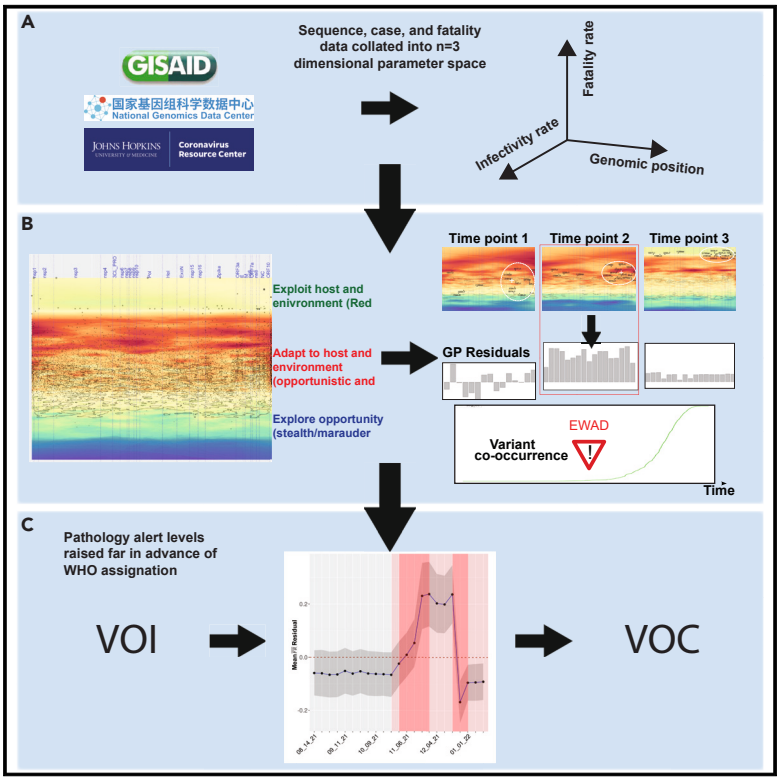
Patterns

Understanding the host-pathogen evolutionary balance through Gaussian process modeling of SARS-CoV-2

Graphical abstract



Highlights

- An early warning detection system for potential SARS-CoV-2 variants of concern
- GP-based spatial covariance applied to whole genome architecture of the virus
- GP residuals and mutation co-occurrences allow for monitoring pathology alert levels

Authors

Salvatore Loguercio, Ben C. Calverley, Chao Wang, ..., Shuhong Sun, G.R. Scott Budinger, William E. Balch

Correspondence

webalch@scripps.edu

In brief

A data-driven early warning detection system is developed for variants of concern in SARS-CoV-2. The model builds on Gaussian process regression and variant co-occurrence, is computationally efficient, and enables the authors to identify variants of concern at times months before their WHO assignation—hence of high interest for real-time variant surveillance. It also gives information about the nature of the variant and its potential fatality impact. This modeling method could easily be applied to other areas of disease and viruses.

Loguercio et al., 2023, Patterns *4*, 100800
August 11, 2023
https://doi.org/10.1016/j.patter.2023.100800

Patterns

CellPress
OPEN ACCESS

Article

Understanding the host-pathogen evolutionary balance through Gaussian process modeling of SARS-CoV-2

Salvatore Loguercio,[1,3,5] Ben C. Calverley,[1,5] Chao Wang,[1,4] Daniel Shak,[1] Pei Zhao,[1] Shuhong Sun,[1] G.R. Scott Budinger,[2] and William E. Balch[1,6,*]

[1]Department of Molecular Medicine, Scripps Research, La Jolla, CA, USA
[2]Division of Pulmonary and Critical Care Medicine, Northwestern University, Chicago, IL, USA
[3]Present address: Scripps Research Translational Institute, La Jolla, California
[4]Present address: Institute of Systems and Physical Biology, Shenzhen Bay Laboratory, Shenzhen, China
[5]These authors contributed equally
[6]Lead contact
*Correspondence: webalch@scripps.edu
https://doi.org/10.1016/j.patter.2023.100800

THE BIGGER PICTURE A critical goal for managing the rapid evolution of the SARS-CoV-2 pandemic is the ability to anticipate in advance the next viral strain that will compromise human health—the "host-pathogen" balance. The drivers of the pandemic are variants of concern (VOCs), virus strains that sequentially achieve dominance using unique patterns of genetic mutations leading to improved fitness. To discover the emergent pattern of VOCs, we developed a new AI tool—early warning anomaly detection (EWAD). EWAD provides a heads-up weeks to months in advance of what the next VOC may look like, helping us to anticipate response measures that tip the host-pathogen balance to favor the host. The pattern recognition algorithm enabling EWAD has important implications beyond the COVID-19 pandemic. It provides us with a "standard model" to understand the emergence of new pandemics as well as to understand mechanistically the genetic variation impacting human health ranging from cancer to neurodegeneration.

Development/Pre-production: Data science output has been rolled out/validated across multiple domains/problems

SUMMARY

We have developed a machine learning (ML) approach using Gaussian process (GP)-based spatial covariance (SCV) to track the impact of spatial-temporal mutational events driving host-pathogen balance in biology. We show how SCV can be applied to understanding the response of evolving covariant relationships linking the variant pattern of virus spread to pathology for the entire SARS-CoV-2 genome on a daily basis. We show that GP-based SCV relationships in conjunction with genome-wide co-occurrence analysis provides an early warning anomaly detection (EWAD) system for the emergence of variants of concern (VOCs). EWAD can anticipate changes in the pattern of performance of spread and pathology weeks in advance, identifying signatures destined to become VOCs. GP-based analyses of variation across entire viral genomes can be used to monitor micro and macro features responsible for host-pathogen balance. The versatility of GP-based SCV defines starting point for understanding nature's evolutionary path to complexity through natural selection.

INTRODUCTION

The coronavirus disease 2019 (COVID-19) caused by SARS-CoV-2 rapidly expanded to a global pandemic that has impacted over 600 million people and led to the death of a projected 6 million individuals.[1] The pathology rate was dominated (>75%) by the over ~60 years age group.[2–14] Given that Alpha, Delta, and Omicron variants of concern (VOCs) generated worldwide social and economic disruption,[15] understanding the evolution of the whole genome architecture (WGA) of SARS-CoV-2 in

response to its global genetic variation and the host biological responses is critical for understanding the fitness balance in host-pathogen relationships that lead to fatality in the aging population.[16–21]

Considerable evolution of the SARS-CoV-2 genome (~30,000 base pairs [bp]) has occurred since its initial emergence, leading to multiple VOCs.[22,23] VOCs are assigned lineage importance based on allele frequency (i.e., a variant found in >75% of strains analyzed) reflecting the dominant spread of a lineage in the worldwide population and therefore potential for impact on pathophysiology.[22] VOC assignments do not necessarily reflect all of the mutational features responsible for real-world pathology in the human population, a concern we now refer to and define hereafter as "variant dark matter," where undesignated variants contribute significantly to the evolution of the many different viral lineages that are traced by hierarchical mapping.[22,24,25] There is a need for a systematic approach toward assessing the entire variational landscape in the context of real-world infection and fatality information, to better understand SARS-CoV-2 and its evolutionary race for host-pathogen dominance.

Gaussian process (GP) is a universal non-parametric regression machine learning (ML) approach used to interpolate a variable over a range when given a sparse collection of known sample *inputs*. The *output* gives a quantitative value and an associated uncertainty for every unknown point across the range sampled for input. GP regression-based interpolation is a tool used widely in a variety of disciplines,[26–28] including geostatistics,[29,30] astronomy,[31] and finite element method mathematical modeling. GP-based covariance relationships provide a computational framework where "distance" separation between points in space can be parameterized by two or more other variables to achieve an understanding of complex environments bounded by the points in space and time.

To address the role of worldwide genetic variation in human inherited genetic disease, we have developed a new approach using GP,[29] referred to as spatial covariance (SCV).[32,33] SCV makes use of mutations in the genome leading to changes in amino acid residues responsible for the protein fold. This helps us understand sequence-to-function-to-structure SCV relationships driving health and disease.[32–37] SCV relationships provide a universal approach to capture probabilistic phenotype outcomes—with assigned uncertainty—that contribute to altered protein function[32,33,37,38] and response to therapeutics[34–36] across the entire protein sequence.

In the evolution of RNA viruses, increased variant frequency in the population can arise from genetic drift due to random events, as well as from positive selection reflecting Darwinian principles of fitness.[39] Connecting these mutations to viral life cycle components is a difficult task due to a lack of system-based approaches that link mutations to functional outcomes, and to diversity in the population.[40] For example, linking large databases that report strictly on the hierarchical mapping of SARS-CoV-2 genomes and their mutations[22,24,25] to those that report experimental and clinical outcomes in the population are in their infancy.[15,41–44]

Using the rapidly evolving collection of mutant alleles contributing to the WGA of SARS-CoV-2 lineages, we apply GP to generate unprecedented "allele phenotype landscapes." Allele phenotype landscapes describe the role of all allele positions in the SARS-CoV-2 genome across the pandemic from the first Wuhan strain to the recent Omicron strains (a total of 5,600,000 sequences over 724 days). They define SCV relationships that link spread to pathology and fatality. These results are evaluated in the context of analysis of co-occurring mutations[45] ("co-occurrence") across the same time frame to extract insights into the evolution of the SARS-CoV-2 WGA. We find that each VOC evolves different GP-based "search" strategies over time. Importantly, a joint analysis of co-occurrences and residuals extracted from the GP-based SCV maps that report on "actual" versus "predicted" changes provides an early warning anomaly detection (EWAD) system for the emergence of VOCs. EWAD provides an unprecedented view of features spanning the entire WGA, from initiating variant dark matter to current Omicron strains, features that drive host-pathogen system dynamics—referred to as the "Red Queen" effect.[46] EWAD enables a fresh view of the host-pathogen dynamics responsible for emergent VOCs in the context of variant dark matter that is hidden from consideration in conventional hierarchical mapping. It provides a performance map to assess progression from pandemic to endemic states by the changing dynamics of spread-fatality covariance.

We posit that knowledge of genome-based SCV relationships[32–34,37,38] linking SARS-CoV-2 genotypic diversity to host phenotypic diversity provides an unanticipated platform to address the "Red Queen" effect[46] responsible for natural selection leading to spread and fatality in the aging population.[17,19–21,47] GP-based SCV provides us with a new view of the pandemic as a collective global phenomenon based on evolutionary conserved covariant relationships that generate biology and drive host-pathogen balance.[32–38,48]

RESULTS

Standard analysis shows no correlation between pathology and genomic position

Despite the abundance of data from many sources, relating genotype to phenotype faces many challenges, since data are overall fragmented, heterogeneous, and miss rigorous connections that incorporate the potential of disease contribution across the entire genome. To address this problem, we first built data processing pipelines that feed directly from both clinical outcomes/phenotypes[49] and viral mutations recorded in sequence files deposited worldwide (Global Initiative on Sharing All Influenza Data [GISAID][47] aggregated at Chinese National Center for Bioinformation 2019 Novel Coronavirus Resource [CNCB][40]). We integrated temporal geospatial data with mutation-specific data between March 2020 and March 2022 capturing the spread of the virus up to the Omicron VOC on a daily basis (see methods).

To understand spread, pathology, and disease management throughout the SARS-CoV-2 virus genome, we used this worldwide dataset of mutations, assessed by allele frequency, along with two composite variables—allele frequency-weighted infectivity rate ($\overline{IR}\,(V)$) and allele frequency-weighted pathology fatality rate ($\overline{FR}\,(V)$) (abbreviated as $\overline{IR}$ and $\overline{FR}$) (see methods). We use this algorithm to understand fitness of the virus in the population, reflecting host-pathogen relationships and thereby what we posit are the conserved evolutionary rules of viral WGA

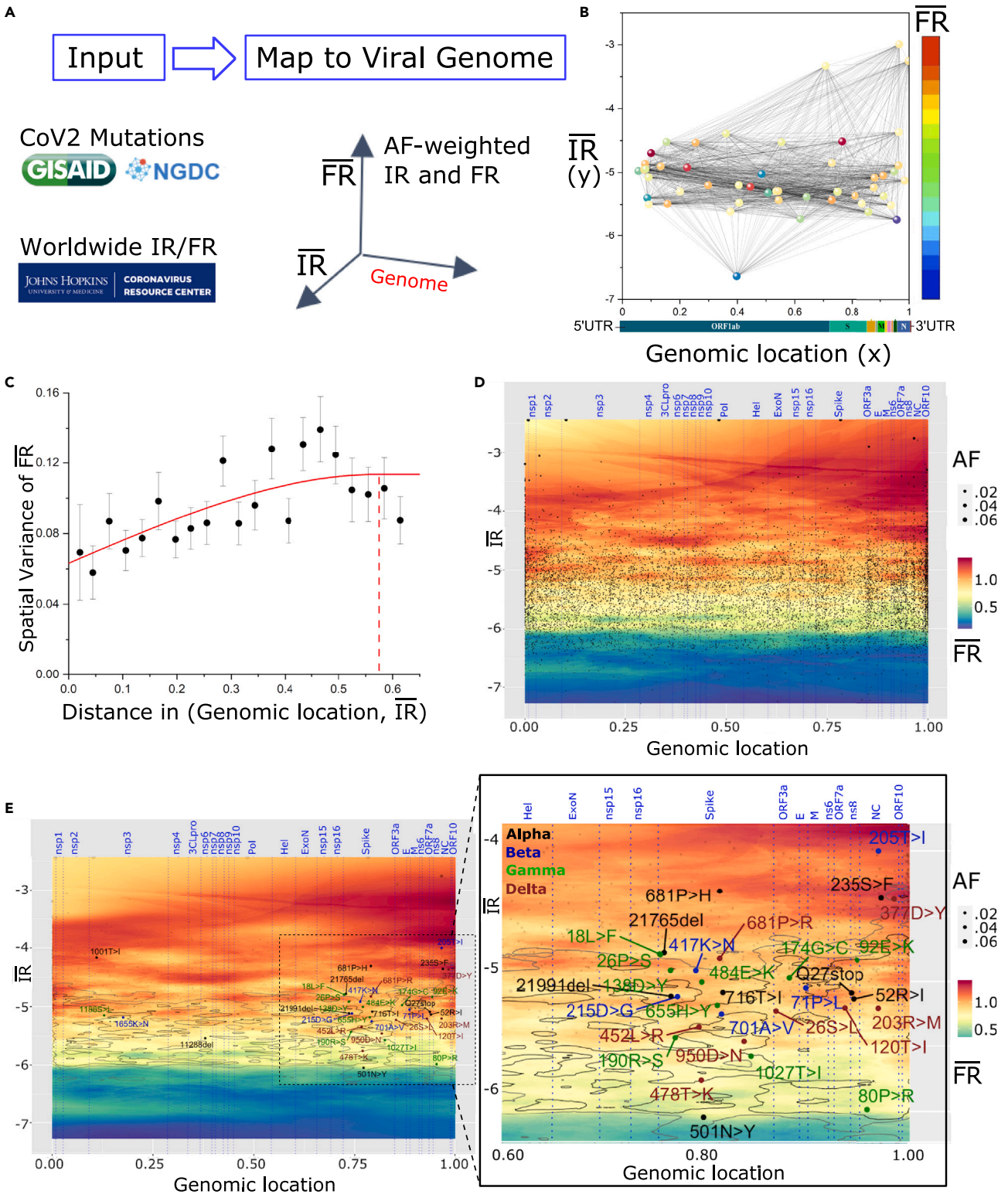


Figure 1. Illustration of the GP regression approach

(A) Data ingestion and pre-processing. Genotypic data from SARS-CoV-2 isolates and phenotypic data (cases and deaths) are gathered daily from NGDC and JH resource, respectively. For each reported SARS-CoV-2 mutation, allele frequency-weighted $\overline{IR}$ and $\overline{FR}$ are computed.

(B) SARS-CoV-2 mutations (alleles) are positioned by their genomic positions (x axis) and $\overline{IR}$ (y axis) and colored by $\overline{FR}$ (z axis). The pairwise spatial relationships (indicated by black lines) are analyzed by GP regression. Shown is a simplified plot showing only 50 mutations for clarity.

(legend continued on next page)

worldwide. This is in contrast to the more mutation-focused hierarchical clustering approaches that supply important information on regional distributions of spread in terms of explicit case counts.[22]

Strikingly, standard correlation analysis between the above variables and genomic position shows that there is no detectable correlation between $\overline{IR}/\overline{FR}$ and genomic position (Figures S1A and S1B), and the two variables themselves are poorly correlated (Figures S1C and S1D, Pearson r = 0.1; log-transformed, Pearson r = 0.4). Therefore, a new systems-based approach is needed to integrate these features and provide a meaningful description of WGA and its contribution to the evolving balance in pathogen and host fitness. This is important to understand disease progression across the worldwide population.

SCV analysis linking SARS-CoV-2 mutations with real-world infection and fatality

To understand the impact of genotype to phenotype relationships from a WGA perspective, we took advantage of our previous work in inherited rare disease.[32–37] Variation distributed in the worldwide population provides a platform to dissect pathology and therapeutic management through spatial covariance (SCV).[32–37] SCV utilizes the sample population to provide a unique lens to focus on the functional impact of a variant in the individual using GP regression ML. GP-based SCV utilizes the sparse distribution of variation in the genome of the worldwide population as *input* to assess as *output* the SCV of relationships found for each protein that ties genotype to phenotype for every residue in the polypeptide sequence. SCV relationships in inherited disease define the strength of covarying functional features encoded by its evolving allele composition through weighted proximity—thereby relating sequence to functional features. SCV relationships allow us to deduce both residue-by-residue and complex residue-residue multi-dimensional interactions that can be used to describe protein function-structure relationships and their contribution to environmental fitness at atomic resolution (see methods).[32–34,38]

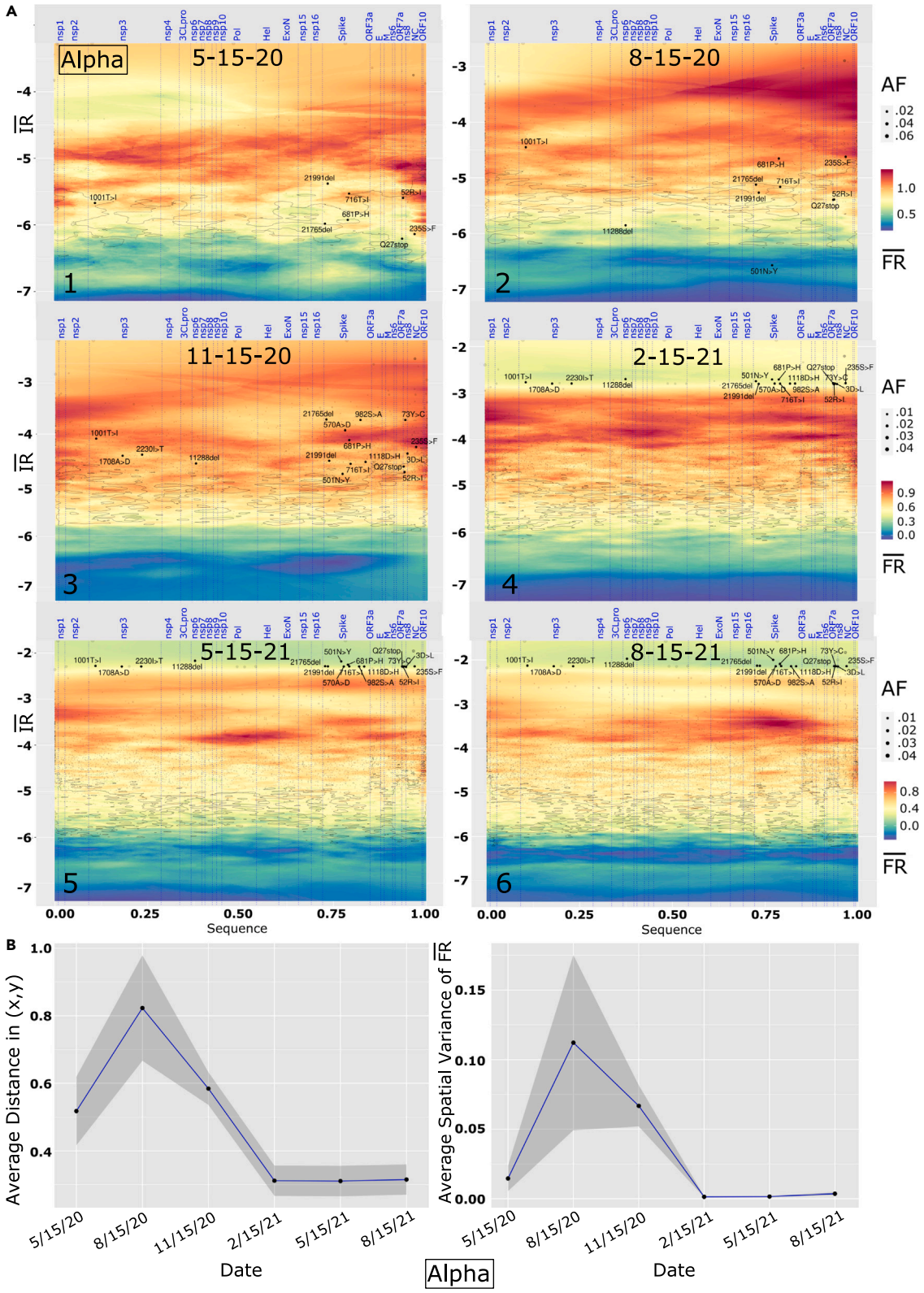
To address the impact of WGA of SARS-CoV-2 on host-pathogen balance driving the spatial-temporal dynamics of spread and pathology, we applied GP to generate allele-based phenotype landscapes that describe on a nucleotide-by-nucleotide basis the SCV relationships linking viral spread to pathology across the entire ~30,000-bp SARS-CoV-2 genome (Figure 1). Figures 1A–1C shows the process for generating the SCV relationships. These SCV relationships are defined by three axes that include (1) allele genome position (*x axis*)), (2) allele frequency-weighted infectivity rate ($\overline{IR}$) (spread) (*y axis*), and (3) allele frequency-weighted fatality pathology rate ($\overline{FR}$) (*z axis*, *color*) (Figure 1A). These three features, when generated in the context of time, allow us to quantitatively image the spatial-temporal features contributing to spread and pathology through the different phases of the pandemic as a covariant collective of the designated VOCs in the universal context of the global mutation load (see methods for full details), visualized in daily SCV landscapes (Figures 1D and 1E, with zoomed inset).

The expanding time frame of SCV relationships

Focusing on the first stages of the pandemic up through to the time of the Delta VOC, we show six tri-monthly snapshots from the time lapses, each taken mid-month and annotated for the set of signature Alpha VOC (Figure 2A, panels 1–6, analysis of distance and variance in Figure 2B) and the Delta VOC (Figure 3A, panels 1–6, analysis of distance and variance in Figure 3B). See supplemental information for Beta and Gamma VOCs (Figures S2 and S3) and supplemental videos 1–5 for the full time lapses for each VOC. Beginning with 5/15/20 for the Alpha VOC (Figure 2A, panel 1), only a few mutations are reported including a mutation in nsp3 (T1001I), four in Spike (21991del, 21765del, P681H, T716I), and two mutations in the nucleocapsid (NC) gene (R52I, S235F), both in areas of relatively high $\overline{FR}$ (Figure 2A, orange to red). The next time point 3 months later (Figure 2A, panel 2) reveals the emergence of an important Spike mutation (501N- > Y) at the bottom of the map with very low $\overline{IR}/\overline{FR}$, while the rest of the defining mutations for Alpha VOC are already migrating to higher $\overline{FR}$ areas. By 11/15/20 (Figure 2A, panel 3), all assigned mutations for Alpha VOC are present, and all in relatively high $\overline{FR}$ regions reflecting their initial impact on pathology as a potential opportunist in a naive host environment, particularly the aging population. From this time point onward, the distribution of mutations on the map increasingly "compacts" where the relative GP defined SCV relationships between mutations get smaller, both at the 5′ and 3′ ends of the SARS-CoV-2 genome (Figure 2A, panels 7 and 8). These VOC-containing clusters, as a covariant collective, migrate toward the top of the allele phenotype landscape with higher $\overline{IR}$ suggestive of cooperation in WGA features impacting both spread and fatality (Figure 2A, panels 4–6). However, their collective migration occurs in the context of the increasing number of variant dark matter variants that retune the GP-based allele phenotype landscape features over time (Figure 2A, black dots). These hidden supporting residues found in the global population contribute to the observed lineage diversification from our GP-based SCV global perspective.

The primary factor responsible for the observed SCV compaction is the increase in the number of Alpha VOC detections in response to population testing leading to higher allele frequency weights in $\overline{IR}$; however, we still detect significant changes in the SCV relationships of all VOC mutations in response to the changing local environments revealed by predicted $\overline{FR}$ (Figure 2A,

(C) As a first step in GP regression modeling, a variogram is computed (an illustrative example is depicted) showing spatial relationships between the separation distance of paired data points in (B) (x axis) and the spatial variance of $\overline{FR}$ relative to $\overline{IR}$ (y axis).
(D and E) GP regression maps of genomic position (x axis), and log-transformed $\overline{IR}$ (y axis) and $\overline{FR}$ (color scale, z axis) for SARS-CoV-2 genome (data in this example are for 9/15/20). $\overline{FR}$ is predicted across the whole landscape according to the variogram computed in (C), where output is an average of surrounding sample points, weighted by a function of distance given by the variogram.
(D) Black dots represent variant input values used to compute GP regression, with dot sizes proportional to the allele frequency of the mutations. Vertical dotted blue lines are boundaries between SARS-CoV-2 proteins, annotated on the top axis.
(E) Input variants are shaded light gray for clarity of VOCs. Contour lines are drawn at 10% and 25% percentiles of global variance estimated for model predictions (C). Labels on the map are signature mutations for Alpha (black), Beta (blue), Gamma (green), and Delta (brown). The zoomed inset shows the region with most VOC mutations in more detail, with all input mutations and contours that are used to train the GP regression model.



(legend on next page)

panels 1–6, z axis). Intriguingly, by 2-15-21 (Figure 2A, panel 4), we already notice that the mutations in the 3′ region of the SARS-CoV-2 genome have migrated to a low $\overline{FR}$ region and this becomes more evident in successive snapshots, where all Alpha mutations lay in a high $\overline{IR}$/low $\overline{FR}$ region at the top of the allele phenotype landscape. This trend likely reflects the beginning of the impact of host countermeasures, including innate and adaptive immune responses, vaccine availability, and physical interventional measures such as masks and social distancing, clearly illustrating the impact on fatality but less so on the global spread of variants seen in the continued increase of $\overline{IR}$ at this time frame (Figure 2A, panels 4–6).

For Delta VOC, the early time point (Figure 3A, panel 1 [(5/15/20]) has two NC mutations already in high $\overline{FR}$ (203R- > M, 377D- > Y), similar to what was observed with Alpha (Figure 2A, panel 1). These mutations are in two of the three disordered domains of NC (LINK and CTD) where most protein-protein and protein-RNA interaction sites occur.[50] By 11/15/20 (Figure 3A, panel 3), almost all signature mutations for Delta VOC are in relatively high $\overline{FR}$ regions and this continues through the next time point (Figure 2B, panel 4: 2/15/21) where the mutations are now more spread over the $\overline{IR}$ axis than was observed in Alpha (compare Figure 2A, panel 4, with Figure 3A, panel 4). There are almost two orders of magnitude of difference in $\overline{IR}$ between 478T- > K (Spike) and 377D- > Y (NC) pointing to the fact that there is a wide difference in spread of signature mutations in Delta VOC reflecting different evolutionary trajectories promoting success. These results, along with the emergent variant dark matter, suggest that the Delta VOC is still evolving in the context of the worldwide population as a "predator," whereas Alpha variants have comparable values at these time points, suggesting that a consolidation of WGA function has been achieved and curiously, appears to be the endpoint in its race for fitness through SCV relationships—reflecting the limitations of its GP evolved WGA.

In the last two snapshots (Figure 3A, panels 5 [(5/15/21] and 6 [(8/15/21]), the Delta VOCs are now located in a high $\overline{IR}$, low $\overline{FR}$ cluster, as seen for Alpha but at an even lower $\overline{FR}$/$\overline{IR}$. Counts for Delta VOC, as of May 2021, were at least 100 times lower than Alpha VOC (Figure S4), reflecting the lower cumulative spread of Delta compared with Alpha VOC at this point in time, although Delta's subsequent surging prevalence in the pandemic dwarfed the real-time values of the Alpha VOC, suggestive of SCV-based optimization for spread. For Beta and Gamma VOCs, we observe more of a mixed behavior over time compared with the trajectories of Alpha VOC and Delta VOC with only few mutations (e.g., 501N- > Y) moving into high $\overline{IR}$, low $\overline{FR}$ clusters at later time points (Figures S2 and S3).

Combined, these results capture the striking VOC divergence in allele phenotype landscape features in the context of the global pandemic. They suggest that these VOCs are progressively channeled into unique collectives of SCV relationships in the context of an increasingly more combative host response environment, limiting their capacity to achieve further prominence on a worldwide scale.

Temporal co-occurrence patterns give a way of tracking VOC emergence

To provide an alternative approach to the patterns of emergence captured through GP analyses and to augment and validate our GP approach, we generated a comprehensive view of mutation co-occurrence across the entire arc of the SARS-CoV-2 pandemic. For each virus isolate, sequence alignment against the reference SARS-CoV-2 genome reports the mutations called on that particular sequence. Such alignments can be mined systematically for co-occurrence of mutations over time, where co-occurrence is simply the count of mutations occurring together on the same viral sequence (see methods). Co-occurrence analyses let us track evolvability, reflecting increasing mutational burden in genome variation events, events that likely contribute to lineage focus on spread and pathology.

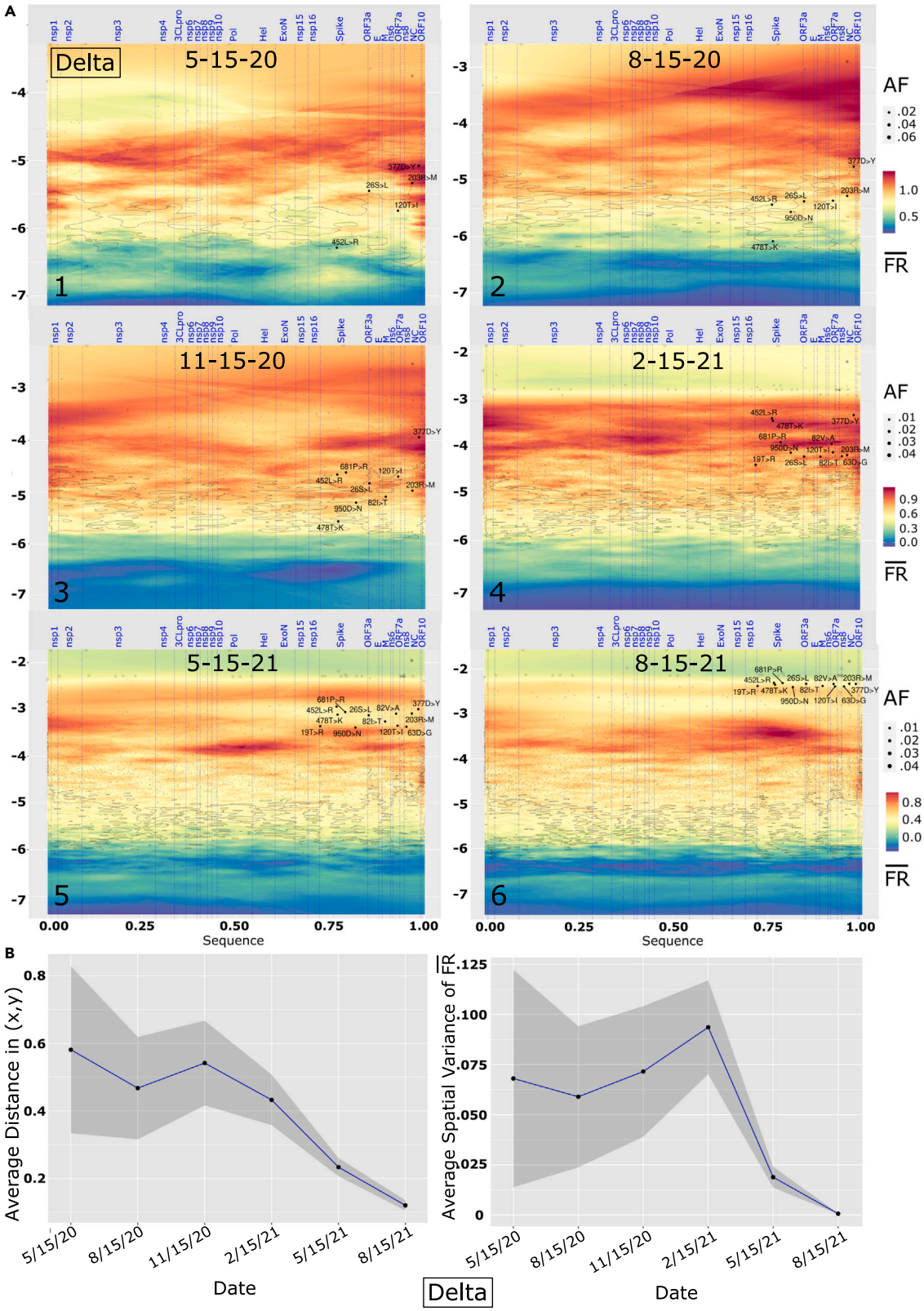
To highlight co-occurrences driving emergence of VOCs Alpha, Beta, Gamma, and Delta, we first tracked cumulative co-occurrences among the defining mutations for every day starting 9/15/21 up to 8/22/21 and computed a daily average co-occurrence by averaging all cumulative co-occurrence values available daily for each of the VOCs. Average co-occurrence over time for the four VOCs is shown (Figure 4A, left panel; zoom of Beta/Gamma/Delta VOC timeline plots shown in the right panel). Differences between VOCs are further evident in the patterning of their representative co-occurrence matrices (Figure 4B). The first VOC to emerge, Alpha, has a uniform distribution of co-occurrences across the signature mutations of this viral strain, suggestive of an opportunistic break in its evolving host-pathogen encounters. This is not the case for later emergent VOCs. See supplementary results for more details.

To track the actual number of co-occurrences between signature mutations of a VOC versus all possible co-occurrences at a specific time point, we defined VOC co-occurrence density as the ratio between the number of non-zero co-occurrences in the VOC co-occurrence matrix (Figure 5, lower panels in each VOC) over the total number of possible co-occurrences. Here, an empty VOC co-occurrence matrix has a density of 0, and one with no zeroes has a density of 1. As an example, we computed co-occurrence density for each of the four VOCs between September 2020 and August 2021 (Figure 5, lower panel for each VOC). Analysis of co-occurrence density curves for each VOC over time reveals that for the two early-onset VOCs (Alpha and Beta), the VOC co-occurrence density has an "all

Figure 2. Time lapse of viral genome allele phenotype landscapes

(A) Alpha VOC showing six tri-monthly time points between May 2020 and August 2021. $\overline{\boldsymbol{IR}}$ (y axis) and $\overline{\boldsymbol{FR}}$ (z axis) are log-transformed, genomic position is scaled to a (0–1) scaled from 5′ to 3′ of the RNA sequence encompassing ~30,000 bp with 5′-end located at the origin of the x axis. Vertical dotted blue lines are boundaries between SARS-CoV-2 proteins, annotated on the top of each figure. Input variants are in shaded color, with dot sizes proportional to allele frequency of the mutations. Contour lines are drawn at 10% and 25% percentile of global variance estimated for model predictions (Figure 1C).

(B and C) (Left) Average distance between Alpha VOC signature mutations defined by x axis (genomic position) and y axis coordinates ($\overline{\boldsymbol{IR}}$) as described in Figure 1B for each of the six time points shown in (A). The gray ribbon marks the 95% confidence interval. (Right) Average spatial variance of $\overline{\boldsymbol{FR}}$ (z axis) between Alpha signature mutations defined by x axis 'Genome position' and y axis coordinates ($\overline{\boldsymbol{IR}}$) as described in Figure 1B for each of the six time points shown in (C). The gray ribbon marks the 95% confidence interval.



(legend on next page)

or none" behavior—going from near zero to one in a single day (Figure 5, lower panels for Alpha and Beta VOC). This event is many weeks ahead of the time where VOC average co-occurrences enter a fast growth phase (Figure 5, upper panels Alpha and Beta, green curves).

In contrast, for the later VOCs (Gamma and Delta), co-occurrence densities exhibit an exploratory behavior where in the early phase few co-occurrence options are explored for a longer time interval before the jump to the full spectrum of co-occurrences, reflecting its predatory behavior. Here, the time at which the jump is observed is nearly proximal to the beginning of the sustained growth phase where the increasing slope of the co-occurrence average curve starts (Figure 4A, e.g., green curve, Gamma). This contrasts with the "all or none" VOC co-occurrence densities for Alpha and Beta (Figure 5, lower Gamma and Delta panels). An intermediate level of acquisition of co-occurrence is particularly evident within the Delta VOC capturing co-occurrence links for months up to 0.4–0.5 (40%–50%) of co-occurrence discovered prior to the jump to the full co-occurrence set. To track the spread and pathology of each VOC in the context of evolving host responses, we performed a joint analysis of all VOCs by tracking their signature mutations through their genomic co-occurrences to provide detailed insight into the convergence of co-occurrence for each possible co-occurrence seen in the pandemic (Figure S5). These results suggest that the virus WGA is evolving improved search strategies over time across a seemingly intractable number of mutation-sensitive co-occurrences influenced by the supporting variant dark matter. Thus, evolving VOC co-occurrence relationships highlight potential pathogen fitness strategies that likely contribute to evolutionary success or failure. These potential strategies are missing from existing perspectives of viral spread and pathology.

GP residuals provide a clear early warning of VOC emergence

Because the co-occurrence analysis alone is not linked to the real-world infection and fatality features, it is not enough to inform whether a certain combination of mutations will eventually result in a VOC that challenges the evolved/evolving host responses that, as a covariant collective, ultimately dictates virus spread.

To track the spread and pathology of each VOC in the context of evolving host responses, we performed a joint analysis of all VOCs by tracking their signature mutations through their genomic co-occurrences (Figure S5) over time in conjunction with GP-based $\overline{IR}$ and $\overline{FR}$ SCV relationships (Figures 2 and 3). Specifically, we examined whether allele phenotype landscapes could be used as an EWAD[51] system for the emergence of VOCs reflected in GP principled relationships dictating global spread and pathology. An EWAD system for spread and pathology is looking for a signal that ideally changes significantly during the early phase of the transition, reaches a maximum, and then quenches when the phenomenon enters a steady state, reflecting accomplishment of the optimized goal across the population, often leading to its diminution in the host population due to the Red Queen effect.[52]

To assess the potential of an emergent EWAD signal as the SARS-CoV-2 evolves in response to host countermeasures, we focused on "$\overline{FR}$ *residuals*." In GP modeling at a given time point in the pandemic (Figure 6A, x axis), *$\overline{FR}$ residuals are defined as the difference between the observed and predicted $\overline{FR}$ values* (Figure 6A, y axis, observed $\overline{FR}$ minus predicted $\overline{FR}$). The observed $\overline{FR}$ for a mutation is the explicit assigned $\overline{FR}$ of that mutation used for the *input* data in GP that does not incorporate the impact of other mutations, while the predicted $\overline{FR}$ for the mutation generated by GP is a proximity weighted average of the observed $\overline{FR}$ value in the context of its surrounding mutations in the phenotype landscape that includes the hidden relationships driven by the variant dark matter. If the observed $\overline{FR}$ is lower than the predicted $\overline{FR}$ for a mutation (i.e., a negative GP residual value [Figure 6A, left panel]), it indicates that the $\overline{FR}$ for the mutation is lower than the $\overline{FR}$ of surrounding mutations in the phenotype landscape, therefore *underperforming* relative to the surrounding VOC/variant dark matter in terms of $\overline{FR}$. In contrast, if the observed $\overline{FR}$ is higher than the predicted $\overline{FR}$ for a mutation (i.e., a positive GP residual value [Figure 6A, left panel]), it indicates that this mutation has a higher $\overline{FR}$ than its surrounding mutations—therefore *overperforming* in its pathology as defined by $\overline{FR}$ relative to the surrounding VOC/variant dark matter. Overperformance is consistent with the interpretation that is emerging as a prominent player in disease pathology as measured by $\overline{FR}$. These relationships we take as our definitions of under- or overperformance, reflecting how a variant's actual pathology ($\overline{FR}$) compares with its GP-based pathology as predicted by the surrounding variant dark matter. In addition, the mean $\overline{FR}$ residual's location above/below the baseline can tell us something about the nature of the impending cluster of variants reflecting their potential impact on pandemic progression. Therefore, the GP-based $\overline{FR}$ residual value represents a real-time monitor of the collective behavior of the global virus system versus the actual sparse measurements defined by a VOC mutation designated at the 75% frequency level used by the epidemiological community to track strain importance.

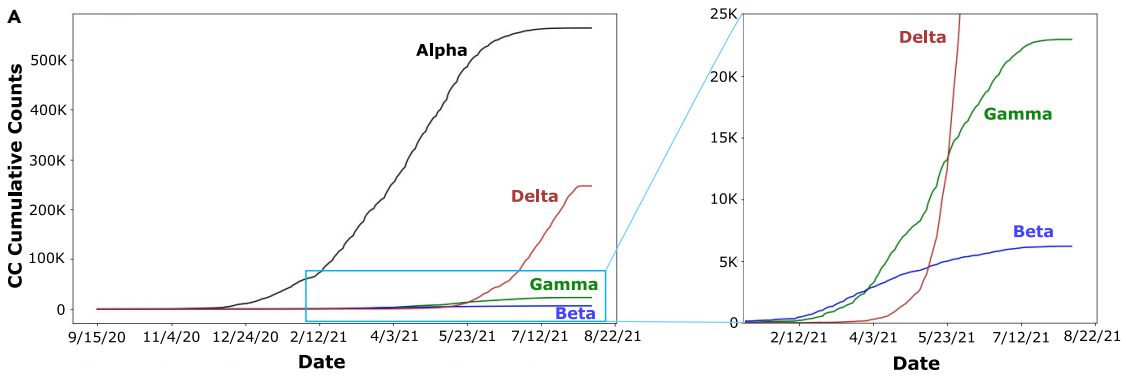
EWAD residuals reveal striking real-time signals of pathology, codified in pathology alert levels

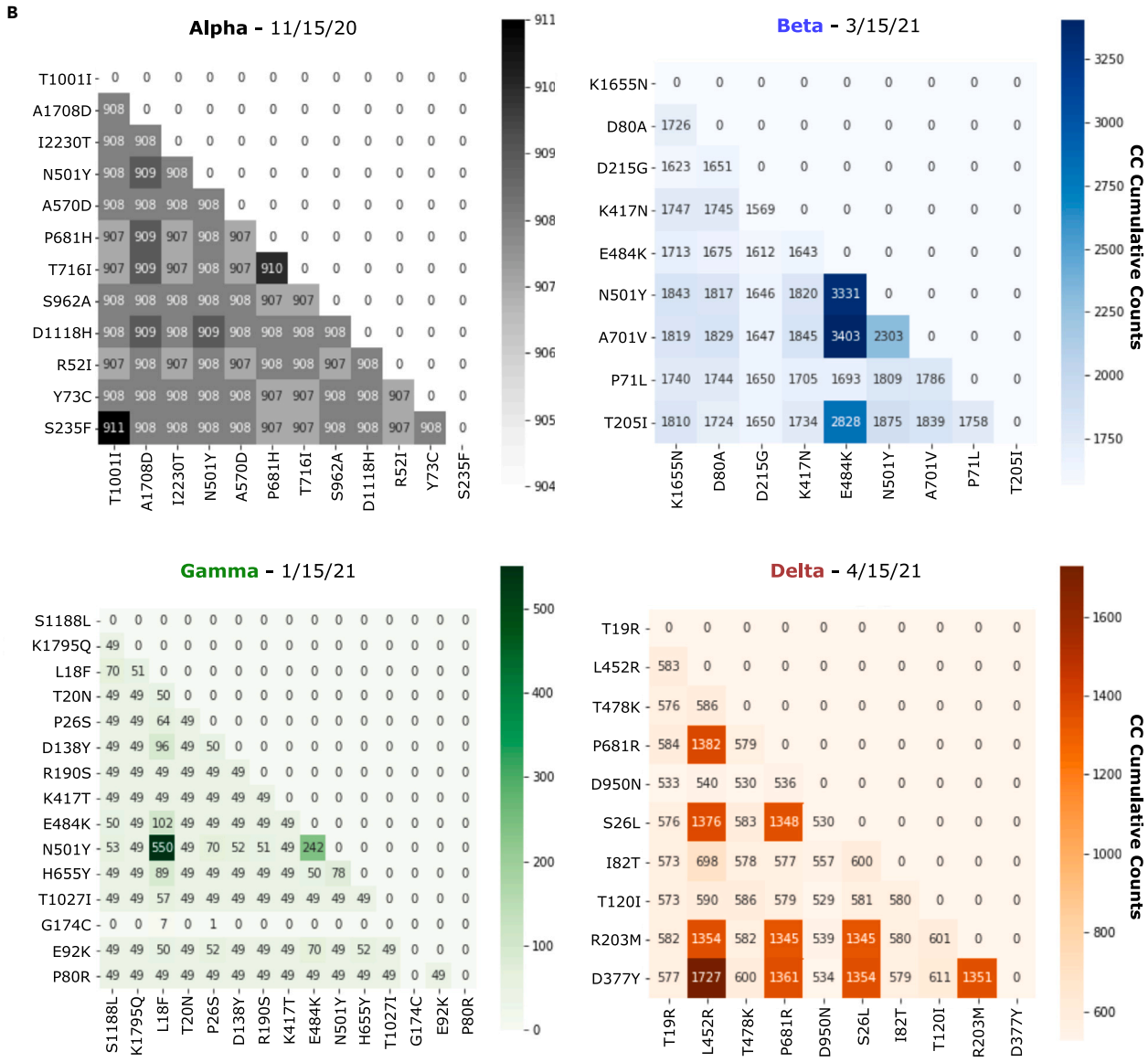
By using as *input* the defining mutations of the Alpha, Beta, Gamma, and Delta VOCs as our sparse collection of measured

Figure 3. Time lapse of viral genome allele phenotype landscapes

(A) Delta VOC showing six tri-monthly time points between May 2020 and August 2021. $\overline{\boldsymbol{IR}}$ (y axis) and $\overline{\boldsymbol{FR}}$ (z axis) are log-transformed, genomic position is scaled to a (0–1) scaled from 5′ to 3′ of the RNA sequence encompassing ~30,000 bp with 5′-end located at the origin of the x axis. Vertical dotted blue lines are boundaries between SARS-CoV-2 proteins, annotated on the top of each figure. Input variants are in shaded color, with dot sizes proportional to allele frequency of the mutations. Contour lines are drawn at 10% and 25% percentile of global variance estimated for model predictions (Figure 1C).

(B) (Left) Average distance between Delta VOC signature mutations defined by x axis (genomic position) and y axis coordinates ($\overline{\boldsymbol{IR}}$) as described in Figure 1B for each of the six time points shown in (A). The gray ribbon marks the 95% confidence interval. (Right) Average spatial variance of $\overline{\boldsymbol{FR}}$ (z axis) between Delta signature mutations defined by x axis 'Genome position' and y axis coordinates ($\overline{\boldsymbol{IR}}$) as described in Figure 1B for each of the six time points shown in (A). The gray ribbon marks the 95% confidence interval.





(legend on next page)

values for GP-based analyses, we reasoned that the calculated *mean* of $\overline{FR}$ residuals, where the mean is the averaged value of the $\overline{FR}$ residuals for a given time frame (see methods), could be used as *output* across the entire pandemic time line to detect potential features in advance reflecting the emergence of VOC from the variant dark matter. Because mean $\overline{FR}$ residuals highlight coordinated changes in real time in the observed $\overline{FR}$ versus the predicted $\overline{FR}$ based on the SCV relationships within the evolving variant dark matter, they provide an EWAD for emerging VOCs.

We first examined the mean $\overline{FR}$ residuals for all the mutations defining Alpha, Beta, Gamma, and Delta VOCs at weekly time points, and plotted the mean $\overline{FR}$ residual value for each VOC as a measure of their changing GP-based relationships to define globally the emergent evolution of SARS-CoV-2 lineages. Initial efforts focused on the average co-occurrence cumulative counts over the time interval between September 2020 and May 2021. We selected six representative time points covering the flat, early, and sustained growth phases for each of the VOCs Alpha, Beta, Gamma, and Delta VOCs for co-occurrence (Figures 6B, 6D, 6F, and 6H) and their mean $\overline{FR}$ residuals (Figures 6C, 6E, 6G, and 6I, blue line) for analyses. We first assign a zero baseline that is set to 0 ± 0.05 by empirical randomization where any overlap of the mean $\overline{FR}$ residuals with the baseline reflects the high probability that the calculated mean relationships are unrelated to the emergent VOC under consideration (Figures 6C, 6E, 6G, and 6I, red dashed line).

To assess potential EWAD signals, we defined two pathology alert levels (PAL): PAL1 (Figures 6C, 6D, 6G, and 6I; light red shade) and PAL2 (Figures 6C, 6D, 6G, and 6I; dark red shade), which considers the degree of change over time, the magnitude of change, and the persistence over time—detailed as follows. PAL1 is defined either as (1) two consecutive points (Figure 6B, x axis) whose combined *change in* mean $\overline{FR}$ residuals (Figure 6B, y axis) is more than 0.05, and/or (2) both mean residual and its 95% confidence interval above or below the zero baseline. PAL2 is defined as three consecutive points whose combined change in mean $\overline{FR}$ residuals is above 0.1. Two empirical alert levels were chosen as they seemed a reasonable trade-off between an overly simplistic single-alert model, and the additional complexity of multiple alert levels (where PAL1 is an invite to watch closely, whereas PAL2 incites to possible action).[51] These alert levels were chosen to show significant deviations from the basal state that could have more rigorous definitions, as our understanding of their root cause evolves in future work. We next examined EWAD development over time for Alpha and Omicron as examples at the extreme ends of the pandemic, with Beta, Gamma, and Delta explained in detail in the supplementary results.

To test the statistical significance of the EWAD PAL system, 15 mutations similar to the size of signature mutations for Alpha and Gamma VOC were randomly chosen and monitored over seven tri-weekly time points between 11/17/20 and 5/18/21. Whereas the mean $\overline{FR}$ residuals corresponding to VOC signature mutations show a coordinated, EWAD signal of potential use for variant surveillance (see the bar plots in Figures S6 and S7), residuals for these randomly chosen mutations show a rather random behavior over time, not correlated with specific temporal events. The full analysis was repeated with thousands of random sets of mutations whose set size is the same as the VOC signature mutations considered here (10–15 mutations) to determine if the EWAD results were statistically significant and unique to VOC, or if they were just a map-wide property that could be observed for any random set of mutations. Remarkably, a great majority of randomly selected mutations did not present any coordinated behavior at the level of $\overline{FR}$ residuals across the selected time window (empirical p values $<10^{-3}$ for all VOCs), providing evidence that the observed EWAD patterns are specifically associated with distinctive VOC mutations. The mutation sets for Alpha, Beta, Gamma, and Delta VOCs were compared with five randomly selected sets of mutations from the data in order to validate the significance of the differences between the VOC mutations and randomly chosen mutation sets. These results showed conclusive differences, as described in full in Figure S8. These results have important implications for the future prediction of unknown VOCs prior to their emergence by focusing on emerging variants in the variant dark matter comprising high spread with either low or high pathology.

An EWAD example: Predicting the Alpha VOC

Alpha VOC was first detected in November 2020 from a sample collected in the United Kingdom in September 2020[53] and declared a VOC on 12/15/20 (Figures 6B and 6C; star). Although our co-occurrence data track down precisely this first Alpha isolate as seen in the co-occurrence density plot for Alpha (Figure 5, transition to full co-occurrence density on 9/16/20 due to a single isolate), co-occurrence data alone at that time point does not reveal that this event would turn into a widespread VOC, as we have few detections and almost no growth indicated by the flat average co-occurrence (Figure 6B; time point #1, 10/15/20). Strikingly, we note on the EWAD plot (Figure 6C) that this point is already marked as PAL1, being significantly below the baseline but with a broad confidence interval, which was activated just 2 weeks after the discovery of the first isolate. Being below the baseline suggests that the mean $\overline{FR}$ residual for the Alpha VOC is below what we might otherwise expect, consistent with an EWAD signal of PAL1. By underperforming relative to surrounding variant dark matter, this suggests the emergent cluster is hidden from view, having a lower fatality rate than expected, but by triggering PAL1 it reveals itself as a potential VOC (Figure 6C).

On 11/14/20 (Figure 6B, time point #2), we see a change in the growth regime, with a rapidly increasing slope transitioning from flat to consistent growth while the number of worldwide detections remains low (Figure 6B, ~200). The EWAD plot (Figure 6C) reveals a remarkable change with a steep decrease in

Figure 4. Co-occurrence over time for VOCs

(A) Timeline plots showing average cumulative co-occurrence (co-occurrence) over time for the four VOCs on the same scale (left), and zoom view on the later VOCs (Beta, Gamma, Delta).

(B) Representative co-occurrence matrices showing co-occurrence counts between the signature mutations of each VOC. For both (A) and (B): Alpha VOC, black; Beta VOC, blue; Gamma VOC, green; Delta VOC, brown.

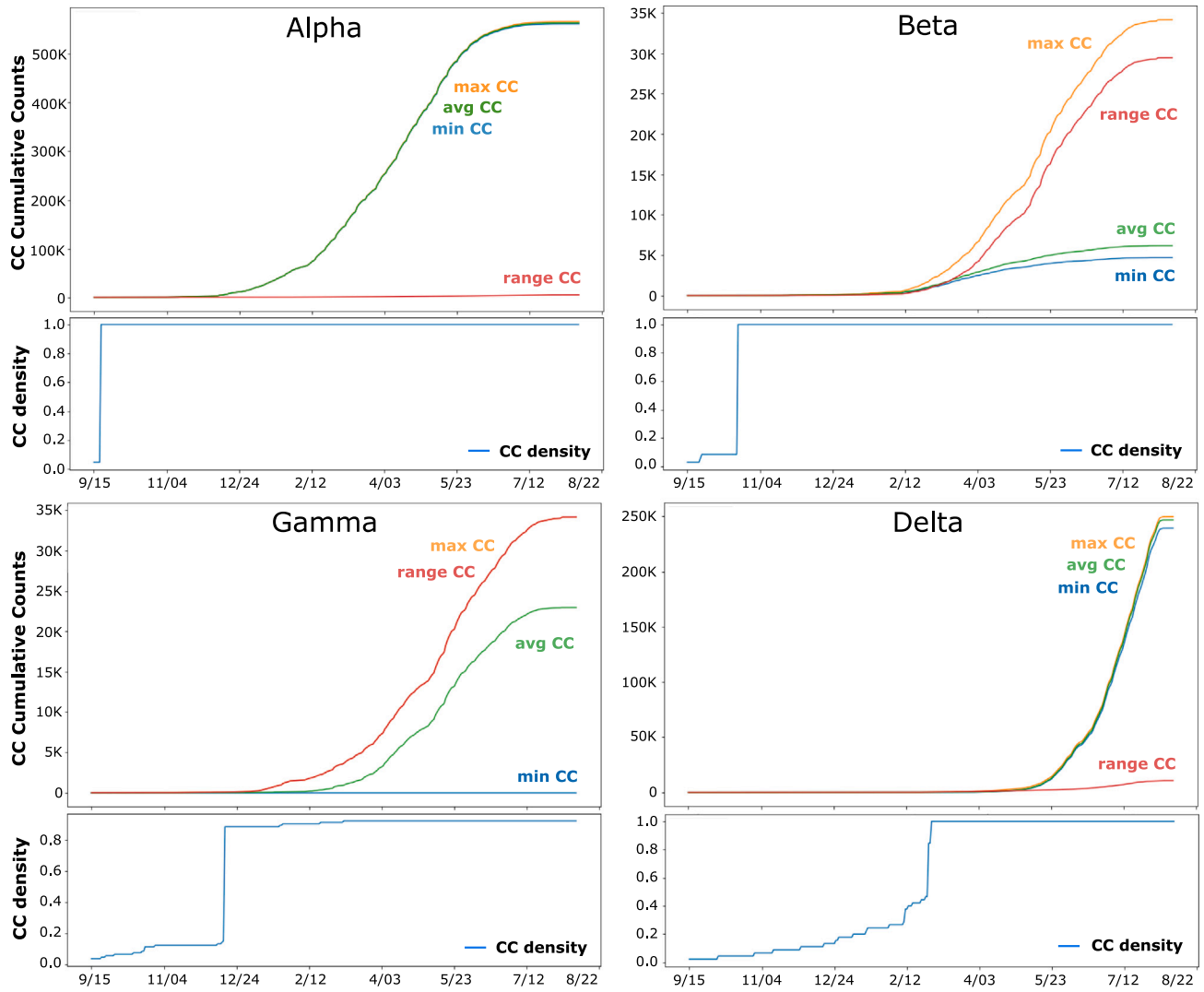
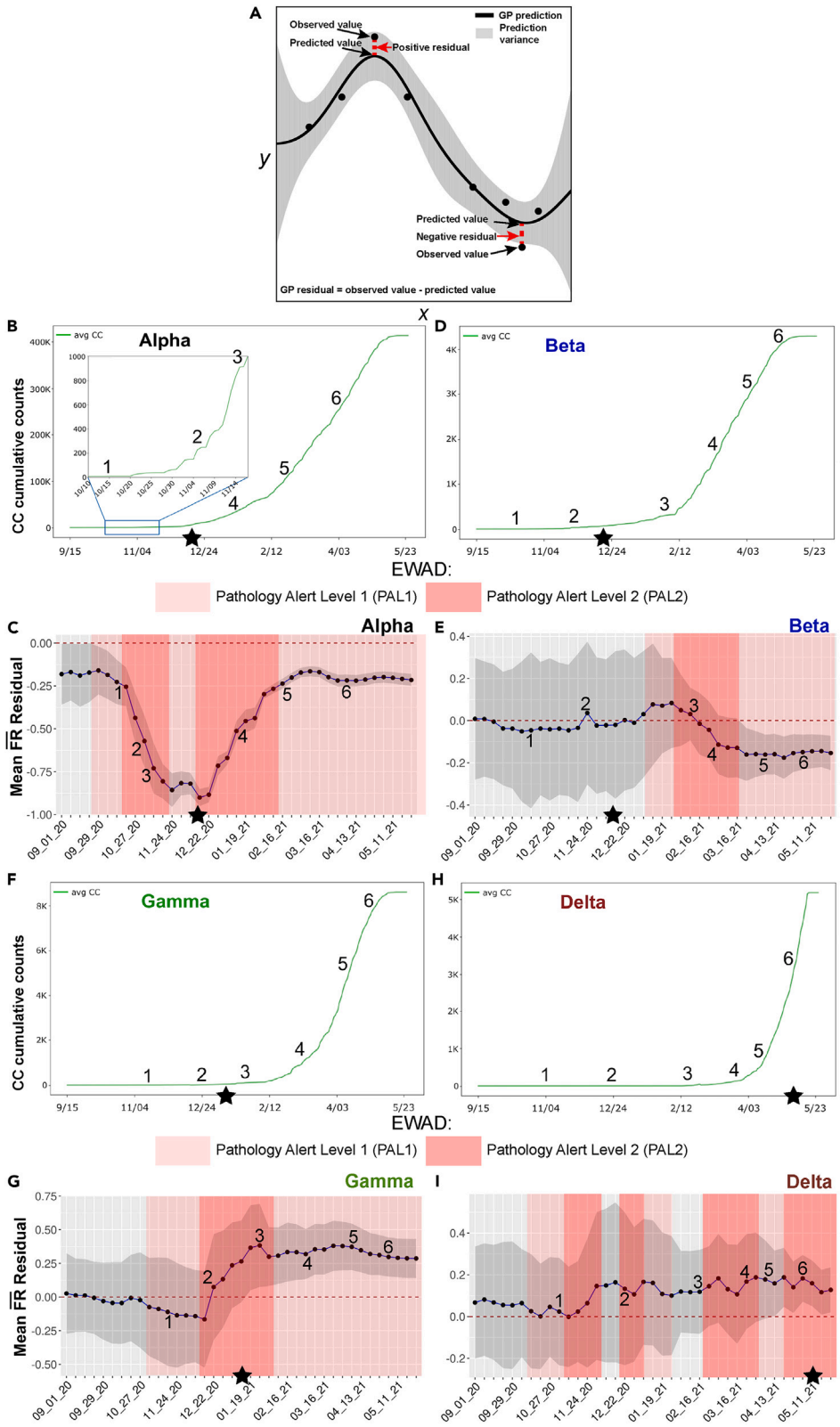


Figure 5. For each VOC (Alpha, Beta, Gamma, Delta), the upper panel reports min co-occurrence, max co-occurrence, average co-occurrence, and range (max co-occurrence minus min co-occurrence) between 9/15 and 8/22

The lower panel shows co-occurrence density, which is the number of non-zero co-occurrences over all possible co-occurrences, standardized for the range of 0–1, for the same time interval. To characterize more precisely co-occurrence patterns emerging for the four VOCs, we tracked max co-occurrence, min co-occurrence, and co-occurrence range over time instead of solely the average co-occurrence described above (upper panel of each). For Beta and Gamma, we observe a "high range" pattern where the difference between max and min co-occurrence (range co-occurrence: red line) increases over time and is above the curve for average co-occurrence (green line). Conversely, Alpha and Delta VOCs show a low range (red line) that is consistently below the average co-occurrence curve (green line) after co-occurrences begin accumulating at a steady pace. Beta and Gamma lineages are thought to boost the immune escape capabilities of the virus, while Alpha and Delta variants are more efficient in enhancing infectivity and spread. Thus, based on the detailed co-occurrence profiles, we can discriminate between different functional classes of VOCs.

mean $\overline{FR}$ residuals exceeding 0.1 and a more compact confidence interval resulting in PAL2 activation beginning on 10/16/20. Looking at the individual $\overline{FR}$ residuals in the allele phenotype landscape (Figure S6), we see that the change is driven by mutations on the map that are compacting in both the 5′ and 3′ regions where the average distances of Alpha VOC mutations decrease between 8/15/20 and 2/15/21 (Figure 3A). The mean $\overline{FR}$ residuals consistently fall to more negative values, suggesting that the Alpha VOC is seeking to optimize covariance relationships between these underperforming driver mutations relative to the supporting passenger variant dark matter. Hence, even with a very low number of detections, the mean $\overline{FR}$ residuals appear to respond quite sensitively to changes in the entire growth regime in a coordinated fashion across the worldwide population. The change would appear negligible and overlooked if examined at the level of hierarchical clustering (mutation counts) or co-occurrences alone, whereas the stark, coordinated change seen within mean $\overline{FR}$ residuals based on GP modeling induces a solid early warning for this set of variants.

The 11/20/21 time point (Figure 6B, time point #3) captures the increasing co-occurrence growth rate with ~800 more detections within 8 days, further showing a coordinated decrease in



(legend on next page)

the mean $\overline{FR}$ residuals (Figure 6C) and accompanied by compaction of mutations on the allele phenotype landscape in the 5′ cluster (Figure 3A). The EWAD signal (Figure 6C) keeps increasing steadily and stays in PAL2, confirming the trend of a strong early warning signal revealed at the previous time point. Again, at this early stage of spread for Alpha VOC, counts alone could be easily overlooked; in contrast, monitoring of mean $\overline{FR}$ residuals provides a clear EWAD signal many weeks ahead of the official (and after-the-fact) designation of these mutations as a VOC.

After an additional 1.5 months (Figure 6B, time point #4, 1/12/21), when the co-occurrence is entering a steady, quasi-exponential growth phase (with over 50,000 viral genomes sequenced bearing the Alpha co-occurrence signature), the average distance between mutations in the 5′ cluster is smaller and at higher values of $\overline{IR}$ in the map corresponding to lower $\overline{FR}$ (Figure S6; panel 4). The mean $\overline{FR}$ residuals exhibit the same consistent pattern of negative values as the previous time point but are more uniform and at a lower magnitude (Figures 7C and S6, panel 4, bar plot). On the EWAD plot (Figure 7C, blue line), mean $\overline{FR}$ residuals are steadily increasing with confidence interval compaction, reflecting improving accuracy of the prediction by GP, still in PAL2. Hence, the collective signal defined by the mean $\overline{FR}$ residuals begin attenuating after reaching the maxima observed in the EWAD phase given its sensitivity to (1) growth rate change and (2) the fact that detections are now in a steady growth regime reflecting balance with supporting variant dark matter in the absence of new competition.

Finally, for the time points #5 and #6 (Figure 6B, sampled on 2/16/21 and 4/5/21, respectively) during the steady growth phase, mean $\overline{FR}$ residuals show no further compaction of the defining Alpha VOC (Figure S6; panels 5 and 6, allele phenotype landscape and bar plots), but rather a collective migration to a higher $\overline{IR}$ and lower $\overline{FR}$ region in both 3′ and 5′ clusters (Figure 6C, blue line; Figure S6, panels 5 and 6, allele phenotype landscape). The mean $\overline{FR}$ residuals modestly increase their magnitude while keeping their uniform, sign pattern (Figure S6; panels 5 and 6, bar plots), thus confirming the attenuated signal configuration initially observed at the previous time point. On the EWAD plot (Figure 6C, blue line), this translates to an approximately flat progression with a narrow 95% confidence interval still in PAL1 since both are well below the baseline. These results suggest a still underperforming variant that could reflect an underestimate of the pathology at this time point, although these results also suggest that Alpha has reached a steady-state equilibrium with the existing supporting variant dark matter and a more effective host environment response to its restricted covariant cluster of mutations.

While we focused on Alpha VOC, EWAD plots can be described that illustrate that different GP-based tactical strategies seen for Beta, Gamma, and Delta VOCs (Figures 6D–6I, see supplemental results). Interestingly, for Beta VOC, the VOC call was made well before PAL1 (Figure 6E) suggesting it was premature, consistent with the fact from a global perspective it remained highly regionalized and fizzled out quickly (outbreak.info/[25]), questioning the utility of the 75% VOC designation as a reliable tool for designating a VOC in the absence of appreciation of the global covariance dictating host-pathogen balance.

Thus, mean $\overline{FR}$ residuals extracted from allele phenotype landscapes have many desirable traits of an EWAD system. The results are statistically significant and unique to each of the VOCs. Importantly, they are not just map-wide properties that could be observed for any random set of mutations, given that if we repeat the analysis with thousands of random sets of mutations whose set size is the same as the Alpha, Beta, Gamma, and Delta VOCs, the majority of the randomly selected mutations will not present any coordinated behavior at the level of mean $\overline{FR}$ residuals across the selected time window (Figures S9 and S10; empirical p value $<10^{-4}$). Consistent with this view, we performed ablation studies for the four VOCs—where the signature mutations of each VOC were removed from the training set prior to GP regression and EWAD analysis. Importantly, for all four ablated sets, we discovered no significant differences between the ablated and the original models in terms of EWAD signal (Figure S11), thus highlighting the robustness of an early warning system with respect to the presence of specific sets of mutations.

These results provide strong evidence that the observed patterns are generated systematically in response to GP-based selection and unknown fitness rules enabling the emergence of a VOC in the context of the large variant dark matter background, providing evidence that the observed EWAD patterns are specifically associated with distinctive VOC mutations. The results

Figure 6. EWAD analysis of Alpha, Beta, Gamma, and Delta VOCs

(A) Graphical explanation of GP regression residuals. GP predictions are covariance-matrix weighted averages of the observed values, so a GP regression prediction is a point comprising the proximity weighted information of its surrounding observed values in the variant dark matter. The GP residual, calculated by using observed value minus the predicted value reports the difference between the mean observed $\overline{FR}$ of that variant and the predicted $\overline{FR}$—the weighted average of its surrounding variants. As illustrated, a positive GP residual indicates that the observed mean $\overline{FR}$ of that variant is higher than the mean weighted averaging of the $\overline{FR}$ for surrounding variants, while a negative GP residual indicates the predicted mean $\overline{FR}$ of that variant is lower than the mean weighted averaging of the $\overline{FR}$ for surrounding variants. GP residual values represent a real-time monitor for the differences of predicted variant $\overline{FR}$ based on SCV analysis.

(B–I) For each VOC (B and C, Alpha; D and E, Beta; F and G, Gamma; H and I, Delta), we report its average co-occurrence (co-occurrence plots: B, D, F, and H) together with the mean $\overline{FR}$ residuals for its signature mutations (mean, blue line; 95% confidence interval, gray shade) computed weekly along the selected time interval (EWAD plots: C, E, G, and I). We examined the time interval between September 20 and May 21, and based on average co-occurrences, we selected six representative time points covering the flat, early, and sustained co-occurrence growth phases for each of the VOCs (see numbered points in graphs B–I). The baseline for EWAD is set to 0 ± 0.05 obtained by empirical randomization (dashed red line at 0 in EWAD plots) (C, E, G, and I) where for a VOC including n signature mutations, we computed the mean $\overline{FR}$ residuals of thousands of random sets of n mutations yielding the interval near zero the random (null) EWAD signal. We then defined two alert levels, pathology alert level 1 (PAL1) (light red shades: C, E, G, and I) and pathology alert level 2 (PAL2) (dark red shades: C, E, G, and I) based on a heuristic that takes into account the degree of change over time, the magnitude of change, and the persistence over time where PAL1 includes two consecutive points whose combined change in mean $\overline{FR}$ residual is above 0.05, and/or where both mean $\overline{FR}$ residual and its 95% confidence interval are above/below zero. PAL2 includes three consecutive points whose combined change in mean $\overline{FR}$ residual is above 0.1. Stars show the date that each variant was designated a VOC by the WHO.

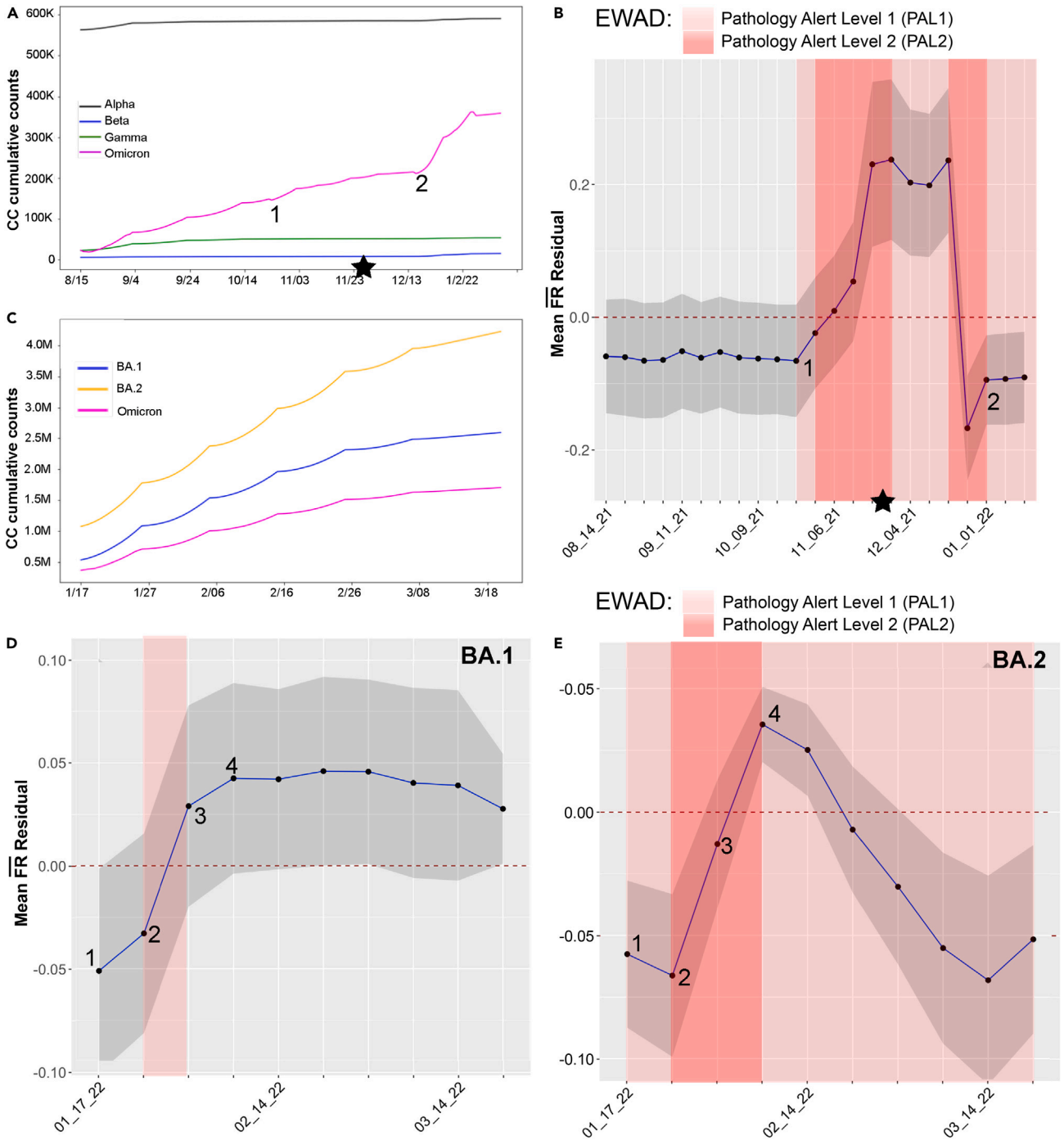


Figure 7. Omicron VOC co-occurrence and EWAD spanning 8/15/21 to 3/20/22

(A) Time line plot showing average cumulative co-occurrence over time for combined Alpha, Beta, Gamma, and Omicron VOC defining mutations (Delta is out of range with values near 1M and is therefore omitted).

(B) EWAD plot for combined Omicron where mean $\overline{FR}$ residuals (blue line with 95% confidence interval, gray shade) for Omicron signature mutations computed weekly along the selected time intervals. The baseline for EWAD is set to 0 ± 0.05 by empirical randomization (dashed red line at 0; details as in Figure 6). Alert levels defined as in Figure 6.

(C) Time line plot showing average cumulative co-occurrence over time for Omicron 1.1.529 and sub-lineages BA.1 and BA.2 for defining mutations.

(D and E) Omicron sub-lineages BA.1 and BA.2 sub-lineages with co-occurrence over time and EWAD analysis between 1/17/22 and 3/20/22. EWAD plots for BA.1 and E. BA.2 where mean of $\overline{FR}$ residuals (blue line; 95% confidence interval, gray shade) signature mutations computed weekly along the selected time interval. The baseline for EWAD is set to 0 ± 0.05 by empirical randomization (dashed red line at 0; details as in Figure 6). PAL defined as in Figure 6.

have important implications for the future prediction of unknown VOCs prior to their emergence by focusing on emerging variants in the variant dark matter comprising high spread with either low or high pathology (Figures S9 and 10). Can this EWAD behavior also be observed with VOCs emergent at the latest stage of the spread and pathology?

A second EWAD example: Predicting the Omicron VOC

To capture the most recent phase of spread and pathology, we updated the GP map from 8/15/20 to 3/20/22 to include emergence of Omicron variants with the peak in cases in January 2021 almost entirely driven by BA.1 and its sub-lineage BA.1.1 where the VOC call was made on 11/26/21 (Figure 7A, star). Here, the Omicron VOC refers to the mutations common to all Omicron lineages found during this time frame[22,24,25] (Figure 7A). When tracking the co-occurrence density for Omicron defining mutations (Figure 7B), we detect a jump in Omicron VOC defining mutations at 10/23/20 (Figure 7B, time point #1) consistent with its robust surge across the worldwide population. An EWAD plot of mean $\overline{FR}$ residuals (Figure 7C, blue line, 95% confidence interval, gray shade) at this time point detects emergence and a rapid transition from PAL1 to PAL2 with the mean $\overline{FR}$ residuals at 10/23/22 to 11/01/22 going above the zero baseline—well before its emergence as a dominant strain in 12/23/22 to 1/20/21 (Figure 7B, time point #2). Not only do we have exceptionally strong EWAD signal, but a rise above the baseline suggests that the observed VOC $\overline{FR}$ data is overperforming the prediction—that is, the observed VOC $\overline{FR}$ data are above what we might otherwise expect in response to the surrounding variant dark matter. These results suggest that $\overline{FR}$ is now being challenged by evolving host responses reflecting the Omicron's "marauder" mode in the face of increasing competition from the host. However, this changes rapidly with the emergence of BA.1 and BA.1.1.1, with a rapid drop in $\overline{FR}$ residuals indicating a rapid evolution to fitness that allows it to dominate the pandemic landscape in the context of the supporting variant dark matter background. A similar sensitivity in early stages of Omicron diffusion is observed on the allele phenotype landscape (Figure S12), full details in the supplementary text.

To gain insight into the evolution of just the emergent Omicron sub-lineages BA.1-BA.2, we added an additional 1.6 M viral sequences between 1/17/22 and 3/20/22 (for a total of over 5.4 M sequences processed) where spread remained a prominent feature in the evolution of SARS-CoV-2 Omicron strains relative to fatality—likely reflecting gains in host immune response and more effective clinical/social management of virus pathology (Figures 7D and 7E).[54–57] The BA.1 sub-lineage, identified on 11/15/21 and becoming dominant worldwide by 1/17/22, was the first VOC suggested to be able to completely escape from neutralizing antibodies induced by vaccination,[57] in essence resetting the host immune response balance. Subsequently, the BA.2 sub-lineage sharply increased heading into February and March of 2022.[55,56] As of March 2022, Omicron BA.2 was the dominant sub-lineage in most countries. BA.1 and BA.2 have many mutations in common (both being sub-lineages of the original Omicron B.1.1.529) but with 21 mutations in the Spike protein, differentiating the two sub-lineages.

We focused on sets of characteristic signature mutations for BA.1 and BA.2 (43 and 49, respectively) that are non-synonymous substitutions or deletions in their encoded viral proteins that occur in >75% of sequences within the overall Omicron lineage[22] (Figure 7C). Once again, co-occurrence counts alone offer limited information regarding the progression and potential risk of the new sub-lineages. We performed separate EWAD analyses for Omicron sub-lineages BA.1 and BA.2 weekly between 1/17/22 and 3/20/22 (Figures 7D and 7E; blue lines, 95% confidence interval [gray shade]) for signature mutations. During this interval, BA.1 and BA.2 sub-lineages already have higher cumulative co-occurrence counts (Figure 7C) compared with the variants common to all Omicron (Figures 7A and 7B), with an accumulation rate increasing with BA.2>BA.1>Omicron. Interestingly, mean $\overline{FR}$ residuals for each of the BA.1 and BA.2 sub-lineages show similar EWAD signals at the early stages of the evolution, but then differentiate suggesting different evolutionary paths are evoked by the change in mutation load in the evolving BA.2. For BA.1, new activity between 1/24/22 and 1/31/22 induces another brief PAL1 (Figure 7D), but then the signal subsides to no-alert with a broad mean $\overline{FR}$ residuals for the following weeks (up to 3/20/22) above the baseline. The mean of the observed values are overperforming our prediction—that is, the predicted VOC fatality rate is above what we might expect, suggesting that the host is gaining the upper hand in mitigating impact, a conjecture supported by the fact that BA.1 counts reach a plateau (Figure 7C, blue line), marking the time point where it begins to lose influence in response to the rising dominance of BA.2 (Figure 7C, orange line). Intriguingly, BA.2 EWAD (Figure 7E) is already in PAL1 during the first week (1/17–1/24) and rapidly escalates to PAL2. These results indicate that the observed data are overperforming our prediction and that the predicted VOC fatality rate is above what we might otherwise expect based on the surrounding variant dark matter, reflecting a stealth/marauder search mode.[58] This rise above baseline quickly drops back to below the baseline, indicating that the observed data are now underperforming our prediction and that the predicted VOC fatality rate is below what we would expect based on surrounding variant dark matter, suggesting a change in response by the host population that challenges fatality. Remarkably, the alerts for BA.2 are triggered in mid-January, again weeks before more official warnings such as the official World Health Organization (WHO) VOC designation in late February.

The overall EWAD pattern for BA.2 (Figure 7D) strikingly resembles the one seen for the originating collection of Omicron variants (Figure 7B). This indicates that EWAD has value not only for newly emerging lineages represented by the strikingly different VOCs, but also for tracking the subsequent evolution of VOC sub-lineages, a current concern of health agencies.[59,60] These results suggest that the integration of genotypic and phenotypic data through GP-based residual analysis provides a temporal and sensitive EWAD output for pandemic performance to anticipate its trajectory across the worldwide population.

DISCUSSION

We have introduced a system-based GP spatial covariance platform[32–34,38] (Figure 8) to plot the role of natural selection and global population fitness in driving host-pathogen relationships.

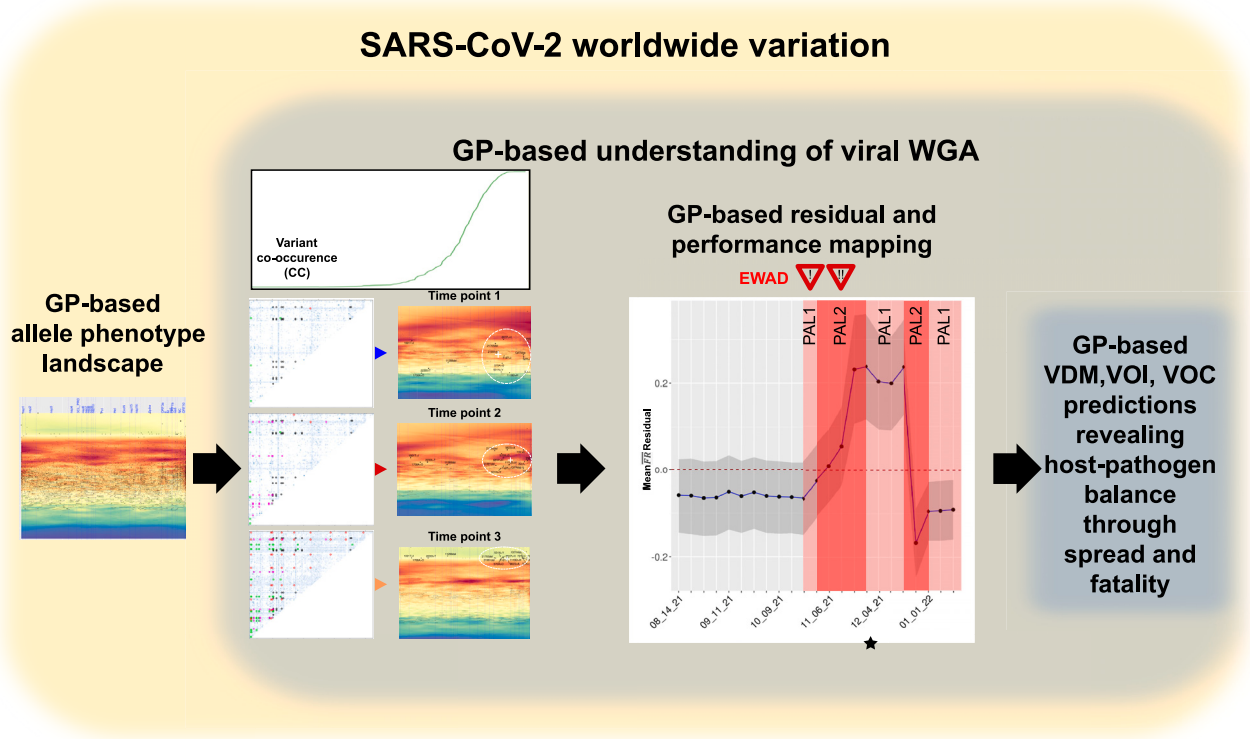


Figure 8. Flow diagram for modeling EWAD and performance using GP

Starting from the GP-based predicted allele phenotype landscape for each VOC (leftmost panel) in combination with the co-occurrences at different time points (second panel from left), GP residuals can be calculated to assign PAL1 and PAL2 danger alerts (third panel from left) predicting the host-pathogen responses weeks to months ahead of the official WHO VOC assignation. The performance characteristics in terms of impact on VOC pandemic features can be estimated by the position of the mean $\overline{FR}$ residuals below or above baseline (dotted line). Shown as an example is the result for Omicron emergence (Figure 7B).

Not only do we capture the distinctive patterns of evolution of each of the VOCs through $\overline{IR}$ and $\overline{FR}$ coordinates in allele phenotype landscapes, but we have shown that tracking these changes through GP reveals a hidden and largely uncharacterized agenda involving the supportive variant dark matter in pandemic progression. GP-based analyses suggest that successful allele changes found in emergent VOCs are using rules defined by covariant weighted proximity that evolves with time and that are predictable based on SCV relationships encompassing the entire collective. This is evidence of the potential for GP-based SCV to provide insights from a high-confidence natural selection view that are rooted in global biological variation.[32–34,38]

It is now apparent that the "collective" strategy defined by our GP-based covariant predictions reveals the ability of the VOCs to differentiate themselves from one another and from the whole in response to the more slowly evolving host genome. These range from the "opportunistic" Alpha VOC trajectory to the "predatory" Delta VOC to the "stealth/marauder" modes of Omicron trajectories. For example, from Alpha VOC onward, the time spent searching for optimal covariant combinations to enhance viral spread increases because of emergent immunological host responses to pathogen aggression. Delta stands out as the first predator VOC, performing a successful extensive GP principled grid search by using a dominant 3′ SCV cluster—one that includes mutations in the Spike protein affecting resistance to immune surveillance, binding, and uptake—as well as the nucleocapsid (NC) protein involved in packaging the virus, contributing to viral load.[50,61] This compact SCV cluster is then extended by the stealth/marauder state of Omicron VOC involving specific mutations not only in Spike and NC, but to additional mutations in viral proteases and envelope proteins, indicating that successive surges may involve increasingly complex covariant search strategies using different viral components to increase spread and thus contribute to the changes in host pathology that may become better targets for therapeutic management. For example, newly emergent strains of Omicron, such as BQ.1, BQ.1.1, and BQ1.5, have gained significant ground with the Omicron subvariant Arcturus (XBB.16) now taking the lead. It is interesting to note that by tracking the differing search strategies, it is the Omicron's "marauder" mode, in the face of increasing competition from the host, that has dominated and become the baseline from which subsequent sub-lineages are emerging, particularly in China as of June 2023. The spread of these sub-lineages, it has been argued, is largely driven by immune response evasion in the Spike protein.[62–64] However, this alone may not be the answer. Covariance across the entire genome is something that will need to be considered in future work by incorporating what is hidden from view (i.e., the emergent features of the variant dark matter such as Q556K in the

Table 1. Table of VOC assignation showing EWAD PAL raised well in advance

VOC	Date of VOC assignation	Date of first PAL1	Date of first PAL2
α	18 Dec 2020	23 Sep 2020	16 Oct 2020
β	18 Dec 2020	5 Jan 2021	27 Jan 2021
γ	11 Jan 2021	26 Oct 2020	12 Dec 2020
δ	11 May 2021	10 Sep 2020	7 Nov 2020
o	26 Nov 2021	23 Oct 2021	30 Oct 2021

ORF1a protein and Y264H in the ORF1b protein that are present in BQ.1 and BQ.1.1 but not BA.1 or BA.5). This can be done through the prism of GP-based SCV and EWAD to advance understanding of viral pathology and/or advancing new therapeutics,[63,65,66] as our GP analysis treats the virus as whole (the collective sum of its covariate parts) in driving the rapid evolution of the pandemic.

The EWAD system showcased here builds upon the SCV method we developed for inherited genetic disease.[32] Instead of focusing on polypeptide sequence changes, it focuses GP-based modeling of allele changes for all ~30,000 bp comprising the SARS-CoV-2 genome, which raises the possibility that in the future could be applied to both coding and non-coding sequences. In addition, it adds the use of co-occurrences and residuals to the analysis allowing for not only the identification of possible VOCs well in advance of their WHO assignation (as can be seen in Table 1), but also giving a covariant-based description of the qualitative/quantitative behavior of those emerging variants. For example, representative heatmaps showing the mutation co-occurrence matrices during the early, mid, and late stages of the pandemic can be seen in Figures S13–S21, and for Delta specifically in Figures S22 and S23. Thus, computational analysis, merging the insights from GP-based mean $\overline{FR}$ residuals with co-occurrence maps, can provide an EWAD framework, where the averaged residuals are the difference between predicted and observed values, to account for host fitness relative to the emergent pathogen aggression. EWAD provides the ability to forecast the emergence of a VOC through both the slope dynamics of the mean $\overline{FR}$ residuals and the spatial-temporal compaction of the cluster reflected in co-occurrence densities. As a novel application of the GP-based SCV approach, the mean $\overline{FR}$ residuals also provide a performance index relative to the EWAD baseline—indicating the status of the VOC collective in response to the evolving (and largely hidden from view) variant dark matter—the evolution of which, for example, can be seen for nsp12 (Figure S24).

While other models have been proposed that attempt to predict the spread of specific SARS-CoV-2 mutations,[67,68] to the best of our knowledge they do not link spread to fatality, or attempt to give qualitative/quantitative information about the potential features of a particular variant. Thus, these features allow us to see in advance the emergence of mutational clusters that could contribute to both spread and/or pathology before the virus advances to clinically stamped VOC status (Figure 8)—a result that does not appear to be tied to the presence of specific sets of signature VOC mutations as the ablation studies confirmed. This illustrates the usefulness of framing the pandemic evolution in terms of SCV relationships defined by the evolving worldwide viral genome as a collective—a view that expands on more traditional approaches such as hierarchical clustering. As such, GP-based SCV analyses of global allele distributions may provide a new way to address the dynamics of spread and fatality in pandemics. For example, Omicron had a strong EWAD signal harboring VOCs a full month before the lineage officially designated a VOC. Here, the lack of compaction of residuals suggests that this lineage is well-tuned to continue to evolve within its strain to succeed in successive waves, as has been captured by GP-based SCV for its sub-lineages highlighted above.[59,60]

Besides the small sets of mostly non-overlapping signature mutations for each VOC lineage, the larger sets of all mutations that comprise the variant dark matter required for the full GP analysis are not generally considered as part of a VOC assignation when defining impact. In contrast, by tracing EWAD potential in terms of performance in the GP residual plots, we learn about the dynamics of viral evolution impacting their trajectory in the worldwide population. Here, the under- or over-powered feature of the prediction can give us a sense of how a given VOC achieved its current position in the context of downstream global spread and pathology. These results suggest that tracking via GP designated "variants being monitored," variants of interest (VOIs), or a larger group of cluster subsets we refer to as covariant clusters with co-evolving high y axis $\overline{IR}$ values linked with $\overline{FR}$, could provide a more quantitative tool to assess risk management of disease from both virus and host perspectives. In general, our GP analysis attests to the need to expand our understanding of the variant dark matter in viral disease to fully appreciate the impact of variation in achieving fitness in host-pathogen race for dominance,[69] particularly the countermoves of the host adaptive and innate immune responses and/or social/clinical/political practices contributing to spread and fatality, particularly of the virus-sensitive aged population (Figure 8). The recently discovered ability to detect emerging variants in sewage may provide a broad and more consistent covariant collective of population behavior in each locale that is amenable to GP-based EWAD analysis.[70]

The method introduced here for VOC early warning and variant surveillance has several features that are worth noting. It is a purely computational method for variant surveillance, using data from publicly available repositories (viral sequences, infectivity, and fatality data updated daily). In its early stage, it applies a novel multimodal data fusion approach across time-resolved genotypic and phenotypic data to obtain the input composite variables for GP modeling. The ML workflow implemented in the method differs from standard supervised ML methods (i.e., classification/regression) since it uses supervised ML (in the form of GP regression) as an intermediate step to generate data instead of an endpoint for the prediction, sharing similarities with generative modeling. Importantly, GP regression acts as an amplifier for small but robust differences through weighted proximity occurring in the phenotypic data over time, that are then exploited for the purpose of VOC anomaly detection.

The limitations in our method lie currently in its reliance on already identified variants in the population to pioneer the SCV framework. It is currently computationally prohibitive to examine all combinations of mutations seen in the SARS-CoV-2 data for

their potential to become a VOC, although emergent variants being monitored and VOIs, as indicated above, provide a focused starting point in evaluating variant dark matter. Given the universal applicability of SCV,[32–38,48] capturing emergence from the total viral variant load may be possible in other settings, including for example, influenza and HIV, where current collections of variant genotypes and associated phenotypes could serve as a collective for GP-based landscape descriptions.[32–38,48,71]

In terms of computational resources, pattern generation by the GP method is highly efficient. Only 10 compute nodes and a few hours were used to image SCV maps for 700+ days of the pandemic. On the other hand, computation of co-occurrences can, in principle, be challenging both in terms of time and memory as the number of known viral mutations grew near-exponentially over time given the massive tracking efforts (for example, see Omicron co-occurrences in Figures S25 and S26). However, adopting a few strategies (detailed in methods) such as sparse data structures, and a cutoff on very low allele frequencies, computation of co-occurrences can be completed in a reasonable time frame on a moderately sized compute resource (e.g., a few days on 20 compute nodes). Code and examples for both GP and co-occurrence calculations is available at the GitHub public repository: https://github.com/balchlab/VSPsnap or Zenodo archival https://doi.org/10.5281/zenodo.8000486. Over recent months, the data coverage and quality of SARS-CoV-2 (such as the mutation frequencies tracked by outbreak.info [Figure S27]) has waned for many reasons—technical, social, and political. As such, our approach will be more challenging to apply. However, the strength of the GP model lies in its versatility and broader applicability given its use of only a sparse collection of variants such has already been applied to multiple human rare diseases.[21,34–37,48] These efforts provide a new paradigm for development of therapeutics impacting genome-encoded sequence-function-structure relationships[71] that can now be applied to viral genomes.

There is a fundamental dichotomy between the pandemic efforts focused on the "microscopic" (genes, mutations, proteins, biochemical and biophysics of virus life cycle, host cell biology) and "macroscopic" (cases, masks, vaccination, hospitalizations, deaths, social behavior, healthcare policies, politics, etc.) issues. Unfortunately, the necessarily integrated rules dictating these diverse covariant relationships remain largely unknown and hidden from view in the simple hierarchical clustering maps used to currently describe spread and pathology progression and guide health policy.[22] In contrast, our GP-based SCV method can integrate both the micro- and macro- at atomic resolution.[32–34,38] For example, the migration to low $\overline{FR}$ regions in Alpha, Delta, and Omicron VOC lineages found at the top of allele phenotype landscapes is strongly correlated with the beginning of vaccination policies across many countries (not shown), indicating improvement in host fitness responses. Moreover, because even a simple estimator extracted from the modeling of GP-based $\overline{FR}$ residuals displays an unanticipated coordinated pattern of future relationships, there might be a number of additional properties that can be extracted that are more actionable than those revealed by simple hierarchical clustering.

In more general terms, we posit that GP-based features could help to elucidate the role of WGA in the context of micro- and macro-complexity observed at the host-pathogen interface *à la* the Red Queen challenge. These relationships are hidden within the variant dark matter and ignored as a collective except as a means of tracking evolutionary trajectories.[69] In contrast, GP-based WGA offers a starting point to begin to understand the complexity of coupled genomic and proteomic architectures, and how evolution uses this coupling through covariance to shape biological function.[32–38,48] As host-pathogen biology moves into a new, rapid phase of management where molecular analysis, designer vaccines, and novel therapeutics can address the immediate need to lower human fatality, particularly in the aging population,[2–14] an understanding of GP-based SCV relationships could allow for a more rapid and expansive exploration of disease at the level of the individual.[32–34,38] GP-based SCV principled ML insights could provide a more generalized approach for understanding pathogen fitness relative to host response (or vice versa) for management of risk in the clinic, given the versatility of phenotype landscapes to quantitatively frame the emergent Red Queen challenge.[72]

EXPERIMENTAL PROCEDURES

Resource availability

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Additional Supplemental Items are available from Mendeley Data	https://doi.org/10.17632/69zm32zvmn.1	
Software and algorithms		
Original code	https://doi.org/10.5281/zenodo.8000486	
Other		
CNCB publicly available individual sequence data	ftp://download.big.ac.cn/GVM/Coronavirus/gff3	
CNCB publicly available individual meta data	https://bigd.big.ac.cn/ncov/release_genome	
Johns Hopkins publicly available data	https://github.com/CSSEGISandData/COVID-19	

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, William E. Balch (webalch@scripps.edu).

Materials availability

All data are available in the main text or the supplementary materials.

Data and code availability

This paper analyzes existing, publicly available data. These accession numbers for the datasets are listed in the Resource availability table.

All original code has been deposited at Zenodo and is publicly available as of the date of publication. DOIs are listed in the Resource availability table.

Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

Method details

Data generation and description

SARS-CoV-2 genome mutations are collected from the Chinese National Center for Bioinformation 2019 Novel Coronavirus Resource (CNCB) website. The

CNCB resource, at the time of the last update (3/16/22), utilizes 3,864,334 genomes from the GISAID database.[24] The CNCB team aligns the sequences to the reference NC_045512 also known as Wuhan-Hu-1 with Muscle (3.8.31) to identify and extract variants. These variants are provided to the public in gff3 file format for each genome and available for download through an ftp server. In total, after parsing all 3,864,334 gff3 files, there are a total of 104,952 unique mutation identifiers. These identifiers include both single nucleotide polymorphisms (SNPs) and insertions/deletions. The SNP mutations follow the IUPAC degenerate base symbol when describing the replacement base. This creates some ambiguity in determining what the mutation is, due to some cases having multiple possible alternatives. However, these degenerate base occurrences are not frequent nor widespread across countries. Mutations are filtered by selecting ones that are found in more than three countries and mask genome positions 1–55 and 29,804–29,903 since these terminal regions are likely to contain sequencing artifacts.

Country-level case/infection rate (IR) and death/fatality rate (FR) counts are collected from the John Hopkins COVID-19 GitHub Data Repository, which provides time series documentation of country-level case and death counts in csv file format.[73] The counts are then used to calculate % cases per 100k people and pathology percentage (deaths/cases) per country.

Generating GP-based SCV landscapes

We began with two host-related features of pathology, IR, and FR, the latter impact largely defined by the aging population, particularly in early stages of the pandemic.[2–14] We then considered two composite variables, allele frequency-weighted infectivity rate ($\overline{IR}\,(V)$) and allele frequency-weighted pathology rate ($\overline{FR}\,(V)$) (abbreviated as $\overline{IR}$ and $\overline{FR}$ in text). $\overline{IR}\,(V)$ reports on spread/cases and ($\overline{FR}$) (V) reports on deaths/fatality in the worldwide population for a variant "(V)." By composite variable we mean a variable made up of two or more variables or measures that are related to one another conceptually or statistically. These allele frequency-weighted

$$\overline{IR}(V) = \frac{\sum_{country} \frac{counts(V, Ctr_i)}{counts(AllSeq, Ctr_i)} \times cases(Ctr_i)}{\sum_{country} population(V, Ctr_i)}$$

$$\overline{FR}(V) = \frac{\sum_{country} \frac{counts(V, Ctr_i)}{counts(AllSeq, Ctr_i)} \times deaths(Ctr_i)}{\sum_{country} cases(V, Ctr_i)}$$

composite variables keep the structure of IR and FR that report on infections over population and deaths over cases, respectively, where the weighting term of allele frequency for a specific variant is summed over countries (as indicated in the numerator). Similarly, denominators are summed over countries where the variant is detected to achieve a balanced worldwide comprehensive view of mutation distribution and density. When analyzed in the context of GP, the $\overline{IR}$ and $\overline{FR}$ allow us to define the relationships between global pathogen and host fitness, capturing the balance defined by the Red Queen effect where the pathogen or host population must continually evolve new adaptations to secure dominance.[69]

The above metrics are essentially weighted sums of IR and FR across countries where the weighting factors are allele frequencies per country. The choice of allele frequency as weighting factors was motivated by the fact that there is a large imbalance among reporting countries, with the top five countries making up for more than 70% of all sequences provided. The imbalance in sequences would result in a skew in the weighted average of cases per 100k people and pathology percentage, heavily favoring countries with the most cases reported.

For most countries, there is a lag of several days between reported deaths and actual deaths, as could be estimated by cross-correlation between daily cases and daily deaths. To correct for the lag factor, first estimate lags for each country are generated by running a cross-correlation function between daily cases and daily deaths—where the optimal lag is the value that maximizes correlation between the two time series (Figure S1). Countries suitable for lag correction were required to have reasonably high cross-correlation and smooth distributions—filtered for cases with cross-correlation at least 0.4 and $|3| \leq \text{lag} < -30$. Sixty-four countries, including the major contributors like the United States and the United Kingdom, passed the criteria. Most common lags found were 10–15 days. For the countries for which computed empirical lags were available, compute the lag-adjusted $\overline{FR}$ according to the following:

$$\overline{FR}(V,t) = \frac{\sum_{country} \frac{counts(V, Ctr_i, t)}{counts(AllSeq, Ctr_i, t)} \times deaths(Ctr_i, t+l)}{\sum_{country} Cases(V, Ctr_i, t)}$$

essentially dividing deaths occurring at time $t+l$ (lag) by cases at time t.

To build allele phenotype landscapes, we regress the *input* genomic position of each allele found in three or more countries (*x axis*) against the corresponding *input* $\overline{IR}$ (*y axis*) reporting on allele frequency with the *input* z axis $\overline{FR}$ values (Figure 1A). Regression prediction of the z axis pathology value across the entire genome in the context of spread here is not the end goal, but rather a tool for unsupervised learning of clustering where phenotype landscapes are used as *output* to mechanistically define patterns inherent to the covariance between spread and pathology. The $\overline{IR}$ (Figure 1B, y axis) and $\overline{FR}$ (Figure 1B, z axis) for each mutation are first positioned in the 2D phenotype landscape defined by their genomic allele positions (Figure 1B, *x axis*). This is followed by a second step where pairwise distances are computed for all mutations (Figure 1B), and the relationship between pairwise distance and variation in $\overline{FR}$ is codified in a variogram[32,34] (Figure 1C). The variogram quantifies the covarying relationships between all pairwise distances incorporating $\overline{IR}$ (Figure 1C, x axis) and their spatial variance with $\overline{FR}$ (Figure 1C, y axis). The data modeling provided by the variogram is used to build the allele phenotype landscape, where $\overline{FR}$ for all points in the landscape is predicted according the the variogram function (Figures 1D and 1E) so that the landscape describes known and unknown $\overline{IR}$ and $\overline{FR}$ SCV relationships for every allele position defined explicitly by the evolving the SARS-CoV-2 genome.[32,34]

Mutation co-occurrence

All Gff3 files available for the time arc studied were downloaded from the CNCB ftp server, resulting in a collection of 3,864,334 unique gff3 files obtained from sequenced viral genomes, covering over 2 years of pandemic in 142 countries. For each day, all binary co-occurrences were recorded in a square matrix whose root is the number of unique mutations—thus obtaining a collection of daily co-occurrence cumulative counts.

The pseudocode used to generate the daily cumulative co-occurrence matrices was the following for each date in the pandemic.

(1) read in all unique mutation descriptors M,
(2) build a squared matrix, M X M for gff3 files in collection,
(3) read all co-occurring mutations,
(4) generate all pairwise combinations $2\binom{c}{2}$,
(5) add each to co-occurrence matrix,
(6) save co-occurrence matrix for that date.

Since pairwise co-occurrence counts are symmetrical (*cc(mut1,mut2)* equal to *cc(mut2,mut1)*), only a half triangular matrix was saved at each iteration. Two key solutions allowed this computation to be carried out within reasonable time and storage requirements. The first was to implement sparse data structures (python/pandas: *pd.SparseDtype*). This step implied a size reduction of up to 1,000x (from ~1 Gb to ~4 Mb) for each matrix, bringing down the time required to save each object significantly. However, sparse data structures do not support cell-wise editing (i.e., adding a co-occurrence value to a specific cell (*df.loc[x,y] + = 1* type of operation)). The second key workaround was to convert only the specific column to be updated to dense, then sparse again at each update.

Variants of concern

Mutation lists for the main VOC (Alpha: B.1.1.7; Beta: B.1.351; Gamma: P.1; Delta: B.1.617.2; Omicron: B.1.1.529, BA.1, BA.2, BA.3) were obtained from PANGO lineages online resource.[74] Alpha VOC (B.1.1.7) was first detected in September 2020 in southeast England and rapidly became the dominant variant in the United Kingdom, possibly owing to its enhanced transmissibility.[53] This strain spread rapidly to more than 50 countries[75] and dominated the early phase of the pandemic with substantial pathology. Subsequently, Beta (B.1.351) and Gamma (P.1) emerged in more localized patterns of dominance.[76] Delta VOC (B.1.617.2) was first identified in India in December 2020.[77] Within a matter of months, this particular variant spread to >200

countries around the world, becoming the dominant variant.[78] Delta was responsible for 99% of COVID-19 cases being reported worldwide by November 2021 (https://outbreak.info/),[25] again with substantial pathology, when it was supplanted by the highly transmissible Omicron VOC BA.1 and its derivative strain BA.2,[78] and more recently by BA.5 strains.[22,24,25,54–57,68,78–80] The Omicron VOC is recognized to be considerably more transmissible than the Alpha and Delta VOCs.[41] The Omicron BA.5 VOC now constitutes a large and growing proportion of cases as of August 2022 with seemingly higher transmissibility, but apparent lower pathology rates requiring hospitalization, likely reflecting changes in host-pathogen balance in immune response and social/clinical management. These results have led to consideration of the pandemic transitioning to an endemic state, albeit one with remaining fatality for the aging population reflecting a lack of immune robustness.[16–21] Understanding what mechanistically drives variant emergence in the context of SARS-CoV-2 WGA and host responses is key to understanding and controlling the trajectory of host-pathogen balance going forward.[68,80]

GP regression

Filtering for mutations observed in three or more countries gave infectivity and pathology rates for 4,663 mutations across the SARS-CoV-2 genome. This set was used to build SCV maps for the full viral genome, and individual proteins. While maps for single proteins had issues related to variable number of mutations and variable (low) accuracy, the whole genome map was more robust in terms of number of *input* mutations and overall accuracy. Consequently, the strategy was to optimize the whole genome map, obtaining an accuracy in the 0.47 range (Pearson *r*, predicted versus observed). The model was expanded on a grid approximately $x \times 1{,}000$, where x equals the number of nucleotides in SARS-CoV-2 genome, thus obtaining single nucleotide resolution on the x axis. GP regression was performed in R 3.2.8 using the *gstat* library for geostatistical computing and rendering of the maps was obtained with *ggplot*. Cross-validation was performed as standard leave-one-out cross-validation. Trained GP models were used to obtain predicted FR values corresponding to minimum uncertainty. For each protein, the value corresponding to each residue was evaluated as the middle nucleotide of each triplet. An RGB color value matching the same value in the allele phenotype landscape was obtained using a color ramp function using the same palette as the map (R library *RColorBrewer*).

Once the whole genome GP model was available, maps for single CoV-2 proteins were obtained as slices of the whole genome map, by subsetting the corresponding kriged output table and replotting the map in R/*ggplot*. Local estimates of accuracy (protein-specific) were obtained by cross-validation limited to sets of protein-specific mutations.

GP residuals

Residuals are defined as the difference between the observed and predicted values of a variable at an *input* sample point. The observed $\overline{FR}$ for a mutation is the explicit assigned $\overline{FR}$ of that mutation used for the *input* data in GP that does not incorporate the impact of other mutations, while the predicted $\overline{FR}$ for the mutation generated by GP is a proximity weighted average of the observed $\overline{FR}$ value of its surrounding mutations in the phenotype landscape as generated through the SCV process. So, if the observed sample value of a variable is 7, for example, and the predicted value is 3, then the residual would be 4.

$$residual = observed - predicted$$

In our context, we are comparing the observed allele frequency-weighted fatality rate $\overline{FR}$, at a given point in the SCV landscape—as defined by genomic location and allele frequency-weighted infectivity rate associated with a particular mutation—with the $\overline{FR}$ predicted for that location by our GP SCV protocol. We calculated the $\overline{FR}$ residuals for all the mutations in each different VOC at different time points. Then for each time point we calculated the mean of all the VOC residues and plotted the mean for each VOC over time (Figures 6 and 7).

Time lapses

Daily resolved datasets for infectivity and pathology rate were built in a cumulative fashion, i.e., for a specific day *d*, values were added from the beginning of the data collection (2/1/20) up to day *d*. Filtering parameters were the same as described before (e.g., three or more countries, etc.). Data were scaled locally, with respect to the current dataset. To automate GP model generation, variogram parameters were automatically fitted through the *autofit* routine available in *gstat*, with the exception of nmin/nmax (set to 5/30) and model type—set to "Exponential" to avoid flat variance across all predicted FR values.

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.patter.2023.100800.

ACKNOWLEDGMENTS

We thank Thomas Stoeger for helpful comments and discussion.

Funding was provided by National Institutes of Health grant HL095524 (W.E.B.); National Institutes of Health grant AG049665 (W.E.B.); and National Institutes of Health grant AG070209 (W.E.B.).

AUTHOR CONTRIBUTIONS

Conceptualization: S.L., W.E.B., C.W., and D.S. Methodology: S.L., B.C.C., C.W., P.Z., S.S., and D.S. Software: S.L., D.S., C.W., and B.C.C. Investigation: S.L., B.C.C., and W.E.B. Supervision: W.E.B. Writing – original draft: S.L., C.W., B.C.C., and W.E.B. Writing – review & editing: B.C.C., S.L., C.W., S.B., and W.E.B. Project administration: W.E.B. Funding acquisition: S.B. and W.E.B.

DECLARATION OF INTERESTS

The authors declare no competing interests. The authors declare no advisory, management or consulting positions. C.W. and W.E.B. have filed a patent application for the SCV methodology (serial no. US2021/0324474). C.W. and W.E.B. have filed a PCT application (serial no. PCT/US2022/039594) for VarC methodology.

INCLUSION AND DIVERSITY

One or more of the authors of this paper self-identifies as an underrepresented ethnic minority in their field of research or within their geographical location. One or more of the authors of this paper self-identifies as a member of the LGBTQIA+ community.

Received: December 7, 2022
Revised: February 22, 2023
Accepted: June 22, 2023
Published: July 21, 2023

REFERENCES

1. WHO Coronavirus (COVID-19) Dashboard. (2022). https://covid19.who.int.
2. Levin, A.T., Hanage, W.P., Owusu-Boaitey, N., Cochran, K.B., Walsh, S.P., and Meyerowitz-Katz, G. (2020). Assessing the age specificity of infection fatality rates for COVID-19: systematic review, meta-analysis, and public policy implications. Eur. J. Epidemiol. *35*, 1123–1138. https://doi.org/10.1007/s10654-020-00698-1.
3. Channappanavar, R., and Perlman, S. (2020). Age-related susceptibility to coronavirus infections: role of impaired and dysregulated host immunity. J. Clin. Invest. *130*, 6204–6213. https://doi.org/10.1172/JCI144115.
4. Niemi, M.E.K., Karjalainen, J., Liao, R.G., Neale, B.M., Daly, M., Ganna, A., Pathak, G.A., Andrews, S.J., Kanai, M., Veerapen, K., et al. (2021). Mapping the human genetic architecture of COVID-19. Nature *600*, 472–477. https://doi.org/10.1038/s41586-021-03767-x.
5. Mendiola-Pastrana, I.R., López-Ortiz, E., Río de la Loza-Zamora, J.G., González, J., Gómez-García, A., and López-Ortiz, G. (2022). SARS-CoV-2 Variants and Clinical Outcomes: A Systematic Review. Life *12*, 170. https://doi.org/10.3390/life12020170.
6. Tazerji, S.S., Shahabinejad, F., Tokasi, M., Rad, M.A., Khan, M.S., Safdar, M., Filipiak, K.J., Szarpak, L., Dzieciatkowski, T., Jurgiel, J., et al. (2022). Global data analysis and risk factors associated with morbidity and

mortality of COVID-19. Gene Rep. *26*, 101505. https://doi.org/10.1016/j.genrep.2022.101505.

7. Rea, I.M., and Alexander, H.D. (2022). Triple jeopardy in ageing: COVID-19, co-morbidities and inflamm-ageing. Ageing Res. Rev. *73*, 101494. https://doi.org/10.1016/j.arr.2021.101494.
8. Singh, J., Alam, A., Samal, J., Maeurer, M., Ehtesham, N.Z., Chakaya, J., Hira, S., and Hasnain, S.E. (2021). Role of multiple factors likely contributing to severity-mortality of COVID-19. Infect. Genet. Evol. *96*, 105101. https://doi.org/10.1016/j.meegid.2021.105101.
9. Alimohamadi, Y., Tola, H.H., Abbasi-Ghahramanloo, A., Janani, M., and Sepandi, M. (2021). Case fatality rate of COVID-19: a systematic review and meta-analysis. J. Prev. Med. Hyg. *62*, E311–E320. https://doi.org/10.15167/2421-4248/jpmh2021.62.2.1627.
10. Dessie, Z.G., and Zewotir, T. (2021). Mortality-related risk factors of COVID-19: a systematic review and meta-analysis of 42 studies and 423,117 patients. BMC Infect. Dis. *21*, 855. https://doi.org/10.1186/s12879-021-06536-3.
11. Boutin, S., Hildebrand, D., Boulant, S., Kreuter, M., Rüter, J., Pallerla, S.R., Velavan, T.P., and Nurjadi, D. (2021). Host factors facilitating SARS-CoV-2 virus infection and replication in the lungs. Cell. Mol. Life Sci. *78*, 5953–5976. https://doi.org/10.1007/s00018-021-03889-5.
12. Cohen, J.F., Korevaar, D.A., Matczak, S., Chalumeau, M., Allali, S., and Toubiana, J. (2020). COVID-19-Related Fatalities and Intensive-Care-Unit Admissions by Age Groups in Europe: A Meta-Analysis. Front. Med. *7*, 560685. https://doi.org/10.3389/fmed.2020.560685.
13. Tisminetzky, M., Delude, C., Hebert, T., Carr, C., Goldberg, R.J., and Gurwitz, J.H. (2022). Age, Multiple Chronic Conditions, and COVID-19: A Literature Review. J. Gerontol. A Biol. Sci. Med. Sci. *77*, 872–878. https://doi.org/10.1093/gerona/glaa320.
14. Meyerowitz-Katz, G., and Merone, L. (2020). A systematic review and meta-analysis of published research data on COVID-19 infection fatality rates. Int. J. Infect. Dis. *101*, 138–148. https://doi.org/10.1016/j.ijid.2020.09.1464.
15. Tracking SARS-CoV-2 Variants. (2022). https://www.who.int/activities/tracking-SARS-CoV-2-variants.
16. Budinger, G.R.S., Kohanski, R.A., Gan, W., Kobor, M.S., Amaral, L.A., Armanios, M., Kelsey, K.T., Pardo, A., Tuder, R., Macian, F., et al. (2017). The Intersection of Aging Biology and the Pathobiology of Lung Diseases: A Joint NHLBI/NIA Workshop. J. Gerontol. A Biol. Sci. Med. Sci. *72*, 1492–1500. https://doi.org/10.1093/gerona/glx090.
17. McQuattie-Pimentel, A.C., Ren, Z., Joshi, N., Watanabe, S., Stoeger, T., Chi, M., Lu, Z., Sichizya, L., Aillon, R.P., Chen, C.I., et al. (2021). The lung microenvironment shapes a dysfunctional response of alveolar macrophages in aging. J. Clin. Invest. *131*, e140299. https://doi.org/10.1172/JCI140299.
18. Misharin, A.V., Morales-Nebreda, L., Reyfman, P.A., Cuda, C.M., Walter, J.M., McQuattie-Pimentel, A.C., Chen, C.I., Anekalla, K.R., Joshi, N., Williams, K.J.N., et al. (2017). Monocyte-derived alveolar macrophages drive lung fibrosis and persist in the lung over the life span. J. Exp. Med. *214*, 2387–2404. https://doi.org/10.1084/jem.20162152.
19. Watanabe, S., Markov, N.S., Lu, Z., Piseaux Aillon, R., Soberanes, S., Runyan, C.E., Ren, Z., Grant, R.A., Maciel, M., Abdala-Valencia, H., et al. (2021). Resetting proteostasis with ISRIB promotes epithelial differentiation to attenuate pulmonary fibrosis. Proc. Natl. Acad. Sci. USA *118*, e2101100118. https://doi.org/10.1073/pnas.2101100118.
20. Hie, B., Zhong, E.D., Berger, B., and Bryson, B. (2021). Learning the language of viral evolution and escape. Science *371*, 284–288. https://doi.org/10.1126/science.abd7331.
21. Loguercio, S., Hutt, D.M., Campos, A.R., Stoeger, T., Grant, R.A., McQuattie-Pimentel, A.C., Abdala-Valencia, H., Lu, Z., Joshi, N., Ridge, K., et al. (2019). Proteostasis and energetics as proteome hallmarks of aging and influenza challenge in pulmonary disease. Preprint at bioRxiv. https://doi.org/10.1101/769737.
22. Hadfield, J., Megill, C., Bell, S.M., Huddleston, J., Potter, B., Callender, C., Sagulenko, P., Bedford, T., and Neher, R.A. (2018). Nextstrain: real-time tracking of pathogen evolution. Bioinformatics *34*, 4121–4123. https://doi.org/10.1093/bioinformatics/bty407.
23. Variants of the Virus. (2022). https://www.cdc.gov/coronavirus/2019-ncov/variants/index.html.
24. Elbe, S., and Buckland-Merrett, G. (2017). Data, disease and diplomacy: GISAID's innovative contribution to global health. Glob. Chall. *1*, 33–46. https://doi.org/10.1002/gch2.1018.
25. Stolberg, S. (2021). The Delta Variant Makes up an Estimated 83 Percent of U.S. Cases, the C.D.C. Director Says. New York Times, 7/20/2021.
26. Hie, B., Bryson, B.D., and Berger, B. (2020). Leveraging Uncertainty in Machine Learning Accelerates Biological Discovery and Design. Cell Syst. *11*, 461–477.e9. https://doi.org/10.1016/j.cels.2020.09.007.
27. Romero, P.A., Krause, A., and Arnold, F.H. (2013). Navigating the protein fitness landscape with Gaussian processes. Proc. Natl. Acad. Sci. USA *110*, E193–E201. https://doi.org/10.1073/pnas.1215251110.
28. Yang, K.K., Wu, Z., and Arnold, F.H. (2019). Machine-learning-guided directed evolution for protein engineering. Nat. Methods *16*, 687–694. https://doi.org/10.1038/s41592-019-0496-6.
29. Chilès, J.-P., and Delfiner, P. (2012). Geostatistics: Modeling Spatial Uncertainty (John Wiley & Sons).
30. Oliver, M.A., and Webster, R. (2015). Basic Steps in Geostatistics: The Variogram and Kriging (Springer International Publishing).
31. Pastorello, N., Forbes, D.A., Foster, C., Brodie, J.P., Usher, C., Romanowsky, A.J., Strader, J., and Arnold, J.A. (2014). The SLUGGS survey: exploring the metallicity gradients of nearby early-type galaxies to large radii. Mon. Not. R. Astron. Soc. *442*, 1003–1039. https://doi.org/10.1093/mnras/stu937.
32. Wang, C., and Balch, W.E. (2018). Bridging Genomics to Phenomics at Atomic Resolution through Variation Spatial Profiling. Cell Rep. *24*, 2013–2028.e6. https://doi.org/10.1016/j.celrep.2018.07.059.
33. Wang, C., Elghobashi-Meinhardt, N., and Balch, W.E. (2022). Covariant Fitness Clusters Reveal Structural Evolution of SARS-CoV-2 Polymerase Across the Human Population. Preprint at bioRxiv. https://doi.org/10.1101/2022.01.07.475295.
34. Wang, C., Anglès, F., and Balch, W.E. (2022). Triangulating variation in the population to define mechanisms for precision management of genetic disease. Structure *30*, 1190–1207.e5. https://doi.org/10.1016/j.str.2022.05.011.
35. Wang, C., Scott, S.M., Subramanian, K., Loguercio, S., Zhao, P., Hutt, D.M., Farhat, N.Y., Porter, F.D., and Balch, W.E. (2019). Quantitating the epigenetic transformation contributing to cholesterol homeostasis using Gaussian process. Nat. Commun. *10*, 5052. https://doi.org/10.1038/s41467-019-12969-x.
36. Wang, C., Scott, S.M., Sun, S., Zhao, P., Hutt, D.M., Shao, H., Gestwicki, J.E., and Balch, W.E. (2020). Individualized management of genetic diversity in Niemann-Pick C1 through modulation of the Hsp70 chaperone system. Hum. Mol. Genet. *29*, 1–19. https://doi.org/10.1093/hmg/ddz215.
37. Anglès, F., Wang, C., and Balch, W.E. (2022). Spatial covariance analysis reveals the residue-by-residue thermodynamic contribution of variation to the CFTR fold. Commun. Biol. *5*, 356. https://doi.org/10.1038/s42003-022-03302-2.
38. Wang, C., Zhao, P., Sun, S., Wang, X., and Balch, W.E. (2022). Profiling genetic diversity reveals the molecular basis for balancing function with misfolding in alpha-1 antitrypsin. Preprint at bioRxiv. https://doi.org/10.1101/2022.03.04.483066.
39. Moya, A., Holmes, E.C., and González-Candelas, F. (2004). The population genetics and evolutionary epidemiology of RNA viruses. Nat. Rev. Microbiol. *2*, 279–288. https://doi.org/10.1038/nrmicro863.
40. Song, S., Ma, L., Zou, D., Tian, D., Li, C., Zhu, J., Chen, M., Wang, A., Ma, Y., Li, M., et al. (2020). The Global Landscape of SARS-CoV-2 Genomes, Variants, and Haplotypes in 2019nCoVR. Dev. Reprod. Biol. *18*, 749–759. https://doi.org/10.1016/j.gpb.2020.09.001.

41. Koelle, K., Martin, M.A., Antia, R., Lopman, B., and Dean, N.E. (2022). The changing epidemiology of SARS-CoV-2. Science *375*, 1116–1121. https://doi.org/10.1126/science.abm4915.

42. MacLean, F. (2021). Knowledge graphs and their applications in drug discovery. Expert Opin. Drug Discov. *16*, 1057–1069. https://doi.org/10.1080/17460441.2021.1910673.

43. Grubaugh, N.D., Petrone, M.E., and Holmes, E.C. (2020). We shouldn't worry when a virus mutates during disease outbreaks. Nat. Microbiol. *5*, 529–530. https://doi.org/10.1038/s41564-020-0690-4.

44. Geoghegan, J.L., and Holmes, E.C. (2018). Evolutionary Virology at 40. Genetics *210*, 1151–1162. https://doi.org/10.1534/genetics.118.301556.

45. Du, X., Wang, Z., Wu, A., Song, L., Cao, Y., Hang, H., and Jiang, T. (2008). Networks of genomic co-occurrence capture characteristics of human influenza A (H3N2) evolution. Genome Res. *18*, 178–187. https://doi.org/10.1101/gr.6969007.

46. Valen, L.V.A.N. (1973). A new evolutionary law. Evol. Theor. *1*. I–30.

47. Shu, Y., and McCauley, J. (2017). GISAID: Global initiative on sharing all influenza data - from vision to reality. Euro Surveill. *22*, 30494. https://doi.org/10.2807/1560-7917.ES.2017.22.13.30494.

48. Wang, C., Zhao, P., Sun, S., Teckman, J., and Balch, W.E. (2020). Leveraging Population Genomics for Individualized Correction of the Hallmarks of Alpha-1 Antitrypsin Deficiency. Chronic Obstr. Pulm. Dis. *7*, 224–246. https://doi.org/10.15326/jcopdf.7.3.2019.0167.

49. COVID-19 Data. (2020). Johns Hopkins University.

50. Cubuk, J., Alston, J.J., Incicco, J.J., Singh, S., Stuchell-Brereton, M.D., Ward, M.D., Zimmerman, M.I., Vithani, N., Griffith, D., Wagoner, J.A., et al. (2021). The SARS-CoV-2 nucleocapsid protein is dynamic, disordered, and phase separates with RNA. Nat. Commun. *12*, 1936. https://doi.org/10.1038/s41467-021-21953-3.

51. Nassif, A.B., Talib, M.A., Nasir, Q., and Dakalbab, F.M. (2021). Machine Learning for Anomaly Detection: A Systematic Review. IEEE Access *9*, 78658–78700. https://doi.org/10.1109/ACCESS.2021.3083060.

52. Saravanan, K.A., Panigrahi, M., Kumar, H., Rajawat, D., Nayak, S.S., Bhushan, B., and Dutt, T. (2022). Role of genomics in combating COVID-19 pandemic. Gene *823*, 146387. https://doi.org/10.1016/j.gene.2022.146387.

53. Galloway, S.E., Paul, P., MacCannell, D.R., Johansson, M.A., Brooks, J.T., MacNeil, A., Slayton, R.B., Tong, S., Silk, B.J., Armstrong, G.L., et al. (2021). Emergence of SARS-CoV-2 B.1.1.7 Lineage - United States, December 29, 2020-January 12, 2021. MMWR Morb. Mortal. Wkly. Rep. *70*, 95–99. https://doi.org/10.15585/mmwr.mm7003e2.

54. Tuekprakhon, A., Nutalai, R., Dijokaite-Guraliuc, A., Zhou, D., Ginn, H.M., Selvaraj, M., Liu, C., Mentzer, A.J., Supasa, P., Duyvesteyn, H.M.E., et al. (2022). Antibody escape of SARS-CoV-2 Omicron BA.4 and BA.5 from vaccine and BA.1 serum. Cell *185*, 2422–2433.e13. https://doi.org/10.1016/j.cell.2022.06.005.

55. Shrestha, L.B., Foster, C., Rawlinson, W., Tedla, N., and Bull, R.A. (2022). Evolution of the SARS-CoV-2 omicron variants BA.1 to BA.5: Implications for immune escape and transmission. Rev. Med. Virol. *32*, e2381. https://doi.org/10.1002/rmv.2381.

56. Rahimi, F., and Bezmin Abadi, A.T. (2022). The Omicron subvariant BA.2: Birth of a new challenge during the COVID-19 pandemic. Int. J. Surg. *99*, 106261. https://doi.org/10.1016/j.ijsu.2022.106261.

57. Lewnard, J.A., Hong, V.X., Patel, M.M., Kahn, R., Lipsitch, M., and Tartof, S.Y. (2022). Clinical outcomes associated with SARS-CoV-2 Omicron (B.1.1.529) variant and BA.1/BA.1.1 or BA.2 subvariant infection in southern California. Nat. Med. *28*, 1933–1943. https://doi.org/10.1038/s41591-022-01887-z.

58. Tiecco, G., Storti, S., Arsuffi, S., Degli Antoni, M., Focà, E., Castelli, F., and Quiros-Roldan, E. (2022). Omicron BA.2 Lineage, the "Stealth" Variant: Is It Truly a Silent Epidemic? A Literature Review. Int. J. Mol. Sci. *23*, 7315. https://doi.org/10.3390/ijms23137315.

59. Le, T.T.B., Vasanthakumaran, T., Thi Hien, H.N., Hung, I.C., Luu, M.N., Khan, Z.A., An, N.T., Tran, V.P., Lee, W.J., Abdul Aziz, J.M., et al. (2023). SARS-CoV-2 Omicron and its current known unknowns: A narrative review. Rev. Med. Virol. *33*, e2398. e2398. https://doi.org/10.1002/rmv.2398.

60. Wiegand, T., Nemudryi, A., Nemudraia, A., McVey, A., Little, A., Taylor, D.N., Walk, S.T., and Wiedenheft, B. (2022). The Rise and Fall of SARS-CoV-2 Variants and Ongoing Diversification of Omicron. Viruses *14*. https://doi.org/10.3390/v14092009.

61. Surjit, M., and Lal, S.K. (2009). The Nucleocapsid Protein of the SARS Coronavirus: Structure, Function and Therapeutic Potential. Molecular Biology of the SARS-Coronavirus *129*, 129–151. https://doi.org/10.1007/978-3-642-03683-5_9.

62. Harvey, W.T., Carabelli, A.M., Jackson, B., Gupta, R.K., Thomson, E.C., Harrison, E.M., Ludden, C., Reeve, R., Rambaut, A., COVID-19 Genomics UK COG-UK Consortium, Peacock, S.J., and Robertson, D.L. (2021). SARS-CoV-2 variants, spike mutations and immune escape. Nat. Rev. Microbiol. *19*, 409–424. https://doi.org/10.1038/s41579-021-00573-0.

63. Willett, B.J., Grove, J., MacLean, O.A., Wilkie, C., De Lorenzo, G., Furnon, W., Cantoni, D., Scott, S., Logan, N., Ashraf, S., et al. (2022). SARS-CoV-2 Omicron is an immune escape variant with an altered cell entry pathway. Nat. Microbiol. *7*, 1161–1179. https://doi.org/10.1038/s41564-022-01143-7.

64. McGrath, M.E., Xue, Y., Dillen, C., Oldfield, L., Assad-Garcia, N., Zaveri, J., Singh, N., Baracco, L., Taylor, L.J., Vashee, S., and Frieman, M.B. (2022). SARS-CoV-2 variant spike and accessory gene mutations alter pathogenesis. Proc. Natl. Acad. Sci. USA *119*, e2204717119. https://doi.org/10.1073/pnas.2204717119.

65. McCallum, M., Czudnochowski, N., Rosen, L.E., Zepeda, S.K., Bowen, J.E., Walls, A.C., Hauser, K., Joshi, A., Stewart, C., Dillen, J.R., et al. (2022). Structural basis of SARS-CoV-2 Omicron immune evasion and receptor engagement. Science *375*, 864–868. https://doi.org/10.1126/science.abn8652.

66. Hossain, A., Akter, S., Rashid, A.A., Khair, S., and Alam, A.S.M.R.U. (2022). Unique mutations in SARS-CoV-2 Omicron subvariants' non-spike proteins: Potential impacts on viral pathogenesis and host immune evasion. Microb. Pathog. *170*, 105699. https://doi.org/10.1016/j.micpath.2022.105699.

67. Obermeyer, F., Jankowiak, M., Barkas, N., Schaffner, S.F., Pyle, J.D., Yurkovetskiy, L., Bosso, M., Park, D.J., Babadi, M., MacInnis, B.L., et al. (2022). Analysis of 6.4 million SARS-CoV-2 genomes identifies mutations associated with fitness. Science *376*, 1327–1332. https://doi.org/10.1126/science.abm1208.

68. Maher, M.C., Bartha, I., Weaver, S., di Iulio, J., Ferri, E., Soriaga, L., Lempp, F.A., Hie, B.L., Bryson, B., Berger, B., et al. (2022). Predicting the mutational drivers of future SARS-CoV-2 variants of concern. Sci. Transl. Med. *14*, eabk3445. https://doi.org/10.1126/scitranslmed.abk3445.

69. Van Valen, L. (1973). A new evolutionary law. Evol. Theor. *1*, 1–30.

70. Karthikeyan, S., Levy, J.I., De Hoff, P., Humphrey, G., Birmingham, A., Jepsen, K., Farmer, S., Tubb, H.M., Valles, T., Tribelhorn, C.E., et al. (2022). Wastewater sequencing reveals early cryptic SARS-CoV-2 variant transmission. Nature *609*, 101–108. https://doi.org/10.1038/s41586-022-05049-6.

71. Sun, S., Wang, C., Zhao, P., Kline, G.M., Grandjean, J.M.D., Jiang, X., Labaudiniere, R., Wiseman, R.L., Kelly, J.W., and Balch, W.E. (2023). Capturing the conversion of the pathogenic alpha-1-antitrypsin fold by ATF6 enhanced proteostasis. Cell Chem. Biol. *30*, 22–42.e5. https://doi.org/10.1016/j.chembiol.2022.12.004.

72. Carroll, L. (1900). Through the Looking-Glass and what Alice Found There (W.B. Conkey Co.), p. 1900. ©1900).

73. CSSE (2021). COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at (Johns Hopkins University). https://github.com/CSSEGISandData/COVID-19.

74. O'Toole, A., Hill, V., Pybus, O.G., Watts, A., Bogoch, I.I., Khan, K., Messina, J.P., et al.; COVID-19 Genomics UK (COG-UK) consortium;

Network for Genomic Surveillance in South Africa (NGS-SA); Brazil-UK CADDE Genomic Network (2021). Tracking the international spread of SARS-CoV-2 lineages B.1.1.7 and B.1.351/501Y-V2. Wellcome Open Res *6*, 121. https://doi.org/10.12688/wellcomeopenres.16661.1.

75. Davies, N.G., Jarvis, C.I., CMMID COVID-19 Working Group, Edmunds, W.J., Jewell, N.P., Diaz-Ordaz, K., and Keogh, R.H. (2021). Increased mortality in community-tested cases of SARS-CoV-2 lineage B.1.1.7. Nature *593*, 270–274. https://doi.org/10.1038/s41586-021-03426-1.

76. cov-lineages.org (2021). Global Report - B.1.351. https://cov-lineages.org/global_report_B.1.351.html.

77. GISAID (2021). Tracking of SARS CoV2 Variants: B.1.617.2. 26 April 2021.

78. WHO (2021). Classification of Omicron (B.1.1.529): SARS-CoV-2 Variant of Concern. https://www.who.int/news/item/26-11-2021-classification-of-omicron-(b.1.1.529)-sars-cov-2-variant-of-concern.

79. Cameroni, E., Bowen, J.E., Rosen, L.E., Saliba, C., Zepeda, S.K., Culap, K., Pinto, D., VanBlargan, L.A., De Marco, A., di Iulio, J., et al. (2022). Broadly neutralizing antibodies overcome SARS-CoV-2 Omicron antigenic shift. Nature *602*, 664–670. https://doi.org/10.1038/s41586-021-04386-2.

80. Telenti, A., Hodcroft, E.B., and Robertson, D.L. (2022). The Evolution and Biology of SARS-CoV-2 Variants. Cold Spring Harb. Perspect. Med. *12*, a041390. https://doi.org/10.1101/cshperspect.a041390.

Patterns, Volume *4*

Supplemental information

Understanding the host-pathogen evolutionary balance through Gaussian process modeling of SARS-CoV-2

Salvatore Loguercio, Ben C. Calverley, Chao Wang, Daniel Shak, Pei Zhao, Shuhong Sun, G.R. Scott Budinger, and William E. Balch

Figure S1

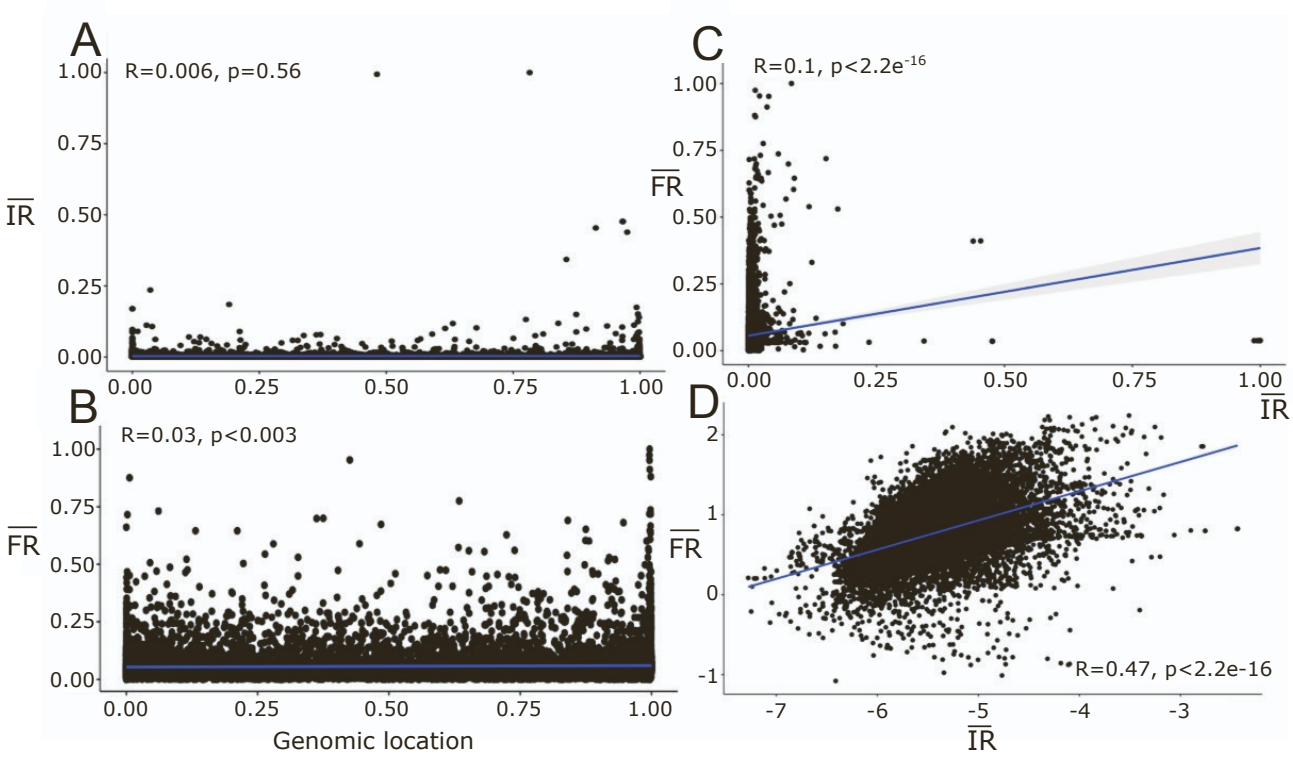

A
R=0.006, p=0.56
IR
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00
B
R=0.03, p<0.003
FR
Genomic location
C
R=0.1, p<2.2e-16
FR
IR
D
R=0.47, p<2.2e-16
FR
IR
2
1
0
-1
-7
-6
-5
-4
-3

Figure S2

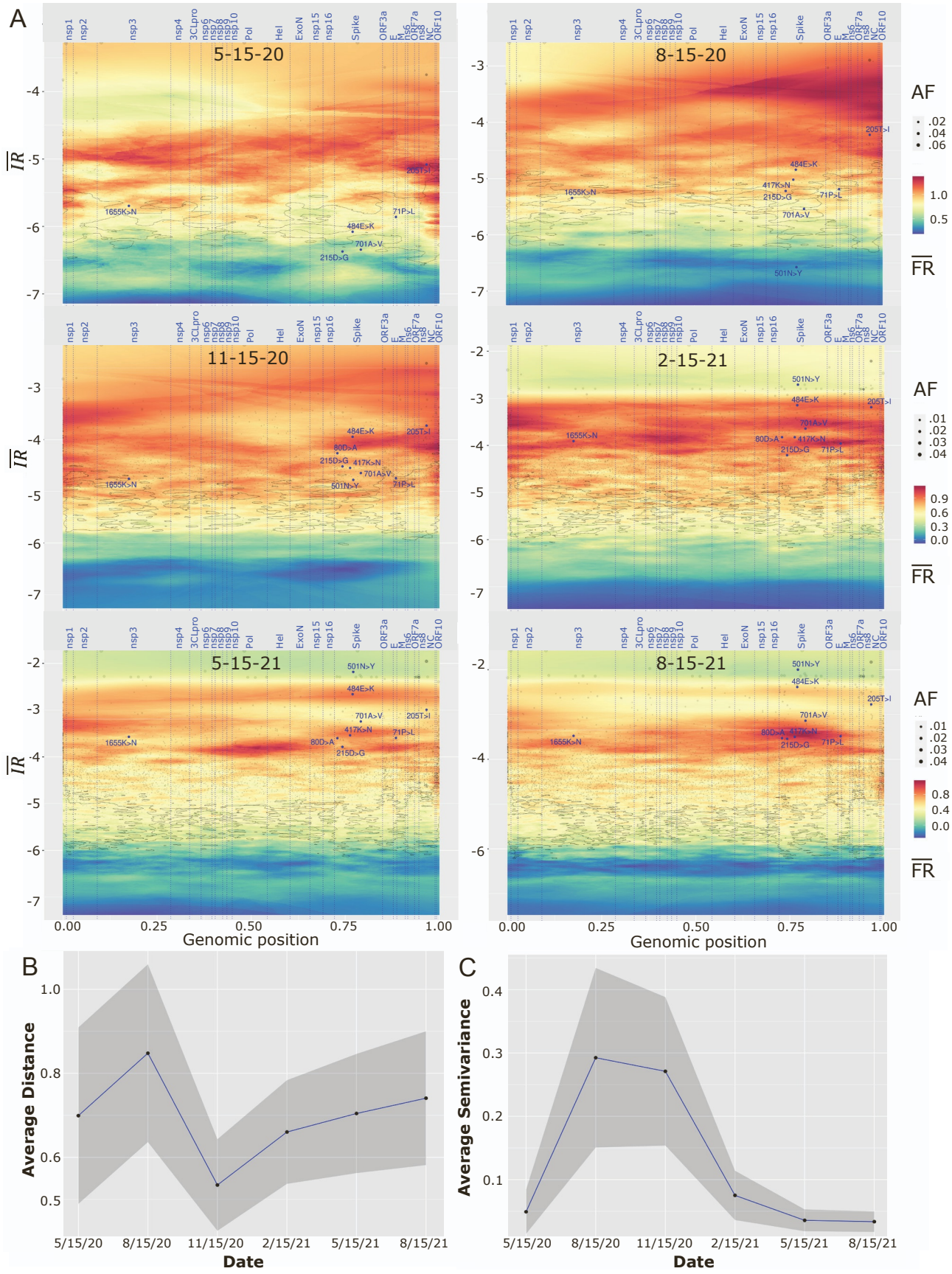

A
5-15-20
8-15-20
11-15-20
2-15-21
5-15-21
8-15-21
nsp1
nsp2
nsp3
nsp4
3CLpro
nsp6
nsp7
nsp8
nsp9
nsp10
Pol
Hel
ExoN
nsp15
nsp16
Spike
ORF3a
E
M
ORF6
ORF7a
ns8
NC
ORF10
1655K>N
205T>I
484E>K
71P>L
701A>V
215D>G
501N>Y
80D>A
417K>N
AF
FR
Genomic position
B
Average Distance
Date
C
Average Semivariance
Date
5/15/20
8/15/20
11/15/20
2/15/21
5/15/21
8/15/21

Figure S3

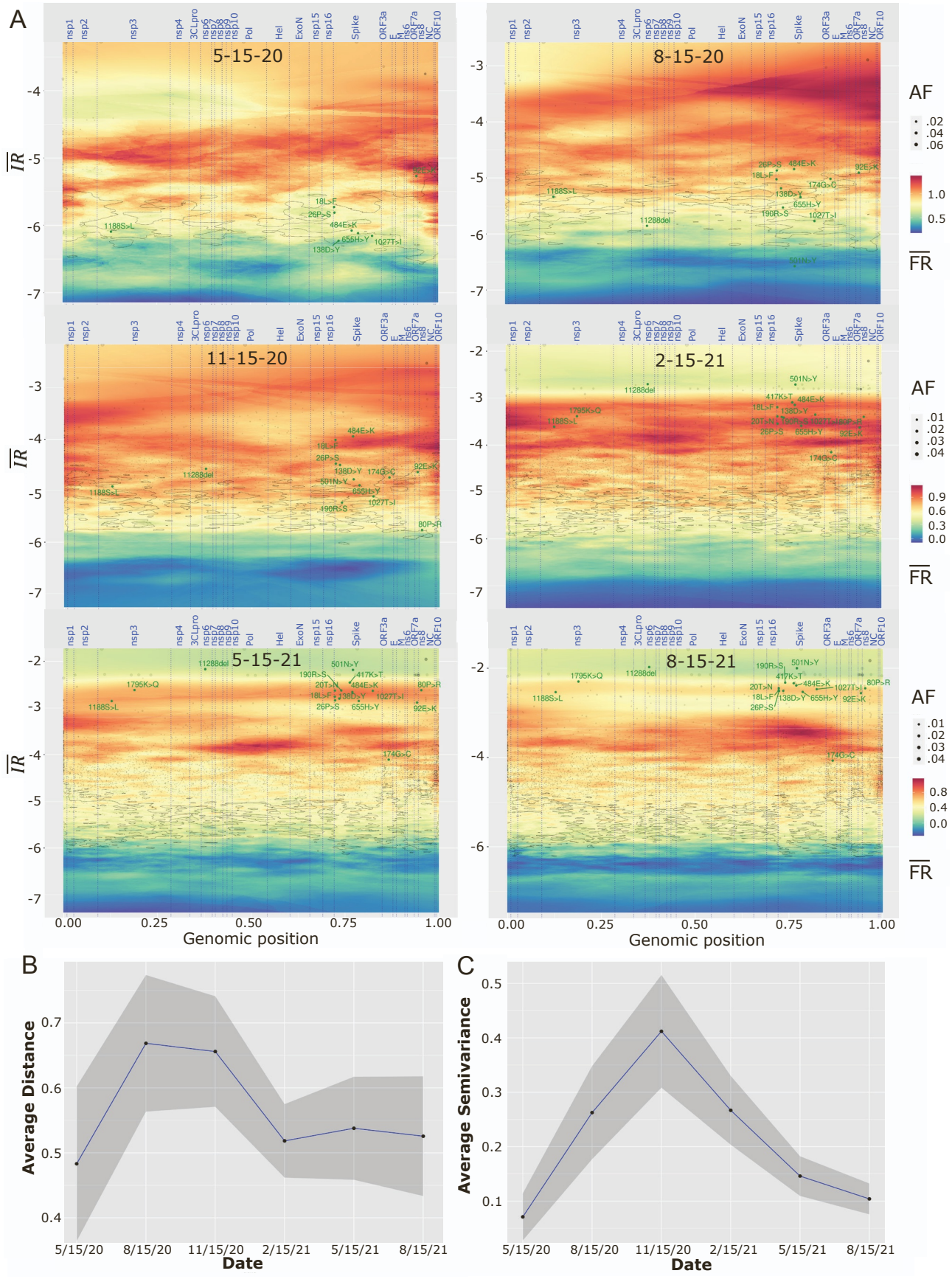

A
5-15-20
8-15-20
11-15-20
2-15-21
5-15-21
8-15-21
nsp1 nsp2 nsp3 nsp4 3CLpro nsp6 nsp7 nsp8 nsp9 nsp10 Pol Hel ExoN nsp15 nsp16 Spike ORF3a E M ORF6 ORF7a ns8 NC ORF10
IR
FR
AF
Genomic position
0.00 0.25 0.50 0.75 1.00
B
Average Distance
Date
5/15/20 8/15/20 11/15/20 2/15/21 5/15/21 8/15/21
C
Average Semivariance
Date
5/15/20 8/15/20 11/15/20 2/15/21 5/15/21 8/15/21

Figure S4

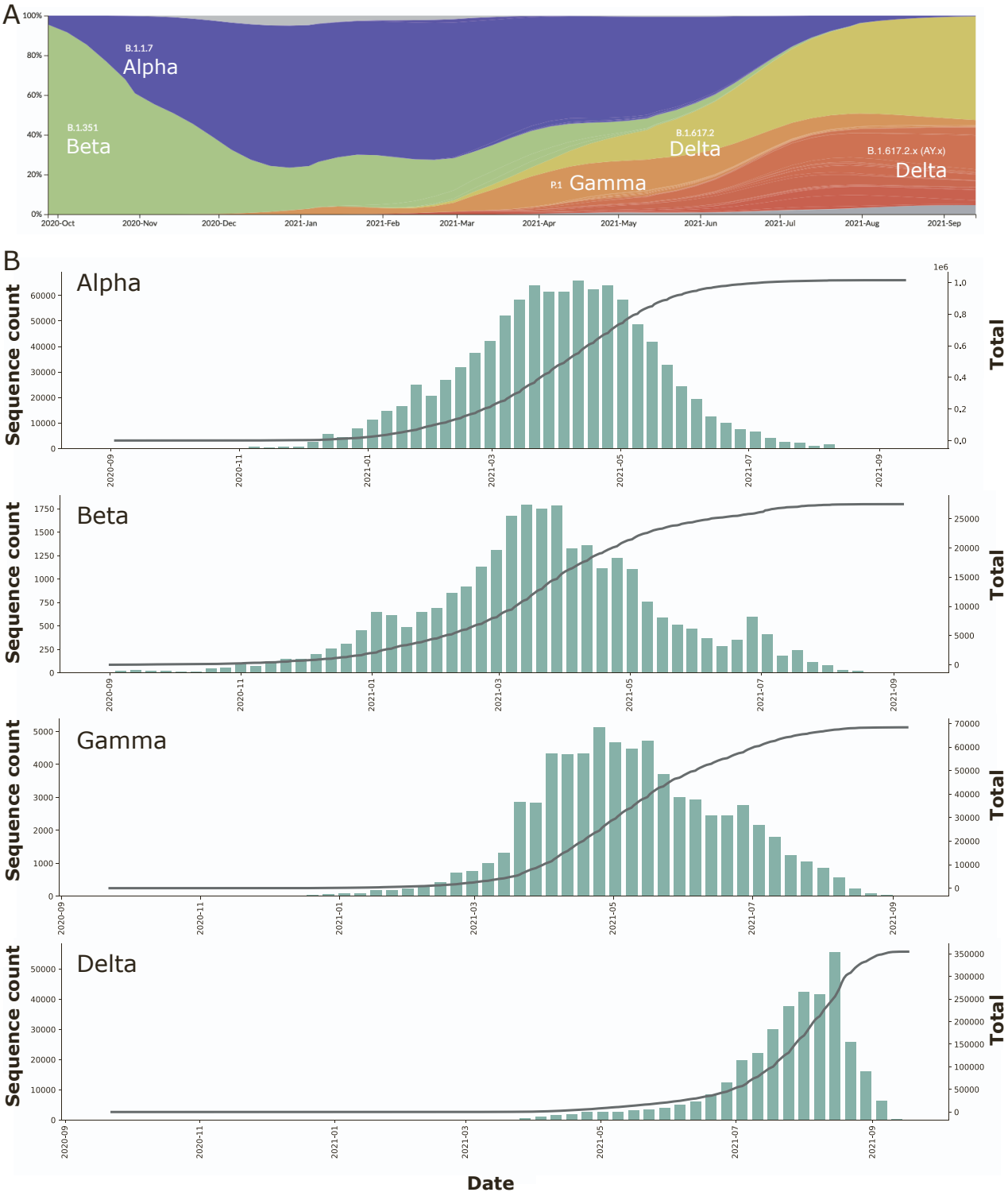

A
100%
80%
60%
40%
20%
0%
B.1.1.7
Alpha
B.1.351
Beta
B.1.617.2
Delta
P.1 Gamma
B.1.617.2.x (AY.x)
Delta
2020-Oct
2020-Nov
2020-Dec
2021-Jan
2021-Feb
2021-Mar
2021-Apr
2021-May
2021-Jun
2021-Jul
2021-Aug
2021-Sep
B
Alpha
Beta
Gamma
Delta
Sequence count
Total
1e6
Date

Figure S5

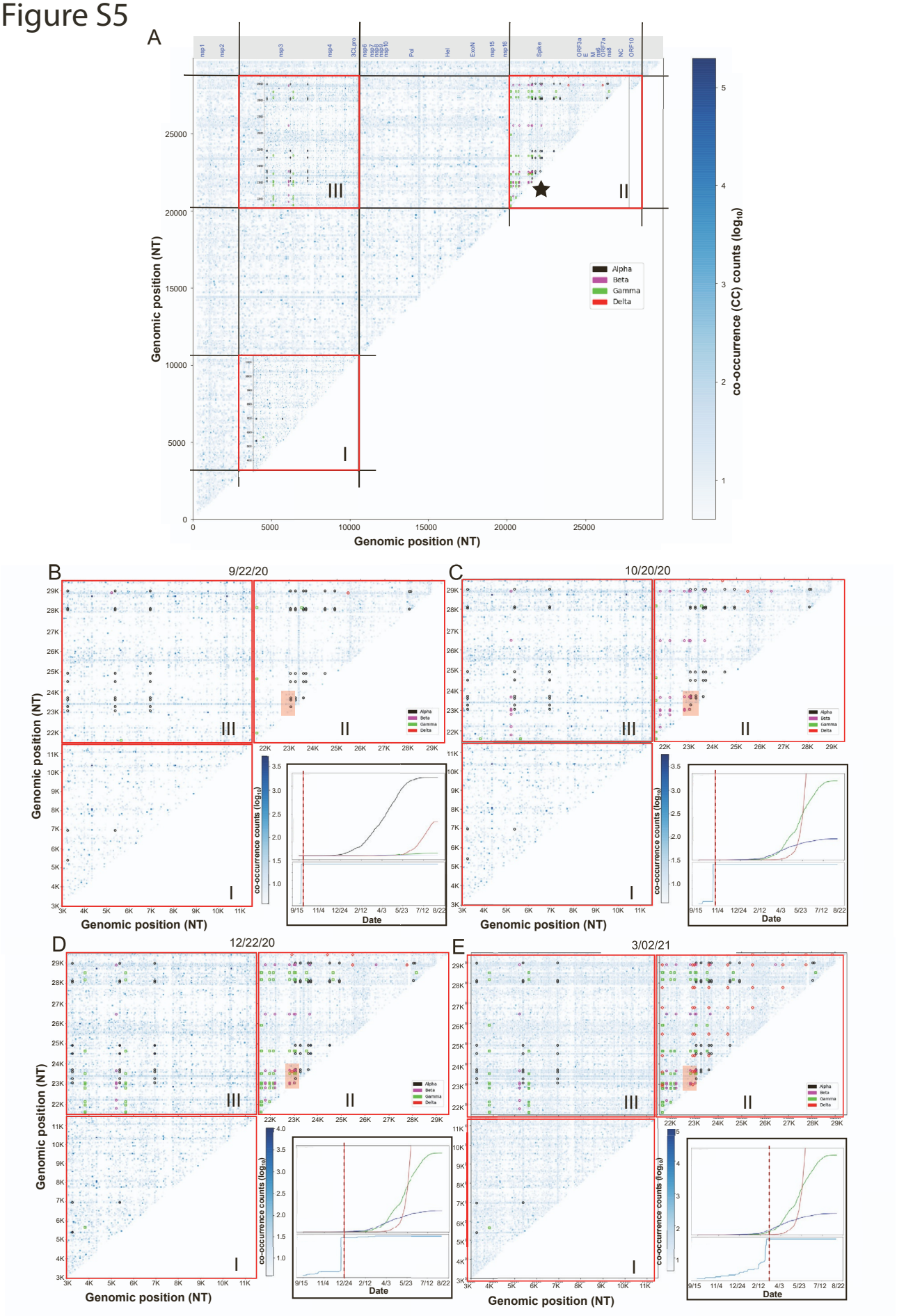

A
nsp1
nsp2
nsp3
nsp4
3CLpro
nsp6
nsp7
nsp8
nsp9
nsp10
Pol
Hel
ExoN
nsp15
nsp16
Spike
ORF3a
E
M
ORF6
ORF7a
ns8
NC
ORF10
Genomic position (NT)
co-occurrence (CC) counts (log10)
Alpha
Beta
Gamma
Delta
I
II
III
B
9/22/20
C
10/20/20
D
12/22/20
E
3/02/21
Date

Figure S6

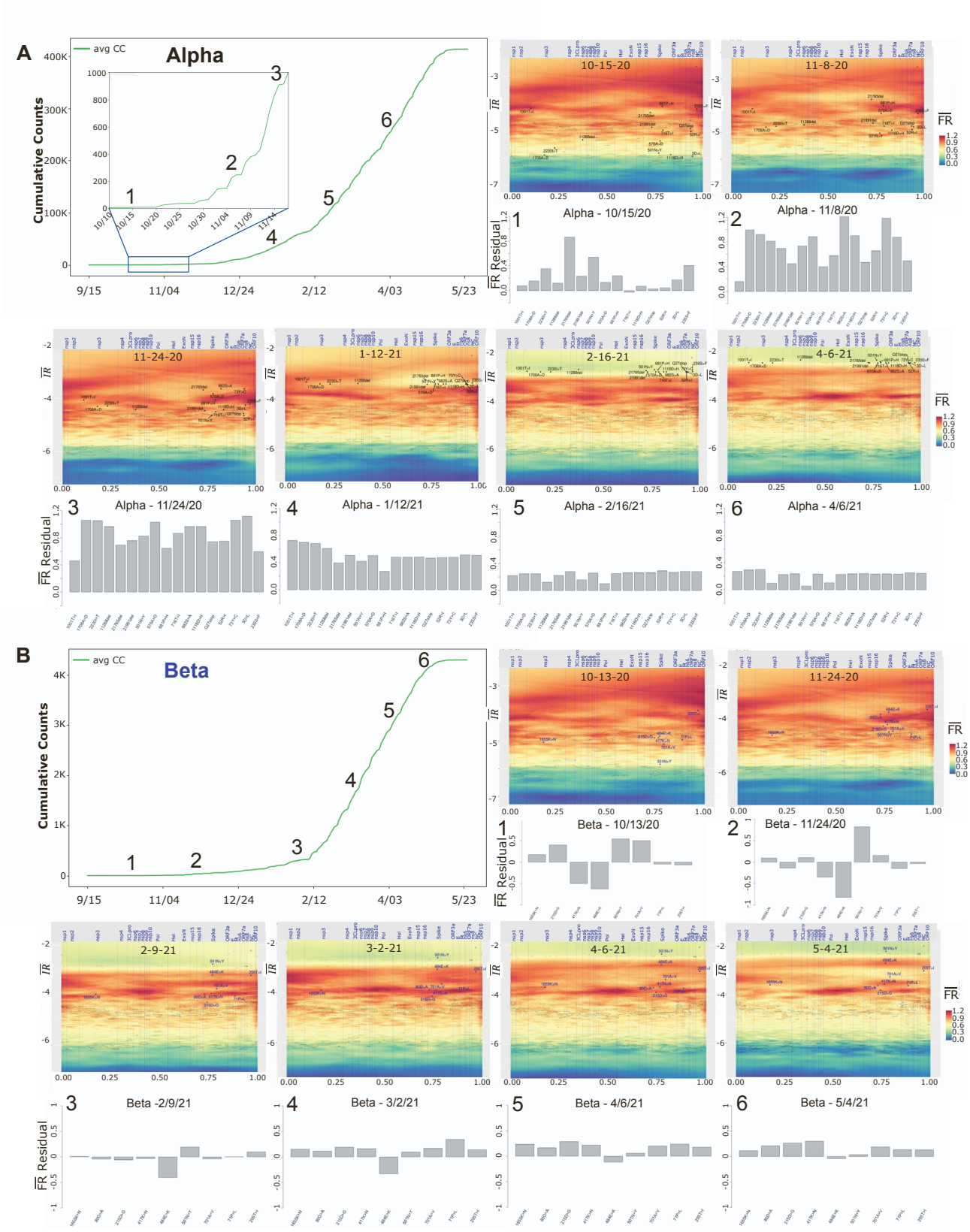

A
avg CC
Alpha
Cumulative Counts
400K
300K
200K
100K
0
9/15
11/04
12/24
2/12
4/03
5/23
10/15
10/20
10/25
10/30
11/04
11/09
11/14
1
2
3
4
5
6
10-15-20
11-8-20
11-24-20
1-12-21
2-16-21
4-6-21
FR
FR Residual
Alpha - 10/15/20
Alpha - 11/8/20
Alpha - 11/24/20
Alpha - 1/12/21
Alpha - 2/16/21
Alpha - 4/6/21
B
avg CC
Beta
Cumulative Counts
4K
3K
2K
1K
0
10-13-20
11-24-20
2-9-21
3-2-21
4-6-21
5-4-21
Beta - 10/13/20
Beta - 11/24/20
Beta -2/9/21
Beta - 3/2/21
Beta - 4/6/21
Beta - 5/4/21

Figure S7

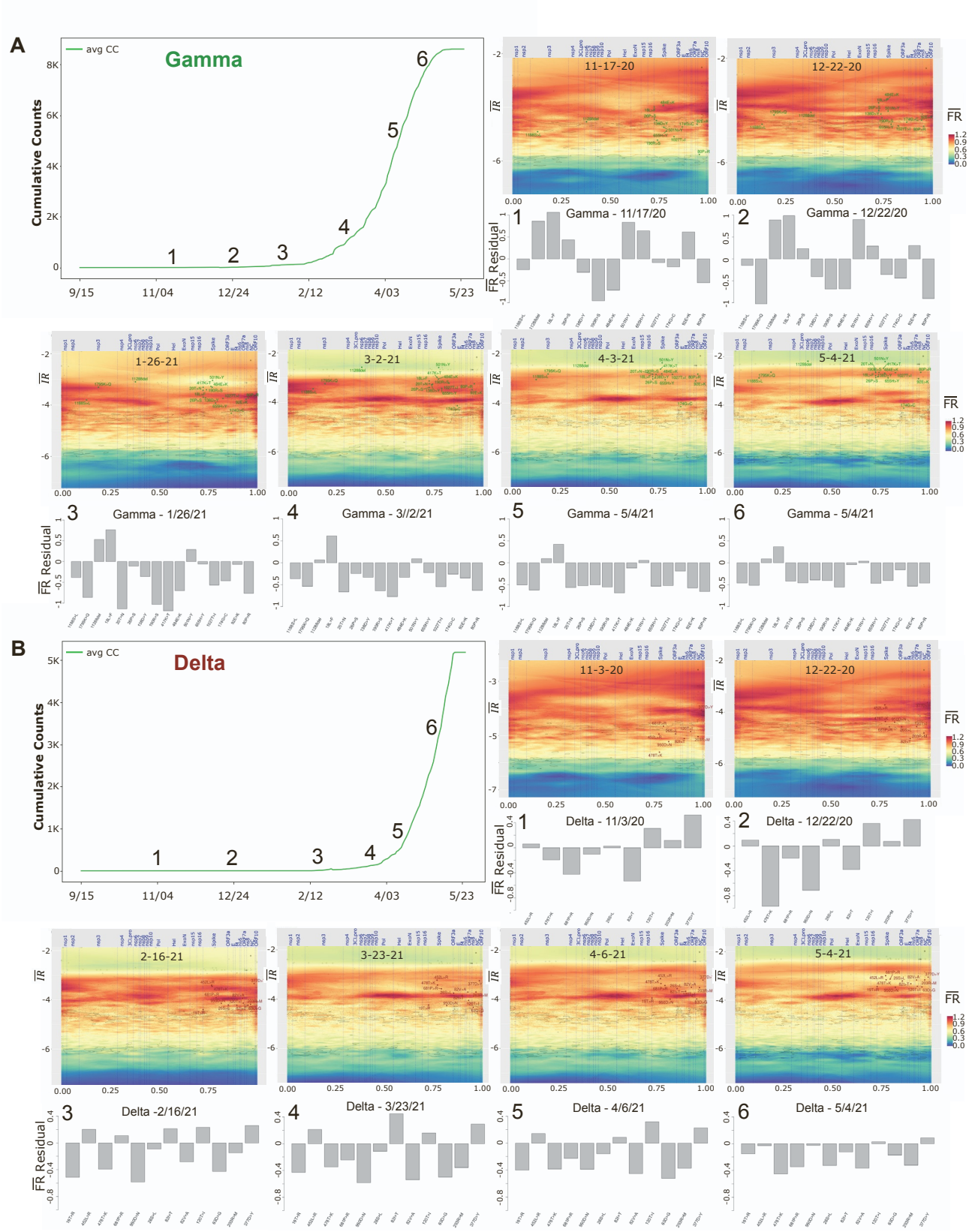
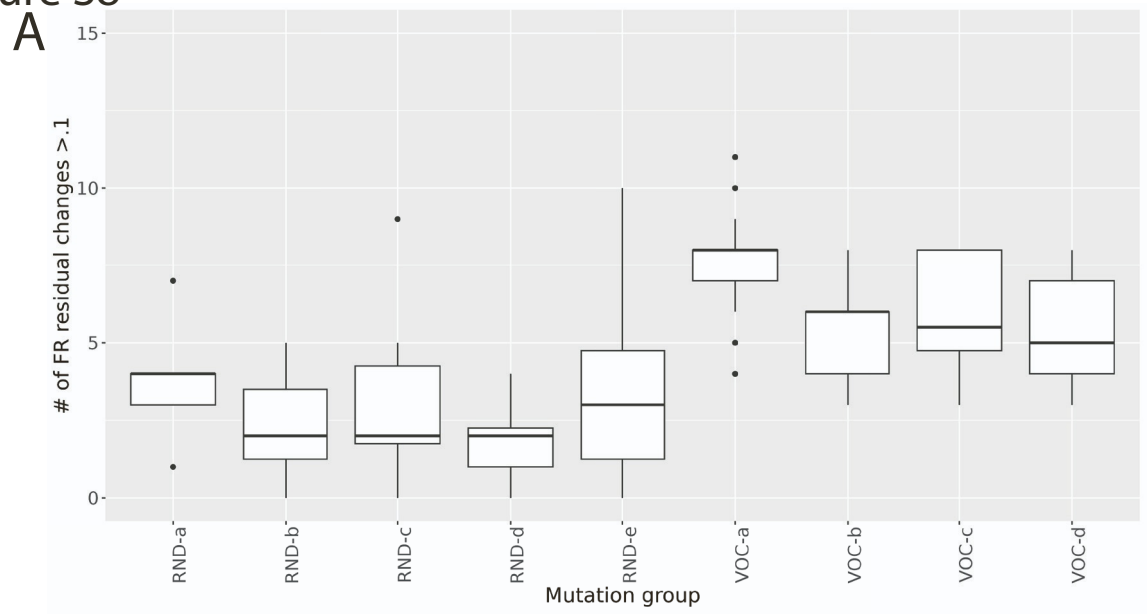
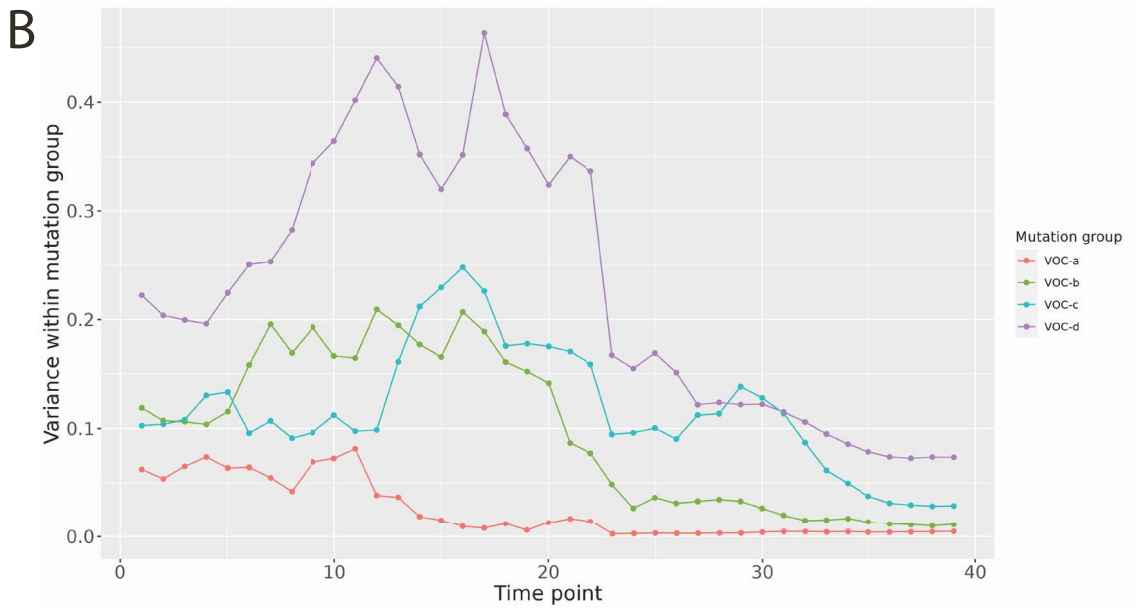

A
avg CC
Gamma
Cumulative Counts
8K
6K
4K
2K
0
9/15
11/04
12/24
2/12
4/03
5/23
1
2
3
4
5
6
11-17-20
12-22-20
1-26-21
3-2-21
4-3-21
5-4-21
FR
FR Residual
Gamma - 11/17/20
Gamma - 12/22/20
Gamma - 1/26/21
Gamma - 3//2/21
Gamma - 5/4/21
Gamma - 5/4/21
B
avg CC
Delta
Cumulative Counts
5K
4K
3K
2K
1K
0
9/15
11/04
12/24
2/12
4/03
5/23
11-3-20
12-22-20
2-16-21
3-23-21
4-6-21
5-4-21
FR
FR Residual
Delta - 11/3/20
Delta - 12/22/20
Delta -2/16/21
Delta - 3/23/21
Delta - 4/6/21
Delta - 5/4/21

Figure S8


A
15
10
5
0
# of FR residual changes >.1
RND-a
RND-b
RND-c
RND-d
RND-e
VOC-a
VOC-b
VOC-c
VOC-d
Mutation group


B
0.4
0.3
0.2
0.1
0.0
Variance within mutation group
0
10
20
30
40
Time point
Mutation group
VOC-a
VOC-b
VOC-c
VOC-d

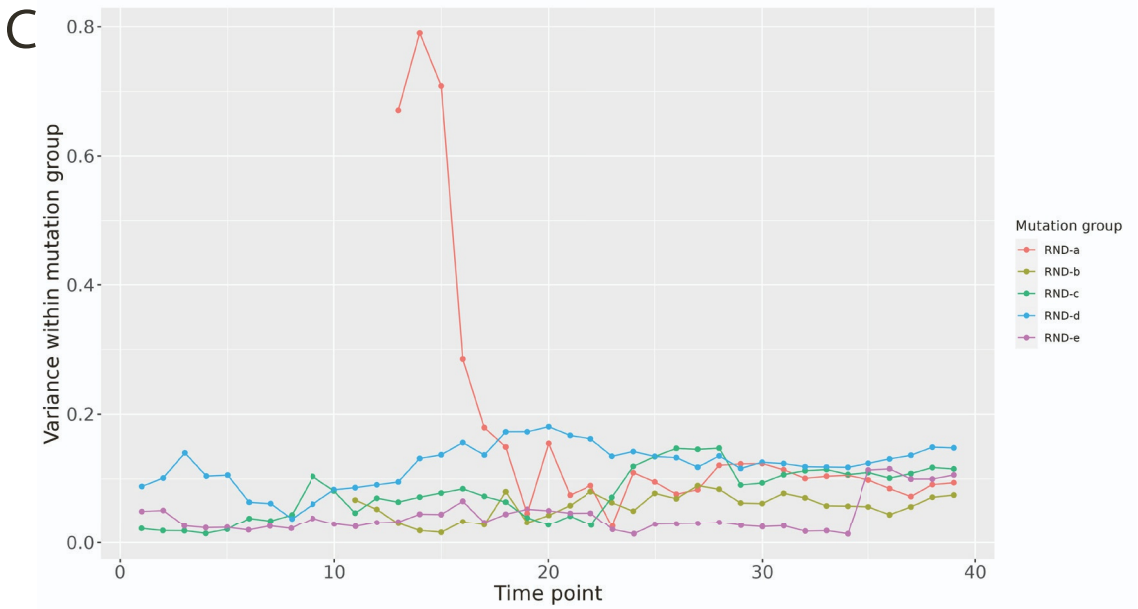

C
0.8
0.6
0.4
0.2
0.0
Variance within mutation group
0
10
20
30
40
Time point
Mutation group
RND-a
RND-b
RND-c
RND-d
RND-e

Figure S9

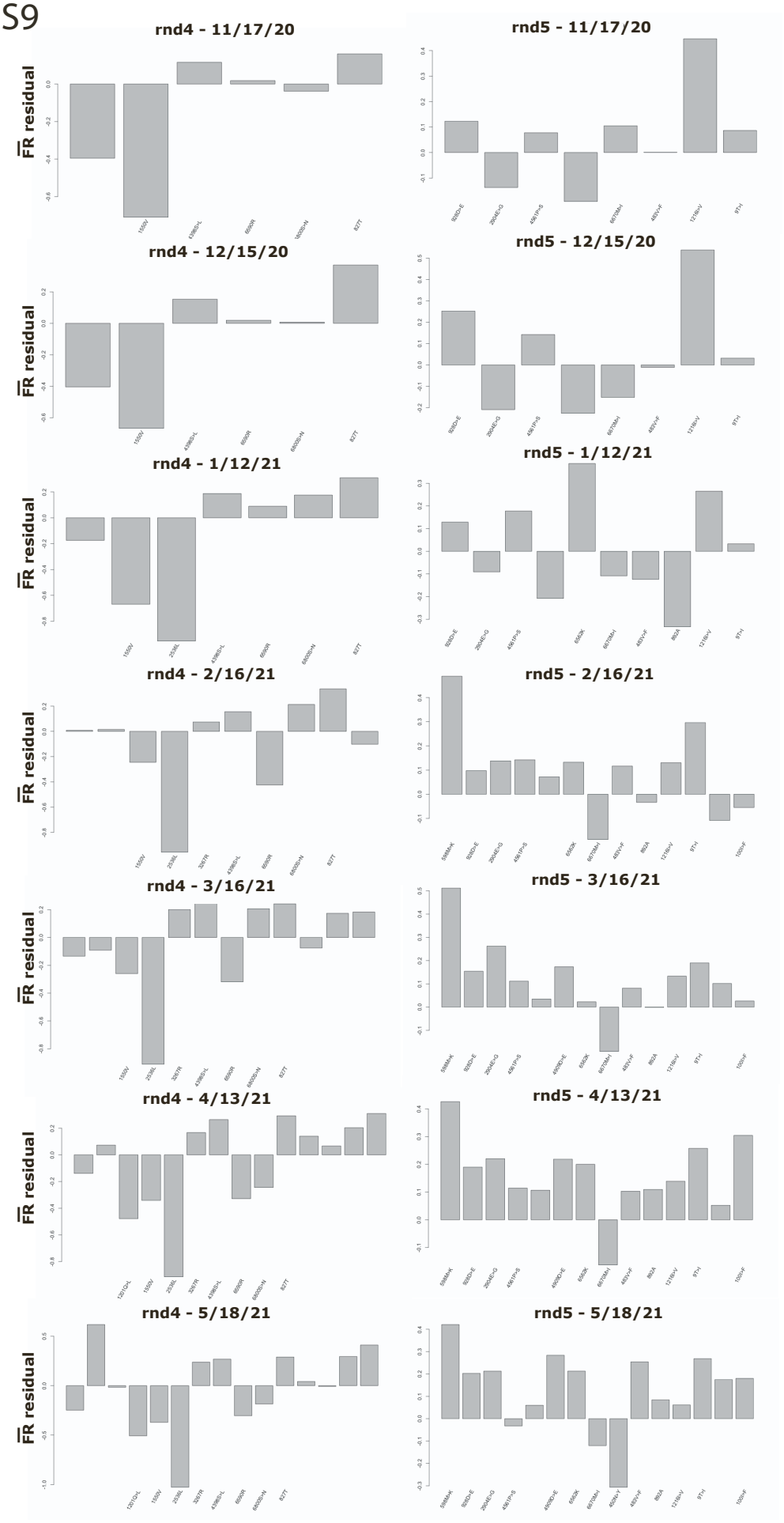

rnd4 - 11/17/20
rnd5 - 11/17/20
rnd4 - 12/15/20
rnd5 - 12/15/20
rnd4 - 1/12/21
rnd5 - 1/12/21
rnd4 - 2/16/21
rnd5 - 2/16/21
rnd4 - 3/16/21
rnd5 - 3/16/21
rnd4 - 4/13/21
rnd5 - 4/13/21
rnd4 - 5/18/21
rnd5 - 5/18/21
FR residual

Figure S10

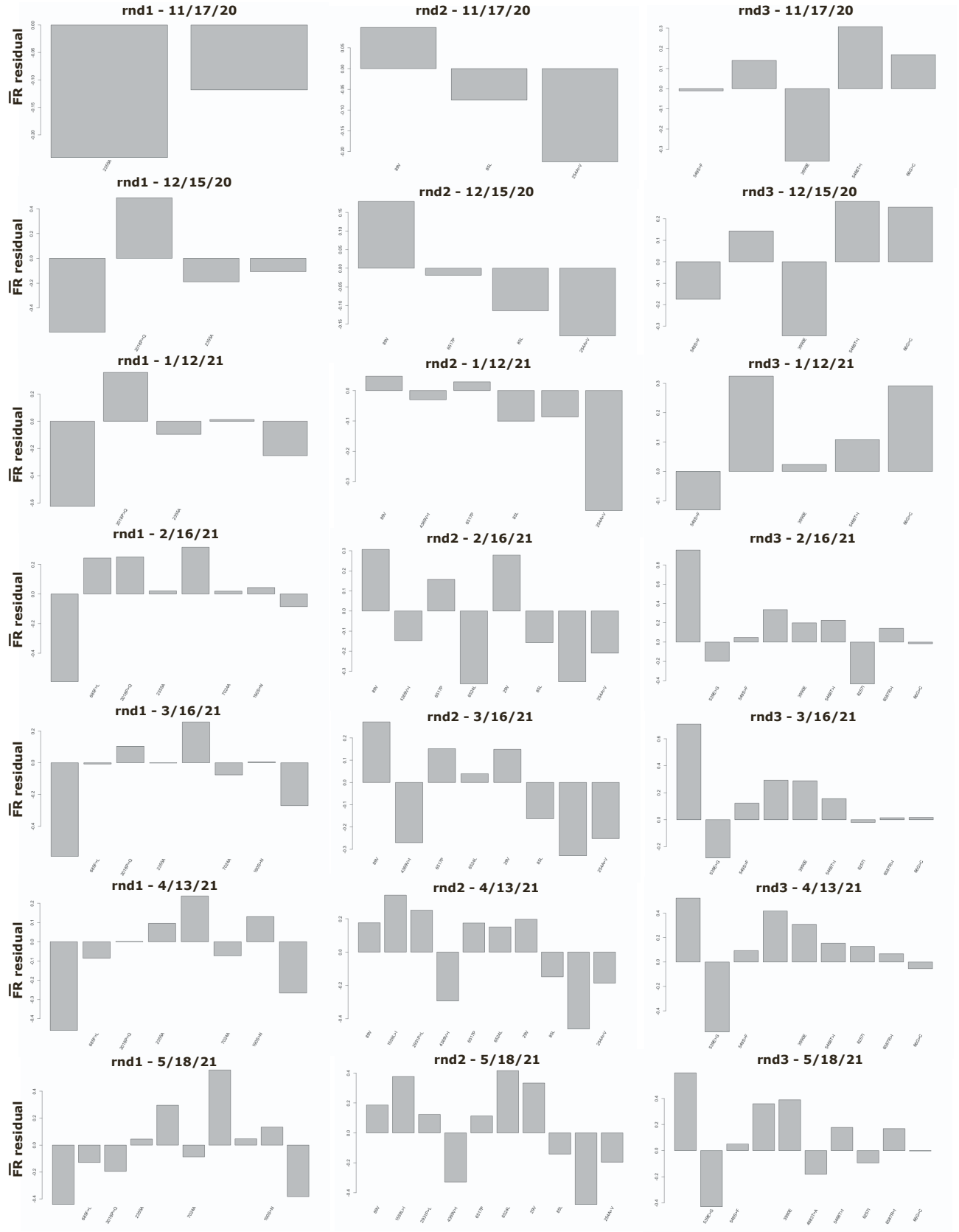

rnd1 - 11/17/20
rnd2 - 11/17/20
rnd3 - 11/17/20
rnd1 - 12/15/20
rnd2 - 12/15/20
rnd3 - 12/15/20
rnd1 - 1/12/21
rnd2 - 1/12/21
rnd3 - 1/12/21
rnd1 - 2/16/21
rnd2 - 2/16/21
rnd3 - 2/16/21
rnd1 - 3/16/21
rnd2 - 3/16/21
rnd3 - 3/16/21
rnd1 - 4/13/21
rnd2 - 4/13/21
rnd3 - 4/13/21
rnd1 - 5/18/21
rnd2 - 5/18/21
rnd3 - 5/18/21
FR residual

Figure S11

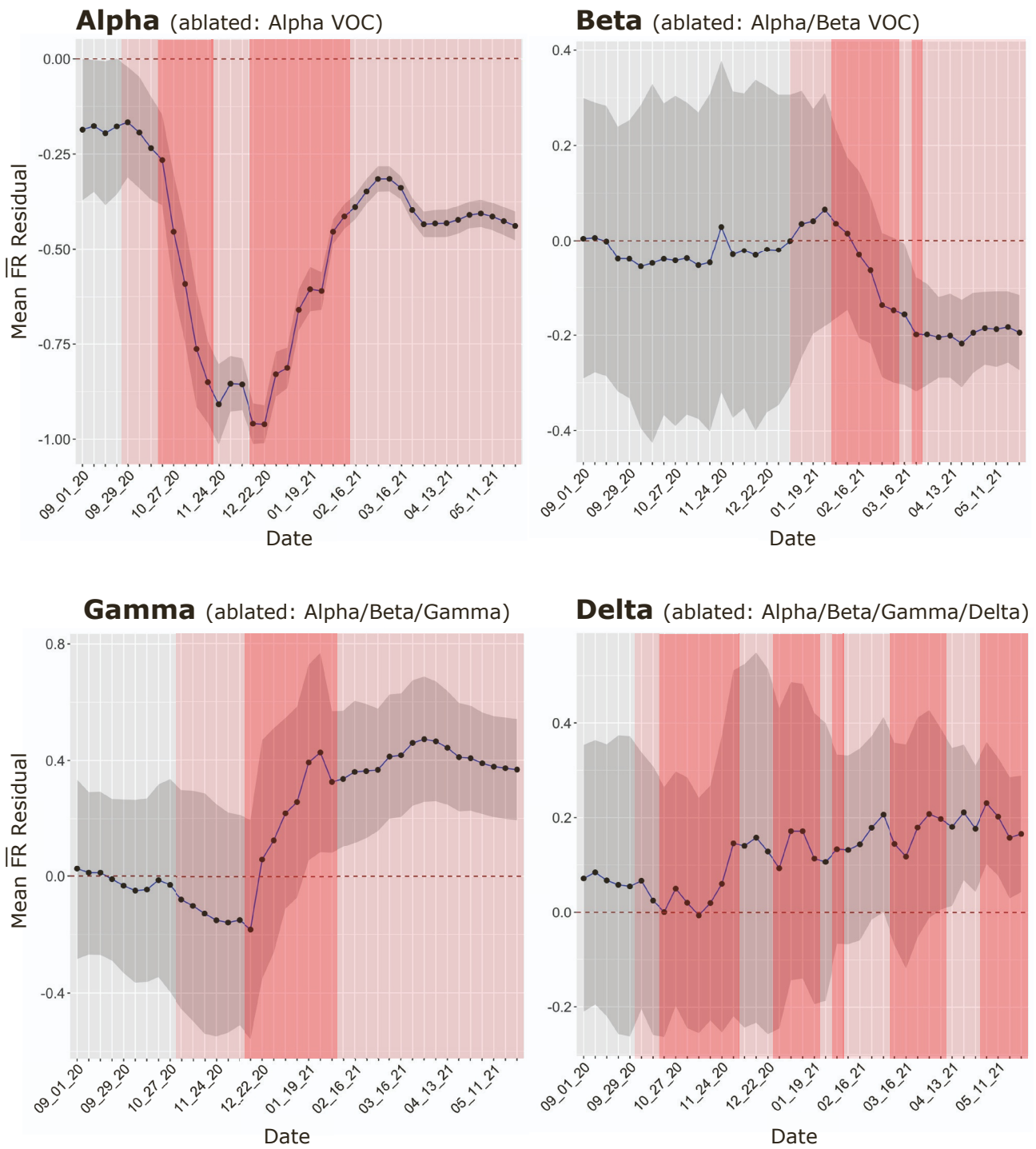

Alpha (ablated: Alpha VOC)
Beta (ablated: Alpha/Beta VOC)
Gamma (ablated: Alpha/Beta/Gamma)
Delta (ablated: Alpha/Beta/Gamma/Delta)
Mean FR Residual
Date
09_01_20
09_29_20
10_27_20
11_24_20
12_22_20
01_19_21
02_16_21
03_16_21
04_13_21
05_11_21

Figure S12

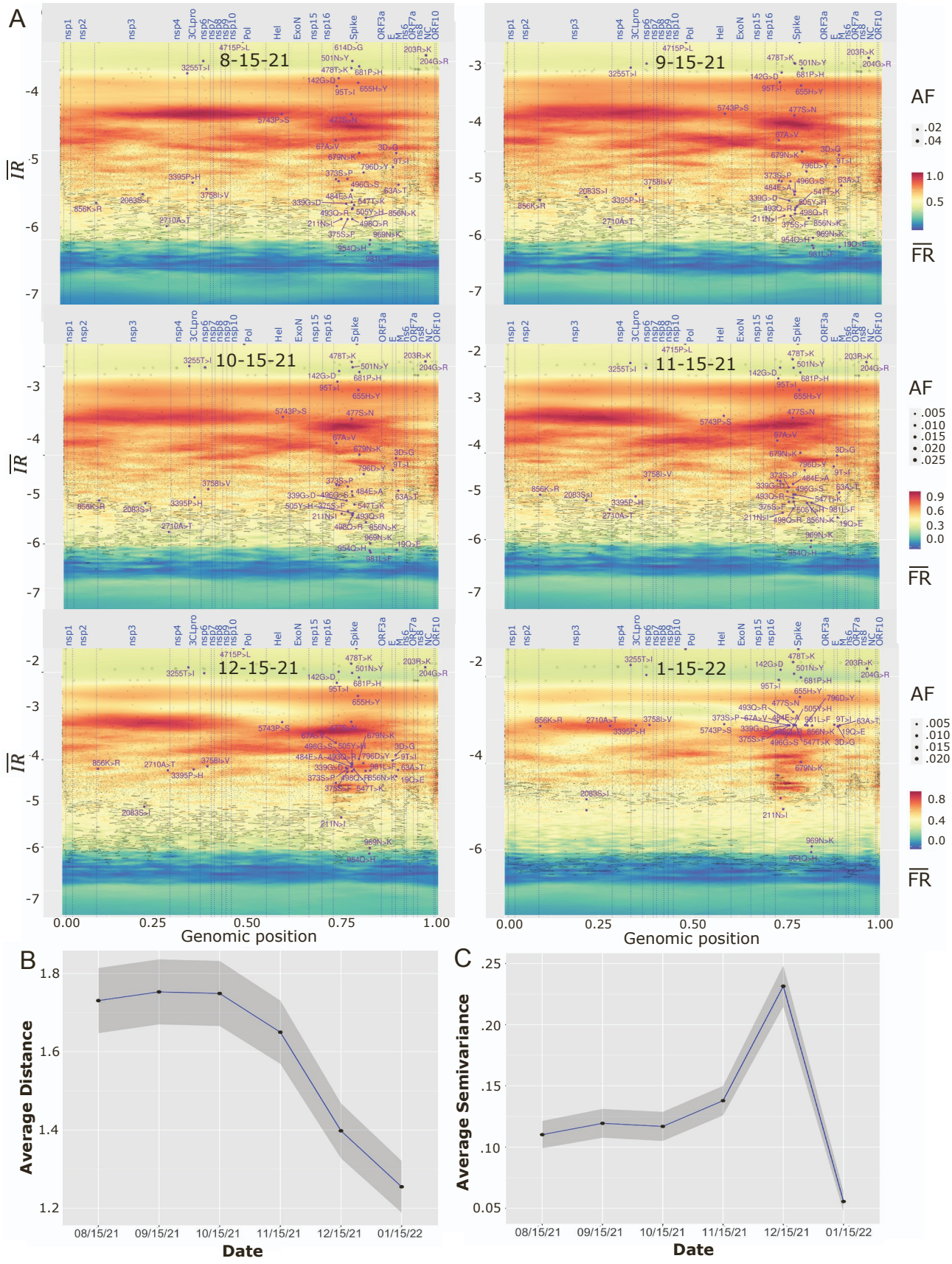

A
8-15-21
9-15-21
10-15-21
11-15-21
12-15-21
1-15-22
IR
FR
AF
Genomic position
0.00
0.25
0.50
0.75
1.00
B
Average Distance
1.8
1.6
1.4
1.2
08/15/21
09/15/21
10/15/21
11/15/21
12/15/21
01/15/22
Date
C
Average Semivariance
.25
.20
.15
.10
0.05
Date

Figure S13

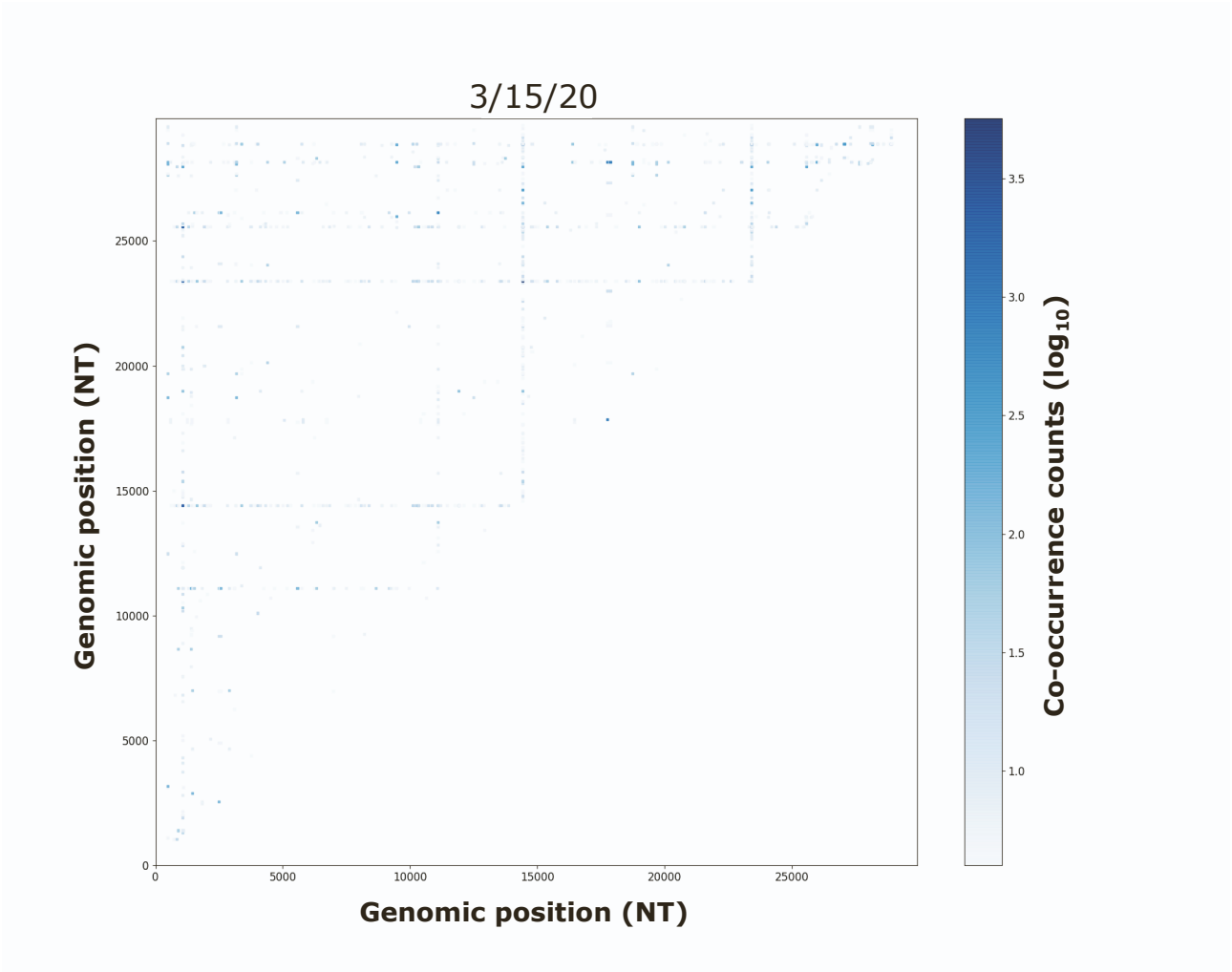

3/15/20
Genomic position (NT)
0
5000
10000
15000
20000
25000
Genomic position (NT)
Co-occurrence counts (log10)
1.0
1.5
2.0
2.5
3.0
3.5

Figure S14

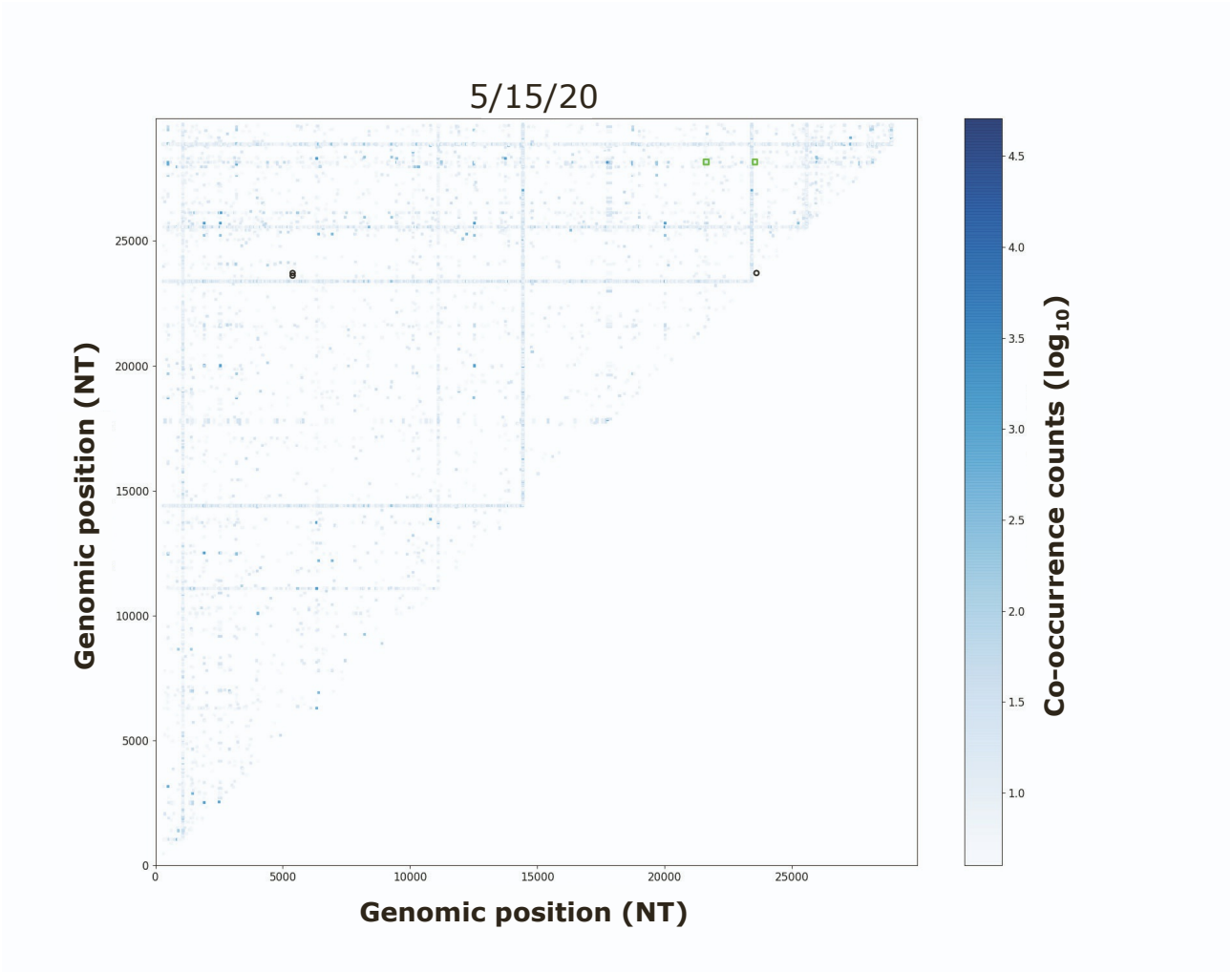

5/15/20
Genomic position (NT)
0
5000
10000
15000
20000
25000
Genomic position (NT)
Co-occurrence counts (log10)
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5

Figure S15

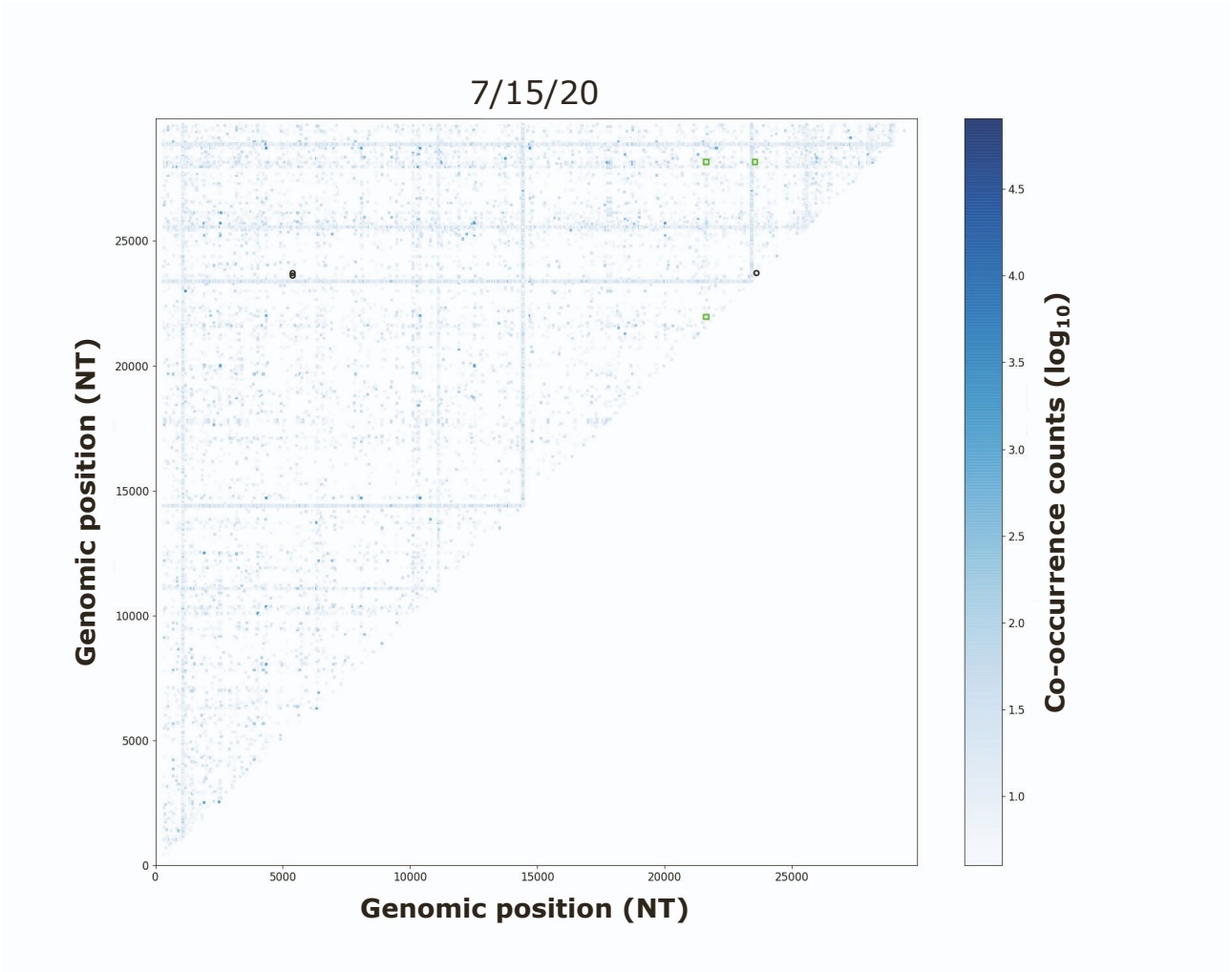

7/15/20
Genomic position (NT)
0
5000
10000
15000
20000
25000
Genomic position (NT)
Co-occurrence counts (log10)
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5

Figure S16

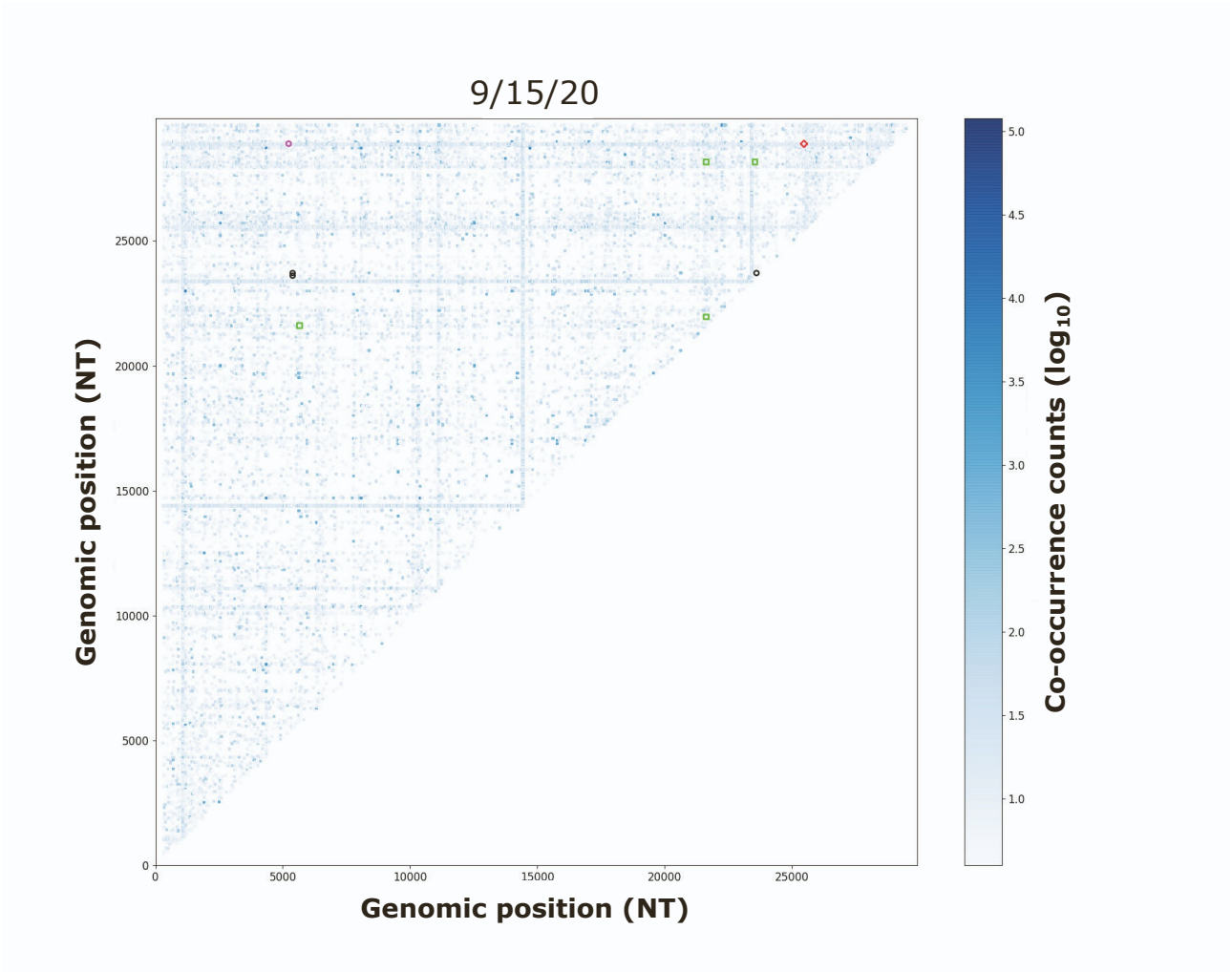

9/15/20
Genomic position (NT)
0
5000
10000
15000
20000
25000
Genomic position (NT)
Co-occurrence counts (log10)
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0

Figure S17

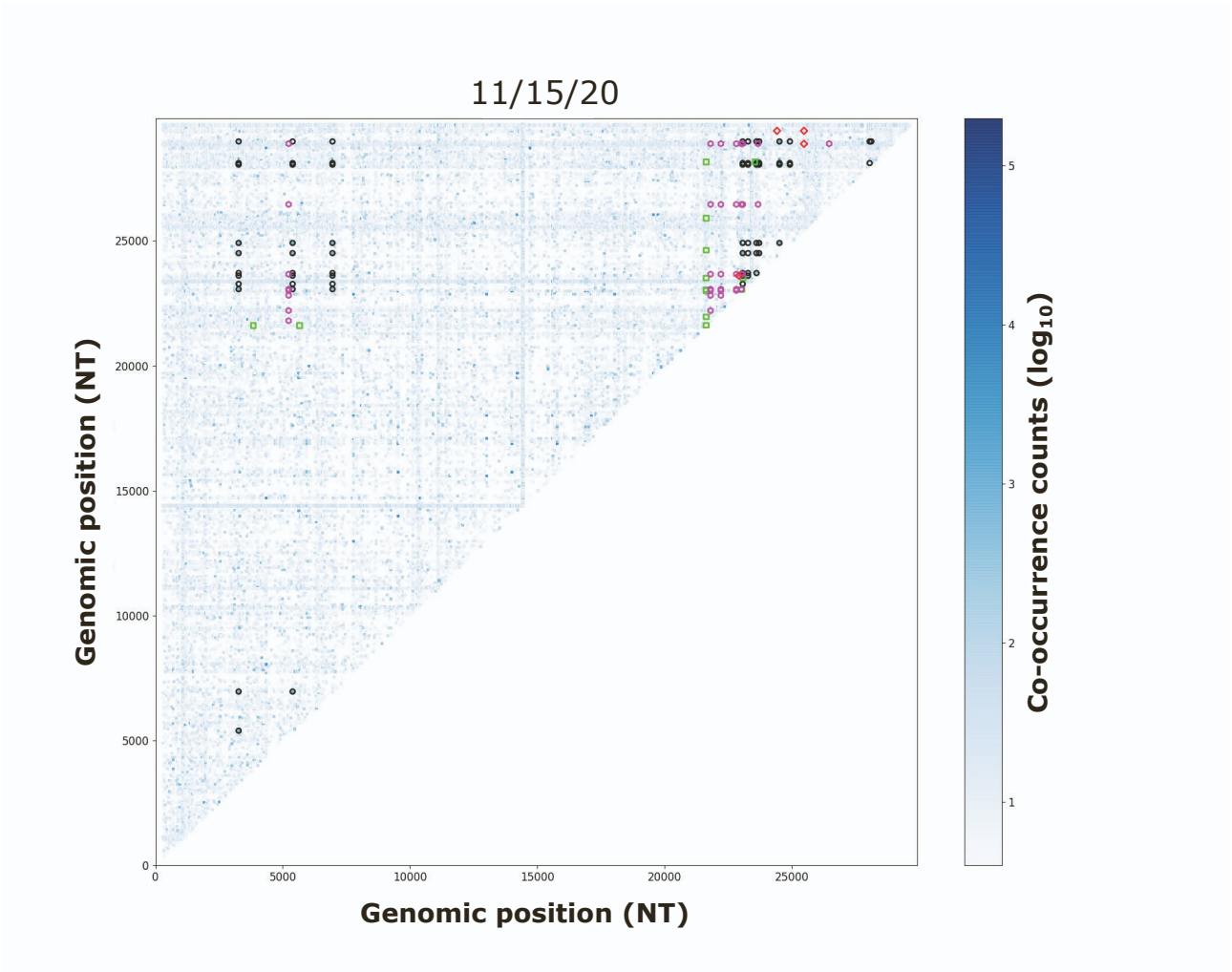

11/15/20
Genomic position (NT)
Genomic position (NT)
Co-occurrence counts (log10)
0
5000
10000
15000
20000
25000
1
2
3
4
5

Figure S18

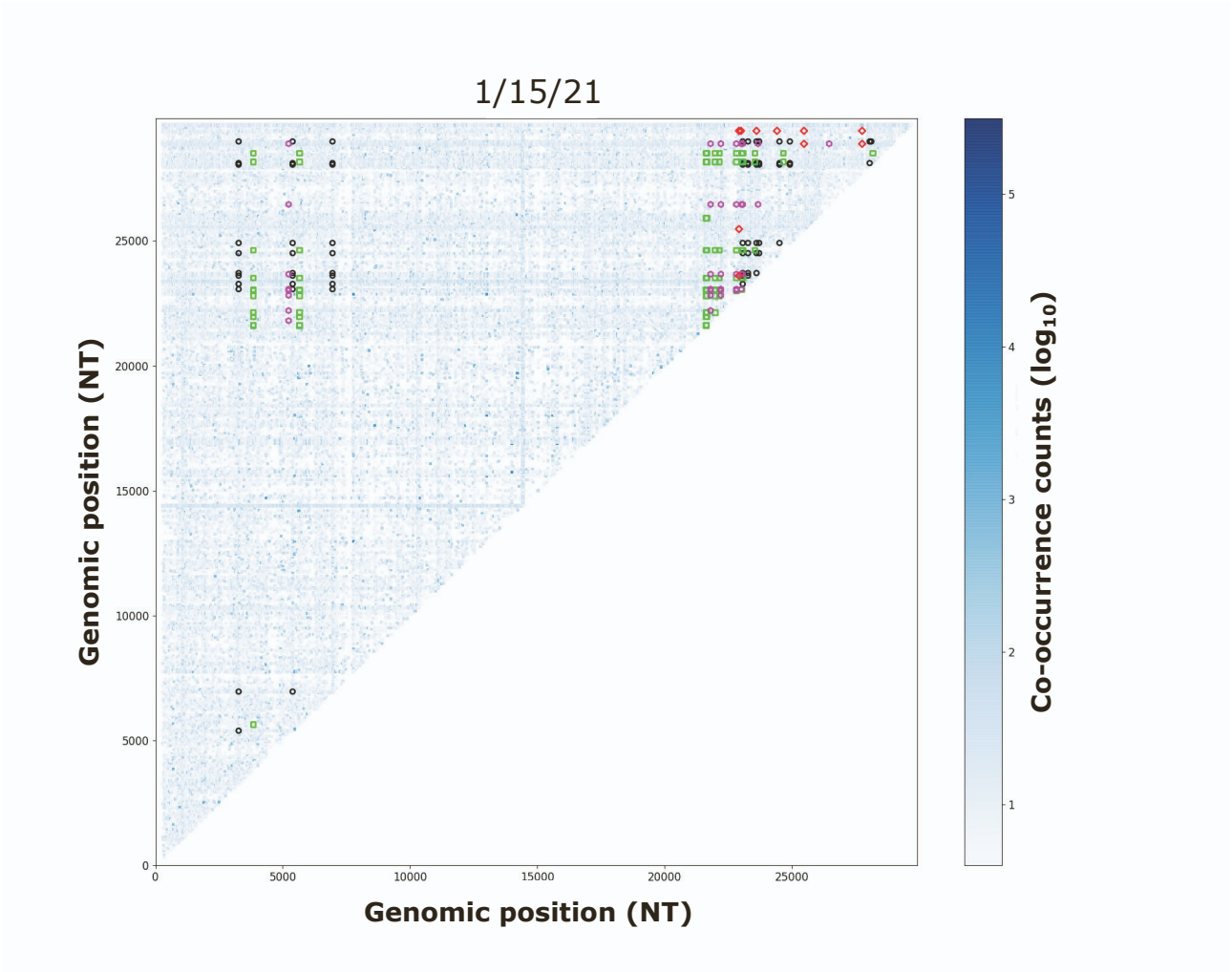

1/15/21
Genomic position (NT)
0
5000
10000
15000
20000
25000
Genomic position (NT)
Co-occurrence counts (log10)
1
2
3
4
5

Figure S19

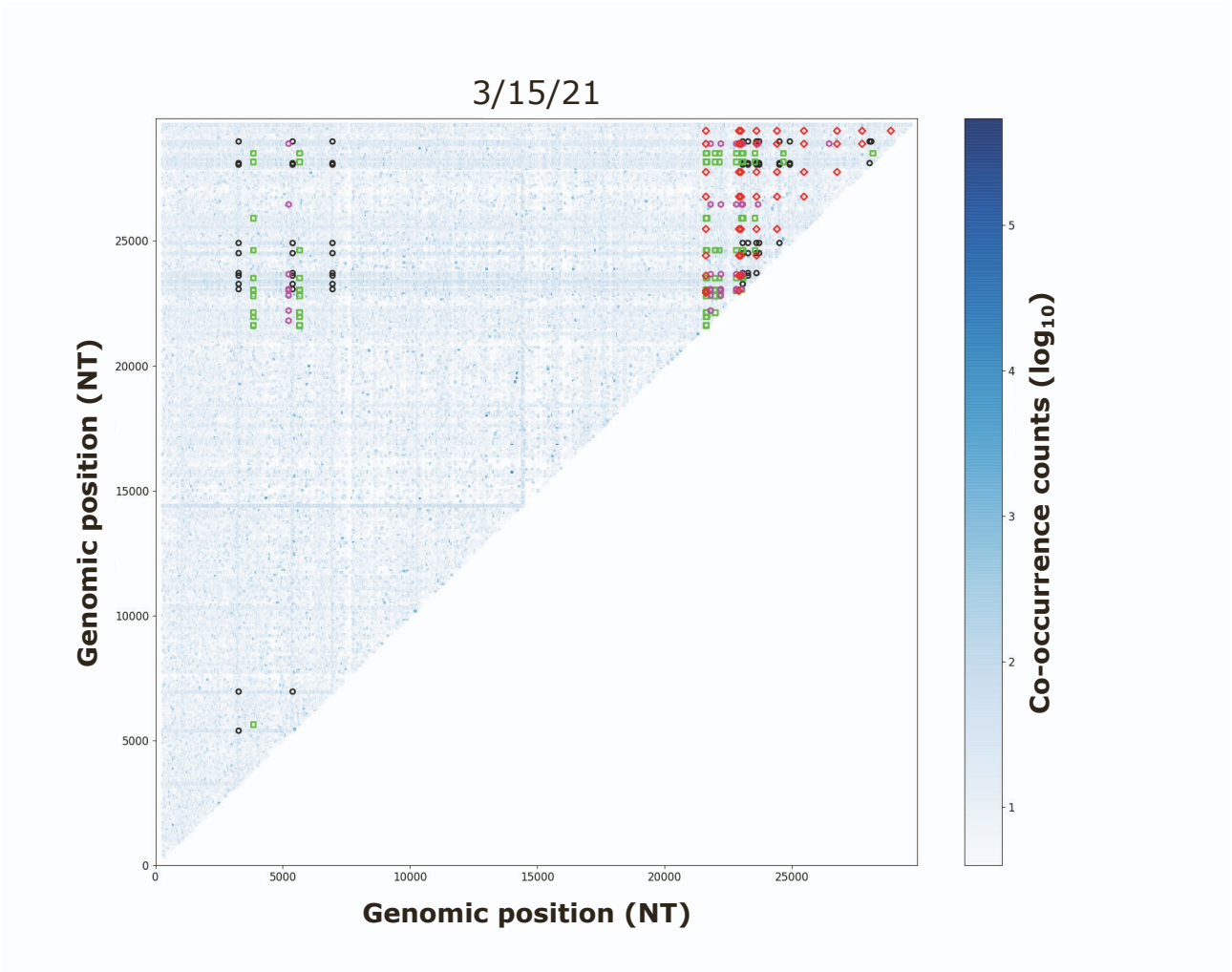

3/15/21
Genomic position (NT)
Genomic position (NT)
0
5000
10000
15000
20000
25000
Co-occurrence counts (log10)
1
2
3
4
5

Figure S20

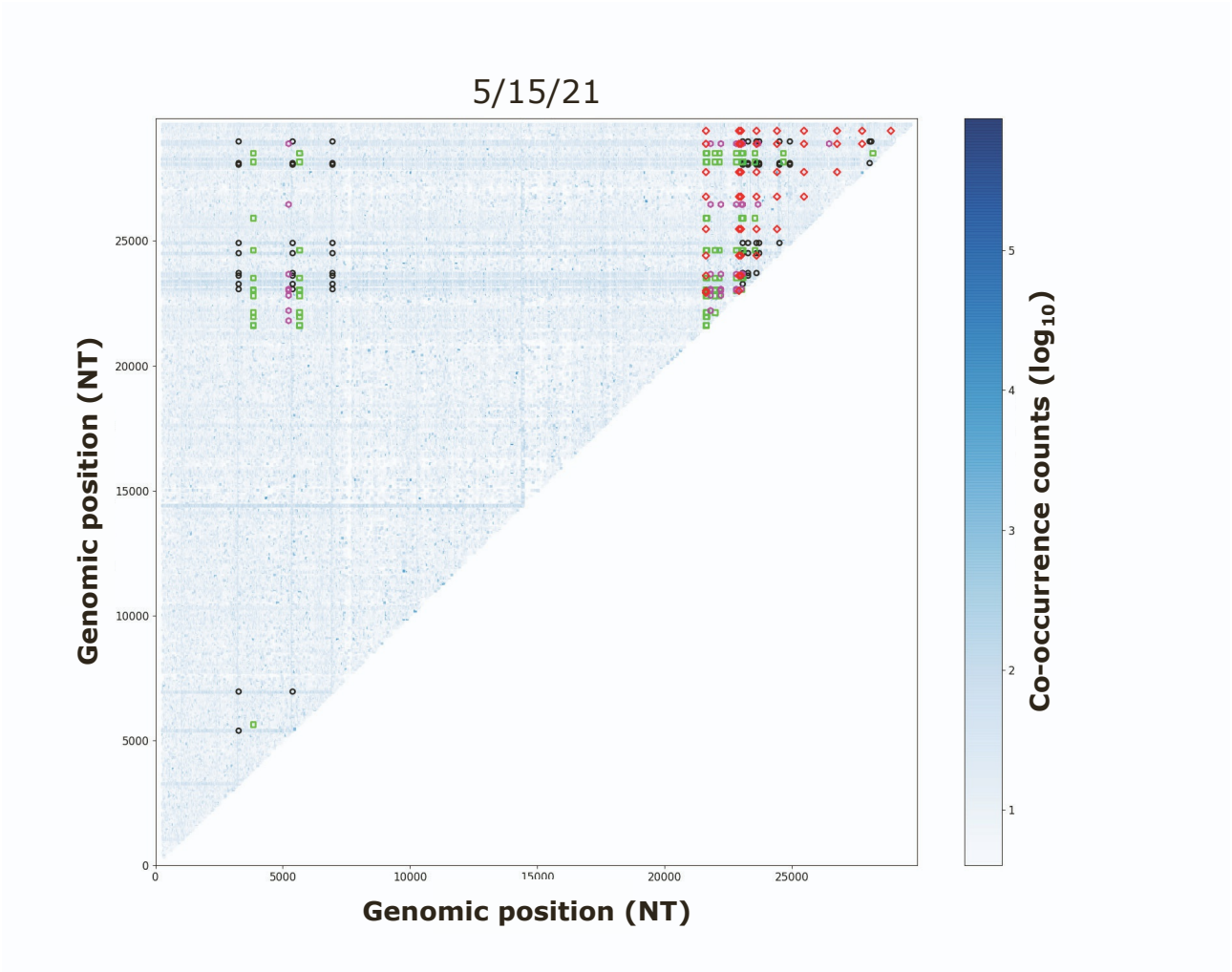

5/15/21
Genomic position (NT)
0
5000
10000
15000
20000
25000
Genomic position (NT)
Co-occurrence counts (log10)
1
2
3
4
5

Figure S21

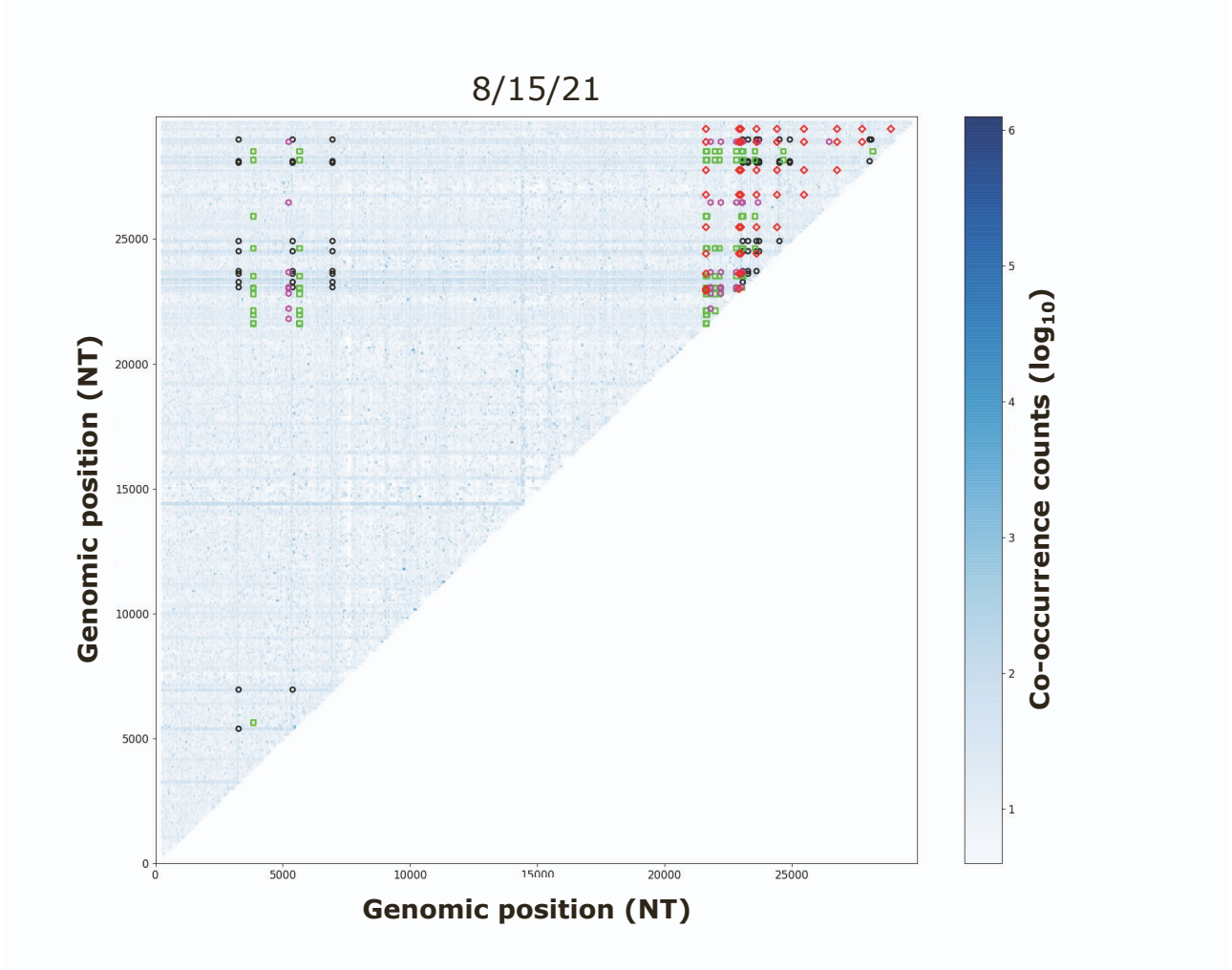

8/15/21
Genomic position (NT)
Genomic position (NT)
Co-occurrence counts (log10)
0
5000
10000
15000
20000
25000
1
2
3
4
5
6

Figure S22

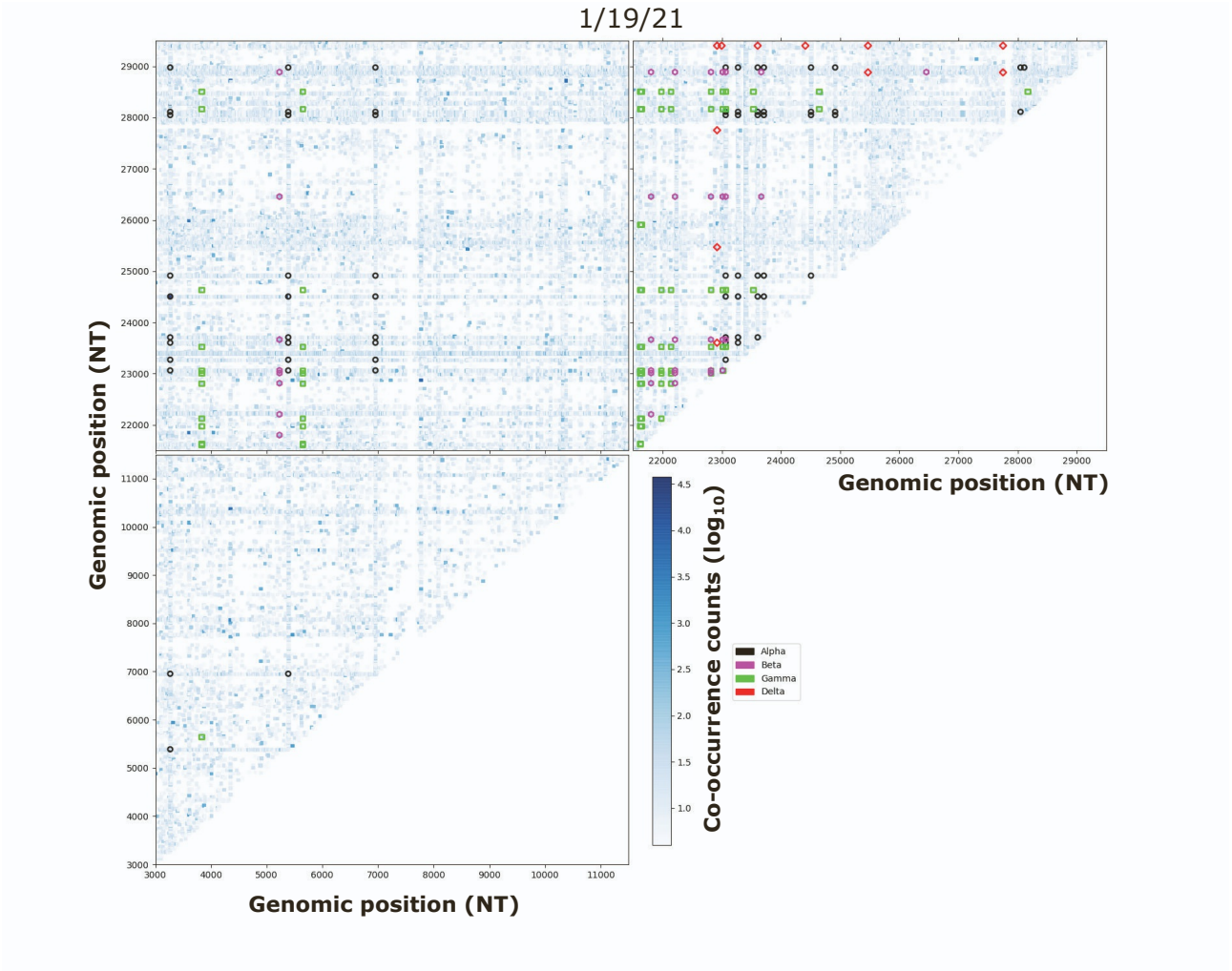

1/19/21
Genomic position (NT)
Genomic position (NT)
Co-occurrence counts (log10)
Alpha
Beta
Gamma
Delta

Figure S23

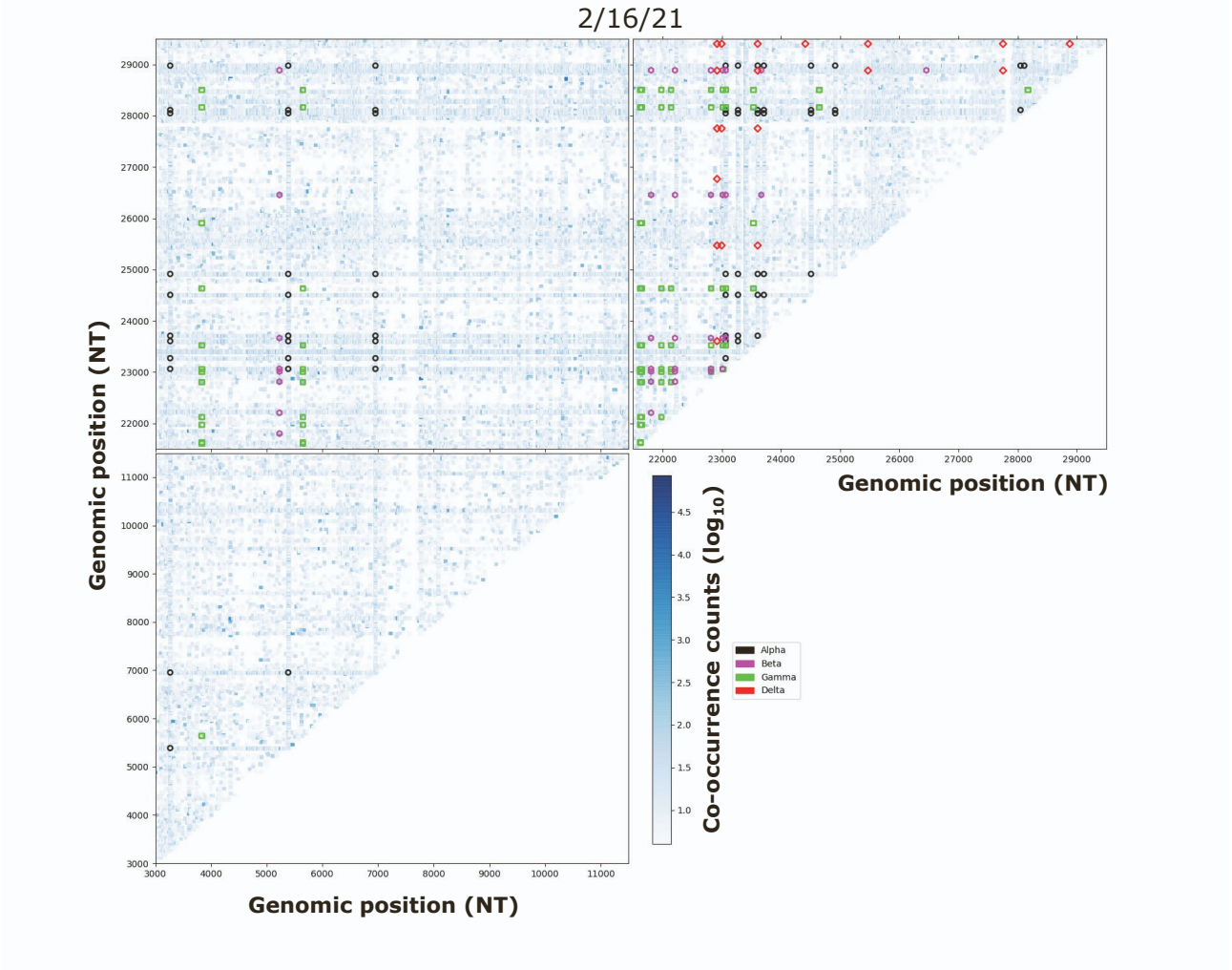

2/16/21
Genomic position (NT)
Genomic position (NT)
Co-occurrence counts (log10)
Alpha
Beta
Gamma
Delta

Figure S24

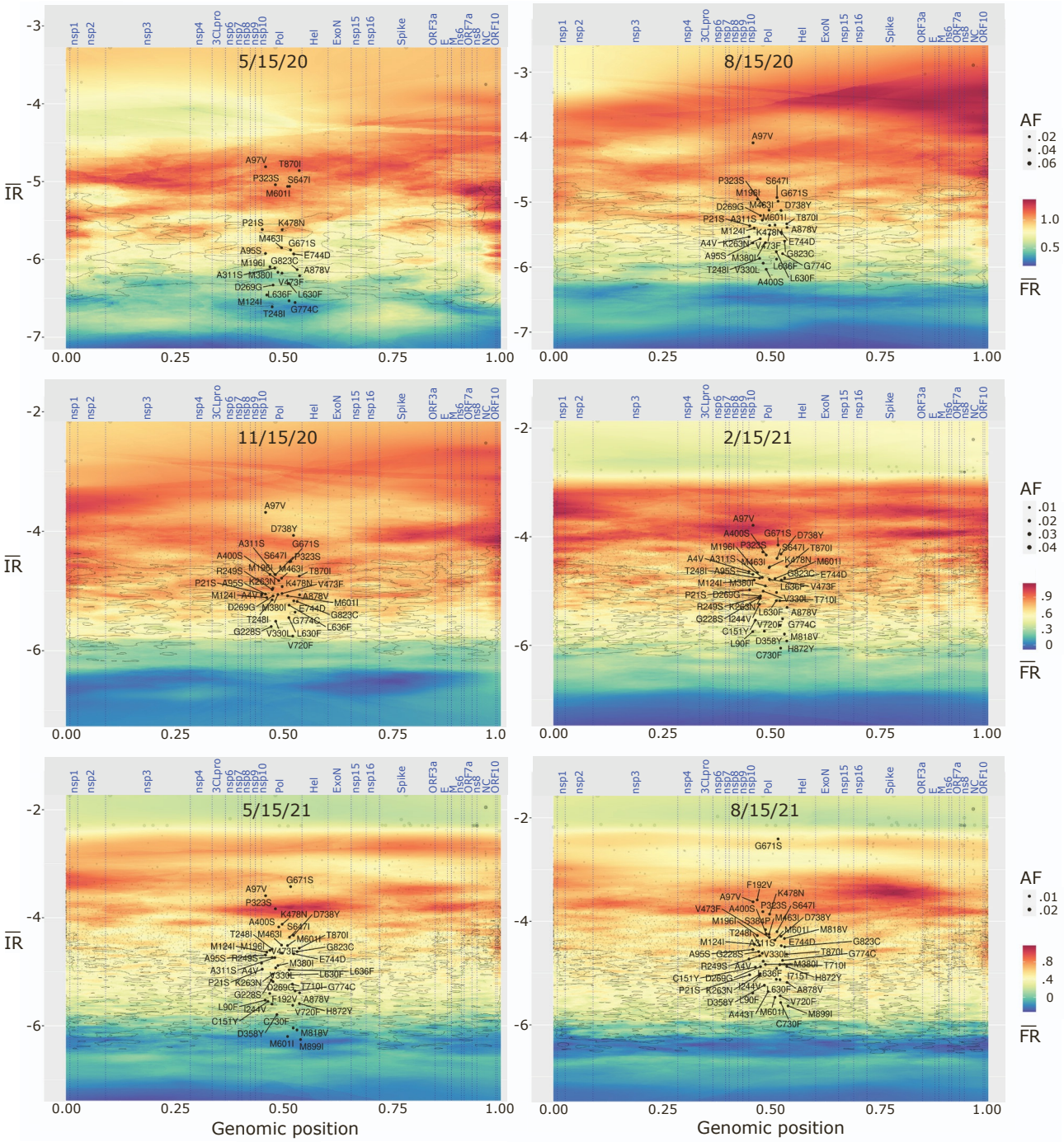

5/15/20
8/15/20
11/15/20
2/15/21
5/15/21
8/15/21
nsp1
nsp2
nsp3
nsp4
3CLpro
nsp6
nsp7
nsp8
nsp9
nsp10
Pol
Hel
ExoN
nsp15
nsp16
Spike
ORF3a
E
M
ns6
ORF7a
ns8
NC
ORF10
IR
FR
AF
Genomic position
0.00
0.25
0.50
0.75
1.00

Figure S25

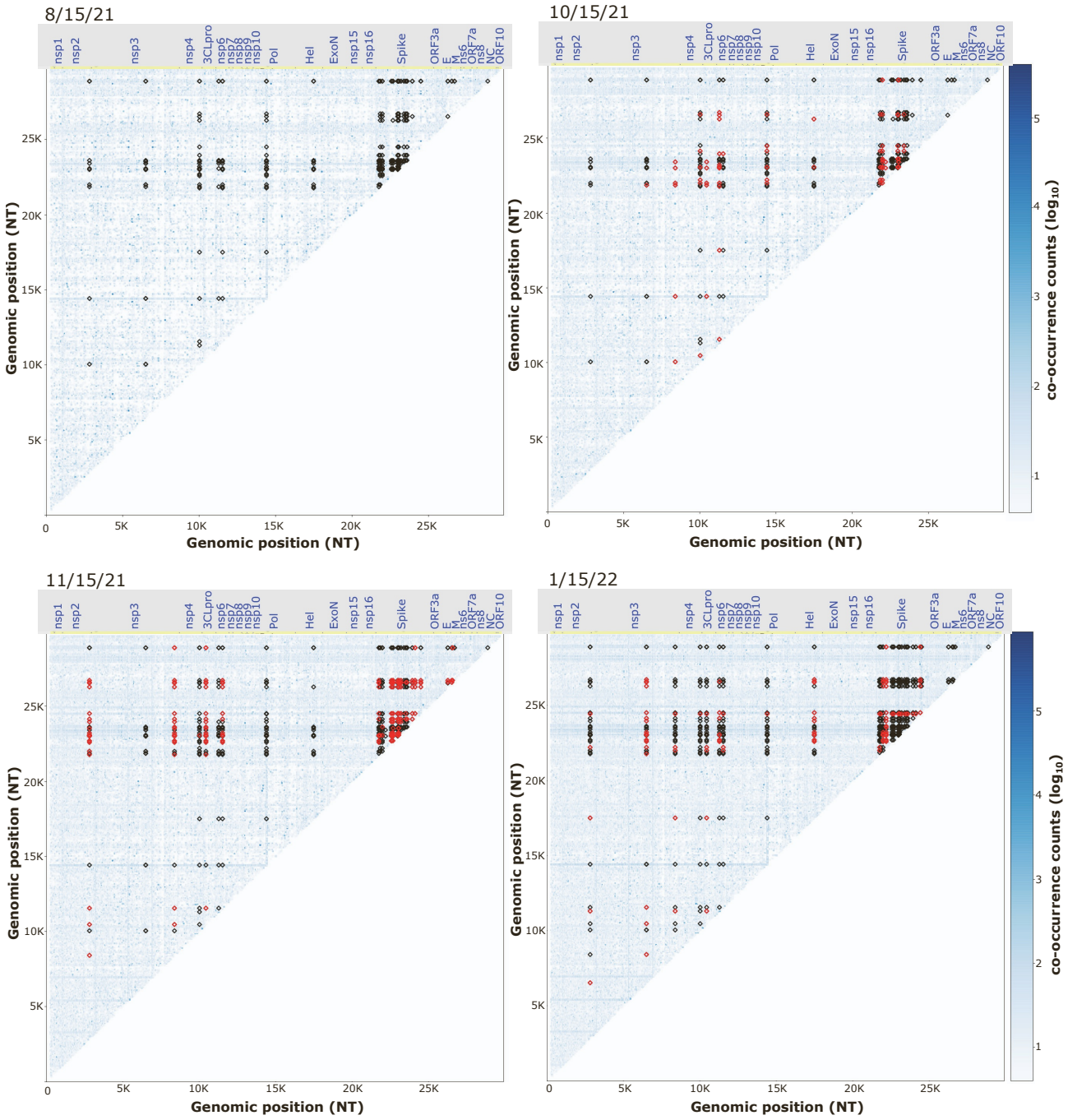

8/15/21
10/15/21
11/15/21
1/15/22
nsp1
nsp2
nsp3
nsp4
3CLpro
nsp6
nsp7
nsp8
nsp9
nsp10
Pol
Hel
ExoN
nsp15
nsp16
Spike
ORF3a
E
M
ns6
ORF7a
ns8
NC
ORF10
Genomic position (NT)
0
5K
10K
15K
20K
25K
co-occurrence counts (log10)
1
2
3
4
5

Figure S26

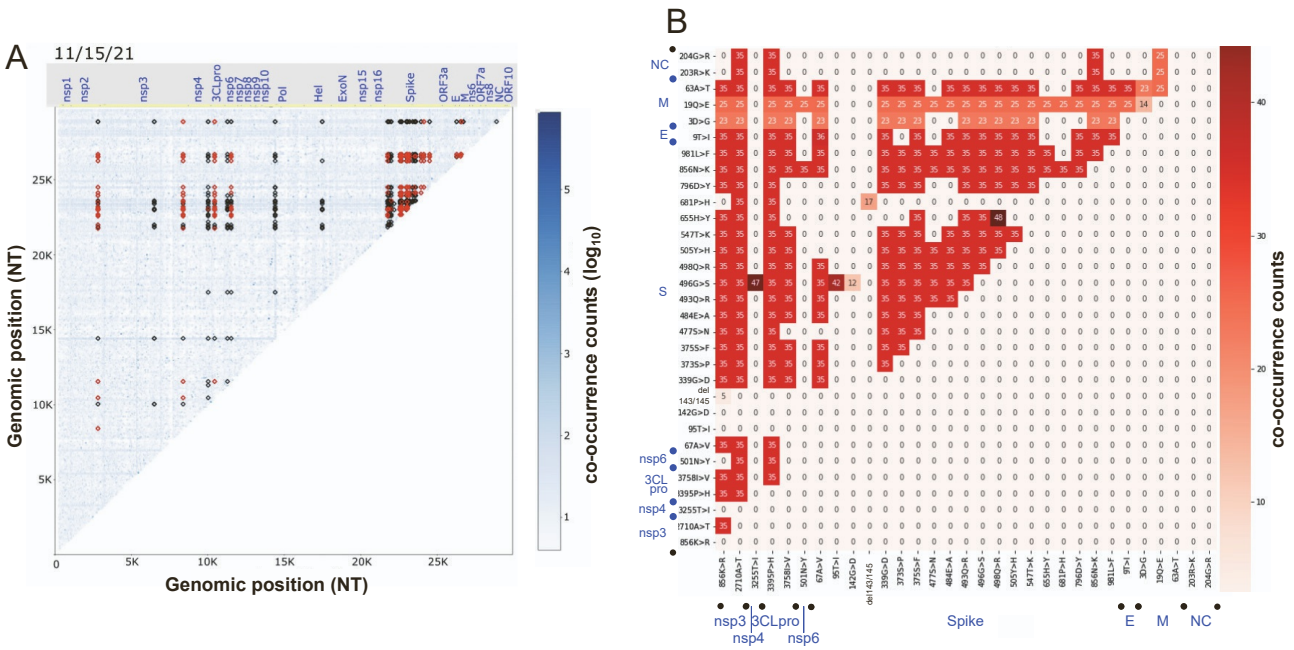

A
11/15/21
nsp1
nsp2
nsp3
nsp4
3CLpro
nsp6
nsp7
nsp8
nsp10
Pol
Hel
ExoN
nsp15
nsp16
Spike
ORF3a
E
M
ORF7a
ns8
N
ORF10
Genomic position (NT)
Genomic position (NT)
co-occurrence counts (log10)
B
NC
M
E
S
nsp6
3CL pro
nsp4
nsp3
co-occurrence counts
nsp3
nsp4
3CLpro
nsp6
Spike
E
M
NC

Figure S27


Mutation prevalence across lineages
Mutations with > 75% prevalence in at least one lineage.
Mutation prevalence in lineage
0
100%
not detected
Variant / Mutation of Concern
Variant / Mutation of Interest
ORF1a
BA.1 (492,244 seqs)
BA.2 (1,197,159)
BA.3 (1,385)
BA.4 (47,220)
BA.5 (48,731)
BA.2.75 (7,844)
ORF1b
S
ORF3a
E
M
ORF6
ORF7b
ORF8
N

SUPPLEMENTARY MATERIALS

Supplementary text

SCV analysis of the whole genome architecture (WGA) of SARS-CoV-2

To define the $\overline{IR}$ and $\overline{FR}$ composite variables for all variants contributing to the evolving spread and pathology, we first considered the number of cases reflecting spread in the population, fatalities, mutation counts and viral sequences on a per-country basis. Given the unprecedented availability of data in response to the world-wide cooperation in understanding onset and progression of the SARS-CoV-2 spread and pathology, as of 3/20/22 the dataset included 189,873 unique variants with 110,876 that had a prevalence in 3 or more countries spanning over twenty-three months of data recorded daily sequences (02/01/2020- 03/20/2022) across 152 countries. Allele frequency-weighted $\overline{IR}$ and $\overline{FR}$ (**Eq. 1, 2** in **Results**) were calculated for each variant through a weighted averaging across country-specific measurements where the weighting was determined by the number of variant counts found in a country (see **Methods**).

SARS-CoV-2 genome-wide Gaussian process (GP)

(Note: unless otherwise indicated, figure numbers refer to order of Results figures)

Traditional phylogenetic tree methods based on genome sequencing used in viral epidemiology [1] have led to considerable progress in mapping the progression of diversity contributing to viral spread [2]. However, they focus on single mutations at a time as branching points, not considering the entirety of the WGA as a collection of coordinated spatial units of functionality in the context of their genomic position that potentially contribute as a collective to viral spread and pathology. When considering $\overline{IR}$ and $\overline{FR}$, this view is indeed supported by lack of correlation between these variables (**Fig. S1**).

Methodologically, to generate allele phenotype landscapes, data are first pre-processed to obtain $\overline{IR}$ and $\overline{FR}$ for each mutation (**Fig. 1A**). Mutations are positioned in a 2D phenotype landscape defined by their genomic allelic positions (**Fig. 1D**, *x-axis*) and calculated (*y-axis*) (**Fig. 1A**, *x- and y-axis* and **Fig. 1A**, colored *z-axis*). Pairwise distances are computed for all mutations (**Fig. 1B**). In the second step of GP-based SCV modelling, a variogram is computed [3,4] quantifying the covarying relationships between the pairwise distances incorporating $\overline{IR}$ (**Fig. 1B**, x-axis) and their spatial variance with $\overline{FR}$ (**Fig. 1C**, y-axis) and used to build the allele phenotype landscape in the third step which describes known and unknown $\overline{IR}$ and $\overline{FR}$ SCV relationships for every sequence position in the SARS-CoV-2 genome [3,4].

Besides the application of GP-based SCV to human rare disease for protein fold architecture for discovery of therapeutics [3], addressing the impact of protein folding proteostasis pathways [5], and the role of epigenetics in proteome architecture[6], GP has also been used in navigating the fitness landscape in directed evolution experiments to guide protein engineering [7-9]. In general, because GP estimates model uncertainty, providing a rigorous and unbiased computational framework lacking the concern of overfitting to assess confidence in the function of a residue encoded by the genome in response to the environment [4].

Examples of allele phenotype landscapes

An example of an allele phenotype landscape for a specific date (9/15/20) is shown (**Fig. 1D**), where we trained our GP regression model on 11,120 mutations present in 3 or more countries at this time-point (Pearson r = 0.5, predicted vs. observed, leave-one-out cross-validation (LCV)). In **Fig. 1D** all *input* mutations used to generate the model are shown as circles with their frequency proportional to their abundance in the population as reported by $\overline{IR}$ (**Fig. 1D**, y-axis). Vertical dotted lines highlight boundaries of the indicated genes encoded in the SARS-CoV-2 genome. These boundaries, for example, illustrate a higher-than-average density of mutations at this time-point in Spike, ORF3a, ORF7a/ns8, and NC found at the 3'-end of the sequence,

whereas the polymerase nsp12 (**Fig. 1D**, Pol) has a lower-than-average density of mutations [10]. The color scale (**Fig. 1D**, z-axis) illustrates whether the predicted response of the genome results in lower (**Fig. 1D**, blue-green) or higher (**Fig. 1D**, increasing yellow-red) pathology predictions based on the surrounding residues, hidden information which we refer to as the variant dark matter that contributes substantially to allele phenotype landscape features. In **Fig. 1E** (right panel), we show the model now annotated with the signature mutations and uncertainty contours of the covariant relationships of the four VOCs driving early stages of spread and pathology including Alpha, Beta, Gamma, and Delta. Here, the large number *input* variant dark matter mutations driving the construction of the allele phenotype landscape are made transparent in order to focus on map structure relevant to VOCs, their defining mutations being those in at least 75% of a designated VOC [1] (see **Supplemental Methods**). The structure of the map shows the distribution of predicted $\overline{FR}$ where some of the VOCs are localized in, for example, high-$\overline{FR}$ areas (**Fig. 1D-E**, red color) while others are not (**Fig. 1D-E**, blue-green color) such as the VOCs found at the 3' end of the genome (**Fig. 1E**, zoom inset). The contours define relationships of VOCs to undesignated variant dark matter alleles that provide the supporting background to interpret covariance in evolving VOC designations. We have documented these evolving relationships for every day since the beginning of the pandemic (**Movie S1**).

In general, GP-based SCV using allele frequency weighted infection rate ($\overline{IR}$) and pathology (as reported by $\overline{FR}$) provides an SCV optimized interpolation of allele frequency variation in SARS-CoV-2 WGA that allows us to focus in an unbiased way on all the covariant relationships contributing to evolving viral fitness in the context of evolving host countermeasures. Because these SCV relationships cannot be captured by simple correlation alone (**Fig. S1**), the GP principled map reveals for the first time SCV relationships hidden within the viral WGA that could globally shape global disease dynamics.

The expanding timeframe of SCV relationships

The example above highlights only a specific time point (9/15/20) where mutation data was collected and $\overline{IR}$ / $\overline{FR}$ cumulatively computed across the world-wide population. To better dissect the differences in allele appearance over the time-span that contributes to the changing face of the spread and pathology, we generated comprehensive time-lapse GP-based SCV maps highlighting as a collective all current VOCs being monitored by the CDC for every day between 2/1/2020, to 3/20/2022 (**Movie S1**) as well as individual VOCs defining distinct Alpha, Beta, Gamma, and Delta lineages (**Movies S2-S5**).

Pairwise relationships dictating Alpha and Delta VOC evolution

To follow the collective evolution of VOC mutations contributing to the allele phenotype landscape over time, we considered all pairwise distances between these mutations (**Fig. 2B,3B**) and reported the *average of all distances*, together with its 95% confidence interval for Alpha VOC (**Fig. 2B1**) and Delta VOC (**Fig. 3B1**). Analogously, we reported the *average predicted* and 95% confidence interval for $\overline{FR}$ for Alpha VOC **(Fig. 3B1**) and Delta VOC (**Fig. 3B2**) for the same mutations as a quantitative way to assess SCV relationships on the z-axis $\overline{FR}$ prediction over time. These line plots provide a convenient way to track the overall relative movements of each of the VOC clusters of high interest (see **Fig. S2, S3** for Beta, Gamma VOCs) illustrating whether their relative positions contract or expand (**Fig 2. B1-2, D1-2**) as the pandemic progresses (**Fig. 2A, 3A**; panels 1-6).

At the first time point all VOC mutations are within the high confidence region (lower 10% total variance) marked by contour lines in the map (**Fig. 2A, B1**, Alpha VOC**; C, D1**, Delta VOC) (5/15/20 date). The high confidence region is where the majority of *input* mutations are located. Subsequently, the average distance in both the Alpha VOC (**Fig. 2B1**, 5/15/20->8/15/20 dates) and Delta VOC map (**Fig. 3B1,** 5/15/20->8/15/20 dates) increases along with the average SCV for $\overline{FR}$ for Alpha VOC (**Fig. 2B2**) and Delta VOC (**Fig. 3B2**). This is largely in response to the

detection of two new mutations (1001T>I (nsp3) and 501N>Y in Spike) distant from the main group, thus adding to the average distance and SCV defining $\overline{FR}$, and increasing their spread revealed by the broader confidence interval for both Alpha VOC and Delta VOC at the 8/15/20 date. These results foreshadow that these variants are potentially prominent driver mutations. By the next time point (**Fig. 2B1, 3B1**; 11/15/20), both average distance (**Fig. 2B1, 3B1**) and average SCV (**Fig. 2B2, 3B2**) are going down as the Alpha and Delta VOC mutations exit the central, dense region of the evolving allele phenotype landscapes and migrate to higher values of $\overline{IR}$ reflecting their increased abundance in the population. Notably, their pairwise distances have decreased and have also become more uniform as reported by their lower 95% confidence interval in average distribution (**Fig. 2B1, 3B1**). A similar pattern is observed for average SCV for $\overline{FR}$ (**Fig. 2B2, 3B2**), illustrating what can visually be seen on the allele phenotype landscapes (**Fig. 2A, 3A**) as compaction of the VOCs over time. Such compaction becomes particularly pronounced in the next time-points where average distance (**Fig. 2B1, 3B1**) and average SCV for $\overline{FR}$ (**Fig. 2B2, 3B2**) strongly decreases as Alpha VOC and Delta VOC signature mutations settle into high $\overline{IR}$ / low $\overline{FR}$ regions.

In general, pairwise relationships detected in the allele phenotype landscapes reveal VOC and variant dark matter coupled SCV relationships defined by $\overline{IR}$ and $\overline{FR}$, highlighting the potential of GP to capture the evolvability of the SARS-CoV-2 WGA as a matrix of spatial-temporal covariant relationships dictating host-pathogen fitness balance across the entire sequence.

Temporal co-occurrence (co-occurrence) patterns of emergent VOC mutations

Starting with a reference set of 16K missense mutations, we first counted cumulative pairwise co-occurrences among them with daily frequency between February 2020 and August 2021. For any given pair of mutations (A and B) their cumulative co-occurrence count (co-occurrence (A, B)) the co-occurrence is increased by one if A and B were found to co-occur on a viral isolate.

Extending this counting to all pairs of mutations considered, for each day we generated a matrix of cumulative co-occurrence counts and a heatmap visualization as a matrix. For example, on 3-15-2020 (**Fig. S13**), the most common mutations are D614G in Spike and P471L in the catalytic subunit of the RNA-dependent RNA polymerase (RdRp / nsp12). These mutations were established early and co-occur with almost every other SARS-CoV-2 mutation, thus appearing as horizontal/vertical bands on the matrix, which otherwise exhibits a rather sparse structure with occasional co-occurrence hotspots. co-occurrence sparsity is not uniform across the viral genome, although as the pandemic proceeds, co-occurrence becomes very noticeable across the entire genome (**Fig. S17-S21**).

To highlight co-occurrences driving emergence of VOCs Alpha, Beta, Gamma, and Delta, we first tracked cumulative co-occurrences among the defining mutations for every day starting 9/15/2021 up to 8/22/2021, and computed a daily average co-occurrence by averaging all cumulative co-occurrence values available daily for each of the VOCs. Average co-occurrence over time for the four VOCs is shown (**Fig. 4A**, left panel; zoom of Beta/Gamma/Delta VOC timeline plots shown in the right panel). All curves enter a quasi-exponential growth phase at different time-points along the timeline. Looking at more quantitative parameters to characterize such behavior, such as initial slope and maximal slope range, we see how the Alpha VOC emerges first with the highest magnitude and range; Beta, Gamma, and Delta VOCs have much smaller ranges that become evident at later times, starting around mid-December 2020 (**Fig. 4A**, right panel zoom). Gamma and Delta have similar, steeper slopes in their quasi-exponential initial growth phase, whereas Alpha and Beta slopes are much less pronounced.

Differences between VOCs are further evident in the patterning of their representative co-occurrence matrices (**Fig. 4B**). The first VOC to emerge, Alpha, has a uniform distribution of co-occurrences across the signature mutations of this viral strain – suggestive of an opportunistic break in its evolving host-pathogen encounters. This is not the case for later emergent VOCs.

For example, Gamma has one variant (501N>Y on Spike) with a co-occurrence 10x higher than average while the nsp3 variant 18L>F is 5X more than average co-occurrence and comparable to Spike's 484E>K. Moreover, 174G>C (Orf3a) shows zero or lower-than-average co-occurrence against most Gamma variants. Beta also shows significant co-occurrence hotspots: Spike (501N>Y, 701A>V) and NC (205T>Y) showing 2x, 2x and 1.5x co-occurrence increase in average co-occurrence, respectively. Given that 501N>Y is common across Alpha, Beta and Gamma VOCs, and 484E>K is shared between Beta and Gamma, higher co-occurrence values for these variants can be partly explained by their commonality. Finally, Delta, the latest VOC to appear in this time frame, shows a more varied pattern suggestive of a predatory behavior – possibly exploring at the atomic level for a breakthrough where four variants including Spike (681P>R), Orf3a (26S>L), NC (203R>M, 377D>Y) have between 2.5x and 3x higher than average co-occurrence counts.

Evolution of the co-occurrence search strategies of Alpha, Beta, Gamma, Delta VOCs

The co-occurrence matrix for Alpha on 9/22/20 (**Fig. S5B**) has few viral isolates with the full set of Alpha signature mutations and associated co-occurrences (**Fig. S5B**, black dots). These mutations belong to three genes: nsp3 (3), Spike (6) and NC (3) all occurring together. Overall, the 3' quadrant (**Fig. S5B**, quadrant II) is more populated, with a cluster between middle- and C-terminal Spike and another covering the beginning and end of NC. The 5' quadrant (**Fig. S5B**, quadrant I) has only three mutations in nsp3 leading to the three vertical co-occurrence bands in the cross-talk quadrant (**Fig. S5B**, quadrant III). Most co-occurrences (**Fig. S5**, nsp3 and Spike) are located in relatively empty areas of the global co-occurrence matrix (**Fig. S16**). At this early point in time, the remaining VOC are present with at least one co-occurrence. For example, for Beta, along the horizontal band at ~29K bp (**Fig. S5B**, quadrant III, intersecting at 5500 bp), and Delta at (25.7K, 29K bp) – near the intersection of two co-occurrence bands. Gamma is present with few co-occurrences in the region near the 5' end of Spike (not overlapping with

Alpha co-occurrence) and one nsp3 mutation in the crosstalk quadrant (**Fig. S5B**, quadrant III). Hence, co-occurrences from different VOCs are generally non-overlapping, with few exceptions. While mutations in Spike and NC have received considerable attention compared to those in nsp3, we now note at this early stage of VOC evolution, the first co-occurrences recorded for Beta and Gamma VOCs involve mutations in nsp3. This protein is a key component of the viral replication and transcription complex[11]. GP suggests it may play an unanticipated key role in the generation of new lineages of concern.

At 10/20/20 (**Fig. S5C**), we first observe a full set of co-occurrences for Beta. Notably, Beta co-occurrences explore additional regions in co-occurrence space compared to Alpha. These include a new region of the Spike gene that complements the region covered by Alpha co-occurrences as well as the end of the E gene (**Fig. S5C**, ~26,500 bp). Both cover relatively depleted areas of the co-occurrence matrix. Only three Beta co-occurrences are near Alpha co-occurrences. In particular, an intra-Spike co-occurrence (**Fig. S5C**, ~23,000 bp and ~23,500 bp) will be a co-occurrence hub for all four VOCs. Gamma added two co-occurrences in the first half of the Spike gene, and Delta added one co-occurrence in a rather uncharted area at the top of quadrant II (**Fig. S5C**). It is interesting to see how, even at this early stage in VOC evolution, initial co-occurrences for each of the VOC tend to sample different regions of the co-occurrence space with minimal overlap, thus favoring exploration versus exploitation which could be an unanticipated mechanism in driving host-pathogen relationships according to the Red Queen concept [12] (**Fig. 8** main text) in which you need to continually evolve to avoid extinction in multiple directions defined by the WGA captured by GP-based SCV relationships.

Two months later, on 12/22/2020 (**Fig. S5D**) we start observing a near-full set of co-occurrences for the Gamma VOC. We notice additional co-occurrences added in the middle to the 3' section of the Spike gene which weren't observed in Alpha or Beta co-occurrence in their debut. In the NC region, two Gamma mutations are placed in between Alpha and Beta VOC defining

mutations, generating co-occurrences non-overlapping with any of the previous VOC. Overall, Gamma co-occurrences appear to cover larger areas in co-occurrence space compared to Alpha and Beta. As previously pointed out (**Fig. 4,5**), co-occurrences for Gamma follow an exploratory phase before acquiring the full spectrum of co-occurrences available at saturation in the spread (**Fig. S5D**). The longer exploratory phase compared to Alpha and Beta VOC could be related to a more extensive field testing of the co-occurrence space reflecting the local environments of the hosts and/or challenges arising from previous infections now accruing in the population limiting its success as a takeover VOC.

Interestingly, even by 12/22/2020 (**Fig. S5D**) Delta VOC is incomplete and continues to sample co-occurrence space by adding co-occurrences at the 3' NC near the top of the 3' quadrant (**Fig. S5D**, quadrant II; 28,000 bp, 29,000 bp, 29,500 bp). The most interesting location in the 3' quadrant (II) (**Fig. S5D**) is a dense and narrow cluster at 23,000 bp and 23,500 bp where we see co-occurrences from all VOC lineages almost overlapping. This is quite unique since we have seen that co-occurrences in different VOCs tend to explore alternative solutions to achieve a higher fitness state. The co-occurrence in this cluster is largely between mutations in the RBD domain and mutations in the furin cleavage site of Spike including 501N>Y, 570A>D, 681P>H and 716T>I for Alpha, 484E>K, 501N>Y and 701A>V for Beta, 484E>K, 501N>Y and 655H>Y for Gamma and 452L>R, 478T>K and 681P>R for Delta. These results highlight the evolutionary advantage of keeping this specific co-occurrence solution while exploring broader areas of co-occurrence space.

By 03/02/2021 (**Fig. S5E**) we witness a full set of co-occurrences between Delta VOC defining mutations. The Delta VOC spends the longest time in exploratory phase of intermediate acquisition of co-occurrences, lasting several months, where only a few co-occurring variants are found on viral isolates before saturating all its co-occurrence space (**Fig. S5E**, inset). Delta is the only VOC of the four that has completely dropped signature mutations in the nsp3 regions,

thus focusing exclusively on the 3' region of the genome for gain of functional fitness (**Fig. S5E**, quadrant **II**). Strikingly, Delta co-occurrence pattern looks as if it is running a systematic grid-search [13] over the co-occurrence space by sampling the entire space at regularly spaced intervals, perhaps reflecting additional challenges in the environment reflected in increased vaccine prevalence and/or 'herd' immunity in the population.

While some Delta co-occurrences, especially on Spike, are recapitulating successful co-occurrence solutions previously discovered by the other VOCs (**Fig. S5E**) (including Beta and Gamma co-occurrence located between the 5' end to the middle of the Spike gene; the RBM-Furin cleavage co-occurrence cluster; Alpha and Gamma co-occurrence located between the middle to the 3' end of the Spike gene; and co-occurrences in NC gene at ~29,000 bp), other Delta co-occurrences are charting previously unexplored regions in a methodic, grid-like fashion illustrated by the co-occurrences at the 3'end of NC gene, the 3' end of ORF7a, the M gene and the ORF3a gene. Interestingly, the full grid-search result of Delta VOC co-occurrences is preceded in the long exploratory phase (**Fig. S22-23**, co-occurrence matrices for 1/19/21 and 2/16/21) by a L-shaped configuration of regularly spaced Delta co-occurrences with one horizontal sequence of co-occurrence at the 3' end of NC and one vertical at 23,000 bp. These results suggest the possibility that Delta is running a systematic preliminary exploration of the co-occurrence space prior to settling on the final stage for the full grid VOC configuration that marks the beginning of the wide-spread predatory behavior of the Delta variant.

Integrating VOC co-occurrences with GP residuals: an early warning anomaly detection (EWAD) system

For each VOC we report its average co-occurrence (**Fig. 6B, D, F, H**; co-occurrence plots) together with the mean of $\overline{FR}$ residuals for its signature mutations reported as the mean ± 95% confidence interval computed weekly along the selected time interval (**Fig. 6C, E, G, I**; EWAD plots). More detailed figures showing how these data are generated are available (**Fig. S6**, Alpha

and Beta; **Fig. S7**, Gamma and Delta) where for each of the above temporal points we report the corresponding allele phenotype landscape with VOC mutations highlighted (**Fig. S6-7**; panels 1-6, upper allele phenotype landscape and lower bar plots) showing the $\overline{FR}$ residuals for the signature mutations of the VOC whose average is reported in the EWAD plot. The baseline for this EWAD experiment was set by empirical randomization to 0 +-0.05 (**Fig. 6C, E, E, I**; dashed red line) where for a VOC including *n* signature mutations, we computed the mean $\overline{FR}$ residuals of thousands of random sets of *n* mutations. Hence, the interval near zero corresponds to a random (null) EWAD signal.

Evolution of the co-occurrence search strategies for Alpha, Beta, Gamma, Delta VOCs

To focus on the emergence of each VOC, we generated a genome-wide co-occurrence matrix spanning the encompassing 30,000 alleles (**Fig. S5A**). To focus on the most prominent emergent viral search strategies, we considered only the genomic ranges in which the majority of sets defining VOCs can be found – primarily in the 5' region (3000 to 11,000 bp) and the 3' region (22,000 to 29,000 bp), although their emergence is contingent on the co-occurrence data for the entire genome **(Fig. S5A**). These broad ranges were further delimited by three smaller subsections – namely, co-occurrences of VOC mutations concentrated within the 5' region (I), those within the 3' region (II) and the 'crosstalk' between the two (III) (**Fig. S5A**). We annotated this subset of the genome-wide co-occurrence matrix with different colors and shapes for the four VOCs (**Fig. S5A**; Alpha (black), Beta (magenta), Gamma (green), and Delta (red)).

We selected four representative time-points, each near the time at which a VOC discovers the full (or near-full) set of possible co-occurrences – i.e., when co-occurrence density nears a value of 1 **(Fig. S5B-E**, lower right panel, dashed red line marking the date of each quadrant). Each panel focuses on the regions of the global co-occurrence matrix that are relevant for the designated VOC signature mutations considered to be drivers of SARS-CoV-2 WGA in spread and pathology in the context of background variant dark matter. Temporal analysis of these

regional signature mutations reveals the evolving search strategies influencing the WGA structure of the Alpha, Beta, Gamma, and Delta VOCs (**Fig. S5A-D,** lower right panels). Strikingly, different VOCs tend to sample different regions of the co-occurrence space with minimal overlap. The few exceptions that occur can be found in the co-occurrence hubs such as the RBD-furin cleavage co-occurrence hub (**Fig. S5C-D**) that appears to be critical for all VOCs. The efficiency and coverage in exploring the space of possible mutational co-occurrences appears to be correlated with the time spent for intermediate acquisition of co-occurrences, with (Alpha, Beta) < Gamma << Delta. Delta in particular displays a systematic, grid-like exploration of mutational co-occurrence space in the crucial Spike-NC region. This is likely related to the capacity of this region to elevate Spike penetration of host cells and/or achieve immune superior evasion with NC region promoting more efficient viral assembly and load reflecting replication activity and packaging for release of mature virus particles. In addition, a GP-based analysis of the nsp12 polymerase subunit reveals that G671S may play an unanticipated role in creating a Delta VOC fitness configuration [10] (**Fig. S5A-D**, G671S; **Fig. S24**) [10].

Age-related EWAD for Beta

Analysis for Beta is shown in **Fig. 6D-E**. Beta was first detected on 12-18-2020 in South Africa. However, our co-occurrence screening shows evidence of an originating genome with all co-occurrences in place by 8/15/2020 (**Fig. 5**, jump to 1 in the co-occurrence density plot). The first two time-points for Beta (**Fig. 6D**, time-points #1 and 2) show a progression near the baseline with broad confidence interval (**Fig. 6E**, EWAD) indicative of a random distribution of $\overline{FR}$ residuals across positive and negative values (**Fig. S6**, Beta, panels 1-2). However, the EWAD signal gains a PAL1 already by 1-19-21. By 2-9-2021 (**Fig. 6D**, time-point #3) detections for Beta are still below 500, but the co-occurrence counts are switching toward faster growth rates; this is readily captured in the EWAD plot where mean $\overline{FR}$ residuals start increasing steadily (PAL2) with narrower spread of the confidence interval (**Fig. 6D**, EWAD). Looking at individual residuals,

time point #3 corresponds to a flipping point where almost all residuals are near zero (**Fig. S6**, Beta, panel 3 (bagplot)). By 3-2-21 (**Fig. 6D**, time-point #4) growth is steady with ~2K detections and the EWAD signal is still positive, reaching a plateau in PAL2, reflecting the fact that $\overline{FR}$ residuals that are moderately positive for nearly all Beta mutations (**Fig. S6**; panel 4, bar plot). The pattern flattens across the last two time-points (F**ig. 6D**, time-points #5 and #6), thereby de-escalating to PAL1 (**Fig. 6E**, EWAD). The EWAD signal for Beta is less early and less pronounced than the one seen with Alpha, possibly because the number of detections for this VOC is almost two orders of magnitude smaller than Alpha yet reveal a coordinated behavior. As seen with Alpha, this behavior is unlikely to be obtained with a random set of mutations of the same size (**Fig. S9-10**, empirical p-value $<10^{-3}$). These results reveal, as described for the Alpha (see **Results**), the performance of the EWAD as an anomaly detection system and the performance of different phases of the VOC affecting its role in the pandemic.

Age-related EWAD for Gamma and Delta

The late VOCs Gamma (**Fig. 6F-G**) and Delta (**Fig. 6H-I**) again reveal a coordinated, early change in mean $\overline{FR}$ residuals – thus triggering appropriate EWAD levels - before increased number of detections are measured through a change in slope of the average co-occurrence curve (**Fig. 6E**,**H**, respectively). However, for these two later VOCs both the mean (**Fig. 6G,I**) and individual $\overline{FR}$ residuals (**Fig. S7A,B**; panels 1-6) switch to negative (i.e., predicted $\overline{FR}$ values < observed $\overline{FR}$ values), instead of the positive values seen in the early Alpha/Beta VOCs (**Fig. S6A,B**, panels 1-6). One possibility could reflect the later time-points at which the EWAD signal (reflecting $\overline{FR}$ residuals) is observed for these VOCs. Since the later landscapes (**Fig. S7A, B**; panels 4-6) show less high-$\overline{FR}$ (red) clusters, most predicted $\overline{FR}s$ for VOC defining mutations have lower values to start with - hence the predicted < observed $\overline{FR}$ values result in negative residuals. For Gamma VOCs, first detected early in 01/01/2021, we see early sequences with full co-occurrences already in mid-December 2020 (**Fig. 5**, lower panel). Early tracking for

Gamma mutations shows a near-random distribution (near the zero baseline) with considerable range (**Fig. 6F,G**). The system however raises an PAL1 by time point #1 (**Fig. 6F, G**; time-point #1, 11/17/20), and a PAL2 by time point #2 (**Fig. 6F, G**; time-point 12/15/20). The steep decrease in mean $\overline{FR}$ residual continues down to 1/26/21 at time point #3) (**Fig. 6F**, 1/26/21) while the number of detections is still low (around 200), then plateaus. Again, as seen for Beta, the transition to flat EWAD signal corresponds to a switch in the distribution of the mean $\overline{FR}$ residuals (**Fig. S7A**; panel 3, bar plot) to mostly negative values. The flat pattern is then conserved at successive time-points (**Fig. 6F**, time-points #5 and #6) but keeping a PAL1 since both the mean signal and its confidence interval (**Fig. 6G**) are significantly different from the baseline following the trend observed in Alpha and Beta VOC (**Fig. 6C, E**).

Of importance, given its impact on human health world-wide, Delta was declared a VOC on 05/06/2021, but co-occurrence data shows gradual accumulation of co-occurring variants for a prolonged period of time preceding this date by several months (**Fig. 5,** lower panel; since October 2020) with a full set of co-occurrences recorded in mid-February 2021 (**Fig. 5**, lower panel). Interestingly, many Alerts (both PAL1 and PAL2) are raised by our EWAD system during this time (**Fig. 6I**), suggesting that EWAD was able to capture (and flag as concerning) this accumulation phase - otherwise hidden in plain sight. Importantly, EWAD was able to raise very early PAL2 warnings even with very low number of detections (i.e., time-points #1, #2, #4). By time point #4, corresponding to the switch point to high co-occurrence accumulation rate (**Fig. 6G**), the PAL2 signal has been seen for a month, and persists in alert mode (switching between PAL1 and PAL2) for the successive time-points. These results reveal, as described for the Alpha (see **Results**), the performance of the EWAD as an anomaly detection system and the performance of different phases of the VOC affecting its role pandemic.

Age-related EWAD for Omicron BA1: an emergent victor in the Red Queen competition for fitness

Omicron (B.1.1.529) was first reported to WHO from South Africa on 24 November 2021, and classified as VOC on 11/26/2021[14]. To simplify analysis, we focused on a set of 48 characteristic mutations for Omicron defined as nonsynonymous substitutions or deletions that occur in > 75% of sequences within the Omicron BA.1 lineage[2]. We monitored the characteristic mutations for Omicron BA.1 as cumulative count of co-occurrence mutations (**Fig. 7A**) and mean $\overline{FR}$ residuals for EWAD (**Fig. 7B**).

As seen before for previous VOCs, while co-occurrence count alone increases monotonically up to mid-December 2021 (**Fig. 7A**), $\overline{FR}$ residuals provide a striking EWAD signal for the ascent of Omicron BA.1 to a global VOC threat (**Fig. 7B**). On 10/21/21, the characteristic mutations for Omicron leaps in the EWAD diagram (**Fig. 7B**) where it enters PAL1 on the same week, and PAL2 on the following week. Notably, the alert is raised one month before Omicron is officially designated as a VOC by the WHO [14]. A similar sensitivity in early stages of Omicron diffusion is observed on the allele phenotype landscape (**Fig. S12A**), and the average distance of mutations on the landscape where in correspondence of the above dates, a sharp decrease in average distance is observed (**Fig. S12B**). Interestingly, the EWAD signal pattern for Omicron consists of an initial dip to negative $\overline{FR}$ residual values where predicted $\overline{FR}$ values are less than observed $\overline{FR}$ values since most mutations are in regions of high $\overline{FR}$ (see **Fig. S12A**, allele phenotype landscape at 11/15/21). This is followed by a sudden jump to positive $\overline{FR}$ residual values corresponding to most Omicron defining mutations on the allele phenotype landscape migrating to low $\overline{FR}$ regions.

In order to precisely map the co-occurring Omicron mutations associated with the anomalous EWAD signal observed around mid-October, we built genome-wide co-occurrence matrices for Omicron characteristic mutations at four time-points using the mid-October mark (**Fig. S25**). In particular, we highlighted the specific co-occurrence mutations added at each time point with red markers. **Fig. S26** shows the co-occurrence matrix for 11/15/21, highlighting the co-occurrence

mutations associated with the leap in $\overline{FR}$ residuals (**Fig. 7B**). It is interesting to note that these co-occurrence mutations involve a very narrow and specific set of SARS-CoV-2 proteins, namely Spike, the structural proteins E, M and NC, the two proteases (nsp3 and 3CLpro), and nsp6. For Spike, there are 18 missense mutation and one deletion including both mutations seen previously and novel ones specific to Omicron. The clusters of added co-occurrences involve mostly mutations around the ACE2 domain including E484A associated with immune evasion [10], Q498R increasing binding affinity to ACE2, and the furin cleavage site H655Y, P681H associated with increased transmissibility and the FP domain[15]. Both nsp3 and 3CLpro proteases have two mutations co-occurring with most of the Spike and envelope mutations including the N-terminal K856R and the C-terminal A2710T for nsp3 and P3395H and I3758V for 3CLPro. Structural envelope proteins are present with three, two and one mutation each (M, NC and, respectively). Notably, only E and M appear to co-occur extensively with Spike and protease mutations. NC has two ancestral mutations that are not new to Omicron but are linked to increased sub-genomic RNA expression including R203K and G204R that selectively co-occur only with protease mutations, M and one Spike mutation including N856K generating an alternative cleavage site for host SKI-1/S1P serine protease. Finally, there are two mutations presenting an exclusive co-occurrence pattern, T3255I on nsp4 that co-occurs only with the Spike/ACE2 G496S mutation, Spike T95I co-occurring with the same ACE2 mutation (G496S), and N501Y in nsp6 only co-occurring with protease mutations, Spike N856K and Q19E in M. The implications of these exclusive co-occurrence patterns are presently unclear and need further investigation.

Overall, the emerging pattern of co-occurrences for Omicron BA.1 combined with EWAD predictions suggest that this unique combination of variants provides the basis for a coordinated attack on the host to improve spread in the population with an associated risk of increased pathology where the emergent pattern is more complex and mutation-rich than previous VOCs. It shows clear regularities that include 1) the selective association of the cluster of Spike mutations with protease and envelope mutations, and 2) the presence of selective co-occurrence

sub-patterns involving specific mutations in different components contributing to the viral lifecycle.

New mutations in BA.4, BA.5, and BA.2.75

At time of submission, BA.1 and BA.2, the most recent dominant strains considered in this paper, have been largely supplanted by two new Omicron sub-lineages; BA.4 and BA.5. It has been suggested that these VOCs are more transmissible and have higher immune evasion capabilities [16], although new lineages of BA.2 – namely BA.2.75 – are gaining attention [17]. In **Fig. S27**, we can see a comparison of the prevalence of all mutations (with > 75% prevalence in at least one of the sub-lineages) in the six main Omicron sub-lineages. In these mutations, and their relations to the host-pathogen response, lies the answer to where the spread and pathology turns next, something EWAD can be an important tool in discovering. See **Discussion**.

Supplementary figure legends

Fig. S1

Linear correlations. A-B. Linear correlation composite variables $\overline{\mathrm{IR}}$ and $\overline{\mathrm{FR}}$ with genomic position of variants (**A, B**). **C-D.** Linear correlation of composite variables $\overline{\mathrm{IR}}$ and $\overline{\mathrm{FR}}$ with themselves (**C**, scaled $\overline{\mathrm{IR}}$ and $\overline{\mathrm{FR}}$; **D**, log-transformed $\overline{\mathrm{IR}}$ and $\overline{\mathrm{FR}}$). Summary: We focused on two important indicators of the pandemic to derive two composite variables, allele frequency (AF)-weighted $\overline{\mathrm{IR}}$ and $\overline{\mathrm{FR}}$, that can be estimated for each viral variant and serve as input for genome-wide SCV analysis (WGA) as output (see Results). A linear correlation analysis between the $\overline{\mathrm{IR}}$ and $\overline{\mathrm{FR}}$ variables and genomic position shows that there is no detectable correlation between allele frequency-weighted $\overline{\mathrm{IR}}$ and $\overline{\mathrm{FR}}$ with genomic position or between themselves. These results indicate that a basic linear correlation fails to provide a meaningful description of genome architecture features contributing to either spread or pathology driving the pandemic.

Fig. S2

Allele phenotype landscapes across the pandemic timeline for all VOCs. Select images annotated with Beta VOC signature mutations from **Movie S3. A.** Six tri-monthly time-points between May 2020 and August 2021 detailed as in Results (**Fig. 2A**). **B-C**. Average distance in the x,y-axis coordinates and average spatial variance of $\overline{\mathrm{FR}}$ (z-axis) coordinate between Beta VOC signature mutations for the six time-points shown in (**A**). Grey ribbons mark the 95% confidence interval.

Fig. S3

Allele phenotype landscapes across the pandemic timeline for all VOCs annotated with Gamma signature mutations from Movie S4. A. Six tri-monthly time-points between May 2020 and August 2021 detailed as in Results (**Fig. 4A**). **B-C.** Average distance in the x,y-axis coordinates and average spatial variance of $\overline{\mathrm{FR}}$ (z-axis) coordinate between Gamma VOC signature mutations for the six time-points shown in (**A**). Grey ribbons mark the 95% confidence interval. Summary: For Beta and Gamma VOCs we observe more of a mixed behavior over time compared to the trajectories of Alpha and Delta (see Results **Fig. 2,3**) with only few mutations (e.g., 501N->Y) moving in high $\overline{\mathrm{IR}}$, low $\overline{\mathrm{FR}}$ clusters at later time-points.

Fig. S4

Distribution of VOC frequencies. A. VOC relative frequencies between 9/27/20 and 9/13/21 (colored by VOC and normalized to 100% at each time point with 4000 sequences randomly sampled from the full distribution). Alpha VOC, Beta VOC, Gamma VOC, Delta VOC. Lower red bands on the right side of graph are Delta sub-lineages (B.1.617.2.x or AY.x). **B.** Cumulative sequence count over time for the four VOCs indicated in (**A**). Bars show the number of new sequences on GISAID over time, binned by epi-week. The line shows the cumulative number of sequences over time.

Fig. S5

Genome-wide co-occurrence matrix enriched in VOCS. A. A co-occurrence matrix was generated for all ~30,000 bp positions in the SARS-CoV-2 genome as described in **Methods**. Values on the axes are genomic coordinates (in nucleotide positions, NT) for the reference SARS-CoV-2 genome. Cumulative co-occurrence counts (in $\log_{10}$ scale) are reported at each location in blue gradient scale. Regions corresponding to viral genes are labeled on top (blue). Locations of VOC co-occurrences observed at this time point are highlighted as colored empty shapes: black circles for Alpha, magenta hexagons for Beta, green squares for Gamma and red diamonds for Delta. Highlighted in red are regions highly relevant for the VOCs: Region I (centered around nsp3), Region II (Spike to NC), Region III ("crosstalk": co-occurrences between genomic positions in Region I and II). Regions I-III include only the genomic ranges in which the sets of defining VOCs occur giving us a more focused insight on the emergence. The co-occurrence matrix shown is for Nov 15, 2020.Reported are four time-points, each corresponding to the first time a full set of co-occurrences is observed for each VOC (**B**, 9/22/20 (Alpha); **C**, 10/20/20 (Beta); **D**, 12/22/20 (Gamma); **ED**, 3/2/21 (Delta)). For each quadrant, we report three subsets of the full co-occurrence matrix that represent regions relevant for the VOCs: Region I (centered around nsp3), Region II (Spike to NC), Region III ("crosstalk": co-occurrences between I and II). The line plots in the lower right corner of each panel are the same as in (**Fig. 4,5**) that show cumulative co-occurrence counts and co-occurrence density over time for the four VOCs and are reported here for temporal reference, with the dashed red line over the line plots marking the date corresponding to each quadrant. ***Summary of results***: Temporal analysis of the global co-occurrence matrix yields many insights for the evolving search strategy developed by current VOCs. In particular, different VOCs tend to sample different regions of the co-occurrence space at different times and in unique temporal patterns with minimal overlap. The few exceptions that occur include the co-occurrence hubs like the RBM-furin cleavage co-occurrence hub (indicated by the star in the figure) are key co-occurring mutations that appear to be critical for all VOCs. The efficiency and coverage in exploring the space of all possible mutational co-occurrences

appear to be correlated with the time spent for intermediate acquisition of co-occurrence, with Alpha-Beta VOCs < Gamma VOC << Delta VOC where the Delta VOC displays a systematic, grid-like predatory exploration of mutational co-occurrences in the crucial Spike-NC region, possibly related to its overwhelming success driving an early stage of the SARS-CoV-2 spread and pathology.

Fig. S6

EWAD analysis for Alpha and Beta VOCs. A. A line plot reporting the average co-occurrence of Alpha VOC signature mutations is shown in the upper left corner, together with numbers (1-6) highlighting the time-points discussed in the Results and Supplementary Results. A zoom inset (**A**) shows the expanded time window between 10/10 and 11/14/20 (points 1-3). For each of the six time-points, a corresponding allele phenotype landscape with signature Alpha VOC mutations is reported together with a bar plot showing $\overline{\mathrm{FR}}$ residuals for the same mutations (1-6). $\overline{\mathrm{FR}}$ residuals for a specific mutation are the difference between the observed and predicted $\overline{\mathrm{FR}}$ values (see Results **Fig. 6**). **B.** Beta follows the same layout as in (**A**) with a line plot of average co-occurrence for Beta VOC signature mutations on top and relevant time-points (1-6) highlighted. For each of the time-points, an allele phenotype landscape annotated with Beta signature mutations and a bagplot of $\overline{\mathrm{FR}}$ residuals for the same mutations is shown. For both (**A**, Alpha VOC) and (**B**, Beta VOC), details for generation of the allele phenotype landscapes are the same as described in Results **Fig. 2,3**. ***Summary of results***: Co-occurrence analysis alone (**Fig. 8** legend) is not able to inform whether a certain combination of mutations will eventually result in a VOC. We performed a joint analysis of the indicated Alpha and Beta VOCs (see **Fig. S7** for Gamma and Delta VOCs) by tracking their mutations simultaneously through allele phenotype landscapes over time to assess whether GP-based SCV relationships defined in the context of co-occurrences (see Results section 'Integrating VOC co-occurrences with GP residuals: an early warning anomaly detection (EWAD) system') could be used as an early

warning anomaly detection (EWAD) monitoring system for variant change surveillance. The 6 time-points selected cover the flat, early, and sustained co-occurrence growth phases for each VOC. For Alpha VOC, even with a low number of detections, mean $\overline{FR}$ residuals appear to respond quite sensitively to changes in the growth regime, with the collective signal defined by the mean $\overline{FR}$ residuals changing significantly during the early phase of the transition (inset, upper left corner, numbers 1, 2), reaching a maximum (inset, upper left corner, 3) then attenuating when the variation enters a steady state. The pattern observed for Alpha VOC is specific and statistically significant. A similar statistically significant EWAD for the Beta VOC is observed (**B**), although less early and less pronounced than with Alpha VOC. Residuals reveal a coordinated behavior which does not occur by chance based on empirical randomization.

Fig. S7

EWAD analysis for Gamma and Delta VOCs. A, B. A line plot reporting the average co-occurrence of signature mutations is shown in the inset in the upper left corner (**A**, Gamma VOC; **B**, Delta VOC), together with numbers (1-6) highlighting the time-points that correspond to the flat, early, and sustained co-occurrence growth phases discussed in the Results and Supplementary Results. For each of the 6 time-points, corresponding allele phenotype landscapes with signature VOC mutations are reported, together with a bar plot showing $\overline{FR}$ residuals for the same mutations. For both (**A**, Gamma VOC) and (**B**, Delta VOC), details for generation of the allele phenotype landscapes are the same as described in Results **Fig. 2,3**. Figure layout and details are the same as for Results **Fig. 6**. ***Summary of results***: The later Gamma and Delta VOCs show a coordinated, early change in signs and magnitude of mean $\overline{FR}$ residuals corresponding to increased number of detections as measured through a change in slope of the average co-occurrence curve (inset, upper left corner (**A, B**). For these 2 later VOCs, the mean $\overline{FR}$ residuals switch to negative where predicted < observed $\overline{FR}$ values, instead of positive values seen in the early Alpha and Beta VOCs. One possibility could reflect the later

time-points at which the EWAD signal reflecting mean $\overline{\mathrm{FR}}$ residuals is detected for these VOCs. Since the later allele phenotype landscapes show fewer high-$\overline{\mathrm{FR}}$ (red) clusters most predicted $\overline{\mathrm{FR}}$s for VOC mutations start with lower values giving rise to the predicted < observed $\overline{\mathrm{FR}}$ values and negative residuals. The combined analysis using GP-based SCV relationships residuals in the context of co-occurrences is an effective EWAD for each the VOCs. Such a system could be implemented into real- time monitoring applications for variant surveillance.

Fig. S8

Comparison of mutation sets for Alpha, Beta, Gamma, and Delta VOCs to five randomly selected sets of mutations. (**A**) We compared each of 5 randomly selected groups of mutations (labelled 'RND[a-e]' on the horizontal axis) against the 4 VOCs (labelled 'VOC-[a-d]' for Alpha, Beta, Gamma, Delta respectively along the horizontal axis). The vertical axis counts how often the FR residues for each mutation set change significantly (>0.1 absolute change in residuals, in line with the PAL definition heuristics) from time point to time point, over the time period examined. This total is calculated for each mutation within the five groups, and their distribution plotted in a boxplot. Statistically, the VOC FR residuals (VOC-[a-d]) change more often than the randomly chosen mutation sets (RND-[a-e]) – with a Welch two sample t-test showing that their median number of significant residual changes per mutation across the time period are significantly higher ($p = 0.00282$) with median VOCs number of changes being 6.125 in contrast to median RNDs being 2.600. These results indicate that the VOC mutations display strong signals as they emerge over time, with consistently higher shifts in the residuals – a key part of the PALs defined here. A random set of mutations does not display the same behavior, with the residuals changing less severely over time. (**B-C**) To assess how the mutation sets differ *at each time point* across mutations, we examined the variance in FR residuals across the mutations within each mutation set. Shown are two plots quantitating time points along the x-axis and FR-residual variance on the y-axis for the four VOCs (**B**) and for the five random mutation sets (**C**).

We can see a clear difference in the qualitative pattern of point distribution where the variance for the VOCs decreases over time (**B**) as the variants emerge, in striking contrast to a lack of a distinguishable pattern for the randomly chosen mutation sets (**C**) with the exception of RND-a, where a high variance is produced initially by mutants that do not appear in the data until later in the time points, thus a lower n. The differences clearly indicate that the residuals of a VOC are markedly different in their behavior from randomly chosen mutations – demonstrating the utility of FR residuals as the foundation for the PALs defined here.

Fig. S9

Bar plots of mean $\overline{\mathrm{FR}}$ residuals used in EWAD analyses (see Results Fig. 6,7). Shown are representative sets of 15 randomly chosen mutations that are similar to the size of signature mutations for Alpha and Gamma VOC. Seven tri-weekly time-points between 11/17/20 and 5/18/21 are reported. Summary: Unlike the mean $\overline{\mathrm{FR}}$ residuals corresponding to VOC signature mutations that show a coordinated, early warning anomaly detection signal of potential use for variant surveillance, residuals here show a rather random behavior over time, not correlated with specific temporal events. The full analysis was repeated with thousands of random sets of mutations whose set size is the same as the VOC signature mutations considered here (10-15 mutations) to determine if the EWAD results were statistically significant and unique to VOC, or if they were just a map-wide property that could be observed for any random set of mutations. Remarkably, a great majority of randomly selected mutations did not present any coordinated behavior at the level of $\overline{\mathrm{FR}}$ residuals across the selected time window (empirical p-values < 10^{-3} for all VOCs), providing evidence that the observed EWAD patterns are specifically associated with distinctive VOC mutations. These results have important implications for the future prediction of unknown VOCs prior to their emergence by focusing on emerging variants in the variant dark matter comprising high spread with either low or high pathology.

Fig. S10

Bar plots of mean $\overline{FR}$ residuals used in EWAD analyses (see Results Fig. 6,7). Shown are representative sets of 10 randomly chosen mutations that are similar to the size of signature mutations for Beta and Delta VOC. Seven tri-weekly time-points between 11/17/20 and 5/18/21 are reported. Summary: Unlike the mean $\overline{FR}$ residuals corresponding to VOC signature mutations that show a coordinated, early warning anomaly detection signal of potential use for variant surveillance, residuals here show a rather random behavior over time, not correlated with specific temporal events. The full analysis was repeated with thousands of random sets of mutations whose set size is the same as the VOC signature mutations considered here (10-15 mutations) to determine if the EWAD results were statistically significant and unique to VOC, or if they were just a map-wide property that could be observed for any random set of mutations. Remarkably, a great majority of randomly selected mutations did not present any coordinated behavior at the level of $\overline{FR}$ residuals across the selected time window (empirical p-values < 10^{-3} for all VOCs), providing evidence that the observed EWAD patterns are specifically associated with distinctive VOC mutations. These results have important implications for the future prediction of unknown VOCs prior to their emergence by focusing on emerging variants in the variant dark matter comprising high spread with either low or high pathology.

Fig. S11

EWAD analysis of ablated Alpha, Beta, Gamma, and Delta VOCs. Legend and details for this figure are identical to the EWAD profile panels (C,E,G,I) in **Fig. 6**. In each case, the mutations for the VOC in question are ablated or removed from the data set, and then the SCV analysis performed, residuals calculated, plotted, and compared to the results generated in the presence of the VOC mutations (**Fig. 6**). For the ablation study of Alpha, all Alpha VOC were omitted from the training set; for Beta, all Alpha and Beta VOC were removed; for Gamma, the VOC corresponding to Alpha, Beta and Gamma were omitted; finally, for Delta, Alpha, beta, Gamma, and Delta were removed from the training set. As can be seen by comparison with **Fig. 6**, the

results are near-identical. This shows that the inclusion of a particular VOC does not affect the predictions and that the model is capable of making valuable early warning predictions for VOCs. Indeed, the strength of the EWAD model lies in its use of *all* of the data at hand to predict the single – the trajectory of the emergent VOC.

Fig. S12

Time-lapse snapshots of viral genome allele phenotype landscapes between 8/15/21 and 1/15/22, annotated with Omicron BA.1 signature mutations. A. 6 bi-monthly time-points between August 2021 and January 2022 are shown. $\overline{\mathrm{IR}}$ and $\overline{\mathrm{FR}}$ are log-transformed, genomic position is scaled to a 0-1 interval. Vertical dotted blue lines are boundaries between SARS CoV-2 proteins annotated on the top axis. Input variants are in shaded color with dot sizes proportional to allele frequency of the mutations. Contour lines are drawn at the 10% and 25% of global variance estimated for model predictions. **B.** (left panel). Average distance between Omicron signature mutations where distance in the x,y-axis coordinates is defined as in Results **Fig. 1B** for the 6 time-points shown in (**A**). The grey ribbon marks the 95% confidence interval. **B**. (right panel). Average spatial variance of $\overline{\mathrm{FR}}$ (z-axis) coordinate between Omicron BA.1 signature mutations for the same time-points. Grey ribbon marks the 95% confidence interval.

Fig. S13

Genome-wide co-occurrence (co-occurrence) matrices across the entire SARS-CoV-2 pandemic timeline, reported on 3/15/20. For each co-occurrence matrix heatmap, values on the axes are genomic coordinates (in bps) for the reference Wuhan SARS-CoV-2 genome. Cumulative co-occurrence counts (in $\log_{10}$ scale) are reported at each location in blue gradient scale. Regions corresponding to viral genes are labeled on top (blue). Locations of VOC co-occurrences observed at this time point are highlighted as colored empty shapes (Alpha, black circles; Beta, magenta hexagons; Gamma, green squares; Delta, red diamonds). ***Summary of results***: The representative genome-wide co-occurrence matrices heatmaps show the mutation

co-occurrence landscape during the early, mid-, and late stages of the pandemic. For example, on 3-15-202, the most common mutations are allele changes encoding D614G in Spike and P471L in the catalytic subunit of the RNA-dependent RNA polymerase (RdRp/nsp12). These mutations were established early and co-occur with almost every other SARS-CoV-2 mutation, thus appearing as horizontal/vertical bands on the heatmap, which otherwise exhibits a rather sparse structure with occasional co-occurrence hotspots. Co-occurrence sparsity is not uniform across the viral genome, although as the pandemic proceeds, co-occurrence becomes very noticeable across the entire genome.

Fig. S14.

Genome-wide co-occurrence (co-occurrence) matrices across the entire SARS-CoV-2 pandemic timeline, reported on 5/15/20. For each co-occurrence matrix heatmap, values on the axes are genomic coordinates (in bps) for the reference Wuhan SARS-CoV-2 genome. Cumulative co-occurrence counts (in $\log_{10}$ scale) are reported at each location in blue gradient scale. Regions corresponding to viral genes are labeled on top (blue). Locations of VOC co-occurrences observed at this time point are highlighted as colored empty shapes (Alpha, black circles; Beta, magenta hexagons; Gamma, green squares; Delta, red diamonds). ***Summary of results***: The representative genome-wide co-occurrence matrices heatmaps show the mutation co-occurrence landscape during the early, mid-, and late stages of the pandemic. For example, on 3-15-2020 (**Fig. S13**), the most common mutations are allele changes encoding D614G in Spike and P471L in the catalytic subunit of the RNA-dependent RNA polymerase (RdRp/nsp12). These mutations were established early and co-occur with almost every other SARS-CoV-2 mutation, thus appearing as horizontal/vertical bands on the heatmap, which otherwise exhibits a rather sparse structure with occasional co-occurrence hotspots. Co-occurrence sparsity is not uniform across the viral genome, although as the pandemic proceeds, co-occurrence becomes very noticeable across the entire genome.

Fig. S15

Genome-wide co-occurrence (co-occurrence) matrices across the entire SARS-CoV-2 pandemic timeline, reported on 7/15/20. For each co-occurrence matrix heatmap, values on the axes are genomic coordinates (in bps) for the reference Wuhan SARS-CoV-2 genome. Cumulative co-occurrence counts (in $\log_{10}$ scale) are reported at each location in blue gradient scale. Regions corresponding to viral genes are labeled on top (blue). Locations of VOC co-occurrences observed at this time point are highlighted as colored empty shapes (Alpha, black circles; Beta, magenta hexagons; Gamma, green squares; Delta, red diamonds). ***Summary of results***: The representative genome-wide co-occurrence matrices heatmaps show the mutation co-occurrence landscape during the early, mid-, and late stages of the pandemic. For example, on 3-15-2020 (**Fig. S13**), the most common mutations are allele changes encoding D614G in Spike and P471L in the catalytic subunit of the RNA-dependent RNA polymerase (RdRp/nsp12). These mutations were established early and co-occur with almost every other SARS-CoV-2 mutation, thus appearing as horizontal/vertical bands on the heatmap, which otherwise exhibits a rather sparse structure with occasional co-occurrence hotspots. Co-occurrence sparsity is not uniform across the viral genome, although as the pandemic proceeds, co-occurrence becomes very noticeable across the entire genome.

Fig. S16

Genome-wide co-occurrence (co-occurrence) matrices across the entire SARS-CoV-2 pandemic timeline, reported on 9/15/20. For each co-occurrence matrix heatmap, values on the axes are genomic coordinates (in bps) for the reference Wuhan SARS-CoV-2 genome. Cumulative co-occurrence counts (in $\log_{10}$ scale) are reported at each location in blue gradient scale. Regions corresponding to viral genes are labeled on top (blue). Locations of VOC co-occurrences observed at this time point are highlighted as colored empty shapes (Alpha, black circles; Beta, magenta hexagons; Gamma, green squares; Delta, red diamonds). ***Summary of***

results: The representative genome-wide co-occurrence matrices heatmaps show the mutation co-occurrence landscape during the early, mid-, and late stages of the pandemic. For example, on 3-15-2020 (**Fig. S3**), the most common mutations are allele changes encoding D614G in Spike and P471L in the catalytic subunit of the RNA-dependent RNA polymerase (RdRp/nsp12). These mutations were established early and co-occur with almost every other SARS-CoV-2 mutation, thus appearing as horizontal/vertical bands on the heatmap, which otherwise exhibits a rather sparse structure with occasional co-occurrence hotspots. Co-occurrence sparsity is not uniform across the viral genome, although as the pandemic proceeds, co-occurrence becomes very noticeable across the entire genome.

Fig. S17

Genome-wide co-occurrence (co-occurrence) matrices across the entire SARS-CoV-2 pandemic timeline, reported on 11/15/20. For each co-occurrence matrix heatmap, values on the axes are genomic coordinates (in bps) for the reference Wuhan SARS-CoV-2 genome. Cumulative co-occurrence counts (in $\log_{10}$ scale) are reported at each location in blue gradient scale. Regions corresponding to viral genes are labeled on top (blue). Locations of VOC co-occurrences observed at this time point are highlighted as colored empty shapes (Alpha, black circles; Beta, magenta hexagons; Gamma, green squares; Delta, red diamonds). ***Summary of results***: The representative genome-wide co-occurrence matrices heatmaps show the mutation co-occurrence landscape during the early, mid-, and late stages of the pandemic. For example, on 3-15-2020 (**Fig. S13**), the most common mutations are allele changes encoding D614G in Spike and P471L in the catalytic subunit of the RNA-dependent RNA polymerase (RdRp/nsp12). These mutations were established early and co-occur with almost every other SARS-CoV-2 mutation, thus appearing as horizontal/vertical bands on the heatmap, which otherwise exhibits a rather sparse structure with occasional co-occurrence hotspots. Co-occurrence sparsity is not uniform across the viral genome, although as the pandemic proceeds, co-occurrence becomes very noticeable across the entire genome.

Fig. S18

Genome-wide co-occurrence (co-occurrence) matrices across the entire SARS-CoV-2 pandemic timeline, reported on 1/15/21. For each co-occurrence matrix heatmap, values on the axes are genomic coordinates (in bps) for the reference Wuhan SARS-CoV-2 genome. Cumulative co-occurrence counts (in $\log_{10}$ scale) are reported at each location in blue gradient scale. Regions corresponding to viral genes are labeled on top (blue). Locations of VOC co-occurrences observed at this time point are highlighted as colored empty shapes (Alpha, black circles; Beta, magenta hexagons; Gamma, green squares; Delta, red diamonds). ***Summary of results***: The representative genome-wide co-occurrence matrices heatmaps show the mutation co-occurrence landscape during the early, mid-, and late stages of the pandemic. For example, on 3-15-2020 (**Fig. S13**), the most common mutations are allele changes encoding D614G in Spike and P471L in the catalytic subunit of the RNA-dependent RNA polymerase (RdRp/nsp12). These mutations were established early and co-occur with almost every other SARS-CoV-2 mutation, thus appearing as horizontal/vertical bands on the heatmap, which otherwise exhibits a rather sparse structure with occasional co-occurrence hotspots. Co-occurrence sparsity is not uniform across the viral genome, although as the pandemic proceeds, co-occurrence becomes very noticeable across the entire genome.

Fig. S19

Genome-wide co-occurrence (co-occurrence) matrices across the entire SARS-CoV-2 pandemic timeline, reported on 3/15/21. For each co-occurrence matrix heatmap, values on the axes are genomic coordinates (in bps) for the reference Wuhan SARS-CoV-2 genome. Cumulative co-occurrence counts (in $\log_{10}$ scale) are reported at each location in blue gradient scale. Regions corresponding to viral genes are labeled on top (blue). Locations of VOC co-occurrences observed at this time point are highlighted as colored empty shapes (Alpha, black circles; Beta, magenta hexagons; Gamma, green squares; Delta, red diamonds). ***Summary of***

results: The representative genome-wide co-occurrence matrices heatmaps show the mutation co-occurrence landscape during the early, mid-, and late stages of the pandemic. For example, on 3-15-2020 (**Fig. S13**), the most common mutations are allele changes encoding D614G in Spike and P471L in the catalytic subunit of the RNA-dependent RNA polymerase (RdRp/nsp12). These mutations were established early and co-occur with almost every other SARS-CoV-2 mutation, thus appearing as horizontal/vertical bands on the heatmap, which otherwise exhibits a rather sparse structure with occasional co-occurrence hotspots. Co-occurrence sparsity is not uniform across the viral genome, although as the pandemic proceeds, co-occurrence becomes very noticeable across the entire genome.

Fig. S20

Genome-wide co-occurrence (co-occurrence) matrices across the entire SARS-CoV-2 pandemic timeline, reported on 5/15/21. For each co-occurrence matrix heatmap, values on the axes are genomic coordinates (in bps) for the reference Wuhan SARS-CoV-2 genome. Cumulative co-occurrence counts (in $\log_{10}$ scale) are reported at each location in blue gradient scale. Regions corresponding to viral genes are labeled on top (blue). Locations of VOC co-occurrences observed at this time point are highlighted as colored empty shapes (Alpha, black circles; Beta, magenta hexagons; Gamma, green squares; Delta, red diamonds). ***Summary of results***: The representative genome-wide co-occurrence matrices heatmaps show the mutation co-occurrence landscape during the early, mid-, and late stages of the pandemic. For example, on 3-15-2020 (**Fig. S13**), the most common mutations are allele changes encoding D614G in Spike and P471L in the catalytic subunit of the RNA-dependent RNA polymerase (RdRp/nsp12). These mutations were established early and co-occur with almost every other SARS-CoV-2 mutation, thus appearing as horizontal/vertical bands on the heatmap, which otherwise exhibits a rather sparse structure with occasional co-occurrence hotspots. Co-occurrence sparsity is not uniform across the viral genome, although as the pandemic proceeds, co-occurrence becomes very noticeable across the entire genome.

748 ***Fig. S21***

749 **Genome-wide co-occurrence (co-occurrence) matrices across the entire SARS-CoV-2**
750 **pandemic timeline, reported on 8/15/21.** For each co-occurrence matrix heatmap, values on
751 the axes are genomic coordinates (in bps) for the reference Wuhan SARS-CoV-2 genome.
752 Cumulative co-occurrence counts (in $\log_{10}$ scale) are reported at each location in blue gradient
753 scale. Regions corresponding to viral genes are labeled on top (blue). Locations of VOC co-
754 occurrences observed at this time point are highlighted as colored empty shapes (Alpha, black
755 circles; Beta, magenta hexagons; Gamma, green squares; Delta, red diamonds). ***Summary of***
756 ***results***: The representative genome-wide co-occurrence matrices heatmaps show the mutation
757 co-occurrence landscape during the early, mid-, and late stages of the pandemic. For example,
758 on 3-15-2020 (**Fig. S13**), the most common mutations are allele changes encoding D614G in
759 Spike and P471L in the catalytic subunit of the RNA-dependent RNA polymerase (RdRp/nsp12).
760 These mutations were established early and co-occur with almost every other SARS-CoV-2
761 mutation, thus appearing as horizontal/vertical bands on the heatmap, which otherwise exhibits
762 a rather sparse structure with occasional co-occurrence hotspots. Co-occurrence sparsity is not
763 uniform across the viral genome, although as the pandemic proceeds, co-occurrence becomes
764 very noticeable across the entire genome.

765 ***Fig. S22***

766 **Evolution of the co-occurrence search strategies of the Delta VOC.** Co-occurrence patterns
767 for Delta highlighted within genome-wide co-occurrence matrices at additional time-point
768 1/19/2021. Figure layout as in Results **Fig. S13-21** with details and legend as described in
769 legends therein). Summary: The full grid of Delta VOC co-occurrences is preceded, during the
770 long exploratory phase by a L-shaped configuration of regularly spaced Delta co-occurrences
771 with one horizontal sequence of co-occurrence at the 3' end of NC and one vertical at ~23,000
772 bp. These results suggest the possibility that the Delta VOC is running a systematic preliminary

exploration of the co-occurrence space prior to settling on the final stage for the full grid VOC configuration - marking the beginning of the wide-spread predatory behavior of the Delta variant.

Fig. S23

Evolution of the co-occurrence search strategies of the Delta VOC. Co-occurrence patterns for Delta highlighted within genome-wide co-occurrence matrices at additional time-point 2/16/2021. Figure layout as in Results **Fig. S13-21** with details and legend as described in legends therein. Summary: The full grid of Delta VOC co-occurrences observed is preceded, during the long exploratory phase by a L-shaped configuration of regularly spaced Delta co-occurrences with one horizontal sequence of co-occurrence at the 3' end of NC and one vertical at ~23,000 bp. These results suggest the possibility that the Delta VOC is running a systematic preliminary exploration of the co-occurrence space prior to settling on the final stage for the full grid VOC configuration - marking the beginning of the wide-spread predatory behavior of the Delta variant.

Fig. S24

Time lapse of viral genome GP maps across the pandemic timeline, annotated with signature mutations for the viral polymerase nsp12. Shown are six tri-monthly time-points between 5/15/2020 and 8/15/2021. $\overline{\mathrm{IR}}$ and $\overline{\mathrm{FR}}$ are log-transformed, genomic position is scaled to a 0-1 interval. Vertical dotted blue lines are boundaries between SARS CoV-2 proteins, annotated on the top axis. Input variants are in shaded color, with dot sizes proportional to allele frequency of the mutations. A set of 45 distinctive Nsp12 mutations (see **Methods**) is annotated with black dots and labels on each GP-based allele phenotype landscape. Summary: GP analysis focusing on distinctive mutations affecting the nsp12 polymerase subunit reveals that G671S plays a prominent role in creating a Delta VOC reconfiguration over time.

Fig. S25

Co-occurrence patterns for Omicron BA.1 highlighted within genome-wide co-occurrence matrices heatmaps. A-D. Reported are 4 time-points: (**A**) 8/15/21 (beginning of Omicron co-occurrence data analysis); (**B**) 10/15/21 (before the leap in co-occurrence density as in Results **Fig. 7B**); (**C**) 11/15/21 (after the leap in co-occurrence density, Results **Fig. 7B**); (**D**) 1/15/22 (last time point available). Values on the axes are genomic positions (bps) for the reference Wuhan SARS-CoV-2 genome. Cumulative co-occurrence counts (in $\log_{10}$ scale) are reported at each location in the blue gradient scale. Regions corresponding to viral genes are labeled on top (blue). For each co-occurrence matrix heatmap, red markers highlight co-occurring mutations that are new with respect to the previous date examined (e.g., the red markers on 10/15/21 matrix are co-occurring mutations whose count is above zero as of 10/15/21 but were zero in the previous iteration (8/15/21)). Black markers highlight co-occurrences that were already present on the previous date.

Fig. S26

Co-occurrence patterns for combined Omicron VOC highlighted over the genome-wide co-occurrence matrix, as of 11/15/21. A. Values on the axes are genomic coordinates (in genomic positions (bp)) for the reference Wuhan SARS-CoV-2 genome. Cumulative co-occurrence counts (in $\log_{10}$ scale) are reported at each location in blue gradient scale. Regions corresponding to viral genes are labeled on top (blue). Locations of Omicron co-occurrences observed up to 10/15/21 are marked as black diamonds; Omicron co-occurrences added between 10/15/21 and 11/15/21 are highlighted as red diamonds. **B.** Representative co-occurrence matrix showing co-occurrence counts between Omicron VOC signature mutations added between 10/15/21 and 11/15/21 (corresponding to red diamonds in the global co-occurrence visualization in (**A**)).

Fig. S27

Mutation comparison across Omicron lineages Figures (adapted from outbreak.info[22]). Showing all mutations in all SARS-CoV-2 proteins with > 75% prevalence in at least one of BA.1, BA.2, BA.3, BA.4, BA.5, or BA.2.75 Omicron lineages. Note new mutations between the earlier (BA.1-3) and later (BA.4-5, BA.2.75) in mutations – Spike L452R is present in both BA.4 and BA.5 but not in BA.1, for example, and is the only mutation designated a VOI or VOC in the BA.4 and BA.5 sublineages. This points to a less than obvious reason why these two sub-lineages have dominated cases, but with a lower fatality, something that EWAD can investigate by focusing on the GP-based mean $\overline{\boldsymbol{IR}}$ and $\overline{\boldsymbol{FR}}$ residuals being capture in response to the surrounding residues impacting, for instance, Spike function. There are 39 mutations in total that are absent in one or both of BA.1 and BA.2, but present in one or more of BA.4, BA.5, and BA.2.75 (outlined in red). Notably, a higher proportion of mutations (when compared to the prevalence of all Omicron mutations) in the three newer sublineages occur in ORF1ab, responsible for transcription and replication (https://www.ncbi.nlm.nih.gov/gene/43740578), suggesting a possible reasons for that these sub-lineages' have higher transmissibility.

SUPPLEMENTAL REFERENCES

1. O'Toole, Á., Scher, E., Underwood, A., Jackson, B., Hill, V., McCrone, J.T., Colquhoun, R., Ruis, C., Abu-Dahab, K., Taylor, B., et al. (2021). Assignment of epidemiological lineages in an emerging pandemic using the pangolin tool. Virus Evolution. 10.1093/ve/veab064.
2. Hadfield, J., Megill, C., Bell, S.M., Huddleston, J., Potter, B., Callender, C., Sagulenko, P., Bedford, T., and Neher, R.A. (2018). Nextstrain: real-time tracking of pathogen evolution. Bioinformatics *34*, 4121-4123. 10.1093/bioinformatics/bty407.
3. Wang, C., Anglès, F., and Balch, W.E. (2022). Triangulating variation in the population to define mechanisms for precision management of genetic disease. Structure *30*, 1190-1207.e1195. https://doi.org/10.1016/j.str.2022.05.011.
4. Wang, C., and Balch, W.E. (2018). Bridging Genomics to Phenomics at Atomic Resolution through Variation Spatial Profiling. Cell Reports *24*, 2013-2028.e2016. 10.1016/j.celrep.2018.07.059.
5. Wang, C., Scott, S.M., Sun, S., Zhao, P., Hutt, D.M., Shao, H., Gestwicki, J.E., and Balch, W.E. (2020). Individualized management of genetic diversity in Niemann-Pick C1 through modulation of the Hsp70 chaperone system. Hum Mol Genet *29*, 1-19. 10.1093/hmg/ddz215.
6. Wang, C., Scott, S.M., Subramanian, K., Loguercio, S., Zhao, P., Hutt, D.M., Farhat, N.Y., Porter, F.D., and Balch, W.E. (2019). Quantitating the epigenetic transformation contributing to cholesterol homeostasis using Gaussian process. Nature Communications *10*, 5052. 10.1038/s41467-019-12969-x.
7. Romero, P.A., Krause, A., and Arnold, F.H. (2013). Navigating the protein fitness landscape with Gaussian processes. Proc Natl Acad Sci U S A *110*, E193-201. 10.1073/pnas.1215251110.
8. Bedbrook, C.N., Yang, K.K., Rice, A.J., Gradinaru, V., and Arnold, F.H. (2017). Machine learning to design integral membrane channelrhodopsins for efficient eukaryotic expression and plasma membrane localization. PLoS Comput Biol *13*, e1005786. 10.1371/journal.pcbi.1005786.
9. Yang, K.K., Wu, Z., and Arnold, F.H. (2019). Machine-learning-guided directed evolution for protein engineering. Nat Methods *16*, 687-694. 10.1038/s41592-019-0496-6.
10. Wang, C., Elghobashi-Meinhardt, N., and Balch, W.E. (2022). Covariant Fitness Clusters Reveal Structural Evolution of SARS-CoV-2 Polymerase Across the Human Population. bioRxiv, 2022.2001.2007.475295. 10.1101/2022.01.07.475295.
11. Angelini, M.M., Akhlaghpour, M., Neuman, B.W., Buchmeier, M.J., and Moscona, A. (2013). Severe Acute Respiratory Syndrome Coronavirus Nonstructural Proteins 3, 4, and 6 Induce Double-Membrane Vesicles. mBio *4*, e00524-00513. doi:10.1128/mBio.00524-13.
12. Van Valen, L. (1973). A new evolutionary law. Evolutionary Theory *1*, 1-30.
13. https://en.wikipedia.org/wiki/Hyperparameter_optimization#Grid_search (2021).
14. WHO (2021). Classification of Omicron (B.1.1.529): SARS-CoV-2 Variant of Concern. https://www.who.int/news/item/26-11-2021-classification-of-omicron-(b.1.1.529)-sars-cov-2-variant-of-concern.
15. Telenti, A., Hodcroft, E.B., and Robertson, D.L. (2022). The Evolution and Biology of SARS-CoV-2 Variants. Cold Spring Harb Perspect Med *12*. 10.1101/cshperspect.a041390.
16. Shrestha, L.B., Foster, C., Rawlinson, W., Tedla, N., and Bull, R.A. (2022). Evolution of the SARS-CoV-2 omicron variants BA.1 to BA.5: Implications for immune escape and transmission. Reviews in Medical Virology, e2381. 10.1002/rmv.2381.
17. Mohapatra, R.K., Kandi, V., Mishra, S., Sarangi, A.K., Pradhan, M.K., Mohapatra, P.K., Behera, A., and Dhama, K. (2022). Emerging novel sub-lineage BA.2.75: The next dominant omicron variant? Int J Surg *104*, 106835. 10.1016/j.ijsu.2022.106835.