Identification of SARS-CoV-2 recombinant genomes

David VanInsberghe[1], Andrew Neish[1], Anice C. Lowen[2,3], Katia Koelle[3,4]

Affiliations:
[1]Department of Pathology, Emory University, Atlanta, GA, USA
[2]Department of Microbiology and Immunology, Emory University, Atlanta, GA, USA
[3]Emory-UGA Center of Excellence for Influenza Research and Surveillance (CEIRS), Atlanta GA, USA
[4]Department of Biology, Emory University, Atlanta, GA, USA

Abstract

Viral recombination has the potential to bring about viral genotypes with modified phenotypic characteristics, including transmissibility and virulence. Although the capacity for recombination among Betacoronaviruses is well documented, SARS-CoV-2 has only been circulating in humans for approximately 8 months and thus has had a relatively short window of opportunity for the occurrence of recombination. The ability to detect recombination has further been limited by the relatively low levels of genetic diversity in SARS-CoV-2. Despite this, two studies have reported recombinants among SARS-CoV-2 strains. Here we first revisit these findings with a new analysis approach, arguing that neither presents a clear case of within-SARS-CoV-2 recombination. Applying this same approach to available SARS-CoV-2 sequences, we then identify five recombinant genomes. Each of these genomes contain phylogenetic markers of two distinct SARS-CoV-2 clades. Further, the predicted parent clades of these recombinant genomes were, with one exception, documented to be co-circulating in the country of infection in the two weeks prior to the sample being collected. Our results indicate that recombination among SARS-CoV-2 strains is occurring, but is either not widespread or often remains undetectable given current levels of viral genetic diversity. Efforts to monitor the emergence of new recombinant genomes should therefore be sustained.

Introduction

Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) emerged in December of 2019 in China but has since spread worldwide, cumulatively infecting more than 17 million people by August, 2020. Laboratories around the world have been sequencing and rapidly sharing SARS-CoV-2 genomes throughout the pandemic, providing researchers the rare opportunity to study the evolution of SARS-CoV-2 in real-time. As of June 20, 2020, 68019 complete viral genomes and 47390 unique genotypes are available on the online repository GISAID (Elbe et al., 2017).

In addition to point mutations and insertions/deletions, coronavirus evolution is heavily driven by recombination (Su et al., 2016). Recombination events create chimeric genotypes between two viral strains that infect the same cell. This process occurs when RNA polymerase prematurely stops replicating the first genotype before reassembling and resuming replication with the second genotype as template. The end result is the unlinking of mutations across the genome, creating novel combinations of existing mutations and allowing selection to operate

more efficiently. The clinical and epidemiological relevance of these new combinations is substantial as they have the potential to create genotypes with unique virulence and transmissibility characteristics.

Measurements of the frequency of this process among coronaviruses in cell culture suggest it is very common (Schaad et al., 1990; Banner et al., 1991). There have further been attempts to detect and measure the magnitude of recombination among naturally circulating SARS-CoV-2 genomes. Based on four single nucleotide polymorphisms (SNPs), an early analysis reported recombinants among the first 85 sequenced SARS-CoV-2 genomes (Yi, 2020). A more recent analysis has reported the identification of recombination among sympatric SARS-CoV-2 strains (Korber et al., 2020). Together, these studies suggest that SARS-CoV-2 recombination may be commonly occurring. In contrast, three reports identified evidence of strong linkage disequilibrium among SARS-CoV-2 polymorphic sites and no disruption of the clonal pattern of inheritance that would accompany widespread recombination (Maio et al., 2020; Nie et al., 2020; Wang et al., 2020).

In an effort to reconcile these findings, we developed a systematic, four-step approach to identify recombinant SARS-CoV-2 genomes. This approach involves (1) characterizing the mutations that define the clonal pattern of inheritance in SARS-CoV-2 clade structure, (2) identifying genomes that violate this pattern, (3) identifying and refining the boundaries of genetic transfer by analyzing lower frequency SNPs shared by predicted parent sequences, and (4) assessing the plausibility of transfer by determining if the predicted parental clades were co-circulating in the country of infection. Using this approach, we find that none of the putative recombinants identified by Korber et al. (2020) contain combinations of clade-defining SNPs that are indicative of recombination. However, by analyzing the 47,390 unique sequences available on GISAID by June 20 2020, we find five viral genomes that contain unique combinations of clade-defining SNPs that are indicative of recombination. Non-clade-defining SNPs present in these genomes further support their chimeric origins and help refine the region of transfer. Finally, the parental clades of four putative recombinants were co-circulating in the country of infection two weeks prior to sampling, while an insufficient number of genomes sequenced in proximity to the fifth putative recombinant precludes this analysis. Ultimately, our results suggest that recombination among SARS-CoV-2 genomes is occurring, but that the resultant genotypes are not widespread.

Results

The limited genome-wide diversity among SARS-CoV-2 strains restricts the ability to detect recombinants

Given the relatively low mutation rate of coronaviruses and the limited amount of time since the pandemic began, only a small amount of genetic variation is currently present in SARS-CoV-2 genomes. This diversity falls largely into 14 monophyletic clades (Fig. 1 A). We screened all biallelic sites in the reference genome alignment to identify clade-defining SNPs, defined as positions where <5% of genomes in at least one clade have the dominant allele, and >95% of genomes in remaining clades have the dominant allele. In total, we identified 37 clade-defining

80 SNPs that reliably distinguish all 14 clades (Fig. 1 ABC). Of the 67167 complete, human-derived strains available on GISAID as of July 20, 2020, fewer than 0.5% of genomes differed from the 14 clade-defining SNP profiles by more than one nucleotide.

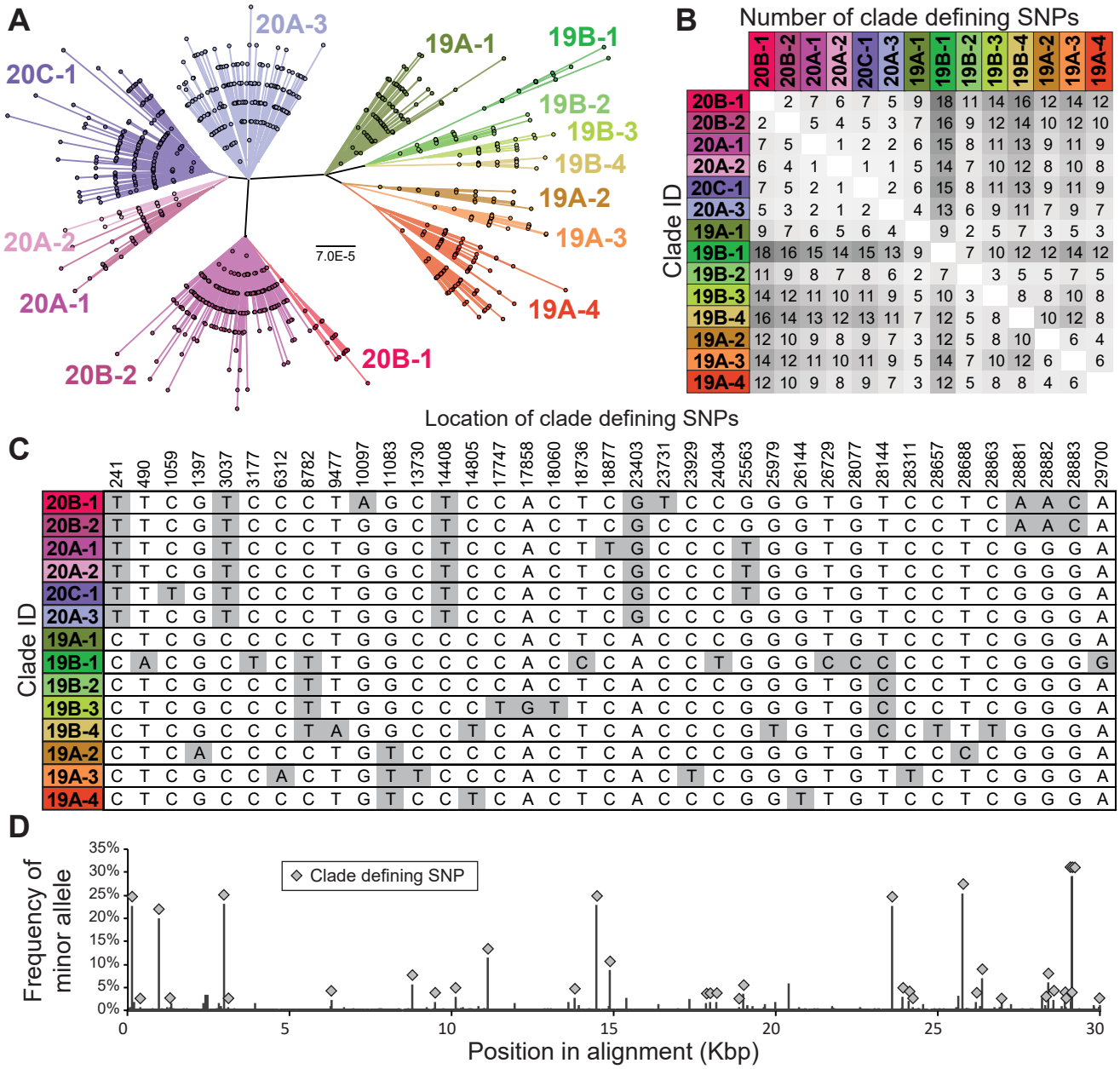

Figure 1. The clade structure of SARS-CoV-2 is structured predominantly by 37 clade-defining SNPs. **(A)** Maximum likelihood phylogeny based on the General Time Reversible model with invariant sites of 9783 high quality unique genome sequences with <1% Ns. 14 monophyletic clades were identified manually. These clades generally correspond to SARS-CoV-2 clades defined in Nextstrain (Hadfield et al., 2018), although a fraction of them are at higher resolution than Nextstrain clades. Clades defined here are named by Nextstrain clade designation (e.g., 20B) followed by a subclade number (e.g., -1). Scale bar is in substitutions per site. **(B)** Pairwise differences between the clade-defining SNP profiles of all 14 clades. **(C)** Location and nucleotide identity of clade-defining SNPs and **(D)** their frequency among SARS-CoV-2 genomes.

Since the clade structure of SARS-CoV-2 is driven by such a limited number of SNPs, it is often impossible to distinguish recombination from *de novo* mutation. Indeed, the limited number of SNPs that distinguish certain clades are also often clustered in short regions of the genome, further restricting our ability to reliably identify recombinant genomes. For instance, clades 20B-2 and 20A-2 are primarily distinguishable based on four SNPs (25563 and 28881-3), but those four positions span only 3.3 kb of the viral genome. As a result, recombination between strains from clades 20B-2 and 20A-2 throughout the first 80% of the genome would not be detectible. Further, a recombination event that unlinks the nucleotides at positions 25563 and 28881-3 would be more parsimoniously explained as a *de novo* T to G mutation at position 25563 in a clade 20A-2 genome or a *de novo* G to T mutation at this position in a clade 20B-2 genome.

Nevertheless, there are many circumstances where recombination could be more reliably detected and distinguished from *de novo* mutation. In particular, all major clades are most strongly differentiated from each other based on 11 SNPs that are distributed throughout the genome (sites 241, 3037, 8782, 11083, 14408, 23403, 25563, 28144, 28881-3). Rearrangement of multiple of these clade-defining markers would be among the strongest indication of recombination between SARS-CoV-2 strains. For instance, the triple mutation GGG to AAC at positions 28881-3 is uniquely found in clade 20B, and would be a strong marker of recombination between clade 20B and clades 19A and 19B when combined with positions 241, 3037, 14408, and 23403.

Although one report has suggested that the triple GGG to AAC mutation at positions 28881-3 is found in multiple clades (Maio et al., 2020), by our analysis this mutation was restricted to clade 20B. We find that any ambiguity about the distribution of AAC is related to the limited number of SNPs outside of positions 28881-3 to distinguish clade 20B from clades 20A and 20C. In fact, without positions 28881-3, none of the clade defining SNPs we identified can distinguish clade 20B-2 from 20A-3, and only one can distinguish 20B-2 from 20A-2 (position G25563T).

Previously identified recombinant genomes have no rearrangements of clade-defining SNPs

The first report of recombination among SARS-CoV-2 genomes was a correspondence article by Huiguang Yi (2020), prepared at a time when there were 84 SARS-CoV-2 genomes in GISAID. This article argued the distribution of 4 SNPs in those early genomes could be explained by multiple recombination events. With such little information, it is difficult to evaluate the strength of these claims. However, three of the four polymorphic sites on which they base their argument are, by our analysis, monophyletic traits. The remaining site, C29095T is not one of the clade-associated SNPs we identified, but is a low frequency allele that is found in multiple clades (Hadfield et al., 2018) and is thus very likely a homoplasy.

A second report used a three-way sequence comparison tool, RAPR, on geographically constrained subsets of GISAID genomes to identify recombination involving strains in the United States (Washington State), in the Netherlands, and in Iceland (Korber et al., 2020). However, all but two of the sequences that Korber et al. (2020) identified as recombinant pairs match our clade SNP profiles (Fig. 2) and thus squarely fall into only one major SARS-CoV-2 clade. The two exceptions (EPI_ISL_422850 and EPI_ISL_422679) each have clade-defining

SNP profiles that differ from the nearest clade by only a single clade-defining SNP substitution (Fig. 2). As such, while these genomes could be explained by recombination, a more parsimonious explanation would be *de novo* mutation.

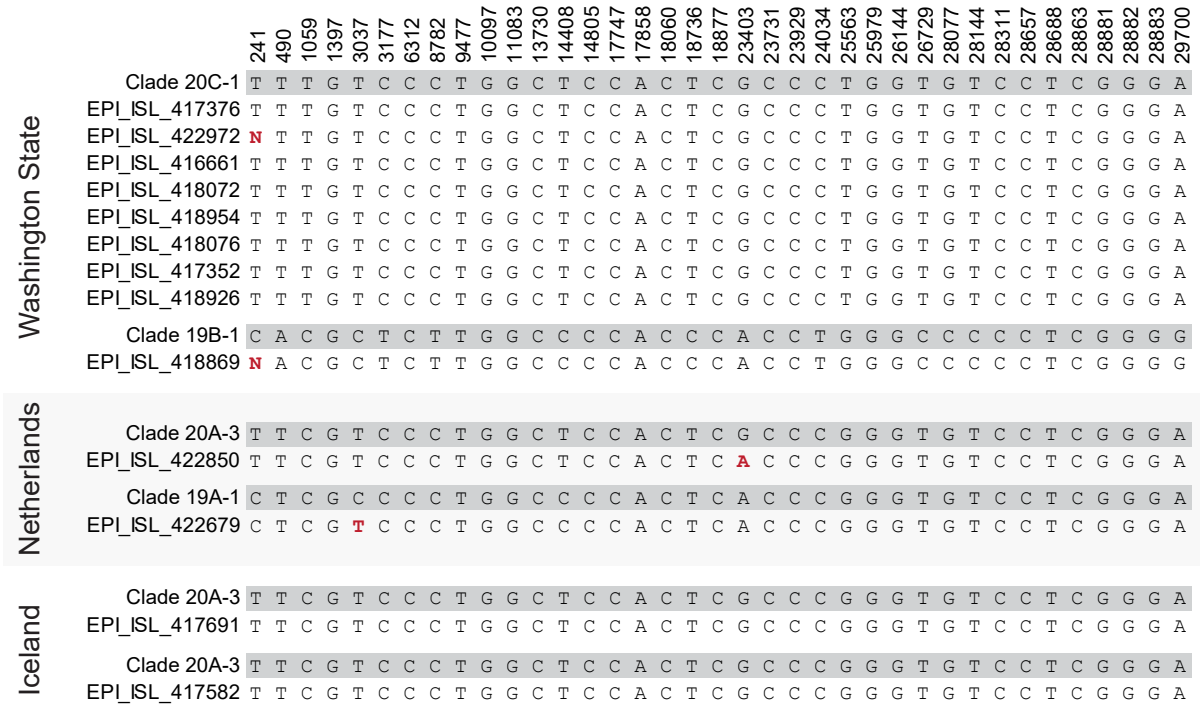

Figure 2. Genomes previously suggested to be recombinant do not show evidence of clade-defining SNP re-arrangement. The clade-defining SNP profiles of each sequence identified as recombinant by Korber et al. (2020) are compared to the SNP profile of the genetically most similar clade (highlighted in grey). Nucleotide differences are highlighted in red. Sequences are named according to GISAID accession number.

Clade-defining and low-frequency SNPs support recombination in five genomes

We next aimed to determine if any of the genomes available on GISAID have combinations of clade-defining SNPs that can be most parsimoniously explained by recombination. In total, we screened 47390 unique genomes and identified five genomes that are strong candidates for having evolved through recombination between two distantly related parental clades (Fig. 3). These sequences were associated with infections from the USA, from the United Kingdom, and from China, and between seven to ten clade-defining SNPs support transfer between two clades. These genomes were sequenced on Illumina Novaseq and NextSeq, and Nanopore GridION and MinION instruments. Although assembly quality information is not consistently indicated in the sequence metadata, three of the five genomes contain no ambiguous nucleotides, one genome (USA/CA-CZB-1437/2020) has 2, and the last genome (England/201090235/2020) has 952 (3.2% of all positions). Each recombinant sequence occurs in the GISAID database only once.

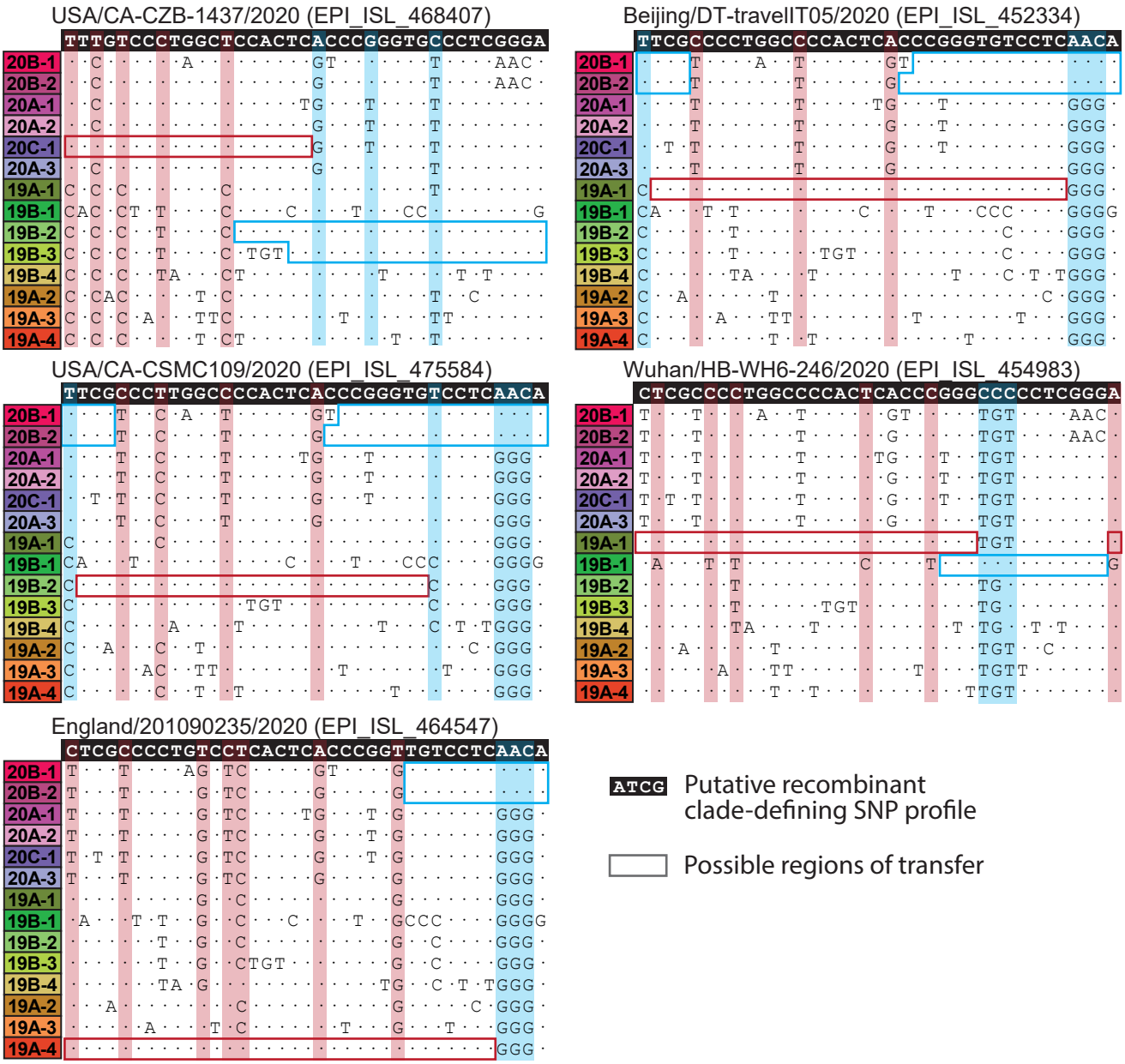

Figure 3. Five sequences parsimoniously generated through recombination between strains from two parental clades. The clade-defining SNP profile of each putative recombinant sequence is compared with the profiles of all 14 clades. Nucleotides that match the putative recombinant are denoted with a dot. Regions boxed in blue and red show potential parental clades, with left and right boundaries indicating potential transfer regions. SNPs that support recombination are highlighted with vertical blue and red windows.

In addition to clade-defining SNPs, minor SNPs that are only present in a subset of strains in each clade support recombination in the five putative recombinant sequences (Fig. 4). The nearest parent sequences of each putative recombinant were identified by searching for genomes
145 that have the fewest nucleotide differences across each proposed region of transfer. The location of these minor SNPs helped further refine the boundaries of transfer. Interestingly, three of the five candidate recombinants require multiple transfers to explain the pattern of SNPs we observe

(Beijing/DT-travelIT05/2020, USA/CA-CSMC109/2020, and Wuhan/HB-WH6-246/2020). While this might suggest that these sequences are not true recombinants, these findings could reflect a high frequency of recombination in co-infected cells. In support of this latter interpretation, a high frequency of transfer was documented in experiments measuring recombination among murine Betacoronaviruses (Schaad et al., 1990; Banner et al., 1991).

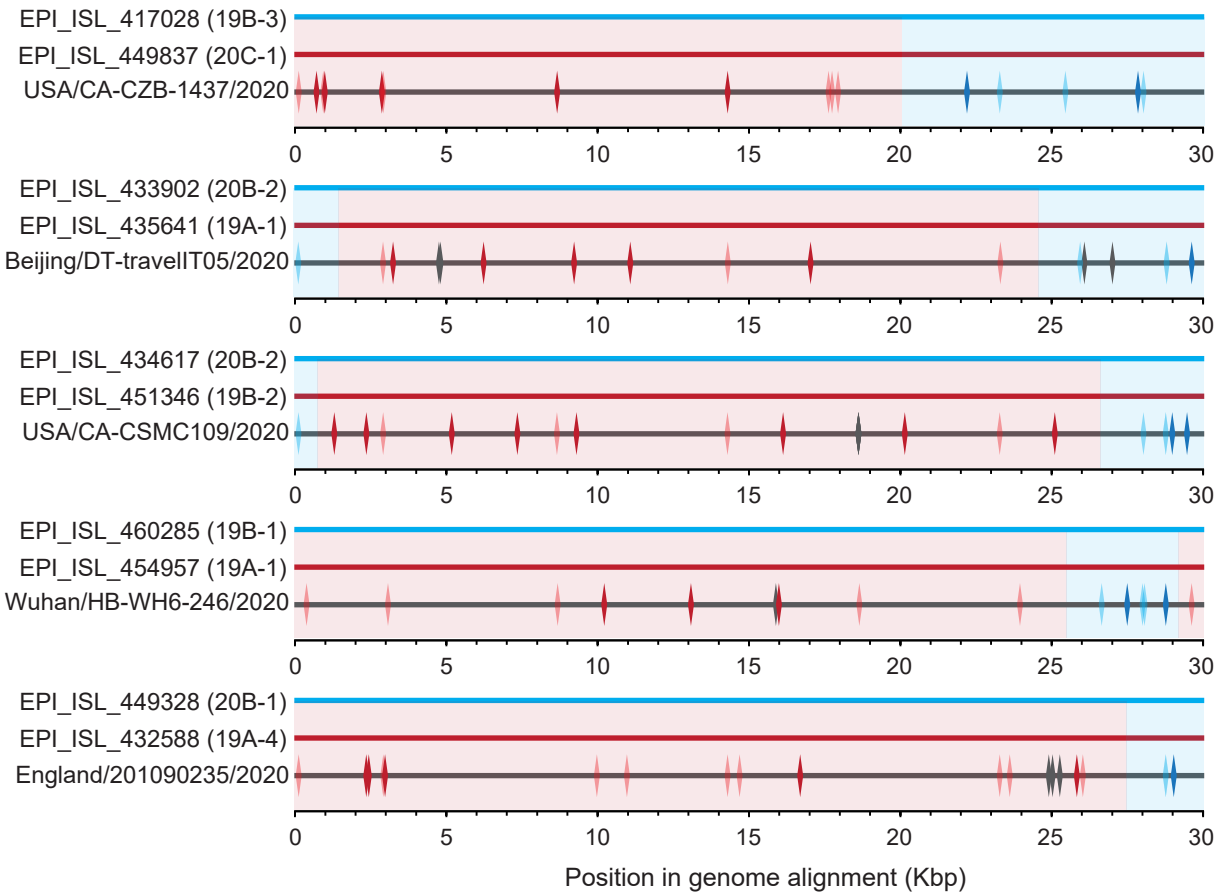

Figure 4. Low frequency polymorphisms support transfer between parental clades to generate identified recombinants. The two parent genomes were chosen as the most closely related sequences to each putative recombinant genome across the shaded regions of transfer. The locations of SNPs that are shared or unique to each strain are indicated by color, where grey indicates a SNP unique to the recombinant. Light red and light blue ticks indicate clade-defining SNPs, while dark red and dark blue ticks indicate low frequency, non-clade defining SNPs. One parent sequence from each region of transfer is shown, but for Beijing/DT-travelIT05/2020, USA/CA-CSMC109/2020, and England/201090235/2020, multiple strains from both clades 20B-1 and 20B-2 are equally distant to the recombinant genomes across the blue regions of transfer. Likewise, multiple genomes from clades 19B-2 and 19B-3 are equally distant to USA/CA-CZB-1437/2020 across the blue region of transfer.

To visualize the empirical support for recombination in the 5 genomes we identified, we performed phylogenetic analysis on subsets of the SARS-CoV-2 viral genome corresponding to stretches of the genome bounded by inferred regions of transfer (Fig5). These trees support the pattern of clade-defining and low-frequency SNPs shared by the most closely related parent sequences

shown in Figure 4. They show that across the regions of transfer, each recombinant genome clusters tightly with the predicted parent clades.

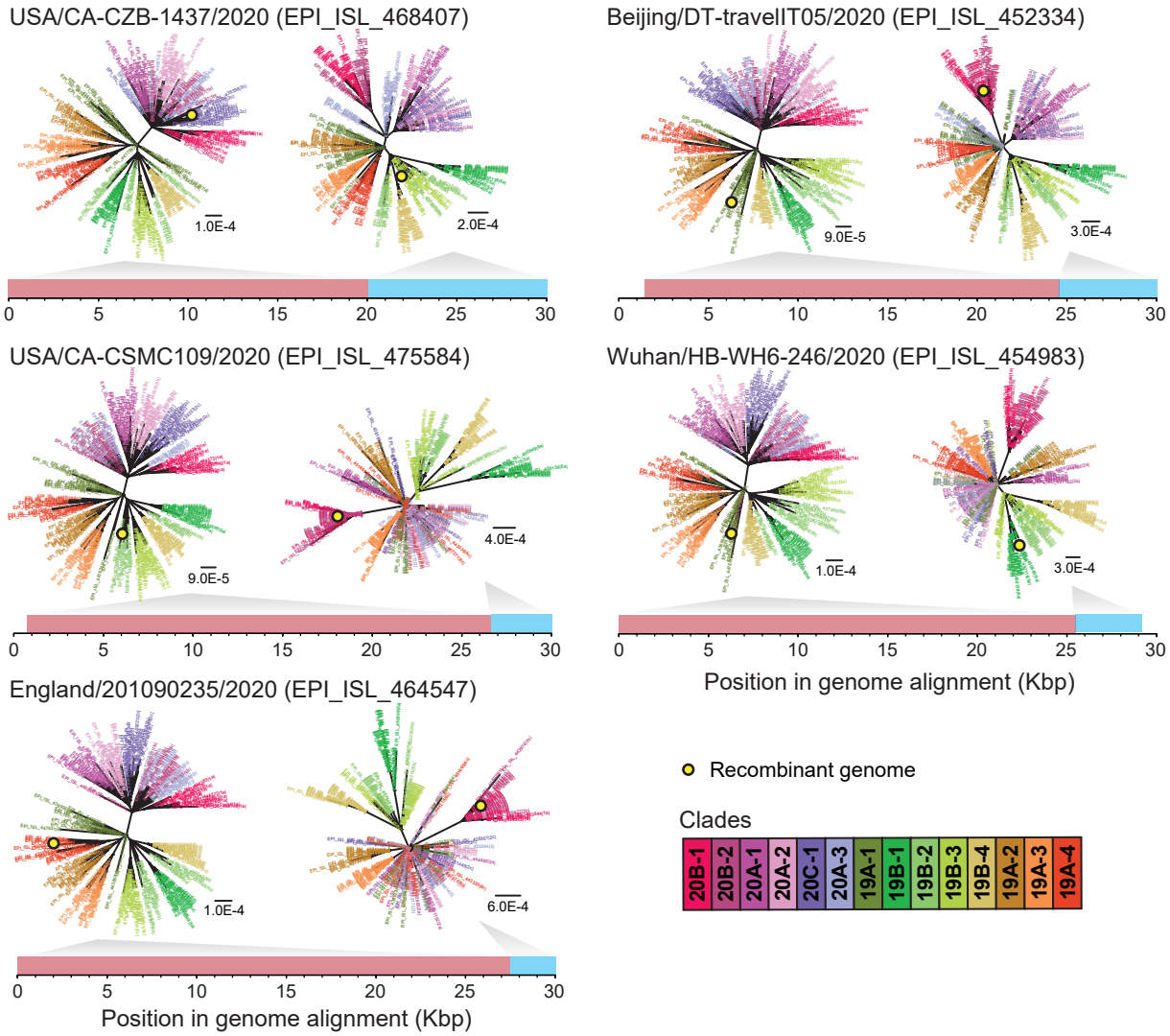

Figure 5. Phylogenies inferred using restricted regions of the SARS-CoV-2 viral genome provide visual support for the five identified recombinants. Maximum likelihood phylogenies were inferred under the General Time Reversible model with invariant sites. For clarity, the shown phylogenies include only 15 representative genomes from each of the 14 clades.

We next sought to assess, based on geographic considerations, the plausibility of transfer between the predicted parental clades to generate the observed recombinants. We counted all genomes that map to the 14 clades that were sequenced in the region of infection in the two weeks prior to the collection date of each recombinant (Fig. 6). For all but one of the putative recombinants, both parental clades were detected, and often at high frequencies. Although only one of the parent clades was detectible for Wuhan/HB-WH6-246/2020, there were only 45 genomes sequenced in China in the two weeks prior to sampling that have a full sample date

listed and passed our quality filters. As such, it is possible that with greater sequencing frequency, clade 19B-1 could have been detected.

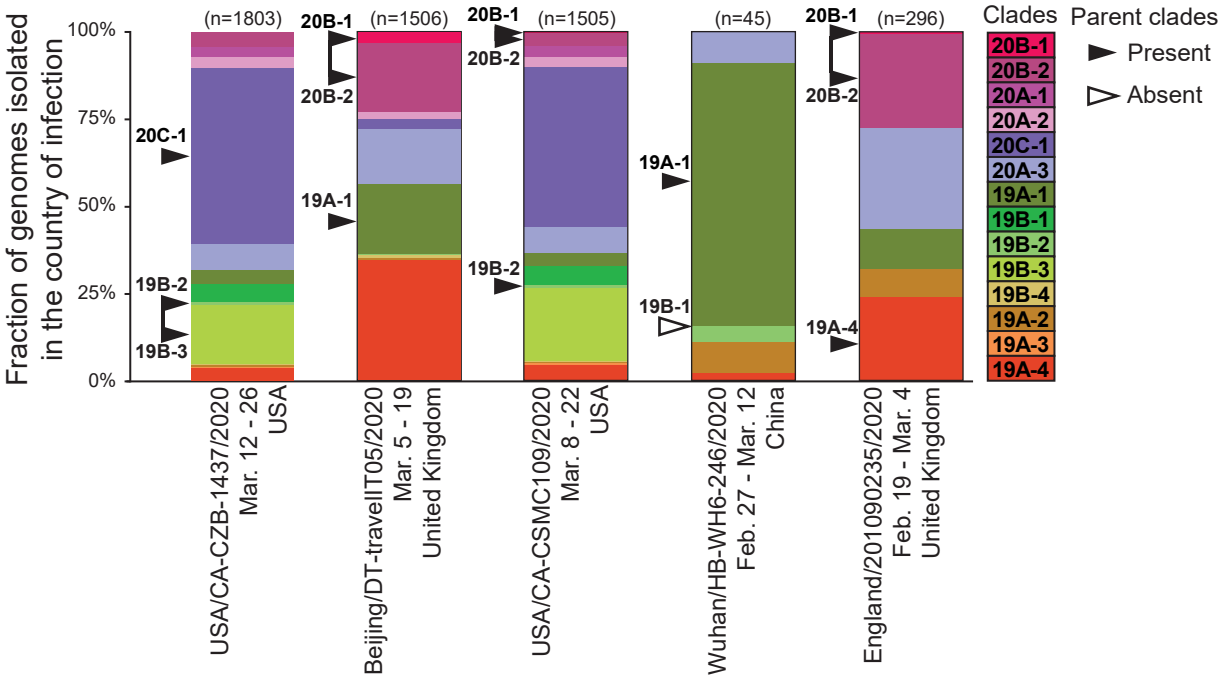

Figure 6. Identified parental clades generally circulated in geographic regions of identified recombinants. Sample Beijing/DT-travelIT05/2020 (EPI_ISL_452334) was isolated in Beijing, but the infection was acquired in the United Kingdom during travel.

Discussion

The small number of polymorphic sites in the SARS-CoV-2 genome that are phylogenetically informative means detecting recombinant genomes is difficult, and highly dependent on the identity of the parent clades. By identifying the nucleotide changes that underpin the clonal phylogeny of SARS-CoV-2, we established criteria for identifying putative recombinant genomes, and for evaluating their plausibility. Although there is limited evidence by these criteria to support transfer among genomes previously identified as recombinant sequences by Yi (2020) and by Korber et al. (2020), we analyzed approximately 68,000 SARS-CoV-2 genomes from GISAID and found five viral sequences that presented strong evidence of recombination. These genomes have rearrangements of between 7 and 10 clade-defining SNPs that point to 2-3 clades as potential parental clades (Fig. 3) Many lower frequency non-clade-defining SNPs resolve which 2 of the identified clades are more likely parents and refine the genomic regions of transfer (Fig. 4). Each of the five recombinants identified here occur only once in GISAID.

It is unlikely that the chimeric sequences we identified are the result of multiple non-recombinant genomes present in mixed infections or of low-quality sequences. In the event of co-infection of a person by two different SARS-CoV-2 strains, if both strains are present at approximately equal levels, polymorphic sites in the final sequencing assembly are most likely to be called as ambiguous nucleotides. If one strain predominates within a mixed infection, then polymorphic

sites in the final assembly will be more likely to match this more abundant strain, and the less abundant strain will be undetectable at the consensus level. Although two of the five putative recombinants have ambiguous nucleotides in their assembly (2 in USA/CA-CZB-1437/2020 and 952 in England/201090235/2020), none of the ambiguous nucleotide sites overlap with clade-defining SNPs.

While we were able to identify five recombinant SARS-CoV-2 genomes, the fraction of recombinant genomes in the set of sequences we analyzed was extremely low (0.007%). This observation supports reports that have found no evidence of widespread recombination among SARS-CoV-2 genomes (Maio et al., 2020; Nie et al., 2020; Wang et al., 2020). Indeed, examining the pattern of clade-defining SNPs suggests none of these lineages emerged through recombination (Fig. 1). The only site that does not strictly follow the pattern of vertical decent is C14805T, which occurs in both clades 19A-4 and 19B-4. However, none of the seven other clade-defining SNPs that differentiate these clades support recombination, suggesting C14805T is a homoplastic trait.

The low frequency of strongly-supported recombinant sequences in GISAID could be due to multiple factors. First, due to the limited genetic diversity of SARS-CoV-2 at this point in time, a large fraction of viral recombinants may not be detectable. This is particularly the case if the parental clades share a large number of clade-defining SNPs. Second, recombinant genomes may be rare because coinfections only rarely occur, or because recombinant genomes that may arise in coinfected individuals rarely transmit. Coinfection may be infrequent for SARS-CoV-2 given the acute nature of the infection and that some regions (but not others) have managed to keep the level of virus circulation low. In instances when coinfections do occur, recombinant genomes may be generated late in the infection or have lower fitness, resulting in rare onward transmission.

Each of the five recombinant SARS-CoV-2 genomes we identified was also a singleton. While this may appear surprising at first, it may be the case that there are recombinant lineages circulating at low frequencies that have gone undetected. Alternatively, the recombinant genomes we identified may not be part of larger, persistent recombinant lineages. Instead, the identified genomes may be fleeting observations into lineages that have gone extinct or not successfully established. This is a clear possibility given the extent of SARS-CoV-2 transmission heterogeneity, with estimates of 5-10% of infected individuals being responsible for upwards of 80% of secondary infections (Bi et al., 2020; Endo et al., 2020; Miller et al., 2020). Transmission heterogeneity results in a lower chance for any given viral infection to establish a persisting lineage. Indeed, with the level of transmission heterogeneity that has been estimated for SARS-CoV-2, only 20% of recombinant genomes that have been transmitted from a coinfected individual to a singularly infected individual would be expected to successfully establish a viral lineage (Lloyd-Smith et al., 2005).

Ultimately, our results suggest that recombination between SARS-CoV-2 strains is occurring, but these chimeric genotypes remain rare. As the pandemic continues to expand, the population genetic diversity of SARS-CoV-2 will increase, making it easier to detect recombinant genomes. With an increasing number of mutations, the possibility for recombinant genomes to have altered

phenotypic characteristics that impact fitness will also increase. Given our finding that recombination is already occurring in SARS-CoV-2, surveillance efforts and real-time analyses to detect recombinants, such as the one here, should be sustained to monitor the circulation and potential spread of high-fitness recombinant genotypes.

Materials and Methods

Genome quality filtering and alignment

Genomes were downloaded from the GISAID genome databases (Elbe et al., 2017), and filtered to exclude low quality sequences. All genomes were trimmed relative to positions 118 and 29740 in the NCBI reference sequence (accession NC_045512) to exclude low coverage sites. Genomes with less than 10% Ns and a final trimmed length greater than 29,610 bp and less than 29,660 bp were included in further analysis. Genomes were aligned to the NCBI reference sequence genome using MAFFT v7.464 (Katoh et al., 2013).

Identifying clade-defining SNPs in SARS-CoV-2 genomes

Clades were identified as monophyletic groups within a maximum likelihood phylogenetic tree built from 9783 unique high quality genome sequences with <1% Ns using PhyML (Guindon et al., 2010). Clade-specific SNPs were subsequently identified as SNPs that are present in >95% of all members of a clade while >95% of the members in remaining clades had another nucleotide at that position. Recombinant genomes were identified by manually screening the SNP profiles of any genome that differed from the nearest clade profile by more than 1 base difference for evidence of reassortment between the profiles of any two clades. In total, 68,019 genomes were screened.

Code availability

All custom computer code necessary to reproduce our results are available on GitHub (https://github.com/davevanins/Sars-CoV-2_CladeSNP).

Acknowledgments

This study was supported by NIAID Centers of Excellence for Influenza Research and Surveillance (CEIRS) grant HHSN272201400004C and an Emory University MP3 seed grant. We gratefully acknowledge all of the authors from the originating laboratories responsible for obtaining the specimens and the submitting laboratories where genetic sequence data were generated and shared via the GISAID Initiative, on which this research is based. In particular, we wish to acknowledge the following laboratories that obtained and sequenced the five genomes which are the primary subject of this manuscript: County of San Luis Obispo Public Health Laboratory, Laboratory of Infectious Diseases Center of Beijing Ditan Hospital, Cedars-Sinai Medical Center Molecular Pathology Laboratory, Wuhan Chain Medical Labs, and the Respiratory Virus Unit of Public Health England.

References

Banner, L. R.; Mc Lai, M., 1991: Random nature of coronavirus RNA recombination in the absence of selection pressure. *Virology*., **185**, 441–445.

Bi, Q.; Wu, Y.; Mei, S.; Ye, C.; Zou, X.; Zhang, Z.; Liu, X.; Wei, L.; Truelove, S. A.; Zhang, T.; Gao, W.; Cheng, C.; Tang, X.; Wu, X.; Wu, Y.; Sun, B.; Huang, S.; Sun, Y.; Zhang, J. et al., 2020: Epidemiology and transmission of COVID-19 in 391 cases and 1286 of their close contacts in Shenzhen, China: a retrospective cohort study. *The Lancet. Infectious diseases*., **20**, 911–919.

Elbe, S.; Buckland-Merrett, G., 2017: Data, disease and diplomacy: GISAID's innovative contribution to global health. *Global Challenges*., **1**, 33–46.

Endo, A.; Abbott, S.; Kucharski, A. J.; Funk, S.; Funk, S., 2020: Estimating the overdispersion in COVID-19 transmission using outbreak sizes outside China. *Wellcome Open Research*., **5**, 67.

Guindon, S.; Dufayard, J. F.; Lefort, V.; Anisimova, M.; Hordijk, W.; Gascuel, O., 2010: New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. *Systematic Biology*., **59**, 307–321.

Hadfield, J.; Megill, C.; Bell, S. M.; Huddleston, J.; Potter, B.; Callender, C.; Sagulenko, P.; Bedford, T.; Neher, R. A., 2018: Nextstrain: real-time tracking of pathogen evolution. (Kelso, J., Ed.)*Bioinformatics*., **34**, 4121–4123.

Katoh, K.; Standley, D. M., 2013: MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Molecular Biology and Evolution*., **30**, 772–780.

Korber, B.; Fischer, W.; Gnanakaran, S. G.; Yoon, H.; Theiler, J.; Abfalterer, W.; Foley, B.; Giorgi, E. E.; Bhattacharya, T.; Parker, M. D.; Partridge, D. G.; Evans, C. M.; Silva, T. de; LaBranche, C. C.; Montefiori, D. C.; Group, S. C. 19 G., 2020: Spike mutation pipeline reveals the emergence of a more transmissible form of SARS-CoV-2. *bioRxiv*., **4**, 2020.04.29.069054.

Lloyd-Smith, J. O.; Schreiber, S. J.; Kopp, P. E.; Getz, W. M., 2005: Superspreading and the effect of individual variation on disease emergence. *Nature*., **438**, 355–359.

Maio, N. De; Walker, C.; Borges, R.; Weilguny, L.; Slodkowicz, G.; Goldman, N., 2020: Issues with SARS-CoV-2 sequencing data. *Virological.org*., 1–19.

Miller, D.; Martin, M. A.; Harel, N.; Kustin, T.; Tirosh, O.; Meir, M.; Sorek, N.; Gefen-Halevi, S.; Amit, S.; Vorontsov, O.; Wolf, D.; Peretz, A.; Shemer-Avni, Y.; Roif-Kaminsky, D.; Kopelman, N.; Huppert, A.; Koelle, K.; Stern, A., 2020: Full genome viral sequences inform patterns of SARS-CoV-2 spread into and within Israel. *medRxiv*., 2020.05.21.20104521.

Nie, Q.; Li, X.; Chen, W.; Liu, D.; Chen, Y.; Li, H.; Li, D.; Tian, M.; Tan, W.; Zai, J., 2020: Phylogenetic and phylodynamic analyses of SARS-CoV-2. *Virus Research*., **287**, 198098.

Schaad, M. C.; Stohlman, S. A.; Egbert, J.; Lum, K.; Fu, K.; Wei, T.; Baric, R. S.; Baric, R. S., 1990: Genetics of mouse hepatitis virus transcription: identification of cistrons which may function in positive and negative strand RNA synthesis. *Virology*., **177**, 634–645.

Su, S.; Wong, G.; Shi, W.; Liu, J.; Lai, A. C. K.; Zhou, J.; Liu, W.; Bi, Y.; Gao, G. F., 2016:

Epidemiology, Genetic Recombination, and Pathogenesis of Coronaviruses. *Trends in Microbiology*., **24**, 490–502.

Wang, H.; Kosakovsky Pond, S. L.; Nekrutenko, A.; Nielsen, R., 2020, May 27: Testing recombination in the pandemic SARS-CoV-2 strains - Novel 2019 coronavirus / nCoV-2019 Genomic Epidemiology - Virological.

Yi, H., 2020: 2019 Novel Coronavirus Is Undergoing Active Recombination. *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America*., **71**, 884–887.