THE ROYAL SOCIETY

26 AUGUST 2020

The SARS-CoV-2 genome: variation, implication and application

This rapid review describes the severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2) genome, its relationship to other coronaviruses, the variation that has occurred since SARS-CoV-2 emerged in Wuhan in late 2019, the implications of these changes and how knowledge of these changes may be utilised.

This pre-print from the Royal Society is provided to assist in the understanding of COVID-19.

Executive summary

- SARS-CoV-2 emerged in late 2019 in Wuhan, China. The genome of many separate virus isolates from early in the Wuhan outbreak are very closely related showing the virus emerged recently in humans.
- The SARS-CoV-2 genome is sufficiently different to all known coronaviruses to refute the assertion that the COVID-19 pandemic arose by deliberate or accidental release of a known virus and make it highly improbable that the virus arose by artificial construction in a laboratory.
- SARS-CoV-2 is most closely related to bat coronaviruses from China, but even the closest of these viruses are too divergent (~97% nucleotide identity across the whole genome) to be the immediate ancestors of SARS-CoV-2. The origin of SARS-CoV-2 is likely directly from bats or via an unknown intermediate mammalian host.
- Genome change in SARS-CoV-2 is slow compared to most RNA viruses but, nonetheless, mutations arise that can be used to trace virus spread and evolution.
- The most variable virus gene encodes the spike (S) protein that mediates virus attachment to and entry into cells, and is the target of neutralising antibodies and a robust T cell response. A D614G mutation in the S1 subunit of the S protein possibly enhances virus transmission and is now dominant in virus strains circulating globally. No genome changes have been identified that are shown to affect virulence.
- Whole genome sequencing is a valuable addition to test, track and trace, and is encouraged. For instance it:
  - has revealed >1350 separate introductions of SARS-CoV-2 into the UK from mid-February to mid-March 2020 arising very largely from Spain, France and Italy, and not from China.
  - can be used to follow transmission within specific communities such as hospitals, schools or factories, and combined with epidemiological data enables routes of transmission to be identified and barriers to transmission implemented.
- Commercial manufacturers of PCR-based diagnostic tests for SARS-CoV-2 should continually consult the open databases of SARS-CoV-2 genome sequences to ensure their tests remain up-to-date and that false negatives do not arise because of genome variation.
- The genome variation seen hitherto is unlikely to enable virus escape from immune responses induced by vaccination or prior infection.

1. The SARS-CoV-2 genome

Seven different types of coronavirus have infected humans. Three of these, severe acute respiratory syndrome coronavirus (SARS-CoV), Middle East respiratory syndrome coronavirus (MERS-CoV) and SARS-CoV-2, have caused serious illness and many deaths, whereas four other human coronavirus (HuCoV) 229E, HKU1, NL63 and OC43, usually are associated with mild common cold-like illness. SARS-CoV (genome size 29.7 kilobases [kb], emerged in 2002), MERS-CoV (genome size 30.1 kb, emerged in 2012) and SARS-CoV-2 (genome size 29.9 kb, emerged in 2019) are all zoonoses introduced into humans from animal reservoirs.

Coronaviruses have a single stranded, positive sense, RNA (ssRNA+) genome with a conserved arrangement of genes. Starting from the 5'-end, most of the genome (~21 kb) codes for the non-structural proteins (nsps) 1a and 1b, and these large polyproteins are cleaved by proteases into several mature proteins. The next gene (~3.8 kb) encodes the spike (S) protein. The other structural proteins (i.e. forming part of the virus particle) envelope (E), membrane (M) and nucleocapsid (N) proteins are encoded towards the 3' end of the genome. Additional accessory proteins are also encoded in this region and generally show more divergence between different coronaviruses.

Coronaviruses have the largest known genomes of animal RNA viruses, ranging from 26-32 kilobases (kb)[1], and this large genome size is made possible by these viruses having a mechanism to identify and correct errors that are introduced during genome replication[2 3]. Without such "proof-reading" the high error rate of RNA-directed RNA polymerases (RdRP) causes too many deleterious mutations to be introduced during each replication cycle and these render the genome non-viable. Nonetheless, despite this proof-reading mechanism, mutations do arise and these may be used to track virus spread and evolution (see below).

SARS-CoV-2 is a member of the genus *Betacoronavirus*, subgenus *Sarbecovirus* of the family *Coronaviridae*. The genomes of multiple SARS-CoV-2 isolates derived from the first COVID-19 patients in Wuhan, China shared 99.98-99.99% nucleotide identity[4] suggesting that the virus had only emerged recently in humans. SARS-CoV-2 is most closely related to coronaviruses isolated from bats within China called SARS-related coronaviruses (SARSr-CoVs). The closest relatives are Bat RaTG13-CoV and Bat RmYN02-CoV, both of which were sampled from horseshoe bats in Yunnan province, China. RaTG13-CoV shares 96.3% nucleotide identity with SARS-CoV-2 across the genome as a whole, while RmYN02-CoV shares ~97% nucleotide identity with SARS-CoV-2 in the long 1ab open reading frame (ORF1ab), although is more divergent in other genes due to widespread recombination[5 6 7 8]. For comparison, SARS-CoV-2 shares ~79% nucleotide identity with SARS-CoV, which is also a member of the *Sarbecovirus* subgenus. However, MERS-CoV, a member of the *Merbecovirus* subgenus, is more divergent with only ~50% nucleotide identity with SARS-CoV-2 [9].

Comparison of SARS-CoV-2 with other coronaviruses showed that the greatest divergence lies within the gene encoding the S protein. This protein forms homotrimers on the surface of the virus particle and mediates binding of virus to target cells and fusion of virus and cell membranes during virus entry. This protein is also the target of antibodies that neutralise virus infectivity and thereby prevent infection and it also elicits a robust T cell response[10]. The S protein is cleaved via cellular proteases into S1 and S2 subunits, which together represent one monomer of the trimer. Cleavage is essential for the S protein to mediate virus entry. Subunit S1 is involved in receptor binding via a defined receptor-binding domain (RBD) and S2 is needed for membrane fusion.

The S1 subunit of SARS-CoV-2 is more divergent than S2 compared to other betacoronaviruses. Nonetheless, the S protein from SARS-CoV and SARS-CoV-2 bind to the same cell receptor, the angiotensin converting enzyme II (ACE-2)[11 12], due to conservation of critical residues within the RBD of the S1 subunit[13 14 15]. Notably, the RBD of S1 from RaTG13-CoV (and RmYN02) and SARS-CoV-2 differ considerably indicating that the SARS-CoV-2 would not have arisen directly from either RaTG13-CoV or RmYN02.

Another important difference between the S protein of SARS-CoV-2 and closely related betacoronaviruses is the insertion in the former of 4 amino acids, PRRA, that creates a polybasic furin cleavage site between the S1 and S2 subunits. Cleavage at this site enables increased exposure of the RBD and, thereby, a high affinity interaction with the human ACE-2 receptor[16]. A furin cleavage site is present in S proteins in MERS-CoV, HuCoV-OC43 and HuCoV-HKU1, but is absent from SARS-CoV, HuCoV-NL63, HuCoV-229E and the SARSr-CoVs[17]. Although RmYN02-CoV also has experienced sequence insertion-deletion events at the S1/S2 cleavage site, these are not polybasic[18].

Other coronaviruses that are closely related to SARS-CoV-2 were isolated from Malayan pangolins smuggled into Guangdong and Guangxi provinces, China. In particular, the virus from the Guangdong pangolins shares 91% nucleotide identity with SARS-CoV-2 and 90.5% identity with RatG13-CoV[19]. Although, over the whole genome, this Pangolin-CoV is less closely related to SARS-CoV-2 than RaTG13-CoV or RmYN02-CoV are, the amino acid sequence of its RBD is much closer to the SARS-CoV-2 RBD than that of any other virus. In particular, within the Guangdong Pangolin-CoV RBD, 5 amino acid residues that are critical for binding to ACE-2 are conserved with SARS-CoV-2, although these are divergent in the RaTG13-CoV and RmYN02 S proteins[20].

Possible origins of SARS-CoV-2

The degree of divergence between SARS-CoV-2 and all other known coronaviruses is sufficient to refute the assertion that the COVID-19 pandemic arose by the deliberate or accidental release of a known virus (e.g. RaTG13) and makes the unsupported claim that SARS-CoV-2 was created artificially in a laboratory highly improbable.

The distinctive and extensive nucleotide sequence differences between the bat viruses RaTG13-CoV and RmYN02 indicate that neither were the immediate ancestor of SARS-CoV-2: rather, these are the most closely related animal viruses sampled to date. This conclusion is also supported by specific changes within the S protein: these are the substantially different RBD, and the insertion into the SARS-CoV-2 S gene of 12 nucleotides that encode a polybasic furin cleavage site at the S1/S2 junction, which are absent in both RaTG13-CoV and RmYN02-CoV[21].

The most likely origins of SARS-CoV-2 are from a bat virus that is more closely (>99% nucleotide sequence identity) related to SARS-CoV-2 than either RaTG13-CoV or RmYN02-CoV, or from a bat virus that was transmitted to humans from an "intermediate" mammalian host species, possibly following evolution via recombination with other coronaviruses[22 23]. It is established firmly that co-infection of the same cell by closely related coronaviruses enables recombination between the virus genomes such that sections of one genome may be replaced by the corresponding region of another, and that recombination has contributed substantially to coronavirus evolution[24 25 26 27]. It is clear that sarbecoviruses have a complex history of recombination, although untangling the exact history of these events is challenging[28]. In addition, the presence of viruses related to SARS-CoV-2 in Malayan pangolins suggests that ongoing surveillance will identify additional coronaviruses in other mammalian species, some of which may fall on the evolutionary pathway to SARS-CoV-2.

How did the polybasic furin cleavage site of SARS-CoV-2 arise? Comparison of the sequence of the 12 inserted nucleotides that encode the PRRA cleavage site with other CoVs identified a very similar sequence (10/12 nucleotides conserved) in the S gene of Bat HKU9-CoV that was isolated from a *Rousettus* fruit bat in Guangdong province, China in 2011. Other similarities upstream and downstream of this sequence show 14/19 identical nucleotides[29]. This substantial similarity may have enabled the replication-transcription complex to switch from one RNA genome template to another and result in insertion of the sequence coding for the polybasic cleavage site into the SARS-CoV-2 genome.

It is important to continue to generate and analyse sequence data from the genomes of human and animal coronaviruses to give additional insight into the origin of SARS-CoV-2.

2. Mutation rate in SARS-CoV-2 genome replication

When comparing mutation rates, a distinction is made between rates calculated as the number of substitutions per nucleotide per cell infection (s/n/c) or substitutions per nucleotide per round of copying (s/n/r). The distinction reflects whether virus genomes replicate via a "stamping machine" model, in which a single template is copied repeatedly, or if replication is semiconservative, in which replicated strands act as templates for additional synthesis. Using the former method, estimated error rates for DNA viruses are 10^{-8} to 10^{-6} s/n/c, whereas RNA viruses range from 10^{-6} to 10^{-4} s/n/c [30 31]. As examples, hepatitis C virus has a mutation rate of 1.2×10^{-4} s/n/c, influenza A virus has a mutation rate of 1.2×10^{-5} and the coronavirus mouse hepatitis virus (MHV) has a mutation rate of 2.5×10^{-6} [32]. The mutation rate for SARS-CoV-2, it is expected to be similar to MHV and hence lower than seen in other RNA viruses.

The rate of virus evolution is expressed as the number of nucleotide substitutions per site per year (s/s/y) and when applied to coronaviruses this gives estimates of $1.5 - 10 \times 10^{-4}$ s/s/y [33 34]. The s/s/y value is, however, a measure of the rate at which mutations accumulate following the action of natural selection and is influenced greatly by different viral life cycles, with viruses that establish latent infections having lower s/s/y values despite having nucleic acid polymerases with comparable fidelity (see below). For this reason, comparisons of mutation rates between different viruses often utilise the s/n/c value.

RNA viruses generally have higher mutation rates than DNA viruses because they lack a proof-reading activity associated with their RNA-directed RNA polymerases (RdRp) and so have lower fidelity. The few RNA viruses with genomes greater than 20 kb have, however, all acquired a proof-reading activity that correlates with the expression of a viral exonuclease (ExoN)[35 36]. In coronaviruses the proof-reading activity is mediated by a 3' to 5' exoribonuclease (ExoN). In SARS-CoV the exonuclease is nsp14 that is found complexed with another virus protein nsp10. Collectively, this complex can detect and excise 3' nucleotide mismatches. Mutation of nsp14/ExoN in MHV[37] or SARS-CoV[38] to remove exonuclease activity caused a 15- or 21-fold decrease in replication fidelity, respectively. The highly conserved nature of nsp14 and nsp10 in SARS-CoV-2 suggest very similar functions.

Coronaviruses engineered to lack the ExoN activity have reduced virulence and have been proposed as live attenuated vaccines, so long as the ExoN protein is changed sufficiently to minimise the risk of reversion to wild type.

3. SARS-CoV-2 genome variation

Since the emergence of SARS-CoV-2 in Wuhan late in 2019, the virus has spread globally and the complete genome sequence of tens of thousands of virus isolates have been determined and deposited in public databases such as the Global Initiative for Sharing All Influenza Data (GISAID) database (https://www.gisaid.org). These data form the basis of phylogenetic analyses of SARS-CoV-2 evolution. Such studies have revealed that as the virus replicated in human cells and transmitted between humans, mutations arose that became fixed into distinct virus lineages and accumulated progressively. Some mutations appear multiple times independently suggesting introduction via a common mechanism, or an advantage for viral growth or transmission. Analyses of the distribution of sequence sampling dates compared to the earliest samples from Wuhan has enabled estimation of the time to the most recent common ancestor (tMRCA) of SARS-CoV-2 in humans, which equates to the start of the COVID-19 epidemic, between 6.10.2019 and 11.12.2019[39]. Although current sequence diversity in SARS-CoV-2 is limited, it has enabled phylogenetic relationships and aspects of viral epidemiology to be studied.

Types of genome variation

Mutations are considered i) globally across the whole genome, including the type and distribution of changes, and ii) within the S gene since the S protein is critical for binding to host cells and is the target for neutralising antibody and T cell responses. Mutations may be single nucleotide polymorphisms (SNPs), which may or may not cause an amino acid substitution (non-synonymous versus synonymous changes), insertions or deletions (indels).

Genome wide changes

i) Single nucleotide polymorphisms

Genetic diversity is accumulating slowly in the SARS-CoV genome. A study of 7666 SARS-CoV-2 genomes that provided good temporal and geographical coverage of the early pandemic up to April 19th 2020 showed that only moderate genetic diversity had arisen and there was an average pairwise difference of 9.6 SNPs between any 2 genomes[40]. A more recent survey (Edward C. Holmes, personal communication) showed that the mean genetic distance (i.e. divergence) from the most recently sampled viruses (10th July 2020) to the earliest viruses from Wuhan (December 2019) is 8 mutations across the viral genome (0.027%) (95% quantile, 1 to 15 mutations; 0.003-0.050%). Similarly, the mean pairwise distance (i.e. diversity) among the currently circulating viruses is 10 (0.033%) mutations (95% quantile, 3 to 22 mutations; 0.010-0.074%). These genetic distances equate to a mean rate of evolutionary change (i.e. the rate at which mutations are fixed in the population) of ~1 x 10^{-3} s/s/y, and hence close to the mean rate observed in many RNA viruses[41]. However, because of the short timescale of sampling, this rate is likely elevated by the presence of transient deleterious mutations and may decline with time. This rate is also likely to have been elevated by C→U hypermutation, see below[42].

As SARS-CoV-2 spread globally it has generated many phylogenetic lineages of differing size, some of which have spread to multiple countries and some of which have already gone extinct[43]. Major virus lineages are generally comprised of viruses isolated from the same continent. However, because the epidemics in most countries were seeded by multiple (and often many) separate introductions of virus, the geographical structure in SARS-CoV genomes is limited. The small lineages unique to some countries are not known to hold any major biological or epidemiological significance. Arguably, the most notable lineage is that designated 'B1' that arose in Italy, spread globally, and contains a high frequency mutation at residue 614 in the S protein (see below).

Genetic diversity is distributed across the SARS-CoV-2 genome, although there may be more diversity within the 3' region, perhaps due to a greater number of accessory genes. A few mutational hotspots were identified in which the same mutation had arisen repeatedly and independently. These include loci within genes encoding nsp6, nsp11, nsp13 and the S protein[44]. At present, there is little evidence for positive selection acting on SARS-CoV-2, including that mediated by the host immune response. However, although most amino acid sites in the SARS-CoV-2 genome are subject to purifying (negative) selection that acts to remove mutations, there are a number of putative positively selected sites that should be monitored. Those with the strongest signal for positive selection are: nsp12-323 (RdRp), accessory protein 3a-57, nsp3-1103, nsp5-108 (3C-pro), nsp13-290 (helicase), S protein-5, S protein-769, and accessory protein 3a-110. Although these all fall in potential epitopes that might be recognised by cytotoxic T lymphocytes, the functional significance of mutations at these sites is unknown. A detailed and ongoing analysis of mutations of interest, including positively selected sites, can be found at: https://observablehq.com/@spond/natural-selection-analysis-of-sars-cov-2-covid-19.

ii) Recombination

Genetic variation in coronaviruses may also be generated by recombination, although this process has been of limited impact for SARS-CoV-2 after its emergence late in 2019 and is hard to detect, at least between different SARS-CoV-2 strains, because these viral genomes are so similar in sequence.

iii) C→U hypermutation

A study of SNPs within ~1000 SARS-CoV-2 genomes isolated up to April 24, 2020, noted that the ratio of non-synonymous to synonymous substitutions per site (dN/dS) was higher (0.57-0.73) than observed in other human coronaviruses (<0.22)[45]. Additionally, almost half of the SNPs were cytidine to uridine (C→U) transition mutations and these were 8-fold more common than the converse U→C. This was the more notable given that U (32.1%) is almost 2-fold more abundant than C (18.4%) in the SARS-CoV-2 genome[46]. C→U transitions are scattered throughout the genome and accounted for almost half the amino acid substitutions observed. The mechanism driving C→U hypermutation is therefore also driving most of the amino acid changes observed, at least in the first 4-5 months of the epidemic. C→U transitions were also dependent upon sequence context, with an upstream or downstream A or U favouring transition compared to C or G at these positions.

These features are reminiscent of the interferon-induced pathway that edits retroviral single stranded DNA (converting C to T by deamination) after its synthesis from the virus ssRNA+ genome by reverse transcription. Cytidine deamination is driven by proteins of the apolipoprotein B mRNA-editing enzyme, catalytic polypeptide-like (APOBEC) family. Several members of this family have antiviral activity against retroviruses. The editing properties of some APOBEC members are also sequence context dependent, as observed in the SARS-CoV-2 genome.

These observations show that a large proportion of sequence change in SARS-CoV-2 are C→U mutations reminiscent of host APOBEC-induced changes and provide evidence for a host-driven antiviral editing mechanism against coronaviruses.

iv) Insertion-deletion mutations (indels)

SARS-CoV-2 replication in humans has generated indels naturally, and these are relatively commonplace among the genomes sequenced to date. However, few have achieved high frequencies in the population and hence are likely of little epidemiological significance. A single codon deletion in ORF1ab is perhaps the most widespread, likely arising in Asia before spreading to Europe, although its fitness effects are uncertain. A large (382 nt) deletion mutation that removed most of ORF8 occurred in viruses sampled from Singapore, although it has not spread further[47]. Finally, although the polybasic (furin) cleavage site at the S1/S2 junction in the S protein has been deleted in several cases[48], these mutations are associated with replication of the virus in cell culture rather than in humans.

Are there genome changes that correlate with altered virulence?

Virulence is a measure of the ability of a specific pathogen to cause disease in a particular host and is determined by the genome of the pathogen. The virulence of a pathogen applies only to a specific pathogen-host combination and a pathogen that is virulent (can cause disease) in one host species may be avirulent (unable to cause disease) in a different host. Often virulence is confused with other terms such as infectiousness or transmissibility. A pathogen can be highly infectious and transmitted easily without being virulent. Conversely, a pathogen might be highly virulent without being very infectious. The outcome of infection by a given pathogen is also influenced by the physiology of the particular host, and in humans several factors enhance the severity of disease caused by SARS-CoV-2, including sex (males are more susceptible), increasing age, obesity, diabetes, hypertension and chronic respiratory disease.

There has been much interest in identifying changes in SARS-CoV-2 that affect virulence. Several SARS-CoV proteins (which are conserved in SARS-CoV-2) have been shown to, or are likely to, contribute to virulence. These include proteins that function inside infected cells to shut down the host innate immune response to infection by, for instance, blocking pathways that lead to the production or action of interferons.

Ultimately, the contribution of a specific mutation to virulence requires observation of the outcome of virus infection in a susceptible host. *In silico* comparisons of genetic changes just generate hypotheses that need testing experimentally.

To date there are no genetic changes proven to alter SARS-CoV-2 virulence. In an attempt to identify virulence factors from human coronaviruses, as measured by a relatively high case-fatality rate (CFR), the genomes of viruses associated with high and low CFRs were compared[49]. Eleven nucleotide sequences that correlated with high CFR were found and mapped to 6 proteins: nsp3, nsp4, nsp14, S protein, membrane glycoprotein (M) and nucleocapsid protein (N). Although, overall, these correlations were weak, the linkage was strongest to S and N genes and corresponded to insertions or deletions in the encoded proteins.

For N the presence of strong nuclear localisation signals (NLS) and nuclear export signals (NES) correlated with a higher CFR[50]. If and how these increasingly strong nuclear import and export signals affects virus virulence is unknown, but it is possible that the increasing positive charge these changes bring affects the localisation of N in the nucleus /nucleoli and/or the interactions of N with other virus molecules such as M and the RNA genome.

Changes in spike protein

An amino acid change that has become dominant globally in most current SARS-CoV-2 strains is the substitution of aspartate 614 (D614) by glycine (G614) near the C terminus of the S1 subunit. Residue 614 is located at the interface of the S1 and S2 subunits and may affect the strength of their interaction. The G614 mutation was absent in most SARS-CoV-2 strains isolated in February 2020, but its prevalence grew rapidly to 26% in March, 70% in May and it now dominant globally[51]. The presence of this mutation has been associated with increased virus loads in COVID-19 patients[52] and pseudotyped lentivirus particles containing an S protein with G614 transduced cells with greater efficiency[53] and grew to higher titres[54] than particles with D614. Convalescent sera from COVID-19 patients bound both S variants to a similar degree[55].

Although the D614G mutation correlates with altered properties of the S protein when studied in isolation from other SARS-CoV-2 proteins, such as within pseudotyped lentiviral particles, the D614G substitution is nearly always accompanied by three other mutations elsewhere in the SARS-CoV-2 genome. These are a C→U mutation at position 241 in the 5' untranslated region, a synonymous C→U mutation at position 3,037 (nsp4), and a C→U mutation at position 14,408 that causes a P323L mutation in the nsp12 (RdRp)[56]. The significance of these individual changes are unknown.

Mutations have also arisen in the RBD of the S protein, including an N439K mutation that is common in Scotland (>400 sequences), and a T478I mutation that is found in England (~100 sequences). The functional significance of these mutations requires experimental investigation.

When comparing the S protein from different coronaviruses two insertions are observed in viruses with higher CFRs (MERS-CoV, SARS-CoV and SARS-CoV-2). The first is within S1 adjacent to the RBD. The second is located in S2 downstream of the hydrophobic fusion peptide, preceding the first of 2 heptad repeats[57]. It has been suggested that the greater length of the loop connecting these 2 domains may enhance flexibility and this might somehow affect membrane fusion efficiency during virus entry.

4. Epidemiology

Whole genome sequencing has greatly aided our understanding of the evolution, epidemiology and spread of SARS-CoV-2. For example, the analysis of thousands of SARS-CoV-2 sequences from UK in the second half of February and March 2020 revealed that there had been separate introductions of at least 1356 genetically distinct SARS-CoV-2 strains into UK. Notably, the great majority of the UK strains originated in three European nations - Spain (34%), France (29%) and Italy (14%) - via incoming international travel. In contrast, less than 1% derived from China[58].

Longitudinal studies can show how widely individual lineages spread and for how long they endure. Some lineages disappeared quickly and may have been eliminated by effective public health control measures such as quarantine of patients and contact tracing, whereas other lineages have endured and spread more widely[59]. Similar studies have been conducted in many other nations, such as The Netherlands[60] and Brazil[61].

Whole genome sequencing of viruses within local settings, such as health care workers in hospitals, can be used to follow the introduction of virus lineages into and spread within these settings[62 63]. One study undertaken between March 13 and April 24 in Addenbrooke's Hospital, Cambridge determined the full genome sequence of SARS-CoV-2 from 299 patients and combined with clinical and epidemiological data identified 35 clusters of identical viruses from 159 patients. The majority of these cases had strong (92 = 58%) or plausible (32 = 20%) epidemiological links and these observations led to the implementation of improved infection control[64 65]. Combined with increased testing of health care workers, identification of asymptomatic infected persons and falling community incidence, these measures led to a dramatic reduction in incidence of infection in this hospital setting between April and May 2020[66 67].

It is probable that in other settings, such as schools or factories, the use of whole genome sequencing, combined with epidemiological data, should be equally effective at illustrating how virus spread occurs and this knowledge will enable the implementation of measures to control the spread of infection.

The beneficial impact of whole genome sequencing is dependent on rapid sequencing after sample isolation.

Sequencing success is influenced by the quality of nucleic acid extracted from test samples. Although providing valuable automation, the introduction into testing centres of integrated diagnostic platforms that extract nucleic acid from samples and undertake reverse transcription quantitative polymerase chain reaction (RT-qPCR), such as the Hologic Panther Fusion Platform and similar, has seen the success rate of sequencing drop below 50%. These platforms do not leave residual nuclei acid, so re-extraction of the sample from a proprietary lysis buffer is required and this likely influences the success rate. Prior to this, nucleic acid extracted by the test laboratory gave a sequencing success rate of >90% for samples with a ct <33. Consequently, currently useful genomics data are being missed.

Whole genome sequencing is a valuable additional control measure; and test, track and trace should become test, sequence, track and trace.

5. Implications of genome change for efficacy of SARS-CoV-2 testing

PCR-based tests

An important question is whether the gradual accumulation of mutations within SARS-CoV-2 genomes might render polymerase chain reaction (PCR)-based detection tests ineffective and lead to false negatives. These tests utilise pairs of short oligonucleotide primers that enable amplification and detection of the intervening genomic region. They are virus-specific and detect the presence of the SARS-CoV-2 genome and by inference the presence of virus. These tests are orders of magnitude more sensitive than other tests to detect the presence of virus proteins, either within infected cells or in virus particles. They do not measure whether someone has been infected previously, and only inform whether, at the time of testing, the virus genome is present.

Theoretically, a mutation in the virus genome at the site to which either oligonucleotide primer should bind, may destabilise the interaction such that, despite the presence of virus genome, a negative result ensues. To mitigate against this possibility, primers are chosen to highly conserved regions of the genomes. In addition, tests can rely on more than one pair of primers that bind to different conserved regions of the genome, so that any single mutation is unable to prevent a positive test result.

Currently, the low level of changes in the SARS-CoV-2 genome make such false negative results improbable. In addition, the availability of new SARS-CoV-2 genome sequences as they are deposited openly in GISAID, enables the commercial manufacturers of test kits to monitor whether the primers chosen for current tests are likely to remain sensitive and specific, or require updating. Therefore, companies will need to continue to monitor SARS-CoV-2 genome diversity and maintain primers that detect the expanding range of genome sequences.

Detection of virus protein and antibodies to SARS-CoV-2 antigens

Virus infection can also be detected using monoclonal antibodies (mAbs) specific to epitopes in SARS-CoV-2 proteins. The protein targets selected should be abundant and conserved. Although these tests are less sensitive than PCR-based tests, they have advantages of simplicity and speed and do not require an equipped laboratory setting. As such, these “point-of-care” tests have utility in field conditions or in developing nations where access to equipped laboratories is more limited.

An epitope recognised by a mAb might be changed by mutation, such that the mAb no longer detects the changed antigen and a false negative may ensue. As with PCR-based tests, the limited genome variation of SARS-CoV-2 makes this less likely than for more highly mutable viruses. Also, tests can be designed to minimise this eventuality by selecting mAbs that recognise epitopes from conserved proteins, such as abundant internal antigens rather than the S protein, and by using combinations of mAbs.

The detection of virus antibodies to determine if someone has been infected previously with SARS-CoV-2 relies upon provision of SARS-CoV-2 antigen(s). These are produced by expression from nucleic acid of defined sequence derived from a SARS-CoV-2 genome, and hitherto, this has usually been the 2019 Wuhan reference virus genome. Although improbable, it is possible that SARS-CoV-2 genomes might emerge that have changed sufficiently to prevent an antibody test directed against a single virus protein or epitope being effective. Strategies to mitigate against this include the use of whole protein rather than single epitopes, use of multiple proteins, and selection of more highly conserved proteins.

6. Implications of SARS-CoV-2 genome change for efficacy of vaccines or anti-viral drugs

Vaccines

Hitherto, infection of humans by coronaviruses has not been prevented by vaccination. In addition, infection by human coronaviruses that produce cold-like symptoms does not prevent reinfection, indicating that prior infection does not induce sterilising immunity. There is some evidence that infection of humans by one coronavirus can induce antibody and T-cell immune responses that cross-react with other coronaviruses. For example, some humans that have not been infected by SARS-CoV-2 contain T-cells that detect SARS-CoV-2 antigens[68], and specifically cross-react between SARS-CoV and SARS-CoV-2[69]. Whether these T-cells are beneficial in reducing disease severity is unknown. However, patients that have recovered from mild COVID-19 have higher levels of CD8$^+$ T cells that recognise SARS-CoV-2 M and N protein peptides than patients that suffered more severe infection, suggesting that T-cell responses might be beneficial[70]. With respect to antibody there is also evidence of cross-reactivity between SARS-CoV-2 and endemic and seasonal betacoronaviruses[71] and some humans have antibodies to SARS-CoV-2 despite not having been infected by this virus[72] [73]. In addition, ~50% of human mAbs against SARS-CoV-2 N and S2 proteins are cross-reactive against human CoV-OC43 (Arthur Huang and Alain R.M. Townsend, personal communication). Nonetheless, despite serological cross-reactivity, cross-neutralisation was not observed[74].

Whether repeat infection with SARS-CoV-2 is possible is currently unknown. To address this knowledge gap, longitudinal studies with cohorts that have or have not been infected previously and that are exposed to risk of reinfection are needed. The high level of confirmed new SARS-CoV-2 infections (currently approximately >250,000 new cases / day) should make it possible to determine if prior infection by SARS-CoV-2 or vaccination against it prevent SARS-CoV-2 infection or disease. This will avoid the need for challenge studies, which, in the absence of effective therapies to prevent disease or cure infection, may be considered unethical.

If vaccination with one of the ~140 SARS-CoV-2 candidate vaccines in development is effective, the question then is whether SARS-CoV-2 will evolve to escape immunity induced by prior infection or vaccination? The answer to this is unknown. However, as the evolutionary rate of SARS-CoV-2 is lower than that of, for example, influenza A virus, any such antigenic evolution is also expected to proceed more slowly. Further, even if the immune response induced by vaccination or prior infection does not induce sterilising immunity, it can still be beneficial in preventing disease and reducing the burden of virus present and so reduce transmission.

Drugs

Drugs that reduce COVID-19 disease by either diminishing virus replication (such as the nucleoside analogue remdesivir) or transmission, or by targeting the excessive inflammatory response of the host to infection, may be affected by SARS-CoV-2 genome variation to differing degrees. Those drugs that target a particular virus protein, such as the active site of a virus-encoded enzyme, or the RBD of the S protein that is needed for virus attachment, might with time cause the selection of drug-resistant mutants with specific amino acid alterations in the target proteins. Again, the relatively low mutation rate of coronaviruses may mean drug-resistance might take longer to emerge, but it is still possible particularly if resistance is induced by a single or small number of mutations. To mitigate against the mergence of drug resistance, drugs are needed that target different stages of the virus life-cycle and these should be used in combination, rather than singly, as employed against HIV in highly active anti-retroviral therapy (HAART).

In contrast, the efficacy of drugs that impair the excessive inflammatory response to infection, which contributes to COVID-19 pathology, are not likely to be affected by virus genome alterations, because these are targeting host functions.

References

1. Lauber C *et al.* 2013 The footprint of genome architecture in the largest genome expansion in RNA viruses. PLoS Pathog, 9, e1003500. (https://doi.org/10.1371/journal.ppat.1003500).
2. *Op. cit.*, note 1.
3. Gorbalenya AE *et al.* 2006 Nidovirales: evolving the largest RNA virus genome. Virus Res, 117, 17-37. (https://doi.org/10.1016/j.virusres.2006.01.017).
4. Lu R *et al.* 2020 Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. Lancet, 395, 565-574. (https://doi.org/10.1101/2020.01.22.914952).
5. *Op. cit.*, note 4.
6. Zhu N *et al.* 2020 A Novel Coronavirus from Patients with Pneumonia in China, 2019. N Engl J Med, 382, 727-733. (https://doi.org/10.1056/NEJMoa2001017).
7. Zhou P *et al.* 2020 A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature, 579, 270-273. (https://doi.org/10.1038/s41586-020-2012-7).
8. Zhou H *et al.* 2020 A Novel Bat Coronavirus Closely Related to SARS-CoV-2 Contains Natural Insertions at the S1/S2 Cleavage Site of the Spike Protein. Curr Biol, 30, 2196-2203 e3. (https://doi.org/10.1016/j.cub.2020.05.023).
9. Su S *et al.* 2016 Epidemiology, Genetic Recombination, and Pathogenesis of Coronaviruses. Trends Microbiol, 24, 490-502. (https://doi.org/10.1016/j.tim.2016.03.003).
10. Mateus J *et al.* 2020 Selective and cross-reactive SARS-CoV-2 T cell epitopes in unexposed humans. Science, eabd3871. (https://doi.org/10.1126/science.abd3871).
11. *Op. cit.*, note 7.
12. Walls AC *et al.* 2020 Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein. Cell, 181, 281-292 e6. (https://doi.org/10.1016/j.cell.2020.02.058).
13. Lan J *et al.* 2020 Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor. Nature, 581, 215-220. (https://doi.org/10.1038/s41586-020-2180-5).
14. Huo J *et al.* 2020 Neutralizing nanobodies bind SARS-CoV-2 spike RBD and block interaction with ACE2. Nat Struct Mol Biol. (https://doi.org/10.1038/s41594-020-0469-6).
15. Wrobel AG *et al.* 2020 SARS-CoV-2 and bat RaTG13 spike glycoprotein structures inform on virus evolution and furin-cleavage effects. Nat Struct Mol Biol, 27, 763-767. (https://doi.org/10.1038/s41594-020-0468-7).
16. *Op. cit.*, note 15.
17. Hoffmann M, Kleine-Weber H, Pohlmann S. 2020 A Multibasic Cleavage Site in the Spike Protein of SARS-CoV-2 Is Essential for Infection of Human Lung Cells. Mol Cell, 78, 779-784 e5. (https://doi.org/10.1016/j.molcel.2020.04.022).
18. *Op. cit.*, note 8.
19. Zhang T, Wu Q, Zhang Z. 2020 Probable Pangolin Origin of SARS-CoV-2 Associated with the COVID-19 Outbreak. Curr Biol, 30, 1578. (https://doi.org/10.1016/j.cub.2020.03.063).
20. *Op. cit.*, note 19.
21. Andersen KG *et al.* 2020 The proximal origin of SARS-CoV-2. Nat Med, 26, 450-452. (https://doi.org/10.1038/s41591-020-0820-9).
22. *Op. cit.*, note 21.
23. Zhang YZ, Holmes EC. 2020 A Genomic Perspective on the Origin and Emergence of SARS-CoV-2. Cell, 181, 223-227. (https://doi.org/10.1016/j.cell.2020.03.035).
24. Woo PC *et al.* 2005 Phylogenetic and recombination analysis of coronavirus HKU1, a novel coronavirus from patients with pneumonia. Arch Virol, 150, 2299-311. (https://doi.org/10.1007/s00705-005-0573-2).
25. Jackwood MW *et al.* 2010 Emergence of a group 3 coronavirus through recombination. Virology, 398, 98-108. (https://doi.org/10.1016/j.virol.2009.11.044).
26. Lau SK *et al.* 2015 Severe Acute Respiratory Syndrome (SARS) Coronavirus ORF8 Protein Is Acquired from SARS-Related Coronavirus from Greater Horseshoe Bats through Recombination. J Virol, 89, 10532-47. (https://doi.org/10.1128/JVI.01048-15).
27. So RTY *et al.* 2019 Diversity of Dromedary Camel Coronavirus HKU23 in African Camels Revealed Multiple Recombination Events among Closely Related Betacoronaviruses of the Subgenus Embecovirus. J Virol, 93, e01236-19. (https://doi.org/10.1128/JVI.01236-19).
28. Boni MF *et al.* 2020 Evolutionary origins of the SARS-CoV-2 sarbecovirus lineage responsible for the COVID-19 pandemic. Nat Microbiol. (https://doi.org/10.1038/s41564-020-0771-4).
29. Gallaher WR. 2020 A palindromic RNA sequence as a common breakpoint contributor to copy-choice recombination in SARS-COV-2. Arch Virol. (https://doi.org/10.1007/s00705-020-04750-z).
30. Sanjuan R *et al.* 2010 Viral mutation rates. J Virol, 84, 9733-48. (https://doi.org/10.1128/JVI.00694-10).
31. Peck KM, Lauring AS. 2018 Complexities of Viral Mutation Rates. J Virol, 92. (https://doi.org/10.1128/JVI.01031-17).
32. *Op. cit.*, note 30.
33. Salemi M *et al.* 2004 Severe acute respiratory syndrome coronavirus sequence characteristics and evolutionary rate estimate from maximum likelihood analysis. J Virol, 78, 1602-3. (https://doi.org/10.1128/JVI.78.3.1602-1603.2004).
34. Cotton M *et al.* 2014 Spread, circulation, and evolution of the Middle East respiratory syndrome coronavirus. mBio, 5. (https://doi.org/10.1128/mBio.01062-13).
35. *Op. cit.*, note 1.
36. *Op. cit.*, note 3.
37. Eckerle LD *et al.* 2007 High fidelity of murine hepatitis virus replication is decreased in nsp14 exoribonuclease mutants. J Virol, 81, 12135-44. (https://doi.org/10.1128/JVI.01296-07).
38. Eckerle LD *et al.* 2010 Infidelity of SARS-CoV Nsp14-exonuclease mutant virus replication is revealed by complete genome sequencing. PLoS Pathog, 6, e1000896.(https://doi.org/10.1371/journal.ppat.1000896).
39. van Dorp L *et al.* 2020 Emergence of genomic diversity and recurrent mutations in SARS-CoV-2. Infect Genet Evol, 83, 104351. (https://doi.org/10.1016/j.meegid.2020.104351).
40. *Op. cit.*, note 39.

41. Duffy S, Shackelton LA, Holmes EC. 2008 Rates of evolutionary change in viruses: patterns and determinants. Nat Rev Genet, 9, 267-76. (https://doi.org/10.1038/nrg2323).

42. Simmonds P. 2020 Rampant C-->U Hypermutation in the Genomes of SARS-CoV-2 and Other Coronaviruses: Causes and Consequences for Their Short- and Long-Term Evolutionary Trajectories. mSphere, 5. (https://doi.org/10.1128/mSphere.00408-20).

43. Rambaut A *et al.* 2020 A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. Nat Microbiol. (https://doi.org/10.1038/s41564-020-0770-5).

44. *Op. cit.*, note 39.

45. *Op. cit.*, note 42.

46. Op. cit., note 42.

47. Su YCF *et al.* 2020 Discovery and Genomic Characterization of a 382-Nucleotide Deletion in ORF7b and ORF8 during the Early Evolution of SARS-CoV-2. mBio, 11. (https://doi.org/10.1128/mBio.01610-20).

48. Liu Z *et al.* 2020 Identification of common deletions in the spike protein of SARS-CoV-2. J Virol. (https://doi.org/10.1128/JVI.00790-20).

49. Gussow AB *et al.* 2020 Genomic determinants of pathogenicity in SARS-CoV-2 and other human coronaviruses. Proc Natl Acad Sci USA, 117, 15193-15199. (https://doi.org/10.1073/pnas.2008176117).

50. *Op. cit.*, note 49.

51. Zhang L *et al.* 2020 The D614G mutation in the SARS-CoV-2 spike protein reduces S1 shedding and increases infectivity. bioRxiv. (https://doi.org/10.1101/2020.06.12.148726).

52. Korber B *et al.* 2020 Tracking Changes in SARS-CoV-2 Spike: Evidence that D614G Increases Infectivity of the COVID-19 Virus. Cell, 182, 812-827 e19. (https://doi.org/10.1016/j.cell.2020.06.043).

53. Daniloski Z, Guo X, Sanjana NE. 2020 The D614G mutation in SARS-CoV-2 Spike increases transduction of multiple human cell types. bioRxiv. (https://doi.org/10.1101/2020.06.14.151357).

54. *Op. cit.*, note 52.

55. Klumpp-Thomas C *et al.* 2020 D614G Spike Variant Does Not Alter IgG, IgM, or IgA Spike Seroassay Performance. medRxiv. (https://doi.org/10.1101/2020.07.08.20147371).

56. *Op. cit.*, note 52.

57. *Op. cit.*, note 49.

58. Pybus O *et al.* 2020 Preliminary analysis of SARS-CoV-2 importation & establishment of UK transmission lineages. See https://virological.org/t/preliminary-analysis-of-sars-cov-2-importation-establishment-of-uk-transmission-lineages/507 (accessed 06 August 2020).

59. *Op. cit.*, note 58.

60. Oude Munnink BB *et al.* 2020 Rapid SARS-CoV-2 whole-genome sequencing and analysis for informed public health decision-making in the Netherlands. Nat Med. (https://doi.org/10.1038/s41591-020-0997-y).

61. Jesus JG *et al.* 2020 Importation and early local transmission of COVID-19 in Brazil, 2020. Rev Inst Med Trop Sao Paulo, 62, e30. (https://doi.org/10.1590/s1678-9946202062030).

62. Meredith LW *et al.* 2020 Rapid implementation of SARS-CoV-2 sequencing to investigate cases of health-care associated COVID-19: a prospective genomic surveillance study. Lancet Infect Dis. (https://doi.org/10.1016/S1473-3099(20)30562-4).

63. Sikkema RS *et al.* 2020 COVID-19 in health-care workers in three hospitals in the south of the Netherlands: a cross-sectional study. Lancet Infect Dis. (https://doi.org/10.1016/S1473-3099(20)30527-2).

64. *Op. cit.*, note 62.

65. Jones NK *et al.* 2020 Effective control of SARS-CoV-2 transmission between healthcare workers during a period of diminished community prevalence of COVID-19. Elife, 9, e59391. (https://doi.org/10.7554/eLife.59391).

66. *Op. cit.*, note 65.

67. Rivett L *et al.* 2020 Screening of healthcare workers for SARS-CoV-2 highlights the role of asymptomatic carriage in COVID-19 transmission. Elife, 9, e58728. (https://doi.org/10.7554/eLife.58728).

68. Grifoni A *et al.* 2020 Targets of T Cell Responses to SARS-CoV-2 Coronavirus in Humans with COVID-19 Disease and Unexposed Individuals. Cell, 181, 1489-1501 e15. (https://doi.org/10.1016/j.cell.2020.05.015).

69. Le Bert N *et al.* 2020 SARS-CoV-2-specific T cell immunity in cases of COVID-19 and SARS, and uninfected controls. Nature, 584, 457-462. (https://doi.org/10.1038/s41586-020-2550-z).

70. Peng Y *et al.* 2020 Broad and strong memory CD4 (+) and CD8 (+) T cells induced by SARS-CoV-2 in UK convalescent COVID-19 patients. bioRxiv. (https://doi.org/10.1101/2020.06.05.134551).

71. Hicks J *et al.* 2020 Serologic cross-reactivity of SARS-CoV-2 with endemic and seasonal Betacoronaviruses. medRxiv. (https://doi.org/10.1101/2020.06.22.20137695).

72. Ng KW *et al.* 2020 Pre-existing and de novo humoral immunity to SARS-CoV-2 in humans. bioRxiv. (https://doi.org/10.1101/2020.05.14.095414).

73. To KK-W *et al.* 2020 Seroprevalence of SARS-CoV-2 in Hong Kong and in residents evacuated from Hubei province, China: a multicohort study. The Lancet Microbe, 1, e111–e118. (https://doi.org/10.1016/S2666-5247(20)30053-7).

74. *Op. cit.*, note 73.

DISCLAIMER

This paper has drawn on the most recent evidence up to 26 August 2020 and has been subject to formal peer-review. Further evidence on this topic is constantly published and the Royal Society may return to this topic in the future. This independent overview of the science has been provided in good faith by subject experts and the Royal Society and paper authors accept no legal liability for decisions made based on this evidence.

THANKS

The Royal Society is grateful to the Leverhulme Trust for its support for the Society's pandemic response work.

Issued: September 2020 DES7139 © The Royal Society