

Arné de Klerk

DKLARN001

MSc (Med) Bioinformatics

Conserved recombination patterns across coronavirus subgenera

Supervised by Associate Professor Darren Patrick Martin

Division Of Computational Biology

Faculty of Health Sciences

University of Cape Town

Anzio Road

7925 Observatory, Cape Town

South Africa



The copyright of this thesis vests in the author. No quotation from it or information derived from it is to be published without full acknowledgement of the source. The thesis is to be used for private study or non-commercial research purposes only.

Published by the University of Cape Town (UCT) in terms of the non-exclusive license granted to UCT by the author.

Declaration

I, Arne de Klerk (DKLARN001), hereby declare that the work on which this dissertation is based is my original work (except where acknowledgements indicate otherwise) and that neither the whole work nor any part of it has been, is being, or is to be submitted for another degree in this or any other university. I empower the university to reproduce for the purpose of research either the whole or any portion of the contents in any manner whatsoever.

Date: 13 December 2021

Signed: 12 April 2022

Signed by candidate

Acknowledgements

Intertwined between the lines of this thesis is the prodigious support from numerous kind, caring and wise individuals, to whom I share my eternal gratitude.

Before I thank my supervisor and peers for their relentless support throughout the years, I would like to thank my parents, for theirs, during the last (nearly) 24. Their collective time-management, analytical thinking, management skills, and unwavered faith and support has consistently and constantly helped guide me, both in my early and future walks of life.

I am deeply indebted to my supervisor, Dr Darren Martin, who so generously allocated tremendous amounts of expertise, time and patience into helping forge every step of this project, contributing enormously to the final product. It has been a humbling experience to work with a person of his calibre and I am privileged to be his student.

Further, I would like to thank Phillip Swanepoel for his assistance in creating computational tools during my project. His mathematical inclination has certainly given an edge to many of the tools developed.

Next, I would like to thank my good friend, Rentia Lourens, for her friendship, consistent laughter and dedication to both her and my own dissertation. "Just one more hour".

I would also like to further extend my sincere thanks to Spyros Lytras, Oscar A MacLean, David Robertson, Sergei L Kosakovsky Pond, Jordan D Zehr, Venkatesh Kumar, Michael J. Stanhope, Gordon Harkins and Ben Murrell, for their extensive knowledge in recombination, evolutionary medicine and biology as a whole.

Another instrumental figure and one who deserves recognition, is that of my brother, Steyn de Klerk, as well as all my other friends and family who have, perhaps unknowingly, contributed to the completion of this work by helping keep me sane throughout the many long days and late evenings.

Lastly, the completion of my dissertation would not have been possible without the support of my loving girlfriend, Kayla Koorts, and her profound belief in my ability to achieve. You're the soft spot in my world - thank you for everything.

Abstract

Recombination contributes to the genetic diversity found in coronaviruses and is known to be a prominent mechanism whereby they evolve. It is apparent, both from controlled experiments and in genome sequences sampled from nature, that patterns of recombination in coronaviruses are non-random and that this is likely attributable to a combination of sequence features that favour the occurrence of recombination breakpoints at specific genomic sites, and selection disfavouring the survival of recombinants within which favourable intra-genome interactions have been disrupted. Here we leverage available whole-genome sequence data for six coronavirus subgenera to identify specific patterns of recombination that are conserved between multiple subgenera and then identify the likely factors that underlie these conserved patterns. Specifically, we confirm the non-randomness of recombination breakpoints across all six tested coronavirus subgenera, locate conserved recombination hot- and cold-spots, and determine that the locations of transcriptional regulatory sequences are likely major determinants of conserved recombination breakpoint hot-spot locations. We find that while the locations of recombination breakpoints are not uniformly associated with degrees of nucleotide sequence conservation, they display significant tendencies in multiple coronavirus subgenera to occur in low guanine-cytosine content genome regions, in non-coding regions, at the edges of genes, and at sites within the Spike gene that are predicted to be minimally disruptive of Spike protein folding. While it is apparent that sequence features such as transcriptional regulatory sequences are likely major determinants of where the template-switching events that yield recombination breakpoints most commonly occur, it is evident that selection against misfolded recombinant proteins also strongly impacts observable recombination breakpoint distributions in coronavirus genomes sampled from nature.

Contents

Declaration	1
Acknowledgements	2
Abstract	3
Contents	4
List of Figures	6
List of Tables	7
Abbreviations	7
Chapter 1: Literature Review	11
1.1 Introduction	11
1.2 Coronavirus disease emergence	11
1.3 The Spike gene	13
1.4 Coronavirus recombination	14
1.5 Rationale	17
1.6 Aim and objectives	18
1.6.1 Aim	18
1.6.2 Objectives	18
Chapter 2: Detection of conserved recombination breakpoint hot- and cold-spots across multiple coronavirus subgenera	21
2.1 Introduction	21
2.2 Methods	21
2.2.1 Data collection	21
2.2.2 Recombination detection	21
2.2.3 Recombination breakpoint hot- and cold-spot tests	22
2.3 Results and discussion: recombination breakpoints	25
2.3.1 Recombination breakpoint detection	25
2.3.2 Breakpoint distributions in and around the S gene are consistent with recombination facilitating host adaptation and/or immune evasion	28
Chapter 3: Detecting underlying associations between sequence features and recombination breakpoint distributions	32
3.2 Methods	32
3.2.1 Data collection	32

3.2.2 Comparing recombination breakpoint counts between pairs of pre-defined genome regions	33
3.2.3 Testing for associations between GC content or pairwise sequence similarity and recombination breakpoint sites	34
3.2.4 Identification of potential transcriptional regulatory sequences	34
3.2.5 Protein folding disruption test	35
3.3 Results and Discussion: underlying associations and sequence features	36
3.3.1 Selection likely disfavours recombinants expressing Spike proteins with disrupted folds	37
3.3.2 Indirect evidence that selection against protein misfolding impacts observable breakpoint distributions throughout coronavirus genomes	38
3.3.3 ORF1a genome regions generally have lower breakpoint densities than other coding regions	40
3.3.4 Breakpoints tend to fall at sites with lower than average GC content	41
3.3.5 It is unclear whether sequence similarity directly influences the locations of recombination breakpoints	42
3.3.6 There is a strong association between recombination breakpoint locations and those of transcriptional regulatory sequences	45
Chapter 4: Limitations, future research and conclusion	48
4.2 Study limitations and future research	48
4.2.1 Data	48
4.2.2 Mutational change and coinfection	49
4.3 Conclusion	50
Appendix	51
References	68

List of Figures

Figure 1.1 The three instances of viral disease emergence. Page 12.

Figure 2.1 Variation across coronavirus genomes in the densities of detectable recombination breakpoints. Page 27.

Figure 2.2 Recombination region count matrices indicating genome regions that are most and least commonly transferred during detectable coronavirus recombination events. Page 31.

Figure 3.1 Regional variations in average pairwise sequence similarity (green) and GC content (blue) across coronavirus genomes. Page 44.

List of Tables

Table 3.1 Spike Protein Data Bank codes assigned to subgenus level from viruses most prominent within datasets. Page 33.

Table 3.2 SCHEMA derived estimates of recombination induced protein fold disruption comparing real and simulated recombination events, with rows in bold indicating subgenera with significantly fewer predicted structural impacts than would be expected under random recombination. Page 38.

Table 3.3 Comparison of detectable breakpoint numbers in non-coding regions and coding regions with rows in bold indicating subgenera with significantly more breakpoints in non-coding regions than would be expected under random recombination. Page 39.

Table 3.4 Breakpoint densities falling in the end 10% (5% each end) of genes vs the middle 90% of genes with rows in bold indicating subgenera with significantly higher numbers of detectable breakpoints in the ending 10% of genes than would be expected under random recombination. Page 39.

Table 3.5 Individual genes and sub-gene regions with significantly lower numbers of detectable breakpoints than would be expected under random recombination. Page 41.

Table 3.6 Associations between decreased GC content and detected recombination breakpoint sites with rows in bold indicating subgenera displaying average GC contents in the vicinity of breakpoint sites that are significantly lower than what would be expected under random recombination. Page 42.

Table 3.7 Association of breakpoint locations with higher/lower degrees of average pairwise sequence similarity with rows in bold indicating significant associations. Page 43.

Table 3.8 Associations between transcription regulatory sequence (TRS) sites and the locations of detected recombination breakpoints with p-values in bold indicating significant associations of TRS sites with higher breakpoint numbers. Page 46.

Table 4.1 Conserved patterns of recombination across various coronavirus subgenera. Page 50.

Abbreviations

Å	Ångström
ACE2	Angiotensin converting enzyme 2
AXL	Tyrosine-protein kinase receptor UFO
BCoV	Bovine Coronavirus
BP/bp	Breakpoint
CCoV	Canine Coronavirus
CCoV2b	Canine Coronavirus (type CCoV2b)
CI	Confidence interval
CoV	Coronavirus
CoVDB	Stanford Coronavirus Antiviral & Resistance Database
COVID-19	Coronavirus disease 19
CSV	Comma separated value
DNA	Deoxyribonucleic acid
DPP4	Dipeptidyl peptidase-4
ER	Endoplasmic reticulum
FCoV	Feline Coronavirus
FDR	False discovery rate
GC	Guanine-Cytosine
HCoV	Human Coronavirus
HIV	Human immunodeficiency virus
HMM	Hidden Markov Model
HR1	Heptad Repeat 1
HR2	Heptad Repeat 2
KB/kb	Kilobase

MAFFT	Multiple Alignment using Fast Fourier Transform
MERS	Middle East respiratory syndrome
MERS-CoV	Middle East respiratory syndrome–related coronavirus
N	Nucleocapsid
N/A	Not applicable
NCBI	National Center for Biotechnology Information
NT/nt	Nucleotide
NTD	N-terminal domain
OC43	Human Coronavirus OC43
ORF	Open reading frame
PCoV	Porcine coronavirus
PDB	Protein data bank
PEDV	Porcine epidemic diarrhea virus
P-val	Probability value
RBD	Receptor binding domain
RDP	Recombination detection program
RdRP	RNA-dependent RNA Polymerase
RNA	Ribonucleic acid
RNA-seq	RNA sequence
S	Spike
S1	Subunit 1
S2	Subunit 2
SARS	Severe acute respiratory syndrome
SARS-CoV-2	Severe acute respiratory syndrome Coronavirus 2
TB	Tuberculosis

TGEV	Transmissible gastroenteritis virus
TRS	Transcriptional Regulatory Sequence
TRS-B	Transcriptional Regulatory Sequence Body
TRS-L	Transcriptional Regulatory Sequence Leader
US	United States
w/o	Without
WHO	World Health Organisation

Chapter 1: Literature Review

1.1 Introduction

In December of 2019, an outbreak of pneumonia cases linked to a live animal market in Wuhan, China (Hubei Province) led local health officials to issue a warning to the Chinese Centre for Disease Control and Prevention and the World Health Organisation (WHO) on the grounds of a potential threat to human life. From 31 December to 3 January 2020, a total of 44 cases of patients with pneumonia of unknown aetiology were reported to the WHO. Several days later, on the 7th of January, the Chinese authorities had classified the new Coronavirus as belonging to the Sarbecovirus Subgenus of Coronaviridae, the same Subgenus as the SARS virus which infected 8,098 people worldwide between the years of 2002 and 2003 (Cherry 2004). By the 20th of January, the virus had been reported to be present in four countries, including China, Thailand, Japan and Korea (Countries where Coronavirus has spread - Worldometer 2021). On January 30th 2020 the World Health Organisation declared the outbreak to be a Public Health Emergency of International Concern. On the 11th of March 2020, WHO Director-General, Dr Tedros Adhanom Ghebreyesus, characterised SARS-CoV-2 (COVID-19) as a pandemic, due to the virus's alarming levels of spread and severity. At this point, there were 118,319 confirmed global cases of COVID-19, with 113 countries/areas outside of China affected (Situation Report-51 SITUATION IN NUMBERS total and new cases in last 24 hours 2020). As of 23 November 2021, 258,514,168 cases, with 5,177,407 deaths have been recorded globally, affecting 223 countries and territories (Countries where Coronavirus has spread - Worldometer 2021).

1.2 Coronavirus disease emergence

Coronaviruses are a family of vertebrate-infecting single-stranded, positive-sense RNA viruses with genomes ~27-32kb in length. The family has four genera - *Alphacoronavirus*, *Betacoronavirus*, *Gammacoronavirus* and *Deltacoronavirus* - each of which has been further subdivided into a number of subgenera such as *Pedacovirus* in the genus *Alphacoronavirus*, and *Merbecovirus*, *Embecovirus*, *Nobecovirus*, and *Sarbecovirus* in the genus *Betacoronavirus*, and *Igacovirus* in the genus *Gammacoronavirus* (Coronaviridae - Positive Sense RNA Viruses - Positive Sense RNA Viruses (2011) - ICTV 2011). Besides SARS-CoV-2, the *Sarbecovirus* member that causes COVID-19, there are four other known *Betacoronavirus* lineages and two known *Alphacoronavirus* lineages that either cause - or have caused - epidemiologically significant disease outbreaks in humans.

These disease-causing outbreaks generally occur when animal-infecting viruses are transmitted into humans and then transmitted between humans in sustained host-to-host transmission events. This process is one of the greatest known risks to global public health as is amply illustrated by the current

COVID-19 pandemic and by the fact that over 70% of human infectious diseases ultimately originated in wildlife (Haider *et al.* 2020).

Generally, viral disease emergence can be categorised into three phases which can lead to successful host switching (Figure 1.1). In phase 1, “spillover” infections occur when individuals in a new host species are infected by a virus, but these individuals fail to transmit the virus to other individuals of the new host species. In phase 2, successful transmissions, called outbreaks, occur between individuals of the new host species but transmission chains are not sustained, and the outbreak fades out. This outbreak fade-out is what happened in 2002-2004 with SARS when health authorities used a combination of testing, quarantining, and travel restrictions to contain the outbreak. A virus transitions to phase 3 of emergence when it achieves sustained transmission within its new host species (Figure 1.1 (Parrish *et al.* 2008)).

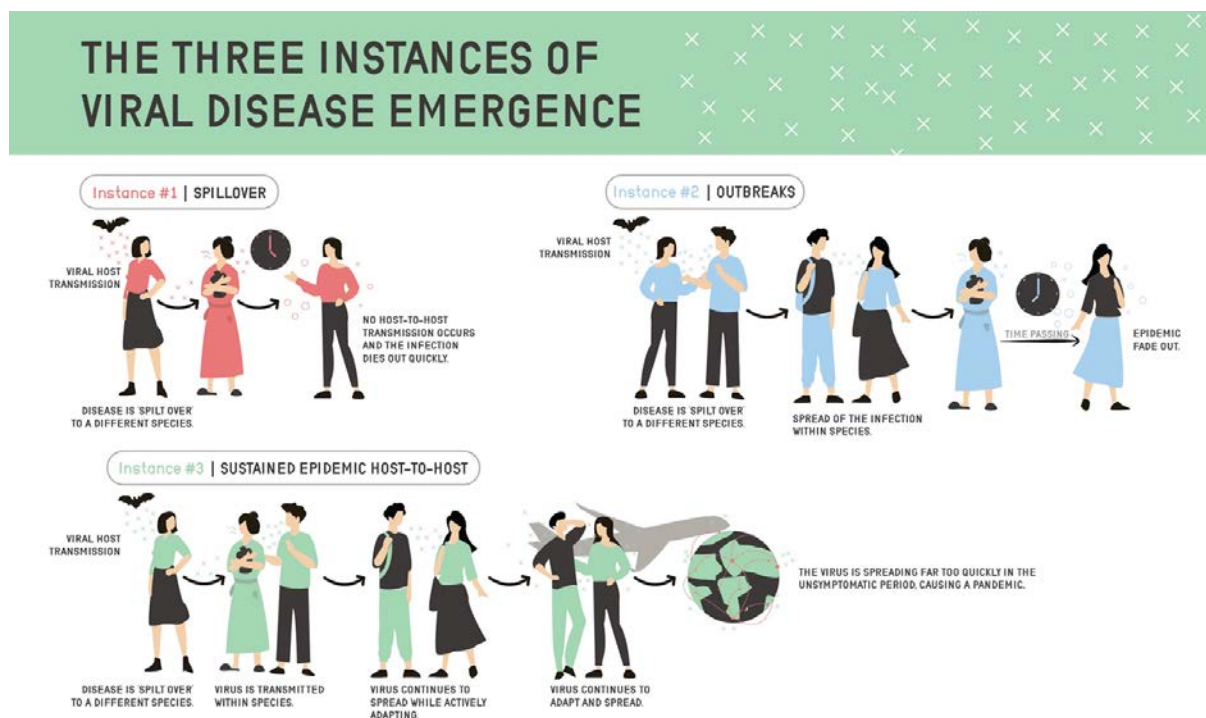


Figure 1.1. The Three Instances of Viral Disease Emergence. Instance #1: “Spillover” events. Disease spilt over from different host species but dies out. Instance #2: Outbreak events. Spreads but ends with epidemic fade out. Instance #3: Sustained Epidemic Host-to-Host transmission. Virus spreads quickly and causes a pandemic.

Over the course of the last 30 years, there have been several cross-species transmission events from multiple different coronaviruses causing lasting sustained endemic or epidemic host-to-host transmission events among both humans and animals. Human coronavirus 229E (HCoV-229E,

Alphacoronavirus) (Corman *et al.* 2018), canine coronaviruses (CCoV, *Alphacoronavirus*) (Guirao *et al.* 2013), feline coronaviruses (FCoV, *Alphacoronavirus*) (Paltrinieri *et al.* 2007), transmissible gastroenteritis viruses (TGEV, *Alphacoronavirus*) (Pensaert and Martelli 2016), bovine coronaviruses (BCoV, *Betacoronavirus*) (Wensman and Stokstad 2020), OC43 (HCoV-OC43, *Betacoronavirus*) and porcine coronaviruses (PCoV, *Deltacoronavirus*) (Ojkic *et al.* 2015; Islam *et al.* 2021) are all examples.

Three especially noteworthy human-infectious coronaviruses are useful for emphasizing the dangers which coronaviruses pose to society. The first is Middle East respiratory syndrome (MERS; a *Betacoronavirus* in the subgenus *Merbecovirus*), which emerged from Saudi Arabia in 2012 (Baharoon and Memish 2019) and continues to spill over from dromedary camels (Reusken *et al.* 2013; Corman *et al.* 2014), causing short-term human host-to-host transmission events but not continuing into stage three sustained epidemic host-to-host transmission (Memish *et al.* 2013, 2013).

The second is SARS (also a *Betacoronavirus* in the subgenus *Sarbecovirus*) which caused sustained host-to-host transmission and emerged in the wet food markets of Guangdong, China. SARS caused a devastating outbreak between the years of 2002 and 2003 where it infected over 8,000 humans and killed 774 (Cherry 2004). A combination of containment tactics implemented by various Asian governments were able to initially slow and ultimately contain the spread of the virus.

The third example is SARS-CoV-2, which occurred in late 2019. Initial SARS-CoV-2 sequences were shown to have high sequence similarity to those of SARS, sharing 79.60% genome-wide sequence identity (Jaimes *et al.* 2020).

Coronavirus genomes generally contain seven to ten genes, often with varying arrangements and compositions (Figure 2.1) (Lai 1996). The largest gene, ORF1ab, encodes multiple non-structural proteins which are involved in viral transcription, replication, proteolytic processing, modulation of host gene expression and the suppression of host immune responses (Emam *et al.* 2021). Directly downstream of ORF1ab in most known coronavirus genomes is the Spike (S) gene (although in *Embecoviruses*, for example, a haemagglutinin-esterase gene separates ORF1ab and the S gene).

1.3 The Spike gene

The extraordinary amount of variation that different coronaviruses display in terms of both tissue tropism and host range is largely due to the coronavirus Spike gene. Spike is the structural glycoprotein found on the outside of coronavirus particles that gives them their iconic crown-like protrusions. Spike binds to cell membrane receptors and mediates virus entry into host cells. It is considered a class I fusion protein in that it contains both a receptor-binding domain (called S1) and a domain for mediating the membrane fusion process (called S2) (White *et al.* 2008; Xia *et al.* 2020). The Spike of

SARS-CoV-2 has become a target in the development of vaccines and therapeutic drugs for COVID-19, due both to its importance during the viral infection cycle and to its being the primary target of host immune responses (Krumm *et al.* 2021).

1.4 Coronavirus recombination

Recombination is a phenomenon common to most positive sense RNA viruses and is a major contributor to both coronavirus variability and rapid coronavirus evolution (Franzo *et al.* 2020). The continuous adaptation of living organisms to changing environments requires that biological species encompass a broad degree of inter-individual genetic variation, since this variation is the raw material moulded by natural selection into the future forms of a species (Aaziz and Tepfer 1999). In the world of RNA viruses to which the coronaviruses belong, high intra-species genetic variability is generally observed, primarily due to three main forces: mutation, reassortment (in viruses with segmented genomes, such as influenza), and recombination (Domingo and Holland 1997).

Although coronaviruses generally have lower mutation rates relative to other single-stranded RNA viruses (Denison *et al.* 2011; Jaroszewski *et al.* 2021), coronavirus species are nevertheless characterized by high degrees of genetic diversity (Liu *et al.* 2017). Much of this genetic diversity is likely generated and maintained by high rates of within-species (Dudas and Rambaut 2016; Su *et al.* 2016; Anthony *et al.* 2017; Forni, Cagliani and Sironi 2020) and between-species genetic recombination (Wesley 1999; Decaro *et al.* 2015; Wang *et al.* 2015, 2020).

The first credible reports of recombination in coronaviruses were made in the mid-to-late 1980s and focused on mixed *in vitro* and *in vivo* infections of different Murine mouse hepatitis virus strains (Makino *et al.* 1986; Keck *et al.* 1988a, 1988b; Banner and Lai 1991). By the year 2000, comparative analyses of coronavirus genomes sampled from natural infections had yielded convincing evidence that recombination, particularly between divergent coronaviruses within individual subgenera, is a major contributor to coronavirus evolution (Kusters *et al.* 1990; Wang, Junker and Collisson 1993; Jia *et al.* 1995; Lee and Jackwood 2000). For example, complex recombinant histories are evident between the Alphacoronavirus-1 species (a canine coronavirus that primarily infects dogs; however, see recent exceptions of canine CoV infections in humans: (Lednicky *et al.* 2021; Vlasova *et al.* 2021; Zehr *et al.* 2021)), transmissible gastroenteritis virus, (infecting pigs) and feline coronavirus (infecting cats) (Herrewegh *et al.* 1998; Decaro *et al.* 2009).

For recombination to occur in RNA viruses there are two main conditions that need to be met: (1) two viruses need to co-infect a single host cell (Barr and Fearn 2010) and (2) a mechanism for recombination needs to exist. The most common mechanism of recombination in coronaviruses (and

many other RNA viruses too) is known as copy-choice, where a viral RNA dependent RNA polymerase (RdRp) is interrupted during replication, drops off the RNA template that it was copying, and re-engages with a different RNA template at a homologous position before resuming replication (Cheng and Nagy 2003). Such template switches during replication yield recombinant daughter genomes with different regions of sequence being derived from two different “parental” genomes. The genome sites at which template switches occur are referred to as recombination breakpoints. The reason that single-stranded negative-sense RNA viruses such as influenza A virus are incapable of frequent recombination, despite very frequent genome segmented reassortment patterns, is because template switching during replication is mechanistically difficult (but not impossible (Simon-Loriere and Holmes 2011)) for (Li *et al.* 2004; Dong *et al.* 2011) these viruses.

This copy-choice type of recombination is called homologous recombination (hereafter simply referred to as recombination), because the transfer sites in parent and offspring sequences are all genetically homologous (i.e. they are derived from the same common ancestral sequence). Recombination involving sequence transfers between sites that are not derived from the same common ancestral sequence in the parent and recombinant sequences is called non-homologous recombination (Barr and Fearn 2010).

Recombination likely provides coronaviruses with more evolutionary options than would be available to them by mutation alone (Cramer *et al.* 1998; Simon-Loriere *et al.* 2009). While it is expected that many newly arising mutations within genetically compact viral genomes (such as those of coronaviruses) will have negative fitness consequences, so too will many of the recombination events that occur between genetically divergent genomes (Drummond *et al.* 2005). By transferring pieces of genomes into genomic backgrounds with which they did not coevolve, recombination will frequently run the risk of disrupting favourable coevolved intra-genome interactions (commonly referred to as epistatic interactions). Examples of favourable coevolved intra-genome interactions that could be disrupted by recombination include those between nucleotides (nt) that base-pair to form biologically functional genomic secondary structures, those between pairs of amino acids that interact to mediate protein folding, those between the binding domains on protein surfaces that mediate multi-protein complex formation, and those between sequence-specific nucleic acid binding domains and nucleotide sequence motifs that mediate gene regulation and genome replication (Martin *et al.* 2005b). However, since recombination generally occurs between fully functioning genomes, the range of potential negative fitness consequences of recombination are, in general, expected to be less extreme than those that might occur due to newly arising mutations (Drummond *et al.* 2005).

Recombination can potentially also be a mechanism of genome repair when it occurs between closely related viruses as it helps defend against the accumulation of mildly deleterious mutations within genomes that, in high enough numbers, might otherwise have serious fitness consequences (Muller 1964; Woo *et al.* 2010; Hussin *et al.* 2015; Goldstein *et al.* 2021). The slow-accumulation of deleterious mutations in the absence of recombination, known as Muller's ratchet, occurs when mildly deleterious mutations build up in the population and irreversibly ratchets the mutation load of genomes upwards until non-recombining species eventually become extinct. It is important to stress though, that although Muller's ratchet has been shown to occur in populations of experimental RNA viruses (Chao 1990; Poon and Chao 2004), it has also been debated whether the population sizes of RNA viruses are small enough for it to occur frequently since the ratchet mostly functions during the early stages of viral population expansion when population size is smaller and extinction is more likely. The sizes of RNA virus populations both within individual infected hosts and globally distributed between thousands of different infected individuals provides RNA virus species with a degree of population-scale genetic robustness, that also helps guard against the accumulation of deleterious mutations which may be the reason that some virus species show very little evidence of recombination (Elena *et al.* 2006).

The potential interchangeability of the S genes to different coronaviruses has also been proposed, based on patterns of recombination seen in circulating lineages (de Haan *et al.* 2008; Graham and Baric 2010). Specifically, it is apparent that (1) the S gene is commonly transferred as a complete unit between genomes by natural genetic recombination; (2) the 5' or 3' part of the S gene (S1 or S2 domains) is frequently conferred as units by recombination; and (3) particular subdomains of the S1 and S2 domains (such as the N-terminal domain or receptor binding domain) are frequently transferred as units by recombination.

1.5 Rationale

Although frequent recombination events have been observed in coronaviruses, little is known about how conserved recombination patterns are across the various coronavirus genera and subgenera.

As recombination processes contribute substantially to the large degrees of genetic diversity seen among coronaviruses found in nature, it is important to both identify conserved patterns of recombination, and determine the underlying biological processes responsible for these patterns.

Studying recombination patterns across various different coronaviruses can improve our understanding of why some regions of coronavirus genomes are more commonly involved in recombination than others and, by extension, aid in the development of therapeutics or vaccines by locating sites of viral genomes that are constrained to not recombine. Therapeutics or vaccines targeting regions which do not commonly recombine would also be less prone to fail due to viruses using recombination to “swap in” versions of these genome regions that are less effectively targeted. Further, knowing the regions of coronavirus genomes that are most prone to recombination will permit us to approximate the relative risks of future recombination events transferring problematic viral traits (such as human cell tropism as is mediated by a well-defined region of the S gene) between potentially dangerous coronaviruses. Lastly, the study contributes to consolidating previous information, as well as aiding future studies in finding common sequence features across current, and by extension, potentially also future, coronavirus genomes that contribute to increased/decreased recombination frequencies in different genome regions.

Here, computational methods are constructed and used, both to locate recombination hot- and cold-spots across various coronavirus subgenera (Chapter 2), and to determine the underlying causes of non-random recombination breakpoint distributions in coronavirus genomes (Chapter 3).

1.6 Aim and objectives

1.6.1 Aim

The aim of this study is to investigate both (1) the recombination breakpoint hot- and cold-spots across multiple coronavirus subgenera, and (2) determine underlying associations between hot and cold-spots and sequence features.

1.6.2 Objectives

I. Collect data

Coronavirus sequence data of various subgenera is to be downloaded from all accessible public databases. In order to keep downstream deletions to a minimum, due to the large variety and quality of sequences on public databases; filters should be applied at the time of downloading.

II. Manipulate data

Downloaded sequence data should next be combined to the subgenus level and have all sequences sharing more than 99% sequence similarity removed. 3D structure files, for analysis with the SCHEMA (Voigt *et al.* 2002) method, are to be downloaded from RCSB (<https://www.rcsb.org/structure>), for all coronavirus proteins for which high resolution atomic coordinate data is available.

III. Detect recombination

Recombination, which forms the basis of this study, is to be detected using the most up-to-date version of RDP5 (Martin *et al.* 2021), a powerful computer program capable of detecting and characterising recombination events using several recombination detection methods in its fully exploratory automated scan.

IV. Test for recombination hot- and cold-spots

Recombination breakpoint distribution maps, and region count matrices should both be created for each of the analysed coronavirus subgenera. These can then be used to find clusters of both hot- and cold-spots of recombination. By creating a stacked figure of the recombination breakpoint distribution maps, recombination breakpoint patterns could easily be observed across subgenera.

V. Comparing recombination breakpoint counts between pairs of pre-defined genomic regions

Recombination breakpoint densities falling at (1) areas of the genome which do not code for proteins vs areas coding for proteins (non-protein-coding sites vs protein-coding sites), (2) a portion of the 5' and 3' ending parts of individual protein-encoding genes vs the centerpiece of these regions, (3) genome sites encoding a particular protein vs those encoding all other proteins within the genome and (4) sites within a specified number of nucleotides of a regulatory sequence vs those in the remainder of the genome, should all be counted, tabulated and investigated.

VI. Testing for associations between GC content and recombination breakpoint sites

As guanine+cytosine (GC) content may influence the positions of detected recombination breakpoints in coronaviruses (as they do in several other viruses), it is possible that we can expect to see correlations between GC content and detected breakpoint patterns across a variety of multiple coronavirus subgenera. To test for this, a modified version of the breakpoint clustering hot- and cold-spot test is to be used to test for associations between breakpoint sites and GC content.

VII. Testing for associations between pairwise sequence similarity and recombination breakpoint sites

Similarly to GC content, it is also possible that pairwise sequence similarity plays a role in detected recombination breakpoint positions across subgenera of coronaviruses. For this reason, sequence similarities within individual coronavirus subgenera should be calculated according to the association of breakpoint locations with higher/lower degrees of average pairwise sequence similarity, and comparatively analysed.

VIII. Identification of potential transcriptional regulatory sequences

As transcriptional regulatory sequences (TRS), a sequence feature, has previously been shown to be strongly correlated with detected recombination breakpoints in coronaviruses (Yang *et al.* 2021), it is imperative that methods are implemented for inferring transcriptional regulatory sequence body (TRS-B) sites. These detected sites should be tested for significant association with detected recombination breakpoint clusters and conservation across multiple coronavirus subgenera.

IX. Protein folding disruption test

Testing for whether observed recombination events were less disruptive than would be expected if recombination breakpoints were randomly distributed, the SCHEMA test should be used, as per its implementation in RDP5, using downloaded PDB, 3D protein structure files.

Chapter 2: Detection of conserved recombination breakpoint hot- and cold-spots across multiple coronavirus subgenera

2.1 Introduction

Here we analyse patterns of recombination evident in whole-genome datasets drawn from one *Alphacoronavirus* subgenus, one *Gammacoronavirus* subgenus and four *Betacoronavirus* subgenera. We confirm previous reports that natural recombination between genetically divergent coronaviruses is common and find strong evidence that detectable recombination breakpoint sites are not randomly distributed across coronavirus genomes. Specifically, we demonstrate the likely occurrence of breakpoint hot- and cold-spots, some of which are conserved across multiple coronavirus groups.

2.2 Methods

2.2.1 Data collection

Two major filters were applied when downloading all publicly available near full-length genomic sequences for viruses in six well-sampled coronavirus subgenera (*Igacovirus*, *Embecovirus*, *Merbecovirus*, *Nobecovirus*, *Pedacovirus* and *Sarbecovirus*); (1) sorting by maximum/minimum genome length +2,000/-2,000 nucleotides according to the reference genome of the most prominent virus in each subgenus and (2) checking the checkbox for nucleotide completeness (when available). Sequences were downloaded from the NCBI Virus (Hatcher *et al.* 2017), CNCB (Song *et al.* 2021), and CoVDB (Zhu *et al.* 2021) databases between February and May of 2021. Each of the six subgenus-level datasets was aligned according to the subgenus level, with MAFFT, using default settings (Kato and Standley 2013). All but one sequence in groups of sequences sharing more than 99% nucleotide sequence identity was removed to yield datasets for recombination analysis containing 412 genome sequences in *Pedacoviruses* (belonging to *Alphacoronavirus* genus), 191 in *Merbecoviruses*, 16 in *Nobecoviruses*, 158 in *Sarbecoviruses*, 181 in *Embecoviruses* (all four belonging to *Betacoronavirus* genus) and 410 in *Igacoviruses* (belonging to *Gammacoronavirus* genus), all sharing $\geq 75\%$ similarity (Supplementary Table 1; Appendix).

2.2.2 Recombination detection

Recombination was detected and analysed using the most recent version of RDP5 (Martin *et al.* 2021) available at the time of analysis (version 5.16), with default settings except that sequences were treated as linear due to coronaviruses being unsegmented, linear genomes (Konings *et al.* 1988). Although several other programs capable of detecting recombination exist (Drouin *et al.* 1999; Shaik *et al.* 2015), RDP5 is a single, highly automated tool capable of detecting and characterising

recombination events using multiple tools with its fully exploratory automated scan. It can also be used without prior user identification of non-recombinant sequences. Each of the six coronavirus datasets were analysed for recombination using the aforementioned fully exploratory automated scan with the RDP (Martin and Rybicki 2000), GENECONV (Sawyer 1989), and MaxChi (Maynard Smith 1992) methods to detect recombination signals (i.e these were used as “primary scanning methods”), and the Bootscan (Martin *et al.* 2005a), Chimaera (Pettersen *et al.* 2004), SiScan (Gibbs, Armstrong and Gibbs 2000) and 3Seq (Lam, Ratmann and Boni 2018) methods to verify the signals (these latter four methods being used as “secondary scanning methods”). From among the individual recombination signals that were each detectable by four or more of these methods, RDP5 refined the positions of detected recombination breakpoints using a hidden Markov model (HMM) based recombination breakpoint detection method, BURT, which is used by RDP5 to both detect the locations of recombination breakpoints and determine identified approximate confidence intervals with each associated estimated position (described in detail in the RDP manual <http://web.cbio.uct.ac.za/~darren/RDP4Manual.pdf>) and determined a plausible near minimal subset of unique recombination events that would be needed to account for all of the detected recombination signals. Each of the unique recombination events detected by RDP5 in each of the six analysed coronavirus subgenera datasets was characterized by: (1) a 5' and 3' pair of maximum likelihood breakpoint locations and their associated probability distributions, (2) a list of one or more sequences carrying evidence of the recombination event (multiple sequences can have evidence of the same recombination event if the event occurred in a common ancestor), and (3) a list of analysed sequences that are closely related enough to the actual parents of the recombinant that they could be used as proxies for the actual parents to detect the recombination events. The overall-recombination patterns in the six subgenera datasets were visualized using recombination region count matrices produced using RDP5. These matrices are an overview of the numbers of detected recombination events that separated individual genome sites from all other genome sites (i.e a visual indication, with colours indicating the number of times recombination events, have separated different parts of the sequences being analysed from one another by recombination events).

2.2.3 Recombination breakpoint hot- and cold-spot tests

For each of the subgenera, a recombination breakpoint distribution map was constructed from the lists of 5' and 3' breakpoint probability distributions associated with each detected recombination event. This was done by sliding a 200 nt window, one nucleotide at a time, along the full length of the analysed alignment, summing the probabilities of all identified breakpoints falling within the window, and plotting these counts at the nucleotide coordinate at the centre of the window. This data, generated by RDP5, was exported in CSV format and used in a custom R script (Supplementary Data

1-2; Appendix) to generate graphs for easier analysis of potentially conserved patterns of recombination. To do this, a custom graph X-axis limit was set to ~33kb, (the length of the Embecovirus alignment analysed which was the longest of the six datasets). The script takes in three files at a time, specifically, a (1) comma separated value (CSV) file containing recombination breakpoint positions, a (2) file containing ORF/gene coordinates, and a file containing (3) both upper and lower 99% and 95% confidence intervals (CI) of recombination breakpoint densities. This data, combined with certain modifiable parameters within the script, such as (1) gene map height randomisation, (2) X and Y axis visibility toggles (useful for appending labels to the figure which will be stacked at the very bottom), (3) gene map plot, including gene/protein names and (4) simple color modification, all enable the script to produce publication-ready graphs of recombination breakpoint distribution plots from RDP5. For our analysis, only the 99% CI was plotted.

Permutation tests for recombination breakpoint clustering implemented in RDP5, account for both (1) site-to-site variations in the detectability of recombination breakpoints and recombination events by the various recombination detection methods employed, and (2) the underlying uncertainties in individual breakpoint position estimates, as determined using the HMM-based BURT method (described previously). These permuted tests were then used to identify recombination breakpoint clustering patterns that varied significantly from expectations under random recombination. Briefly, this test involved: (1) randomly shuffling the breakpoint locations of each of the observed recombination events to maintain the spacing between 5' and 3' breakpoint pairs with respect to the numbers of polymorphic nucleotide sites separating the breakpoint pairs within the triplets of analysed sequences, used to detect the recombination event; (2) ensuring that in instances where individual sequences contained evidence of multiple independent recombination events, the regions bounded by 5' and 3' breakpoint pairs for those events did not overlap to a greater or lesser degree those observed in the actual recombinants (i.e. the spacings of all the 5' and 3' breakpoint locations of all overlapping events within a single sequence were maintained); (3) ensuring that in instances where breakpoints were flagged as having undetermined positions in the actual dataset (such as breakpoints called at the start/end of the alignment or at sites that were overprinted by subsequent recombination events), these were excluded from breakpoint counts; (4) making recombination breakpoint distribution maps for each permuted dataset using the exact same approach as that used for the actual dataset; and (5) identifying unusually high or low degrees of breakpoint clustering in the actual dataset as those window coordinates where the breakpoint probability sums of the actual dataset fell outside the bounds of those determined at that coordinate for 99% of the permuted datasets. With this test, unusually high degrees of breakpoint clustering (i.e. greater than 99% of the permuted datasets at a given genome site) are suggestive of recombination hot-spots, whereas

unusually low degrees of clustering (i.e. lower than 99% of the permuted datasets at a given genome site) are suggestive of recombination cold-spots.

It is important to stress, that this breakpoint clustering test is not conservative; given the lengths and degrees of diversity of the analysed coronavirus genomes, it is expected that one or two hot-spot-like clusters of breakpoints would be detectable in each of the datasets even under completely random recombination (Lytras *et al.* 2021). We, therefore, referred to hot-spots detected by this test in individual datasets as “potential hot-spots” and required that for a particular genome site to be defined as an actual statistically-supported hot-spot, potential hot-spots needed to be detectable at a homologous site in two or more of the different analysed subgenus datasets.

2.3 Results and discussion: recombination breakpoints

2.3.1 Recombination breakpoint detection

Using a combination of recombination detection methods implemented in RDP5, we identified 416 unique recombination events in the *Pedacovirus* dataset, 255 in *Embecovirus*, 65 in *Merbecovirus*, 107 in *Nobecovirus*, 282 in *Sarbecovirus*, and 1703 in *Igacovirus*. It is important to also note that the highly variable numbers of detected recombination events between datasets should not be considered evidence that the viruses in some subgenera recombine more than others. Rather, the variable numbers reflect differences in both the numbers of analysed sequences in each dataset (e.g. the *Pedacovirus* dataset had the most sequences) and the genetic diversity of the sequences in the different datasets. Within the individual analysed datasets the mean pairwise similarity was 98% for the *Pedacoviruses*, 92% for the *Merbecoviruses*, 75% for the *Nobecoviruses*, 86% for the *Sarbecoviruses*, 78% for the *Embecoviruses* and 89% for the *Igacoviruses* (Supplementary Table 1; Appendix).

To visualise the recombination breakpoints associated with these events in each subgenus, recombination breakpoint distribution plots (Figure 2.1), and recombination region count matrices (Figure 2.2) were constructed. The breakpoint distribution plots revealed clusters of breakpoints that were either more or less densely packed with recombination events at individual genome sites than those observed at corresponding sites in 99% of permuted datasets where recombination breakpoint positions were randomly distributed (Figure 2.1). Potential recombination hot-spots were detected in all of the analysed subgenera (indicated by red shading in Figure 2.1) and recombination cold-spots in three of them; namely in *Pedacoviruses*, *Sarbecoviruses* and *Igacoviruses* (indicated by blue shading in Figure 2.1, or blue vertical lines in Figure 2.2).

In all the subgenera other than *Embecovirus* and *Merbecovirus*, potential recombination hot-spots were detected within 300 nucleotides of the 5' end of the genome. This non-coding region is upstream of ORF1ab, where the transcription and replication initiation sites are. These initiation sites prime the transcription of subgenomic mRNAs and contain extensive secondary structures that are partially conserved amongst the viruses belonging to a given coronavirus genus (reviewed in (Yang and Leibowitz 2015) (Siegfried *et al.* 2014; Manfredonia *et al.* 2020)). In various other viruses such as HIV, recombination breakpoints tend to colocalize with highly structured genome regions (Simon-Loriere *et al.* 2010) and it is therefore plausible that recombination hot-spots detected at the 5' end of *Pedacovirus*, *Nobecovirus*, *Sarbecovirus* and *Igacovirus* genomes might also be attributable to secondary-structure induced template-switching during replication.

Multiple other potential recombination hot-spots were detected near the boundaries of various different genes in the 3' genome regions of *Sarbecoviruses*, *Nobecoviruses*, and *Igacoviruses*. In *Sarbecoviruses*, four potential recombination hot-spots were detected in the 3' genome regions, between the M gene and ORF6, between ORF7AB and ORF8AB, and between ORF8AB and N. In *Nobecoviruses*, potential recombination hot-spots were detected in the 3' genome regions, between the E gene and the M gene, and near the centre of the N gene. In *Igacoviruses* there were several potential hot-spots in the 3' genome region, the clearest of which fell towards the 3' end of N.

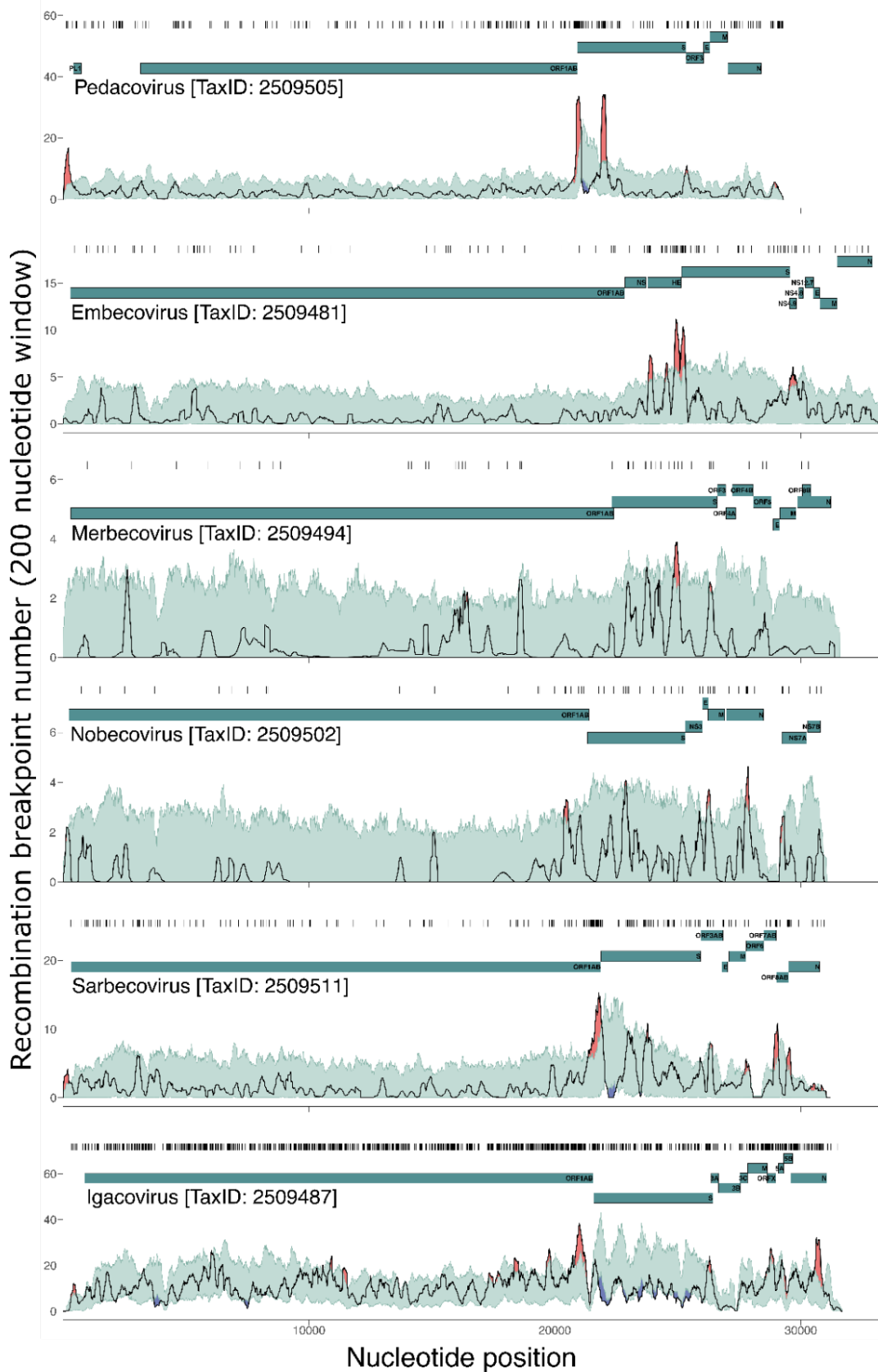


Figure 2.1. Variation across coronavirus genomes in the densities of detectable recombination breakpoints. All detected breakpoint positions are indicated directly above each graph with black vertical lines. A gene-map is shown as teal-coloured lines. The light green areas indicate 99% bounds of expected degrees of breakpoint clustering under random recombination. Areas where the black lines (breakpoint numbers per 200 nucleotide window) have emerged above the green areas, are considered potential recombination hot-spots, and are marked in red. Areas where the black lines drop below the green areas are considered potential recombination cold-spots, and are marked in blue.

2.3.2 Breakpoint distributions in and around the S gene are consistent with recombination facilitating host adaptation and/or immune evasion

Most noteworthy of all the detected potential breakpoint hot-spots were those falling within 800 nucleotides upstream of the S gene start codon in all subgenera other than the *Merbecoviruses*. The conserved arrangement of recombination breakpoint clusters in relation to the S gene likely underlies the observation that, according to our analyses, the S gene has been frequently transferred in its entirety during recombination events in *Igacoviruses*, *Sarbecoviruses* and *Embecoviruses* (note red diagonals associated with the S genes of these subgenera in Figure 2.2). However, in all analysed coronavirus groups other than the *Embecoviruses* and *Igacoviruses*, potential recombination hot-spots were also detected within the 5' half of the S gene (S1 domain), further suggesting that the 3' half of the gene (S2 domain) is the portion that is most commonly transferred during recombination events as a complete module (note the red diagonals associated with the 3' part of the S gene in Figure 2.2). The locations of the detected recombination hot-spots in, and immediately adjacent to, the S gene suggest that either the complete S gene or its 3' half, has been frequently transferred during recombination. This is not to say, that exceptions to this undoubtedly occur, including for example a likely instance in Alphacoronavirus-1 (a group not included in this current analysis, because of lower complete genome numbers) involving a section of the S2 domain of the newly described canine coronavirus HuPn-2018 (Zehr *et al.* 2021).

The Spike proteins that are encoded by the S gene are composed of an amino-terminal subunit 1 (S1) and a carboxyl-terminal subunit 2 (S2) (Wrapp *et al.* 2020). The S1 contains the N-terminal domain (NTD) and a receptor-binding domain (RBD) which mediates the binding of viral particles to host cell surface receptors. Different coronaviruses bind to different receptors. For example, the *Merbecovirus*, MERS-CoV, binds dipeptidyl peptidase-4 (DPP4), the *Pedacovirus*, PEDV, binds aminopeptidase N, and the *Sarbecoviruses* SARS-CoV and SARS-CoV-2 bind to angiotensin-converting enzyme 2 (ACE2) (Yeager *et al.* 1992; Li, Ge and Li 2007; Belouzard *et al.* 2012; Raj *et al.* 2013; Reusken *et al.* 2016; Wan *et al.* 2020). It is also likely that in some coronaviruses the NTD of Spike also interacts with cell surface receptors. For example, the NTD of the SARS-CoV-2 Spike interacts with the tyrosine-protein kinase receptor UFO (AXL) which appears to function as a co-receptor for human cell entry (Wang *et al.* 2021). The S2 subunit contains a heptad repeat region (including subregions HR1 and HR2) which mediate the fusion of the virion envelope with the host cell membrane during viral entry (Liu *et al.* 2004; Cui, Li and Shi 2019).

Being responsible for receptor binding and cellular entry, the evolution of the S gene is therefore key to host adaptation. It may be beneficial for coronaviruses to exchange either entire S genes, S1 subunit

encoding portions of S genes, or smaller subdomains within the N-terminal domains of S1 during recombination, both because Spike is the main target of neutralising antibodies (Ou *et al.* 2020) and because the S gene is the main determinant of host species and host cell-type specificity (Lu, Wang and Gao 2015). Although recombination frequently transfers the entire S1 encoding region of the gene, it is not uncommon in particular groups of viruses for it to transfer smaller subsections of the S1 (as can be seen with the red diagonals associated with the S genes of *Pedacoviruses* and *Merbecoviruses* in Figure 2.2). In the *Alphacoronaviruses*, for example, recombination has involved transfer of the 5' half of the NTD of S1 from transmissible gastroenteritis virus into canine coronavirus (type CCoV2b) (Decaro *et al.* 2009; Licitra, Duhamel and Whittaker 2014).

The S gene is also the only gene in which recombination cold-spots were detected in our breakpoint distribution analyses. Most noteworthy is that the 3' 500 nucleotides of the S gene is the site of a conserved cold-spot detected in the *Igacoviruses*, *Pedacoviruses* and *Sarbecoviruses*. In the *Igacoviruses*, the coronavirus group with the richest full genome dataset in terms of both numbers of analysed sequences and their diversity, and within which the highest numbers of recombination breakpoints were detected (n=1703), our power to detect recombination cold-spots was greatest. Accordingly, recombination cold-spots were additionally detectable in the region of the S gene encoding the receptor-binding domain and dispersed throughout the 3' half of the gene encoding the S2 subunit.

This arrangement of recombination cold-spots suggests that either basal recombination rates are suppressed within the NTD and S2 encoding regions of the S gene, or that recombination breakpoints falling within these regions tend to yield S genes that encode defective chimaeric Spike proteins. The NTD encoding region of the S gene is among the most genetically variable regions of coronavirus genomes and this alone might explain the relative absence of recombination breakpoints near the 5' end of the S genes of *Igacoviruses*, *Nobecoviruses*, *Pedacoviruses*, and *Sarbecoviruses* (Archer *et al.* 2008; Boni *et al.* 2020). Similarly, the S2 encoding region of the S gene also tends to be more variable than most other coronavirus genome regions. However, the S2 subunit of Spike also contains multiple coevolved intra-protein amino acid interactions that are crucial for the cell-fusion functions of Spike (Bosch *et al.* 2003; Tang *et al.* 2020). It is also plausible, therefore, that the relative absence of detectable recombination breakpoints in the 3' half of the S gene might be because recombinants carrying breakpoints falling within this region commonly express defective Spike proteins, which then in turn are incapable of binding to receptors on the cell surface of their normal hosts, and quickly die out in nature. In this regard, the S2 encoding region of the S gene may be a functional module which, while tending to retain its functionality when transferred by recombination as a complete unit into divergent genomic backgrounds (Wege *et al.* 1998), might be highly sensitive to recombination-

induced disruptions of co-evolved amino acid interactions within S2 whenever recombination breakpoints fall within its boundaries.

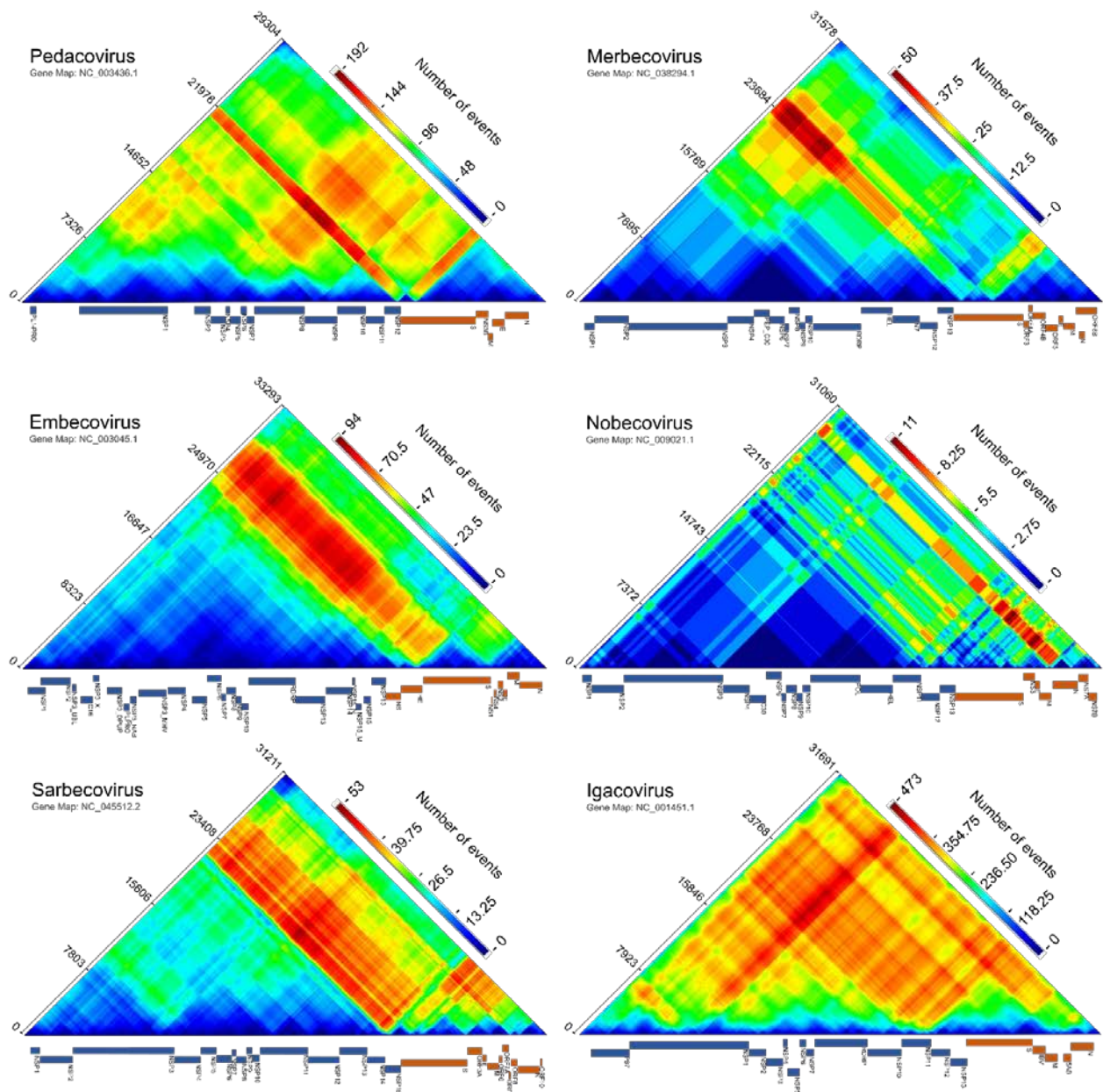


Figure 2.2. Recombination region count matrices indicating genome regions that are most and least commonly transferred during detectable coronavirus recombination events. Unique recombination events for six coronavirus subgenera, mapped onto recombination region count matrices based on determined breakpoint positions. Each cell in the matrix represents a pair of genome sites with the colours of cells indicating the numbers of times recombination events separated the represented pairs of sites. Reference sequence gene maps of the most prevalent virus in each subgenus were obtained from the NCBI nucleotide database (<https://www.ncbi.nlm.nih.gov/nucleotide>) and are plotted alongside each matrix. Nucleotide positions are plotted according to full analysed nucleotide sequence alignments (Supplementary material). Genome maps indicate the coding regions of individual protein products. Non-structural proteins encoded by ORF1ab are indicated in blue and other genes are indicated in orange.

Chapter 3: Detecting underlying associations between sequence features and recombination breakpoint distributions

3.1 Introduction

Here, we find detectable underlying associations across multiple different coronavirus subgenera between detected recombination breakpoint locations and various sequence features that might impact the mechanistic predisposition of certain genome sites to recombine more than others (such as decreased guanine-cytosine content, and the locations of transcriptional regulatory sequences). We also find evidence across multiple subgenera that selection differentially favours the survival of recombinants based on the genome sites at which breakpoints occur (such as at the edges of genes or in intergenic regions relative to the middle portions of genes) and look for evidence of events being less disruptive of protein folding than would be expected if recombination breakpoints were randomly distributed.

3.2 Methods

3.2.1 Data collection

Predominantly for analyses involving the SCHEMA method, 3D protein structure data was downloaded from RCSB (<https://www.rcsb.org/structure>) throughout June of 2021. Protein structure data was located according to one or two of the most prominent viruses within each subgenus dataset (for example, MERS-CoV in *Merbecovirus*, SARS-CoV-2 in *Sarbecovirus*, HKU1 or OC43 in *Embecovirus* or avian infectious bronchitis virus in *Igacovirus*; Table 3.1). It is important to note that although initially all PDB (protein data bank) files available for each gene of each virus in each subgenus were downloaded (Supplementary Tables 4-9; Appendix), only the S gene had sufficient data for analysis of protein folding disruption. It is for this reason that future research (assuming an increase in future PDB availability) is essential for discovering more recombination events potentially less disruptive of protein folding than would be expected if recombination breakpoints were randomly distributed.

Table 3.1: Spike Protein Data Bank codes assigned to subgenus level from viruses most prominent within datasets

Subgenus	Virus	PDB Code(s)
<i>Pedacovirus</i>	Porcine epidemic diarrhea virus	6VV5
<i>Embecovirus</i>	Human coronavirus HKU1/Human coronavirus OC43	5I08/6OHW
<i>Merbecovirus</i>	Middle East respiratory syndrome	5X5C
<i>Nobecovirus</i>	Severe acute respiratory syndrome coronavirus 2	6VYB*
<i>Sarbecovirus</i>	Severe acute respiratory syndrome coronavirus 2	6VYB
<i>Igacovirus</i>	Avian infectious bronchitis virus	6CV0

*For viruses with insufficient data at time of study, SARS-CoV-2 data was substituted in.

3.2.2 Comparing recombination breakpoint counts between pairs of pre-defined genome regions

We used a version of the breakpoint clustering hot- and cold-spot test that compared observed breakpoint numbers in two preselected groups of sites in an analysed alignment (Lefeuvre *et al.* 2009). Since the original recombination breakpoint distribution test determined whether the numbers of breakpoints observed in 200 nt sliding windows were greater or lesser than chance under random recombination, the test relied on the detection of sufficient breakpoints for statistically implausible clusters of breakpoints to emerge. As the number of detected recombination breakpoints varied widely between the different coronavirus datasets (ranging from 65 for the *Merbecoviruses* and 1703 for the *Igacoviruses*; Supplementary Table 1; Appendix), the power of the test varied substantially, largely due to the low numbers of detected recombination breakpoints in each sliding window for some of the datasets. In an adapted version of the test, we partitioned the sites in each of the six datasets into two large subsets and directly compared observed breakpoint numbers in each of the site subsets to those expected under random recombination. More specifically, we compared densities of breakpoints falling at: (1) non-protein-coding sites vs protein-coding sites, (2) the beginning and ending 5% of sites within individual protein-encoding regions vs the middle 90% of these regions (in the case of ORF1ab we defined protein encoding-regions as those encoding individual post-translational protein cleavage products), (3) genome sites encoding a particular protein vs those encoding all other proteins within the genome and (4) sites within a specified number of nucleotides (2, 9, 21, 46) of a transcriptional regulatory sequence vs those in the remainder of the genome.

3.2.3 Testing for associations between GC content or pairwise sequence similarity and recombination breakpoint sites

Both GC content and sequence similarity has been shown to be associated with recombination breakpoint locations (Magiorkinis *et al.* 2003; Kiktev *et al.* 2018). Therefore, a further modification of the breakpoint clustering hot- and cold-spot test was used to test for associations between breakpoint sites (specifically breakpoint probability distributions) and: (1) guanine+cytosine (GC) content; and (2) pairwise sequence similarity (Simon-Loriere *et al.* 2010). In this test average GC proportions or pairwise sequence similarities of sites between a specified number of nucleotides (either 10 or 20) of every site in the genome across all possible sequence pairs were determined. Breakpoint probabilities at each site were multiplied with the GC proportion or pairwise similarity associated with that site and summed across all sites. These sums for the real datasets were then compared with the corresponding sums from the permuted datasets. For each analysed subgenus dataset the proportion of permuted datasets with sums higher than or equal to the real dataset were reported as the probability that there was no association between breakpoint positions and either higher GC proportions or higher degrees of pairwise similarity. Conversely, the proportion of permuted datasets with sums lower than or equal to those determined for the real dataset was reported as the probability that there was no association between breakpoint positions and either lower GC proportions or lower degrees of pairwise sequence similarity.

3.2.4 Identification of potential transcriptional regulatory sequences

SuPER was used to detect transcriptional regulatory sequence leader (TRS-L) sites without RNA-seq data. This was done by supplying SuPER with annotation files and reference sequence files (downloaded from NCBI in September 2021 for the best-sampled species in each of the six coronavirus subgenus datasets) in order for the algorithm implemented within SuPER to infer the subgenomic mRNA positions. Furthermore, a custom Python (Rossum and Drake 2010) script was customised to infer transcriptional regulatory sequence body (TRS-B) sites, as detected by SuPER, and previously confirmed (Yang *et al.* 2021). This custom Python script, namely “trsb-finder” (<https://github.com/phillipswanepoel/trsb-finder>) was used to search for potential TRS-B sites in each subgenus dataset, following the methodology used in SuPER (but using Levenshtein distance, as opposed to a hamming distance), which involved searching for all occurrences of sub-sequences with a specified Levenshtein distance (a string metric for measuring differences between sequences that can be used to determine or enforce the minimum required number of changes to a string of characters for one to match the other) of one or zero from the TRS-Leader sequence (Yang *et al.* 2021). These potential TRS-B sites were then filtered by; (1) removing all the sites not conserved across at

least 75% of the analysed sequences, (2) removing upstream sites when multiple sites were found in close proximity 5' of the start of the same ORF and, (3) removing all except the most downstream site if multiple resided within individual intergenic regions. This filtered siteset was then tested for association with breakpoint positions in each of the six analysed subgenera datasets using RDP5.

Given that neither the TRS distributions nor the breakpoint distributions were random in any of the analysed datasets and that both TRS sites and recombination breakpoint clusters occurred at the edges of coronavirus genes, we anticipated that the association test could have a high false-positive rate. To estimate the false discovery rate (FDR) of the breakpoint association test, another custom Python script, namely "random_trsb" (<https://github.com/phillipswanepoel/trsb-finder>) was used to generate randomly permuted versions of TRS-B site locations, for each of the subgenera alignments. As input files, the script takes in a coronavirus subgenus alignment file and another file containing associated TRS-B sites. The script then outputs an RDP5 readable siteset file containing the permuted nucleotide positions. These positions are calculated by collectively shifting all the TRS-B sites by some number of nucleotides (which preserves their spacing), varied randomly between one and the length of the analysed alignment. If a new shifted TRS position was beyond the end of the genome, the position was "wrapped" around to the other end of the genome. Two hundred permuted TRS-B sitesets were tested and the average estimated FDR across all datasets for the association between breakpoint and TRS sites was 17.27% (i.e. 17.27% of the analyses with "shifted" TRS-B sites yielded a significant association - with a p-value < 0.05 - between these sites and observed breakpoint positions). Given that the FDRs for individual subgenera datasets ranged from 5% to 29%, we only considered associations detected between TRS-B and breakpoint sites as being significant if they were detected in multiple datasets.

3.2.5 Protein folding disruption test

To test whether the observed recombination events were less disruptive of protein folding than would be expected if recombination breakpoints were randomly distributed, the SCHEMA test (Meyer *et al.* 2003; Lefeuvre *et al.* 2007), implemented in RDP5, was used to examine all protein-coding regions with associated publicly available high resolution atomic coordinate data (obtained from the Protein Data Bank; <https://www.rcsb.org/> (Berman, Henrick and Nakamura 2003)) and within which more than ten recombination breakpoints were detected. These stipulations were required to ensure that the test would have sufficient power to detect whether observed recombinants displayed significantly lower degrees of protein folding disruption with the SCHEMA test than would be expected under random recombination. Of all 56 unique encoded proteins for which structural data was available (across all subgenera; Supplementary Tables 4-9; Appendix), only Spike was amenable to further

analysis (Table 3.1). Specifically, four subgenera (*Pedacovirus*, *Merbecovirus*, *Sarbecovirus* and *Igacovirus*) had both available Spike atomic coordinate structural data and >10 detected recombination breakpoints in the portion of the S gene corresponding to the structural data.

The SCHEMA test involves identifying potential interactions that occur between amino acid residues within folded proteins (in our case pairs of non-hydrogen atoms from different amino acids within 4.5 ångströms ((Å): equal to one ten-billionth (US) of a meter (Ångström (unit) Definition 2021)) of one another) and counting the numbers of interacting amino acid pairs within a chimaera of two parental amino acid sequences, where the chimaera has a different pair of amino acids than both parents. The 4.5 Å interaction cut-off (the default setting) corresponds to approximately five to eight potential pairwise interactions per residue. The counts of potentially altered pairwise amino acid interactions (called the disruption or E-score) that the SCHEMA method calculates has been shown to strongly correlate with observed degrees of fold disruption within chimaeric proteins (Meyer *et al.* 2003). To determine whether observed recombinants expressed chimeric proteins with significantly lower E-scores than expected under random recombination, we used the permutation-based recombinant protein simulation approach of Lefeuvre *et al.* (2009).

3.3 Results and Discussion: underlying associations and sequence features

3.3.1 Selection likely disfavours recombinants expressing Spike proteins with disrupted folds

We used the SCHEMA method (Voigt *et al.* 2002; Lefeuvre *et al.* 2007) to more directly test for evidence of the inferred coronavirus recombination breakpoint distributions in the S gene having been impacted by natural selection disfavoring the survival of recombinants that express chimeric Spike proteins with disrupted folds. The only coronavirus proteins for which high resolution atomic coordinate data were available, and for which sufficient recombination breakpoint numbers were detected within their associated genome sites to perform the SCHEMA folding disruption test, were those of sequences in the *Merbecovirus*, *Sarbecovirus*, *Pedacovirus* and *Igacovirus* datasets.

We found that in the *Igacoviruses* and *Sarbecoviruses*, potential amino acid interactions within the Spike proteins expressed by observed recombinants have significantly fewer predicted structural impacts than would be expected under random recombination ($p < 0.05$; SCHEMA permutation test; Table 3.2). It is noteworthy that the test result for the *Pedacoviruses* also approached significance ($p = 0.079$; Table 3.2) but that for the *Merbecoviruses* displayed no such tendencies ($p = 0.875$; although it should be noted that, of the four datasets tested, this dataset had the lowest number of detected breakpoints in the S gene; Table 3.2). This implies that, as has been suggested previously with *in vitro* recombination experiments involving the *Embecovirus*, murine coronavirus (Banner and Lai 1991), the *Igacoviruses* and *Sarbecoviruses* (and possibly also the *Pedacoviruses*) display lower degrees of predicted recombination-induced protein folding disruption in their expressed Spike proteins than would be expected under random recombination in the absence of selection. It should be stressed that our power to detect such “avoidance of protein folding disruption” signals was restricted to Spike and that it remains plausible that, given enough additional sequence data and more extensive atomic-resolution 3D structure information for other coronavirus proteins, many of these proteins might also display such signals.

Table 3.2: SCHEMA derived estimates of recombination induced protein fold disruption comparing real and simulated recombination events, with rows in bold indicating subgenera with significantly fewer predicted structural impacts than would be expected under random recombination

Subgenus	Alignment From	Alignment To	Breakpoint Num	P. Val
Pedacovirus	21017	24947	23	0.07916
Embecovirus	25208	28924	8	0.24545
Merbecovirus	22362	26144	13	0.87464
Nobecovirus	N/A	N/A	N/A	N/A
Sarbecovirus	22001	25542	22	<0.001
Igacovirus	22094	27534	92	<0.001

* N/A = Not applicable (data unavailable at time of testing)

3.3.2 Indirect evidence that selection against protein misfolding impacts observable breakpoint distributions throughout coronavirus genomes

It would be expected that if natural selection tended to disfavour recombinants with misfolded proteins then breakpoints would tend to be found more frequently per non-coding nucleotide site than per amino acid encoding nucleotide site (Drummond *et al.* 2005). Also, it might be expected that, of the recombination breakpoints falling at amino acid encoding sites within genes, those falling at the edges of genes (for example in the first and last 5% of the coding sequence of a particular protein) might be less disruptive of coevolved intra-protein amino acid contacts that were crucial for correct folding than breakpoints falling within the middle regions of genes (Lefeuvre *et al.* 2007). If selection against misfolded proteins was impacting the distributions of recombination breakpoints throughout coronavirus genomes we would therefore expect that observed breakpoints might tend to fall more commonly: (1) in non-coding regions than in coding regions, and (2) at the edges of genes than in the middle parts of genes.

Accordingly, we found that the intergenic regions of the *Pedacoviruses*, *Embecoviruses*, *Nobecoviruses* and *Sarbecoviruses* all had significantly higher breakpoint densities ($p < 0.05$; permutation test; Table 3.3) than those in the protein-coding regions. Similarly, we detected that in the *Pedacoviruses*, *Embecoviruses*, *Sarbecoviruses* and *Igacoviruses*, detectable breakpoint densities were significantly higher in the beginning and ending 5% of coding regions than in the middle 90% of these regions ($p < 0.05$; permutation test; Table 3.4) with marginal significance observed in *Nobecoviruses* ($p = 0.054$; permutation test; Table 3.4).

Table 3.3: Comparison of detectable breakpoint numbers in non-coding regions and coding regions with rows in bold indicating subgenera with significantly more breakpoints in non-coding regions than would be expected under random recombination

Subgenus	BPs* in non-coding regions	BPs in coding regions	Permutation p-val
<i>Pedacovirus</i>	32	392	<0.001
<i>Embecovirus</i>	11	73	<0.001
<i>Merbecovirus</i>	1	66	0.660
<i>Nobecovirus</i>	4	79	0.012
<i>Sarbecovirus</i>	7	307	<0.001
<i>Igacovirus</i>	30	1683	0.650

* BPs = Breakpoints.

Table 3.4: Breakpoint densities falling in the end 10% (5% each end) of genes vs the middle 90% of genes with rows in bold indicating subgenera with significantly higher numbers of detectable breakpoints in the ending 10% of genes than would be expected under random recombination

Subgenus	BPs* in the end 10% of genes	BPs in the middle 90% of genes	Permutation p-val
<i>Pedacovirus</i>	68	507	<0.001
<i>Embecovirus</i>	25	195	0.003
<i>Merbecovirus</i>	5	127	0.810
<i>Nobecovirus</i>	12	112	0.054
<i>Sarbecovirus</i>	47	612	0.007
<i>Igacovirus</i>	191	3369	0.004

* BPs = Breakpoints.

Taken together the lower densities of breakpoints both within genes than in intergenic regions, and within the middle parts of genes than in the ends of genes, is reminiscent of similar breakpoint distribution patterns detected in HIV (Simon-Loriere *et al.* 2010) and the members of various single-stranded DNA virus families (Lefevre *et al.* 2009) and is consistent with the hypothesis that in

coronaviruses natural selection generally disfavours the survival of recombinants that express chimeric proteins with disrupted folds.

3.3.3 ORF1a genome regions generally have lower breakpoint densities than other coding regions

There was a significantly lower density of breakpoints detected in ORF1a than in other coding regions of the genome for all six of the analysed subgenera ($p < 0.05$; permutation test; Table 3.5). The relatively low numbers of recombination events involving transfers of sequence fragments within this region is most notable in three of the *Betacoronaviruses* subgenera: *Embecovirus*, *Nobecovirus* and *Sarbecoviruses* (note the blue/cyan/green triangles associated with most of ORF1ab in these subgenera in Figure 2.2). Our results here are therefore consistent with previous observations that there is a significant tendency for recombination breakpoints to fall outside ORF1a in the human-infecting coronaviruses OC43 (an *Embecovirus*) and NL63 (an *Alphacoronavirus* in the subgenus *Setracovirus*) (Pollett *et al.* 2021).

Within ORF1a the regions encoding the nonstructural proteins (nsps) nsp3 (a papain-like cysteine protease), and nsp4 have particularly low densities of identified breakpoints in multiple different subgenera (*Nobecovirus*, *Sarbecovirus* and *Igacovirus* for nsp3 and *Merbecoviruses* and *Igacoviruses* for nsp4). Together with nsp6, nsp3 and nsp4 cooperatively modify the endoplasmic reticulum (ER) of coronavirus infected cells into vesicles with double membranes to which viral replication complexes are tethered (Knoops *et al.* 2008; Hartenian *et al.* 2020; Klein *et al.* 2020; Mohan and Wollert 2021).

It is plausible that the relatively low numbers of recombination events detectable in ORF1a are attributable to the high degree to which these components interact with one another (Stark *et al.* 2006; Li *et al.* 2021). It is expected that these interactions might rely on coevolved interaction motifs and that these proteins might therefore not function optimally if transferred into a genomic background within which they did not coevolve (Jain, Rivera and Lake 1999; Martin *et al.* 2005b).

Table 3.5: Individual genes and sub-gene regions with significantly lower numbers of detectable breakpoints than would be expected under random recombination

Subgenus	Genome region	BPs* region	inside BPs region	outside permutation p-val
<i>Pedacovirus</i>	ORF1a	114	278	0.001
<i>Embecovirus</i>	ORF1a	43	130	0.001
<i>Merbecovirus</i>	ORF1a	43	130	0.001
<i>Nobecovirus</i>	ORF1a	14	65	<0.001
<i>Sarbecovirus</i>	ORF1a	94	213	<0.001
<i>Igacovirus</i>	ORF1a	667	1016	0.024
<i>Nobecovirus</i>	plpro (nsp3)	7	72	0.039
<i>Sarbecovirus</i>	plpro (nsp3)	49	258	0.031
<i>Igacovirus</i>	plpro (nsp3)	282	1401	0.016
<i>Merbecovirus</i>	nsp4	0	66	0.035
<i>Igacovirus</i>	nsp4	78	1605	0.002

* BPs = Breakpoints.

3.3.4 Breakpoints tend to fall at sites with lower than average GC content

To further our understanding of why, irrespective of selection, some coronavirus genomic sites might be more mechanistically predisposed to recombination than others, we tested breakpoint positions detected in each of the six analysed coronavirus datasets for associations with local GC contents (i.e. calculated proportions of all nucleotide residues that were G or C between 10 or 20 nucleotide sites up and downstream of detected breakpoint locations). High GC content is expected to potentially impact the frequencies at which recombination breakpoints occur in various ways such as (1) predisposing genome regions to form stable secondary structures that could cause pausing of RNA-Dependent RNA polymerase (RdRP) (Stark *et al.* 2006) (Experimental Evidence Codes | BioGRID 2021), (2) increasing the energy needed to break base-pairs during replication, and increasing the amount of time taken for RdRP to traverse these regions (Petes and Merker 2002; Sershen *et al.* 2011) and, if RdRPs disengages during replication, (3) increasing the probability of re-engagement through annealing with the same or a different template molecule (Lai 1990).

Contrary to expectations, but consistent with a recent report on recombination in coronaviruses (Pollett *et al.* 2021), we found that GC content within 20 nucleotides of breakpoint positions (Table 3.6) tended to be lower than expected under random recombination in all of the coronavirus datasets:

significantly so in the *Sarbecovirus* and *Pedacovirus* datasets ($P < 0.05$; permutation test). When we repeated the test only considering GC contents within 10 nucleotides of recombination breakpoints (20 nt window in Table 3.6), the significant associations between breakpoint positions and lower GC content in *Sarbecoviruses* and *Pedacoviruses* were strengthened, and additionally, marginally significant associations with lower GC contents ($0.05 < p < 0.1$; permutation test) were detected in *Merbecoviruses*, *Embecoviruses* and *Igacoviruses*.

Table 3.6: Associations between decreased GC content and detected recombination breakpoint sites with rows in bold indicating subgenera displaying average GC contents in the vicinity of breakpoint sites that are significantly lower than what would be expected under random recombination

Subgenus	Within 20 nt of breakpoint site		Within 10 nt of breakpoint site	
	p-val.	Significant	p-val	Significant
<i>Pedacovirus</i>	0.047	Yes	0.019	Yes
<i>Embecovirus</i>	0.322	No	0.080	Marginal
<i>Merbecovirus</i>	0.590	No	0.051	Marginal
<i>Nobecovirus</i>	0.791	No	0.693	No
<i>Sarbecovirus</i>	0.005	Yes	0.004	Yes
<i>Igacovirus</i>	0.693	No	0.099	Marginal

3.3.5 It is unclear whether sequence similarity directly influences the locations of recombination breakpoints

It has been previously found in other viruses that recombination breakpoint sites tend to occur more commonly at genome sites with elevated degrees of sequence conservation (van Vugt *et al.* 2001; Dazza *et al.* 2005; Archer *et al.* 2008). We therefore tested whether this pattern held for the six analyzed coronavirus subgenera.

Although recombination breakpoints in the *Sarbecovirus* and *Igacoviruses* datasets displayed a significant tendency to occur in genome regions displaying elevated degrees of average pairwise similarity among the analysed sequences ($p < 0.007$; permutation test; Table 3.7), for the *Pedacoviruses* and *Embecoviruses* the opposite was the case. In these subgenera, detectable recombination breakpoints have tended to fall in genome regions with lower degrees of average pairwise sequence similarity ($p < 0.005$; permutation test; Table 3.7). It is therefore unclear from our

test whether pairwise sequence similarity within 10 or 20 nucleotides of prospective recombination breakpoint sites is a direct determinant of where breakpoints occur within coronavirus genomes.

It is noteworthy in this regard that there are substantial variations in the degrees of sequence conservation across the analysed sequence datasets with, for example, the genome regions corresponding to the recombination breakpoint hot-spot immediately upstream of the S gene in the *Nobecovirus*, *Sarbecovirus* and *Igacovirus* datasets (all with a tendency for breakpoints to fall at more conserved sites) displaying among the highest degrees of sequence conservation within these datasets (Figure 3.1). Conversely, for the *Pedacoviruses* and *Embecoviruses* datasets (both with a tendency for breakpoints to fall at less conserved sites) the corresponding recombination hot-spots upstream of the S gene start codon fall at genome sites that have among the lowest degrees of genome-wide conservation in these datasets (Figure 3.1). It is therefore likely that, for this conserved hot-spot at least, sequence similarity has not been a primary determinant of where breakpoints have occurred.

Table 3.7: Association of breakpoint locations with higher/lower degrees of average pairwise sequence similarity with rows in bold indicating significant associations

Subgenus	Within 20 nt of breakpoint site		Within 10 nt of breakpoint site	
	Association with higher/lower similarity	p-val	Association with higher/lower similarity	p-val
<i>Pedacovirus</i>	Lower	<0.001	Lower	<0.001
<i>Embecovirus</i>	Lower	0.006	Lower	0.007
<i>Merbecovirus</i>	Higher	0.184	Higher	0.192
<i>Nobecovirus</i>	Higher	0.465	Higher	0.475
<i>Sarbecovirus</i>	Higher	0.005	Higher	0.005
<i>Igacovirus</i>	Higher	<0.001	Higher	<0.001

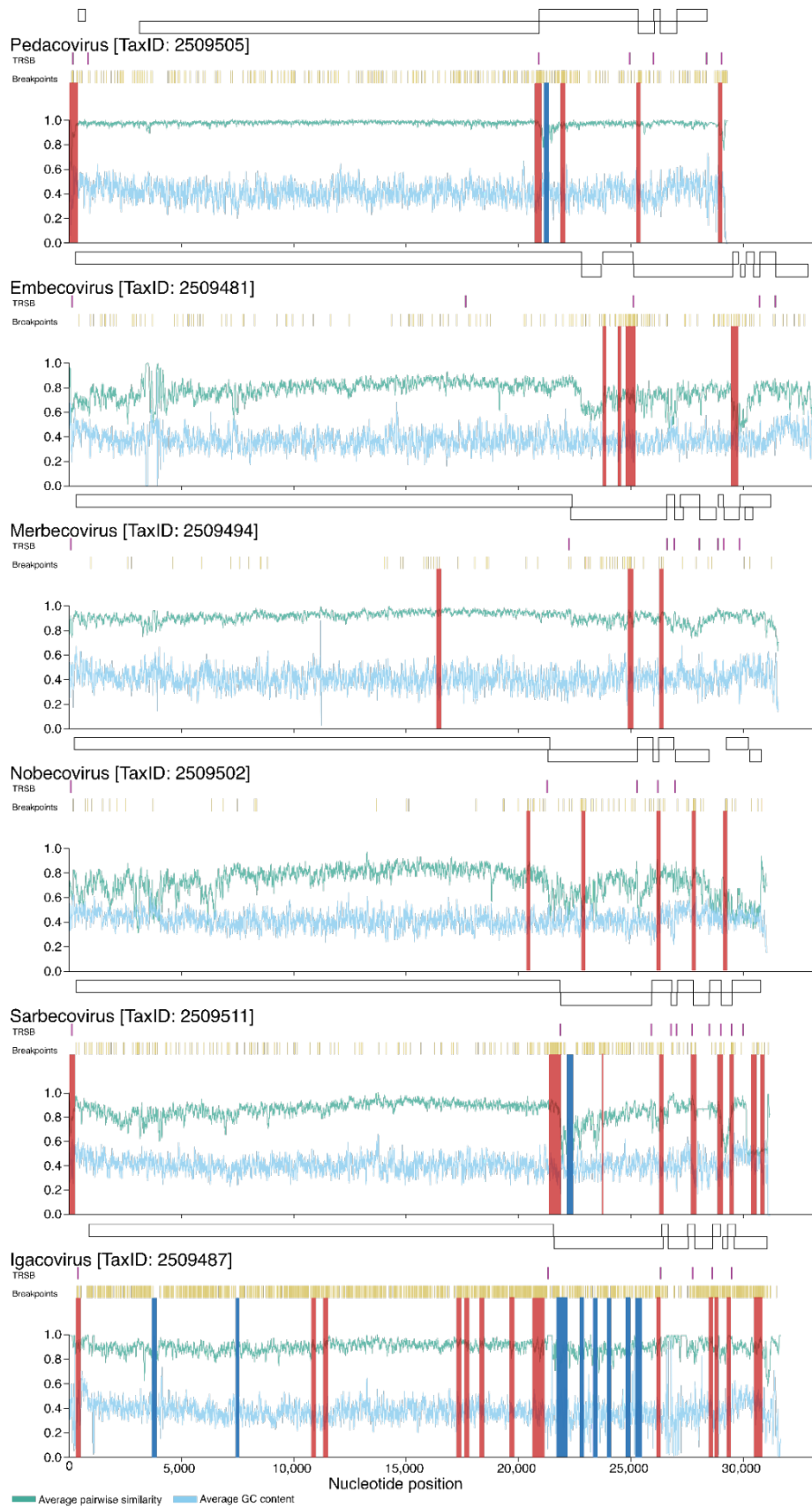


Figure 3.1. Regional variations in average pairwise sequence similarity (green) and GC content (blue) across coronavirus genomes. The plotted values indicate the pairwise sequence similarities and GC proportions within a moving 40 nucleotide window. Also indicated are the locations of the main genes (above each graph), transcriptional regulatory sequences (TRSs; in purple), identified breakpoint locations (in mustard), potential recombination hotspots (in red) and potential recombination cold-spots (in blue).

3.3.6 There is a strong association between recombination breakpoint locations and those of transcriptional regulatory sequences

Coronavirus transcription involves template switching at specific genome sites, called transcriptional regulatory sequences (TRSs) (Yang *et al.* 2021), previously called the intergenic sequence (Alonso *et al.* 2002). A possible link between template switching during gene expression and the genomic sites where recombination breakpoints occur during genome replication has been noted previously for coronaviruses in general (Zúñiga *et al.* 2004; Sola *et al.* 2015) and SARS-CoV specifically (Graham *et al.* 2018). Template switching is prone to occur during transcription of coronavirus negative genome strands whenever RdRp encounters the TRS sequences that are commonly found upstream of various genes. Because these “body TRS” (or TRS-B;) (Alonso *et al.* 2002; Sola *et al.* 2015) sites are involved in frequent template switching during transcription, it has been suggested that these sites might also promote template switching during genome replication (Graham *et al.* 2018) and, therefore, that they might colocalize with recombination hot-spots (Yang *et al.* 2021).

We used the SuPER method (Yang *et al.* 2021) to detect potential TRS-B sites in each of our six coronavirus datasets. Although SuPER is able to use RNA-seq data to precisely locate TRS-B sites, we were limited by not having RNA-Seq data for our specific datasets. We therefore made use of the first section of the SuPER workflow, involving supplying SuPER with a reference genome and annotation files to infer TRS-L sites. These inferred sites were then exported and compared with previously identified TRS-L sequences (Yang *et al.* 2021) to find and annotate likely TRS-B sites within the six analysed coronavirus datasets, using our custom “trsb-finder” script, in place of SuPER’s.

We found strong evidence for associations between the locations of conserved TRS-B sites (i.e. those detected in >75% of the analysed sequences in each dataset) and the locations of detected recombination breakpoints in the *Pedacoviruses*, *Igacoviruses*, *Embecoviruses* and *Sarbecoviruses* ($p < 0.05$; permutation test; Table 3.8 and Supplementary Figures 1, 2, 5 and 6; Appendix). These associations were detectable when we varied the required proximity between breakpoints and potential TRS-B sites to be considered a match from between 2 and 46 nucleotides.

Table 3.8: Associations between transcription regulatory sequence (TRS) sites and the locations of detected recombination breakpoints with p-values in bold indicating significant associations of TRS sites with higher breakpoint numbers

Subgenus	Within 46 nts p-val	Within 21 nts p-val	Within 9 nts p-val	Within 2 nts p-val
<i>Pedacovirus</i>	<0.001	<0.001	<0.001	<0.001
<i>Embecovirus</i>	<0.001	<0.001	<0.001	<0.001
<i>Merbecovirus</i>	0.117	0.178	0.210	0.806
<i>Nobecovirus</i>	0.014	0.039	0.020	0.478
<i>Sarbecovirus</i>	<0.001	<0.001	<0.001	0.005
<i>Igacovirus</i>	<0.001	<0.001	<0.001	<0.001

Given the large detectable recombination breakpoint hot-spots directly upstream of the S gene in most of the analysed subgenera datasets and the TRS-B sequences that map near these hot-spots, it was possible that the associations detected between TRS locations and breakpoint positions could have been attributable entirely to the TRS-B sites upstream of the Spike gene. To determine if this was the case, we repeated the association test (25nt window size) but this time with the TRS upstream of Spike removed from the analysis. We observed a minimal decrease in the significance of the association between TRS-B sites and recombination breakpoint positions, indicating that the initial result was not simply being driven by the coincidental colocalization of the S gene associated TRS-B site and the conserved recombination hot-spot upstream of the S gene in most of the analysed datasets.

We reran the TRS-B association tests with the positions of the TRS-B sites randomly shifted along the genome. The script takes as input the alignment file for each of the six datasets and places five to ten “false” TRS-B sites across each genome (the exact number corresponding for each subgenus dataset to the “true” TRS-B number for that dataset). When considering breakpoint probability distributions and an analysis window of 25 nucleotides, there was a significant absence of breakpoints within 12 nucleotides of TRS-B sites in the *Embecovirus* and *Sarbecovirus* datasets and neither significantly more or less breakpoints in close proximity to TRS-B sites in any of the other datasets. Both these results, along with our previous tests, are strong evidence that recombination breakpoints in coronaviruses generally tend to cluster at TRS-B sites.

However, given we have found that detectable recombination breakpoints tend to fall near the edges of genes in the same four subgenera in which we detected an association between TRS-B locations and recombination breakpoints (*Pedacovirus*, *Embecovirus*, *Sarbecovirus* and *Igacovirus*), this association between breakpoint locations and TRS-B sites might simply be attributable to the fact that TRS-B sites also tend to fall at the edges of genes. We therefore attempted to determine whether the association between breakpoint locations and TRS-B sites were still evident if we controlled for the colocalization of these sites at the edges of genes (Supplementary Figures 1, 2, 5 and 6). We were specifically interested in whether the presence/absence of a TRS-B site immediately upstream of a gene was associated with the presence/absence of a recombination breakpoint hot-spot upstream of the gene. Considering only the TRS-B sites and recombination breakpoint hot-spots falling either in intergenic regions or within 300 nucleotides of the beginning of genes we found a significant association between the presence of a TRS-B site near the beginning of a gene and the presence of a hot-spot near that location ($p = 0.0392$, Chi-square test with N-1 correction). Therefore suggesting that, for the *Pedacovirus*, *Sarbecovirus*, *Igacovirus* and *Embecovirus* datasets at least, the significant association we found between TRS-B sites and recombination breakpoint locations was not merely attributable to a coincidental tendency for breakpoints and TRS-B sites to colocalize near the edges of genes.

Chapter 4: Limitations, future research and conclusion

4.2 Study limitations and future research

There are several limitations of this study which should be considered. The results in this study highlight the need for more research to be directed at recombination analysis across other (potentially larger) groups of coronaviruses in some of the subgenera that were not considered here. Furthermore, when high-resolution maps of amino acid contacts within coronavirus protein complexes become available, and when the conserved nucleotide interactions within biologically functional RNA structural elements in a diverse enough array of coronavirus genomes have been identified, it should also be possible to determine the degree to which selection acting over the short- and/or long-terms to preserve these other categories of coevolved intra-genome interactions have impacted observable coronavirus recombination patterns.

4.2.1 Data

The publicly available full-length coronavirus datasets comprise sequences that have so far only been sampled for a miniscule portion of the time that coronaviruses have existed. As time passes and more data accumulates there will be far more opportunities to detect arising recombinants near to the times when they originated which, besides providing optimal power to detect signals of recombination breakpoints, will also increase the power of tests to detect sequence features that might impact the mechanistic predisposition of certain genome sites to recombine more than others.

The recent boom in coronavirus sequencing, has exposed growing issues relating to the quality and diversity of the coronavirus sequence datasets that are emerging. While the numbers of viral genomes being sequenced has increased dramatically, the average quality of individual sequences has decreased. With the continuous development of better sequencing technologies and greater awareness of coronavirus recombination, an increase in both the total number, and the quality of recombinant coronaviruses being sequenced are bound to follow.

Lastly, there are large differences in the numbers of recombination breakpoints that are detectable in different subgenus datasets. These differences largely reflect both the numbers of sequences analysed per dataset (e.g. the *Pedacovirus* dataset contained 410 sequences where *Nobecovirus* only contained 16 sequences) and the diversity of the datasets (e.g. *Nobecovirus* had only 75% similarity, where *Pedacovirus* had 98% similarity). Hopefully, with an increase in sampling, high quality uploads of genomic raw read and alignment data, and the constant maintenance of large public databases, these limitations are likely to diminish through time.

4.2.2 Mutational change and coinfection

As of late November 2021 (26/11/2021), 1117 unique allelic variants of SARS-CoV-2 (the virus causing COVID-19) have been identified in South Africa alone (of 2987 samples analysed) (<https://observablehq.com/@spond/intrahost-dashboard>; (Intrahost allelic variant SARS-CoV-2 analysis 2021)). With just over 25 million vaccine doses given and 14.1 million people having been fully vaccinated, South Africa is sitting with only 23.8% of its population fully vaccinated against COVID-19 (Countries where Coronavirus has spread - Worldometer 2021). These low vaccination numbers compound with the high numbers of immunocompromised individuals (such as those living with HIV or Tuberculosis (TB)). Rapidly mutating and recombining SARS-CoV-2 lineages potentially makes South Africa a hotbed for future SARS-CoV-2 diversification. Future research, based locally, can therefore help in further determining the long-term consequences of co-infections involving HIV, TB, SARS-CoV-2 and other coronaviruses (Dookie, Padayatchi and Naidoo 2020; Makoti and Fielding 2020).

4.3 Conclusion

Table 4.1: Conserved patterns of recombination across various coronavirus subgenera. Rows each contain the result of either a statistical test or the presence/absence of a particular characteristic of recombination (such as the presence of a hot-spot at a specific genome location): BP = breakpoint; blue = significant association or presence of characteristic; light blue = marginally significant association; pink = no significant association or absence of characteristic; white = untested

	<i>Pedacovirus</i>	<i>Sarbecovirus</i>	<i>Nobecovirus</i>	<i>Embecovirus</i>	<i>Igacovirus</i>	<i>Merbecovirus</i>
Hot-spot at 5' end of genome	Blue	Blue	Blue	Pink	Blue	Pink
Hot-spot upstream of S gene	Blue	Blue	Blue	Blue	Blue	Pink
Hot-spot in 5' half of S gene	Blue	Blue	Blue	Pink	Pink	Blue
Cold-spot at 5' end of S gene	Blue	Blue	Pink	Pink	Blue	Pink
More BPs in intergenic regions	Blue	Blue	Blue	Blue	Pink	Pink
Fewer BPs in middle of genes	Blue	Blue	Light blue	Blue	Blue	Pink
Higher BP density in lower GC regions	Blue	Blue	Pink	Light blue	Pink	Light blue
Higher BP density near TRS-B sites	Blue	Blue	Light blue	Blue	Blue	Pink
Avoidance of S gene fold disruption	Light blue	Blue	White	White	Blue	Pink

Across all the tests that we performed, viruses in the different analysed coronavirus genera displayed similar patterns of recombination (Table 4.1). The most strikingly similar of these patterns were those observed in the *Sarbecoviruses* (members of the *Betacoronavirus* genus) and the *Pedacoviruses* (members of the *Alphacoronavirus* genus). These mostly concordant patterns indicate that the processes that yield and select recombinant coronaviruses are likely broadly conserved across the three analysed coronavirus genera.

The subgenus dataset displaying the least concordant recombination patterns was that of the *Merbecoviruses*. It is unclear to us why the *Merbecovirus* dataset displays recombination breakpoint patterns that differ from the other analysed datasets: it is not an outlier among the datasets in terms of the numbers of sequences analysed or the average pairwise similarities of these sequences, but the dataset does have the lowest number of detectable recombination events. It is therefore possible that either the processes that generates recombinant genomes, the genetic factors that determine the viability of recombinants, or the epidemiological and evolutionary processes that impact the survival of recombinants, might differ somewhat between the *Merbecoviruses* and most other coronaviruses.

Nevertheless, the non-random and mostly conserved recombination patterns that we and others have detected in various coronavirus subgenera are likely shaped both by evolutionarily conserved variations in the mechanistic predispositions of different genome regions to recombination and by shared selective processes disfavoring the survival of recombinants that express improperly folded proteins. There are two non-exclusive explanations for why coronavirus genome sites that are mechanistically predisposed to recombination (such as those of TRS-B sequences) tend to coincide with sites where recombination seems to have had a minimal impact on protein folding: (1) negative selection over the short-term may be so efficient at purging all viral variants with recombination-induced protein misfolding that such variants are only rarely sequenced; and/or (2) longer-term selection, possibly acting since the most recent common ancestor of all known coronaviruses, may have yielded coronavirus genomes that are configured such that they are mechanistically predisposed to only recombine at sites where recombination breakpoints are minimally disruptive of protein folding.

Appendix

Supplementary Table 1: Coronavirus whole genome nucleotide sequence dataset characteristics

Subgenus	Genus	Common animal hosts	Number of pairwise sequences	Mean similarity
<i>Pedacovirus</i>	<i>Alphacoronavirus</i>	Bat, Porcine	412	98%
<i>Merbecovirus</i>	<i>Betacoronavirus</i>	Bat, Camel, Human	191	92%
<i>Nobecovirus</i>	<i>Betacoronavirus</i>	Bat	16	75%
<i>Sarbecovirus</i>	<i>Betacoronavirus</i>	Bat, Pangolin, Human	158	86%
<i>Embecovirus</i>	<i>Betacoronavirus</i>	Rodent, Bovine, Porcine, Rabbit, Deer, Human	181	78%
<i>Igacovirus</i>	<i>Gammacoronavirus</i>	Pigeon, Fowl, Pheasant	410	89%

Supplementary Table 2: TRS-B association test

	Gene boundaries without TRS-Bs	hot-spots	Gene boundaries with TRS-Bs	hot-spots
<i>Sarbecovirus</i>	1	1	9	6
<i>Pedacovirus</i>	3	1	5	3
<i>Igacovirus</i>	6	0	6	5
<i>Embecovirus</i>	7	3	4	1
total	17	5	24	15
N (boundary regions) 41				

Supplementary Table 3: Contingency table for TRS-Bs and hot-spots at gene boundaries

	w/o TRS-B	With TRS-B	Marginal Totals
No hot-spot	12	9	21
Hot-spot	5	15	20
Marginal Totals	17	24	41

Supplementary Table 4: Sarbecovirus exhaustive protein data bank key for SCHEMA protein fold disruption testing. Virus column denotes virus for which the most accurate protein data was available. Grey text in the virus column displays substituted data. Protein column displays the protein for which the PDB code is for. PDB Code column indicates the Protein Data Bank file accession code from <https://www.rcsb.org/>. A blank PDB code indicates no data at time of downloading

Virus	Protein	PDB Code
SARS-CoV	NSP1	2HSX
SARS-CoV-2	NSP2	7MSW
SARS-CoV-2	NSP3	6WOJ
Mouse hepatitis virus A59	NSP4	3VC8
SARS-CoV-2	NSP5	6LU7
SARS-CoV-2	NSP6	
SARS-CoV-2	NSP7-NSP8	7JLT
SARS-CoV-2	NSP9	6WXD
SARS-CoV	NSP10&NSP16	3R24
SARS-CoV-2	NSP11	
SARS-CoV-2	NSP12	6YYT
SARS-CoV-2	NSP13	7NIO
SARS-CoV-2	NSP14	5C8U
SARS-CoV-2	NSP15	6WXC
SARS-CoV-2	NSP16	
SARS-CoV-2	Spike	6VYB
SARS-CoV-2	ORF3A	6XDC
SARS-CoV	E-Protein	5X29
SARS-CoV-2	M-Protein	
SARS-CoV-2	ORF6	
SARS-CoV-2	ORF7A	7CI3
SARS-CoV-2	ORF7B	
SARS-CoV-2	ORF8	7JTL
SARS-CoV-2	N-Protein	6VYO

Supplementary Table 5: Embecovirus exhaustive protein data bank key for SCHEMA protein fold disruption testing. Virus column denotes virus for which the most accurate protein data was available. Grey text in the virus column displays substituted data. Protein column displays the protein for which the PDB code is for. PDB Code column indicates the Protein Data Bank file accession code from <https://www.rcsb.org/>. A blank PDB code indicates no data at time of downloading

Virus	Protein	PDB Code
SARS-CoV	NSP1	2HSX
SARS-CoV-2	NSP2	7MSW
SARS-CoV-2	NSP3	6WOJ
Mouse hepatitis virus A59	NSP4	3VC8
SARS-CoV-2	NSP5	6LU7
SARS-CoV-2	NSP7-NSP8	7JLT
SARS-CoV-2	NSP9	6WXD
SARS-CoV	NSP10&NSP16	3R24
SARS-CoV-2	NSP12	6YYT
SARS-CoV-2	NSP13	7NIO
SARS-CoV-2	NSP14	5C8U
SARS-CoV-2	NSP15	6WXC
HKU1	Main Protease	3D23
HKU1	RBD	5KWB
HKU1	HE	6Y3Y
OC43	N-NTD	4LI4
OC43	B6 Fab	7M51
OC43	HE	5N11
HKU1	Spike	5I08
OC43	Spike	6OHW
SARS-CoV-2	ORF3A	6XDC
SARS-CoV	E-Protein	5X29
SARS-CoV-2	ORF7A	7CI3
SARS-CoV-2	ORF8	7JTL
SARS-CoV-2	N-Protein	6VYO

Supplementary Table 6: Igacovirus exhaustive protein data bank key for SCHEMA protein fold disruption testing. Virus column denotes virus for which the most accurate protein data was available. Grey text in the virus column displays substituted data. Protein column displays the protein for which the PDB code is for. PDB Code column indicates the Protein Data Bank file accession code from <https://www.rcsb.org/>. A blank PDB code indicates no data at time of downloading

Virus	Protein	PDB Code
SARS-CoV	NSP1	2HSX
Avian infectious bronchitis virus	NSP2	3LD1
Avian infectious bronchitis virus	NSP3	3EJF
Mouse hepatitis virus A59	NSP4	3VC8
SARS-CoV-2	NSP7-NSP8	7JLT
Avian infectious bronchitis virus	NSP9	5C94
SARS-CoV	NSP10&NSP16	3R24
SARS-CoV-2	NSP12	6YYT
SARS-CoV-2	NSP13	7NIO
SARS-CoV-2	NSP14	5C8U
SARS-CoV-2	NSP15	6WXC
Bronchitis Cov	Spike	6CV0
SARS-CoV-2	ORF3A	6XDC
SARS-CoV	E-Protein	5X29
SARS-CoV-2	ORF7A	7CI3
SARS-CoV-2	ORF8	7JTL
Avian infectious bronchitis virus	N-Protein	2BTL

Supplementary Table 7: Pedacovirus exhaustive protein data bank key for SCHEMA protein fold disruption testing. Virus column denotes virus for which the most accurate protein data was available. Grey text in the virus column displays substituted data. Protein column displays the protein for which the PDB code is for. PDB Code column indicates the Protein Data Bank file accession code from <https://www.rcsb.org/>. A blank PDB code indicates no data at time of downloading

Virus	Protein	PDB Code
Porcine epidemic diarrhea virus	NSP1	6LPA
Porcine epidemic diarrhea virus	NSP1	5XBC
SARS-CoV-2	NSP2	7MSW
SARS-CoV-2	NSP3	6WOJ
Porcine epidemic diarrhea virus	NSP4	5Y4L
SARS-CoV-2	NSP5	6LU7
SARS-CoV-2	NSP7-NSP8	7JLT
Porcine epidemic diarrhea virus	NSP9	5YM6
SARS-CoV	NSP10&NSP16	3R24
SARS-CoV-2	NSP12	6YYT
SARS-CoV-2	NSP13	7NIO
SARS-CoV-2	NSP14	5C8U
SARS-CoV-2	NSP15	6WXC
Porcine epidemic diarrhea virus	Spike	6VV5
Porcine epidemic diarrhea virus	3CLpro with GC376	6L70
Porcine epidemic diarrhea virus	Protease 2	6NOZ
SARS-CoV-2	ORF3A	6XDC
SARS-CoV	E-Protein	5X29
SARS-CoV-2	ORF7A	7CI3
SARS-CoV-2	ORF8	7JTL
SARS-CoV-2	N-Protein	6VYO

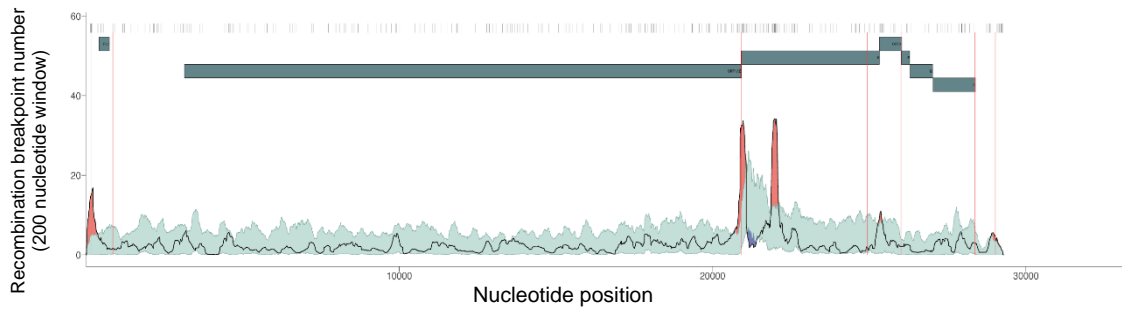
Supplementary Table 8: Merbecovirus exhaustive protein data bank key for SCHEMA protein fold disruption testing. Virus column denotes virus for which the most accurate protein data was available. Grey text in the virus column displays substituted data. Protein column displays the protein for which the PDB code is for. PDB Code column indicates the Protein Data Bank file accession code from <https://www.rcsb.org/>. A blank PDB code indicates no data at time of downloading

Virus	Protein	PDB Code
SARS-CoV	NSP1	2HSX
SARS-CoV-2	NSP2	7MSW
SARS-CoV-2	NSP3	6WOJ
Mouse hepatitis virus A59	NSP4	3VC8
MERS-CoV	NSP5	4RSP
SARS-CoV-2	NSP6	
SARS-CoV-2	NSP7-NSP8	7JLT
SARS-CoV-2	NSP9	6WXD
MERS-CoV	NSP10-16	5YN5
MERS-CoV	NSP15	5YVD
MERS-CoV	RBD	6L8Q
MERS-CoV	Spike	5X5C
MERS-CoV	S2 Fusion core	4NJL
MERS-CoV	ORF1a	5WKL
MERS-CoV	B6 Fab	7M55
MERS-CoV	N terminal Domain	6PXH
MERS-CoV	Nucleoprotein	6LZ8
MERS-CoV	Macro Domain	5HIH

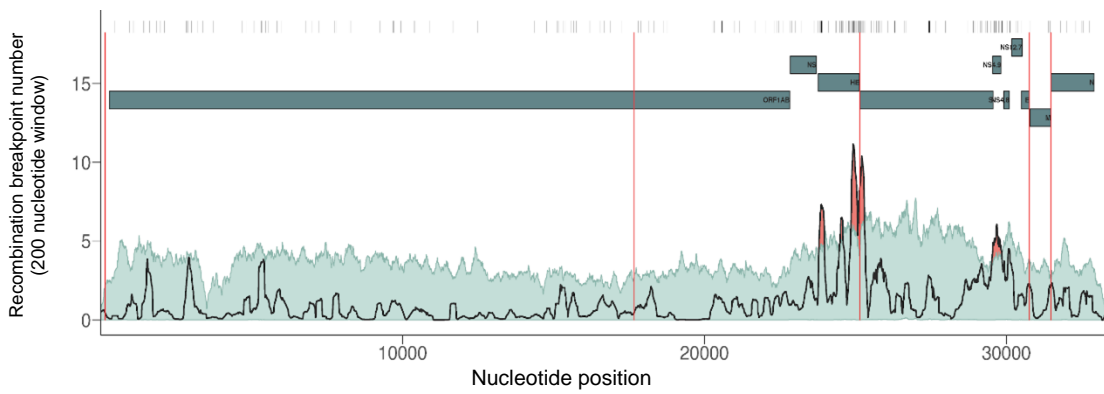
Supplementary Table 9: Nobecovirus exhaustive protein data bank key for SCHEMA protein fold disruption testing. Virus column denotes virus for which the most accurate protein data was available. Grey text in the virus column displays substituted data. Protein column displays the protein for which the PDB code is for. PDB Code column indicates the Protein Data Bank file accession code from <https://www.rcsb.org/>. A blank PDB code indicates no data at time of downloading

Virus	Protein	PDB Code
SARS-CoV	NSP1	2HSX
SARS-CoV-2	NSP2	7MSW
HKU9	NSP3	5UTV
HKU9	RBD	5GYQ
Mouse hepatitis virus A59	NSP4	3VC8
SARS-CoV-2	NSP5	6LU7
SARS-CoV-2	NSP6	
SARS-CoV-2	NSP7-NSP8	7JLT
SARS-CoV-2	NSP9	6WXD
SARS-CoV	NSP10&NSP16	3R24
SARS-CoV-2	NSP11	
SARS-CoV-2	NSP12	6YYT
SARS-CoV-2	NSP13	7NIO
SARS-CoV-2	NSP14	5C8U
SARS-CoV-2	NSP15	6WXC
SARS-CoV-2	NSP16	
SARS-CoV-2	Spike	6VYB
SARS-CoV-2	ORF3A	6XDC
SARS-CoV	E-Protein	5X29
SARS-CoV-2	M-Protein	
SARS-CoV-2	ORF6	
SARS-CoV-2	ORF7A	7CI3
SARS-CoV-2	ORF7B	
SARS-CoV-2	ORF8	7JTL
SARS-CoV-2	N-Protein	6VYO

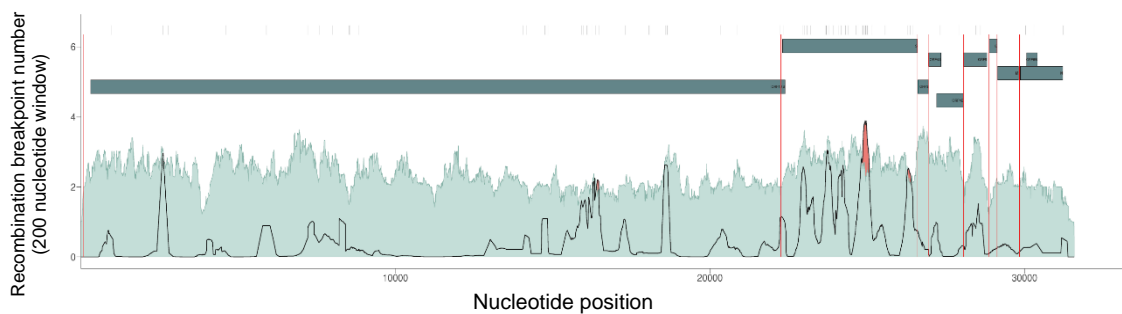
Supplementary Figure 1: Pedacovirus TRSB locations generated by adding filtered TRSB sites on top of Recombination Breakpoint Plots, generated in RDP5_RBDP_Rgrapher.



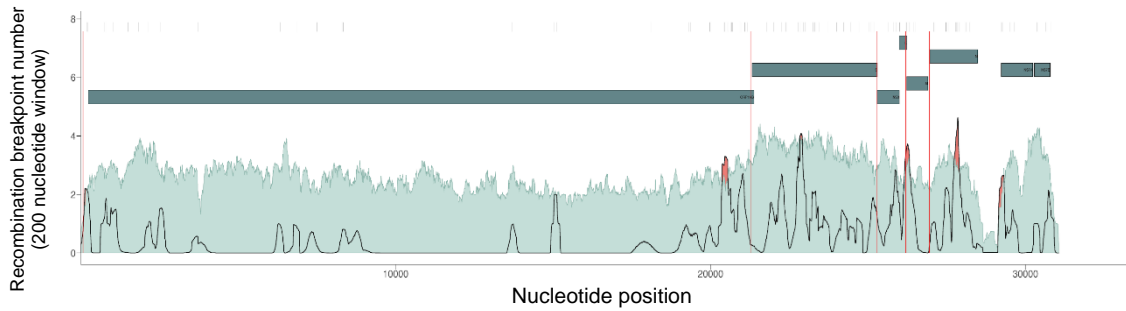
Supplementary Figure 2: Embecovirus TRSB locations generated by adding filtered TRSB sites on top of Recombination Breakpoint Plots, generated in RDP5_RBDP_Rgrapher.



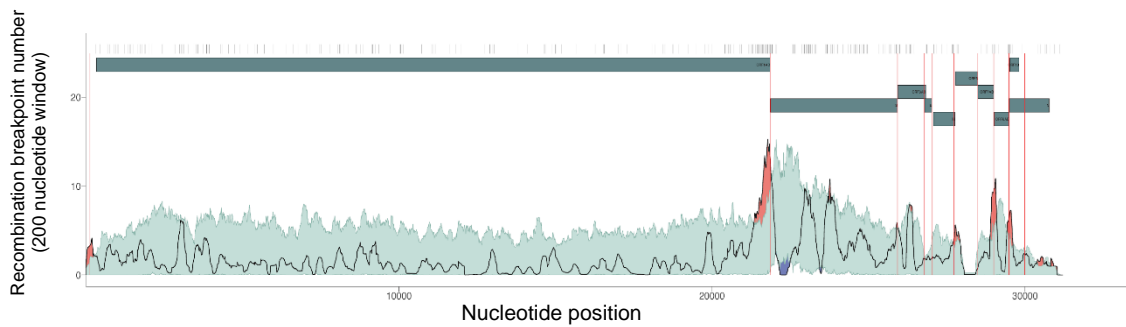
Supplementary Figure 3: Merbecovirus TRSB locations generated by adding filtered TRSB sites on top of Recombination Breakpoint Plots, generated in RDP5_RBDP_Rgrapher.



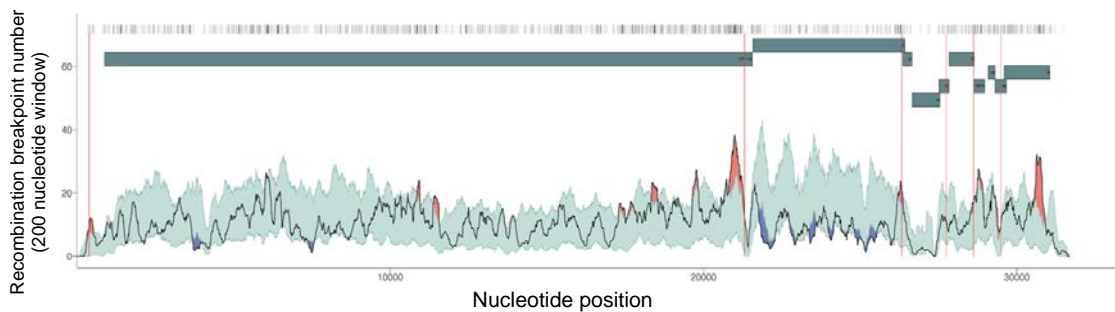
Supplementary Figure 4: Nobecovirus TRSB locations generated by adding filtered TRSB sites on top of Recombination Breakpoint Plots, generated in RDP5_RBDP_Rgrapher.



Supplementary Figure 5: Sarbecovirus TRSB locations generated by adding filtered TRSB sites on top of Recombination Breakpoint Plots, generated in RDP5_RBDP_Rgrapher.



Supplementary Figure 6: Igacovirus TRSB locations generated by adding filtered TRSB sites on top of Recombination Breakpoint Plots, generated in RDP5_RBDP_Rgrapher.



Supplementary Data 1: RDP5 RBDP Rgrapher guide. The program takes RDP5 Recombination Breakpoint Distribution Plot data and converts it into more legible graphs, fit for use in journal article figures, using the 99% (or 95%) confidence intervals.

"How to" guide

Initial setup

1. Clone/Download the code.
2. Open up `MainFile.R` in R studio (or similar IDE).
3. Allow R Studio (recommended) to install required libraries and packages (yellow popup).

What data to use as input

- RDP5 Beta version 5.16 and higher is required. Download (older) RDP here: <http://web.cbio.uct.ac.za/~darren/rdp.html>, or get the hidden, most updated version here: <http://web.cbio.uct.ac.za/~darren/mysetup.exe>
- Within RDP5 > 5.16, create a Breakpoint Distribution Plot and export the data by right-clicking the graph and hitting "Save CSV".
- RDBP_Grapher uses the following data - move it into the root directory of RDBP_Grapher: Breakpoint Distribution Data, ORFCoords and BreakpointPositions.

Data cleaning

1. Some data will require minor modification, here we guide you through the steps:
 - a. Open your Breakpoint Distribution Data (this will be the one without either csvORFCoords or csvBreakpointPositions appended).
 - b. Add a new column in G, titled `Bottom`.
 - c. Use the following formula, applied to all cells with contents to the left: `=MIN(B2:F2)`.
 - d. Modify both position one, and the final position of the dataset to have 0 values in all but their "Position in alignment" field. Like so, for both the first and last positions:

	A	B	C	D	E	F	G
1	Position in alignment	Recombination breakpoint number (200nt win)	Upper 99% CI	Lower 99% CI	Upper 95% CI	Lower 95% CI	Bottom
2	1	0	0	0	0	0	0
3	3	0.8206201	0.9228217	0	0.421387	0	0
4	5	0.9177571	0.9228283	0	0.4218889	0	0
5	7	1.014894	0.9228349	0	0.4223908	0	0
6	9	1.112031	0.9228414	0	0.4228927	0	0

- e. Save and exit.
2. Open your ORFPositions file.

3. Look carefully at the Gene Symbol column and purge duplicates. Be careful, as you need to take the smallest Start position and the largest Stop position for each unique Gene Symbol. Keep the formatting (leave gapless).

Data in R

Data import

There are just a few more things we need to modify or tweak, depending on the dataset.

1. In the `MainFile.R`, look under the `Importing the data` comment;
 - a. Modify the name of each of the 3 CSV files, according to your own data. In the example it is using Sarbecovirus or Nobecovirus (Change the green text).
 - b. Double check that you changed breakpointData, geneMap and breakpointDotPos to now read your own data.

Data Modifiers

- virusName: Output title used both on top of the graph, and underneath on the X axis.
- taxID: TaxID according to NCBI (<https://www.ncbi.nlm.nih.gov/taxonomy>).

More Modifiers

- fontSizeForGeneMap: How big must the text be in the gene map? Remember that when exporting, the text does not scale linearly compared to the preview image.
- fontSizeMultiplier: How much would you like to multiply the font size by.
- breakPointLineLength: How long should the breakpoint lines be (minimum of 1).

Optional Modifiers

There are many optional modifiers, but they require a bit more digging through the code. If you want something gone, simply comment it out and re-run.

A few notable optional modifiers include:

- The ability to remove titles (and spacing for them)
- Changing the height of the gene
- Individual font size modification
- Changing to 95% upper and lower confidence intervals
- Color changes
- Theme modding

Running the Code

- Press `CTRL + A` (to select all)
- Press `CTRL + ENTER` (to run selected code)

Saving the graph

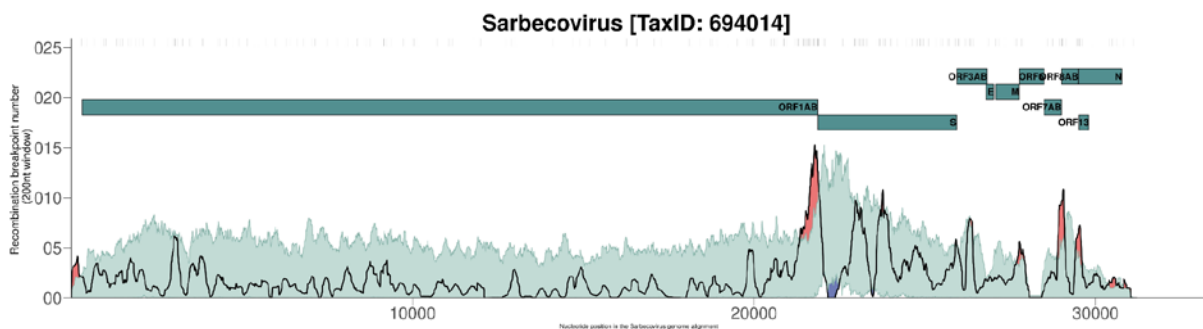
- To save the graph, click on Export -> Save as PDF
- If you save as a PNG, the image will likely be jagged

Saving settings:

PDF Size: 30 x 8 (but will depend on your own use case, so mess around with it)

1. Portrait
2. Pick a directory
3. Pick a File Name
4. Save

Example output:



Missing features

- Better randomisation to avoid overlapping gene maps (currently, you may need to generate a figure a couple of times to get the output you want where no gene maps overlap).

RDP5 RBDP R grapher also available at https://github.com/Arnedeklerk/RDP5_RBDP_Rgrapher

Supplementary Data 2: RDP5 RBDP Rgrapher code. R language code for the Rgrapher

```

# loading libraries and downloading packages                                # Rec (sarbecovirus 1.75*8 and 1.95) ()

library(ggplot2)                                                         yAxisEnable <- FALSE # Leave this modifier alone!

library(RCurl)

library(gridExtra)                                                       # More modifiers

library(plyr)                                                            fontSizeForGeneMap <- 6 # How big must the text be in the
                                                                         gene map?

library(tidyverse)                                                       fontSizeMultiplier <- 2 # How much would you like to times
                                                                         the font size by?

library(readxl)                                                         breakpointLineLength <- 1.2 # How long should the
                                                                         breakpoint lines be? (value greater than 1)

library(ggfittext)

# library(grid)

# library(viridis)

                                                                         # Program code: for those who know what they're doing:

# If you do not have these installed, remove the comment                # Making life easier by using default settings for the csv files
hash and install them.

# install.packages("viridis")

# install.packages("ggfittext")

                                                                         # Importing the data

breakpointData <-
read.csv("example_files/Sarbecovirus.csv")

geneMap <-
read.csv("example_files/Sarbecovirus.csvORFCoords.csv")

breakpointDotPos <-
read.csv("example_files/Sarbecovirus.csvBreakpointPositio
ns.csv")

# Modifiers (Edit here, as well as the above data for input!)

spikeOrFullLength <- "" # Leave this if you are not using
coronavirus data

virusName <- "Sarbecovirus"

taxId <- "694014"

                                                                         upper <- breakpointData$Upper.99..CI

                                                                         lower <- breakpointData$Lower.99..CI

# This bottom needs to be manually added in the
spadsheet (I could add code for this but that's a lot of
effort for something that is simple in excel)

bottom <- breakpointData$Bottom

# Getting the definite highest number and adding 10% to
move it a bit away

highestGene <- max(lower,rbn, upper)

highestGeneVal <- highestGene*1.1

                                                                         geneMapIndividualHeight <- (highestGene *0.1) # How
thick should the individual maps be?

```

```

geneMapHeight <- geneMapIndividualHeight *4 # Size of
potential random gaps

# The main graph, ordered correctly to show hot and cold
recombination breakpoint positions.

p <- ggplot() +

  geom_polygon(data = breakpointData,
aes(posInAlignment, rbn), fill = "#E85E5D", alpha=l(0.85)) +

  geom_polygon(data = breakpointData,
aes(posInAlignment, upper), fill = "#C2DBD5", alpha=l(1)) +

  geom_polygon(data = breakpointData,
aes(posInAlignment, lower), fill = "#5F6AB1", alpha=l(0.85))
+

  geom_polygon(data = breakpointData,
aes(posInAlignment, bottom), fill = "white") +

  geom_line(data = breakpointData, aes(posInAlignment,
upper), color = "#87B4A9") +

  geom_line(data = breakpointData, aes(posInAlignment,
lower), color = "#87B4A9") +

  geom_line(data = breakpointData, aes(posInAlignment,
rbn), size = 1.05, color = "black")

# Data Frame 1

geneMapDataFrame <-data.frame(geneMap$Gene.symbol,
geneMap$Start, geneMap$Stop)

# geneMapDataFrame

# Function for generating a random number (between two
numbers)

RANDBETWEEN <-

function(bottom,top, number = 1) {

  (runif(number,min = bottom, max = top))

}

# Data Frame 2

```

```

df2 <- data.frame(

  xMinimum = geneMapDataFrame$geneMap.Stop,

  xMaximum = geneMapDataFrame$geneMap.Start,

  allNames = geneMapDataFrame$geneMap.Gene.symbol,

  randomOne =

  round(RANDBETWEEN((highestGeneVal),(highestGeneVal+g
eneMapHeight),nrow(geneMapDataFrame))/(geneMapIndi
vidualHeight))*(geneMapIndividualHeight)

)

# Data Frame 3 (for bpps)

df3 <- data.frame(

  bppXMin = breakpointDotPos$Breakpoint.position

)

# Adjusting the look of the graphs

mytheme = list(

  theme_classic()+

  theme(panel.background =

  element_blank(),strip.background =

  element_rect(colour=NA, fill=NA),panel.border =

  element_rect(fill = NA, color = "black"),

  legend.title =

  element_blank(),legend.position="bottom", strip.text =

  element_text(face="italic", size=20 * fontSizeMultiplier),

  axis.text=element_text(size = 15 *

  fontSizeMultiplier),axis.title.y = element_text(size = 10*

  fontSizeMultiplier),axis.title.x = element_text(size = 10*

  fontSizeMultiplier),plot.title = element_text(face = "bold",

  hjust = 0.5,size=20* fontSizeMultiplier)) +

  theme(axis.ticks.length=unit(0.5,"cm")), # This is for the
length of "whiskers" on ggplot

  theme(

  # axis.title.x=element_blank(),

  # axis.title.y=element_blank(),

```

```

# axis.text.x = element_blank(), # TOGGLE COMMENT
ENABLE HERE TO MAKE X COORDINATES SHOW OR NOT
),
expand_limits(x=33300, y=0), # adding white space to the
graph
theme(
  panel.grid = element_blank(),
  panel.border = element_blank()
)
# theme(plot.caption = element_text(vjust = -3, hjust = 1,
size=(12*fontSizeMultiplier)))
# Gene Mapping (the color modification now lies here, as it
caused problems when within the data frame)
p + geom_rect(data=df2, aes(xmin = xMinimum, xmax =
xMaximum, ymin = randomOne, ymax = randomOne +
geneMapIndividualHeight), fill = "#085F63", alpha = 0.7,
color = "black") +
  geom_text(data=df2, aes(x = (xMinimum), y = randomOne
+ (geneMapIndividualHeight/2), label= toupper(allNames)),
size=fontSizeForGeneMap, fontface = "bold", hjust = 1,
color = "black") +
  # Breakpoint Pos Mapping
  geom_rect(data=df3, aes(xmin = bppXMin, xmax =
(bppXMin+1), ymin =
(highestGeneVal+geneMapHeight+(geneMapIndividualHeig
ht+geneMapIndividualHeight)), ymax =
((highestGeneVal+(geneMapIndividualHeight)+(geneMapHe
ight+(geneMapIndividualHeight/2))))), alpha = 0.7, color =
NA, fill = "black" ) +
  # Titles
  ggtitle(paste0(virusName, " [TaxID: ", taxId, "]" ,
spikeOrFullLength)) + # Toggle this comment to allow for
title generation
  xlab(paste0("Nucleotide position in the " , virusName, "
genome alignment " , spikeOrFullLength)) +
  ylab("Recombination breakpoint number \n (200nt
window)") +
  # labs(caption = "Bottletest") +
  mytheme +
  scale_x_continuous(expand = c(0,0)) +
  scale_y_continuous(expand = c(0,0), labels = function(x)
paste0("0",x)) # Toggle these if there must be no leading 0
# scale_y_continuous(expand = c(0,0)) print("You have
successfully printed the graph. If you get warnings do not
worry!")

```

*Comments have been added (# = comment) for code clarity.
RDP5 RBDP Rgrapher is available at https://github.com/Arnedeklerk/RDP5_RBDP_Rgrapher

References

- Aaziz R, Tepfer M 1999. Recombination in RNA viruses and in virus-resistant transgenic plants. *J Gen Virol* 1999;**80**:1339–46.
- Alonso S, Izeta A, Sola I *et al.* Transcription Regulatory Sequences and mRNA Expression Levels in the Coronavirus Transmissible Gastroenteritis Virus. *J Virol* 2002;**76**:1293–308.
- Ångström (unit) Definition. *Oxf Engl Dict* 2021.
- Anthony SJ, Gilardi K, Menachery VD *et al.* Further Evidence for Bats as the Evolutionary Source of Middle East Respiratory Syndrome Coronavirus. Schultz-Cherry S (ed.). *mBio* 2017;**8**, DOI: 10.1128/mBio.00373-17.
- Archer J, Pinney JW, Fan J *et al.* Identifying the Important HIV-1 Recombination Breakpoints. *PLOS Comput Biol* 2008;**4**:e1000178.
- Baharoon S, Memish ZA. MERS-CoV as an emerging respiratory illness: A review of prevention methods. *Travel Med Infect Dis* 2019;**32**:101520.
- Banner LR, Lai MM. Random nature of coronavirus RNA recombination in the absence of selection pressure. *Virology* 1991;**185**:441–5.
- Barr JN, Fearn R 2010. How RNA viruses maintain their genome integrity. *J Gen Virol* 2010;**91**:1373–87.
- Belouzard S, Millet JK, Licitra BN *et al.* Mechanisms of Coronavirus Cell Entry Mediated by the Viral Spike Protein. *Viruses* 2012;**4**:1011–33.
- Berman H, Henrick K, Nakamura H. Announcing the worldwide Protein Data Bank. *Nat Struct Biol* 2003;**10**:980.
- Boni MF, Lemey P, Jiang X *et al.* Evolutionary origins of the SARS-CoV-2 sarbecovirus lineage responsible for the COVID-19 pandemic. *Nat Microbiol* 2020 511 2020;**5**:1408–17.
- Bosch BJ, van der Zee R, de Haan CAM *et al.* The Coronavirus Spike Protein Is a Class I Virus Fusion Protein: Structural and Functional Characterization of the Fusion Core Complex. *J Virol* 2003;**77**:8801–11.
- Chao L. Fitness of RNA virus decreased by Muller's ratchet. *Nature* 1990;**348**:454–5.
- Cheng C-P, Nagy PD. Mechanism of RNA Recombination in Carmo- and Tombusviruses: Evidence for Template Switching by the RNA-Dependent RNA Polymerase In Vitro. *J Virol* 2003;**77**:12033.
- Cherry JD. The chronology of the 2002-2003 SARS mini pandemic. *Paediatr Respir Rev* 2004;**5**:262–9.
- Corman VM, Jores J, Meyer B *et al.* Antibodies against MERS Coronavirus in Dromedary Camels, Kenya, 1992–2013. *Emerg Infect Dis* 2014;**20**:1319–22.
- Corman VM, Muth D, Niemeyer D *et al.* Chapter Eight - Hosts and Sources of Endemic Human Coronaviruses. In: Kielian M, Mettenleiter TC, Roossinck MJ (eds.). *Advances in Virus Research*. Vol 100. Academic Press, 2018, 163–88.
- Coronaviridae - Positive Sense RNA Viruses - Positive Sense RNA Viruses (2011) - ICTV. *ICTV* 2011.
- Countries where Coronavirus has spread - Worldometer. 2021.
- Crameri A, Raillard S, Bermudez E *et al.* DNA shuffling of a family of genes from diverse species accelerates directed evolution. *Nature* 1998;**391**:288–91.
- Cui J, Li F, Shi Z-L. Origin and evolution of pathogenic coronaviruses. *Nat Rev Microbiol* 2019;**17**:181–92.
- Dazza M-C, Ekwawanga M, Nende M *et al.* Characterization of a Novel *vpu* -Harboring Simian Immunodeficiency Virus from a Dent's Mona Monkey (*Cercopithecus mona denti*). *J Virol* 2005;**79**:8560–71.
- Decaro N, Mari V, Campolo M *et al.* Recombinant Canine Coronaviruses Related to Transmissible Gastroenteritis Virus of Swine Are Circulating in Dogs. *J Virol* 2009;**83**:1532–7.
- Decaro N, Mari V, Elia G *et al.* Full-length genome analysis of canine coronavirus type I. *Virus Res* 2015;**210**:100–5.
- Denison MR, Graham RL, Donaldson EF *et al.* Coronaviruses: An RNA proofreading machine regulates replication fidelity and diversity. *RNA Biol* 2011;**8**:270–9.

- Domingo E, Holland JJ. RNA virus mutations and fitness for survival. *Annu Rev Microbiol* 1997;**51**:151–78.
- Dong G, Luo J, Zhang H *et al.* Phylogenetic Diversity and Genotypical Complexity of H9N2 Influenza A Viruses Revealed by Genomic Sequence Analysis. Brown J (ed.). *PLoS ONE* 2011;**6**:e17212.
- Dookie N, Padayatchi N, Naidoo K. Tuberculosis Elimination in the Era of Coronavirus Disease 2019 (COVID-19): A Moving Target. *Clin Infect Dis* 2020, DOI: 10.1093/cid/ciaa1400.
- Drouin G, Prat F, Ell M *et al.* Detecting and characterizing gene conversions between multigene family members. *Mol Biol Evol* 1999;**16**:1369–90.
- Drummond DA, Silberg JJ, Meyer MM *et al.* On the conservative nature of intragenic recombination. *Proc Natl Acad Sci* 2005;**102**:5380–5.
- Dudas G, Rambaut A. MERS-CoV recombination: implications about the reservoir and potential for adaptation. *Virus Evol* 2016;**2**:vev023.
- Elena SF, Carrasco P, Daròs J *et al.* Mechanisms of genetic robustness in RNA viruses. *EMBO Rep* 2006;**7**:168–73.
- Emam M, Oweda M, Antunes A *et al.* Positive selection as a key player for SARS-CoV-2 pathogenicity: Insights into ORF1ab, S and E genes. *Virus Res* 2021;**302**:198472.
- Experimental Evidence Codes | BioGRID. *BioGRID* 2021.
- Forni D, Cagliani R, Sironi M. Recombination and Positive Selection Differentially Shaped the Diversity of Betacoronavirus Subgenera. *Viruses* 2020;**12**:1313.
- Franzo G, Drigo M, Legnardi M *et al.* Bovine Coronavirus: Variability, Evolution, and Dispersal Patterns of a No Longer Neglected Betacoronavirus. *Viruses* 2020;**12**:1285.
- Gibbs MJ, Armstrong JS, Gibbs AJ. Sister-Scanning: a Monte Carlo procedure for assessing signals in recombinant sequences. *Bioinformatics* 2000;**16**:573–82.
- Goldstein SA, Brown J, Pedersen BS *et al.* Extensive Recombination-Driven Coronavirus Diversification Expands the Pool of Potential Pandemic Pathogens., 2021:2021.02.03.429646.
- Graham RL, Baric RS. Recombination, Reservoirs, and the Modular Spike: Mechanisms of Coronavirus Cross-Species Transmission. *J Virol* 2010;**84**:3134–46.
- Graham RL, Deming DJ, Deming ME *et al.* Evaluation of a recombination-resistant coronavirus as a broadly applicable, rapidly implementable vaccine platform. *Commun Biol* 2018;**1**:179.
- Guirao MP, Souza SP, Jerez JA *et al.* Phylogeny of canine coronavirus (CCoV) from Brazilian dogs based on membrane protein partial sequences. *Arq Bras Med Veterinária E Zootec* 2013;**65**:1887–90.
- de Haan CAM, Haijema BJ, Masters PS *et al.* Manipulation of the coronavirus genome using targeted RNA recombination with interspecies chimeric coronaviruses. *Methods Mol Biol Clifton NJ* 2008;**454**:229–36.
- Haider N, Rothman-Ostrow P, Osman AY *et al.* COVID-19—Zoonosis or Emerging Infectious Disease? *Front Public Health* 2020;**8**:596944.
- Hartenian E, Nandakumar D, Lari A *et al.* The molecular virology of coronaviruses. *J Biol Chem* 2020;**295**:12910–34.
- Hatcher EL, Zhdanov SA, Bao Y *et al.* Virus Variation Resource - improved response to emergent viral outbreaks. *Nucleic Acids Res* 2017;**45**:D482–90.
- Herrewegh AAPM, Smeenk I, Horzinek MC *et al.* Feline Coronavirus Type II Strains 79-1683 and 79-1146 Originate from a Double Recombination between Feline Coronavirus Type I and Canine Coronavirus. *J Virol* 1998;**72**:4508–14.
- Hussin JG, Hodgkinson A, Idaghdour Y *et al.* Recombination affects accumulation of damaging and disease-associated mutations in human populations. *Nat Genet* 2015 474 2015;**47**:400–4.
- Intrahost allelic variant SARS-CoV-2 analysis. 2021.
- Islam A, Ferdous J, Islam S *et al.* Evolutionary Dynamics and Epidemiology of Endemic and Emerging Coronaviruses in Humans, Domestic Animals, and Wildlife. *Viruses* 2021;**13**:1908.
- Jaimes JA, André NM, Chappie JS *et al.* Phylogenetic Analysis and Structural Modeling of SARS-CoV-2 Spike Protein Reveals an Evolutionary Distinct and Proteolytically Sensitive Activation Loop. *J*

- Mol Biol* 2020;**432**:3309–25.
- Jain R, Rivera MC, Lake JA. Horizontal gene transfer among genomes: The complexity hypothesis. *Proc Natl Acad Sci* 1999;**96**:3801–6.
- Jaroszewski L, Iyer M, Alisoltani A *et al.* The interplay of SARS-CoV-2 evolution and constraints imposed by the structure and functionality of its proteins. Punta M (ed.). *PLOS Comput Biol* 2021;**17**:e1009147.
- Jia W, Karaca K, Parrish CR *et al.* A novel variant of avian infectious bronchitis virus resulting from recombination among three different strains. *Arch Virol* 1995;**140**:259–71.
- Katoh K, Standley DM. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Mol Biol Evol* 2013;**30**:772.
- Keck JG, Matsushima GK, Makino S *et al.* In vivo RNA-RNA recombination of coronavirus in mouse brain. *J Virol* 1988a;**62**:1810–3.
- Keck JG, Soe LH, Makino S *et al.* RNA recombination of murine coronaviruses: recombination between fusion-positive mouse hepatitis virus A59 and fusion-negative mouse hepatitis virus 2. *J Virol* 1988b;**62**:1989–98.
- Kiktev DA, Sheng Z, Lobachev KS *et al.* GC content elevates mutation and recombination rates in the yeast *Saccharomyces cerevisiae*. *Proc Natl Acad Sci* 2018;**115**:E7109–18.
- Klein S, Cortese M, Winter SL *et al.* SARS-CoV-2 structure and replication characterized by in situ cryo-electron tomography. *Nat Commun* 2020;**11**:5885.
- Knoops K, Kikkert M, Worm SHE van den *et al.* SARS-Coronavirus Replication Is Supported by a Reticulovesicular Network of Modified Endoplasmic Reticulum. *PLOS Biol* 2008;**6**:e226.
- Konings DAM, Bredenbeek PJ, Noten JFH *et al.* Differential premature termination of transcription as a proposed mechanism for the regulation of coronavirus gene expression. *Nucleic Acids Res* 1988;**16**:10849–60.
- Krumm ZA, Lloyd GM, Francis CP *et al.* Precision therapeutic targets for COVID-19. *Viol J* 2021 **181** 2021;**18**:1–22.
- Kusters JG, Jager EJ, Niesters HG *et al.* Sequence evidence for RNA recombination in field isolates of avian coronavirus infectious bronchitis virus. *Vaccine* 1990;**8**:605–8.
- Lai MM. Coronavirus: organization, replication and expression of genome. *Annu Rev Microbiol* 1990;**44**:303–303.
- Lai MMC. Recombination in large RNA viruses: Coronaviruses. 1996.
- Lam H, Ratmann O, Boni M. Improved Algorithmic Complexity for the 3SEQ Recombination Detection Algorithm. *Mol Biol Evol* 2018;**35**:247–51.
- Lednicky JA, Tagliamonte MS, White SK *et al.* Isolation of a Novel Recombinant Canine Coronavirus from a Visitor to Haiti: Further Evidence of Transmission of Coronaviruses of Zoonotic Origin to Humans. *Clin Infect Dis* 2021, DOI: 10.1093/cid/ciab924.
- Lee CW, Jackwood MW. Evidence of genetic diversity generated by recombination among avian coronavirus IBV. *Arch Virol* 2000;**145**:2135–48.
- Lefevre P, Lett J-M, Reynaud B *et al.* Avoidance of Protein Fold Disruption in Natural Virus Recombinants. *PLOS Pathog* 2007;**3**:e181.
- Lefevre P, Lett J-M, Varsani A *et al.* Widely Conserved Recombination Patterns among Single-Stranded DNA Viruses. *J Virol* 2009;**83**:2697–707.
- Li BX, Ge JW, Li YJ. Porcine aminopeptidase N is a functional receptor for the PEDV coronavirus. *Virology* 2007;**365**:166–72.
- Li J, Guo M, Tian X *et al.* Virus-Host Interactome and Proteomic Survey Reveal Potential Virulence Factors Influencing SARS-CoV-2 Pathogenesis. *Med* 2021;**2**:99-112.e7.
- Li KS, Guan Y, Wang J *et al.* Genesis of a highly pathogenic and potentially pandemic H5N1 influenza virus in eastern Asia. *Nature* 2004;**430**:209–13.
- Licitra B, Duhamel G, Whittaker G. Canine Enteric Coronaviruses: Emerging Viral Pathogens with Distinct Recombinant Spike Proteins. *Viruses* 2014;**6**:3363–76.
- Liu P, Shi L, Zhang W *et al.* Prevalence and genetic diversity analysis of human coronaviruses among

- cross-border children. *Virology* 2017;141:1–8.
- Liu S, Xiao G, Chen Y *et al.* Interaction between heptad repeat 1 and 2 regions in spike protein of SARS-associated coronavirus: implications for virus fusogenic mechanism and identification of fusion inhibitors. *The Lancet* 2004;363:938–47.
- Lu G, Wang Q, Gao GF. Bat-to-human: spike features determining “host jump” of coronaviruses SARS-CoV, MERS-CoV, and beyond. *Trends Microbiol* 2015;23:468–78.
- Lytras S, Hughes J, Martin D *et al.* Exploring the natural origins of SARS-CoV-2 in the light of recombination. 2021:2021.01.22.427830.
- Magiorkinis G, Paraskevis D, Vandamme A-M *et al.* In vivo characteristics of human immunodeficiency virus type 1 intersubtype recombination: determination of hot spots and correlation with sequence similarity. *J Gen Virol* 2003;84:2715–22.
- Makino S, Keck J, SA S *et al.* High-frequency RNA recombination of murine coronaviruses. *J Virol* 1986;57:729–37.
- Makoti P, Fielding BC. HIV and Human Coronavirus Coinfections: A Historical Perspective. *Viruses* 2020;12:937.
- Manfredonia I, Nithin C, Ponce-Salvatierra A *et al.* Genome-wide mapping of SARS-CoV-2 RNA structures identifies therapeutically-relevant elements. *Nucleic Acids Res* 2020;48:12436–52.
- Martin D, Posada D, Crandall KA *et al.* A Modified Bootscan Algorithm for Automated Identification of Recombinant Sequences and Recombination Breakpoints. <https://home.liebertpub.com/aid> 2005a;21:98–102.
- Martin D, Rybicki E. RDP: detection of recombination amongst aligned sequences. *Bioinformatics* 2000;16:562–3.
- Martin D, Varsani A, Roumagnac P *et al.* RDP5: a computer program for analyzing recombination in, and removing signals of recombination from, nucleotide sequence datasets. *Virus Evol* 2021;7:veaa087.
- Martin D, van der Walt E, Posada D *et al.* The Evolutionary Value of Recombination Is Constrained by Genome Modularity. *PLoS Genet* 2005b;1:e51.
- Maynard Smith J. Analyzing the mosaic structure of genes. *J Mol Evol* 1992 342 1992;34:126–9.
- Memish ZA, Zumla AI, Al-Hakeem RF *et al.* Family Cluster of Middle East Respiratory Syndrome Coronavirus Infections. *N Engl J Med* 2013;368:2487–94.
- Meyer MM, Silberg JJ, Voigt CA *et al.* Library analysis of SCHEMA-guided protein recombination. *Protein Sci* 2003;12:1686–93.
- Mohan J, Wollert T. Membrane remodeling by SARS-CoV-2 – double-enveloped viral replication. *Fac Rev* 2021;10:17.
- Muller H. The relation of recombination to mutational advance. *Mutat Res* 1964;106:2–9.
- Ojkic D, Hazlett M, Fairles J *et al.* The first case of porcine epidemic diarrhea in Canada. *Can Vet J* 2015;56:149–52.
- Ou X, Liu Y, Lei X *et al.* Characterization of spike glycoprotein of SARS-CoV-2 on virus entry and its immune cross-reactivity with SARS-CoV. *Nat Commun* 2020;11:1620.
- Paltrinieri S, Metzger C, Battilani M *et al.* Serum α 1-acid glycoprotein (AGP) concentration in non-symptomatic cats with feline coronavirus (FCoV) infection. *J Feline Med Surg* 2007;9:271–7.
- Parrish CR, Holmes EC, Morens DM *et al.* Cross-Species Virus Transmission and the Emergence of New Epidemic Diseases. *Microbiol Mol Biol Rev* 2008;72:457–70.
- Pensaert MB, Martelli P. Porcine epidemic diarrhea: A retrospect from Europe and matters of debate. *Virus Res* 2016;226:1–6.
- Petes TD, Merker JD. Context Dependence of Meiotic Recombination Hotspots in Yeast: The Relationship Between Recombination Activity of a Reporter Construct and Base Composition. *Genetics* 2002;162:2049–52.
- Pettersen E, Goddard T, Huang C *et al.* UCSF Chimera—a visualization system for exploratory research and analysis. *J Comput Chem* 2004;25:1605–12.
- Pollett S, Conte MA, Sanborn M *et al.* A comparative recombination analysis of human coronaviruses

- and implications for the SARS-CoV-2 pandemic. *Sci Rep* 2021;**11**:17365.
- Poon A, Chao L. Drift Increases the Advantage of Sex in RNA Bacteriophage Φ 6. *Genetics* 2004;**166**:19–24.
- Raj VS, Mou H, Smits SL *et al*. Dipeptidyl peptidase 4 is a functional receptor for the emerging human coronavirus-EMC. *Nature* 2013;**495**:251–4.
- Reusken CB, Haagsma BL, Müller MA *et al*. Middle East respiratory syndrome coronavirus neutralising serum antibodies in dromedary camels: a comparative serological study. *Lancet Infect Dis* 2013;**13**:859–66.
- Reusken CB, Raj VS, Koopmans MP *et al*. Cross host transmission in the emergence of MERS coronavirus. *Curr Opin Virol* 2016;**16**:55–62.
- Rossum G van, Drake FL. *The Python Language Reference*. Release 3.0.1 [Repr.]. Hampton, NH: Python Software Foundation, 2010.
- Sawyer S. Statistical tests for detecting gene conversion. *Mol Biol Evol* 1989;**6**:526–38.
- Sershen CL, Mell JC, Madden SM *et al*. Superhelical Duplex Destabilization and the Recombination Position Effect. Lustig AJ (ed.). *PLoS ONE* 2011;**6**:e20798.
- Shaik JS, Khan A, Beverley SM *et al*. REDHORSE-REcombination and Double crossover detection in Haploid Organisms using next-geneRation SEquencing data. *BMC Genomics* 2015;**16**:133.
- Siegfried NA, Busan S, Rice GM *et al*. RNA motif discovery by SHAPE and mutational profiling (SHAPE-MaP). *Nat Methods* 2014;**11**:959–65.
- Simon-Loriere E, Galetto R, Hamoudi M *et al*. Molecular Mechanisms of Recombination Restriction in the Envelope Gene of the Human Immunodeficiency Virus. *PLoS Pathog* 2009;**5**:1000418.
- Simon-Loriere E, Holmes EC. Why do RNA viruses recombine? *Nat Rev Microbiol* 2011;**9**:617–26.
- Simon-Loriere E, Martin D, Weeks K *et al*. RNA structures facilitate recombination-mediated gene swapping in HIV-1. *J Virol* 2010;**84**:12675–82.
- Situation Report-51 SITUATION IN NUMBERS total and new cases in last 24 hours. 2020.
- Sola I, Almazán F, Zúñiga S *et al*. Continuous and Discontinuous RNA Synthesis in Coronaviruses. *Annu Rev Virol* 2015;**2**:265–88.
- Song S, Ma L, Zou D *et al*. The Global Landscape of SARS-CoV-2 Genomes, Variants, and Haplotypes in 2019nCoV. *Genomics Proteomics Bioinformatics* 2021, DOI: 10.1016/J.GPB.2020.09.001.
- Stark C, Breitkreutz B-J, Reguly T *et al*. BioGRID: a general repository for interaction datasets. *Nucleic Acids Res* 2006;**34**:D535–9.
- Su S, Wong G, Shi W *et al*. Epidemiology, Genetic Recombination, and Pathogenesis of Coronaviruses. *Trends Microbiol* 2016;**24**:490–502.
- Tang T, Bidon M, Jaimes JA *et al*. Coronavirus membrane fusion mechanism offers a potential target for antiviral development. *Antiviral Res* 2020;**178**:104792.
- Vlasova AN, Diaz A, Damtie D *et al*. Novel Canine Coronavirus Isolated from a Hospitalized Patient With Pneumonia in East Malaysia. *Clin Infect Dis* 2021, DOI: 10.1093/cid/ciab456.
- Voigt CA, Martinez C, Wang Z-G *et al*. Protein building blocks preserved by recombination. *Nat Struct Biol* 2002;**9**:553–8.
- van Vugt JJFA, Storgaard T, Oleksiewicz MB *et al*. High frequency RNA recombination in porcine reproductive and respiratory syndrome virus occurs preferentially between parental sequences with high similarity. *J Gen Virol* 2001;**82**:2615–20.
- Wan Y, Shang J, Graham R *et al*. Receptor Recognition by the Novel Coronavirus from Wuhan: an Analysis Based on Decade-Long Structural Studies of SARS Coronavirus. Gallagher T (ed.). *J Virol* 2020;**94**, DOI: 10.1128/JVI.00127-20.
- Wang L, Junker D, Collisson EW. Evidence of natural recombination within the S1 gene of infectious bronchitis virus. *Virology* 1993;**192**:710–6.
- Wang S, Qiu Z, Hou Y *et al*. AXL is a candidate receptor for SARS-CoV-2 that promotes infection of pulmonary and bronchial epithelial cells. *Cell Res* 2021;**31**:126–40.
- Wang W, Lin X-D, Guo W-P *et al*. Discovery, diversity and evolution of novel coronaviruses sampled from rodents in China. *Virology* 2015;**474**:19–27.

- Wang W, Lin X-D, Zhang H-L *et al.* Extensive genetic diversity and host range of rodent-borne coronaviruses. *Virus Evol* 2020;**6**:veaa078.
- Wege H, Stühler A, Lassmann H *et al.* Coronavirus infection and demyelination: Sequence conservation of the S-gene during persistent infection of Lewis-rats. *Adv Exp Med Biol* 1998;**440**:767–73.
- Wensman JJ, Stokstad M. Could Naturally Occurring Coronaviral Diseases in Animals Serve as Models for COVID-19? A Review Focusing on the Bovine Model. *Pathogens* 2020;**9**:991.
- Wesley RD. The S gene of canine coronavirus, strain UCD-1, is more closely related to the S gene of transmissible gastroenteritis virus than to that of feline infectious peritonitis virus. *Virus Res* 1999;**61**:145–52.
- White JM, Delos SE, Brecher M *et al.* Structures and Mechanisms of Viral Membrane Fusion Proteins: Multiple Variations on a Common Theme. *Crit Rev Biochem Mol Biol* 2008;**43**:189–219.
- Woo PCY, Huang Y, Lau SKP *et al.* Coronavirus Genomics and Bioinformatics Analysis. *Viruses* 2010;**2**:1804–20.
- Wrapp D, Wang N, Corbett KS *et al.* Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. *Science* 2020;**367**:1260–3.
- Xia S, Liu M, Wang C *et al.* Inhibition of SARS-CoV-2 (previously 2019-nCoV) infection by a highly potent pan-coronavirus fusion inhibitor targeting its spike protein that harbors a high capacity to mediate membrane fusion. *Cell Res* 2020;**30**:343–55.
- Yang D, Leibowitz JL. The structure and functions of coronavirus genomic 3' and 5' ends. *Virus Res* 2015;**206**:120–33.
- Yang Y, Yan W, Hall AB *et al.* Characterizing Transcriptional Regulatory Sequences in Coronaviruses and Their Role in Recombination. *Mol Biol Evol* 2021;**38**:1241–8.
- Yeager CL, Ashmun RA, Williams RK *et al.* Human aminopeptidase N is a receptor for human coronavirus 229E. *Nature* 1992;**357**:420–2.
- Zehr JD, Kosakovsky Pond SL, Martin DP *et al.* *Recent Zoonotic Spillover and Tropism Shift of a Canine Coronavirus Is Associated with Relaxed Selection and Putative Loss of Function in NTD Subdomain of Spike Protein.* *Evolutionary Biology*, 2021.
- Zhu Z, Meng K, Liu G *et al.* A database resource and online analysis tools for coronaviruses on a historical and global scale. *Database* 2021;**2020**, DOI: 10.1093/DATABASE/BAAA070.
- Zúñiga S, Sola I, Alonso S *et al.* Sequence Motifs Involved in the Regulation of Discontinuous Coronavirus Subgenomic RNA Synthesis. *J Virol* 2004;**78**:980–94.