Article

Not peer-reviewed version

Emergence of recombinant SARS-CoV-2 variants in California, 2020-2022

Rahil Ryder , Emily Smith , Deva Borthwick , Jesse Elder , Mayuri Panditrao , Christina Morales , Debra A. Wadford *

Posted Date: 27 June 2024

doi: 10.20944/preprints202406.1937.v1

Keywords: SARS-CoV-2; genomic epidemiology; genomic surveillance; recombination; whole-genome sequencing; California COVIDNet

Article

Emergence of Recombinant SARS-CoV-2 Variants in California, 2020–2022

Rahil Ryder [1,†], Emily Smith [2,†], Deva Borthwick [3], Jesse Elder [1], Mayuri Panditrao [3], Christina Morales [1] and Debra A. Wadford [1,*]

[1] Viral and Rickettsial Disease Laboratory, Center for Laboratory Sciences, California Department of Public Health (CDPH), Richmond, CA, USA
[2] Theiagen Genomics, Highlands Ranch, CO, USA
[3] COVID Control Branch, Division of Communicable Disease Control, CDPH, CA, USA
* Correspondence: Debra.Wadford@cdph.ca.gov; 510-307-8624
† These authors contributed equally to this work and share first authorship.

Abstract: The detection, characterization, and monitoring of SARS-CoV-2 recombinant variants is a challenge to public health authorities worldwide. Recombinant variants, composed of two or more SARS-CoV-2 lineages, often have unknown impacts on transmission, immune escape, and virulence in the early stages of emergence. We examined 4,213 SARS-CoV-2 recombinant SARS-CoV-2 genomes collected between 2020 and 2022 from California to describe regional and statewide trends in prevalence. Many of these recombinant genomes, such as those belonging to the XZ lineage, or novel recombinant lineages, likely originated within the state of California. We discuss the challenges and limitations surrounding Pango lineage assignments, the use of publicly available sequence data, and adequate sample sizes for epidemiologic analyses. Although these challenges will continue as SARS-CoV-2 sequencing volumes decrease globally, this study enhances our understanding of SARS-CoV-2 recombinant genomes to date while providing a foundation for future insights into emerging recombinant lineages.

Keywords: SARS-CoV-2; genomic epidemiology; genomic surveillance; recombination; whole-genome sequencing; California COVIDNet

1. Introduction

SARS-CoV-2 recombinants present unique challenges to public health genomic surveillance systems regarding detection, characterization, and possible impact on the phenotype of the virus. Recombinant SARS-CoV-2 genomes arise through the recombination of at least 2 different SARS-CoV-2 lineages and generally emerge following the co-circulation of multiple lineages at high prevalence. Recombination events likely occur within a single patient co-infected with co-circulating lineages [1–3]. However, there is also genomic evidence to suggest that recombination between non-co-circulating viruses may occur within long-term infected individuals whereby the original lineage recombines with a more recent lineage from a subsequent infection [4].

Recombinant genomes are important to track because they contain a novel combination of mutations not seen previously in a single lineage, such as the XBC recombinant, which includes mutations from both the Delta and Omicron (BA.2) variants (https://outbreak.info/compare-lineages). Known evolutionary mechanisms for other viral pathogens suggest that pandemic disease events are often preceded by genetic recombination [5,6]. This is perhaps most evident for influenza, whereby recombination and reassortment between different host-adapted gene segments has resulted in several influenza pandemics [7,8]. While recombination between currently circulating SARS-CoV-2 lineages is unlikely to result in the type of major antigenic shift that precedes influenza pandemics, it can result in the creation of novel strains with the potential to outcompete other lineages and evade host immunity, which has implications for both vaccine development and

 doi:10.20944/preprints202406.1937.v1

treatment of SARS-CoV-2 [9,10]. It is also thought that recombination between different coronaviruses played a role in the origin of SARS-CoV-2, further demonstrating the importance of monitoring recombinant viruses in a public health context [11–13]. Prior studies have found higher rates of hospitalization and lower neutralization titers for certain SARS-CoV-2 recombinant genomes [14,15]. However, the lack of available epidemiological metadata in public sequence repositories makes a large-scale analysis of these associations a significant challenge.

Worldwide, the first recognized SARS-CoV-2 recombinant lineage, designated as XA, occurred between lineages B.1.1.7 and B.1.177 and emerged in Europe in early 2021 following co-circulation of those lineages in that geographic region (https://virological.org/t/recombinant-sars-cov-2-genomes-involving-lineage-b-1-1-7-in-the-uk/658). However, a previous study revealed that SARS-CoV-2 recombinants were likely circulating at low levels early in the pandemic prior to the very first recombinant Pango lineage designation [16]. Through the end of 2022, 60 recombinant lineages were designated, some of which disseminated worldwide, while others remained as local clusters.

Throughout the pandemic, many designated recombinant SARS-CoV-2 lineages and several novel recombinant lineages have been identified from California genomic surveillance data. The high sequencing volume of SARS-CoV-2 positive specimens in California allows for the detection of emerging lineages not yet at high prevalence [17].

The objective of this study was to provide an overview of the genomic landscape of SARS-CoV-2 recombinants in California from the beginning of the pandemic through December 31, 2022 in combination with epidemiologic metadata to gain insight into how SARS-CoV-2 recombinant lineages emerge and spread within a population. Additionally, this study highlights persistent challenges of monitoring the emergence of novel SARS-CoV-2 lineages via genomic surveillance.

2. Materials and Methods

2.1. Genome Inclusion Criteria

The recombinant genomes examined in this study were obtained from the California COVIDNet sequence database [17] hosted in Terra (https://terra.bio/), a cloud-based bioinformatics platform used for genomic surveillance across many different public health laboratories in California [18], and GISAID (https://gisaid.org/) [19–21]. A genome was included in this study if it was collected in California prior to January 1, 2023 and determined to be a recombinant of 2 or more SARS-CoV-2 lineages, either manually by identifying groups of mutations that define two or more clades (https://github.com/pha4ge/pipeline-resources/blob/main/docs/sc2-recombinants.md) or systematically using the version of Pangolin that was most up-to-date at the time of genome assembly [22,23]. Genomes from the California COVIDNet sequence database were included if they had greater than 85% breadth of coverage relative to Wuhan-1 (NC_045512) as determined by the TheiaCoV workflows within the Public Health Bioinformatics Github repository (https://github.com/theiagen/public_health_bioinformatics). Duplicate genomes in both the California COVIDNet sequence database and GISAID were removed by matching internal identifiers to the Virus Name in GISAID. Some genomes were in the California COVIDNet sequence database but not GISAID either due to the location information being considered identifiable, or because GISAID had rejected the submitted genomes.

2.2. Genomic Epidemiology

To match SARS-CoV-2 sequences to epidemiologic information, data were obtained from four different California Department of Public Health (CDPH) databases: the COVID-19 Hospitalization Registry, the COVID-19 Case Registry, the Vaccine Registry, and the Integrated Genomic Epidemiology Dataset (IGED). The COVID-19 Hospitalization Registry includes data reported to CDPH following an All Facilities Letter (AFL) that required all hospitals in California to report specific patient-level information for each hospitalized patient who tested positive for COVID-19 on July 27, 2021 or later. The COVID-19 Case Registry includes SARS-CoV-2 laboratory results that are reported electronically or manually by laboratories, healthcare providers, and local health

 doi:10.20944/preprints202406.1937.v1

departments. The Vaccine Registry database includes vaccine information reported to the California Immunization Registry (CAIR). The IGED is a database of California SARS-CoV-2 lineages derived from whole genome sequencing (WGS) along with case demographic and epidemiologic information from the COVID-19 Case Registry. SARS-CoV-2 sequence data is required to be reported to CDPH per the updated California Code of Regulations Title 17, Section 2505, subsection (q).

A case was considered vaccinated if a dose of COVID-19 vaccine was received 14 days or more before the earliest known date associated with a specimen. An unknown vaccine status or unvaccinated status were not distinguishable and therefore not included. The denominator used to calculate percentages of cases that matched with the aforementioned CDPH databases was the total count of recombinant genomes, while the denominator used to determine percentage of hospitalizations and deaths was the total number of matched cases.

To compare total California sequences of all SARS-CoV-2 lineages in the same study period, analyses were performed for all California de-duplicated sequences using the above-mentioned databases by the following variables of interest: Specimen Collection Month, Vaccination Dose Count, Deaths, and Hospitalization. Specimen Collection Month was derived from the earliest date associated with each specimen and taken from the WGS dataset along with Pango Lineage and WHO variant classification (https://www.who.int/activities/tracking-SARS-CoV-2-variants). Total case counts were taken from the California COVID-19 Case Registry and included only confirmed cases. A confirmed case was defined as an individual with detection of SARS-CoV-2 ribonucleic acid (RNA) in a clinical or post-mortem specimen using a diagnostic molecular amplification test performed by a Clinical Laboratory Improvement Amendments (CLIA)-certified provider, or detection of SARS-CoV-2 RNA in a clinical or post-mortem specimen by genomic sequencing. A map visualization of the number of recombinants per region was created using ArcGIS Pro (version 3.0.0), with population information from worldpopulationreview.com.

2.3. Phylogeny and Recombinant Sites

Phylogenetic trees of select recombinant lineages, including both California and non-California genome assemblies, were obtained by uploading GISAID accessions to UShER for phylogenetic placement [24]. Separate analyses were performed specifically on California genomes using the Augur workflows (https://github.com/theiagen/public_health_bioinformatics) with default settings on Terra.bio for phylogenetic tree construction [25]. Auspice was used for tree visualization to demonstrate spread throughout the state [23]. Genomes were annotated according to the corresponding California Public Health Officer Region if location information was available (https://www.cdph.ca.gov/Programs/CID/DCDC/Pages/COVID-19/Order-of-the-State-Public-Health-Officer-Hospital-Health-Care-System-Surge-FAQ.aspx).

Recombinant sites within the genome were identified from the original Github issue on the pango-designation repository (https://github.com/cov-lineages/pango-designation/) for designated recombinant lineages or were determined empirically by manually assessing the mutation pattern for novel recombinant lineages. Microsoft PowerPoint was used to create a figure approximating the recombinant sites across the SARS-CoV-2 genome for recombinant lineages included in this study.

3. Results

3.1. Landscape of SARS-CoV-2 Recombinants in California

Through December 31, 2022, 4,213 SARS-CoV-2 recombinant genome sequences were identified that originated from specimens collected in California (Supplementary Table S1). Of these, 2,889 (68.6%) were hosted on Terra in the California COVIDNet sequence database and 1,324 (31.4%) were identified from GISAID. Among the recombinants from the California COVIDNet sequence database and GISAID, 3,932 (93.3%) belonged to a designated SARS-CoV-2 recombinant lineage, while the remainder belonged to novel recombinant lineages. Major identified recombinant lineages (>30 sequences) were primarily composed of Omicron parental lineages except for XB (B.1.634 x B.1.631), a recombinant of two lineages that did not belong to one of the WHO variant classifications.

 doi:10.20944/preprints202406.1937.v1

The first recombinant lineage detected in California was XB, collected on May 3, 2021 when the prevalent lineages were primarily Alpha (B.1.1.7) and Epsilon (B.1.427/B.1.429) variants (Figure 1). XB circulated in California through September 2021, at which time Delta had become the dominant variant in the state. Interestingly, none of the designated Delta x Omicron (BA.1) recombinant lineages, XD or XF, were identified in California, but two novel Delta x BA.1 recombinant genomes were identified between December 2021 and February 2022. In the following months, BA.1 was rapidly replaced by BA.2, leading to the designation of more than 30 recombinant BA.1 x BA.2 lineages, most of which were ultimately detected in California. XE was the most frequently identified of these BA.1 x BA.2 recombinants, followed by XZ. At least three novel recombinant BA.1 x BA.2 lineages were also detected in California, but none of those lineages grew beyond 50 genomes so they did not receive a Pango lineage designation. Not long after BA.2 became the dominant lineage in California, BA.5 emerged, giving rise to a new set of recombinants with BA.2 and BA.5 parental lineages. XAS was the most frequently identified BA.5 x BA.2 recombinant lineage to date in California.

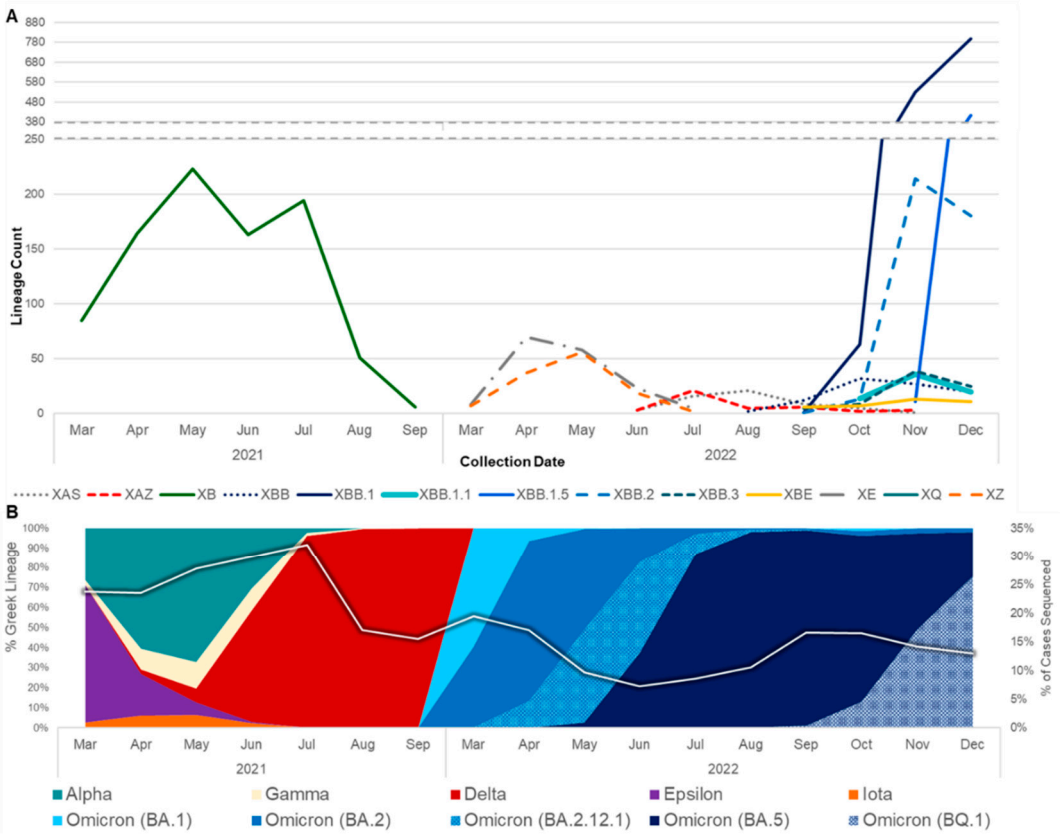

Figure 1. Prevalence of SARS-CoV-2 recombinants relative to SARS-CoV-2 variants from March 2021 to December 2022. A) The number of genomes belonging to major recombinant lineages (n≥30) in California. B) The prevalence of WHO variants and Pango lineage groupings in California data available from the California COVIDNet sequence database in Terra (left y-axis). The white line represents the percent of COVID-19 cases sequenced in California over time (right y-axis).

The rise of second-generation BA.2 lineages such as BA.2.75, BA.2.3.20, and BA.2.10.1, which trailed the initial BA.2 surge earlier in 2022 by several months, gave rise to another set of BA.2 and BA.5 recombinants. This included XBD, XBF, XBJ, and others, all of which have been identified in California and remained in circulation through the end of 2022. Similarly, two second-generation BA.2 lineages, BA.2.75 and BA.2.10.1 recombined to form the XBB lineage, of which 2,044 (48.5% of recombinants surveyed) were identified in California. Diversification of the XBB recombinant gave rise to XBB.1.5, a sublineage which was projected to become dominant in the United States shortly after emergence, of which 419 (9.9% of recombinants surveyed) were identified in California through

 doi:10.20944/preprints202406.1937.v1

the end of 2022. Two Delta x BA.2 recombinant lineages, XAY and XBC, were identified in California beginning in September 2022 and remained in circulation through December 2022.

SARS-CoV-2 recombinants were spread throughout all regions of California (Figure 2). In Northern California, the Rural Association of Northern California Health Officers (RANCHO) and the Association of Bay Area Health Officials (ABAHO) had XBB as the largest proportion of recombinants (36.4 - 37.9%). For the inland regions, represented by Greater Sacramento and the San Joaquin Valley Consortium (SJVC), the majority of recombinants were XB (31.6 – 32.9%) and XBB (25.4 - 27.0%). In the Southern California region, XBB accounted for 53.2% of all recombinants.

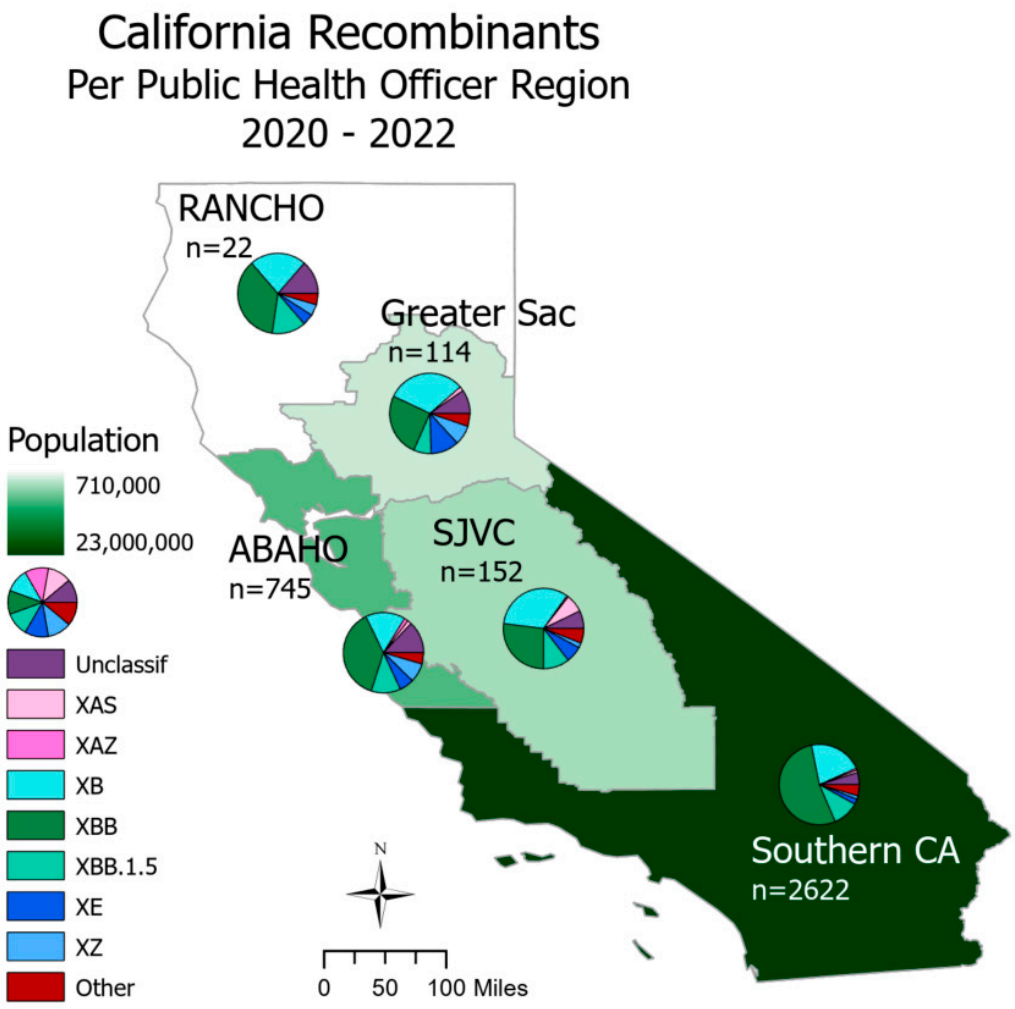

Figure 2. Recombinant lineages in California from 2020 to 2022 by Public Health Officer Region (PHO). Total sequence count (n) is split using a pie chart by major recombinant lineages and the "Other" category includes lineages with counts less than 30. PHO regions are colored by 2023 population size. There are 558 sequences with unknown county information which are not depicted on this map.

3.2. Epidemiological Data from CDPH Databases

From the California SARS-CoV-2 recombinant genomes, 2,523 (59.9%) matched to CDPH COVID-19 Hospitalization Registry and COVID-19 Case Registry databases (Table 1). Of these matched cases, 228 (5.4%) were younger than age 17 years, 1,524 (36.2%) were age 18-49, 521 (12.4%) were age 50-64, and 219 (5.2%) above age 65. From the Vaccine Registry, 1,433 (34.0%) vaccination statuses were identified. Vaccinations of those infected with a recombinant were as follows: 64 with one dose (4.5%), 358 with two doses (25.0%), 584 with three doses (40.8%), 179 with four doses (12.5%), 71 with five doses (5.0%), and 177 were vaccinated after infection or not within 14 days of

 doi:10.20944/preprints202406.1937.v1

earliest onset date (12.4%). It is important to note, however, that the timing and availability of vaccine doses changed throughout the pandemic (Supplemental Figure 1). Of those matched to the COVID-19 Hospitalization Registry and COVID-19 Case Registry, hospitalizations were reported for 51 (2.0%) cases infected with recombinant genomes, which included lineages XB, XBB, and XBB.1.5. Sixteen (0.6%) deaths occurred in cases with the recombinant lineages XAS, XB, XBB, XBB.1.5, and XE.

Table 1. Epidemiologic information associated with recombinant lineages in California.

Lineage	Dates Circulating in CA	Recomb^ Lineages	N	Epi Data Available* (%)	Hospitalized (%)	Died (%)	Died w/ 2 or more vaccine doses	Age Range of Hospitalized or Died
All Recomb^	--	--	4,213	2,523 (59.9%)	51 (2.0%)	16 (0.6%)	6	11-96
XB	Mar-Sept 2021	B.1.634 x B.1.631	886	497 (56.1%)	25 (5.0%)	7 (1.4%)	Unmatched	11-78
XE	Mar-July 2022	BA.1 x BA.2	161	106 (65.8%)	3 (2.8%)	3 (2.8%)	2	88-91
XZ	Mar-July 2022	BA.2 x BA.1	114	85 (74.6%)	2 (2.4%)	0	--	53-62
XAS	Jun-Nov 2022	BA.5 x BA.2	54	37 (68.5%)	1 (2.7%)	1 (2.7%)	1	75
XBB	Aug 2022-	BA.2 x BA.2	2,044	1103 (54.0%)	9 (0.8%)	3 (0.2%)	2	69-96
XBB.1.5	Nov 2022-	BA.2 x BA.2	419	318 (75.9%)	6 (1.9%)	2 (0.6%)	1	29-84
Other Recomb^	--	--	535	377 (70.5%)	5 (1.3%)	0	--	24-94

^ Recomb: abbreviation for "Recombinant". *Epi Data Available indicates the number of samples that matched to the COVID-19 Hospitalization Registry or COVID-19 Case Registry (epidemiologic data).

Comparatively, in California for all SARS-CoV-2 genomes (n= 801,534) sequenced prior to January 1, 2023, 24,483 (3.1%) of hospitalizations and 4,389 (0.5%) of deaths occurred. Of these cases, 467,395 (58.3%) were unvaccinated or had an unknown vaccination status, and 19,099 (2.4%) had 4 vaccine doses. It's important to note that these numbers only represent sequenced SARS-CoV-2 and not all cases.

3.3. Known Recombinant Lineages with Early Emergence in California

Both XZ (n=114) and XAS (n=54) demonstrate early emergence in California. From March to July 2022, 122 genomes of the BA.2 x BA.1 recombinant XZ, with breakpoints identified in Figure 3, were identified in California. It is possible that this recombinant lineage originated in California, as many of the sequences with the earliest collection dates and which were present on the most ancestral branch were from California (Figure 4a), and sequences from California made up more than 50% of all XZ genomes from the United States. Within California, some of the most ancestral sequences came from multiple geographic regions (Figure 4b), indicating that this lineage was already widespread before it began to diversify. This particular lineage was also eventually identified in many other US states in March and April 2022, followed by subsequent spread to Canada, Denmark, England, France, Germany, Mexico, and Japan in May through July 2022.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 June 2024 doi:10.20944/preprints202406.1937.v1

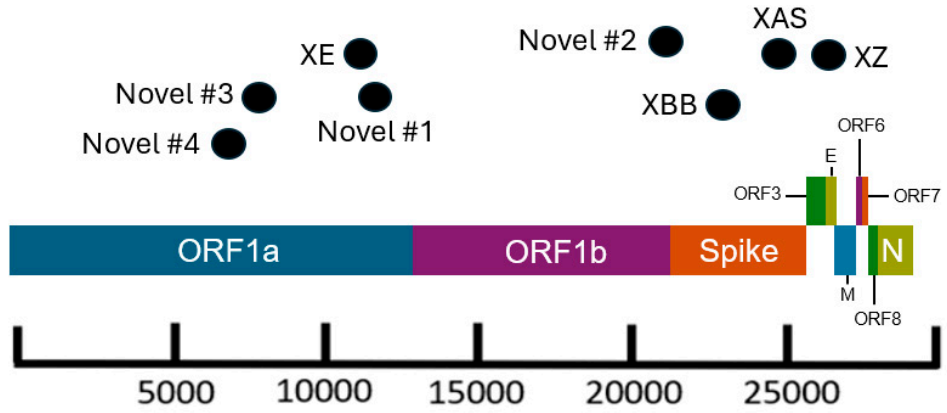

Figure 3. Schematic of the SARS-CoV-2 genome showing approximate breakpoint locations of novel and major recombinant lineages found in California. The horizontal location of the black dots represents the approximate region of the genome where the recombination occurred for the corresponding recombinant lineages.

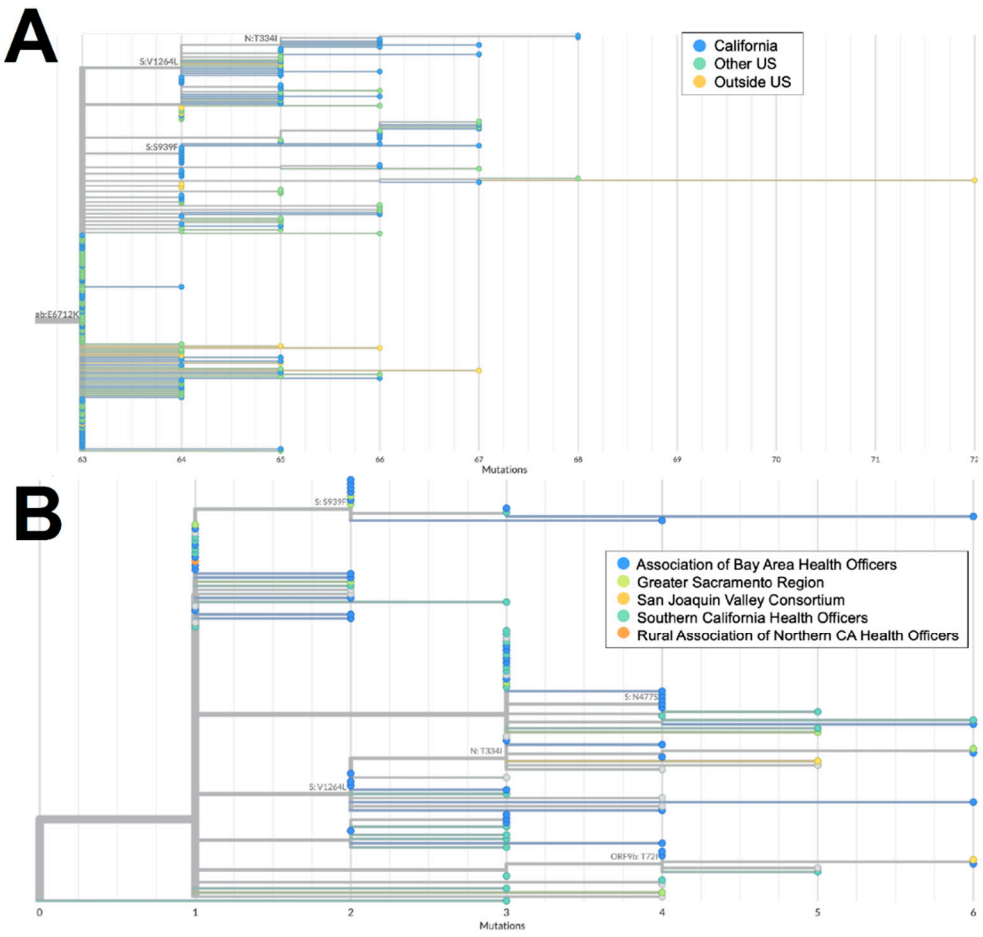

Figure 4. Phylogeny of XZ recombinant genomes. A) Global UShER tree of XZ genomes from GISAID. Genomes are colored according to geographical location in California (blue), outside of California but within the US (teal), or outside the US (orange). B) Phylogenetic tree of XZ genomes within California from Terra and GISAID. Genomes are colored according to regional location in the Association of Bay Area Health Officers (blue), Greater Sacramento (green), San Joaquin Valley Consortium (light

 doi:10.20944/preprints202406.1937.v1

orange), South California Health Officers (teal), or Rural Association of Northern California Health Officers (dark orange). Genomes for which regional location information was not available are shown in gray.

The XAS recombinant lineage had likely origins in Canada based on a large number of genomes from that country with early collection dates, followed by spread to countries in North and South America, as well as Europe. The sequences from California accounted for more than 40% of all the XAS genomes in the US and are disseminated throughout the global XAS tree (Figure 5a), indicating that the state played a key role in the propagation of this lineage within the US. Based on the phylogeny of sequences from California (Figure 5b), early emergence within the state occurred in the San Joaquin Valley Region, with subsequent spread through Southern California and the Bay Area.

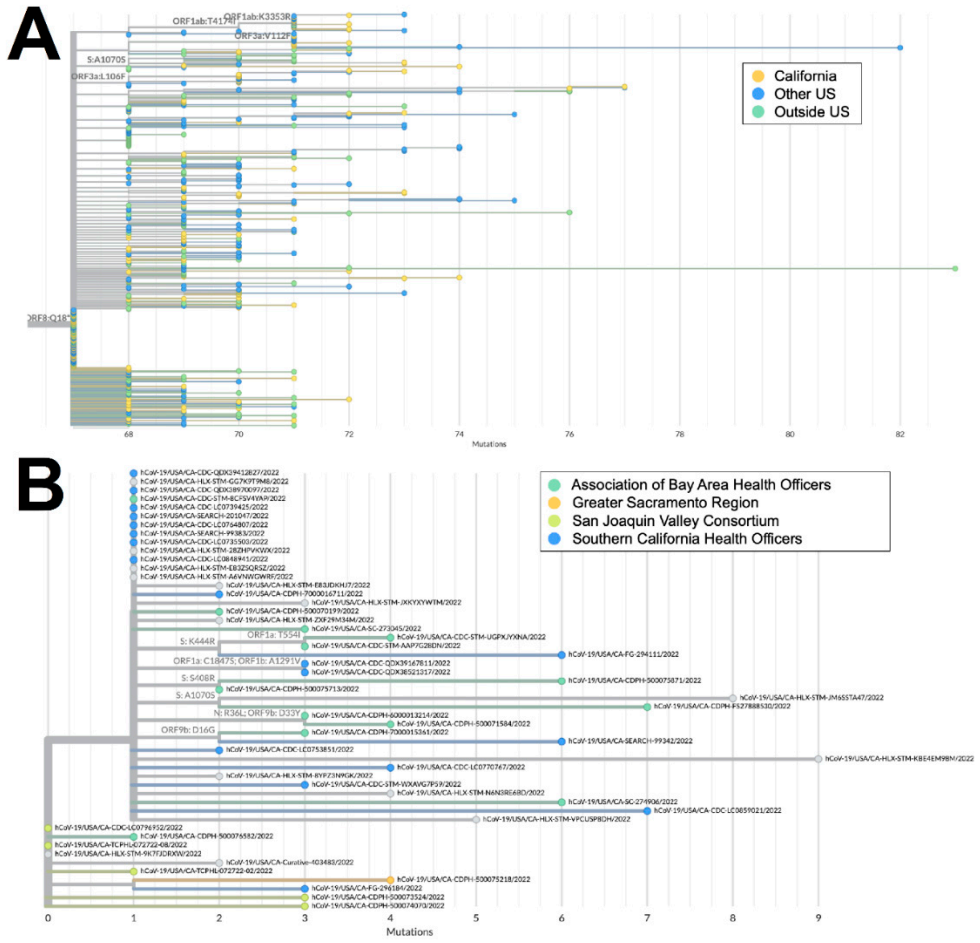

Figure 5. Phylogeny of XAS recombinant genomes. A) Global UShER tree of XAS genomes from GISAID. Genomes are colored according to geographical location in California (orange), outside of California but within the US (blue), or outside the US (teal). B) Phylogenetic tree of XAS genomes within California from GISAID. Genomes are colored according to regional location in the Association of Bay Area Health Officers (teal), Greater Sacramento (orange), San Joaquin Valley Consortium (green), or South California Health Officers (blue).

3.4. Novel Recombinant Lineages with Early Emergence in California

There are several recombinant lineages with early emergence in California that never grew beyond a small number of sequences and were either not proposed for or never received a Pango lineage designation. One of those novel recombinant lineages in California, hereafter designated as Novel #1, was first observed as a cluster of 11 BA.5 x XBC.1 sequences from four different sequencing laboratories (Figure 6). Novel #1 genomes had a recombinant site within ORF1a (Figure 3), and the

 doi:10.20944/preprints202406.1937.v1

XBC.1 portion of the genome also contained an additional amino acid substitution in the N gene, N:M210T. The first two of these genomes originated from specimens collected November 30, 2022, and genomes belonging to this novel recombinant lineage were still being identified through the end of 2022. This recombinant lineage was exclusive to sequences from California through the end of 2022. For the three sequences with granular location information available, all came from the Southern California Health Officers Region, but from different counties within that region.

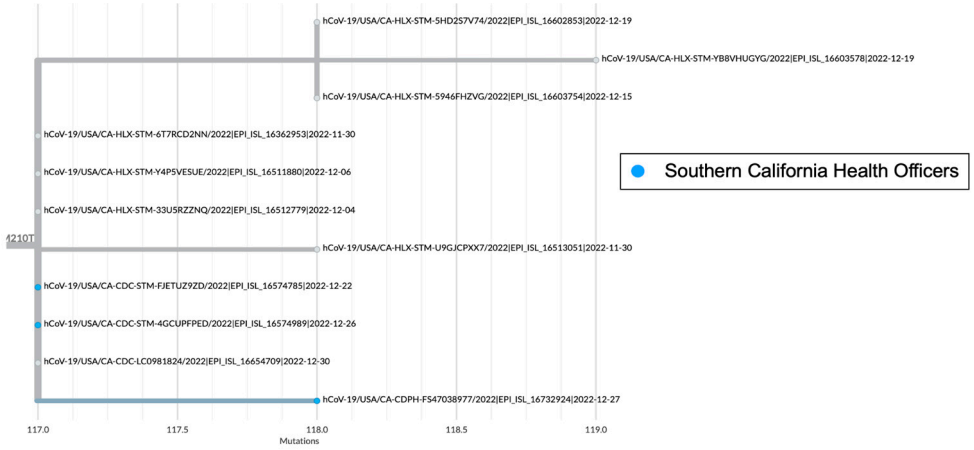

Figure 6. Global UShER tree of BA.5 x XBC.1 recombinant genomes from GISAID. Genomes from specimens originating in the Southern California Health Officers Region are shown in blue, and those for which regional location information was not available are shown in gray.

Earlier in 2022, two novel and distinct BA.1 x BA.2 recombinant lineages were identified. The first cluster, designated as Novel #2, contained 11 BA.1 x BA.2 sequences from California, collected February 15 - March 18, 2022, and two sequences from Nevada and Tennessee. The recombinant site for this cluster fell either in late ORF1b or early in the Spike protein based on the absence of synonymous nucleotide mutation A20055G and presence of S:T19I (Figure 3). A different BA.1 x BA.2 recombinant lineage, Novel #3, consisted of 18 genomes from California and one from Idaho which were collected between February 18 and March 21, 2022. The recombinant site for this cluster fell within ORF1a based on the presence of ORF1a:L2084I and absence of ORF1a:A2710T (Figure 3).

Another recombinant cluster, Novel #4, contained a single Delta x Omicron recombinant genome from California (EPI_ISL_10378301), but 11 other genomes were identified from at least four other states within the US. The recombinant site was located within ORF1a, with the early part of the genome belonging to the Delta variant, and potentially the AY.44 lineage within Delta based on the presence of synonymous nucleotide mutations early in the genome (Figure 3). The Omicron portion of the genome was determined to belong to the BA.1.1 lineage based on the presence of S:R346K amongst the many other Omicron-specific mutations. The California Novel #4 sequence, originated from a sample collected on February 13, 2022 in Sacramento County that had an additional amino acid substitution, N:P364L, differentiating it from the 11 other genomes from Massachusetts, Utah, New Mexico, and Hawaii.

3.5. Other Major Recombinant Lineages

The two largest recombinant lineages identified in California were XB (n=886) and XBB (n=2,044), both of which likely had international origins. XB genomes were identified from March through September of 2021, and this was the only recombinant lineage identified in California during the study period that did not have at least one Omicron parental lineage.

The first XBB genome in California was collected on July 27, 2022, and this lineage remained in circulation through the end of 2022. This lineage diversified as it expanded globally, resulting in the XBB.1.5 sublineage, of which 419 genomes were identified in California during the study period.

 doi:10.20944/preprints202406.1937.v1

4. Discussion

Geographic regions with higher population density and numbers of SARS-CoV-2 sequences in public repositories can provide insight into recombination events and dissemination of recombinant lineages worldwide. On a global scale, recombinants were identified at a higher frequency later in the pandemic attributing to the increased co-circulation of divergent lineages [26]. Notably in this study, although the percentage of cases sequenced decreased over time in California, there was an increase in recombinants detected (Figure 1). At least 30 different SARS-CoV-2 recombinant lineages were identified from California sequences collected through the end of 2022. California may be an ideal setting for identification of emerging lineages and recombination events due to the population density of several metropolitan regions, the influx of international travelers via cross-border ports of entry and major international airports, and the co-circulation of multiple lineages. However, another possibility is that the high volume of sequencing performed in the state due to the California COVIDNet Initiative [17] increased the likelihood of detecting recombinant lineages in California compared to other geographic locations.

Of the recombinants with matched case information, more than half originated in Southern California. This is not surprising considering that more than 22 million people (> 50% of the population) reside in Southern California and that SARS-CoV-2 genomic surveillance of that region has been well-represented. Interestingly, hospitalizations and deaths among individuals infected with recombinant genomes primarily occurred in the first 18 months of the pandemic, as seen with the lineage XB, which circulated from March through September of 2021. It is unclear if this observed difference is due to the lineage specifically, or due to its emergence before COVID-19 vaccines were widely available. Had sample sizes been sufficient for statistical analyses, a comparison of COVID-19 severity between recombinant lineages, and the parental lineages that gave rise to the recombinants may have helped to answer these questions. Regardless of the severity of COVID-19 outcomes, recombinants are important to surveil as their contribution to large mutational changes could lead to functional differences in transmissibility, immune escape, or pathogenesis.

Identifying specific combinations of mutations that contribute to the success of recombinant lineages has major implications for forecasting which lineages might become dominant or that may be considered as targets for vaccine or drug development. The recombinant lineage XBB, initially detected in India and projected to become dominant worldwide at the end of 2022, contained many amino acid substitutions in the Spike protein significantly associated with BA.2 breakthrough infection and enhanced ACE2 binding affinity [27]. However, contrary to predictions, XBB alone was not very advantageous on a global scale; it further diversified into XBB.1.5 with the additional S:G252V and S:F486P mutations that provided a substantial advantage against other lineages circulating at the time [28]. The recombinant lineage XAS had a partial BA.5 Spike protein, and perhaps combined with the truncation of ORF8 from the BA.2 donor, a genomic feature hypothesized to be under positive selection, conferred an advantage over other circulating lineages for a brief period of time (https://virological.org/t/repeated-loss-of-orf8-expression-in-circulating-sars-cov-2-lineages/931). Although the XZ recombinant contained the entire BA.2 Spike protein, it also contained the positively selected M:D3G mutation from BA.1; the combination thereof may have conferred an advantage over the canonical BA.2 early in 2022 [29].

Tracking emerging lineages, including recombinants, that have not yet been assigned a Pango lineage presents a challenge for public health authorities trying to monitor the increase and decrease of certain lineages over time. There are also occasions when a lineage has been designated but is unable to be assigned by the latest version of lineage-calling tools. Recently, there have been efforts to overcome this challenge, and web-based tools like UShER and Nextclade that are updated frequently have been transformative for SARS-CoV-2 genomic epidemiology [23,24]. However, many novel recombinant lineages that do not grow beyond a certain number of sequences remain undesignated. Retrospectively, it is understandable not to name every cluster of sequences that diverge from an ancestor, but the challenge remains when these sequences appear in near real-time.

Analysis of wastewater has become increasingly critical for SARS-CoV-2 epidemiology following the shifting landscape of SARS-CoV-2 testing towards at-home rapid antigen tests instead

 doi:10.20944/preprints202406.1937.v1

of PCR tests. Metagenomic analyses of SARS-CoV-2 in wastewater relies on detecting a specific set of mutations in designated lineages known as barcodes (https://github.com/andersen-lab/Freyja). It is also more difficult to identify novel lineages including recombinants amidst the background signal of other SARS-CoV-2 lineages in wastewater samples, emphasizing a continued need for WGS data from clinical samples in combination with metagenomic data from wastewater.

Some important limitations of this work apply to any study that aims to combine pathogen sequence data with epidemiologic data. Although epidemiologic information was included in this study, it was not available for every recombinant genome. When combined with the changes in reporting requirements as well as vaccine availability and recommendations throughout the pandemic, our ability to conduct statistical analyses and draw robust conclusions on associations between specific SARS-CoV-2 lineages and epidemiologic information was limited. While 59.9% of SARS-CoV-2 recombinant genomes matched to the CDPH databases, not all accompanying epidemiological information was available for each matched case due to data entry omissions and errors, staffing limitations, or changes in reporting requirements during the SARS-CoV-2 pandemic. The large portion of sequences that did not match to CDPH databases may be due to differences between sample identifiers in the various databases. For example, Virus Names in GISAID that did not match identifiers in CDPH databases may have resulted in duplicate entries of samples or the inability to link to epidemiologic information for corresponding samples. Of the recombinant genomes in this study that did not match to the CDPH databases, 787 (46.6%) were identified from GISAID. Therefore, we were unable to link more than half of all samples pulled from GISAID to the CDPH databases. Streamlining the naming schemes and reporting of testing results has proven to be difficult given the decentralized sequencing efforts in California. In the future, a standardized and automated system for linking sequence information from public repositories with epidemiologic records may improve public health surveillance efforts and allow for near real-time analyses of potential functional differences between SARS-CoV-2 lineages, including recombinants.

There are additional limitations to consider when drawing conclusions based on the prevalence of SARS-CoV-2 recombinants in California including 1) lack of uniformity in sequencing capacity throughout the study period (Figure 1b) that may have affected detection of recombinants, 2) sequence quality was not confirmed for samples from GISAID, which may impact the Pango lineage assignments, 3) retrospective submissions to GISAID after the study period with collection dates spanning 2020-2022 may have resulted in their exclusion from this study, 4) California Case Registry hospitalization and death statuses do not differentiate between hospitalizations and deaths with or for COVID-19, and 5) we were unable to differentiate between unvaccinated cases and cases with an unknown vaccination status.

Beyond the time period for this study, additional recombinants have emerged and public health scientists have continued to assess the potential impact of new and divergent lineages. Recently, the world has seen JN.1, a descendant of a second-generation BA.2 lineage, outcompete almost every other currently circulating lineage in a matter of months. However, previously circulating lineages recombined with JN.1 to form new recombinants. Recently, XDP, a recombinant between JN.1 and FL.15 (XBB.1.9.1.15) has emerged and is now increasing in prevalence worldwide. While FL.15 is no longer in circulation and never grew beyond a few thousand sequences, recombining with JN.1 in March 2024 gave rise to XDP, which was considered to have a growth advantage over JN.1. California COVIDNet continues to monitor XDP, as well as all SARS-CoV-2 recombinant lineages within the state.

This report highlights the successful detection and genomic characterization of recombinant SARS-CoV-2 lineages in California, as well as remaining challenges to public health on how best to detect, monitor, and respond to novel lineages. These challenges will likely persist on a global scale particularly in light of: 1) decreasing sequencing volumes following the expiration of public health emergencies worldwide, 2) the prevalence of at-home testing, 3) increasing number of co-circulating lineages resulting in the potential for more recombinants (https://cov-lineages.org/lineage_list.html), and 4) continued convergent evolution of SARS-CoV-2 making it more difficult to distinguish between lineages. These challenges highlight the need to maintain SARS-CoV-2 genomic surveillance

 doi:10.20944/preprints202406.1937.v1

at a sufficient level to detect emerging lineages and recombinants, to monitor changes in the virus, and to inform public health response and pharmaceutical interventions.

Supplementary Materials: The following supporting information can be downloaded at the website of this paper posted on Preprints.org, Figure S1: Percent Vaccinations over time; Table S1: List of available metadata for all samples.

Author Contributions: Conceptualization, Rahil Ryder and Emily Smith; Data curation, Rahil Ryder, Emily Smith and Deva Borthwick; Formal analysis, Rahil Ryder, Emily Smith and Deva Borthwick; Funding acquisition, Mayuri Panditrao and Debra A. Wadford; Investigation, Rahil Ryder and Emily Smith; Methodology, Rahil Ryder and Emily Smith; Project administration, Rahil Ryder and Emily Smith; Resources, Rahil Ryder, Emily Smith, Deva Borthwick, Jesse Elder, Christina Morales and Debra A. Wadford; Supervision, Mayuri Panditrao and Debra A. Wadford; Validation, Rahil Ryder and Emily Smith; Visualization, Rahil Ryder and Emily Smith; Writing – original draft, Rahil Ryder, Emily Smith and Deva Borthwick; Writing – review & editing, Rahil Ryder, Emily Smith, Deva Borthwick, Jesse Elder, Mayuri Panditrao, Christina Morales and Debra A. Wadford. All authors have read and agreed to the published version of the manuscript.

Funding: CDPH/COVIDNet genomic surveillance work was funded in part by the Centers for Disease Control and Prevention, Epidemiology and Laboratory Capacity for Infectious Diseases, Cooperative Agreement Number 5 NU50CK000539.

Institutional Review Board Statement: Ethical review and approval were waived for this work due to public health surveillance considerations during a public health emergency and deemed exempt by the Committee for the Protection of Human Subjects (Project number 2023-103) issued under the California Health and Human Services Agency's Federal Wide Assurance #00000681 with the Office of Human Research Protections.

Informed Consent Statement: Patient consent was waived due to public health surveillance considerations during a public health emergency and an IRB exemption for this work by the Committee for the Protection of Human Subjects (Project number 2023-103) as described above.

Data Availability Statement: Available sequencing data presented in the study are openly available in GISAID (https://gisaid.org/). GISAID Accessions are included in the Supplemental Table 1. Sequences not available in GISAID are available on request from corresponding author due to the location information being considered identifiable, or because GISAID had rejected the submitted genomes.

Acknowledgments: We gratefully acknowledge all data contributors, i.e., authors from the originating laboratories and the submitting laboratories who generated and shared genetic sequence data and metadata via GISAID, on which this report is based. We thank the California Association of Public Health Laboratory Directors, public health laboratorians, and California COVIDNet lab partners for their contributions to California COVIDNet. We thank the CDPH CalREDIE and CDPH Data Teams for their stewardship of COVID data throughout the pandemic emergency. We are grateful to John Bell, Esther Lim, and the CDPH COVID Clinical and CDPH Epidemiology Teams for their valuable feedback. Lastly, we thank the international team of volunteers that identifies and names new SARS-CoV-2 Pango lineages to benefit the world.

Disclaimer/Publisher's Note: The findings and conclusions in this article are those of the authors and do not necessarily represent the views or opinions of the California Department of Public Health or the California Health and Human Services Agency.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Bolze, A., Basler, T., White, S., Dei Rossi, A., Wyman, D., Dai, H., ... & Luo, S. (2022). Evidence for SARS-CoV-2 Delta and Omicron co-infections and recombination. *Med*, *3*(12), 848-859.
2. Pangilinan, E. A. R., Egana, J. M. C., Mantaring, R. J. Q., Telles, A. J. E., Tablizo, F. A., Lapid, C. M., ... & Saloma, C. P. (2023). Analysis of SARS-CoV-2 Recombinant Lineages XBC and XBC. 1 in the Philippines and Evidence for Delta-Omicron Co-infection as a Potential Origin. bioRxiv, 2023-04.
3. Perez-Florido, J., Casimiro-Soriguer, C. S., Ortuño, F., Fernandez-Rueda, J. L., Aguado, A., Lara, M., ... & Lepe, J. A. (2023). Detection of High Level of Co-Infection and the Emergence of Novel SARS CoV-2 Delta-Omicron and Omicron-Omicron Recombinants in the Epidemiological Surveillance of Andalusia. *International Journal of Molecular Sciences*, *24*(3), 2419.
4. Roemer, C. H., Hisner, R., Frohberg, N., Sakaguchi, H., Gueli, F., & Peacock, T. (2022). SARS-CoV-2 evolution, post-Omicron. *Virological. org*, *564*.

5. Eden, J. S., Tanaka, M. M., Boni, M. F., Rawlinson, W. D., & White, P. A. (2013). Recombination within the pandemic norovirus GII. 4 lineage. *Journal of virology*, *87*(11), 6270-6282.
6. Laver, G., & Garman, E. (2001). The origin and control of pandemic influenza. *Science*, *293*(5536), 1776-1777.
7. Simonsen, L. (2009). Geographic Dependence, Surveillance, and Origins of the 2009 Influenza A (H1N1) Virus. *Morb Mortal Wkly Rep*, *58*, 453-8.
8. Michaelis, M., Doerr, H. W., & Cinatl, J. (2009). Novel swine-origin influenza A virus in humans: Another pandemic knocking at the door. *Medical microbiology and immunology*, *198*, 175-183.
9. Nelson, M., Holmes, E. (2007). The evolution of epidemic influenza. Nat Rev Genet 8, 196–205.
10. Webster, R. G., & Laver, W. G. (1972). The origin of pandemic influenza. Bulletin of the World Health Organization, 47(4), 449.
11. Patiño-Galindo, J. Á., Filip, I., Chowdhury, R., Maranas, C. D., Sorger, P. K., AlQuraishi, M., & Rabadan, R. (2021). Recombination and lineage-specific mutations linked to the emergence of SARS-CoV-2. Genome medicine, 13(1), 124. https://doi.org/10.1186/s13073-021-00943-6.
12. Tang, X., Wu, C., Li, X., Song, Y., Yao, X., Wu, X., ... & Lu, J. (2020). On the origin and continuing evolution of SARS-CoV-2. National science review, 7(6), 1012-1023.
13. Lytras, S., Hughes, J., Martin, D., Swanepoel, P., de Klerk, A., Lourens, R., ... & Robertson, D. L. (2022). Exploring the natural origins of SARS-CoV-2 in the light of recombination. Genome Biology and Evolution, 14(2), evac018.
14. Bal, A., Simon, B., Destras, G., Chalvignac, R., Semanas, Q., Oblette, A., ... & Josset, L. (2022). Detection and prevalence of SARS-CoV-2 co-infections during the Omicron variant circulation, France, December 2021-February 2022. *Medrxiv*, 2022-03.
15. Evans, J. P., Qu, P., Zeng, C., Zheng, Y. M., Carlin, C., Bednash, J. S., ... & Liu, S. L. (2022). Neutralization of the SARS-CoV-2 deltacron and BA. 3 variants. *New England Journal of Medicine*, *386*(24), 2340-2342.
16. VanInsberghe, D., Neish, A. S., Lowen, A. C., & Koelle, K. (2021). Recombinant SARS-CoV-2 genomes circulated at low levels over the first year of the pandemic. Virus Evolution, 7(2), veab059.
17. Wadford, D. A., Baumrind, N., Baylis, E. F., Bell, J. M., Bouchard, E. L., Crumpler, M., Foote, E. M., Gilliam, S., Glaser, C. A., Hacker, J. K., Ledin, K., Messenger, S. L., Morales, C., Smith, E. A., Sevinsky, J. R., Corbett-Detig, R. B., DeRisi, J., & Jacobson, K. (2023). Implementation of California COVIDNet - a multi-sector collaboration for statewide SARS-CoV-2 genomic surveillance. Frontiers in public health, 11, 1249614. https://doi.org/10.3389/fpubh.2023.1249614.
18. Smith, E. A., Libuit, K. G., Kapsak, C. J., Scribner, M. R., Wright, S. M., Bell, J., ... & Wadford, D. A. (2023). Pathogen genomics in public health laboratories: Successes, challenges, and lessons learned from California's SARS-CoV-2 Whole-Genome Sequencing Initiative, California COVIDNet. *Microbial Genomics*, *9*(6), 001027.
19. Elbe, S. and Buckland-Merrett, G. (2017) Data, disease and diplomacy: GISAID's innovative contribution to global health. *Global Challenges*, 1:33-46. https://doi.org/10.1002/gch2.1018 PMCID: 31565258.
20. Khare, S., Gurry, C., Freitas, L., Schultz, M. B., Bach, G., Diallo, A., Akite, N., Ho, J., Lee, R. T., Yeo, W., Curation Team, G. C., & Maurer-Stroh, S. (2021) GISAID's Role in Pandemic Response. *China CDC Weekly*, *3*(49): 1049-1051.
21. Shu, Y. and McCauley, J. (2017) GISAID: From vision to reality. *EuroSurveillance*, 22(13).
22. O'Toole, Á., Scher, E., Underwood, A., Jackson, B., Hill, V., McCrone, J. T., ... & Rambaut, A. (2021). Assignment of epidemiological lineages in an emerging pandemic using the pangolin tool. *Virus evolution*, *7*(2), veab064.
23. Hadfield, J., Megill, C., Bell, S. M., Huddleston, J., Potter, B., Callender, C., ... & Neher, R. A. (2018). Nextstrain: Real-time tracking of pathogen evolution. *Bioinformatics*, *34*(23), 4121-4123.
24. Turakhia, Y., Thornlow, B., Hinrichs, A. S., De Maio, N., Gozashti, L., Lanfear, R., ... & Corbett-Detig, R. (2021). Ultrafast Sample placement on Existing tRees (UShER) enables real-time phylogenetics for the SARS-CoV-2 pandemic. *Nature Genetics*, *53*(6), 809-816.
25. Huddleston, J., Hadfield, J., Sibley, T. R., Lee, J., Fay, K., Ilcisin, M., ... & Hodcroft, E. B. (2021). Augur: A bioinformatics toolkit for phylogenetic analyses of human pathogens. *Journal of open source software*, *6*(57).
26. Singh P, Sharma K, Shaw D, Bhargava A, Negi SS. (2022). Mosaic Recombination Inflicted Various SARS-CoV-2 Lineages to Emerge into Novel Virus Variants: A Review Update. Indian J Clin Biochem. 1-8.
27. Tamura, T., Ito, J., Uriu, K., Zahradnik, J., Kida, I., Anraku, Y., ... & Sato, K. (2023). Virological characteristics of the SARS-CoV-2 XBB variant derived from recombination of two Omicron subvariants. *Nature communications*, *14*(1), 2800.
28. Uriu, K., Ito, J., Zahradnik, J., Fujita, S., Kosugi, Y., Schreiber, G., ... & Sato, K. (2023). Enhanced transmissibility, infectivity and immune resistance of the SARS-CoV-2 Omicron XBB. 1.5 variant. *bioRxiv*, 2023-01.
29. Bloom, J. D., & Neher, R. A. (2023). Fitness effects of mutations to SARS-CoV-2 proteins. *bioRxiv*, 2023-01.
30. Ghafari, M., Hall, M., Golubchik, T., Ayoubkhani, D., House, T., MacIntyre-Cockett, G., ... & Lythgoe, K. (2024). Prevalence of persistent SARS-CoV-2 in a large community surveillance study. *Nature*, 1-8.

 doi:10.20944/preprints202406.1937.v1

31. Chaguza, C., Hahn, A. M., Petrone, M. E., Zhou, S., Ferguson, D., Breban, M. I., ... & Grubaugh, N. D. (2023). Accelerated SARS-CoV-2 intrahost evolution leading to distinct genotypes during chronic infection. *Cell Reports Medicine*, *4*(2).
32. Gonzalez-Reiche, A. S., Alshammary, H., Schaefer, S., Patel, G., Polanco, J., Carreño, J. M., ... & van Bakel, H. (2023). Sequential intrahost evolution and onward transmission of SARS-CoV-2 variants. *Nature Communications*, *14*(1), 3235.
33. Pickering, B., Lung, O., Maguire, F., Kruczkiewicz, P., Kotwa, J. D., Buchanan, T., ... & Bowman, J. (2022). Divergent SARS-CoV-2 variant emerges in white-tailed deer with deer-to-human transmission. *Nature Microbiology*, *7*(12), 2011-2024.
34. Sparrer, M. N., Hodges, N. F., Sherman, T., VandeWoude, S., Bosco-Lauth, A. M., & Mayo, C. E. (2023). Role of Spillover and Spillback in SARS-CoV-2 Transmission and the Importance of One Health in Understanding the Dynamics of the COVID-19 Pandemic. *Journal of Clinical Microbiology*, e01610-22.