COVID-19 Genomics UK (COG-UK) Consortium

Report #7 - 28th May 2020

This report is provided at the request of SAGE and includes information on the ongoing state of the research being carried out. It should not be considered formal or informal advice. The conclusions of the ongoing scientific studies may be subject to change as further evidence becomes available and as such any firm conclusions would be premature.

Executive Summary

- With 17 sequencing centres now active, COG-UK has passed a significant milestone by sequencing and analysing more than 20K SARS-CoV-2 genomes to date, corresponding to 56% of the global total number of genomes.
- COG-UK data is being used increasingly for genomic epidemiological investigation of SARS-CoV-2 outbreaks at local and regional levels across the UK and is providing valuable insight regarding virus introductions and the associations between cases in potential outbreaks that would not be possible using non-genomic epidemiological approaches.

COG-UK update

Two additional sequencing centres have been brought online over the past fortnight (Northumbria and Portsmouth) bringing the total number of active sequencing sites to 17.

By the data cut off for this report, the total number of virus genomes sequenced by COG-UK stands at 20,637 (Figure 1), which is a significant milestone and a highly notable achievement in just 11 weeks since the establishment of the consortium.

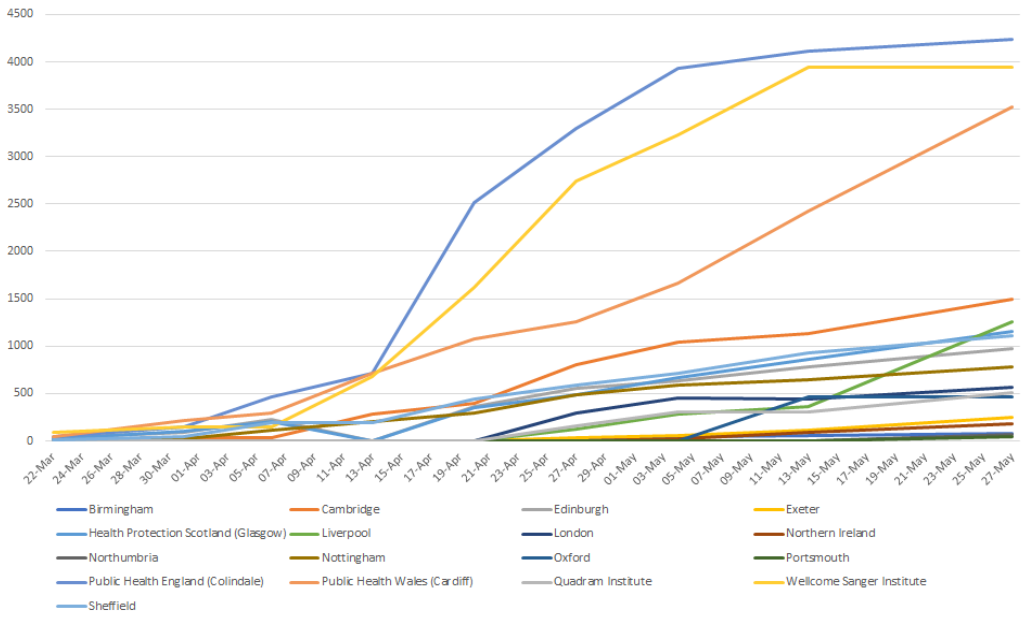

Figure 1: Number of SARS-CoV-2 genomes sequenced and analysed by the COG-UK centres by date. By the 27th May, a total of 20,637 genomes have been sequenced across all COG-UK centres. Occasional decreases owe to increased stringency in quality control parameters affecting inclusion of genomes.

The number of SARS-CoV-2 genomes generated by COG-UK continues to grow at a faster rate than reported by any other country and our data now accounts for 56% of the total number of genomes reported globally (Figure 2).

COG-UK data is now being used by researchers and epidemiologists nationally and internationally to understand the patterns underpinning the spread of SARS-CoV-2 and to provide the context for local genomic epidemiology efforts (see "SARS-CoV-2 genomic epidemiology in the UK" below).

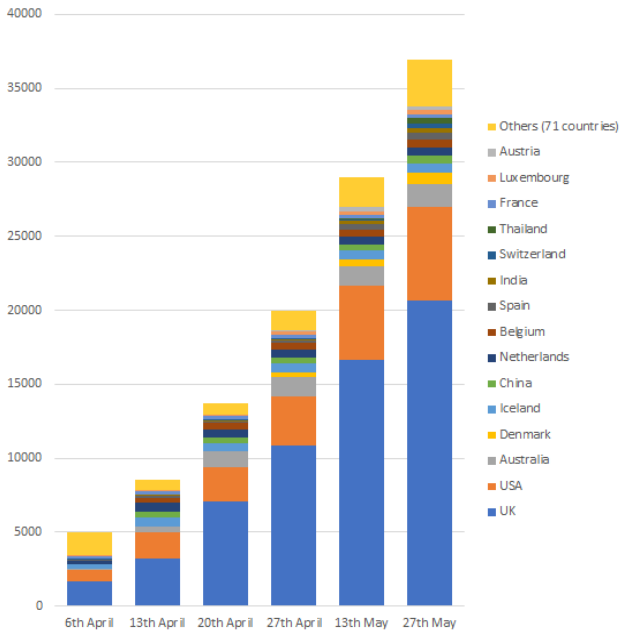

Figure 2: Number of SARS-CoV-2 genomes sequences reported (in MRC CLIMB and GISAID). Data shown up to 27th May.

Analysis updates

SARS-CoV-2 genomic epidemiology in the UK

Lead author: Nick Loman (University of Birmingham)

There is a great deal of interest in using genomics to help investigate potential SARS-CoV-2 outbreak clusters, particularly in institutional settings such as hospitals, care homes and prisons. Genomic epidemiology has been widely used to investigate microbial outbreaks identified by traditional epidemiological studies. In addition, genome data alone has been used more recently to identify and define outbreaks for further epidemiological investigation, for example for outbreaks of *Salmonella* food poisoning and tuberculosis in the UK.

However, the COVID-19 pandemic differs from other outbreaks in several important ways. There is uncertainty about how effective the various applications of genomic epidemiology of SARS-CoV-2 will be and how best to interpret the data generated. This is an active field of research and COG-UK aims to generate reliable actionable findings to aid local public health teams and clinicians in tackling COVID-19.

This summary will briefly describe how genome data can be reliably used now, and identify areas still in need of further improvements, drawing from experiences with genomic epidemiology of other pathogens, and early findings from SARS-CoV-2 (including case studies from several UK regions).

Similarities and differences between COVID-19 and other epidemics:

- All SARS-CoV-2 genomes share a very recent common ancestor (around December 2019). This limited timeframe means that only a relatively small amount of genetic diversity has accrued among SARS-CoV-2 lineages. The level genetic diversity is even lower than that observed in some pandemic influenza viruses, because of the lower rate of mutation in SARS-CoV-2. Other pathogens that have been circulating in human populations for a long time (e.g. *Salmonella* sp. or *Mycobacterium tuberculosis*) typically exhibit much higher levels of genetic diversity.
- The total number of cases is extremely large and internationally distributed (as expected of a pandemic virus), but sampling is biased towards a small number of countries (see Figure 2).
- SARS-CoV-2 has, at present, a moderately high rate of homoplasy (apparently repeated mutations in separate lineages), possibly as a consequence of RNA editing (a process by which some viruses modify their mRNA during transcription to express multiple proteins from the same gene) or owing to technical artifacts. In this manner it is different from other recent viral epidemics such as Ebola. Homoplasy can complicate some applications of genomic epidemiology.
- The widespread response to the outbreak (e.g. introduction of lockdown measures) is in contrast to other outbreaks, such as those associated with tuberculosis or recent pandemic influenza.

Principles of genomic epidemiology:

- Genetic similarity between pathogen genomes in samples collected from patients can reveal useful information about the degree to which the patients are or aren't epidemiologically linked.

- Phylogenetic reconstructions provide additional information about population-level and international trends in transmission and virus evolution.
- Viruses and other pathogens evolve according to a molecular clock (i.e. their genomes accrue genetic variation at a rate that is relatively constant through time, such that the genetic difference between any two isolates is proportional to the time since they last shared a common ancestor).
- Underlying changes in the relatedness and genetic diversity of the virus population relate to changes in the progression of the epidemic, and these can be investigated and characterised using evolutionary and population genetics theory.

Questions that can be answered readily with genomic epidemiology:

- Are two cases unlikely to be part of the same transmission chain? (belong to separate lineages)
- What is the minimum number of independent introductions into a location (identification of multiple lineages in same place)

Lessons already learned from genomic epidemiology:

- It is much easier to rule-out transmission linkage among cases using virus genomes than it is to confirm or prove linkage; the chances of two genomes in two genetically-distinct virus lineages being part of the same transmission chain are mathematically extremely low. However, for a virus population with low genetic diversity (such as SARS-CoV-2 at the present time), the chance that two epidemiologically-unlinked isolates share a similar or even identical genomes can be quite high. This is particularly problematic for SARS-CoV-2 genomes where any given SARS-CoV-2 evolves at a rate of approximately 2.5 mutations/month. Across the entire SARS-CoV-2 epidemic we observe virus genomes that are identical to their ancestors even after weeks of transmission has occurred (mean time between identical genomes 14 days, 95th percentile 47 days). Further, extensive national and international spread means that identical genomes can be observed in different countries. Although they share a recent common ancestor, these cases are not epidemiologically linked in a meaningful way (i.e. it is unlikely that *e.g.* any given patient in Australia infected someone in Canada, although the virus may have travelled between those countries via a direct or indirect route and be indistinguishable genetically).
- More complex mathematical frameworks can be employed to enable the estimation of epidemiological parameters from sequence data, but to be most effective these methods require large sample sizes and detailed information on the sampling protocol to ensure that spurious signals are not introduced. COG-UK has been set up to enable appropriate sampling to be undertaken and to support these sorts of analyses.
- Effective interpretation of genomic data requires appropriate metadata and the task of acquiring and linking metadata on a national level is critical - by working as a single consortium COG-UK simplifies this process and provides a platform for data sharing and analysis.
- The presence of a cluster of related isolates does not prove that cases are linked by a transmission chain, or prove direction of transmission. Clusters can only be interpreted if the “background” cases are well sampled, through random, systematic and dense sampling of cases. This justifies the COG-UK approach of sequencing a very high number of genomes. However the COG-UK ‘background’ sample mainly relies on sequencing samples taken for clinical and public health reasons and as such

is not intrinsically systematic (e.g. community cases are highly underrepresented). We are attempting to find ways of producing a more representative sample set.

The following case studies highlight how genomic epidemiology is already being applied in multiple regions of the UK to tackle COVID-19 by COG-UK consortium members.

East of England

https://doi.org/10.1101/2020.05.08.20095687

Note that this study has since been peer reviewed. Some minor changes have been made to specific numbers quoted below, but the substance of the manuscripts is unchanged.

Study leads

Ian Goodfellow (University of Cambridge) and Estée Török (University of Cambridge, Cambridge University Hospitals)

Question addressed

What is the utility of rapid sequencing of SARS-CoV-2 combined with detailed epidemiological analysis to investigate healthcare-associated COVID-19 infections?

Finding summary

Between the 13th of March and the 24th of April 2020, a prospective surveillance study at Cambridge University Hospitals NHS Foundation Trust (CUH) was undertaken using rapid SARS-CoV-2 genome sequencing from PCR-positive diagnostic samples at CUH and 17 hospitals in the East of England.

From the 5191 confirmed COVID-19 patients in the East of England during this period, 1000 samples were sent for sequencing, producing 747 complete genomes after quality control and de-duplication.

The frequency and location of mutations across the genomes were examined and the viruses assigned to lineages accordingly. Most belonged to global lineage B.1. Phylogenetic trees were constructed to explore potential clustering and to assess correlation with ward location and/or suspected cases of hospital-acquired infection (HAI). By contrast to samples from patients presenting to the emergency department, which were phylogenetically dispersed (reflecting likely unconnected transmission events), some samples collected within CUH and among outpatients formed clusters, suggestive of linked transmission chains within healthcare settings.

A combined epidemiological and genomic analysis of 299 patients identified 26 genomic clusters involving 114 patients. The clusters were defined as groups of identical genomes and the two largest each contained 15. Based on assessment of patient medical records (including address, social setting, clinical details and ward movements) 63 cases (55.3%) had a strong epidemiological evidence to support recent transmission,

27 (23.7%) had intermediate epidemiological evidence and 24 (21.1%) had no evidence of connected transmission. The detected clusters included both those for which a connection was already suspected and those which were not previously suspected of being linked, but prompted more detailed investigations which uncovered potential epidemiological links.

The data obtained were fed back to clinical, infection control and hospital management teams in weekly meetings resulting in reviews of infection control procedures and patient safety.

For example, for one cluster of assumed HAI on a single ward, confirmation of a connected cluster supported the assumption of recent ward-based transmission. A cluster among patients on a different ward could be linked with infection of three healthcare workers, only two of which had worked on the ward (the third having had professional contact with the other HCWs), prompting a review of infection control and PPE measures. In another example, a cluster of six patients with identical genomes were revealed by epidemiological investigation to have attended an outpatient renal dialysis unit on the same days of the week. This led to a review of infection control procedures which identified shared patient transport as the most likely risk factor. Although there were concerns about transmission between the dialysis unit and inpatient renal wards, this was ruled out by genomics, which demonstrated that the two clusters belonged to two distinct viral lineages. Finally, a cluster of cases identified links between healthcare workers and a care home which had not been detected by clinicians or infection control. In summary, the genomic data provided evidence to support or refute transmission between epidemiologically linked cases.

Other UK regions

Study leads

Various COG-UK consortium members.

Question addressed

How are consortium members attempting to use genome data to help outbreak investigations locally?

Finding summary

One COG-UK consortium member described how they had used genome data to investigate 5 hospital wards. For two of the wards there were a variety of SARS-CoV-2 lineages detected, indicating multiple introductions. One ward experienced a cluster of cases of a single lineage that was rare in the UK population (138 out of >17K sequenced cases by 21st May) suggesting that it might have resulted from a single introduction linked to the initial outbreak of this lineage in the region. This ward is associated with care homes and investigations are ongoing to determine the predominant lineages in those homes. The remaining two wards had infections with the UK5 SARS-CoV-2 lineage (B.1.1.1 global lineage) which is the predominant lineage in the UK. Interestingly, an infected patient was moved from the ward with the rare lineage onto one of these other wards, but there was no evidence for onward transmission of the rare lineage.

Another COG-UK consortium member is using genome data to undertake a detailed study using hospital ward locations and epidemiological knowledge of where outbreaks were thought to be occurring. In most cases, the observed clustering pattern of SARS-CoV-2 lineages supported the expectation of an outbreak resulting from a single transmission event, but in some cases the presence of several lineages have revealed multiple coincident smaller outbreaks rather than a single large one. Samples from healthcare workers have contributed around 40-45% of the viral sequences being looked at by this consortium member, and work is underway to determine how the relatedness between isolates from patients and healthcare works fits with the overall outbreak patterns observed.

To mitigate the highlighted difficulties in determining whether clusters of similar sequences really represent recent transmissions between staff and/or patients, the COG-HOCI study are developing a reporting tool that can be used to help infection control teams better identify where transmissions have occurred. This reporting tool will be assessed in the next few weeks within the COG-HOCI trial with the aim of making it widely available for general use for management of hospital onset SARS-CoV-2 infections.

Appendix

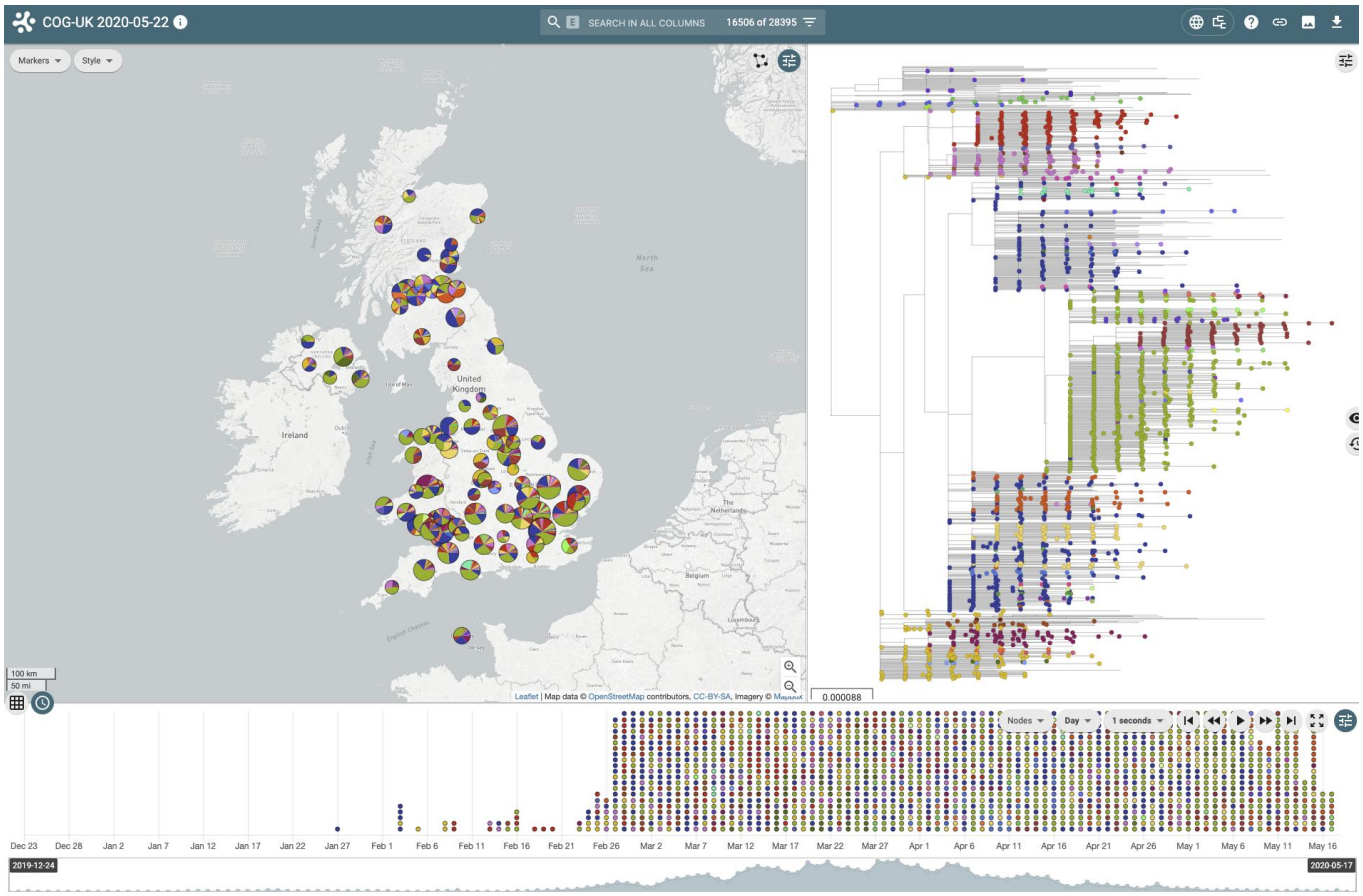

Figure S1 | Data linked and delivered through Microreact - Distribution of lineages are indicated by location and UK lineages are contextualised within the global phylogenetic tree. The lower timeline can be used to investigate spread and location of lineages over time. https://microreact.org/project/cogconsortium