike glycoprotein of SARS-CoV-2 on virus entry and its immune cross-  
906 reactivity with SARS-CoV. *Nat. Commun.* 11, 1620.
- 907 Piñeiro-Yáñez, E., Reboiro-Jato, M., Gómez-López, G., Perales-Patón, J., Troulé, K.,  
908 Rodríguez, J.M., Tejero, H., Shimamura, T., López-Casas, P.P., Carretero, J., et al. (2018).  
909 PanDrugs: a novel method to prioritize anticancer drug treatments according to individual  
910 genomic data. *Genome Med.* 10, 41.
- 911 Quach, H., Rotival, M., Pothlichet, J., Loh, Y.-H.E., Dannemann, M., Zidane, N., Laval, G.,  
912 Patin, E., Harmant, C., Lopez, M., et al. (2016). Genetic Adaptation and Neandertal Admixture  
913 Shaped the Immune System of Human Populations. *Cell* 167, 643–656.e17.

- 914 Quintana-Murci, L. (2019). Human Immunology through the Lens of Evolutionary Genetics. *Cell*  
915 177, 184–199.
- 916 Richman, D.D., Whitley, R.J., and Hayden, F.G. (2020). *Clinical Virology* (John Wiley & Sons).
- 917 Roberts, G.H.L., Park, D.S., Coignet, M.V., McCurdy, S.R., Knight, S.C., Partha, R., Rhead, B.,  
918 Zhang, M., Berkowitz, N., Baltzell, A.K.H., et al. (2020). AncestryDNA COVID-19 Host Genetic  
919 Study Identifies Three Novel Loci. medRxiv.
- 920 Sabeti, P.C., Schaffner, S.F., Fry, B., Lohmueller, J., Varilly, P., Shamovsky, O., Palma, A.,  
921 Mikkelsen, T.S., Altshuler, D., and Lander, E.S. (2006). Positive natural selection in the human  
922 lineage. *Science* 312, 1614–1620.
- 923 Sattar Naveed, McInnes Iain B., and McMurray John J.V. (2020). Obesity Is a Risk Factor for  
924 Severe COVID-19 Infection. *Circulation* 142, 4–6.
- 925 Sawyer, S.L., Wu, L.I., Emerman, M., and Malik, H.S. (2005). Positive selection of primate  
926 TRIM5 $\alpha$  identifies a critical species-specific retroviral restriction domain. *Proc. Natl. Acad. Sci.*  
927 *U. S. A.* 102, 2832–2837.
- 928 Scarpone, C., Brinkmann, S.T., Große, T., Sonnenwald, D., Fuchs, M., and Walker, B.B. (2020).  
929 A multimethod approach for county-scale geospatial analysis of emerging infectious diseases: a  
930 cross-sectional case study of COVID-19 incidence in Germany. *Int. J. Health Geogr.* 19, 32.
- 931 Schrider, D.R. (2020). Background Selection Does Not Mimic the Patterns of Genetic Diversity  
932 Produced by Selective Sweeps. *Genetics* 216, 499–519.
- 933 Siepel, A., Bejerano, G., Pedersen, J.S., Hinrichs, A.S., Hou, M., Rosenbloom, K., Clawson, H.,  
934 Spieth, J., Hillier, L.W., Richards, S., et al. (2005). Evolutionarily conserved elements in  
935 vertebrate, insect, worm, and yeast genomes. *Genome Res.* 15, 1034–1050.
- 936 Speidel, L., Forest, M., Shi, S., and Myers, S.R. (2019). A method for genome-wide genealogy  
937 estimation for thousands of samples. *Nat. Genet.* 51, 1321–1329.
- 938 Stern, A.J., Wilton, P.R., and Nielsen, R. (2019). An approximate full-likelihood method for  
939 inferring selection and allele frequency trajectories from DNA sequence data. *PLoS Genet.* 15,  
940 e1008384.
- 941 Stern, A.J., Speidel, L., Zaitlen, N.A., and Nielsen, R. (2020). Disentangling selection on  
942 genetically correlated polygenic traits using whole-genome genealogies. bioRxiv.
- 943 Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P.,  
944 Green, J., Landray, M., et al. (2015). UK biobank: an open access resource for identifying the  
945 causes of a wide range of complex diseases of middle and old age. *PLoS Med.* 12, e1001779.
- 946 Szpiech, Z.A., and Hernandez, R.D. (2014). selscan: an efficient multithreaded program to  
947 perform EHH-based scans for positive selection. *Mol. Biol. Evol.* 31, 2824–2827.
- 948 Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA  
949 polymorphism. *Genetics* 123, 585–595.
- 950 The COVID-19 Host Genetics Initiative (2020). The COVID-19 Host Genetics Initiative, a global  
951 initiative to elucidate the role of host genetic factors in susceptibility and severity of the SARS-

- 952 CoV-2 virus pandemic. *Eur. J. Hum. Genet.* 28, 715–718.
- 953 Uricchio, L.H., Petrov, D.A., and Enard, D. (2019). Exploiting selection at linked sites to infer the  
954 rate and strength of adaptation. *Nat Ecol Evol* 3, 977–984.
- 955 Voight, B.F., Kudaravalli, S., Wen, X., and Pritchard, J.K. (2006). A map of recent positive  
956 selection in the human genome. *PLoS Biol.* 4, e72.
- 957 Wong, A.C.P., Li, X., Lau, S.K.P., and Woo, P.C.Y. (2019). Global Epidemiology of Bat  
958 Coronaviruses. *Viruses* 11.
- 959 World Health Organization (2019). Middle East respiratory syndrome coronavirus (MERS-CoV).
- 960 Zeberg, H., and Pääbo, S. (2020). The major genetic risk factor for severe COVID-19 is  
961 inherited from Neandertals.
- 962