COVID-19 Genomics UK (COG-UK) Consortium

Report #6 - 14th May 2020

Executive Summary

- All 15 COG-UK sites are now active and have sequenced and analysed 16,670 SARS-CoV genomes to date. Sequencing capacity now outstrips the availability of samples in some areas owing to the decline of the first wave of infections.
- A recent computational modelling study implicated the D614G variant in the SARS-CoV-2 spike protein in increasing affinity of the virus for its human cell receptor ACE2 and speculated about an impact on infection and transmission. A preliminary analyses of the D614G variant in currently available COG-UK data indicate that no significant association between the D614G variant in the spike protein and changes in SARS-CoV-2 transmission potential, or patient sex could be detected. However, further detailed analysis is required.
- Another SARS-CoV-2 genome variant (C26340T) has been linked to failure of the cobas® SARS-CoV-2 PCR diagnostic assay from Roche. Our analyses indicate that the variant is present at low frequency among UK and global lineages and that there is likely to be little impact on the use of this assay in diagnostic laboratories at present.

COG-UK update

An additional sequencing centre has been brought online over the past fortnight (University of Oxford) bringing the total number of active sequencing sites to 15.

By the data cut off for this report, the total number of virus genomes available is 16670 (Table 1), which continues to account for more than half of the SARS-CoV-2 genomes reported globally (Figure 1).

While the number of genomes being sequenced continues to grow, the number of positive samples received for sequencing each week is decreasing as the first wave of infections declines. As such the sequencing capacity available now exceeds current demand, but COG-UK is well-placed to cope as the challenge of ongoing infection continues.

Figure 1: Number of SARS-CoV-2 genomes sequences reported (in MRC CLIMB or GISAID). Data shown up to 13th May.
Highlighted findings with public health implications

- The D614G variant in the SARS-CoV-2 spike protein has been increasing in frequency in publicly available genome sequences and in a recent preprint viruses whose genomes carry this variant were suggested to be a ‘more transmissible form’. However, preliminary analyses of COG-UK data have found no significant association between the D614G variant and changes in SARS-CoV-2 transmission potential or patient sex. A significant association with patient age was observed, although the basis for this association is unclear at present, warranting further investigation.

- It was recently reported that a C26340T variant in the SARS-CoV-2 genome may be linked with failure of the cobas® SARS-CoV-2 diagnostic assay from Roche. Analysis of COG-UK and global data containing >23K genomes identified the variant in just 19 genomes (12 from the UK and 7 from Belgium, Switzerland, Turkey and Australia).

Table 1: Number of SARS-CoV-2 genomes sequenced and analysed by the COG-UK centres. *Some figures are lower than previous week owing to increased stringency in quality control parameters affecting inclusion of genomes.
Analysis updates

Relevance of the D614G spike protein mutation

Analysis leads
Tom Connor and colleagues, Public Health Wales and Cardiff University

Question addressed
One of the aims of COG-UK is to undertake surveillance and determine whether new mutations observed in the genomic dataset are relevant to any detectable changes in the behaviour of SARS-CoV-2. We have been monitoring the prevalence of a number of variants across the SARS-CoV-2 genome since early April. One such variant - an amino acid change from aspartate (D) to glycine (G) in the spike protein has recently been highlighted in a number of papers, which predict that this change may impact the transmissibility of the virus. This update describes an assessment of this mutation.

Methodology
Preliminary assessment of sampling proportions, phylogenetic distribution, and the relationship between D/G at position 614 of the spike protein and RT-PCR Ct value, patient age and patient sex on Welsh COVID-19 data. (See Appendix 2.1 for detail).

Finding Summary
There are a number of mutations across the SARS-CoV-2 genome that are seen at a high frequency around the world, however it is unclear whether these mutations are under positive selection (i.e. have become established as they improve some aspect of viral performance). An increase in mutation frequency through time is a necessary condition for it to be considered as being under positive selection, but is not sufficient. The nature of a virus pandemic (rapidly, growing, spatially expanding, hierarchically structured) means the simplest and most likely explanation for spatial and temporal changes in mutation frequency is expected to be random chance processes until proven otherwise. It is also important to note that changes that increase the predicted infectivity/transmissibility of the virus in cellular models or computer simulations may not translate into a meaningful advantage for the virus in the real world. Furthermore, the seeding of the virus in the UK, or other countries who are contributing the majority of sequenced samples, is unlikely to be from a uniform sample of the global virus population. Any bias in the number of introductions of one variant or the other can have a large effect on the observed frequencies of the variants in sequence databases which has no relation to the fitness of the viruses themselves.

The reference genome sequence has an aspartic acid (D) residue at position 614 of the spike protein, which is known to interact with the ACE2 host cell receptor. A change at this position, to a glycine (G) residue, has been increasing in frequency in publicly available SARS-CoV-2 genomes, prompting speculation and investigation about the potential relevance of this mutation. One recent preprint (https://www.biorxiv.org/content/10.1101/2020.04.29.069054v1) described computational modelling that indicated that this mutation might alter the affinity of the spike protein for ACE2, and made a number of predictions around what this might mean for infection and transmission.
A preliminary analysis by Tom Connor and colleagues at Public Health Wales of 1515 SARS-CoV-2 genomes sequenced (for which the assignment of D/G at position 614 was unambiguous) found no significant association for D614G with either cycle threshold value (Ct), a parameter that some groups have claimed is an indicator of viral load in a sample (and hence transmissibility), or sex of the individual from which the sample was taken. There was a significant difference relating to age of the individual from which the sample was taken, with the G variant associated with a lower average age (mean and median) compared to the D variant. However, the basis for this age difference is unclear and there are a number of potential variables (such as demographic/geographic-related variables, hospital outbreaks, founder effects, different responses to lockdown/control measures by age group) that could affect the observed distribution in ages. From this analysis, while there are unambiguously more cases that have the G variant, there is limited signal in the data to suggest that the G variant produces a meaningful/detectable increase in transmissibility, within the context of the current pandemic. See Appendix 2 for the full analysis.

**Key conclusions**

While no significant association between the D614G variant in the spike protein and changes in SARS-CoV-2 transmission potential or patient sex could be detected, these analyses do not unequivocally demonstrate that there is no effect relating to these factors, only that from the present data there is not a difference significant enough to raise concern. An association between D614G and age of the individual from which the sample was taken was observed, although the basis for this association is unclear at present, warranting further investigation.

**Proposed next steps**

In addition to this preliminary work, further detailed analysis is currently being undertaken to investigate any potential association between D614G SARS-CoV-2 transmission and clinical disease severity.

Irrespective, continued monitoring of D614G is prudent, in particular in genomes derived from community sampling efforts that may avoid potential sampling bias influencing analyses of age and clinical severity.

As part of our ongoing surveillance efforts we will continue to monitor other variants above a certain frequency, and investigate variants that have been shown elsewhere to be associated with more severe outcomes. Access to more detailed clinical and epidemiological data will enable more in-depth analyses to rule out confounding factors.
Screening of diagnostic primers

Analysis leads
Richard Myers, Eileen Gallagher, Natalie Groves, David Williams (Public Health England)

Question addressed
Is there any evidence of genomic changes potentially affecting common diagnostic tests (in particular the Roche cobas® SARS-CoV-2 PCR diagnostic assay)?

Methodology
Information gathering from UK diagnostic laboratories and sequence analysis of COG-UK and GISAID genome datasets for the variant reported as a possible target site for Roche cobas® SARS-CoV-2 PCR diagnostic assay.

Finding Summary
To better ensure the comprehensivity of ongoing monitoring of variation in SARS-CoV-2 genome regions targeted by primers and probes in diagnostic assays, PHE have been collecting information from UK diagnostic laboratories on the in-house and commercial tests being used. Responses are still awaited from some laboratories but the list of genome regions to monitor is now more thorough.

For one of these commercial tests (the cobas® SARS-CoV-2 PCR diagnostic assay from Roche, which is presently used within some UK labs), a recent publication linked failure of the assay to a C to T mutation at position 26340 (E gene) of the SARS-CoV-2 genome, albeit based on a limited set of data (four samples). Analysis of >23K genomes from the current COG-UK and GISAID combined datasets identified 19 genomes containing the C26340T variant, 12 of which were from the UK. The presence of the variant at a low frequency (0.08%) and distributed across a range of global and UK specific lineages, indicates that there is likely to be little impact on the use of the cobas® assay in a diagnostic setting. Identifying the prevalence of genomic variants is one component of assay monitoring and testing laboratories will follow standard practices to monitor the overall performance of testing procedures. Where issues are detected these will be reported and investigated using established channels.

Key conclusions
A more comprehensive view of the diagnostic tests at use in UK diagnostic assays will allow for more thorough monitoring of variation in the relevant regions of the SARS-CoV-2 genome.

Identification of specific lineages bearing the C26340T will allow their frequency to be monitored so that estimates of its impact on the reliability of the cobas® PCR assay can be revised over time.

Proposed next steps
The findings and conclusions outlined above will be integrated into ongoing monitoring efforts.
Current population structure of SARS-CoV-2 in UK

**Who did the analysis**
This is a brief update from the latest UK lineages summary report (8th May 2020) generated by Andrew Rambaut and colleagues, University of Edinburgh.

**Question being answered**
SARS-CoV-2 infections worldwide are classified into a number of “lineages” according to differences in the virus genome, likely the result of independent introductions into the UK followed by ongoing transmission. Although there is no evidence these lineages have different biological properties, they are useful in tracking the number and size of different chains of transmission.

**Finding Summary**
The analysis was undertaken on 14,277 SARS-CoV-2 genomes and identified 279 lineages (containing at least 5 genomes). Of these lineages 79 are pending extinction (i.e. have not been sampled by COG-UK for 3 weeks) and 12 are considered extinct (having not been seen for more than four weeks). An additional 135 lineages have not been sampled in the past week leaving 47 lineages known to have been circulating continuously in the UK.

Updated views of the geographic distribution and frequency of SARS-CoV-2 lineages in the UK dataset can be found in Appendix 1 (Figures S1 and S2).

In terms of the number of counties in which they are present, the ten most sampled lineages increased until late March and have been declining since early April (Figure 2), a pattern similar to the overall decrease in diversity of SARS-CoV lineages described in report #5.

![Figure 2](image-url): Number of administrative regions (counties) in which the ten most-sampled SARS-CoV-2 lineages were present between 22nd February and 22nd April 2020.
Appendix 1

Figure S1 | A. Latest geographic distribution of main SARS-CoV-2 lineages in the UK visualised using Microreact. B. Expanded view of distribution of lineages in Wales. Note that the large central pie chart displays samples for which no administrative region (county) was recorded.
Figure S2 | Recent view of data visualised using Microreact. Upper left panel displays geographic distribution of main SARS-CoV-2 lineages in the UK. Upper right panel displays lineage frequency over time. Lower panels display a timeline of the sampling for viral genomes (dots) and total sample numbers (grey graph). Live link to the view in the above screenshot: https://microreact.org/project/cogconsortium/
Appendix 2

SARS-CoV-2 Spike Protein Mutation D614G in Wales: Preliminary Findings

Sara Rey¹, Joel Southgate¹², Nicole Pacchiarini¹, Amy Gaskin¹, Matt Bull¹, Tom Connor¹²

¹: Bioinformatics Team, Pathogen Genomics, Public Health Wales Microbiology
²: Cardiff University School of Biosciences

Following reports of increased fitness and prevalence of the D614G mutant, we performed an assessment of sampling proportions, phylogenetic distribution, and the relationship between D/G at position 614 and RT-PCR Ct value, patient age and patient sex on Welsh COVID-19 data. We found, in summary:

- The reference SARS-CoV-2 sequence from Wuhan has an aspartic acid residue (D) at position 614 of the spike protein. It has been observed that a change at this position - to a Glycine (G) - has been increasing in frequency in public sequence databases. In Wales, this pattern is evident amongst sequenced samples, with the majority of samples sequenced now having a G at this position (Figure 1).

- No independent D614G mutations have been observed in the Welsh phylogenetic trees as yet, (Figure 2). If these had been found, they may have been signatures of positive selection or recombination.

- There has been a great deal of interpretation of the meaning of this mutation. Various approaches have been utilised to examine if there is theoretical evidence for the D614G variant to have an effect on viral transmission or outcomes. Wetlab work is also being undertaken by various groups to look at the relevance of the mutation in vitro. Many of the larger analyses have looked at data sources such as GISAID. These collections of data are unlikely to be representative of what is happening in the community, and may contain sampling or collection artefacts that are difficult to predict or mitigate.

- In Wales we have been sequencing every sample with a Ct less than 30, beginning with our first sample. We have also observed that increasing Ct correlates with increasingly poor sequencing quality. Concerned that the process of selection of samples for sequencing and quality steps put in place prior to upload to GISAID, we examined our complete sequenced dataset for which we had clear D/G calls (n= 1515) for any of the signatures recorded elsewhere.

- It has been postulated that the G mutant may have higher fitness/transmissibility, leading to a higher viral load, and therefore lower Ct values. We examined this in our data, but found no significant difference.

- A shift in the distribution of Ct values between the aspartate (D) and glycine (G) variants was possibly present, but the signal was weak if present at all. (Figure 3). The difference that was present did not meet the significance threshold (p = 0.08, Mann-Whitney U test; p = 0.35, two-sample Kolmogorov-Smirnov test). This does not support what has been found elsewhere (such as at https://github.com/blab/ncov-D614G), where the Ct associated with G was found to be significantly lower. Data from Wales, which is extensively sampled, suggests Ct differences should be interpreted with caution, with the understanding that confounding factors may be responsible for the differences observed. One caveat should be noted, that if there is a propensity for D variants to report higher Ct values, this may mean that D is under-represented in sequencing data (and so in any frequency estimates), and so more or less of a variance may exist in the complete dataset.
A shift in the distribution of age between D and G variants has been postulated to indicate differences in transmissibility, with increased frequency in lower age groups being held to be indicative of increased transmission potential (Figures 4-5). There was a statistical difference between the age distributions observed between the D and G variant in the Welsh dataset.

We examined our age distributions and there was a significant difference between them (p=0.006, Student’s t test; p = 0.006, Mann-Whitney U test; p = 0.006, two-sample Kolmogorov-Smirnov test). This is in agreement with other work (such as at https://github.com/blab/ncov-D614G), where a difference in age in patients infected with the G variant was found to be significantly lower than for those infected with the D variant. This is an interesting signature, but requires much more investigation. The change in frequency of the two types over time is important, as sampling strategy has been changed over time, and this and other variables (e.g. demographic/geographic aspects, hospital outbreaks, founder effects, different responses to lockdown/control measures by age group) could impact this distribution.

We also looked for association between sex and the presence of the D or G variant at position 614 (Figure 6). The hypothesis testing did not meet the significance threshold (p = 0.74, Fisher Exact Test; p=0.76, Chi Squared test).

The Ct value was plotted against sample collection date to allow for visual inspection of any patterns in the data which may be associated with changes in laboratory methodology occurring at a particular time (Figure 7). At present we use a mix of platforms (and tests), with the variation in test performance/reporting potentially being evident from the shapes of the histogram of sample frequencies. The variation in platforms should be instructive as to the danger of over-interpreting Ct values.

Occurrences of the D and G variant were plotted against local authority region (Figure 8). Plots of percentage of cases in each local authority which belong to the D and G clades (Figure 9). While providing no useful
conclusions, they demonstrate that clusters of cases including both D and G variants were/are circulating across Wales. Figure 10 also emphasises the fact that as one moves further into Wales from England the frequency of G relative to D increases, which is a temporal artefact, as local case numbers in the east of Wales rose before those in the west. This may imply something about possible founder effects or importation of cases, however this would require further investigation and analysis.

Conclusion

Examination of a cleaned, high quality sequence dataset generated in Wales, encompassing over 10% of all confirmed Welsh cases to date showed no significant difference in the distribution of D614G variants with respect to Ct or sex. There was a significant difference in the distribution of D614G variants with respect to the age of patients. The basis for this difference is unknown, but there are a number of potential variables that could affect the observed distribution of ages, which makes any larger inference of the meaning for this variation unsafe. This difference could be explained by changes in transmission, or other factors including demography, geography, founder effects, changes in testing strategy and lockdown compliance (which may vary by age group). While there are now unambiguously more cases that have the G variant, we therefore find limited signal in our data to clearly suggest that the G variant produces a meaningful/detectable increase in transmissibility, within the context of the current pandemic. Ranges of Ct values at a given time point are comparable between the two variants, with the main changes in distribution of Cts most likely being related to changes in the platforms used for testing.

Differences in the age distribution are interesting, but additional analyses would be required to understand the basis of that difference. Further work could be performed by examining the trajectories of imported lineages of each type, and their dynamics. However, such an analysis would require unbiased samples and robust lineage designations. Further analyses including comorbidity information could provide more information, while analysis of eHealth records might provide more of an indication of any variables that correlate with variants in the population. This type of work is currently being initiated in Wales at present.

One issue remains, whether Ct relates to transmission or not, which is that if one is interested in the frequency of variants across all confirmed cases, then there is likely to be a natural bias with samples that do not sequence well. This could relate to biological factors (including low viral load) or laboratory/human ones (sample site, skill of individual taking sample, degradation of sample, mutations in amplification primer sites). Generating estimates that relate to frequency of a given variant in the population using a biased sample is likely to be problematic, and without extensive community screening over time, it is difficult to identify extant datasets that might enable this type of investigation.

Lastly, it should also be noted that questions of transmission have a large number of confounders, and that most of the sequencing performed by centres in the UK is focused on confirmed cases, often using diagnostic residual samples. As testing priorities change, and as cases in different segments of the population fluctuate, signals may emerge that are due to policy rather than biology. The only way to properly investigate questions of transmission or similar questions is through well constructed prospective studies. Our results, informed by extensive metadata and based on more than just the higher quality sequences submitted to GISAID, point towards a lack of signal which would call into question some of the claims made from GISAID data around this mutation.
Figure 1. Histograms showing sampling count over time. The proportion of the G variant increased to surpass D at the end of March/beginning of April.

Figure 2. Maximum likelihood phylogeny (HKY+G, rooted by early Wuhan sequence MN908947.3). No independent D614G mutations were observed, nor were any back-mutations.
Figure 3. Frequency histograms for D/G variants for Ct with KDE.

Samples where the amino acid at position 614 was not recorded and a sample with a Ct value of 68 were excluded.

Figure 4. Box plot for D/G variants for Ct.

Samples (n=1515) where the amino acid at position 614 was not recorded and a sample with a Ct value of 68 were excluded.
Figure 5. Frequency histograms for D/G variants for age.

Samples where either the amino acid at position 614 or the age was not recorded were excluded.
Figure 6. Plot of counts for D/G variants by sex.

Samples where either the amino acid at position 614 or the sex was not recorded were excluded.
Figure 7. Box plot for D/G variants for Ct by date

Samples where no sample collection date was recorded or the amino acid at position 614 was not recorded and a sample with a Ct value of 68 were excluded.
Figure 8. Counts for D/G variants for Local Authority region.

Samples where either the amino acid at position 614 or the local authority was not recorded were excluded.

Figure 9. Percentage of cases of D/G variants for Local Authority region.

Samples where either the amino acid at position 614 or the local authority was not recorded were excluded.
Figure 10. Pie charts showing the proportion of D/G variants overlaid on the relevant Local Authority region.

Samples where either the amino acid at position 614 or the local authority was not recorded were excluded.
Appendix 3

Primer Summary

Introduction

We have been collecting information on in-house and commercial tests being used in UK diagnostic laboratories. This information is not yet complete. We have also looked into the number of sequences within COG-UK that have a variant in the E gene which may result in false negatives for the E gene portion of the Roche cobas SARS-CoV-2 test.

Variant in the SARS-CoV-2 Genome Reported as Impacting on the cobas Diagnostic Assay

A recent publication (1) has linked failure of the cobas assay to a C→T mutation at position 26340 (E gene) of the SARS-CoV-2 genome. This variant is postulated to affect one of the two targets that this assay uses as part of the diagnostic assay. Results in four samples containing this variant were negative for the E gene diagnostic PCR, but were positive in the Orf1ab assay. It is therefore important to note that even though the presence of the 26340T variant may affect the cobas assay, it is predicted to lead to a positive/negative result rather than a genuine false-negative result.

It should also be noted that the conclusion that the C→T variant at 26340 impacts on the E gene component of the cobas assay is based on a very limited set of data (four samples). This finding should be validated by a larger study before this observation can be treated with any certainty. The variant results in a synonymous change codon GCC -> GCT, Alanine at reference envelope position 32.

The current COG and GISAID combined dataset contains 23,079 SARS-CoV-2 genomes (18,147 distinct sequences) of which 19 genomes were identified as containing the C26340T variant. Twelve of the genomes containing the E gene variant were from samples sequenced within the UK: seven from UK lineage 53 (total size 19) and one from UK lineage 1661 (total size one). A further four sequences from England do not have a UK lineage assigned. One falls in global lineage B.1.X, and the other three fall in global lineage B.2.1 with the other seven sequences in UK lineage 53. The remaining seven sequences containing this variant were from Belgium, Switzerland, Turkey and Australia. The 19 sequences containing the variant at 26340 were distributed across a range of B lineages (B = 2, B.1.5 = 2, B.2.1 = 10, B.1.X = 1, B.3 = 3, B.4 =1). The presence of the variant at a low frequency and distributed across a range of global and UK specific lineages, indicated that there was likely to be little impact on the use of the cobas assay in a diagnostic setting. The frequency and lineage of this variant will be monitored so that estimates of its impact can be revised over time. Identifying the prevalence of genomic variants is one component of assay monitoring and laboratories will also follow standard practices to monitor the overall performance of assays. Where issues are detected these will be reported and investigated using established channels.

1. Failure of the cobas® SARS-CoV-2 (Roche) E-gene assay is associated with a C-to-T transition at position 26340 of the SARS-CoV-2 genome: Maria Artesi, Keith Durkin et al (medRxiv preprint doi: https://doi.org/10.1101/2020.04.28.20083337)