Genomic epidemiology reveals multiple introductions of SARS-CoV-2 from mainland Europe into Scotland

Ana da Silva Filipe[1,#], James G Shepherd[1,#] , Thomas Williams[2,#], Joseph Hughes[1,#], Elihu Aranday-Cortes[1], Patawee Asamaphan[1], Shirin Ashraf[1], Carlos Balcazar[3], Kirstyn Brunker[1], Alasdair Campbell[4], Stephen Carmichael[1], Chris Davis[1], Rebecca Dewar[5], Michael D. Gallagher[6], Rory Gunson[7,8], Verity Hill [9], Antonia Ho[1], Ben Jackson[9], Edward James[10], Natasha Jesudason[1], Natasha Johnson[1], E Carol McWilliam Leitch[1], Kathy Li[1], Alasdair MacLean[7], Daniel Mair[1], David A. McAllister[11,12], John T McCrone[9], Sarah E. McDonald[1], Martin P. McHugh[5,13], A. Keith Morris[14], Jenna Nichols[1], Marc Niebel[1], Kyriaki Nomikou[1], Richard J Orton[1], Áine O'Toole[9], Massimo Palmarini[1], Benjamin J Parcell[15], Yasmin A. Parr [1], Andrew Rambaut[9], Stefan Rooke[8], Sharif Shaaban[11,] Rajiv Shah[1] , Joshua B. Singer[1], Katherine Smollett[1], Igor Starinskij[7], Lily Tong[1], Vattipally B. Sreenu[1], Elizabeth Wastnedge[5], The COVID-19 Genomics UK (COG-UK) consortium[16] , Matthew T.G. Holden[11,13,#], , David L Robertson[1,#], Kate Templeton[5,#], Emma C. Thomson[1,17,#,*]

[1]MRC-University of Glasgow Centre for Virus Research (CVR)

[2]MRC Institute of Genetics and Molecular Medicine, University of Edinburgh

[3]Queens Medical Research Institute, University of Edinburgh, UK

[4]Royal Hospital for Children and Young People, Edinburgh, UK

[5]Virology Department, Royal Infirmary of Edinburgh, UK

[6]Roslin Institute, University of Edinburgh, UK

[7]West of Scotland Specialist Virology Centre, Glasgow Royal Infirmary, UK

[8]Institute of Infection Immunity & Inflammation, University of Glasgow, UK

[9]Institute of Evolutionary Biology, University of Edinburgh, UK

[10]Borders General Hospital, Melrose, UK

[11]Public Health Scotland

[12]Institute of Health and Wellbeing, University of Glasgow, UK

[13]School of Medicine, University of St Andrews, UK

[14]Victoria Hospital, Kirkcaldy, UK

[15]Ninewells Hospital & Medical School, Dundee, UK

[16] The COVID-19 Genomics UK (COG-UK) consortium. A full list of members and their affiliations appears in the Supplementary Information.

[17]Department of Clinical Research, London School of Hygiene and Tropical Medicine, UK

[#]Joint first and last authors; all other authors are listed alphabetically.

*Corresponding author

Abstract

Coronavirus disease-2019 (COVID-19) was first diagnosed in Scotland on the 1st of March 2020. During the first month of the outbreak, 2641 cases of COVID-19 led to 1832 hospital admissions, 207 intensive care admissions and 126 deaths. We aimed to identify the source and number of introductions of SARS-CoV-2 into Scotland using a combined phylogenetic and epidemiological approach. Sequencing of 1314 SARS-CoV-2 viral genomes from available patient samples enabled us to estimate that SARS-CoV-2 was introduced to Scotland on at least 283 occasions during February and March 2020. Epidemiological analysis confirmed that early introductions of SARS-CoV-2 originated from mainland Europe (the majority from Italy and Spain). We identified subsequent early outbreaks in the community, within healthcare facilities, and at an international conference. Community transmission occurred after 2nd March, three weeks before control measures were introduced. Earlier travel restrictions or quarantine measures, both locally and internationally, would have reduced the number of COVID-19 cases in Scotland. The risk of multiple reintroduction events in future waves of infection remains high in the absence of population immunity.

Introduction

The pandemic virus severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) has spread rapidly throughout the world following its emergence in Wuhan, China in December 2019 (1-3). SARS-CoV-2 is a highly transmissible *Betacoronavirus*, related to the first SARS virus (4). It causes the clinical syndrome coronavirus disease-2019 (COVID-19), characterised by nonspecific respiratory or gastrointestinal viral symptoms and anosmia. In severe cases, acute respiratory distress syndrome, cardiovascular disease, neurological manifestations, thrombosis and renal failure may occur (5, 6). A rare Kawasaki-like disease has been described in children (7). Despite the mobilisation of significant resources to contain the outbreak, COVID-19 was declared a Public Health Emergency of International Concern (PHEIC) by the World Health Organisation on 30th January 2020 and a pandemic on 11th March 2020 (8, 9). Many countries are now responding to large outbreaks triggering unprecedented social and economic disruption and challenges to local healthcare systems.

The WHO has declared a PHEIC on five occasions since 2009, all as a result of RNA virus outbreaks (H1N1 influenza, Zika, polio and Ebola). Pathogen genomic sequencing is now established as a core component of the modern epidemiological response to such outbreaks, driven by modern nucleic acid sequencing technologies that can rapidly yield entire pathogen genomes from clinical samples (10). The integration of viral genomic data with spatial, temporal and other metadata in a genomic epidemiology framework has allowed enhanced inference of the origin and transmission dynamics of disease outbreaks (11-13). Such an approach is particularly applicable to RNA viruses, as their relatively low-fidelity replication cycle generates mutations in the viral genome at a rate observable over the rapid time scale of an outbreak (14).

In this study, we sequenced laboratory-confirmed cases of COVID-19 in Scotland, and analysed them alongside available international data, in order to estimate the number of

introduction events and early spread of SARS-CoV-2 in the country. During the sampling period, 2641 positive cases of COVID-19 were detected (https://www.opendata.nhs.scot, https://statistics.gov.scot), associated with 1832 hospital admissions, 207 intensive care admissions and 126 deaths. Applying a genomic epidemiology approach to the data, we demonstrate the outbreak is the result of multiple separate introductions of the virus associated with international travel, and that community transmission was quickly established in Scotland, well before the introduction of lockdown countermeasures on the 23rd March 2020.

Results

Multiple introductions of SARS-CoV-2 in Scotland. 1314 SARS-CoV-2 genomes were generated with >90% genome coverage, representing 49% of laboratory-confirmed Scottish COVID-19 cases (Extended Data Fig. 1**)**. The median Ct value was 28.92 (interquartile range 25.81-32.06). Of 1314 individuals, 976 (74%) reported no travel, and 114 (9%) reported travel outside Scotland in the two weeks preceding the onset of symptoms. No travel history was available for 224 (17%) of individuals. Countries visited included Italy (n=41), Spain (n=28), Austria (n=6), Switzerland (n=4), France (n=4), England (n=9), Wales (n=1), Germany (n=1), The Netherlands (n=1), Ireland (n=1), Poland (n=1), Cyprus, (n=1), Turkey (n=1), Argentina (n=1), Egypt (n=1), Tunisia (n=1), Canada (n=1), USA (n=2) and Thailand (n=1). Seven individuals returned from Caribbean cruise holidays. Samples were drawn from all regions of Scotland, with the exception of Orkney Islands, which had only three cases detected during the study period (Extended Data Fig. 2).

The first case of COVID-19 in Scotland (CVR01) was a 51-year-old male from Tayside with mild respiratory symptoms who was tested on the 28th February 2020 (and reported positive on the 1st March 2020). He had returned from Italy after attending a rugby match 9 days earlier (15). The first confirmed case who had not travelled occurred three days later, on 2nd March 2020 (CVR02). Reflecting the change from returning travellers to an older disease-susceptible demographic, the median age of cases increased from 44 (IQR 32-51) in the first week of the epidemic to 62 (IQR 47-76) in the fourth, as infections moved from travel-associated to local community transmission **(**Figure 1; Kruskal-Wallis test, $p<0.001$). A large proportion of sampled cases were healthcare workers and this increased from 5.9% during week 1 to 13.9% by week 4 of the outbreak (Extended Data Fig. 3**)**.

To determine the relationship of the viruses identified in Scotland to global SARS-CoV-2 variants, we inferred an evolutionary tree using viruses sequenced in this study against all complete genomes available from GISAID sampled prior to the 31st of March 2020 (GISAID = 9091, UK-ENG = 7801, UK-SCT = 1314, UK-WLS = 965, UK-NIR = 199) (Figure 2; Extended Data Fig. 1).

While overall limited variability in the genome was observed in keeping with the lower evolutionary rate of coronaviruses compared to other RNA viruses and the recent introduction of the virus into the human population, the strains introduced into Scotland were diverse, representing much of the global distribution of lineages, with the majority fitting within lineage B (16) (Figure 2). Lineage A2 (n = 53) and A5 (n = 22) were the most prevalent A lineages; both are prevalent in Spain (47% of global A2 and 56% of A5 lineage sequences on GISAID derive from Spain). However, low sequencing coverage from some regions of the world (e.g., China) is likely to limit this assessment. Scottish sequences occurred throughout the B lineage with the largest number of sequences (n = 287) found within the B.1 lineage. Most Scottish lineages have also been detected in other parts of the UK, except for the marked absence of B.1.13 (UK sequences = 165 but 0 in Scotland). An increase in the number of sequences within the B.1, B.1.1, B.2 and B lineages coincided with the number of returning travellers from Italy and other parts of Europe in the first half of March, while the second half of March showed a rise in sequences within the A.2 lineage, coinciding with returning travellers from Spain and other parts of the UK (Extended Data Fig. 4).

The 1314 Scottish genomes displayed an average of 3.6 nonsynonymous and 1.9 synonymous nucleotide substitutions in comparison to Wuhan-Hu-1 (**Figure 3**). Two common amino acid

replacements encoding 614D/G and 323P/L were observed in spike and nsp12, respectively and increased in prevalence over the first month of the epidemic (**Figures 3A, 3B, 3C**). Another spike replacement, N439K (reported by the CoVsurver receptor binding surveillance GISAID, https://www.gisaid.org/covsurver) occurred in addition to D614G in a Scottish cluster of 12 sequences. D614G has been hypothesised to be associated with increased transmissibility and escape from neutralisation (17) while N439K occurs at a predicted ACE2 binding site. The 614G spike variant was introduced into Scotland an estimated 191 times in comparison to 91 for 614D and was the predominant variant sampled in March, with 807 sequences compared to 504 for 614D. Three sequences lacked sufficient coverage at residue 614 for the variant present to be determined, one of these being a singleton introduction.

Introductions estimated based on phylogeny alone ranged from 234 (del_sct_lineage) to 1035 (del_sct_introduction) if no attempt was made to merge sibling lineages at the polytomies. We combined the phylogeny, the travel history and date of sampling in order to refine these estimates. Merging sibling lineages provided a conservative phylogenetic estimate of introductions, but failed to identify lineages comprising multiple introductions, of which there are 17 examples within our dataset (Extended Data **Fig.1**). Conversely, failure to merge sibling lineages resulted in the inference of a large number of introductions that were unsupported by the available epidemiological data, for example, 7 individuals with no travel history who were known to have been infected through a superspreader event at a conference were counted as 7 separate introductions when the unmerged phylogenetic algorithm was applied (Extended Data **Fig. 5**). Combining the more conservative del_sct_lineage estimate, with the travel and temporal history permitted the confident identification of 283 independent introductions of SARS-CoV-2 into Scotland. In the majority of cases the phylogenetic lineage placement correlated well with the available epidemiological information for

introduction events. For example, all cases from two known superspreader events clustered together appropriately within single lineages, as did 8/12 of known household contacts. Interestingly, two household contacts (EDB006 and EDB035) with common exposure history through travel to Italy in early March had differing lineage assignments, suggesting acquisition of separate variants whilst travelling in a high-prevalence area at the time of exposure. A second pair of household contacts who had not travelled also had differing lineage assignments (EDB032 and EDB041) associated with 3 unique SNPs, again suggesting an alternative source of infection.

Many of the introductions of COVID-19 into Scotland were from known returning travellers from Europe, mostly Italy (41 out of 283 introductions). 140/234 (59.8%) of phylogenetic lineages were single cases not linked with further cases over time. One hundred and eight of these singleton lineages were not associated with travel, likely corresponding to undetected introductions and community transmission clusters. There were 94 phylogenetic lineages of at least two individuals associated with transmission in varied community settings **(**Figures 4A-C**),** for example, in the Shetland Islands with travel links to Italy, associated with a care home facility, community transmission across central Scotland, and transmission related to an international conference event in Edinburgh at the end of February prior to the first documented SARS-CoV-2 case in Scotland (Extended Data Fig. 5). The latter demonstrates the extent to which conferences and other large gatherings act as superspreader events and contribute to intensified spreading of the virus, supporting findings from similar events in China and Singapore (18, 19). Time-scaled trees were carried out in order to estimate the timing of undetected introductions in Scotland in lineages with no associated epidemiological travel history (Figure 5). These inferred first introductions dated back as far as the 19th February 2020, indicating that community transmission was likely to have occurred

undetected up to 1-2 weeks earlier than the first detected cases. This is in keeping with previously published seroprevalence data from the Scottish blood transfusion service (20).

Shift to community transmission. Ninety introductions (32%) were linked epidemiologically to travel in Europe, seven (2.5%) were linked to Caribbean cruises and seven (2.5%) to travel to the rest of the world. The first case of documented community transmission occurred on the 2nd March 2020 and community transmission was well-established by the 11th March (Figures 1, 4A-C). Figure 4A represents a large cluster which spread across the Scottish central belt between the 13th March and 31st March with no known associated travel. A cluster occurring between the 6th and 30th March focused on the Shetland Islands contained two index cases with travel history to Italy (Figure 4B) but all subsequent cases in the cluster did not report travel. This shift to community transmission is evident in other example clusters: Figure 4C represents a cluster from a care home. We also investigated the possibility of local transmission chains in healthcare settings. As described above, healthcare workers were noted to represent some of the earliest cases, with an increase in cases over time (Extended Data Fig. 3). However, at the beginning of the outbreak, not all cases were community acquired and it was possible to exclude cases of potential nosocomial transmission by comparing sequences from patients on the ward with that of the HCW. For example, a HCW, CVR10 (Extended Data Fig. 1), showed evidence of infection from a virus strain distinct from other samples from the same hospital within lineage B.2.2, whilst the ward patient sequences were from lineages B.1.5(CVR76), B.2.1(CVR07) and B.1.10 (CVR79), indicating community rather than nosocomial infection. The increase in cases over time is likely to reflect a combination of increased testing in HCWs and nosocomial infection. Further studies of genetic epidemiology in healthcare settings are indicated.

Discussion

Our study indicates that SARS-CoV-2 entered the Scottish population through at least 283 separate travel-related introductions. This estimate was calculated using a combination of phylogenetic lineage and epidemiological data. In isolation, each dataset would have resulted in an under-estimate of the number of introductions due to the slow rate of evolution of the SARS-CoV-2 virus and absence of detection of travel-related incident cases respectively. There were 234 phylogenetic lineages, of which 94 (40%) were associated with sustained community transmission (containing at least 2 Scottish sequences without a history of travel) and 140 (60%) singleton sequences with no evidence of onward transmission. 34/94 (36%) of phylogenetic lineages associated with onward transmission involved individuals with a known history of international travel. The majority returned to Scotland from Europe at the end of February and early March following travel to Italy and less commonly to Spain, Austria, Switzerland, France, England, Ireland, Poland, the Caribbean and Thailand. While the first positive case occurred on the 1st March 2020, evidence of community transmission during late February 2020 was supported by epidemiological data and by time-scaled phylogenetic analysis. A shift from travel-associated infection in younger adults to community transmission in older adults and in healthcare workers was noted throughout the first month of the epidemic.

On the 28th January, the UK government recommended against all but essential travel to China and for returning travellers to self-isolate for two weeks upon their return regardless of symptoms (21). However, in this study, cases with direct links to Southeast Asia were rare (only one case associated with travel to Thailand was detected through epidemiological analysis). In contrast, travel to continental Europe in February and March 2020, by then the

epicentre of the global COVID-19 pandemic, was a clear driver of the Scottish outbreak; the majority of the lineages detected in this study were lineage B and related to European sequences. Lineage A, a lineage with a distribution more limited to China at the beginning of the outbreak, was introduced to Scotland on at least ten occasions. The travel history and phylogenetic analysis of these cases indicated that the majority of these occurred via Spain. One introduction was potentially attributable to importation from China (and one from the USA) based on phylogenetic evidence alone. Despite evidence of local transmission in Italy as early as February 21st, the advice from the Scottish Government for returning travellers from Italy to self-isolate was issued only on the 25th of February and was limited to those having returned from specific lockdown areas (22-24). By the time this advice was extended to all travellers on 10th March, the COVID-19 outbreak within Scotland was already being driven by community transmission. A lack of robust measures to manage ingress of high numbers of infected travellers from rapidly emerging pandemic hotspots may have accelerated the course of the outbreak in Scotland and the UK as a whole.

Our data demonstrate that SARS-CoV-2 was introduced to Scotland on many hundreds of occasions, so no single event can be considered to have 'sparked' the epidemic in the rest of the country. One notable introduction event at the beginning of the outbreak occurred at an international conference held in Edinburgh during late February, several days before the first Scottish case was confirmed. Several cases were linked to this event, with the last case within the cluster in the UK occurring on the 27th March, showing that the local public health response was effective in controlling spread. However, the geographical distribution of related sequences is striking, spanning four continents and ten countries. The role of this event in local dispersal of the virus, before a single case had been identified in Scotland, demonstrates that governments should be wary of prematurely relaxing restrictions on large

gatherings and international travel. Importantly, the directionality of transmission cannot be inferred from the phylogeny; SARS-CoV-2 could have been introduced from the Netherlands and other countries to Scotland, exported from Scotland or both.

In parallel with the above introduction, we identified other viral lineages with no epidemiological link to travel as early as three days after the first detection of infection, suggesting earlier introduction to Scotland than the first detected case, reported on the 1st March 2020 (15). This is supported by the analysis of time-scaled phylogenies, which infers a common ancestor from mid-late February for some Scotland-specific clades. The majority of individuals had no recorded link to travel by the 12th March, only 2 days after substantial travel and physical distancing restrictions were put in place. Importantly these data are suggestive of introductions and community spread well before initial detection. The epidemic in the UK expanded rapidly, prompting the government to respond with restrictive public health measures or "lockdown" to disrupt transmission. These phylogenetic data will provide a baseline for granular real-time sequencing of infections as cases rise and fall over time and will be used as a measure of the success of current measures with the potential to contribute to decision-making around the easing (or tightening) of public health measures. In this study, the use of epidemiology or phylogenetic lineage assignment alone would have resulted in an under-estimate of the number of introductions of the virus (114 and 234 respectively) due to undetected cases reaching the attention of public health authorities and due to the slow rate of evolution of the virus. Integration of genomic sequence data with traditional case-finding and contact tracing has the potential to enhance descriptive epidemiology and deliver more targeted control measures. However previous interventions of this type have been largely retrospective. The challenge will be to develop a framework where phylogenetic information

is delivered in real-time in an easily actionable format to public health and infection control teams.

This study has some limitations. We sampled only around half of positive samples detected during the initial outbreak, therefore some introduction events will have been missed. Further, while laboratory and hospital case notes were available for review, public health records were not accessed to record linkage with sequence data. Our analysis is likely to have been affected by a shift in sampling all symptomatic individuals to hospitalised patients and healthcare workers only during mid-March. Some travel-related introductions may therefore not have been detected (as evidenced by several clusters with no evidence of travel related infection after this time) and healthcare workers may have been over-sampled in the analysis. While introduction of erroneous variation during sequencing or genome assembly is possible, we estimated experimentally that such rates were extremely low and should not affect clustering patterns in phylogenetic analyses. This analysis, based on the available phylogeny, is most likely to be an under-estimate, despite dense sampling of the beginning of the outbreak, as many events may be linked to an identical sequence due to the slow evolution of SARS-CoV-2. Several events that we counted as introductions were based on epidemiological history but could not be resolved by phylogeny alone. The slow evolutionary rate of the virus also means that linkage of sequences does not categorically prove that transmission events have occurred and require correlation with epidemiological information. Exclusion of linkage may be inferred with more certainty. Nevertheless, as the variation present within the global diversity is well-represented in Scotland, reflecting a high number of near simultaneous introductions, tracking of the outbreak is feasible and can be used to refine public health interventions. Finally, as further sequence data becomes available from other countries, the topology of the global tree is predicted to change, which is likely to have the effect of a slight increase in our estimate of introductions.

In summary, the first month of the COVID-19 outbreak in Scotland was associated with multiple introductions related to returning travellers from Europe, early community transmission, and clusters related to large indoor events and healthcare facilities. An earlier lockdown from countries with a high burden of cases such as Italy and other measures such as quarantine of travellers from high-risk areas might have prevented escalation of the outbreak and multiple clusters of ongoing community transmission. Combining genomic data with epidemiological data has the potential to inform public health intervention policy. Multiple travel-associated introductions during the first wave of infection in Scotland highlight this as a significant risk for re-introduction in future waves.

Data availability statement

All consensus genomes are available from the GISAID database (https://www.gisaid.org), the COG-UK consortium website (https://www.cogconsortium.uk/data/) and BAM files from the European Nucleotide Archive's Sequence Read Archive service, BioProject PRJEB37886 (https://www.ebi.ac.uk/ena/data/view/PRJEB37886). See Source Data Table 1 for IDs and dates of sampling.

Code availability statement

Reads were size filtered, demultiplexed and trimmed with Porechop (https://github.com/rrwick/Porechop v0.2.4). The ARTIC bioinformatic pipeline v1.1.3 (https://github.com/artic-network/artic-ncov2019) was used for generation of consensus sequences. The grapevine pipeline (https://github.com/COG-UK/grapevine) was used for generating the phylogeny based on all data available on GISAID and COG-UK up until 23-08-2020. Sequence typing was carried out using PANGOLIN (https://github.com/hCoV-

2019/pangolin v2.0.5). Scottish clusters were investigated following the DELTRAN implementation method using clusterfunk (https://github.com/cov-ert/clusterfunk v0.0.3).

Online Methods

Samples. Up to 300 samples per week were selected prospectively following ethical approval from the relevant national biorepository authorities (16/WS/0207NHS and10/S1402/33) between 1st March and 1st April 2020. This work was conducted as part of the UK Government funded Covid-19 Genomics (COG-UK) consortium, which was set up in March 2020 and aims to provide representative, large-scale and rapid whole genome virus sequencing across the United Kingdom (25). 50% of samples were randomly selected to achieve a representative target for all Scottish health boards and 50% to cover suspected healthcare-related nosocomial infections as they occurred. Health boards with a small population size were reported at a minimum of 5 sequences per region to avoid deductive disclosure. The Royal Infirmary of Edinburgh (RIE) and West of Scotland Specialist Virology Centre, NHSGGC conducted diagnostic real-time RT-PCR to detect SARS-CoV-2 positive samples, following nucleic acid extraction utilising the NucliSENS® EasyMag® and Roche MG96 platforms (26). Residual nucleic acid from 1314 samples underwent whole genome next generation sequencing at the MRC-University of Glasgow Centre for Virus Research (CVR) and the RIE. Clinical details including recent travel history were obtained from assay request forms submitted to the diagnostic laboratory, and where available electronic patient records and local public health databases. A travel history was defined as travel from any country (including England, Wales and Northern Ireland) by any means of transport including by air, ferry, train and car within the two weeks prior to the onset of

symptoms. Statistical comparisons (two-tailed) were carried out using R version 3.6.3 by Kruskal-Wallis test.

Rapid sequencing protocol using Oxford Nanopore Technologies (ONT) Following extraction, libraries were prepared utilising protocols developed by the ARTIC network (v1 and v2) https://artic.network/ncov-2019. 50 fmol of library pools were loaded onto each flow cell (R9.4.1). Sequencing was conducted in MinKNOW version 19.12.5. Raw FAST5 files were basecalled using Guppy version 3.4.5 in high accuracy mode using a minimum quality score of 7. RAMPART v1.0.5 was used to visualise read mapping in real-time. Reads were size filtered, demultiplexed and trimmed with Porechop (https://github.com/rrwick/Porechop v0.2.4), and mapped against reference strain Wuhan-Hu-1 (MN908947). Variants were called using Nanopolish 0.11.3 and accepted if they had a log-likelihood score of greater than 200 and minimum read coverage of 20 following the ARTIC bioinformatic pipeline v1.1.3 (https://github.com/artic-network/artic-ncov2019). This protocol was used by CVR and RIE sites.

High throughput sequencing protocol using Illumina MiSeq. Amplicons were generated as described above. DNA fragments were cleaned using AMPURE beads (Beckman Coulter) and 40ng used to prepare Illumina sequencing libraries with a DNA KAPA library preparation kit (Roche). Indexing was carried out with NEBNext multiplex oligos (NEB), using 7 cycles of PCR. Libraries were pooled in equimolar amounts and loaded on a MiSeqV2 cartridge (500 cycles). Reads were trimmed with trim_galore (http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/ v0.6.5) and mapped with BWA (27) to the Wuhan-Hu-1 (MN908947) reference sequence, followed by primer trimming and consensus calling with iVar v1.2.2 (28) and a minimum read coverage of 10.

This protocol was followed by the CVR site whereas samples sequenced at the Wellcome Sanger Institute (labelled with the prefix GCVR) follow a diverse Illumina library preparation method and use the NovaSeq instrument. Consensus sequence data was highly correlated between different sequencing platforms.

Sequence Data. Consensus sequences with >90% coverage were included. All consensus genomes are available from the GISAID database (https://www.gisaid.org), the COG-UK consortium website (https://www.cogconsortium.uk/data/) and BAM files from the European Nucleotide Archive's Sequence Read Archive service, BioProject PRJEB37886 (https://www.ebi.ac.uk/ena/data/view/PRJEB37886). See ource data Table 1 for IDs and dates of sampling. Amplicon-based sequencing using Illumina instruments has previously been applied to RNA viruses for accurate identification of intra-host variants, with performance comparable to metagenomics (28). A similar protocol has been applied to dengue virus using multiplex PCR tiling on Oxford Nanopore sequencer, resulting in 99.69-99.92% consensus identity when compared to those produced by Illumina (29). In order to minimise consensus level errors in regions of lower genome coverage and in samples with very low viral load (28), we opted to exclude samples with poor genome coverage (<90%), typically associated to low viral load while a) sequencing at a high depth with an average of 5000X per sample (CVR), or b) actively excluding low viral load samples (Ct>30) while sequencing at an average depth of 500X (RIE).

Phylogenetic analysis. The grapevine pipeline (https://github.com/COG-UK/grapevine) was used for generating the phylogeny based on all data available on GISAID and COG-UK up until 23-08-2020. Briefly, the pipeline cleans and filters reads based on quality and consensus

coverage, aligns the sequences to the reference, types the sequences using PANGOLIN (https://github.com/hCoV-2019/pangolin v2.0.5), merges the COG-UK sequences with the GISAID alignment, masks homoplasies (position 11083 relative to Wuhan-Hu1) and reconstructs the maximum likelihood phylogeny using FastTree v2.1.11 (30). Wuhan/WH04/2020 was used as an outgroup for the phylogeny. Finally, the phylogeny was pruned to keep only sequences prior to 31-March-2020 to retrospectively investigate the introductions into Scotland alongside known travel history and epidemiological information. Scottish clusters were investigated following the DELTRAN implementation method (31) with the following four steps using clusterfunk (https://github.com/cov-ert/clusterfunk v0.0.3): 1) Scottish and non-Scottish sequences were coded as a binary trait and annotated on the tree, 2) The ancestral state of the Scottish trait was reconstructed on the phylogeny, 3) Transition nodes from non-Scottish to Scottish were identified as "DELTRAN Scottish introductions" on the tree (del_sct_introductions), 4) Transitions were then merged such that sibling introductions are clustered together ensuring that identical sequences were given the same "DELTRAN Scottish lineage" number (del_sct_lineage). This is a fully automated approach which has some caveats: due to the slow evolutionary rate of the virus, the tree has many polytomies and thus the parsimony reconstruction of a trait on the tree becomes ambiguous, hence, the need to merge sibling introductions together. Additionally, where there are exports and subsequent re-introductions (unlikely during the first month of the outbreak), the DELTRAN approach would label these as the ancestral introduction.

Due to the relatively low evolutionary rate of SARS-CoV2, it is often difficult to differentiate introductions on the basis of the phylogeny alone, as sequences from distinct introductions may be identical or cluster on the phylogeny (thereby under-estimating introductions) and incorporation of the assumption that identical sequences at the base of the tree represent distinct introductions may result in an over-estimate. In order to refine the phylogenetic

estimate of the number of times the virus was introduced into Scotland we used the more conservative del_sct_lineage as a framework upon which we layered available epidemiological and temporal data. Lineages with no known travel related cases were counted as a single introduction regardless of the number of taxa present. Multiple travel related cases within a lineage were each classed as separate introductions. Where a single travel related case was recorded within a phylogenetic lineage this was counted as a single introduction if it occurred up to 14 days after the first case in the cluster. If the travel related case returned to Scotland after the lineage had already been detected, or if the travel related case was sampled more than 14 days following the first detection of the lineage within Scotland, it was counted as a further introduction.

Seven large UK lineages that were predominantly sampled from Scotland were investigated further using time-scaled trees with treetime (32) (clock-rate of 0.001 and 10 iterations) in order to give an indication of the detection lag (time between the most recent common ancestor and the first sequenced sample).

References

1 Wu, F. et al. A new coronavirus associated with human respiratory disease in China. *Nature* **579**, 265-269, doi:10.1038/s41586-020-2008-3 (2020).

2 Zhou, P. et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* **579**, 270-273, doi:10.1038/s41586-020-2012-7 (2020).

3 Zhu, N. et al. A Novel Coronavirus from Patients with Pneumonia in China, 2019. *N Engl J Med* **382**, 727-733, doi:10.1056/NEJMoa2001017 (2020).

4 Coronaviridae Study Group of the International Committee on Taxonomy of Viruses. The species Severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2. *Nature Microbiology* **5**, 536-544, doi:10.1038/s41564-020-0695-z (2020).

5 Huang, C. et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *Lancet* **395**, 497-506, doi:10.1016/S0140-6736(20)30183-5 (2020).

6 Mao, L. et al. Neurologic Manifestations of Hospitalized Patients With Coronavirus Disease 2019 in Wuhan, China. *JAMA Neurol* **77**, 683-690, doi:10.1001/jamaneurol.2020.1127 (2020).

7 Verdoni, L. et al. An outbreak of severe Kawasaki-like disease at the Italian epicentre of the SARS-CoV-2 epidemic: an observational cohort study. *Lancet* **395**, 1771-1778, doi:10.1016/S0140-6736(20)31103-X (2020).

8 WHO. Statement on the second meeting of the International Health Regulations (2005) Emergency Committee regarding the outbreak of novel coronavirus (2019-nCoV) (World Health Organisation, 30 January 2020).

9 WHO. WHO Director-General's opening remarks at the media briefing on COVID-19 - 11 March 2020 (World Health Organisation, 11 March 2020).

10 Quick, J. et al. Real-time, portable genome sequencing for Ebola surveillance. *Nature* **530**, 228-232, doi:10.1038/nature16996 (2016).

11 Grubaugh, N. D. et al. Genomic epidemiology reveals multiple introductions of Zika virus into the United States. *Nature* **546**, 401-405, doi:10.1038/nature22400 (2017).

12 Grubaugh, N. D. et al. Tracking virus outbreaks in the twenty-first century. *Nature Microbiology* **4**, 10-19, doi:10.1038/s41564-018-0296-2 (2019).

13 Kafetzopoulou, L. E. et al. Metagenomic sequencing at the epicenter of the Nigeria 2018 Lassa fever outbreak. *Science* **363**, 74-77, doi:10.1126/science.aau9343 (2019).

14 Biek, R., Pybus, O. G., Lloyd-Smith, J. O. & Didelot, X. Measurably evolving pathogens in the genomic era. *Trends Ecol Evol* **30**, 306-313, doi:10.1016/j.tree.2015.03.009 (2015).

15 Hill, K. J. et al. The index case of SARS-CoV-2 in Scotland. *J Infect* **81**, 147-178, doi:10.1016/j.jinf.2020.03.022 (2020).

16 Rambaut, A. et al. A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nature Microbiology*, 2020.2004.2017.046086, doi:10.1038/s41564-020-0770-5 (2020).

17 Korber, B. et al. Tracking Changes in SARS-CoV-2 Spike: Evidence that D614G Increases Infectivity of the COVID-19 Virus. *Cell* **182**, 812-827 e819, doi:10.1016/j.cell.2020.06.043 (2020).

18 Tong, Z. D. et al. Potential Presymptomatic Transmission of SARS-CoV-2, Zhejiang Province, China, 2020. *Emerg Infect Dis* **26**, 1052-1054, doi:10.3201/eid2605.200198 (2020).

19 Pung, R. et al. Investigation of three clusters of COVID-19 in Singapore: implications for surveillance and response measures. *Lancet* **395**, 1039-1046, doi:10.1016/S0140-6736(20)30528-6 (2020).

20 Thompson, C. P. et al. Detection of neutralising antibodies to SARS coronavirus 2 to determine population exposure in Scottish blood donors between March and May 2020. *medRxiv*, 2020.2004.2013.20060467, doi:10.1101/2020.04.13.20060467 (2020).

21 Foreign and Commonwealth Office. Foreign Office advises against all but essential travel to China. (28 January 2020). at https://www.gov.uk/government/news/fco-advises-against-all-but-essential-travel-to-mainland-china

22 Scottish Government. Preparations for coronavirus stepped up. (25 February 2020). at https://www.gov.scot/news/preparations-for-coronavirus-stepped-up

23 ECDC. Communicable disease threats report; 16-22 February 2020, week 8. (European Centre for Disease Prevention and Control, Stockholm, 21 Feb 2020).

24 ECDC. Outbreak of novel coronavirus disease 2019 (COVID19): situation in Italy – 23 February 2020. (European Centre for Disease Prevention and Control, Stockholm, 23 Feb 2020).

25 The COVID-19 Genomics UK (COG-UK) consortium. An integrated national scale SARS-CoV-2 genomic surveillance network. *The Lancet Microbe* **1**, e99-e100, doi:10.1016/s2666-5247(20)30054-9 (2020).

26 Corman, V. M. et al. Detection of 2019 novel coronavirus (2019-nCoV) by real-time RT-PCR. *Euro Surveill* **25**, 2000045, doi:10.2807/1560-7917.ES.2020.25.3.2000045 (2020).

27 Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* (Oxford, England) **25**, 1754-1760, doi:10.1093/bioinformatics/btp324 (2009).

28 Grubaugh, N. D. et al. An amplicon-based sequencing framework for accurately measuring intrahost virus diversity using PrimalSeq and iVar. *Genome Biol* **20**, 8, doi:10.1186/s13059-018-1618-7 (2019).

29 Stubbs, S. C. B. et al. Assessment of a multiplex PCR and Nanopore-based method for dengue virus sequencing in Indonesia. *Virol J* **17**, 24, doi:10.1186/s12985-020-1294-6 (2020).

30 Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2--approximately maximum-likelihood trees for large alignments. *PLoS One* **5**, e9490, doi:10.1371/journal.pone.0009490 (2010).

31 Farris, J. S. Methods for Computing Wagner Trees. *Systematic Zoology* **19**, 83-&, doi:Doi 10.2307/2412028 (1970).

32 Sagulenko, P., Puller, V. & Neher, R. A. TreeTime: Maximum-likelihood phylodynamic analysis. *Virus Evol* **4**, vex042, doi:10.1093/ve/vex042 (2018).

Correspondence and requests for materials should be addressed to Professor Emma C. Thomson, Sir Michael Stoker building, MRC-University of Glasgow Centre for Virus Research, 464 Bearsden Road, Glasgow, UK G61 1QH emma.thomson@glasgow.ac.uk

Acknowledgments

We thank all Scottish NHS virology laboratories who provided samples for sequencing and the global researchers who have shared genome data on GISAID, https://www.gisaid.org. We thank Richard Kuo, Tim Regan and Amanda Warr (Roslin Institute, University of Edinburgh) for sequencing reagents and Scott Arkison for HPC maintenance. **Funding**: MRC (MC UU 1201412), Wellcome Trust Collaborator Award (206298/Z/17/Z – ARTIC Network, Wellcome Trust Award 204802/Z/16/Z (TCW), University of Glasgow. COG-UK is supported by funding from the Medical Research Council (MRC) part of UK Research & Innovation (UKRI), the National Institute of Health Research (NIHR) and Genome Research Limited, operating as the Wellcome Sanger Institute. CLIMB is funded by the Medical Research Council (MRC) through grant MR/L015080/1.

Competing Interests

The authors declare no competing interests.

Author contributions

C.B., R.D., M.D.G., M.P.McH., S.R., E.W., K.T., B.J.P., E.J., K.M., R.G., I.S., A.M. K.L., J.S. and N.J acquired the clinical samples, performed extractions and quantitative PCR. T.W., R.D., M.P.McH., K.T., B.J.P., E.J., K.M., P.A., S.A., C.D., M.N., S.E.M, R.S., K.L., J.S., N.J. and E.C.T. were responsible for the selection, reception, verification and organisation of clinical samples. J.S., T.W., E.W., A.C., R.G., A.H., E.J., N.J., K.L., D.A.McA., K.M., B.J.P., R.S., I.S., M.T.G.H., K.T., E.C.T. collected and analysed the clinical and epidemiological data. T.W., R.D., M.D.G., V.H., B.J., J.T.McC., M.P.McH., A.O'T., A.R., S.R., K.T., E.A.C., K.B., S.C., N.J., E.C.M.L., D.M., J.N., K.N., Y.A.P., K.S., L.T. and A.d.S.F. acquired the sequencing data and optimised protocols. M.D.G., V.H., B.J., J.T.McC., A.O'T., A.R., S.R., J.H., R.J.O., V.B.S. and D.L.R. were responsible for the genome alignment and variant

calling. J.H., J.B.S., E.C.T., J.S. and D.L.R. performed phylogenetic analysis, integration of epidemiological analysis and genome mutation analysis. A.d.S.F., J.S., T.W., J.H., D.L.R., M.T.G.H., S.S., T.W., K.T., M.P. and E.C.T analysed and interpreted the data. A.d.S.F., J.S., T.W., J.H., D.L.R., and E.C.T. drafted the manuscript and all authors reviewed and contributed to the final version. A.d.S.F., J.S., T.W., J.H., D.L.R., A.E., K.T. and E.C.T. were responsible for the conception and design of the study. M.P., E.C.T. and A.R. acquired funds to support the study.

All authors approved the submitted version and have agreed both to be personally accountable for the author's own contributions and to ensure that questions related to the accuracy or integrity of any part of the work, even ones in which the author was not personally involved, are appropriately investigated, resolved, and the resolution documented in the literature .

Consortium Information

The following authors are members of the COG-UK consortium.

Ana da Silva Filipe, James G Shepherd, Thomas Williams, Joseph Hughes, Elihu Aranday-Cortes, Patawee Asamaphan, Carlos Balcazar, Kirstyn Brunker, Stephen Carmichael, Rebecca Dewar, Michael D. Gallagher, Verity Hill, Daniel Mair, John T McCrone, Martin P. McHugh, Jenna Nichols, Marc Niebel, Kyriaki Nomikou, Richard J Orton, Áine O'Toole, Yasmin A. Parr , Andrew Rambaut, Stefan Rooke, Rajiv Shah, Joshua B. Singer, Katherine Smollett, Igor Starinskij, Lily Tong, Vattipally B. Sreenu, Matthew T.G. Holden, David L Robertson, Kate Templeton, Emma C. Thomson

Figure Legends

Figure 1. Spatial and demographic features of sequenced cases. (A) Histogram of daily cases stratified by confirmed travel history. (B) Ages of positive cases referred for diagnostic testing to the WOSSVC by week of the outbreak in Scotland (showing median (middle bar), interquartile range (box), range (vertical line) and outliers). Median age increased from 44 in week one (n=12, IQR 18.8) to 50 in week 2 (n=52 IQR 25.8), 61 in week 3 (n = 183 IQR 30.5), and 62 (n=516 IQR 29) in week 4 (Kruskal-Wallis rank sum test, H = 27.47, df = 3, $p < 0.001$). (C) Spatial distribution and associated travel history over the first four weeks of the outbreak. Testing of community cases ceased on day a1 of week 3 (14th March) and lockdown occurred on day 2 of week 4 (23rd March).

Figure 2. **Phylogenetic relationships of Scottish genomes to all SARS-CoV-2 genomes.** Known travel histories are indicated (see keys). Global sequence data were available from GISAID. Non-Scottish sequences were subsampled for presentation purposes by keeping a single sequence per date/country/lineage reducing the total number of non-Scottish sequences from 17578 to 5389. The scale bar indicates substitutions per nucleotide site.

Figure 3. **Detection of polymorphisms within the SARS-CoV-2 genome** (A) Frequency of amino acid residue at position 614 in spike protein for Scottish sequences in March (D - aspartate, G - glycine, X - undetermined). (B) Amino acid residue at spike 614 for each Scottish lineage introduction by date of first detection. (C) A visualisation of the genetic variation across the entire genome observed in the 1314 SARS-CoV-2 genomes in Scotland. Nonsynonymous (pink) and synonymous (green) substitutions (with respect to Wuhan-Hu-1, GenBank accession number MN908947) are represented in colour in each row. The mutations are plotted in a grid format where each row is a sample and each column is a unique genome position; mutations have been filtered to only display those observed in more

than two samples. Mutation labels have been added into the heatmap showing the genome position and ORF name with amino acid number/mutation (non-synonymous mutation details are highlighted in bold). Labels for each of the ORFs of the SARS-CoV-2 genome are shown in the lower panel. The plot was created using the d3heatmap package in R with samples (rows) ordered according to Ward's clustering.

Figure 4. **Selected phylogenetic clusters associated with introduction events** (A) A spatially distributed outbreak across the Central Scotland belt without any known links to travel, (B) A cluster focused on Shetland associated with two linked cases with travel to Italy, (C) a focal outbreak in a residential facility not associated with any known travel.

Figure 5: Time-scaled trees in lineages with no associated epidemiological travel history Time scaled trees were produced with treetime (clock-rate of 0.001 and 10 iterations) for 7 established lineages in order to give an indication of the detection lag (time between the most recent common ancestor and the first sequenced sample). Scottish lineage numbers are prefixed by SCT and equivalent UK lineages are shown in brackets.