**Universidade de São Paulo**
**Faculdade de Medicina de Ribeirão Preto**

# Representações numéricas e técnicas livres de alinhamento de sequências como ferramentas de agrupamento não supervisionado: aplicações em filogenia de coronavírus e linhagens brasileiras de SARS-CoV-2.

**Murilo Henrique Anzolini Cassiano**

**Ribeirão Preto**
**2023**

**MURILO HENRIQUE ANZOLINI CASSIANO**

# Representações numéricas e técnicas livres de alinhamento de sequências como ferramentas de agrupamento não supervisionado: aplicações em filogenia de coronavírus e linhagens brasileiras de SARS-CoV-2.

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Biologia Celular e Molecular da Faculdade de Medicina de Ribeirão Preto/USP para obtenção do Título de Mestre em Ciências.

Área de Concentração: Bioinformática.

**Orientador:** Prof. Dr. Eurico Arruda
**Orientador:** Dr.Daniel Macedo de Melo Jorge

Ribeirão Preto-SP
2023

Anzolini-Cassiano, Murilo Henrique

Representações numéricas e técnicas livres de alinhamento de sequências como ferramentas de agrupamento não supervisionado: aplicações em filogenia de coronavírus e linhagens brasileiras de SARS-CoV-2. Ribeirão Preto, 2023.

83 p. : il. ; 30cm.

Dissertação de Mestrado, apresentada à Faculdade de Medicina de Ribeirão Preto/USP – Área de concentração: Bioinformática.

Orientador: Arruda, Eurico

Co-orientador: Jorge, Daniel Macedo de Melo

1. Análise de sequências. 2. Análise livre de alinhamento. 3. SARS-CoV-2. 4. Estudos filogenéticos. 5. Representações numéricas de genomas.

**MURILO HENRIQUE ANZOLINI CASSIANO**

# Representações numéricas e técnicas livres de alinhamento de sequências como ferramentas de agrupamento não supervisionado: aplicações em filogenia de coronavírus e linhagens brasileiras de SARS-CoV-2.

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Biologia Celular e Molecular da Faculdade de Medicina de Ribeirão Preto/USP para obtenção do Título de Mestre em Ciências.

Área de Concentração: Bioinformática

Aprovado em: ____/____/____.

**Banca Examinadora**

Prof. Dr.

_____

Instituição: _____ Assinatura:

_____

Prof. Dr.

_____

Instituição: _____ Assinatura:

_____

Prof. Dr.

_____

Instituição: _____ Assinatura:

_____

**A minha mãe Fátima, a minha vó Neide, ao meu irmão Leonardo e a minha *soulmate* Letícia, com carinho**.

Agradeço a CAPES por conceder o suporte financeiro: *o presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001.*

## Resumo

ANZOLINI-CASSIANO, M. H. **Representações numéricas e técnicas livres de alinhamento de sequências como ferramentas de agrupamento não supervisionado: aplicações em filogenia de coronavírus e linhagens brasileiras de SARS-CoV-2**. 2023. 83f. Dissertação (Mestrado). Faculdade de Medicina de Ribeirão Preto – Universidade de São Paulo, Ribeirão Preto, 2022.

A pandemia da SARS-CoV-2 se espalhou pelo mundo causando quase 700 milhões de casos confirmados, sendo 34 milhões apenas no Brasil. Os coronavírus têm um dos maiores genomas entre todos os vírus RNA e, embora codifiquem exonucleases corretoras de erros, ainda hoje, novas linhagens ainda emergem, criando uma diversidade significativa de genomas. Nesse sentido, os esforços para o rastreamento de linhagens emergentes de SARS-CoV-2 geraram um volume expressivo e sem precedentes de dados públicos referentes à sequências genômicas. Todavia, lidar com tamanha quantidade de dados com técnicas convencionais dependentes de alinhamento é impraticável computacionalmente. Visando lidar com grandes conjuntos de dados e, ao mesmo tempo, com algumas das limitações das técnicas baseadas em alinhamentos, diversas metodologias foram propostas para codificação numérica e subsequente comparação de distâncias evolutivas entre genomas completos. Apesar da diversidade de técnicas disponíveis, há uma escassez de comparações criteriosas das metodologias existentes. Neste sentido, a grande disponibilidade de sequências de SARS-CoV-2 oferece uma oportunidade para aplicação de representações numéricas de genomas completos desenvolvidas nos últimos anos com foco em comparação de sequências virais. Neste trabalho testamos as representações numéricas baseadas em K-mer: Triplet Frequency, K-mer Natural Vector, Fast Vector, e Magnus Genomic Representation com sequências de coronaviridae (curadas e publicadas) e aproximadamente 86 mil genomas sequenciados no Brasil, obtidos do banco de dados GISAID EpiCov. Para cada dataset, comparamos i) medidas que sumarizam características estruturais, ii) correlações cofenéticas e iii) distâncias, entre as árvores feitas com as distâncias euclidianas das representações numéricas e a árvore construída a partir de alinhamento múltiplo de sequências com conseguinte estimação filogenética por máxima-verossimilhança. Também avaliamos a capacidade de cada representação testada em carregar consigo informações biológicas sabidas das sequências, como grupo taxonômico ou linhagem viral, via técnicas de redução de dimensionalidade. Vimos que no geral todas as representações numéricas revelaram algum padrão biológico esperado para agrupamento dos genomas virais e, embora as técnicas aqui exploradas, juntamente com uma das melhores e mais acuradas ferramentas publicada para comparação de sequências livre de alinhamento falhem em recuperar características globais da árvore filogenética de SARS-CoV-2, vimos que seu uso como entrada para o algoritmo neighbor-joining resultou em árvores que mantém a estrutura local, sendo aptas para separação de linhagens virais. Esperamos que estes resultados, juntamente com os códigos construídos para implementar a metodologia possam servir como base tanto para o desenvolvimento de ferramentas como para melhoria das técnicas de comparações genômicas livres de alinhamento.



Palavras-chave: análise de sequências, análise livre de alinhamento, SARS-CoV-2, estudos filogenéticos, representações numéricas de genomas.

## Abstract

ANZOLINI-CASSIANO, M. H. **Numeric representations and alignment-free techniques for sequence clustering as tools for unsupervised grouping: applications in coronavirus phylogeny and Brazilian lineages of SARS-CoV-2.** . 2023. 83f. Dissertation (Master). Faculdade de Medicina de Ribeirão Preto – Universidade de São Paulo, Ribeirão Preto, 2023.

The SARS-CoV-2 pandemic spread throughout the world causing nearly 700 million confirmed cases, with 34 million just in Brazil. Coronaviruses have one of the largest genomes among all RNA viruses, and although they encode error-correcting exonucleases, new lineages are still emerging, creating a significant diversity of genomes. In this sense, efforts to track emerging SARS-CoV-2 lineages have generated an unprecedented and substantial amount of public data regarding genomic sequences. However, dealing with such a large amount of data with conventional alignment-dependent techniques is computationally impractical. Aiming to deal with large datasets and, at the same time, with some of the limitations of alignment-based techniques, several methods have been proposed for numerical encoding and subsequent comparison of evolutionary distances between complete genomes. Despite the diversity of techniques available, there is a scarcity of rigorous comparisons of existing methods. In this sense, the large availability of SARS-CoV-2 sequences offers an opportunity for the application of numerical representations of complete genomes developed in recent years with a focus on viral sequence comparison. In this work, we tested the K-mer-based numerical representations: Triplet Frequency, K-mer Natural Vector, Fast Vector, and Magnus Genomic Representation with coronaviridae sequences (cured and published) and approximately 86 thousand sequenced genomes in Brazil, obtained from the GISAID EpiCov database. For each dataset, we compared i) measures that summarize structural characteristics, ii) cofeneic correlations, and iii) distances between the trees made with the Euclidean distances of the numerical representations and the tree built from multiple sequence alignment and subsequent phylogenetic estimation by maximum likelihood. We also evaluated the ability of each tested representation to carry biological information known from the sequences, such as taxonomic group or viral lineage, through dimensionality reduction techniques. We saw that overall all the numerical representations revealed some expected biological pattern for grouping viral genomes, and although the techniques explored here, along with one of the best and most accurate published tools for alignment-free sequence comparison, fail to recover global characteristics of the SARS-CoV-2 phylogenetic tree, we saw that its use as input to the neighbor-joining algorithm resulted in trees that maintain the local structure, being suitable for separating viral lineages. We hope that these results, along with the codes built to implement the methodology, can serve as a basis both for the development of tools and for the improvement of alignment-free genomic comparison techniques.



Keywords: sequence analysis, alignment-free analysis, SARS-CoV-2, phylogenetic studies, numerical representations of genomes.

## Lista de Figuras

## Lista de Tabelas

## Sumário

## 1 Introdução

Cerca de 700 milhões de casos e aproximadamente 7 milhões mortes confirmadas ocorreram desde o início da pandemia da COVID-19 até fevereiro de 2023 (COVID-19 Dashboard https://www.arcgis.com/apps/dashboards/bda7594740fd40299423467b48e9ecf6, acesso em 11 de fevereiro de 2023). Segundo o Ministério da Saúde, somente no Brasil foram reportados aproximadamente 37 milhões de casos confirmados, 700 mil vidas perdidas, com mortalidade definida em 1,9% (Secretarias Estaduais de Saúde. Brasil, 2020. https://covid.saude.gov.br, acesso em 11 de fevereiro de 2023).

Os primeiros casos de COVID-19 datam de 8 de dezembro de 2019, com pacientes apresentando sintomas de pneumonia grave causada por vírus. Dia 9 de janeiro de 2020, praticamente um mês após os primeiros relatos, a China anunciou a descoberta do agente etiológico causador da doença, via análise de metagenomas de pacientes (ZHOU et al., 2020). Somente em 11 de março de 2020 a Organização Mundial da Saúde declarou a COVID-19 como pandemia (HU et al., 2021).

O agente etiológico foi um novo vírus que ficou conhecido como *Severe Acute Respiratory Syndrome Coronavirus* 2, ou SARS-CoV-2. Esse vírus pertence à ordem Nidovirales, família Coronaviridae, sub-família Coronavirinae, gênero betacoronavírus, que inclui os vírus SARS-CoV e MERS-CoV, que também foram alvo de preocupações para os órgãos de saúde nos anos de 2003 e 2012, respectivamente. SARS-CoV e SARS-CoV-2 têm maior proximidade entre si, e pertencem ao subgênero sarbecovírus (GORBALENYA et al., 2020).

A subfamília dos Coronavirinae é composta por 4 gêneros: alfacoronavírus, betacoronavírus, gamacoronavírus e deltacoronavírus. Dentre estes, alguns pertencentes a classificação alfa e betacoronavírus infectam mamíferos, assim como alguns pertencentes ao gama e deltacoronavírus. Não é novidade que estes circulam e possuem potencial patogênico (KHAN et al., 2020), podendo causar infecções intestinais e doença respiratórias em humanos, no entanto, até o início do século XXI os casos eram reportados isoladamente em pessoas que apresentam algum tipo de comorbidade. Alguns exemplos de coronavírus humanos endêmicos circulantes há muitos anos, mas que causam infecção branda, são os HCoV-OC43, HCoV-229E e HKU1 (CUI; LI; SHI, 2019).

O genoma de SARS-CoV-2 é um RNA de fita simples não-segmentado, com polaridade positiva, que é o maior genoma dentre os vírus de RNA, contendo 29.903 nucleotídeos. Destes, dois terços codificam um único open read frame (orf), a orf1ab, que é traduzida em um único peptídeo, cujo terço 3' proximal codifica as 4 proteínas estruturais: S (spike), M (membrana), E (envelope) e N (nucleoproteína) (JUNGREIS, I., SEALFON, R. & KELLIS, 2021). O peptídeo da orf1ab dá origem a 16 proteínas não estruturais, dentre as quais destacam-se a nsp12, que é a RNA polimerase dependente de RNA (RdRp), necessária e suficiente para transcrição e replicação do genoma, e a nsp14, que é uma exonuclease 3', com função revisora de erros, necessária para a estabilidade dos longos genomas de RNA (OGANDO et al., 2020).

A hipótese mais aceita para a origem do SARS-CoV-2 (DEIGIN; SEGRETO, 2021) é a de que este seja originário de animais, sendo altamente provável que tenha sido transmitido a seres humanos por hospedeiros intermediários, como visto para o SARS-CoV e para o MERS-CoV. A espécie animal mais comumente associada como hospedeiro intermediário é o morcego, pois estudos filogenéticos demonstraram altos níveis de similaridade com genomas de morcegos, nomeadamente o vírus RaTG13, que chega a 96% de identidade de genoma (ZHOU et al., 2020). Foi constatada alta similaridade da região do Receptor Binding Domain (RBD) da proteína spike com coronavírus isolados de pangolim (97,4% de identidade em nível de aminoácido) (LAM et al., 2020) e, como esta região tem papel chave no tropismo do vírus pelo receptor ACE2 humano, se hipotetiza que o pangolim tenha sido hospedeiro (e não reservatório natural) de um ancestral comum do SARS-CoV-2 que provavelmente evoluiu sofrendo diversas recombinações genômicas (SINGH; YI, 2021), antes de ser transmitido aos humanos.

O enfrentamento da pandemia de SARS-CoV-2 colocou em evidência a grande importância da bioinformática não só na compreensão de filogenia e evolução de SARS-CoV-2 e suas variantes, quanto na obtenção de pistas quanto à sua origem, e constatação de hospedeiros intermediários, tais como cobras, ratos, hamsters, pangolins, visto que o índice de adaptação de códons do SARS-CoV-2 para com o uso de códon desses animais e de humanos demonstrou-se altíssimo (DILUCCA et al., 2020). Adicionalmente, a bioinformática, por análise de sequência, pode ser usada para desvendar padrões menos explícitos nos genomas, como por exemplo quando foi visto que o sítio de clivagem de furina na proteína spike do SARS-CoV-2, um

importante e determinante elemento para o espalhamento deste vírus em seus antecessores (JOHNSON et al., 2021), provavelmente é fruto de um mecanismo molecular presente na estrutura secundária do genoma viral e não uma criação humana (SANDER et al., 2022).

A quantidade de dados disponíveis sobre o vírus SARS-CoV-2 é muito maior do que a quantidade de dados disponíveis sobre outros vírus, como a influenza (CARVALHO et al., 2020). Isso se deve em parte ao fato de a pandemia de COVID-19 ter atraído uma enorme quantidade de atenção, recursos para a pesquisa e formado redes de colaboração para o monitoramento da doença. Em consequência disso, a disponibilidade de sequências genômicas completas e dados epidemiológicos sobre SARS-CoV-2 tem sido amplamente divulgada e acessível a pesquisadores de todo o mundo. Por instância, quando pesquisamos por 'Influenza' e por 'SARS-CoV-2' na base de dados do NCBI, vemos que a quantidade de arquivos de sequenciamento no segundo caso é mais de 100 vezes maior do que o primeiro (https://www.ncbi.nlm.nih.gov/search/all/?term=Influenza e https://www.ncbi.nlm.nih.gov/search/all/?term=SARS-CoV-2, acesso em 11 de fevereiro de 2023).

Na base da maioria esmagadoras das análises biológicas atuais estão os algoritmos de alinhamento múltiplo de sequências, sejam elas DNA, RNA ou proteínas. O alinhamento é feito para que, não só a similaridade e conservação das sequências fiquem evidente, mas também para que cada base ou resíduo em cada posição represente a mesmo significado evolutivo, assim relações evolutivas entre organismos, genes ou pistas sobre estrutura-função são descobertas (CHATZOU et al., 2016). Acompanhado pela grande disponibilidade de sequenciamento, o alinhamento múltiplo se tornou uma tarefa cada vez mais complexa e desafiadora. Isso levou ao desenvolvimento de diversas heurísticas e algoritmos eficientes para lidar com esse problema, com estes não garantindo a solução ótima, mas soluções aproximadamente acuradas e mais rápidas (KEMENA; NOTREDAME, 2009) e a escolha destas soluções afeta diretamente os resultados de análises posteriores (WARNOW, 2021).

Em uma revisão complexa e detalhada do assunto (ZIELEZINSKI et al., 2017), muitas situações problemáticas das análises de sequências baseadas em

alinhamento foram discutidas, dentre as quais, para nosso contexto de análise do coronavírus SARS-CoV-2, valem mencionar que: os programas de alinhamento assumem que as sequências possuem colinearidade, isto é, assumem que as sequencias homólogas são compostas por uma série de trechos de sequências lineares com um certo grau de conservação; estas abordagens requerem enormes quantidades de recurso e tempo computacionais, especialmente memória, levando a estudos de dados em escala de multigenômica ficarem inviáveis e infactíveis; e, por último, a maioria dos programas de alinhamento utilizam diversas suposições a priori sobre os processos evolutivos que moldam as sequências em questão, o que levam a resultados caracteristicamente distintos e dependentes das ferramentas (WONG; SUCHARD; HUELSENBECK, 2008).

De forma direta, a colinearidade não é garantida, uma vez que os vírus realizam frequentes recombinações. Há evidências experimentais de que que, além da natureza da replicação dos Nidovirales ser aninhada (‘nested’), o que torna os coronavírus propensos a recombinação, a proteína nsp14 tem papel direto nesse processo, o que aumenta a complexidade e imprevisibilidade em entender o surgimento de variantes e linhagens deste coronavírus (GRIBBLE et al., 2021; YANG et al., 2021).

Para entender a não-factibilidade, tomemos por exemplo sequências do gene 16s rRNA de procariotos, que contém por volta de 1.500 pares de base e são tidos como válidas marcações para distinguir espécies (CLARRIDGE, 2004). Foi descrito que o RAxML requer aproximadamente um mês de tempo de CPU para processar aproximadamente 28.000 destes rRNA e com bootstrap de 444 amostragens, esse tempo seria acrescido em 5,6 anos de tempo de CPU (LIU; LINDER; WARNOW, 2011). Temos que, no geral, os genomas de coronavírus são aproximadamente 20 vezes maiores do que o mencionado no caso acima.

Por definição, as análises livres de alinhamento são aquelas em que se quantifica similaridade ou dissimilaridade de um conjunto de sequências sem que haja o passo intermediário de alinhamento. Como estas técnicas não dependem das técnicas de programação dinâmica e suas variantes, comumente associadas aos algoritmos de alinhamento para a busca das melhores soluções de alinhamento dentro do espaço total de soluções do problema, elas tendem a ser mais eficientes para lidar com grandes quantidades de dados (VINGA, 2014). A maioria destas abordagens consiste em contar a frequência de cada subsequência de tamanho definido (os

chamados k-mers) que ocorrem em um dado conjunto de sequencias no qual se quer aferir relações de proximidade evolutiva e, a partir destas medidas, aplicar algumas séries de tratamentos estatísticos ou fórmulas de distância diversas para a quantificação de similaridade ou dissimilaridade (BONHAM-CARTER; STEELE; BASTOLA, 2014).

A partir dos k-mers foram desenvolvidas diversas representações numéricas de material genético ao longo dos anos para aplicações de bioinformática e de reconhecimento de padrões, cada uma com seus méritos e deméritos, carregando informação bioquímica/biofísica, codificando propriedades matemáticas complexas ou ainda propriedades da estrutura primária do material genético (KWAN; ARNIKER, 2009) .

Várias destas já foram exploradas como métricas para cálculo de similaridade entre sequências genômicas e indicadas para aumentar a eficiência e comparação de genomas em grandes bancos de dados biológicos (MENDIZABAL-RUIZ et al., 2017; ZHANG et al., 2017). Contudo, ainda há a necessidade da criação de novos métodos que, principalmente, combinem diferentes propriedades na mesma representação para codificar genomas numericamente e assim extrair propriedades biológicas das suas sequências (YU; LI; YU, 2018).

Maior talvez do que a necessidade da criação de novos métodos (uma vez que que está área é relativamente nova na bioinformática/análise de sequência), existe uma chamada na literatura para que as técnicas sejam compreensivamente testadas (ZIELEZINSKI et al., 2019) e feitas públicas, com seus códigos implementados e disponíveis para uso da comunidade, explorando dados de aplicações “do mundo real” e de tamanho razoáveis (RANDHAWA; HILL; KARI, 2019).

Mais do que somente gerar os dados genômicos, as redes de colaboração da pandemia geraram metadados e classificações para as sequências levando em conta informações filogenéticas e epidemiológicas, curadas por *experts* no assunto e discutidas coletivamente (RAMBAUT et al., 2020). Assim, acreditamos que os dados genômicos de SARS-CoV-2, por sua vasta disponibilidade e por seus curados metadados sejam uma fonte beneficiosa para testes com técnicas não supervisionadas de agrupamento focados na averiguação do quanto as representações genômicas e técnicas livres de alinhamento conseguem refletir estas informações biológicas levantadas e construídas por especialistas.

Portanto, a presente dissertação de mestrado tem como objetivo avaliar três diferentes abordagens de representação de sequências livres de alinhamento. Mais especificamente, neste trabalho foram testadas três abordagens de representação que foram desenvolvidas nos últimos anos, bem como uma representação básica de k-mers. É importante ressaltar que as abordagens avaliadas se destacam por carregar a informação posicional das sub palavras, além de sua contagem. Dessa forma, esperamos que este estudo contribua para a compreensão das vantagens e limitações dessas técnicas de representação em análises de sequências biológicas, em especial, de sequências de SARS-CoV-2.

## 2 Objetivo Geral

Analisar e comparar métodos de representação numérica e filogenética para análise genômica de sequências de SARS-CoV-2 brasileiras obtidas durante a pandemia e representantes da subfamília Orthocoronavirinae.

## 2.1 Objetivos Específicos

- Desenvolver um dataset das sequências brasileiras de SARS-CoV-2 associados a metadados;
- Implementar códigos e scripts para cada uma das metodologias de representação numérica, visando a possível integração entre elas;
- Analisar comparativamente representações numéricas do genoma de SARS-CoV-2 e dos quatros gêneros da subfamília Orthocoronavirinae e comparar com os resultados melhor ferramenta livre de alinhamento da literatura;
- Determinar a melhor forma de codificar genomas numericamente e extrair propriedades biológicas de suas sequências;
- Comparar métricas quantitativas de árvores geradas a partir de sequências alinhadas ou não alinhadas;

# 3 Material e Métodos

## 3.1 Sequências genômicas da subfamília Orthocoronavirinae

Para análise de sequências em um nível taxonômico de subfamília foi utilizado um dataset publicado de 69 genomas completos de representantes da subfamília Orthocoronavirinae, que estavam disponíveis no trabalho de Kirichenko et al (KIRICHENKO et al., 2022). Destes, 66 genomas completos de coronavírus foram baixados do NCBI com o uso da ferramenta *E-utilities*, sob os ids NC_022103.1, NC_048216.1, NC_046964.1, NC_032107.1, NC_028833.1, NC_028824.1, NC_028814.1, NC_028811.1 ,NC_032730.1, NC_018871.1, NC_025217.1, NC_030886.1, MN996532, NC_034440.1, NC_014470.1, NC_010438.1, NC_010437.1, NC_009988.1, NC_009021.1, NC_009020.1, NC_009019.1, NC_009657.1, NC_005831.2, NC_002645.1, NC_045512.2, NC_006213.1, NC_038294.1, NC_019843.3, NC_006577.2, NC_004718.3, NC_034972.1, NC_028752.1, MT121216.1, MT040336.1, NC_039207.1, NC_026011.1, NC_017083.1, NC_012936.1, NC_003045.1, KX432213.1, JX860640.1, NC_010646.1, NC_038861.1, NC_030292.1, NC_028806.1, NC_023760.1, NC_002306.3, NC_003436.1, NC_039208.1, NC_010800.1, NC_048214.1, NC_048213.1, NC_046965.1, NC_001451.1, NC_011547.1, NC_016992.1, NC_016991.1, NC_016996.1, NC_016995.1, NC_016994.1, NC_016993.1, NC_011550.1, NC_011549.1, NC_001846.1, NC_048217.1 e AC_000192.1. As adicionais 03 sequências de Sars-Cov-2 foram baixadas da base de dados do GISAID EpiCoV database (KHARE et al., 2021), sendo: Beijing.IVDC_01, Beijing.IVDC_02 e Beijing.IVDC_03.

## 3.2 Sequências genômicas brasileiras de SARS-CoV-2 e seus tratamentos iniciais

Para análises no nível taxonômico de espécies, obtivemos 85.691 sequências de genomas completos também provenientes do GISAID EpiCov database e estes, por sua vez foram filtrados por genomas completos (aproximadamente 30kb), vírus que foram extraídos de amostras humanas, com alta cobertura. Dentre as informações obrigatórias no metadado para seleção da sequência está a data da coleta completa.

Os isolados compreendem desde o início da pandemia no Brasil, sendo o primeiro registro de coleta em 25 de fevereiro de 2020, até 2 de setembro de 2022, data em que atualizamos o *dataset* do laboratório pela última vez.

Os genomas obtidos foram alinhados à referência (variante Wuhan) e classificados usando a nomenclatura de clados do Nextrain (https://nextstrain.org/blog/2022-04-29-SARS-CoV-2-clade-naming-2022, acesso em 11 de fevereiro de 2023) de acordo com um conjunto de mutações-assinatura bem definidas, com o uso da ferramenta Nextclade Versão 1.11.0 (AKSAMENTOV et al., 2021).

Com o objetivo de reduzir o tempo computacional da construção da árvore filogenética de máxima-verossimilhança de referência, visto a sua impraticável inferência no conjunto completo, foi utilizada a ferramenta *filter tool* do software Augur (HUDDLESTON et al., 2021) para fazer uma subamostragem por data, selecionando sequências por mês-ano, imponto o limite de 3 mil genomas.

Vimos que foram feitas abordagens semelhantes em outros trabalhos disponíveis na literatura, com o mesmo intuito de redução de base de dados (NASCIMENTO et al., 2020; GRÄF et al., 2021; ZIMERMAN et al., 2022), pois assim, dependendo da análise, têm-se tanto a inferência confiável de relógios moleculares e sinais temporais como evita-se a perda de genomas de baixa diversidade, iniciais a uma pandemia (BOLYEN et al., 2020; MARINI et al., 2022). Adicionalmente, outros projetos em nosso laboratório demonstraram que a subamostragem que leva em consideração somente as datas de coleta produziram datasets que preservam sinais filogenéticos (dados não mostrados).

O sinal evolutivo da subamostragem foi realizado usando o algoritmo do *likelihood mapping*, utilizando a ferramenta IQTREE v2.2.0 (MINH et al., 2020), com modelo GTR+I+G e 75.000 amostragens de quartetos, seguindo recomendações do manual de referência dos autores (http://www.iqtree.org/doc/Assessing-Phylogenetic-Assumptions, acesso em 11 de fevereiro de 2023).

### 3.3 Alinhamento Múltiplo de Sequências e Análise Filogenética

Os 69 genomas completos foram alinhados utilizando o MAFFT v7.490 (KATOH; STANDLEY, 2013), com o uso do parâmetro LINSI (--maxiterate 1000 --localpair) para aumentar a acurácia. Em seguida, o alinhamento múltiplo foi trimado com o software Trimal v1.2 (CAPELLA-GUTIÉRREZ; SILLA-MARTÍNEZ; TONI

GABALDÓN, 2009), para remover de forma automatizada as posições pobres e pouco informativas no alinhamento.

As árvores filogenéticas de máxima-verossimilhança foram obtidas com o software IQTREE v2.2.0, com o uso do parâmetro Ultrafast Bootstrap no valor de 1000 (HOANG et al., 2018) e 1000 réplicas do teste SH-aLRT (NASER-KHDOUR et al., 2019). Utilizamos o modelo de substituição rico em parâmetros GTR+I+G, pois há evidências de que a escolha desse modelo leva a análises filogenéticas equiparáveis às análises feitas com as estratégias atuais de seleção de modelo evolutivo (ABADI et al., 2019). Todas as representações gráficas das árvores filogenéticas consenso foram construídas com as bibliotecas ggtree (YU et al., 2017) e treeio (WANG et al., 2020), através de scripts em R (R CORE TEAM, 2014).

## 3.4 Representações numéricas

O conjunto de sequências selecionadas foram utilizadas para obtenção de representações numéricas com o uso de k-mer e as 4 representações com suas variantes, testadas neste trabalho.

### 3.4.1 K-mer

Considerando uma sequência arbitrária s de tamanho L, s = '$N_1N_2…N_L$', em que $N_l \in \{A,C,G,T\}$ e l = 1, 2, ..., L, um k-mer pode ser definido como uma subpalavra de k nucleotídeos consecutivos dentro desta sequência. Dado um k arbitrário, temos $4^k$ possíveis permutações de k-mer.

### 3.4.2 'Triplet Frequency'

No trabalho de Kirichenko et. al. esta representação numérica foi utilizada para comparação entre métodos livres de alinhamento e baseados em alinhamento múltiplo e demonstrou que, mesmo sendo uma codificação simples, possui base promissora para ser utilizada juntamente com os métodos tradicionais no estudo de eventos evolutivos dramáticos em sequências de coronavírus (KIRICHENKO et al., 2022). Esta medida consiste na lista de contagem da ocorrência de cada um dos 64 trímeros (3-mer) possíveis, de AAA à TTT, medidos com uma janela deslizante por todo o

genoma, deslocando-se três nucleotídeos por vez até o fim da sequência, isto é, não há sobreposição na contagem dos 3-mer. Para uso, essa contagem foi transformada em frequência, dividindo-se cada valor pelo comprimento total da sequência genômica.

### 3.4.3 'K-mer Natural Vector'

O k-mer natural vector foi desenvolvido por Wen et. al. e é uma representação que codifica tanto a composição dos k-mers em uma determinada sequência, como sua relação da posicional (WEN et al., 2014). Esta representação é composta pela concatenação de três vetores que carregam três propriedades diferentes dos k-mers numa dada sequência.

O primeiro vetor desta representação consiste na contagem absoluta dos k-mers numa dada sequência s: $n_{(s,k)} = (n_{s[1]}, n_{s[2]}, \ldots, n_{s[4^K]})$.

O segundo vetor codifica a distância média de cada k-mer $\mu_{(s,k)} = (\mu_{[1]},\mu_{[2]},\ldots\mu_{[4^k]})$, em que cada $\mu_{[i]}$ sendo a média das distâncias de cada k-mer [i] à primeira base da sequência. Se um dado k-mer [i] não ocorre na sequência, $\mu_{[i]}$ é definido como zero.

Por último, o terceiro vetor é o momento central normalizado $D_{(s,k)} = (D_m^{[1]}, D_m^{[2]}, \ldots, D_m^{[4^k]})$. Para um dado momento m, este vetor geralmente é definido como:

$$D_m^{[i]} = \sum_{j=1}^{n[i]} \frac{(s[i][j]-\mu[i])^m}{(L-K+1)^{m-1}},$$

com s[i][j] sendo a distância do do j-ésimo k-mer [i] à primeira base.

Todos os valores para os k-mer são medidos com uma janela deslizante por todo o genoma, deslocando-se um nucleotídeo por vez até o fim da sequência. No total, o vetor numérico é composto por $3 \times 4^k$.

Esta representação demonstrou potencial para inferir filogenias com classificações taxonômicas corretas de genomas mitocondriais de animais (WEN et al., 2014) e, recentemente um outro estudo demonstrou que genomas de SARS-CoV-2 poderiam ser comparados com esta codificação, utilizando como momento central m = 2 (PEI; YAU, 2021). Neste trabalho testamos k-mer com k = 4, 6 e momento central m = 2, referidos posteriormente como 4-mer natural vector e 6-mer natural vector, respectivamente.

### 3.4.4 ‘Cumulative K-mer Natural Vector’

Testamos uma representação variada da definida acima em que não utilizamos somente o K-mer para calcular os três vetores ($n_{(s,k)}$, , $\mu_{(s,k)}$ e $D_{(s,k)}$) mas também os 1-mer, 2-mer, … e (K-1)-mer. No total, o vetor numérico é composto por $3 \times \sum_{i=1}^{k} 4^i$ valores. Similarmente, testamos para K = 4, 6 e momento central m=2.

### 3.4.5 ‘Fast Vector’

O fast vector (LI et al., 2017) divide as quatro bases em dois grupos, dentro de três propriedades físico-químicas diferentes. Na primeira, agrupam-se A e G, purinas, e C e T, pirimidinas. Na segunda, A e C; e G e T; são divididas em amino e ceto, respectivamente. Na última, as bases são divididas pelo número de ligações de hidrogênio que fazem, sendo A e T duas ligações e G e C 3 ligações. Para cada uma destas três classes de propriedades, é calculado o número de ocorrências, a média posicional com relação a primeira base de cada sequência e a variação de cada média posicional, de forma análoga ao k-mer natural vector descrito acima.

Optamos por testar essa codificação numérica pois seus autores demonstraram eficiência em reproduzir clados para genomas virais semelhantes aos obtidos pelas técnicas padrões de filogenia (LI et al., 2017) e, além disso, essa métrica combina as propriedades citadas acima com a estratégia dos k-mer natural vectors de codificar as relações posicionais dos nucleotídeos.

### 3.4.6 ‘Magnus Genomic Representation’

Esta codificação, também baseada em frequência de k-mer, atenta-se por carregar informação dos sub k-mers e de sua posição ao longo do genoma (WU et al., 2019). Nesta representação, uma sequência genética é dividida em N-mers, sem sobreposição, e, para cada N-mer, são computados a ocorrência dos k-mers, com k = 1, 2, …, N. A quantidade dos sub k-mers é colocada em ordem lexicográfica, por exemplo, para o caso em que N = 2, as quantidades seriam ordenadas pelo número de ocorrências de A, C, G, T, AA, AC, AG, AT, CA, CC, CG, CT, GA, GC, GG e GT, respectivamente. No final, uma média de cada ocorrência em todas as janelas de N nucleotídeos sem sobreposição no genoma é gerada. Utilizamos uma janela de 5 nucleotídeos, conforme recomendado pelos autores e testado para genomas virais e

bacterianos (WU et al., 2019). Posteriormente vamos nos referir a esta representação como ‘magnus representation’.

### 3.5 Medidas patrísticas, de morfologia de árvore filogenética e de distância entre as sequências

Antes do cálculo de distância entre as sequências, normalizamos todas as representações numéricas supracitadas pelo maior e menor valor de cada dimensão, em cada representação. Assim, a distância entre duas sequências foi calculada pela distância euclidiana entre os dois vetores numéricos normalizados destas.

Para comparar estas distâncias às geradas pelos métodos dependentes de alinhamento, utilizamos a biblioteca de R Castor (LOUCA; DOEBELI, 2018) para extrair da árvore filogenética gerada pelo IQTREE duas distâncias patrísticas: a soma do comprimento dos ramos e o número de nós conectando duas folhas da árvore.

Nós extraímos medidas morfológicas das árvores geradas com a biblioteca de R phyloTop (KENDALL; BOYD; COLIJN, 2023), sendo:

- Average ladder: tamanho médio das escadas na árvore, sendo escada definida como uma série de nós internos conectados com uma folha descendente;
- Cherry Number: número de ‘cerejas’ em uma árvore, sendo uma cereja considerada um par de pontas irmãs;
- IL number: número de nós internos com um único filho de ponta;
- Maximum tree height: Altura máxima da árvore;
- Number of pitchforks: um pitchfork é considerado um clado com três pontas;
- Colless number: calculada com base na soma dos valores absolutos da diferença do tamanho dos clados filhos em cada nó da árvore. Maiores valores representam árvores com maior desbalanceamento(HEARD, 1992);
- Sackin Index (SACKIN, 1972): calculado com base na soma do número de ancestrais para cada ponta da árvore. Quanto menos balanceada, maior é o índice de Sackin de uma árvore;
- Stairs (stair 1 e stair 2): definidas em (NORSTRÖM et al., 2012), sendo stair 1 a porção de sub-árvores que são desequilibradas e stair 2 sendo

a média do mínimo sobre o máximo do número de pontas descendentes dos lados esquerdo e direito em um nó, para todos os nós internos de uma árvore.

### 3.6 Técnicas de redução de dimensionalidade

Optamos por utilizar técnicas de redução de dimensionalidade por algumas razões: i) com elas podemos visualizar nosso dado de alta dimensionalidade em duas dimensões, mas preservando aspectos e relações importantes da natureza das amostras que compõe o nosso conjunto de dados, facilitando assim a compreensão dos resultados, ii) a aplicação desses métodos pode evitar sobreajustes, lidando com ruídos e redundância nos atributos, iii) para aplicações em análise de microbiomas, foi visto que algumas dessas transformações podem incorporar informações filogenéticas do dado original (ARMSTRONG et al., 2022), e iv) foi demonstrado que essas técnicas podem ser úteis em tarefas de agrupamento de sequências de coronavírus e especificamente de SARS-CoV-2, formando clusters que carreiam informação biológica (HOZUMI et al., 2021; KIRICHENKO et al., 2022).

Como essas técnicas possuem características e particularidades diferentes, neste trabalho exploramos três diferentes métodos (descritos abaixo) para visualizar e avaliar as representações numéricas dos genomas virais:

- PCA (análise de componentes principais) é um método estatístico popular usado em estudos exploratórios, com o objetivo de encontrar as direções da variância máxima em dados de alta dimensão e projetá-los em um novo subespaço, os chamados componentes principais, que são algumas combinações lineares das variáveis originais (GREENACRE et al., 2022). Uma das vantagens dessa técnica é a interpretabilidade dos componentes resultantes e a manutenção da estrutura global do dado.
- MDS (MultiDimensional Scaling) é uma técnica usada para analisar dados focando em sua similaridade ou dissimilaridade (distâncias), tentando remodelar os dados de modo a preservar essas distâncias em espaços geométricos com menores números de dimensões (BORG; GROENEN, 2005). Adicionamos esta técnica, pois os resultados costumam manter tanto a estrutura local quanto as estruturas globais do dado.
- UMAP (uniform manifold approximation and projection) é uma técnica desenvolvida recentemente, muito aplicada no campo de análise de single-cell

RNA Sequencing (BECHT et al., 2019), que representa os dados de alta dimensão em baixa dimensão, sem restrições de relações lineares entre os dados (diferentemente das técnicas citadas acima) e que pode preservar as tanto as estruturas locais e globais dos dados (https://umap-learn.readthedocs.io/en/latest/parameters.html, acesso em 2 de fevereiro de 2023), dependendo da escolha dos parâmetros de seu algoritmo (MCINNES; HEALY; MELVILLE, 2018).

### 3.7 Construção e avaliação das árvores utilizando as distâncias entre as representações numéricas

Construímos as árvores a partir das distâncias euclidianas dos dados normalizados dos genomas, utilizando o algoritmo neighbor-joining (NJ) implementado na biblioteca ape (PARADIS; SCHLIEP, 2019), com 1000 repetições de bootstrap utilizando a mesma biblioteca. Optamos por utilizar o algoritmo NJ pois este foi criado para trabalhar com distâncias, é menos sensível à escolha da medida de dissimilaridade, e foi amplamente utilizado porque provê medidas acuradas dos comprimentos dos ramos e é eficiente para lidar com grandes quantidades de dados (SAITOU, N., & NEI, 1987; PEER; SALEMI, 2003).

Para avaliarmos quantitativamente as árvores geradas, utilizamos quatro critérios: i) comparamos a distribuição de valores de bootstrap para cada caso, ii) medimos a correlação cofenética entre as sequências nas árvores geradas e as distâncias entre as sequências no espaço de distâncias originais, iii) a correlação cofenética e iv) a distância de Robinson-Foulds (RF) entre as árvores geradas sem alinhamento e a gerada pelo alinhamento acurado seguida de estimação filogenética por máxima-verossimilhança, implementados, respectivamente, nas biblioteca de R dendextend (GALILI, 2015) e phangorn (SCHLIEP, 2011).

A correlação cofenética (SOKAL; MICHENER, 1985) é um método utilizado para medir a similaridade entre dois diferentes agrupamentos hierárquicos e baseia-se na comparação entre a distância entre duas sequências na árvore e a distância observada entre as sequências no espaço de distância original. Como esta é sensível aos comprimentos dos ramos, utilizamos a distância RF (ROBINSON; FOULDS, 1981) que, por sua vez, é baseada no número de bipartições (splits) que as árvores compartilham. Quanto menor a distância RF, mais similares são as árvores.

As árvores e tanglegramas foram criados com os pacotes de R ggtree (YU et al., 2017), ggpmisc (APHALO; SLOWIKOWSKI; MOUKSASSI, 2022) e tidyverse (WICKHAM et al., 2019).

### 3.8 Comparação com nova ferramenta para estimação de distância genética

Foi utilizada a ferramenta Phylonium (https://github.com/EvolBioInf/phylonium), com os parâmetros padrão, para estimar distâncias evolutivas para o conjunto de genomas do Brasil. Esta ferramenta foi escolhida pois, em um benchmark extensivo, foi visto que ela forneceu as filogenias mais próximas entre genomas bacterianos completos, com distâncias RF próximas a 0,04 (ZIELEZINSKI et al., 2019). Embora não haja testes na literatura com genomas virais, acreditamos que a Phylonium ainda será eficaz para nossos dados, tendo em vista a proximidade dos genomas de SARS-CoV-2. A Phylonium foi desenvolvida para lidar com genomas próximos evolutivamente e, embora seja considerada uma ferramenta "alignment-free", sua estratégia envolve a criação de alinhamentos locais com base em âncoras exatamente idênticas de tamanho máximo entre as sequências. Os mismatches flanqueando estas regiões são então considerados polimorfismos e utilizados para calcular o número final de substituições por sítio (KLÖTZL; HAUBOLD, 2020).

### 3.9 Disponibilidade dos códigos utilizados nas análises

As linguagens Python (PYTHON CORE TEAM, 2015), R e shell script foram usadas para executar todas as análises e criar as figuras deste trabalho. Os códigos escritos na linguagem Python para a representação numérica de sequências genômicas e notebooks em linguagem Python, Shell e R contendo os passos de execução deste projeto estão disponíveis publicamente através do Github (https://github.com/MuriloACassiano/Viral_Genomic_Numeric_Representations ).

# 4 Resultados e Discussão

## 4.1 Testes quantitativos de métodos para comparar sequências genômicas de Orthocoronavirinae usando um conjunto de sequências públicas

### 4.1.1 Análise filogenética de 69 sequências alinhadas

A reconstrução filogenética por máxima verossimilhança dos 69 genomas completos de coronavírus resultou em uma árvore filogenética com alto suporte estatístico de *bootstrap* ao longo de todos os nós internos da árvore, exceto pelo nó único que separa os alfacoronavírus dos demais (Figura 1). Os quatro gêneros de coronavírus alfa, beta, delta e gama, respectivamente com 20, 24, 6 e 8 genomas cada, foram agrupados corretamente. Dos 9 genomas sem classificação, 4 se agruparam com os alfacoronavírus e 5, com os betacoronavírus (Figura 1A). De forma semelhante, ao marcarmos o hospedeiro principal de onde cada genoma foi extraído, vemos que os subclados foram formados adequadamente (Figura 1B). Assim, prosseguimos a seguir com as comparações de medidas evolutivas desta árvore com as métricas derivadas das comparações das representações numéricas.

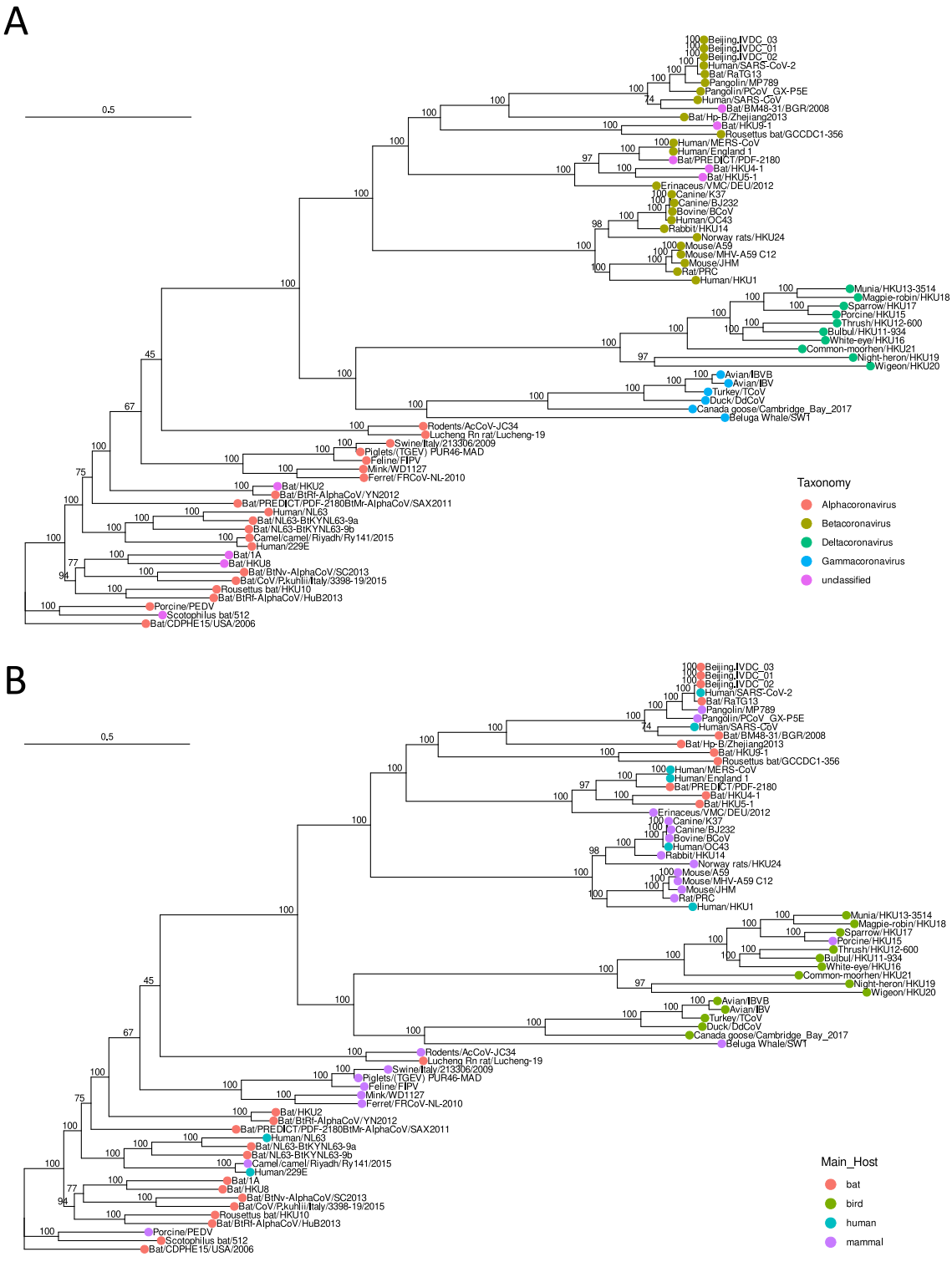



**Figura 1 -** Árvore filogenética de máxima-verossimilhança das 69 sequências genômicas de orthocoronavirinae. A figura acima está colorida para ilustrar o nível taxonômico do genoma. A figura abaixo utiliza as cores para a representação do principal hospedeiro do qual os genomas foram isolados.

### 4.1.2 Redução da dimensionalidade dos vetores numéricos

Antes de prosseguirmos para comparações entre as distâncias patrísticas da árvore filogenética e a distância euclidiana entre as representações numéricas, nós quisemos avaliar a capacidade de cada representação para formar corretamente agrupamentos em que seus componentes sejam membros de um mesmo nível taxonômico. Para isso, avaliamos a PCA (e a variância representada em duas dimensões), o MDS (e a sua medida de stress), e três execuções do UMAP, variando o número de vizinhos considerados pelo cálculo do algoritmo para 68, 30 e 17 vizinhos, de modo que tanto características globais como locais da estrutura do dado pudessem ser capturadas. As métricas da PCA e do MDS estão sumarizadas na tabela abaixo (Tabela 1).

**Tabela 1** - Número de atributos, valores de variância contidos nos 2 componentes principais da PCA e valor do strees gerado pelo MDS.

| Representação | Número de atributos | PCA 1 | PCA 2 | Variância em 2D | Stress do MDS |
|---|---|---|---|---|---|
| Fast | 18 | 0,90 | 0,05 | 0,95 | 11,17 |
| Triplets | 64 | 0,35 | 0,25 | 0,60 | 582,16 |
| 4-mer Natural Vector | 768 | 0,30 | 0,10 | 0,40 | 13.244,51 |
| Cumulative 4-mer Natural Vector | 1.020 | 0,37 | 0,10 | 0,47 | 14.934,83 |
| Magnus | 1.364 | 0,14 | 0,10 | 0,24 | 35.130,13 |
| 6-mer Natural Vector | 12.288 | 0,08 | 0,07 | 0,15 | 436.675,61 |
| Cumulative 6-mer Natural Vector | 16.380 | 0,10 | 0,08 | 0,18 | 527.662,05 |

Analisando os mapas em 2D dos genomas codificados na forma de triplets podemos notar tanto na PCA como no MDS que nenhum grupo taxonômico foi separado totalmente, e não se pôde notar nenhum padrão específico sendo formado com relação ao hospedeiro principal. No entanto, é possível notar somente uma tendência de separação dos genomas de alfacoronavírus com a UMAP de 17 vizinhos (Figura 2, última coluna).


**Figura 2** - Mapas em 2D das reduções de dimensionalidade da representação numérica triplets. Na linha superior os pontos representando os genomas estão coloridos pela taxonomia e na inferior por seu hospedeiro principal. Note que para a PCA temos no topo dos gráficos as quantidades de variância que representam cada dimensão e nos gráficos de MDS temos o valor total de estresse.

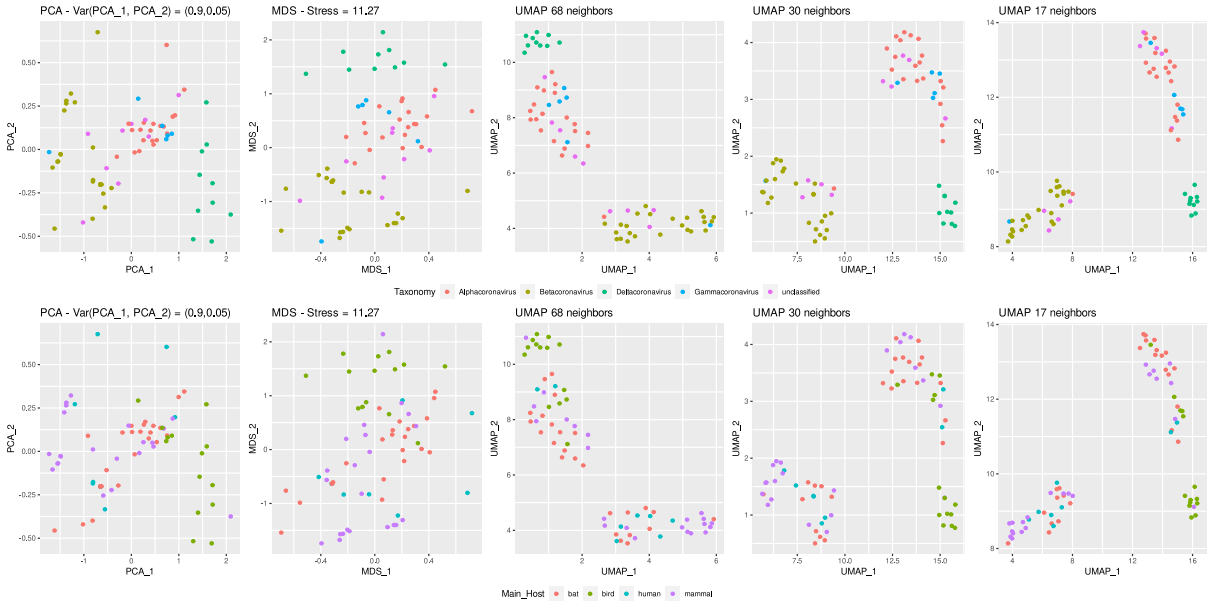



**Figura 3** - Mapas em 2D das reduções de dimensionalidade da representação numérica fast vector. Na linha superior os pontos representando os genomas estão coloridos pela taxonomia e na inferior por seu hospedeiro principal. Note que para a PCA temos no topo dos gráficos as quantidades de variância que representam cada dimensão e nos gráficos de MDS temos o valor total de estresse.

Já para o Fast vector, vemos separações entre alfa-, beta- e deltacoronavírus em todas as técnicas de redução de dimensionalidade (Figura 3). Vale destacar que os genomas com classificação taxonômica desconhecida ficaram agrupados quatro com os alfas e cinco com os betacoronavírus, ao contrário do que vimos com a árvore filogenética de referência (Figura 1). Também vale destacar que, em duas dimensões, a transformação por PCA conseguiu abranger 95% da variabilidade do dado (Tabela 1).

Com a representação de Magnus (Figura 4), o que pôde ser observado foi um comportamento de dispersão dos pontos representando os genomas. Nenhuma tendência de formação de clusters foi visualizada nos mapas 2D de nenhuma técnica de redução de dimensionalidade para esta representação, como visto para o triplet (Figura 2) e mais notavelmente para o Fast vector (Figura 3).

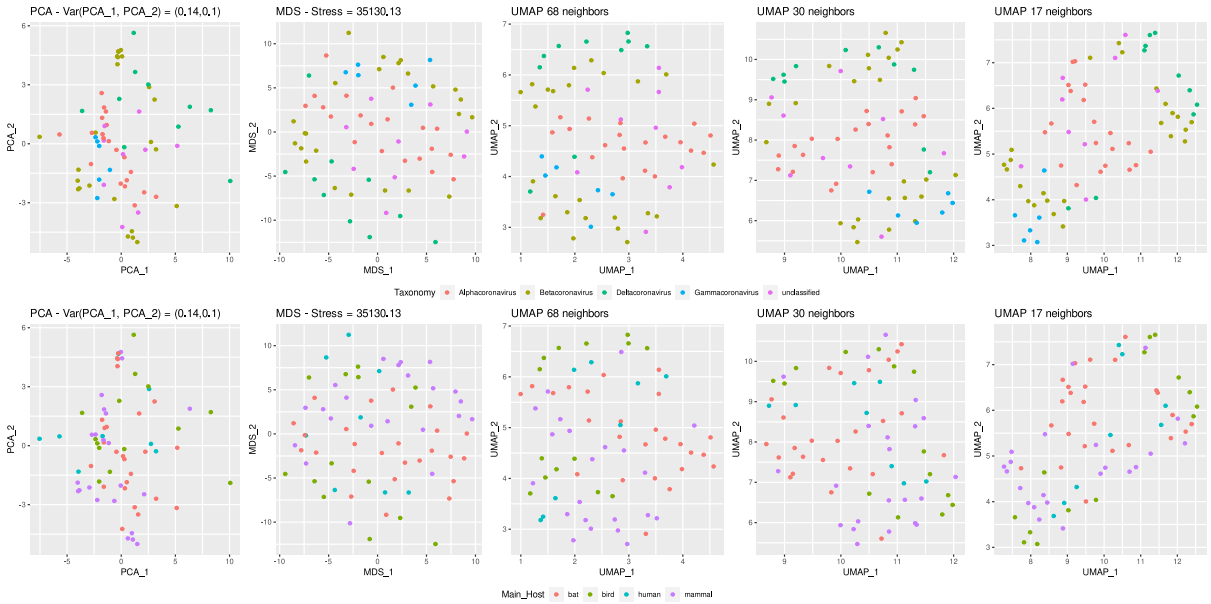



**Figura 4** - Mapas em 2D das reduções de dimensionalidade da representação numérica magnus representation. Na linha superior os pontos representando os genomas estão coloridos pela taxonomia e na inferior por seu hospedeiro principal. Note que para a PCA temos no topo dos gráficos as quantidades de variância que representam cada dimensão e nos gráficos de MDS temos o valor total de estresse.

Nos mapas 2D para o 4-mer natural vector, vemos uma separação clara por nível taxonômico nos três gráficos UMAP e uma distribuição no gráfico do MDS semelhante à obtida com a magnus representation (Figura 5).

O cumulative 4-mer natural vector, por sua vez, quando mapeado em duas dimensões, demonstrou a formação de grupamentos em todos os casos e um padrão de separação mais consistente com os 4 gêneros de coronavírus presentes nas

sequências (Figura 6). Interessantemente, vemos maior quantidade de variância na PCA nesta forma cumulativa do 4-mer natural vector, mesmo possuindo maior número de dimensões quando comparada com a sua forma original (Tabela 1). Vale ressaltar que, tanto para o cumulative 4-mer natural vector como para o 4-mer natural vector, podemos observar uma tendência na formação de dois clusters para betacoronavírus, como visto na árvore filogenética. Adicionalmente, vemos que os clusters formados pelo UMAP parecem ter relação não somente com o grupo taxonômico, como também pelo hospedeiro principal de cada vírus.

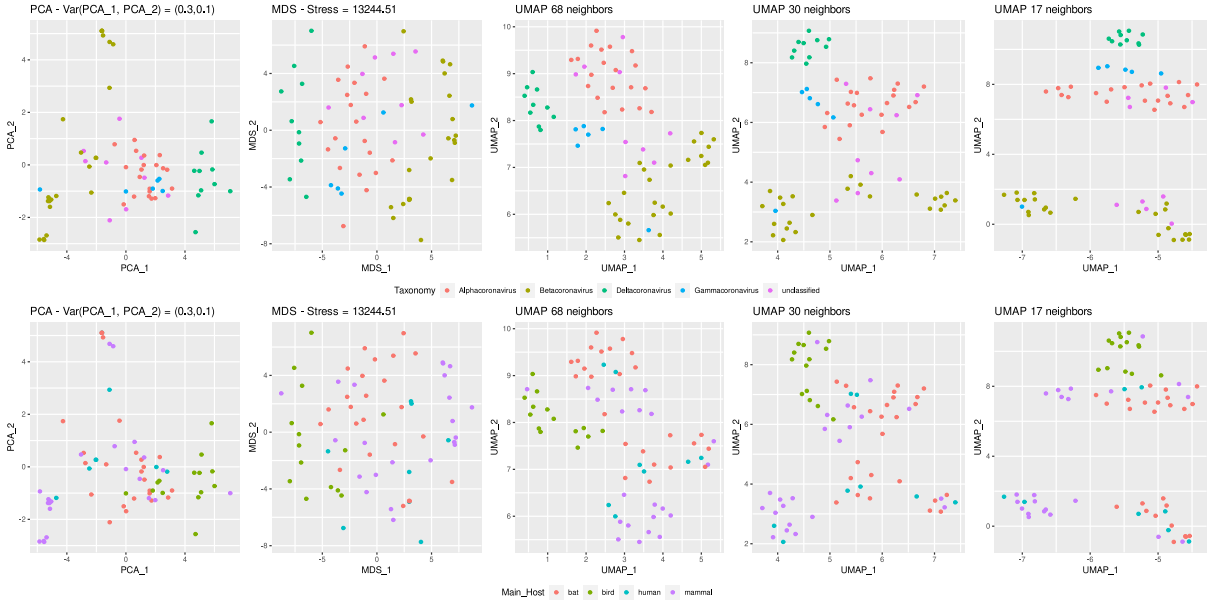



**Figura 5** - Mapas em 2D das reduções de dimensionalidade da representação numérica 4-mer natural vector. Na linha superior os pontos representando os genomas estão coloridos pela taxonomia e na inferior por seu hospedeiro principal. Note que para a PCA temos no topo dos gráficos as quantidades de variância que representam cada dimensão e nos gráficos de MDS temos o valor total de estresse.

Este último achado fica ainda mais evidente para o caso do 6-mer natural vector (Figura 7) e do cumulative 6-mer-natural vector (Figura 8). Contudo, nesses casos, o grupo dos gamacoronavírus fica mais evidentemente separado dos alfacoronavírus do que visto em todas as representações numéricas anteriores, especialmente com a transformação UMAP, sobretudo para o caso de 17 vizinhos.

Tendo em mente estes resultados, concluímos que, mesmo o Magnus possuindo um alto número de atributos (dimensões), com as técnicas de redução de dimensionalidade testadas, sua capacidade de separação assemelha-se à da representação em teoria mais simplista de triplets. As representações de natural k-mer e fast vector demonstraram-se úteis para separação por gênero de coronavírus

e, em especial, os natural k-mers parecem relacionar-se também à informação do hospedeiro principal dos genomas virais.

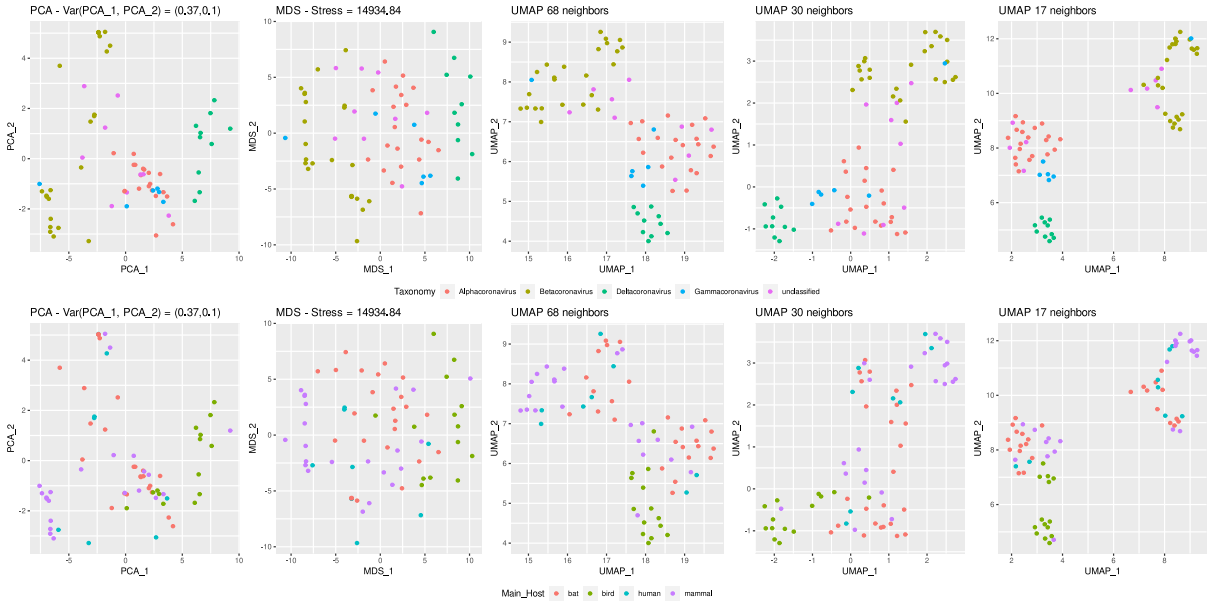



**Figura 6** - Mapas em 2D das reduções de dimensionalidade da representação numérica cumulative 4-mer natural vector. Na linha superior os pontos representando os genomas estão coloridos pela taxonomia e na inferior por seu hospedeiro principal. Note que para a PCA temos no topo dos gráficos as quantidades de variância que representam cada dimensão e nos gráficos de MDS temos o valor total de estresse.

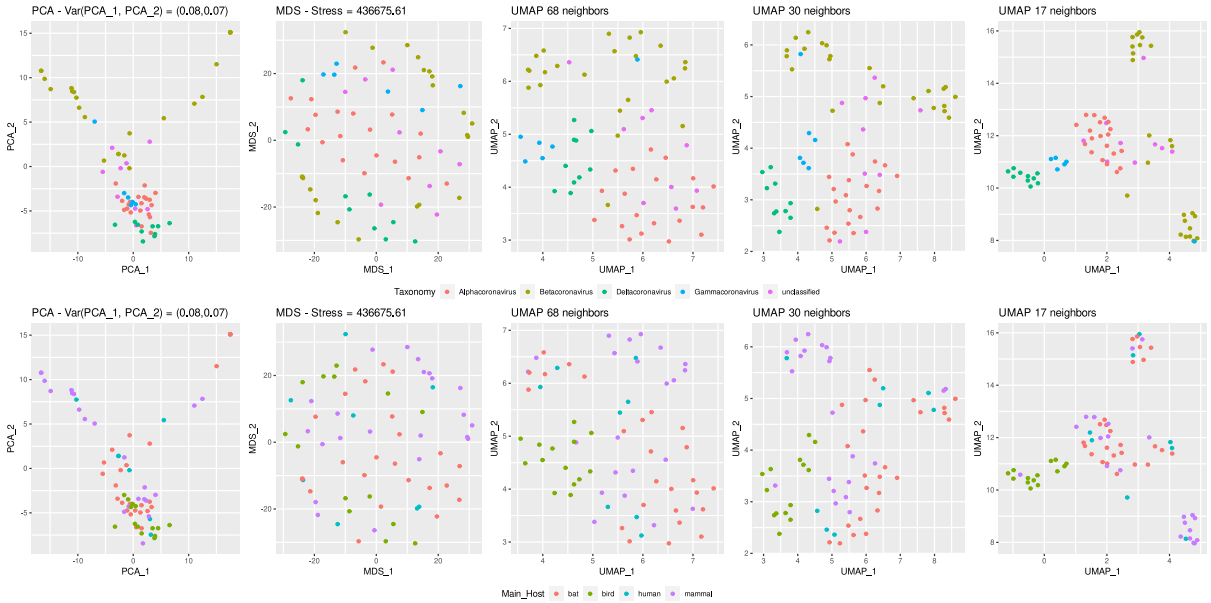



**Figura 7** - Mapas em 2D das reduções de dimensionalidade da representação numérica 6-mer vector. Na linha superior os pontos representando os genomas estão coloridos pela taxonomia e na inferior por seu hospedeiro principal. Note que para a PCA temos no topo dos gráficos as quantidades de variância que representam cada dimensão e nos gráficos de MDS temos o valor total de estresse.

Os números de atributos e, por consequência, o custo das representações baseadas em k-mer natural vector, sugerem que as métricas 4-mer natural vector, cumulative 4-mer natural vector e 6-mer natural vector possuem potencial para capturar informação biológica relevante das sequências, visto que o cumulative 6-mer natural vector não demonstrou melhores distinções dos genomas em seus grupos em relação a sua versão não cumulativa.

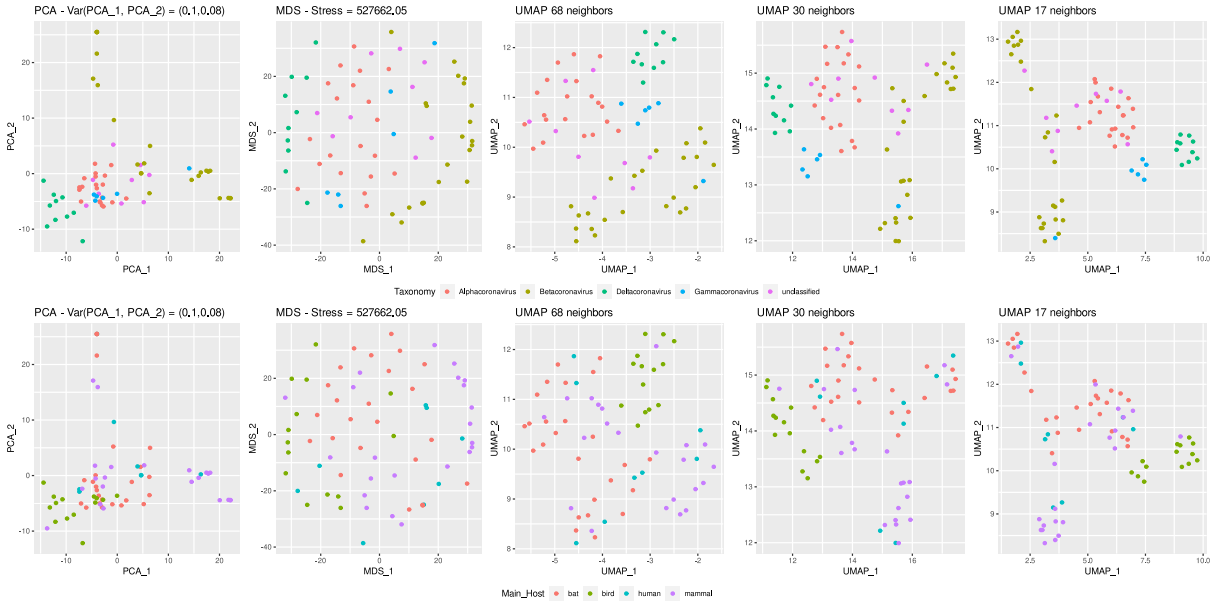



**Figura 8** - Mapas em 2D das reduções de dimensionalidade da representação numérica cumulative 6-mer natural vector. Na linha superior os pontos representando os genomas estão coloridos pela taxonomia e na inferior por seu hospedeiro principal. Note que para a PCA temos no topo dos gráficos as quantidades de variância que representam cada dimensão e nos gráficos de MDS temos o valor total de estresse.

### 4.1.3 Comparando as distâncias patrísticas às distâncias euclidianas das representações numéricas

Como a soma dos ramos que conectam dois táxons (aqui mencionados como patristic length) e o número de nós entre um táxon e outro (aqui mencionados como patristic edges) estão indiretamente relacionados com a morfologia de uma árvore filogenética, nós hipotetizamos que medindo a correlação entre as distâncias euclidianas calculadas entre as representações numéricas dos genomas e as distâncias patrísticas, podemos ter uma ideia da possível árvore filogenética formada

a partir das distâncias euclidianas e também da possível informação filogenética que estas representações podem carregar.

Na Figura 9, temos os valores das distâncias patrísticas nos eixos y e as distâncias euclidianas de cada representação numérica plotadas nos eixos x. Para cada caso, separamos os valores de distância conforme o pertencimento das duas amostras ao mesmo gênero de coronavírus ou a gêneros diferentes. Realizamos regressões lineares para avaliar o coeficiente de correlação $R^2$ entre as distâncias euclidianas e as patrísticas para cada caso.

Verificamos que não há relação entre as distâncias euclidianas e o número de nós conectando genomas de gêneros de coronavírus diferentes. Além disso, a correlação entre distâncias euclidianas e patristic edge conectando genomas de mesmo gênero é geralmente menor do que a correlação com patristic length (Figura 9, linha superior). Considerando somente as distâncias entre genomas dentro do mesmo gênero, o 4-mer natural vector e sua versão cumulativa apresentaram a melhor correlação ($R^2$ = 0,33). Para as relações entre as distâncias euclidianas e a patristic length, 4-mer e 6-mer natural vectors e suas versões cumulativas performaram melhor do que as outras três representações, alcançando um $R^2$ de 0,54 para genomas de mesmo grupo e 0,27 para genomas de grupos distintos.


**Figura 9** - Gráfico de dispersão em que nos eixos x estão representadas as distâncias euclidianas entre cada um dos valores de representação numérica dos genomas de orthocoronavirinae. No eixo y da primeira linha estão representadas as distâncias patristic lenght e na segunda linha estão representadas as distâncias patristic edge. Os pontos estão coloridos em ciano para representar quando as distâncias correspondentes são entre genomas pertencentes a níveis taxonômicos

diferentes e em vermelho quando os pontos representam a distância entre genomas de um mesmo nível taxonômico. Para cada caso a um $R^2$ e um p valor associado à regressão linear.

Similarmente, fazendo a mesma análise só que agrupando pelas distâncias entre genomas de mesmo grupo taxonômico, temos que somente os 4-mer e 6-mer natural vectors não falham na regressão, seja com o $R^2$ próximo a 0 ou com o p-valor da regressão maior que 0,05 (Figura 10). Notavelmente, as distâncias euclidianas e o patristic lenght dos gamas e betacoronavírus tiveram maiores valores de $R^2$, chegando a 0.98 para gama- e 0.64 para betacoronavírus. Novamente, as correlações com patristic length foram maiores do que as com patristic edge para todos os casos.

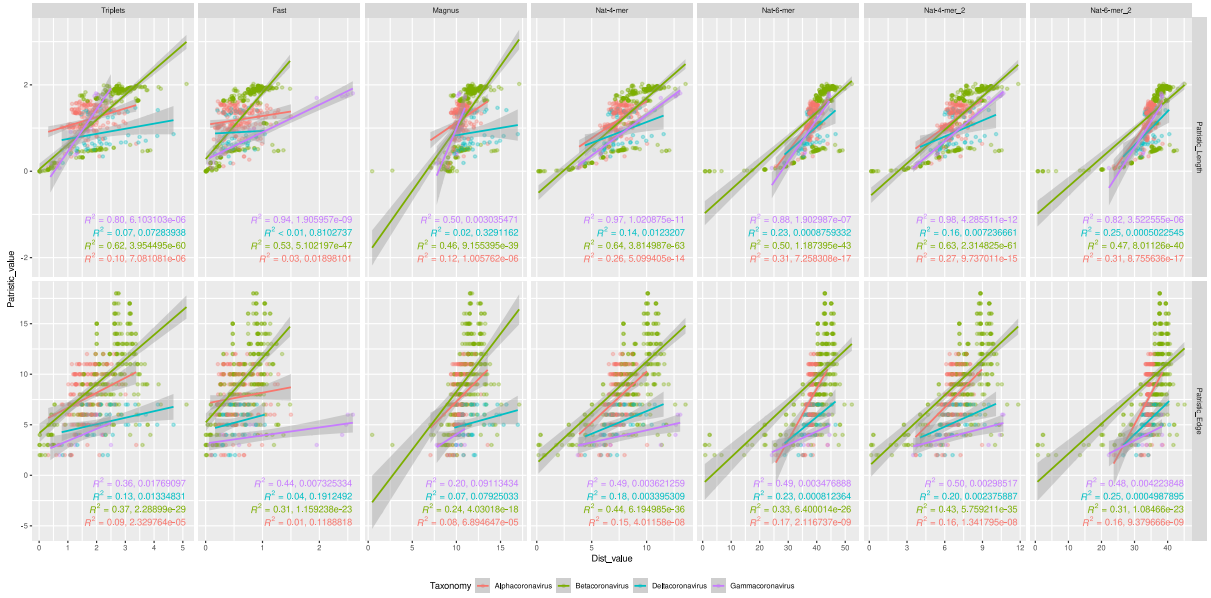



**Figura 10** - Este gráfico de dispersão é semelhante ao da figura anterior (figura 9), contudo, representamos os pontos com base nas distâncias entre genomas de grupo taxonômico semelhante.

Seguindo a mesma linha de raciocínio de (KIRICHENKO et al., 2022), também agrupamos as distâncias pelo hospedeiro principal de cada vírus, separando as distâncias de genomas dentro de um mesmo hospedeiro e entre hospedeiros diferentes (Figura 11) e em seguida, analisamos em detalhe os 4 tipos de hospedeiros (Figura 12).

Surpreendentemente, nas duas análises citadas acima, os métodos que se sobressaíram em correlação com as distâncias patrísticas foram 4-mer natural vector, cumulative 4-mer natural vector e fast vector. De forma geral, todas as distâncias entre as sequências agrupadas pelo hospedeiro correlacionaram melhor com as distâncias patrísticas do que as agrupadas anteriormente por taxonomia.

Adicionalmente, as representações de triplets e magnus falharam em algumas das regressões (seja pelo p-valor > 0,05 ou $R^2 < 0.1$). Os resultados sugerem que as distâncias obtidas pelos métodos de representação testados sirvam melhor para comparações de sequências com maior relação evolutiva entre si, seja dentro de um grupo taxonômico ou pertencentes a um mesmo hospedeiro. Nossos achados são corroborados pelo trabalho e Kirichenko et. al., em que todos os coeficientes de correlação intragrupo foram maiores que os intergrupo (KIRICHENKO et al., 2022).

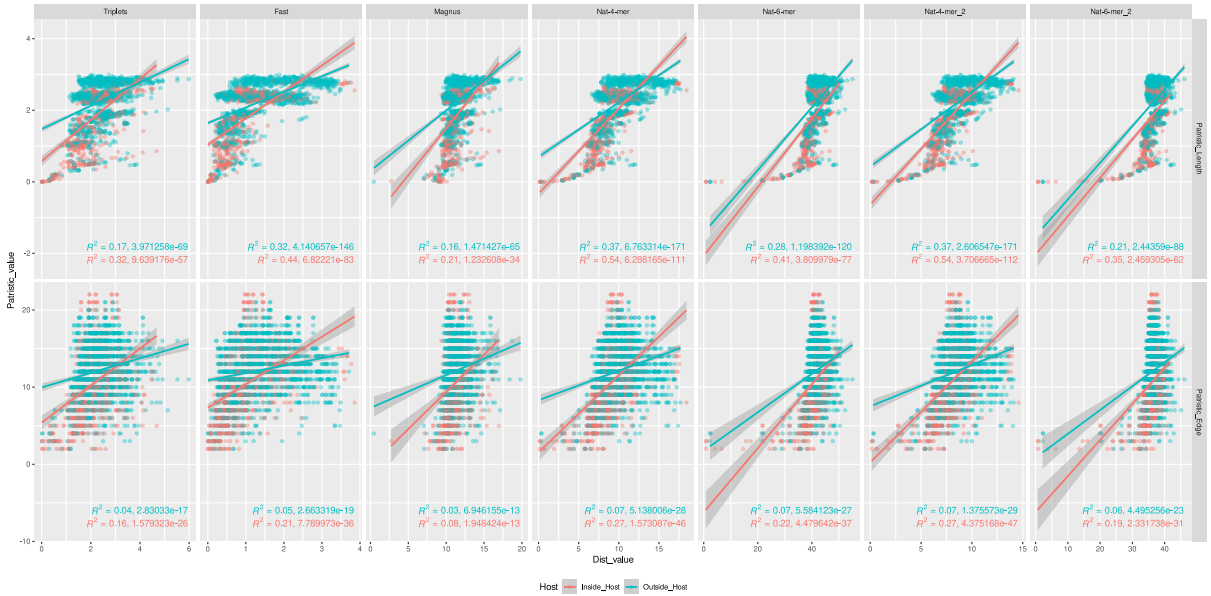



**Figura 11** - Gráfico de dispersão em que nos eixos x estão representadas as distâncias euclidianas entre cada um dos valores de representação numérica dos genomas de orthocoronavirinae. No eixo y da primeira linha estão representadas as distâncias patristic lenght e na segunda linha estão representadas as distâncias patristic edge. Os pontos estão coloridos em ciano para representar quando as distâncias correspondentes são entre genomas isolados de diferentes hospedeiros e em vermelho quando os pontos representam a distância entre genomas de um mesmo tipo de hospedeiro. Para cada caso a um $R^2$ e um p valor associado à regressão linear.

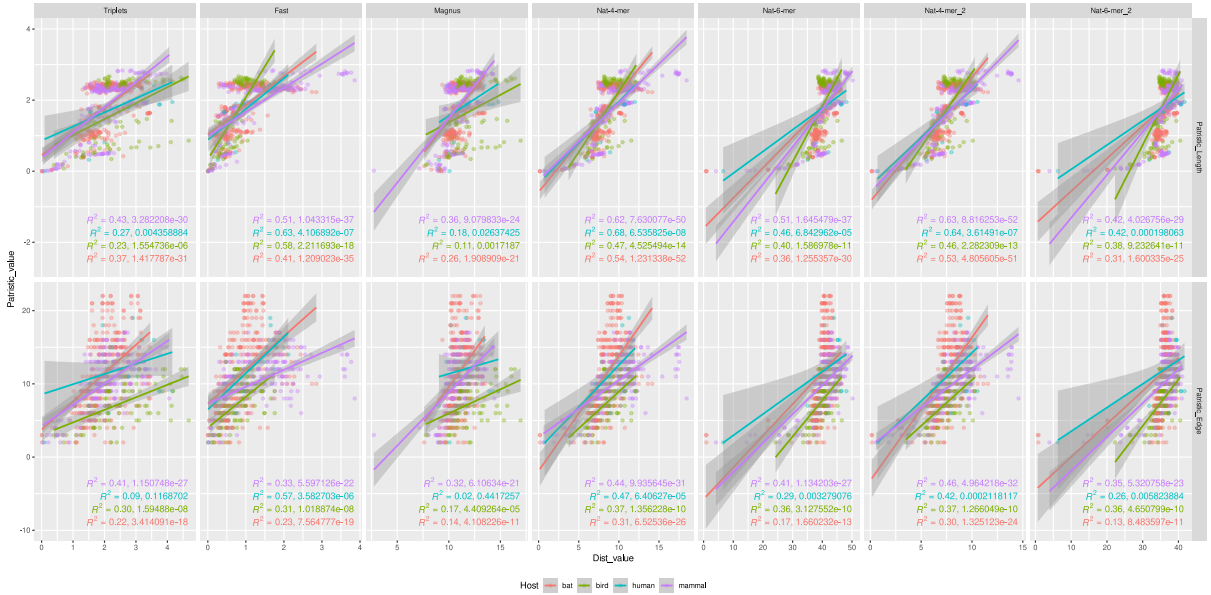



**Figura 12** - Este gráfico de dispersão é semelhante ao da figura anterior (Figura 11), contudo, representamos os pontos com base nas distâncias entre genomas que compartilham hospedeiro em comum.

### 4.1.4 Clusterização das sequências com base na distância euclidiana

O próximo passo foi avaliar o uso das distâncias pelo algoritmo NJ, com 1000 repetições de bootstrap, para construção de árvores filogenéticas putativas. A tabela a seguir apresenta as métricas de morfologia geral de estrutura, obtidas com cada árvore gerada (Tabela 2).

A árvore formada pela distância dos triplets (Figura 13) não foi capaz de recuperar o agrupamento entre as sequências de grupos taxonômicos semelhantes, apresentando desbalanço no comprimento dos ramos comparativamente menor do que a árvore de máxima-verossimilhança (menores valores de Colless e Sackin index. No entanto, ela apresenta uma maior proporção de subárvores desbalanceadas (stair1, Tabela 2). Os valores de suporte estatístico de bootstrap obtidos ficaram distribuídos uniformemente entre 100% e valores próximos a 0% (Apêndice 1).

**Tabela 2** - Valores de medidas morfológicas que sumarizam a estrutura das árvores filogenéticas analisadas nesta seção.

| | Average ladder | Cherry number | Colless Index | IL number | Altura máxima | Number of pitchforks | Sackin Index | Stairs 1 | Stairs 2 |
|---|---|---|---|---|---|---|---|---|---|
| Triplets | 3,375000 | 18 | 408 | 33 | 16 | 13 | 692 | 0,7205882 | 0,5393930 |
| Fast | 3,636364 | 13 | 881 | 43 | 29 | 7 | 1107 | 0,8088235 | 0,3569931 |
| Magnus | 4,000000 | 22 | 239 | 25 | 13 | 8 | 547 | 0,6176471 | 0,6174779 |
| Cumulative 4-mer nat vec 2 | 3,500000 | 16 | 476 | 37 | 18 | 14 | 748 | 0,7205882 | 0,5184127 |
| 4-mer nat. vec. | 3,000000 | 16 | 304 | 37 | 15 | 10 | 600 | 0,7500000 | 0,5203503 |
| 6-mer nat. vec. | 3,571429 | 19 | 233 | 31 | 14 | 11 | 541 | 0,6911765 | 0,5827255 |
| Cumulative 6-mer nat. vec. | 3,666667 | 19 | 423 | 31 | 16 | 11 | 701 | 0,6764706 | 0,5584302 |
| MAFFT + IQTREE2 | 4,000000 | 22 | 542 | 25 | 20 | 8 | 810 | 0,6470588 | 0,5746997 |

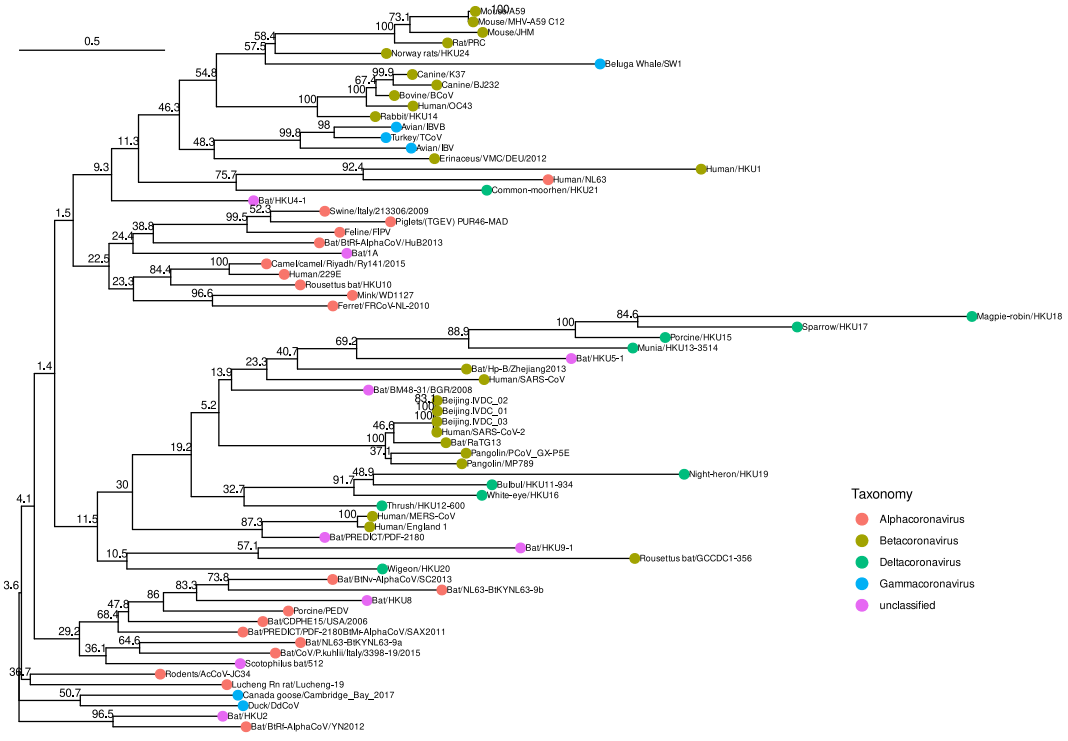



Figura 13 - Árvore gerada pelo algoritmo NJ a partir das distâncias euclidianas da representação numérica triplets. As cores representam cada grupo taxonômico e fizemos 1000 rodadas de bootstrap para aferir a confiabilidade estatística dos nós internos.

Notamos uma distribuição semelhante aos valores bootstrap para a árvore do fast vector, contudo, houve uma leve concentração de valores acima do 50% (Apêndice 1). Esta árvore (Figura 14) ficou visivelmente desbalanceada, apresentando os maiores índices de Colless e Sackin e maior valor de stair 1, apesar da média entre o mínimo e o máximo do número de pontas dos nós internos das árvores ser o menor entre todas as árvores geradas (stair 2) (Tabela 2). Isto se deve a todos os nós internos praticamente terem um nó filho que é folha e o outro ser um nó interno, levando a um padrão de escada com comprimentos de ramos curtos (Figura 14).

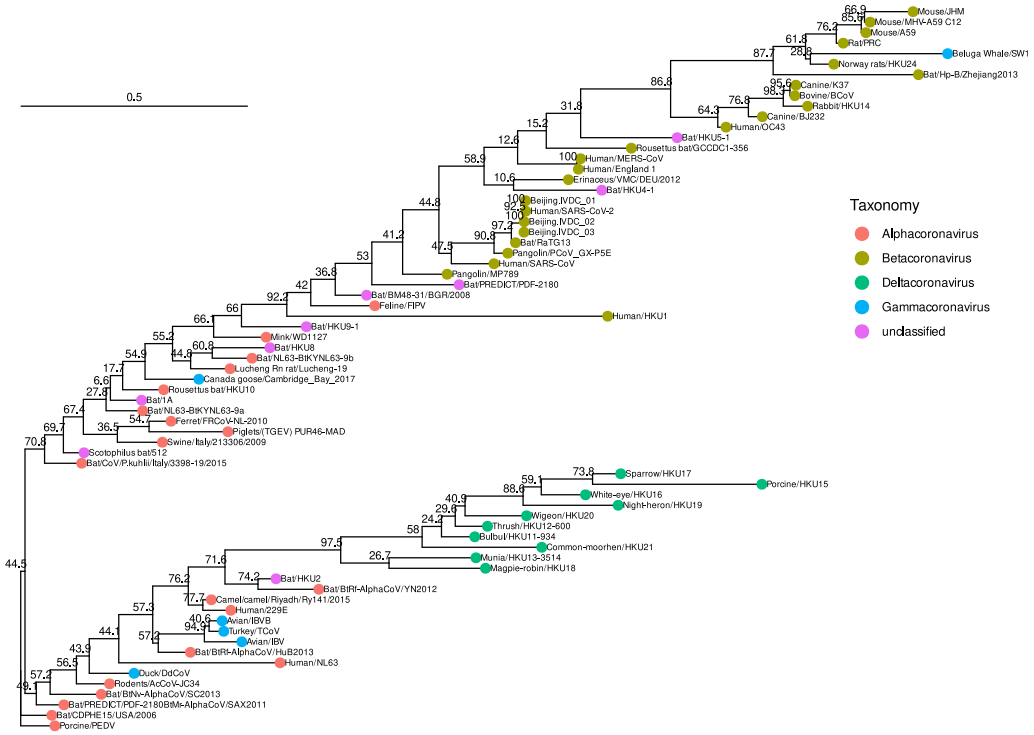



**Figura 14 -** Árvore gerada pelo algoritmo NJ a partir das distâncias euclidianas da representação numérica fast vector. As cores representam cada grupo taxonômico e fizemos 1000 rodadas de bootstrap para aferir a confiabilidade estatística dos nós internos.

Assim como na árvore dos triplets, na árvore do Magnus representation (Figura 15) não houve agrupamento algum dos genomas por taxonomia e os valores de bootstrap para os nós mais internos da árvore ficaram todos com valores abaixo de 50%. Curiosamente, vários valores que sumarizam a estrutura geral da árvore ficaram parecidos com os da árvore de máxima-verossimilhança (Tabela 2).

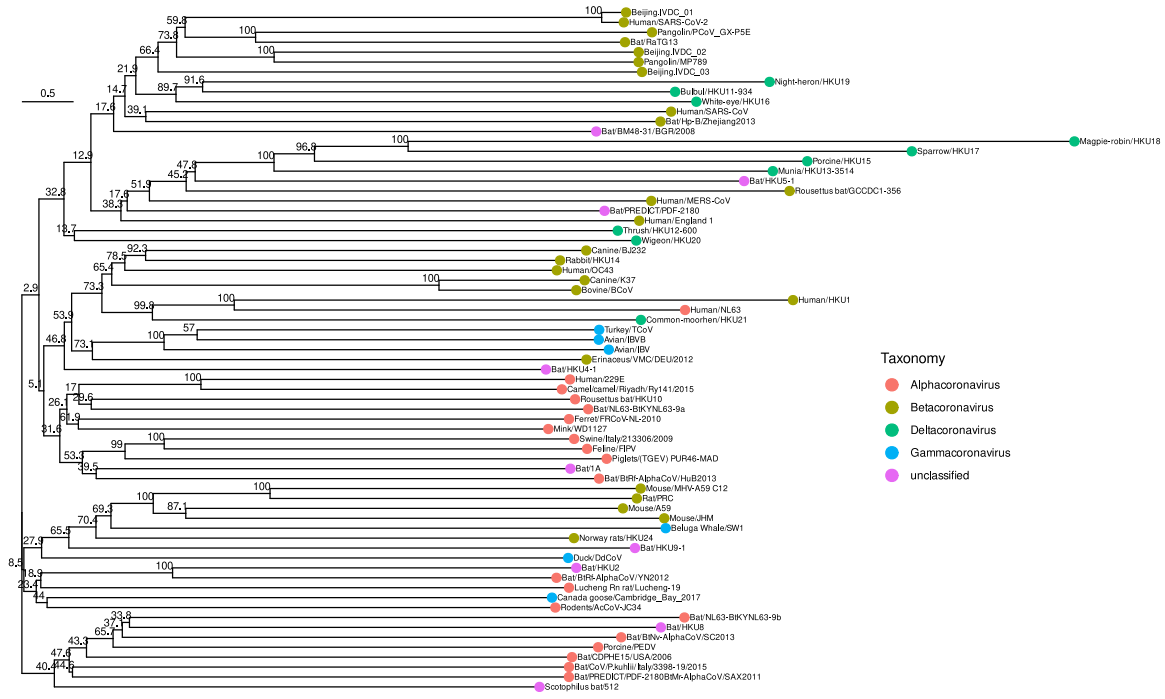



**Figura 15** - Árvore gerada pelo algoritmo NJ a partir das distâncias euclidianas da representação numérica magnus representation. As cores representam cada grupo taxonômico e fizemos 1000 rodadas de bootstrap para aferir a confiabilidade estatística dos nós internos.

Para a árvore dos 4-mer natural vectors (Figura 16) e de sua versão cumulativa (Figura 17), vimos que alfa, beta e deltacoronavírus formaram agrupamentos distintos, e, apesar do valor de bootstrap do nó que divergem os dois clados estarem próximos de 60%, os dois principais clusters de betacoronavírus foram formados, um contendo o SARS-CoV-2, SARS-CoV, MERS-CoV e os relacionados a SARS, e o outro contendo os coronavírus mais próximos do HKU1 e OC43 humanos.

Com relação às medidas que resumem a estrutura, o cumulative 4-mer natural vector gerou a árvore mais diferente da árvore gerada pelo IQTREE, enquanto o 4-mer natural vector gerou a mais similar (Tabela 2, Apêndice 2). Estas árvores apresentaram menos desbalanceamento em comparação com a nossa referência, porém mantiveram a estrutura de “escada” (ver Stairs 1 e 2 na tabela 2).


**Figura 16** - Árvore gerada pelo algoritmo NJ a partir das distâncias euclidianas da representação numérica 4-mer natural vector. As cores representam cada grupo taxonômico e fizemos 1000 rodadas de bootstrap para aferir a confiabilidade estatística dos nós internos.

A árvore dos 6-mer natural vectors (Figura 18) e de sua versão cumulativa (Figura 19) foram as únicas que melhor separaram o grupo dos gamacoronavírus. Como no caso das duas representações anteriores, os comprimentos dos ramos da árvore parecem estar mais uniformes e a estrutura geral, mais balanceada do que a da referência, ainda assim, preservando a estrutura de escada (stair 1 aproximadamente 0,7 - Tabela 2). Adicionalmente, vemos que para o cumulative 6-mer natural vector os valores de bootstrap foram os mais altos, com a concentração da distribuição em torno de 100% (Apêndice 1).

O oposto do que foi notado nos casos do 4-mer, foram as estatísticas descritivas de estruturas da versão cumulativa do 6-mer natural vector que melhor se assemelhou aos da árvore gerada por alinhamento com subsequente estimação de máxima-verossimilhança (Apêndice 2). É importante ressaltar também que somente na árvore do Magnus representation não foi possível ver o agrupamento próximo das

sequências de SARS-CoV-2 com o RATG13 e com os genomas de coronavírus de pangolins.


**Figura 17** - Árvore gerada pelo algoritmo NJ a partir das distâncias euclidianas da representação numérica cumulative 4-mer natural vector. As cores representam cada grupo taxonômico e fizemos 1000 rodadas de bootstrap para aferir a confiabilidade estatística dos nós internos.

**Tabela 3** - Medidas de correlação cofenética entre: i) a árvore gerada e sua própria distância, ii) a árvore gerada e a de máxima-verossimilhança de referência (Preferívies valores próximos a 1). Medidas de Distância RF entre a árvore gerada e a de máxima-verossimilhança de referência (Preferíveis valores próximos a 0).

| | Triplets | Fast | Magnus | 4-mer nat. vec. | Cumulat. 4-mer nat. vec. | 6-mer nat. vec. | Cumulat. 6-mer nat. vec. |
|---|---|---|---|---|---|---|---|
| Correlação cofenética com as próprias distâncias | 0,8521199 | 0,9817585 | 0,8609456 | 0,9129655 | 0,9082372 | 0,8869883 | 0,9039771 |
| Correlação Cofenética com a árvore Referência | 0,4486002 | 0,5952053 | 0,4295419 | 0,5567296 | 0,556154 | 0,3793219 | 0,4145955 |
| Distância de RF à árvore Referência | 0,6666667 | 0,8181818 | 0,8636364 | 0,5606061 | 0,5757576 | 0,4545455 | 0,4848485 |

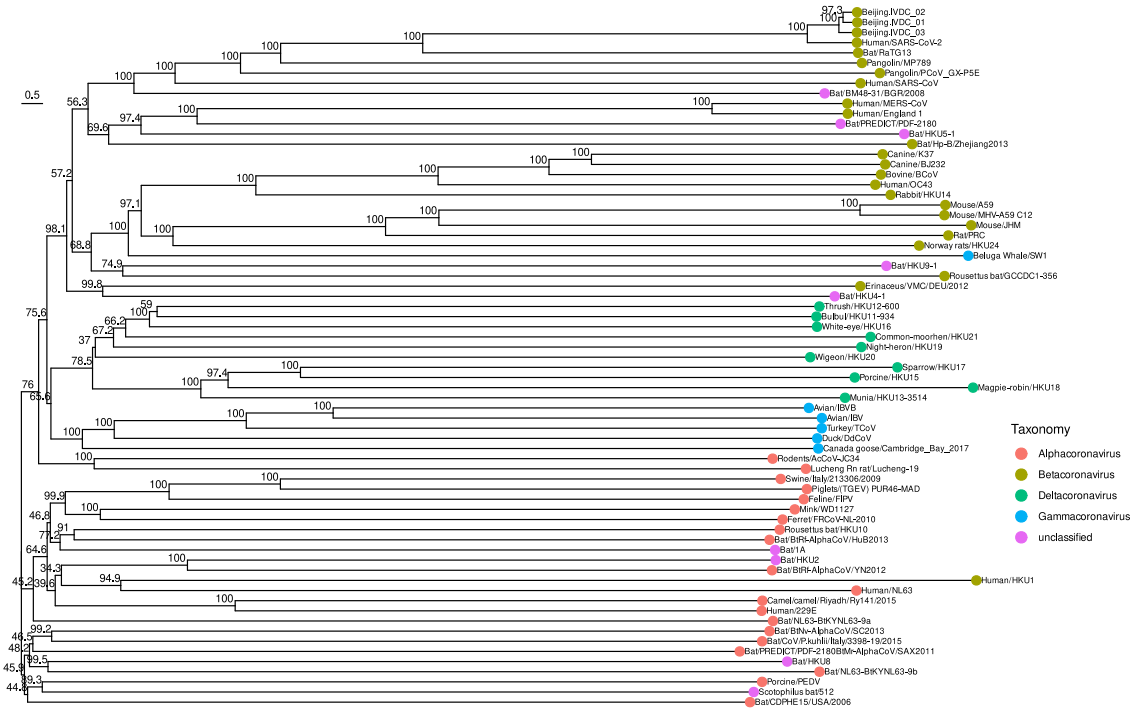



**Figura 18** - Árvore gerada pelo algoritmo NJ a partir das distâncias euclidianas da representação numérica 6-mer natural vector. As cores representam cada grupo taxonômico e fizemos 1000 rodadas de bootstrap para aferir a confiabilidade estatística dos nós internos.

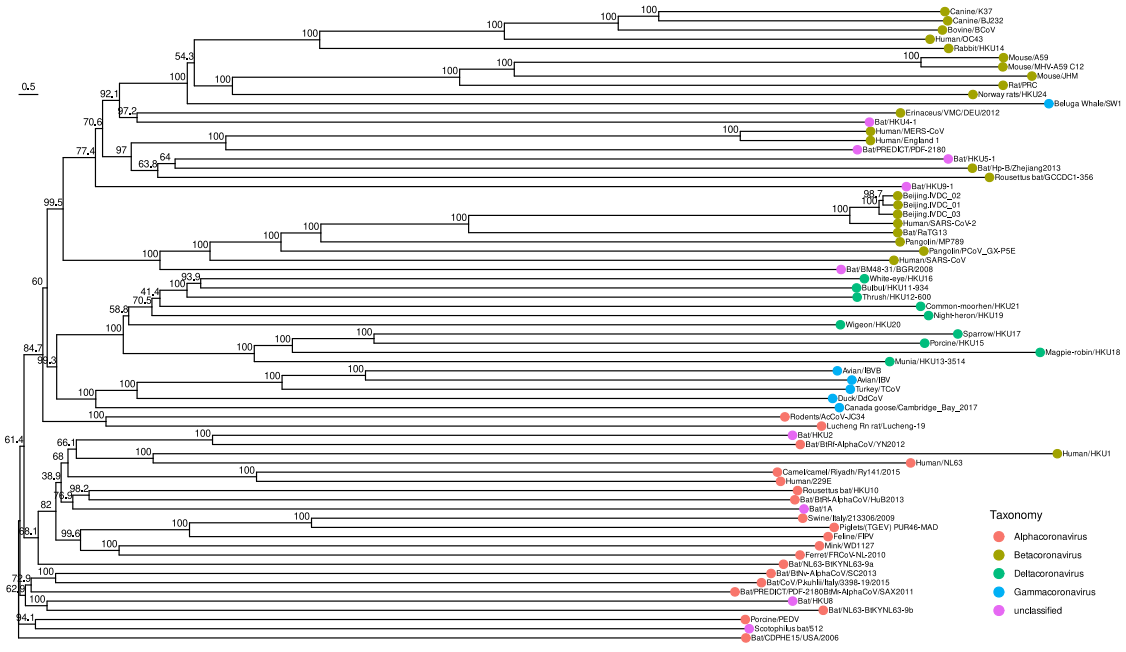



**Figura 19** - Árvore gerada pelo algoritmo NJ a partir das distâncias euclidianas da representação numérica cumulative 6-mer natural vector. As cores representam cada grupo taxonômico e fizemos 1000 rodadas de bootstrap para aferir a confiabilidade estatística dos nós internos.

Nós, também, analisamos para cada árvore o quanto estas se correlacionam com suas próprias distâncias originais e com a árvore de máxima-verossimilhança, via correlação cofenética (Tabela 3). Vimos que as árvores que melhor representaram suas distâncias originais foram Fast vector, 4-mer e cumulative 4-mer natural vectors, respectivamente, com valores de correlação cofenética acima de 0,9. Essas três representações também, na mesma ordem, obtiveram os melhores valores de correlação com a árvore gerada com alinhamento múltiplo de sequência. No entanto, comparando as distâncias RF, as representações 6-mer, cumulative 6-mer e 4-mer natural vector apresentaram menor distância à nossa árvore-referência, aproximadamente 0,45, 0,48 e 0,57, respectivamente (Tabela 3).


**Figura 20** - Tanglegrama conectando os táxons na esquerda da árvore de máxima-verossimilhança aos seus correspondentes na árvore da direita, criada com o 4-mer natural vector. A cor vermelha representa os genomas de alfacoronavírus, a bege os betacoronavírus, o verde os deltacoronavírus e o azul os gamacoronavírus.

Por fim, em vista dos melhores valores de correlação comparados, de estrutura global e comportamento de grupamentos, fizemos quatro tanglegramas, um para cada representação dos k-mer natural vectors (Figuras de 20 a 23). Vimos grande semelhança na formação dos grupamentos para o caso dos 4-mer e cumulative 4-mer natural vectors entre si e com a árvore gerada pelo IQTREE, exceto que os deltacoronavírus não derivam de um mesmo nó, de um possível ancestral comum (Figuras 20 e 21). Vale destacar que o grupamento das sequências de genomas de alfacoronavírus ficou mais consistente com a árvore referência na árvore do cumulative 4-mer natural vectors (Figura 21).

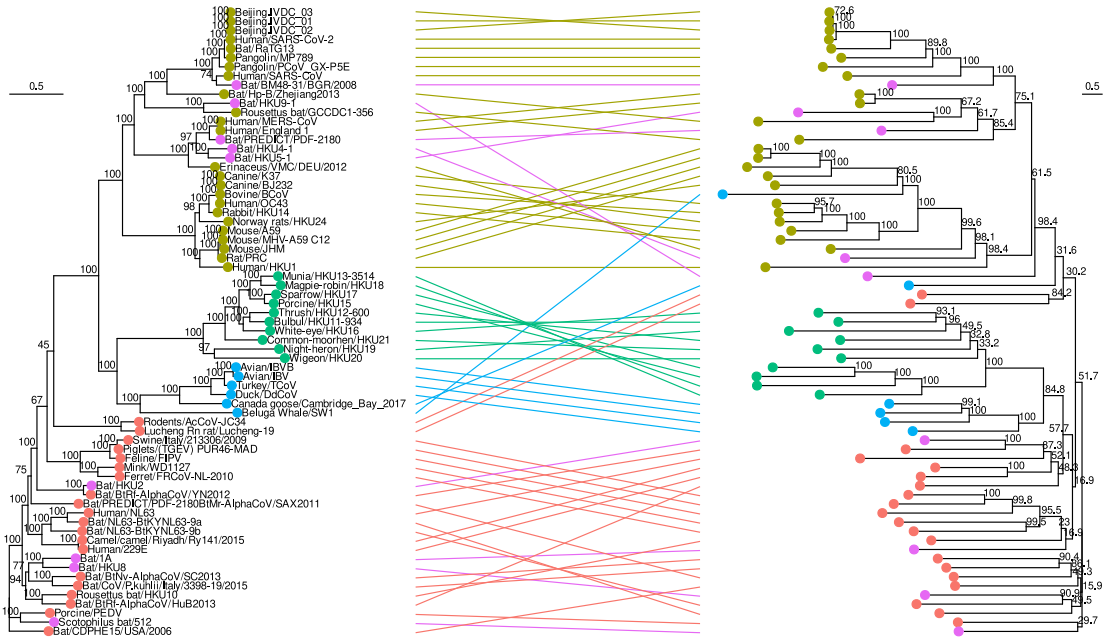



**Figura 21** - Tanglegrama conectando os táxons na esquerda da árvore de máxima-verossimilhança aos seus correspondentes na árvore da direita, criada com o cumulative 4-mer natural vector. A cor vermelha representa os genomas de alfacoronavírus, a bege os betacoronavírus, o verde os deltacoronavírus e o azul os gamacoronavírus.

No caso dos 6-mer, beta-, delta e gamacoronavírus compartilham do mesmo nó-ancestral comum (Figuras 22 e 23) como ocorre na árvore de referência. Em ambos os casos vemos que o betacoronavírus HKU1 humano agrupou-se aos alfacoronavírus, próximo ao NL63 humano. Contudo, podemos observar que para o 6-mer natural vector (Figura 22) o embaralhamento dos ramos é menos evidente do que em comparação com sua versão cumulativa (Figura 23).

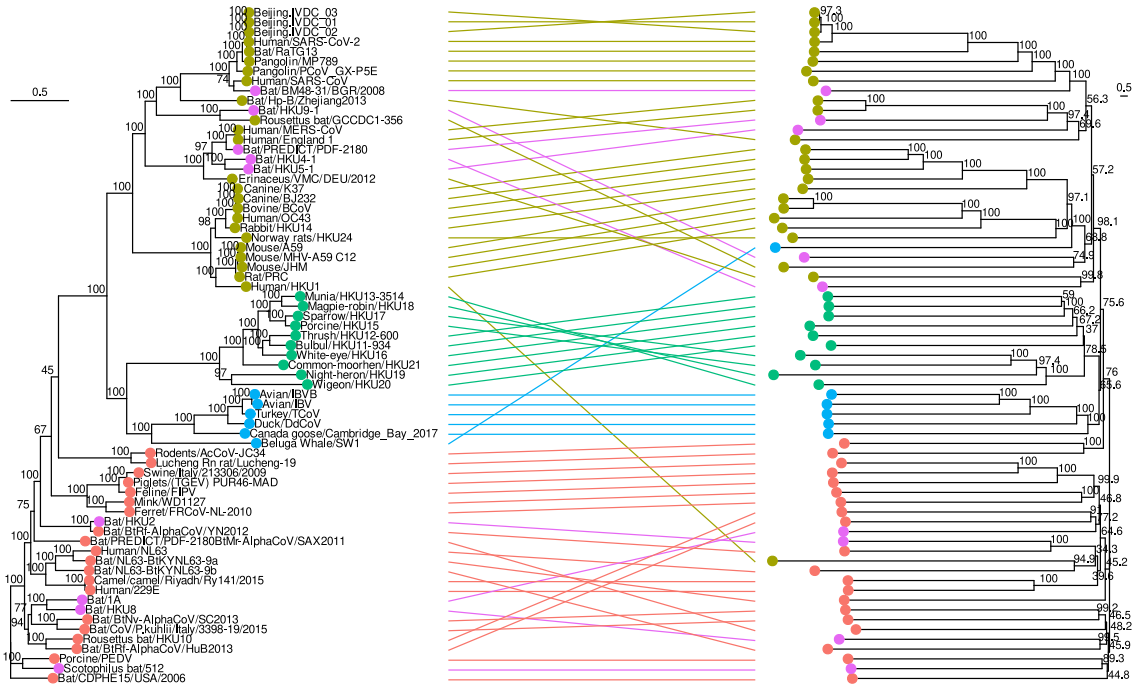



**Figura 22** - Tanglegrama conectando os táxons na esquerda da árvore de máxima-verossimilhança aos seus correspondentes na árvore da direita, criada com o 6-mer natural vector. A cor vermelha representa os genomas de alfacoronavírus, a bege os betacoronavírus, o verde os deltacoronavírus e o azul os gamacoronavírus.

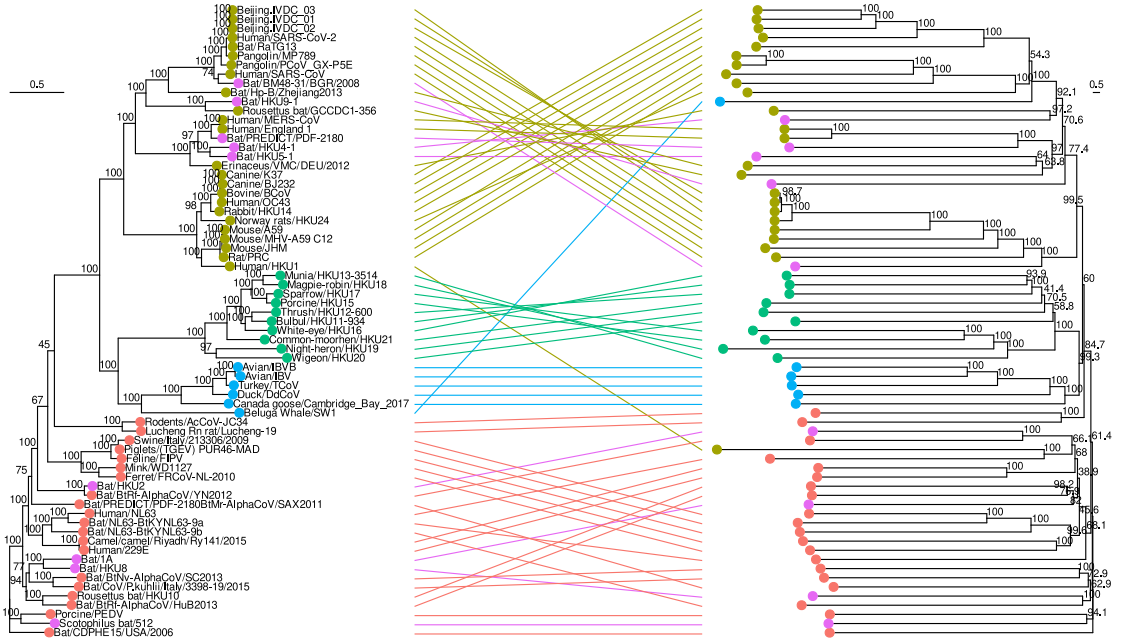



**Figura 23** - Tanglegrama conectando os táxons na esquerda da árvore de máxima-verossimilhança aos seus correspondentes na árvore da direita, criada com o cumulative 6-mer natural vector. A cor vermelha representa os genomas de alfacoronavírus, a bege os betacoronavírus, o verde os deltacoronavírus e o azul os gamacoronavírus.

Com isso, vemos que as representações baseadas em k-mer natural vectors apresentam potencial para lidar com os genomas de coronavírus próximos entre si, uma vez que os grupamentos puderam ser reconhecidos corretamente embora a estrutura global das árvores não tenha capturado grande similaridade comparada com nossa árvore-referência.

## 4.2 Aplicando as técnicas em conjunto genomas brasileiros de SARS-CoV-2 publicamente disponíveis: analisando a aplicação das metodologias em situações reais

### 4.2.1 Sequências genômicas brasileiras de SARS-CoV-2

O dataset original, que contém 85.691 genomas completos, possui exemplares de 20 clados do nextrain, sendo estes 19A, 19B, 20A, 20B, 20C, 20D, 20E (EU1), 20G, 20H (Beta,V2), 20I (Alpha,V1), 20J (Gamma,V3), 21A (Delta), 21D (Eta), 21G (Lambda), 21H (Mu), 21I (Delta), 21J (Delta), 21K (Ômicron), 21L (Ômicron), 21M (Ômicron), 22B (Ômicron), 22A (Ômicron) e 22C (Ômicron). Além de algumas sequências recombinantes. As três linhagens com mais exemplares são 20J (Gamma, V3), 21J (Delta) e 20B, com 43.051, 32.610 e 7.928 genomas, respectivamente. Vale destacar que estas sequências são genomas massivamente coletados entre março e dezembro de 2021 e que a região sudeste concentra grande quantidade do total depositado (Apêndice 3).

A subseleção de sequências deste grande dataset com a ferramenta augur resultou em 2.951 sequências e, destas, a ferramenta de classificação e alinhamento rápido nextclade retirou uma, devido à baixa qualidade. Assim, a subamostragem final possui 2.950 genomas, distribuídos de maneira mais homogênea (Apêndice 4).

Para determinar se as sequências na subamostra continham informações evolutivas adequadas para inferência filogenética, realizamos uma análise de likelihood-mapping (Apêndice 5). Verificamos que 88,7% dos pontos analisados apresentaram sinais tree-like, portanto menos de 20% foram classificados como sinais star-like ou network-like. Assim, concluímos que os dados estão aptos para continuarmos a inferência da árvore (MORRISON, 2005).

Nesta subamostragem final, 15 clados do nextrain foram mantidos, sendo 19A, 19B, 20A, 20B, 20C, 20G, 20J (Gamma, V3), 21A (Delta), 21I (Delta), 21J (Delta), 21K (Ômicron), 21L (Ômicron), 22A (Ômicron), 22B (Ômicron), 22C (Ômicron). Os clados mais representados em decorrência de sua distribuição temporal foram 20B, 20J

(Gamma, V3) e 21J (Delta), com 1.116, 649 e 529 genomas, respectivamente. Genomas de ômicron estavam divididos em 5 clados, totalizando 540 sequências. Adicionalmente, 29 genomas recombinantes estavam entre os selecionados pelo augur filter tool.

### 4.2.2 Árvore de máxima-verossimilhança dos genomas brasileiros de SARS-CoV-2

A Figura 24 demonstra a árvore de máxima-verossimilhança gerada pela ferramenta IQTREE2, com os genomas alinhados pela ferramenta Nextclade. Na parte basal da árvore vemos sequências de genomas iniciais, em sua maioria os do clado 20A (representado em verde-claro), o qual é parental a todos os demais (ver https://nextstrain.org/blog/2022-04-29-SARS-CoV-2-clade-naming-2022, acesso em 17 de fevereiro de 2023). Apesar de poucas sequências pertencentes ao clado 20I (Alpha, v1), temos a formação de um pequeno cluster ao centro da árvore. É relevante observar que as sequências de 20B dividiram-se em três clusters centrais.

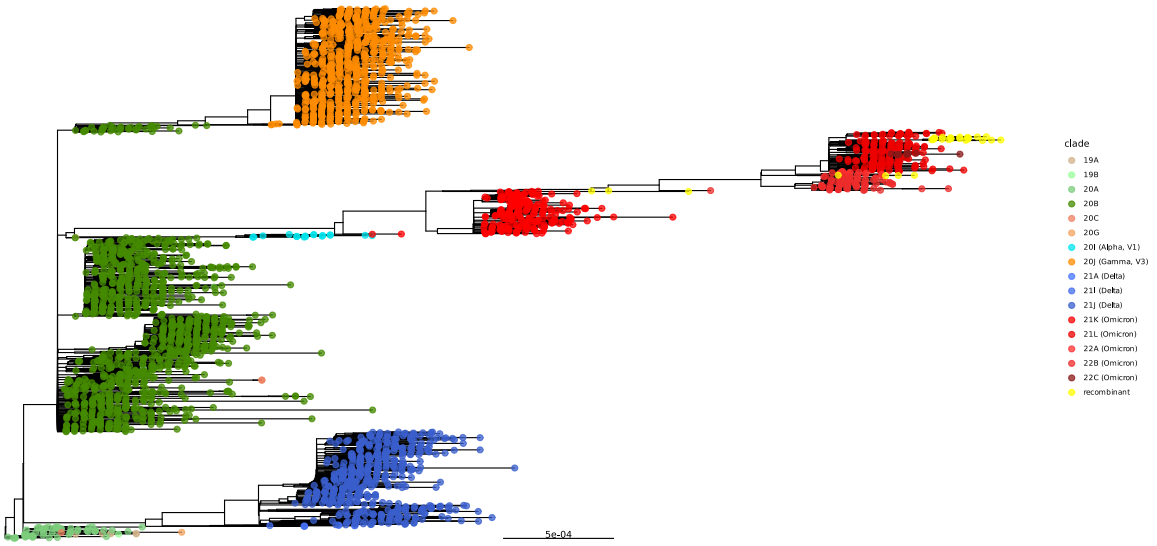



**Figura 24** - Árvore filogenética de máxima-verossimilhança das aproximadamente 3000 sequências de genomas brasileiros de SARS-CoV-2 baixados da base de dados GISAID EpiCOV database. As cores representam cada um dos clados criados pelo Nextrain para designar nomes as linhagens virais.

Como o tamanho da árvore impossibilita a plotagem dos valores de bootstrap em cada nó interno, mostramos na Apêndice 6 um histograma contendo a distribuição destes valores. Notamos que, os valores gerados pelo Ultrafast Bootstrap do

IQTREE2 concentraram-se em 100% e em 25%, sendo que uma quantidade pequena, porém uniforme distribuiu-se ao longo de valores entre 0 e 100% (Apêndice 6).

### 4.2.3 Redução da dimensionalidade dos vetores numéricos

De forma similar ao que fizemos para os 69 genomas de coronaviridae, verificamos se a representação simplificada dos dados revelaria padrões relacionados às linhagens, aqui representadas pelos clados do nextclade. Como a quantidade de genomas é maior para este caso, adequamos os números de vizinhos nas três execuções do UMAP, ajustando para 2.500, 300 e 30 vizinhos, com o mesmo objetivo de representar tanto características globais como locais da estrutura do dado. As quantidades de variância da PCA e o valor de stress do MDS encontram-se na tabela abaixo (Tabela 4). As representações reduzidas encontram-se nas Figuras 25-28.

**Tabela 4** - Número de atributos, valores de variância contidos nos 2 componentes principais da PCA e valor do strees gerado pelo MDS.

| Representação | Número de atributos | PCA 1 | PCA 2 | Variância em 2D | Stress do MDS |
|---|---|---|---|---|---|
| 4-mer Natural Vector | 768 | 0,20 | 0,14 | 0,34 | 6.651.377,73 |
| Cumulative 4-mer Natural Vector | 1.020 | 0,23 | 0,13 | 0,36 | 7.793.229,29 |
| 6-mer Natural Vector | 12.288 | 0,20 | 0,15 | 0,35 | 71.247.246,02 |
| Cumulative 6-mer Natural Vector | 16.380 | 0,20 | 0,15 | 0,35 | 106.632.001,36 |

De maneira ampla, todas as reduções de dimensionalidade resultaram em grupamentos similares, para as três técnicas, com os quatro conjuntos de dados. Embora os valores de stress do MDS aumentaram conforme o número de dimensões de cada representação, os valores da variância total mantida em duas dimensões da PCA mantiveram-se similares (Tabela 4).

Quatro agrupamentos foram formados em todos os casos: i) um contendo majoritariamente o clado 20B, juntamente com as linhagens iniciais da pandemia (19A, 19B, 20A, 20C, 20G e 20I - Alfa); ii) outro contendo em sua maioria genomas de Gama (20J); iii) outro contendo genomas da linhagem Delta, o que inclui os clados

21A, 21I e 21J; e iv) o último contendo Ômicron (21K, 21L, 22A, 22B e 22C) e sequências genômicas classificadas como recombinantes.

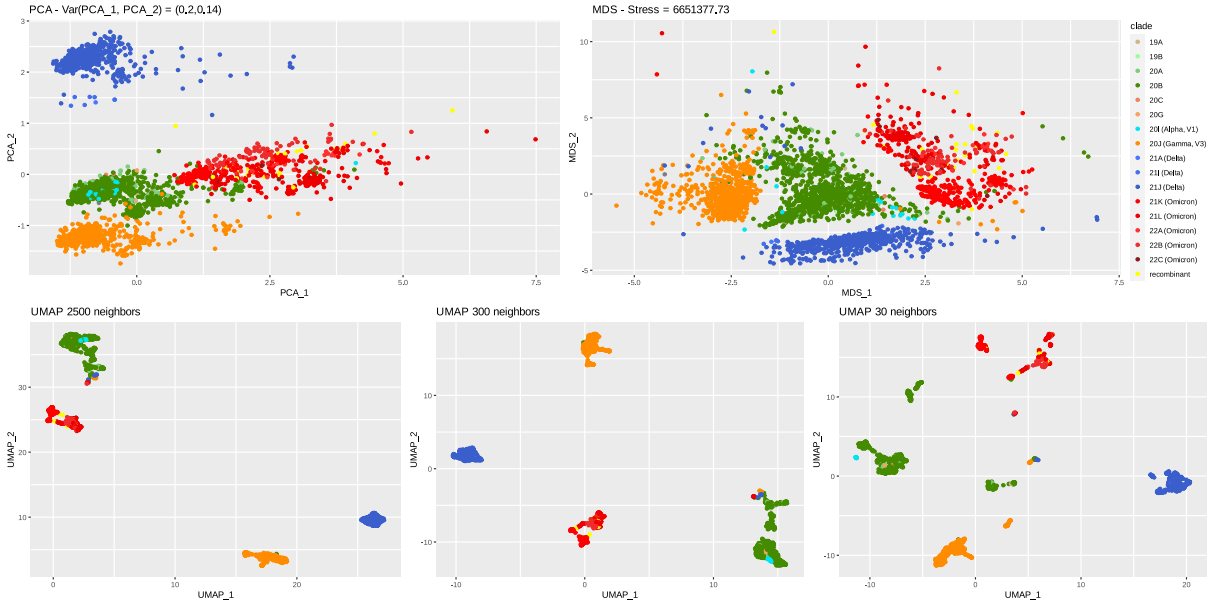



**Figura 25** - Gráficos das análises de redução de dimensionalidade das sequências genômicas do SARS-CoV-2 brasileiras representadas pelo 4-mer natural vector. Nesta análise cada ponto foi colorido representando o clado do Nextrain a qual cada genoma pertence.


**Figura 26** - Gráficos das análises de redução de dimensionalidade das sequências genômicas do SARS-CoV-2 brasileiras representadas pelo cumulative 4-mer natural vector. Nesta análise cada ponto foi colorido representando o clado do Nextrain a qual cada genoma pertence.

As reduções das representações 6-mer natural vector (Figura 27) e cumulative 6-mer natural vector (Figura 28) obtiveram separações claras com relação aos

agrupamentos supracitados, sobretudo quando utilizamos a técnica PCA. Nas UMAPs geradas a partir das representações 4-mer natural vector (Figura 25) e cumulative 4-mer natural vector (Figura 26), vimos que, mesmo na transformação com 2.500 vizinhos que demonstrariam aspectos globais do conjunto de dados, genomas de 20B estão presentes no agrupamento de genomas da linhagem Gama e também notamos pequenos grupamentos formados por genomas de diversos clados.

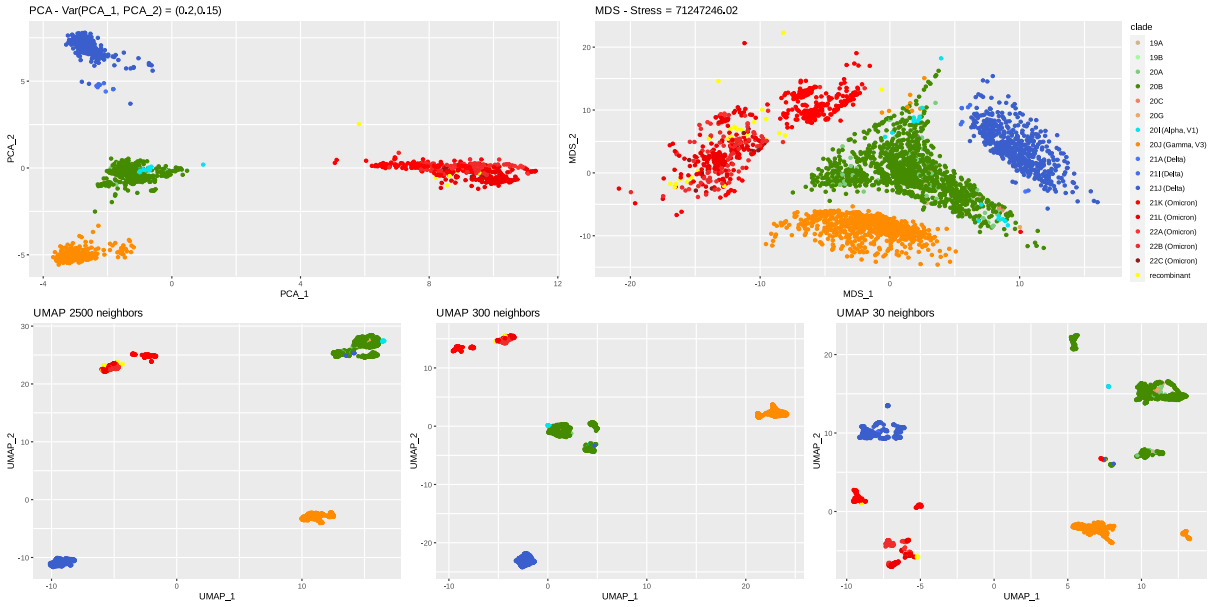



**Figura 27** - Gráficos das análises de redução de dimensionalidade das sequências genômicas do SARS-CoV-2 brasileiras representadas pelo 6-mer natural vector. Nesta análise cada ponto foi colorido representando o clado do Nextrain a qual cada genoma pertence.

Dado o que foi apresentado, verificou-se que houve grupamentos similares para as três técnicas de análise utilizadas e que as representações numéricas foram capazes de correlacionar com as principais linhagens circulantes do vírus no Brasil. As representações 6-mer natural vector e cumulative 6-mer natural vector obtiveram melhores separações dos agrupamentos, enquanto as representações 4-mer natural vector e sua versão cumulativa apresentaram algumas dificuldades em separar todos os agrupamentos, com genomas de diferentes linhagens presentes em um mesmo agrupamento.

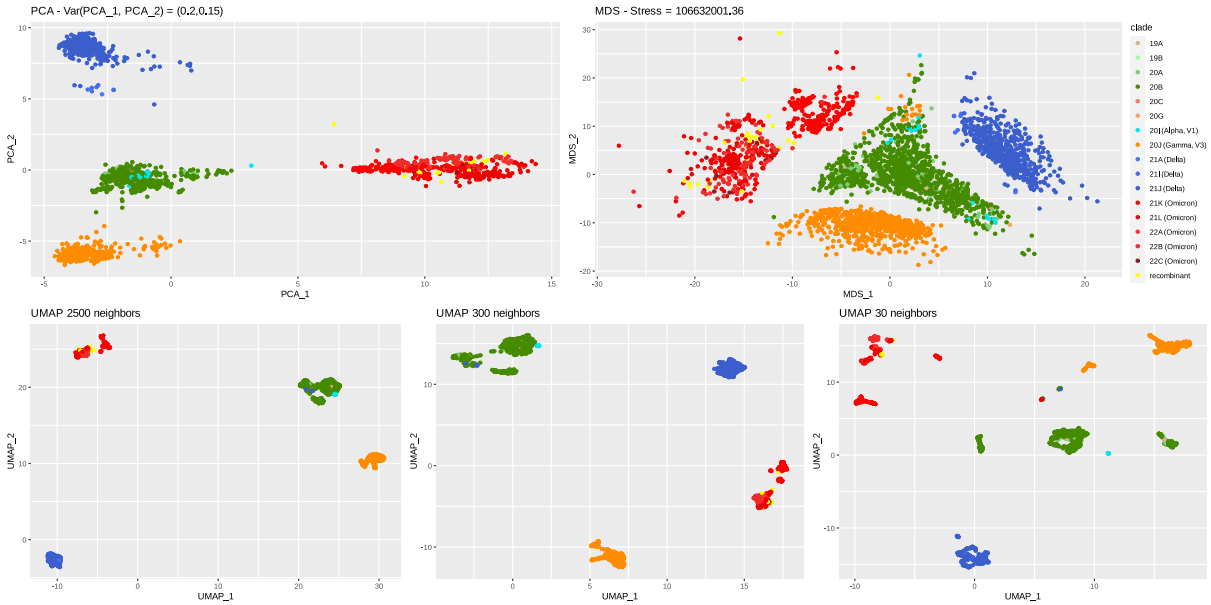



**Figura 28 - Gráficos das análises de redução de dimensionalidade das sequências genômicas do** SARS-CoV-2 brasileiras representadas pelo cumulative 6-mer natural vector. Nesta análise cada ponto foi colorido representando o clado do Nextrain a qual cada genoma pertence.

### 4.2.4 Construção e comparações das árvores com sequências de genomas brasileiros de SARS-CoV-2

Construímos árvores filogenéticas utilizando as distâncias euclidianas e o algoritmo NJ. Avaliamos, para estes dados, menos rodadas de bootstraps devido às limitações da metodologia testada implementada na linguagem R. Devido à falta de ocorrências e de variabilidade de alguns atributos, após normalização, 431 e 432 atributos foram removidos, respectivamente, dos 6-mer natural vectors e de sua versão cumulativa para calcularmos as distâncias.

Em contraste com nossa abordagem anterior, desta vez as distâncias foram avaliadas apenas com base na correlação, sem uma avaliação gráfica. Analisamos cerca de 3.000 genomas de SARS-CoV-2 brasileiros e comparamos as quatro árvores de referência visualmente com tanglegramas, utilizando métricas de correlação e distância entre árvores. Abaixo apresentamos uma tabela com as métricas de morfologia geral de estrutura (Tabela 5) e outra com as medidas de correlação e distância, por árvore gerada e por métrica analisada (Tabela 6).

**Tabela 5** - Valores de medidas morfológicas que sumarizam a estrutura das árvores filogenéticas analisadas nesta seção.

| | Average ladder | Cherry number | Colless Index | IL number | Altura máxima | Number of pitchforks | Sackin Index | Stairs 1 | Stairs 2 |
|---|---|---|---|---|---|---|---|---|---|
| 4-mer nat. vec. | 2,817881 | 866 | 79.304 | 1.218 | 60 | 414 | 99.770 | 0,6768396 | 0,5340445 |
| Cumulative 4-mer nat. vec. | 2,882550 | 858 | 75.721 | 1.234 | 57 | 410 | 96.843 | 0,6805697 | 0,5329912 |
| 6-mer nat. vec. | 3,110345 | 835 | 71.663 | 1.280 | 64 | 407 | 91.909 | 0,6863344 | 0,5172351 |
| Cumulative 6-mer nat. vec. | 3,111111 | 829 | 77.899 | 1.292 | 63 | 387 | 98.229 | 0,6887080 | 0,5172254 |
| MAFFT + IQTREE2 | 2,756972 | 935 | 132.305 | 1.080 | 138 | 452 | 151.005 | 0,6476772 | 0,5615945 |

Olhando para as árvores das representações 4-mer (Figura 29) e cumulative 4-mer natural vectors (Figura 30) notamos que somente os clados de alfa, gama, delta e ômicron permaneceram juntos, enquanto as linhagens iniciais da pandemia e a linhagem 20B ficaram espalhadas ao longo de toda a árvore, o que é interessante pois em nenhum caso das representações por redução de dimensionalidade o clado Alfa tinha ficado evidentemente separado. Notamos uma dispersão um pouco mais acentuada no caso do 4-mer natural vector, em relação à sua versão cumulativa.

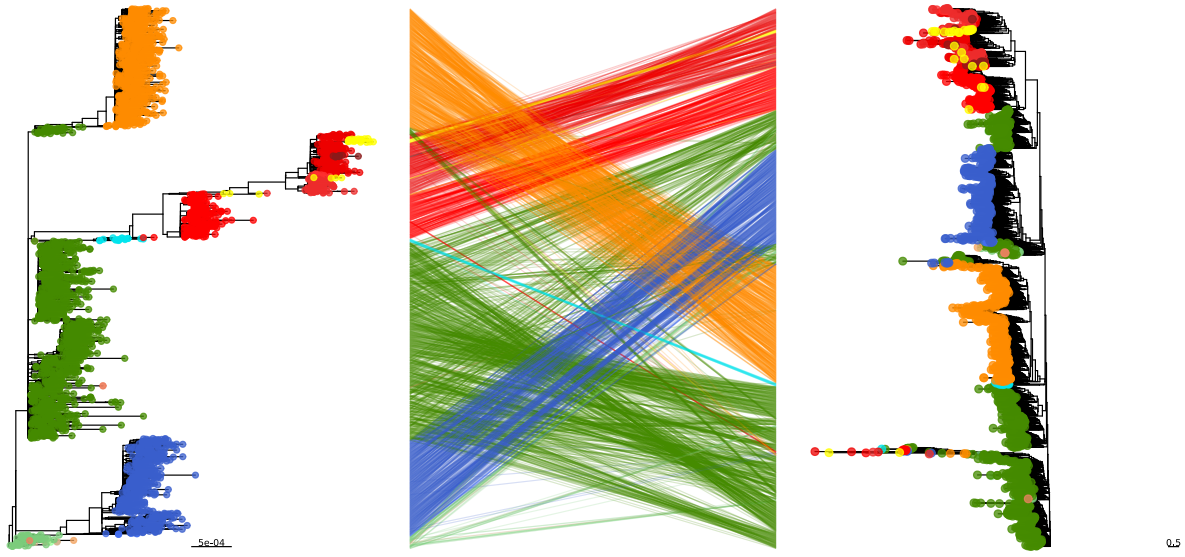



**Figura 29** - Tanglegrama conectando os aproximados 3.0000 genoams de SARS-CoV-2 na esquerda da árvore de máxima-verossimilhança aos seus correspondentes na árvore da direita, criada com o cumulative 4-mer natural vector. As cores usadas representam os clados do Nextclade, seguindo o mesmo esquema de cores das Figuras 25-28.

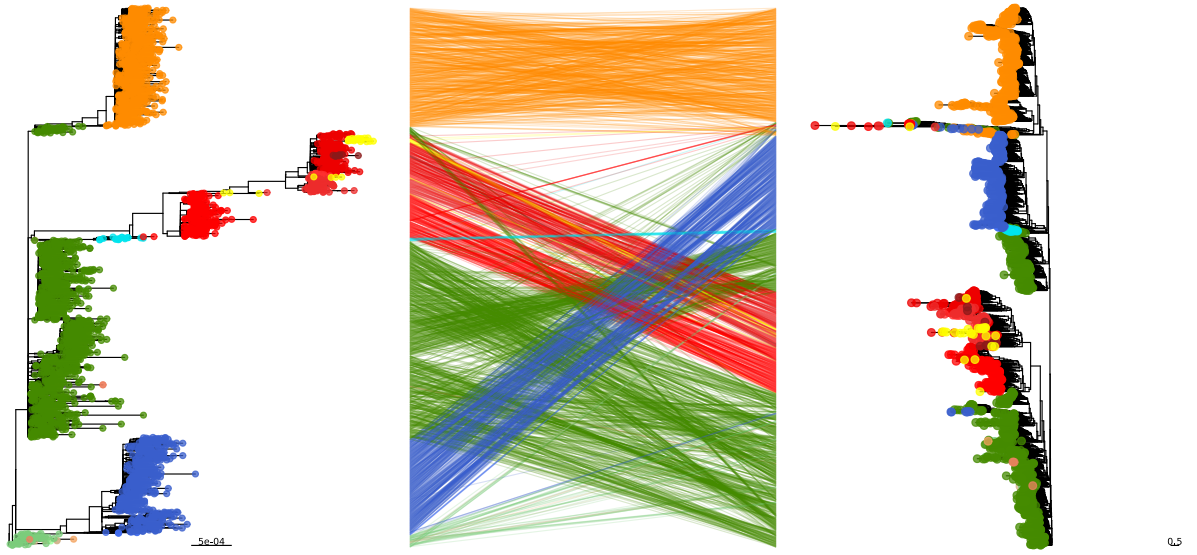



**Figura 30** - Tanglegrama conectando os aproximados 3.0000 genoams de SARS-CoV-2 na esquerda da árvore de máxima-verossimilhança aos seus correspondentes na árvore da direita, criada com o cumulative 4-mer natural vector. As cores usadas representam os clados do Nextclade, seguindo o mesmo esquema de cores das Figuras 25-28.

Adicionalmente, vale destacar a formação, nos dois casos, de um clado com múltiplas linhagens, na parte inferior da árvore dos 4-mer natural vectors (Figura 29), e na superior, da árvore do cumulative 4-mer natural vector (Figura 30). Ao explorarmos os metadados gerados pela ferramenta nextclade vimos que este cluster possui sequências recombinantes e com a maior quantidade de bases faltantes com relação à referência (Apêndice 7).

Apesar dos valores de ‘average ladder’ e stair 1 e 2 estarem semelhantes aos da árvore referência, os índices de Colless e de Sackin indicaram um menor desbalanço na estrutura geral das árvores em comparação com a árvore gerada pelo IQTREE (Tabela 5).

Já as árvores das representações 6-mer (Figura 31) e cumulative 6-mer (Figura 32) natural vectors apresentarem valores de ‘average ladder’ maiores com relação à árvore referência, o que pode ser visto em ambas as árvores de maneira que cada nova bifurcação da direita pra esquerda origina um novo grande clado. Em ambos os casos, o agrupamento com genomas de diversos clados também foi visto, embora tenham tido menor tamanho em comparação com os resultados acima para os 4-mers, e ambos ficaram próximos aos genomas de ômicron. É importante salientar que, na árvore dos 6-mer natural vectors uma melhor organização dos genomas do clado 20B foi alcançada, o que está diretamente alinhado com os resultados de correlação cofenética apresentados na Tabela 6.

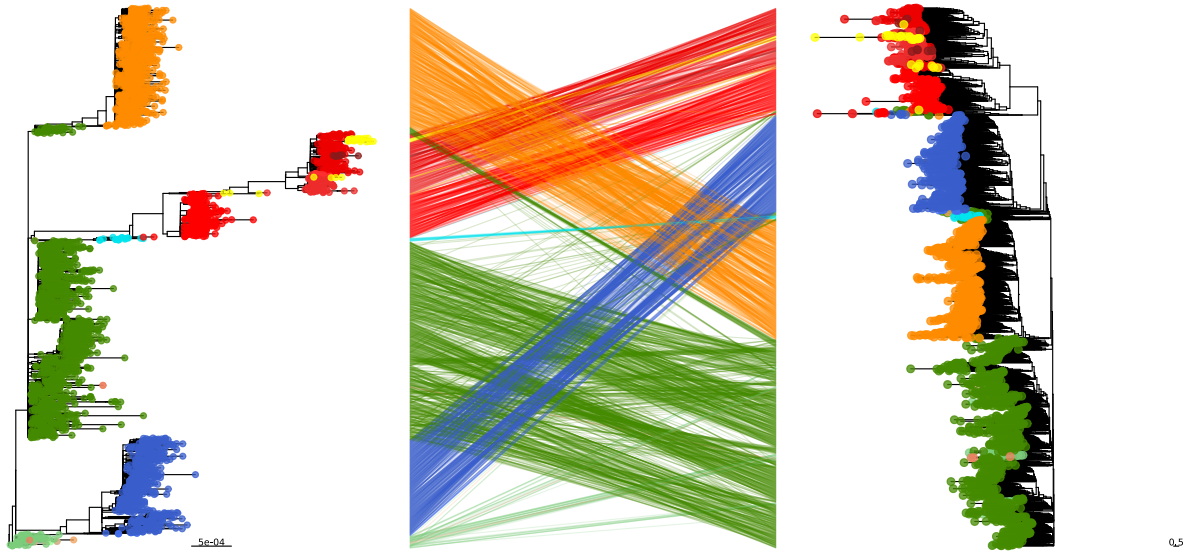



**Figura 31** - Tanglegrama conectando os aproximados 3.0000 genoams de SARS-CoV-2 na esquerda da árvore de máxima-verossimilhança aos seus correspondentes na árvore da direita, criada com o cumulative 6-mer natural vector. As cores usadas representam os clados do Nextclade, seguindo o mesmo esquema de cores das Figuras 25-28.

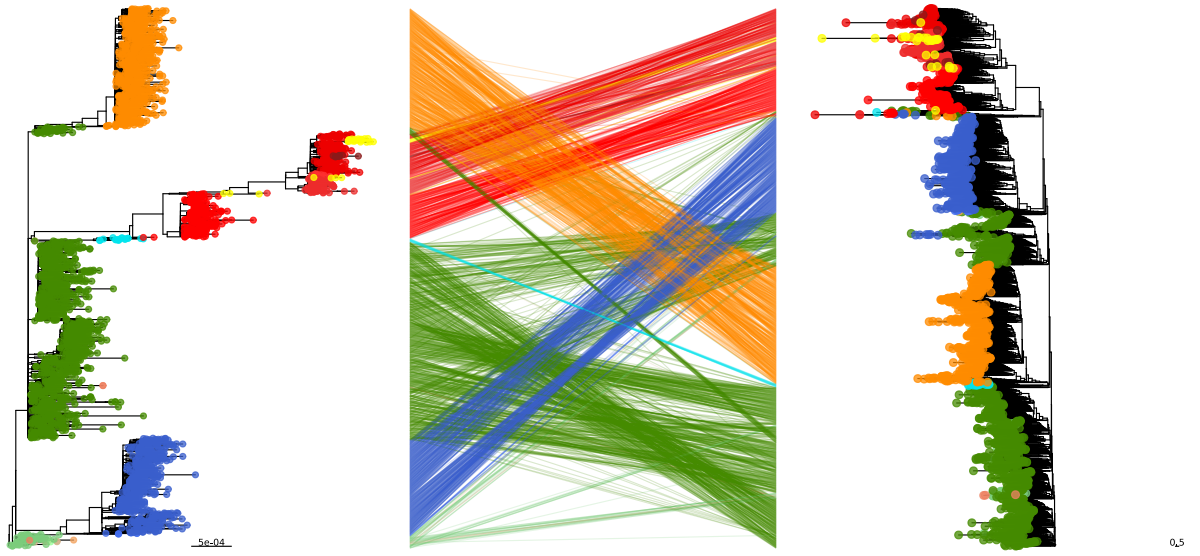



**Figura 32 -** Tanglegrama conectando os aproximados 3.0000 genoams de SARS-CoV-2 na esquerda da árvore de máxima-verossimilhança aos seus correspondentes na árvore da direita, criada com o cumulative 6-mer natural vector. As cores usadas representam os clados do Nextclade, seguindo o mesmo esquema de cores das Figuras 25-28.

Os valores de bootstrap destas árvores apresentaram distribuição similar ao visto para os da árvore de máxima-verossimilhança (Apêndice 6), embora entre os quatro resultados, os valores de bootstrap da árvore dos cumulative 6-mer natural vectors tenha obtido uma média relativamente maior (Apêndice 8).

**Tabela 6** -Medidas de correlação cofenética entre: i) a árvore gerada e sua própria distância, ii) a árvore gerada e a de máxima-verossimilhança de referência (Preferívies valores próximos a 1). Medidas de Distância RF entre a árvore gerada e a de máxima-verossimilhança de referência (Preferíveis valores próximos a 0).

| | 4-mer nat. vec. | Cumulat. 4-mer nat. vec. | 6-mer nat. vec. | Cumulat. 6-mer nat. vec. |
|---|---|---|---|---|
| Correlação cofenética com as próprias distâncias | 0,9559421 | 0,9513473 | 0,978355 | 0,9783453 |
| Correlação Cofenética com a árvore Referência | 0,7607726 | 0,7280986 | 0,8833394 | 0,8645622 |
| Distância de RF à árvore Referência | 0,9202176 | 0,9245694 | 0,8737987 | 0,8839529 |

Em oposição ao resultado visto para os 69 genomas de Orthocoronavirinae, aqui as correlações entre as distâncias dos 6-mer natural vectors e suas árvores foram melhores que as correlações vistas para os 4-mer natural vectors. Todos os valores de correlação obtidos com as árvores dos genomas brasileiros foram consideravelmente maiores do que os vistos anteriormente. Curiosamente, contrariamente ao esperado devido aos altos valores de correlação, as distâncias RF observadas para essas árvores foram todas superiores a 0,87.

Assim, como a distância RF compara a topologia das duas árvores, enquanto a correlação cofenética mede a similaridade dos grupos taxonômicos em duas árvores, temos que as representações empregadas e as formações das árvores pelo algoritmo NJ foram capazes de realizar agrupamentos similares à árvore de máxima-verossimilhança, mas não foram capazes de representar a estrutura global dos clados.

Tendo em conta estes e os resultados apresentados anteriormente, começamos especular sobre os motivos que levaram: i) aos agrupamentos pobres para os alfacoronavírus (ver Figuras 20-23) e ii) à incapacidade de separação dos genomas dos clados iniciais da pandemia (Figuras 29-32).

No primeiro caso, obtivemos uma árvore de máxima-verossimilhança com bom suporte estatístico para os ramos e, no geral, observamos boas métricas de correlação com as árvores geradas por NJ. Embora o conjunto de dados possua poucos exemplares de alfacoronavírus, em comparação com o demais, quando aumentamos o valor de k de quatro para seis observamos melhora nos agrupamentos. De fato, resultados de comparações extensivas envolvendo o tamanho do genoma, o

comprimento k e a quantidade de informação dependente da relação dessas duas características, apontam que para comparações baseadas somente em k-mer de um genoma do tamanho dos de coronavírus seria indicado a escolha de um k = 9 (ZHANG et al., 2017). Também da literatura, algumas investigações levaram a resultados que indicam um valor de k dependente do logaritmo na base 4 do comprimento total do genoma (SIMS et al., 2009), assim, um k-mer ideal seria próximo a um 7-mer. Por outro lado, é importante salientar que esse aumento em k aumentaria também a complexidade computacional, uso de recursos, e que isso potencialmente poderia causar *overfitting* das representações, pois o número de atributos gerado potencialmente ultrapassa o número informação primária (número de bases), o que é mais discutido no campo do aprendizado de máquina (GREENER et al., 2022). Adicionalmente, como a distribuição e ocorrência dos k-mer nos genomas não é aleatória e sim guiada por restrições biofísicas e evolutivas, foi demonstrado que pequenos k-mer já contemplam relevantes atributos para se treinar modelos de aprendizado de máquina (ALAM; CHOWDHURY, 2020).

No segundo caso, até mesmo um dos métodos mais acurados e bem estabelecidos de máxima-verossimilhança e bootstrap (IQTREE, UltrafastBootstrap) não pôde recuperar bons valores de suporte estatístico para os ramos. Reconstruções filogenéticas do período inicial da pandemia demonstraram que as linhagens iniciais de SARS-CoV-2 não se agrupam em perfis monofiléticos e que diferentes métodos levaram a diferentes agrupamentos destes genomas (BUKIN et al., 2021), pois pouca diversidade de mutações entre as sequências era observada, o que levava a poucos sítios polimórficos e subsequente baixa capacidade de gerar métricas baseadas em reamostragem como o bootstrap (WERTHEIM; STEEL; SANDERSON, 2022).

Logo, como os mapas 2D puderam separar bem vários clados e somente não os clados iniciais, acreditamos que esta limitação está atrelada à natureza deste dado pandêmico pouco divergentes no início das linhagens SARS-CoV-2, e não à falta de informação de k-mers. Ademais, a dispersão do clado 20B pode também ser explicada pelo mesmo motivo acima, somado ao fato de que, provavelmente, no início da pandemia múltiplas introduções dessa linhagem foram feiras ao redor do globo (BUKIN et al., 2021) e inclusive com evidências pra isso em estudos brasileiros (PAIVA et al., 2020; GIOVANETTI et al., 2022).

### 4.2.5 Construção de uma árvore filogenética com os genomas brasileiros de SARS-CoV-2 com distâncias geradas pela ferramenta Phylonium

Com esta excelente ferramenta designada a estimar distâncias evolutivas de amostras de genomas similares, obtivemos uma árvore filogenética (Figura 33) com o algoritmo NJ, que pôde separar melhor os genomas de cada clado e, principalmente, formar um clado com as sequências de genomas iniciais da pandemia, algo que não foi possível observar anteriormente com as técnicas testadas neste trabalho.

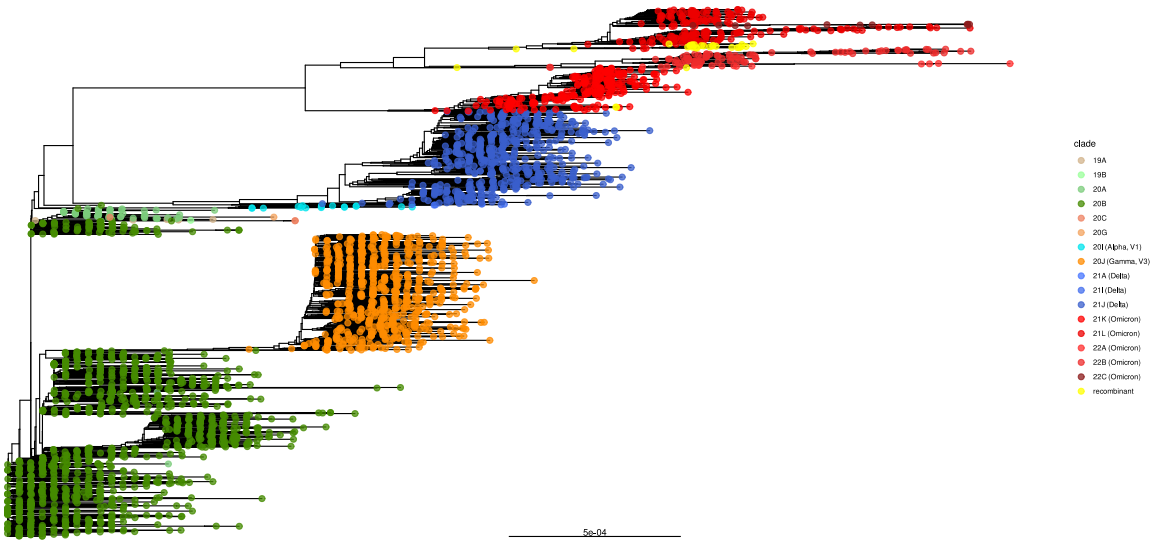



**Figura 33** - Árvore filogenética de baseada nas distâncias estimadas pela ferramenta Phylonium das aproximadamente 3000 sequências de genomas brasileiros de SARS-CoV-2. As cores representam cada um dos clados criados pelo Nextrain para designar nomes as linhagens virais.

Em contrapartida, apesar do bom agrupamento gerado e melhor dispersão dos táxons ao longo da árvore (Figura 34), as distâncias estimadas também falharam em recuperar a estrutura global da filogenia, tal qual observada na árvore de máxima-verossimilhança (Figura 24). Notamos maior desbalanceamento geral (índices de Colless e Sackin, Tabela 7) e maiores valores de padrões ladder-like (Average ladder, Stairs 1, Tabela 7).

Em um estudo recente sobre a dinâmica evolutiva dos coronavírus endêmicos HCoV-229E e HCoV-OC43, foram vistos valores de ladder length entre 2,44 e 3,11, e staircase-ness (aqui chamado de Stair 1) entre 0,69 e 0,72 (JO; DROSTEN; DREXLER, 2021). A maioria das árvores que geramos tiveram valores de staircase-ness por volta de 0,7, o que ressalta a importância de analisar outros parâmetros das árvores e verificarmos os agrupamentos.  No caso da medida de average ladder, todas

as árvores de SARS-CoV-2 brasileiros geraram valores semelhantes aos vistos para os coronavírus endêmicos, sendo a árvore gerada pelas distâncias da ferramenta phylonium a que apresentou maior average ladder juntamente com maior desbalanço na estrutura. Também vale ressaltar que as árvores de genomas de orthocoronavirinae apresentaram maiores médias para estrutura de escada.

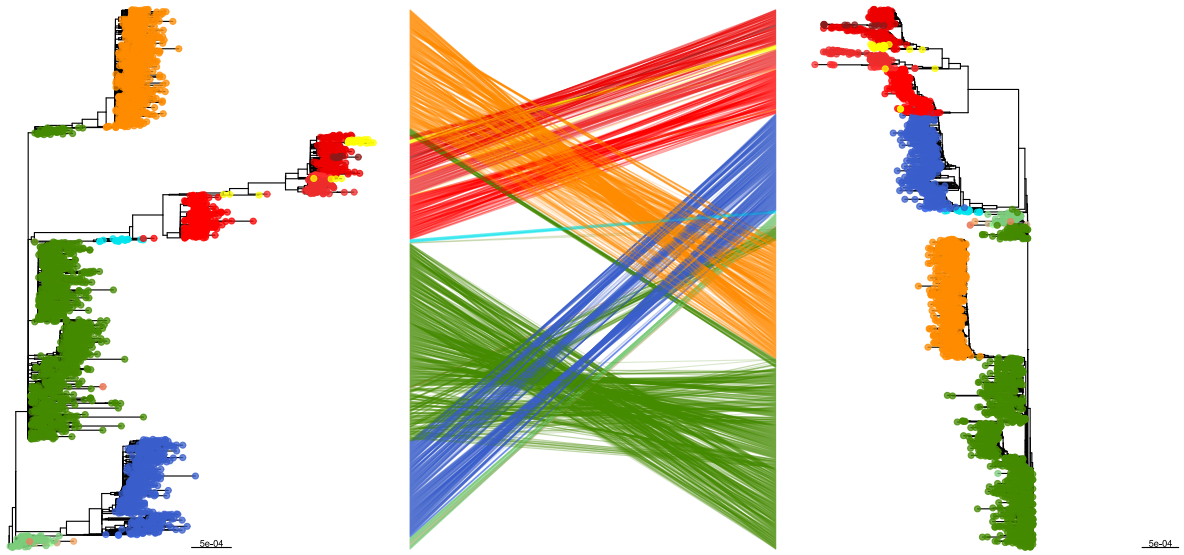



**Figura 34 -** Tanglegrama conectando os aproximados 3.0000 genoams de SARS-CoV-2 na esquerda da árvore de máxima-verossimilhança aos seus correspondentes na árvore da direita, baseada nas distâncias estimadas pela ferramenta Phylonium. As cores usadas representam os clados do Nextclade, seguindo o mesmo esquema de cores da Figuras 33.

Em concordância com o bom agrupamento e, sobretudo, melhor dispersão das sequências do clado 20B, a correlação cofenética das distâncias originais da ferramenta com sua própria árvore gerada foi 0,989334, e a da árvore gerada com a árvore de referência foi 0,9725789, as maiores obtidas para os testes com as sequências brasileiras. Similarmente ao visto para os testes das representações numéricas, a distância de RF à árvore Referência foi alta, sendo 0,6540345, ainda que este valor tenha sido mais baixo do que o observado para nosso melhor resultado, com o 6-mer natural vector (Tabela 6).

Tabela 7 - Valores de medidas morfológicas que sumarizam a estrutura das árvores filogenéticas analisadas nesta seção.

| | Average ladder | Cherry number | Colless Index | IL number | Altura máxima | Number of pitchforks | Sackin Index | Stairs 1 | Stairs 2 |
|---|---|---|---|---|---|---|---|---|---|
| phylorium | 3,154839 | 802 | 145.257 | 1.346 | 112 | 410 | 165.565 | 0,7056629 | 0,4689577 |
| MAFFT + IQTREE2 | 2,756972 | 935 | 132.305 | 1.080 | 138 | 452 | 151.005 | 0,6476772 | 0,5615945 |

O Phylonium pode lidar com inserções aleatórias e, devido à metodologia de comparação com uma única referência em seu conjunto de dados de entrada, possivelmente pode lidar com a falta de algumas regiões homólogas entre as sequências (KLÖTZL; HAUBOLD, 2020). Provavelmente por isso lidou melhor com os genomas com mais dados faltantes. O fato de a heurística utilizada para estimação da distância genética estar intimamente relacionada com um modelo de substituição pode ter favorecido a geração de uma escala que se aproxima melhor a arvore filogenética de referência. Em contrapartida, como este método somente gera as distâncias e não uma representação para cada genoma, a implementação de bootstrap para testar a confiabilidade dos ramos é dificultada.

## 5 Conclusões e Perspectivas Futuras

Ao melhor de nossa compreensão, este trabalho foi pioneiro em comparar genomas virais com um mesmo conjunto de sequências e analisar os resultados de cada representação sob a óptica de métricas diversas. Nós desenvolvemos um dataset, a partir de genomas e metadados de SARS-CoV-2, para ser usado como referência para estudos no nível taxonômico de espécie.

Vários trabalhos disponíveis na literatura ainda continuam usando conjuntos de sequências diferentes, bem como suas métricas. Neste sentido trabalhos como que visam criar benchmarks compreensivos destas técnicas e algoritmos fornecem boas inspirações para novos trabalhos.

Os scripts e pipelines desenvolvidos permitem uma análise comparativa entre métodos de representação numérica. As técnicas de representação numérica, por sua vez, permitem análises sem conceitos a priori e permitem produzir árvores com suporte estatístico.

Em comparação com as filogenias geradas pelos métodos usuais (alinhamento múltiplo de sequência seguida da inferência da árvore filogenética) as técnicas testadas neste trabalho ainda apresentam algumas diferenças e limitações, não possuindo o mesmo resultado. Entretanto novos parâmetros podem ser adicionados para melhorar a aproximação da análise de dados reais.

É importante ressaltar que as abordagens avaliadas se destacaram no sentido de gerar uma base para aplicações de aprendizado de máquina não supervisionado e clusterização de sequências biológicas. Esperamos e que a metodologia desenvolvida pode ser aplicada tanto para a análise de outros vírus como para testes de novas metodologias a serem criadas e que, em futuras versões dos nossos códigos, possamos fornecer para a comunidade ferramentas prontas e de fácil utilização.

**Referências**

ABADI, S. et al. Model selection may not be a mandatory step for phylogeny reconstruction. **Nature Communications**, v. 10, n. 1, p. 934, 25 Fev. 2019. Disponível em: <https://www.nature.com/articles/s41467-019-08822-w>.

AKSAMENTOV, I. et al. Nextclade: clade assignment, mutation calling and quality control for viral genomes. **Journal of Open Source Software**, v. 6, n. 67, p. 3773, 30 Nov. 2021. Disponível em: <https://joss.theoj.org/papers/10.21105/joss.03773>.

ALAM, M. N. U.; CHOWDHURY, U. F. Short k-mer abundance profiles yield robust machine learning features and accurate classifiers for RNA viruses. **PLOS ONE**, v. 15, n. 9, p. e0239381, 18 Set. 2020. Disponível em: <https://dx.plos.org/10.1371/journal.pone.0239381>.

APHALO, P. J.; SLOWIKOWSKI, K.; MOUKSASSI, S. ggpmisc: Miscellaneous Extensions to «ggplot2». 2022. Disponível em: <https://cran.r-project.org/package=ggpmisc>.

ARMSTRONG, G. et al. Applications and Comparison of Dimensionality Reduction Methods for Microbiome Data. **Frontiers in Bioinformatics**, v. 2, 24 Fev. 2022. Disponível em: <https://www.frontiersin.org/articles/10.3389/fbinf.2022.821861/full>.

BECHT, E. et al. Dimensionality reduction for visualizing single-cell data using UMAP. **Nature Biotechnology**, v. 37, n. 1, p. 38–44, 3 Jan. 2019. Disponível em: <http://www.nature.com/articles/nbt.4314>.

BOLYEN, E. et al. Reproducibly sampling SARS-CoV-2 genomes across time, geography, and viral diversity. **F1000Research**, v. 9, p. 657, 28 Out. 2020. Disponível em: <https://f1000research.com/articles/9-657/v2>.

BONHAM-CARTER, O.; STEELE, J.; BASTOLA, D. Alignment-free genetic sequence comparisons: a review of recent approaches by word analysis. **Briefings in Bioinformatics**, v. 15, n. 6, p. 890–905, 1 Nov. 2014. Disponível em: <https://academic.oup.com/bib/article-lookup/doi/10.1093/bib/bbt052>.

BORG, I.; GROENEN, P. J. F. The Four Purposes of Multidimensional Scaling. Em: **Modern Multidimens«The Four Purposes of Multidimensional Scaling» (sem data) em Modern Multidimensional Scaling. New York, NY: Springer New York, pp. 3–18. doi: 10.1007/0-387-28981-X_1.ional Scaling**. New York, NY: Springer New York, [s.d.]p. 3–18.

BUKIN, Y. S. et al. Phylogenetic reconstruction of the initial stages of the spread of the SARS-CoV-2 virus in the Eurasian and American continents by analyzing

genomic data. **Virus Research**, v. 305, p. 198551, Nov. 2021. Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/S0168170221002586>.

CAPELLA-GUTIÉRREZ, S.; SILLA-MARTÍNEZ, J. M.; TONI GABALDÓN. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. **Bioinformatics**, v. 25, n. 15, p. 1972–1973, 2009.

CARVALHO, T. A. et al. The scientific production during 2009 swine flu pandemic and 2019/2020 COVID-19 pandemic. **Pulmonology**, v. 26, n. 6, p. 340–345, Nov. 2020. Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/S2531043720301768>.

CHATZOU, M. et al. Multiple sequence alignment modeling: methods and applications. **Briefings in Bioinformatics**, v. 17, n. 6, p. 1009–1023, Nov. 2016. Disponível em: <https://academic.oup.com/bib/article/2606431/Multiple>.

CLARRIDGE, J. E. Impact of 16S rRNA Gene Sequence Analysis for Identification of Bacteria on Clinical Microbiology and Infectious Diseases. **Clinical Microbiology Reviews**, v. 17, n. 4, p. 840–862, Out. 2004. Disponível em: <https://journals.asm.org/doi/10.1128/CMR.17.4.840-862.2004>.

CUI, J.; LI, F.; SHI, Z.-L. Origin and evolution of pathogenic coronaviruses. **Nature Reviews Microbiology**, v. 17, n. 3, p. 181–192, 10 Mar. 2019. Disponível em: <http://www.nature.com/articles/s41579-018-0118-9>.

DEIGIN, Y.; SEGRETO, R. SARS-CoV-2′s claimed natural origin is undermined by issues with genome sequences of its relative strains. **BioEssays**, v. 43, n. 7, p. 2100015, 27 Jul. 2021. Disponível em: <https://onlinelibrary.wiley.com/doi/10.1002/bies.202100015>.

DILUCCA, M. et al. Codon Usage and Phenotypic Divergences of SARS-CoV-2 Genes. **Viruses**, v. 12, n. 5, p. 498, 30 Abr. 2020. Disponível em: <https://www.mdpi.com/1999-4915/12/5/498>.

GALILI, T. dendextend: an R package for visualizing, adjusting and comparing trees of hierarchical clustering. **Bioinformatics**, v. 31, n. 22, p. 3718–3720, 15 Nov. 2015. Disponível em: <https://academic.oup.com/bioinformatics/article/31/22/3718/240978>.

GIOVANETTI, M. et al. Genomic epidemiology of the SARS-CoV-2 epidemic in Brazil. **Nature Microbiology**, v. 7, n. 9, p. 1490–1500, 18 Ago. 2022. Disponível em: <https://www.nature.com/articles/s41564-022-01191-z>.

GORBALENYA, A. E. et al. The species Severe acute respiratory syndrome-related

coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2. **Nature Microbiology**, v. 5, n. 4, p. 536–544, 2 Mar. 2020. Disponível em: <https://www.nature.com/articles/s41564-020-0695-z>.

GRÄF, T. et al. Identification of a novel SARS-CoV-2 P.1 sub-lineage in Brazil provides new insights about the mechanisms of emergence of variants of concern. **Virus Evolution**, v. 7, n. 2, 1 Set. 2021. Disponível em: <https://academic.oup.com/ve/article/doi/10.1093/ve/veab091/6462077>.

GREENACRE, M. et al. Principal component analysis. **Nature Reviews Methods Primers**, v. 2, n. 1, p. 100, 22 Dez. 2022. Disponível em: <https://www.nature.com/articles/s43586-022-00184-w>.

GREENER, J. G. et al. A guide to machine learning for biologists. **Nature Reviews Molecular Cell Biology**, v. 23, n. 1, p. 40–55, 13 Jan. 2022. Disponível em: <https://www.nature.com/articles/s41580-021-00407-0>.

GRIBBLE, J. et al. The coronavirus proofreading exoribonuclease mediates extensive viral recombination. **PLOS Pathogens**, v. 17, n. 1, p. e1009226, 19 Jan. 2021. Disponível em: <https://dx.plos.org/10.1371/journal.ppat.1009226>.

HEARD, S. B. PATTERNS IN TREE BALANCE AMONG CLADISTIC, PHENETIC, AND RANDOMLY GENERATED PHYLOGENETIC TREES. **Evolution**, v. 46, n. 6, p. 1818–1826, Dez. 1992. Disponível em: <https://academic.oup.com/evolut/article/46/6/1818/6870474>.

HOANG, D. T. et al. UFBoot2: Improving the Ultrafast Bootstrap Approximation. **Molecular Biology and Evolution**, v. 35, n. 2, p. 518–522, 1 Fev. 2018. Disponível em: <https://academic.oup.com/mbe/article/35/2/518/4565479>.

HOZUMI, Y. et al. UMAP-assisted K-means clustering of large-scale SARS-CoV-2 mutation datasets. **Computers in Biology and Medicine**, v. 131, p. 104264, Abr. 2021. Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/S0010482521000585>.

HU, B. et al. Characteristics of SARS-CoV-2 and COVID-19. **Nature Reviews Microbiology**, v. 19, n. 3, p. 141–154, 6 Mar. 2021. Disponível em: <https://www.nature.com/articles/s41579-020-00459-7>.

HUDDLESTON, J. et al. Augur: a bioinformatics toolkit for phylogenetic analyses of human pathogens. **Journal of Open Source Software**, v. 6, n. 57, p. 2906, 7 Jan. 2021. Disponível em: <https://joss.theoj.org/papers/10.21105/joss.02906>.

JO, W. K.; DROSTEN, C.; DREXLER, J. F. The evolutionary dynamics of endemic

human coronaviruses. **Virus Evolution**, v. 7, n. 1, 20 Jan. 2021. Disponível em: <https://academic.oup.com/ve/article/doi/10.1093/ve/veab020/6157737>.

JOHNSON, B. A. et al. Loss of furin cleavage site attenuates SARS-CoV-2 pathogenesis. **Nature**, v. 591, n. 7849, p. 293–299, 11 Mar. 2021. Disponível em: <http://www.nature.com/articles/s41586-021-03237-4>.

JUNGREIS, I., SEALFON, R. & KELLIS, M. SARS-CoV-2 gene content and COVID-19 mutation impact by comparing 44 Sarbecovirus genomes. **Nat Commun**, v. 12, n. 2642, 2021.

KATOH, K.; STANDLEY, D. M. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. **Molecular Biology and Evolution**, v. 30, n. 4, p. 772–780, 2013.

KEMENA, C.; NOTREDAME, C. Upcoming challenges for multiple sequence alignment methods in the high-throughput era. **Bioinformatics**, v. 25, n. 19, p. 2455–2465, 1 Out. 2009. Disponível em: <https://academic.oup.com/bioinformatics/article/25/19/2455/180922>.

KENDALL, M.; BOYD, M.; COLIJN, C. phyloTop: Calculating Topological Properties of Phylogenies. 2023. Disponível em: <https://cran.r-project.org/package=phyloTop>.

KHAN, M. et al. COVID-19: A Global Challenge with Old History, Epidemiology and Progress So Far. **Molecules**, v. 26, n. 1, p. 39, 23 Dez. 2020. Disponível em: <https://www.mdpi.com/1420-3049/26/1/39>.

KHARE, S. et al. GISAID's Role in Pandemic Response. **China CDC Weekly**, v. 3, n. 49, p. 1049–1051, 2021. Disponível em: <http://weekly.chinacdc.cn/en/article/doi/10.46234/ccdcw2021.255>.

KIRICHENKO, A. D. et al. Comparative analysis of alignment-free genome clustering and whole genome alignment-based phylogenomic relationship of coronaviruses. **PLOS ONE**, v. 17, n. 3, p. e0264640, 8 Mar. 2022. Disponível em: <https://dx.plos.org/10.1371/journal.pone.0264640>.

KLÖTZL, F.; HAUBOLD, B. Phylonium: fast estimation of evolutionary distances from large samples of similar genomes. **Bioinformatics**, v. 36, n. 7, p. 2040–2046, 1 Abr. 2020. Disponível em: <https://academic.oup.com/bioinformatics/article/36/7/2040/5650408>.

KWAN, H. K.; ARNIKER, S. B. Numerical representation of DNA sequences. Em: 2009 IEEE International Conference on Electro/Information Technology, **Anais**...IEEE, Jun. 2009. Disponível em:

<http://ieeexplore.ieee.org/document/5189632/>.
LAM, T. T.-Y. et al. Identifying SARS-CoV-2-related coronaviruses in Malayan pangolins. **Nature**, v. 583, n. 7815, p. 282–285, 9 Jul. 2020. Disponível em: <http://www.nature.com/articles/s41586-020-2169-0>.
LI, Y. et al. A novel fast vector method for genetic sequence comparison. **Scientific Reports**, v. 7, n. 1, p. 12226, 22 Set. 2017. Disponível em: <https://www.nature.com/articles/s41598-017-12493-2>.
LIU, K.; LINDER, C. R.; WARNOW, T. RAxML and FastTree: Comparing Two Methods for Large-Scale Maximum Likelihood Phylogeny Estimation. **PLoS ONE**, v. 6, n. 11, p. e27731, 21 Nov. 2011. Disponível em: <https://dx.plos.org/10.1371/journal.pone.0027731>.
LOUCA, S.; DOEBELI, M. Efficient comparative phylogenetics on large trees. **Bioinformatics**, v. 34, n. 6, p. 1053–1055, 15 Mar. 2018. Disponível em: <https://academic.oup.com/bioinformatics/article/34/6/1053/4582279>.
MARINI, S. et al. Optimizing viral genome subsampling by genetic diversity and temporal distribution (TARDiS) for phylogenetics. **Bioinformatics**, v. 38, n. 3, p. 856–860, 12 Jan. 2022. Disponível em: <https://academic.oup.com/bioinformatics/article/38/3/856/6407117>.
MCINNES, L.; HEALY, J.; MELVILLE, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. 2018.
MENDIZABAL-RUIZ, G. et al. On DNA numerical representations for genomic similarity computation. **PLOS ONE**, v. 12, n. 3, p. e0173288, 21 Mar. 2017. Disponível em: <https://dx.plos.org/10.1371/journal.pone.0173288>.
MINH, B. Q. et al. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. **Molecular Biology and Evolution**, v. 37, n. 5, p. 1530–1534, 1 Mai. 2020. Disponível em: <https://academic.oup.com/mbe/article/37/5/1530/5721363>.
MORRISON, D. A. The Phylogenetic Handbook: A Practical Approach to DNA and Protein Phylogeny.—Marco Salemi and Anne-Mieke Vandamme (editors). 2003. Cambridge University Press, Cambridge, UK. 406 pp. ISBN 0-521-80390-X. US$75 (hardcover). **Systematic Biology**, v. 54, n. 6, p. 984–986, 1 Dez. 2005. Disponível em: <https://academic.oup.com/sysbio/article/54/6/984/1631992>.
NASCIMENTO, V. A. do et al. Genomic and phylogenetic characterisation of an imported case of SARS-CoV-2 in Amazonas State, Brazil. **Memórias do Instituto**

**Oswaldo Cruz**, v. 115, 2020. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0074-02762020000100421&tlng=en>.

NASER-KHDOUR, S. et al. The Prevalence and Impact of Model Violations in Phylogenetic Analysis. **Genome Biology and Evolution**, v. 11, n. 12, p. 3341–3352, 1 Dez. 2019. Disponível em: <https://academic.oup.com/gbe/article/11/12/3341/5571717>.

NORSTRÖM, M. M. et al. PhyloTempo: A Set of R Scripts for Assessing and Visualizing Temporal Clustering in Genealogies Inferred from Serially Sampled Viral Sequences. **Evolutionary Bioinformatics**, v. 8, p. EBO.S9738, 11 Jan. 2012. Disponível em: <http://journals.sagepub.com/doi/10.4137/EBO.S9738>.

OGANDO, N. S. et al. The Enzymatic Activity of the nsp14 Exoribonuclease Is Critical for Replication of MERS-CoV and SARS-CoV-2. **Journal of Virology**, v. 94, n. 23, 2020.

PAIVA, M. H. S. et al. Multiple Introductions Followed by Ongoing Community Spread of SARS-CoV-2 at One of the Largest Metropolitan Areas of Northeast Brazil. **Viruses**, v. 12, n. 12, p. 1414, 9 Dez. 2020. Disponível em: <https://www.mdpi.com/1999-4915/12/12/1414>.

PARADIS, E.; SCHLIEP, K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. **Bioinformatics**, v. 35, n. 3, p. 526–528, 1 Fev. 2019. Disponível em: <https://academic.oup.com/bioinformatics/article/35/3/526/5055127>.

PEER, Y. Van der; SALEMI, M. Phylogeny Inference Based on Distance Methods. Em: **The Phylogenetic Handbook: A Practical Approach to DNA and Protein Phylogeny**. [s.l: s.n.]

PEI, S.; YAU, S. S.-T. Analysis of the Genomic Distance Between Bat Coronavirus RaTG13 and SARS-CoV-2 Reveals Multiple Origins of COVID-19. **Acta Mathematica Scientia**, v. 41, n. 3, p. 1017–1022, 19 Mai. 2021. Disponível em: <https://link.springer.com/10.1007/s10473-021-0323-x>.

PYTHON CORE TEAM. Python: A dynamic, open source programming language. **Python Software Foundation**, 2015. Disponível em: <https://www.python.org/.>.

R CORE TEAM. **R: A language and environment for statistical computing.** Disponível em: <http://www.r-project.org>.

RAMBAUT, A. et al. A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. **Nature Microbiology**, v. 5, n. November, 2020.

Disponível em: <http://dx.doi.org/10.1038/s41564-020-0770-5>.
RANDHAWA, G. S.; HILL, K. A.; KARI, L. ML-DSP: Machine Learning with Digital Signal Processing for ultrafast, accurate, and scalable genome classification at all taxonomic levels. **BMC Genomics**, v. 20, n. 1, p. 267, 3 Dez. 2019. Disponível em: <https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-019-5571-y>.
ROBINSON, D. F.; FOULDS, L. R. Comparison of phylogenetic trees. **Mathematical Biosciences**, v. 53, n. 1–2, p. 131–147, Fev. 1981. Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/0025556481900432>.
SACKIN, M. J. «Good» and «Bad» Phenograms. **Systematic Biology**, v. 21, n. 2, p. 225–226, 1 Jul. 1972. Disponível em: <https://academic.oup.com/sysbio/article-lookup/doi/10.1093/sysbio/21.2.225>.
SAITOU, N., & NEI, M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. **Molecular Biology and Evolution**, Jul. 1987. Disponível em: <https://academic.oup.com/mbe/article/4/4/406/1029664/The-neighborjoining-method-a-new-method-for>.
SANDER, A.-L. et al. Genomic determinants of Furin cleavage in diverse European SARS-related bat coronaviruses. **Communications Biology**, v. 5, n. 1, p. 491, 30 Mai. 2022. Disponível em: <https://www.nature.com/articles/s42003-022-03421-w>.
SCHLIEP, K. P. phangorn: phylogenetic analysis in R. **Bioinformatics**, v. 27, n. 4, p. 592–593, 15 Fev. 2011. Disponível em: <https://academic.oup.com/bioinformatics/article/27/4/592/198887>.
SIMS, G. E. et al. Alignment-free genome comparison with feature frequency profiles (FFP) and optimal resolutions. **Proceedings of the National Academy of Sciences**, v. 106, n. 8, p. 2677–2682, 24 Fev. 2009. Disponível em: <https://pnas.org/doi/full/10.1073/pnas.0813249106>.
SINGH, D.; YI, S. V. On the origin and evolution of SARS-CoV-2. **Experimental & Molecular Medicine**, v. 53, n. 4, p. 537–547, 16 Abr. 2021. Disponível em: <http://www.nature.com/articles/s12276-021-00604-z>.
SOKAL, R. R.; MICHENER, C. D. A statistical method for evaluating systematic relationships. **University of Kansas Science Bulletin**, v. 38, p. 1409–1438, 1985.
VINGA, S. Editorial: Alignment-free methods in computational biology. **Briefings in Bioinformatics**, v. 15, n. 3, p. 341–342, 1 Mai. 2014. Disponível em: <https://academic.oup.com/bib/article-lookup/doi/10.1093/bib/bbu005>.
WANG, L.-G. et al. Treeio: An R Package for Phylogenetic Tree Input and Output

with Richly Annotated and Associated Data. **Molecular Biology and Evolution**, v. 37, n. 2, p. 599–603, 1 Fev. 2020. Disponível em: <https://academic.oup.com/mbe/article/37/2/599/5601621>.

WARNOW, T. Revisiting Evaluation of Multiple Sequence Alignment Methods. Em: [s.l: s.n.]p. 299–317.

WEN, J. et al. K-mer natural vector and its application to the phylogenetic analysis of genetic sequences. **Gene**, v. 546, n. 1, p. 25–34, Ago. 2014. Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/S0378111914006064>.

WERTHEIM, J. O.; STEEL, M.; SANDERSON, M. J. Accuracy in Near-Perfect Virus Phylogenies. **Systematic Biology**, v. 71, n. 2, p. 426–438, 10 Fev. 2022. Disponível em: <https://academic.oup.com/sysbio/article/71/2/426/6353043>.

WICKHAM, H. et al. Welcome to the Tidyverse. **Journal of Open Source Software**, v. 4, n. 43, p. 1686, 21 Nov. 2019. Disponível em: <https://joss.theoj.org/papers/10.21105/joss.01686>.

WONG, K. M.; SUCHARD, M. A.; HUELSENBECK, J. P. Alignment Uncertainty and Genomic Analysis. **Science**, v. 319, n. 5862, p. 473–476, 25 Jan. 2008. Disponível em: <https://www.science.org/doi/10.1126/science.1151532>.

WU, C. et al. Magnus representation of genome sequences. **Journal of Theoretical Biology**, v. 480, p. 104–111, Nov. 2019. Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/S0022519319303169>.

YANG, Y. et al. Characterizing Transcriptional Regulatory Sequences in Coronaviruses and Their Role in Recombination. **Molecular Biology and Evolution**, v. 38, n. 4, p. 1241–1248, 13 Abr. 2021. Disponível em: <https://academic.oup.com/mbe/article/38/4/1241/5955840>.

YU, G. et al. <scp>ggtree</scp> : an <scp>r</scp> package for visualization and annotation of phylogenetic trees with their covariates and other associated data. **Methods in Ecology and Evolution**, v. 8, n. 1, p. 28–36, 22 Jan. 2017. Disponível em: <https://onlinelibrary.wiley.com/doi/10.1111/2041-210X.12628>.

YU, N.; LI, Z.; YU, Z. Survey on encoding schemes for genomic data representation and feature learning—from signal processing to machine learning. **Big Data Mining and Analytics**, v. 1, n. 3, p. 191–210, Set. 2018. Disponível em: <https://ieeexplore.ieee.org/document/8361572/>.

ZHANG, Q. et al. Viral Phylogenomics Using an Alignment-Free Method: A Three-Step Approach to Determine Optimal Length of k-mer. **Scientific Reports**, v. 7, n. 1,

p. 40712, 19 Jan. 2017. Disponível em: <https://www.nature.com/articles/srep40712>.

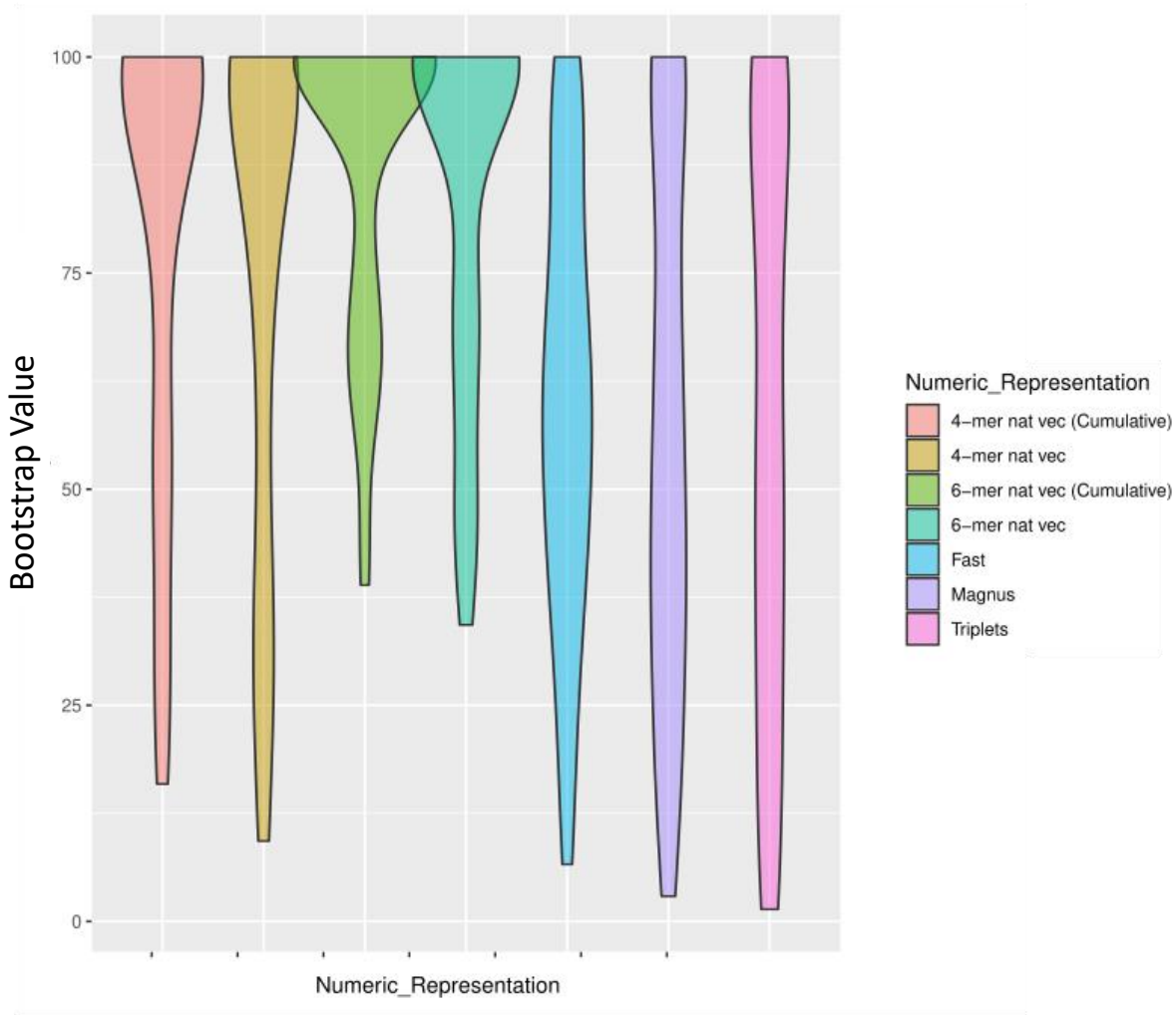
ZHOU, P. et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. **Nature**, v. 579, n. 7798, p. 270–273, 12 Mar. 2020. Disponível em: <http://www.nature.com/articles/s41586-020-2012-7>.

ZIELEZINSKI, A. et al. Alignment-free sequence comparison: benefits, applications, and tools. **Genome Biology**, v. 18, n. 1, p. 186, 3 Dez. 2017. Disponível em: <http://genomebiology.biomedcentral.com/articles/10.1186/s13059-017-1319-7>.

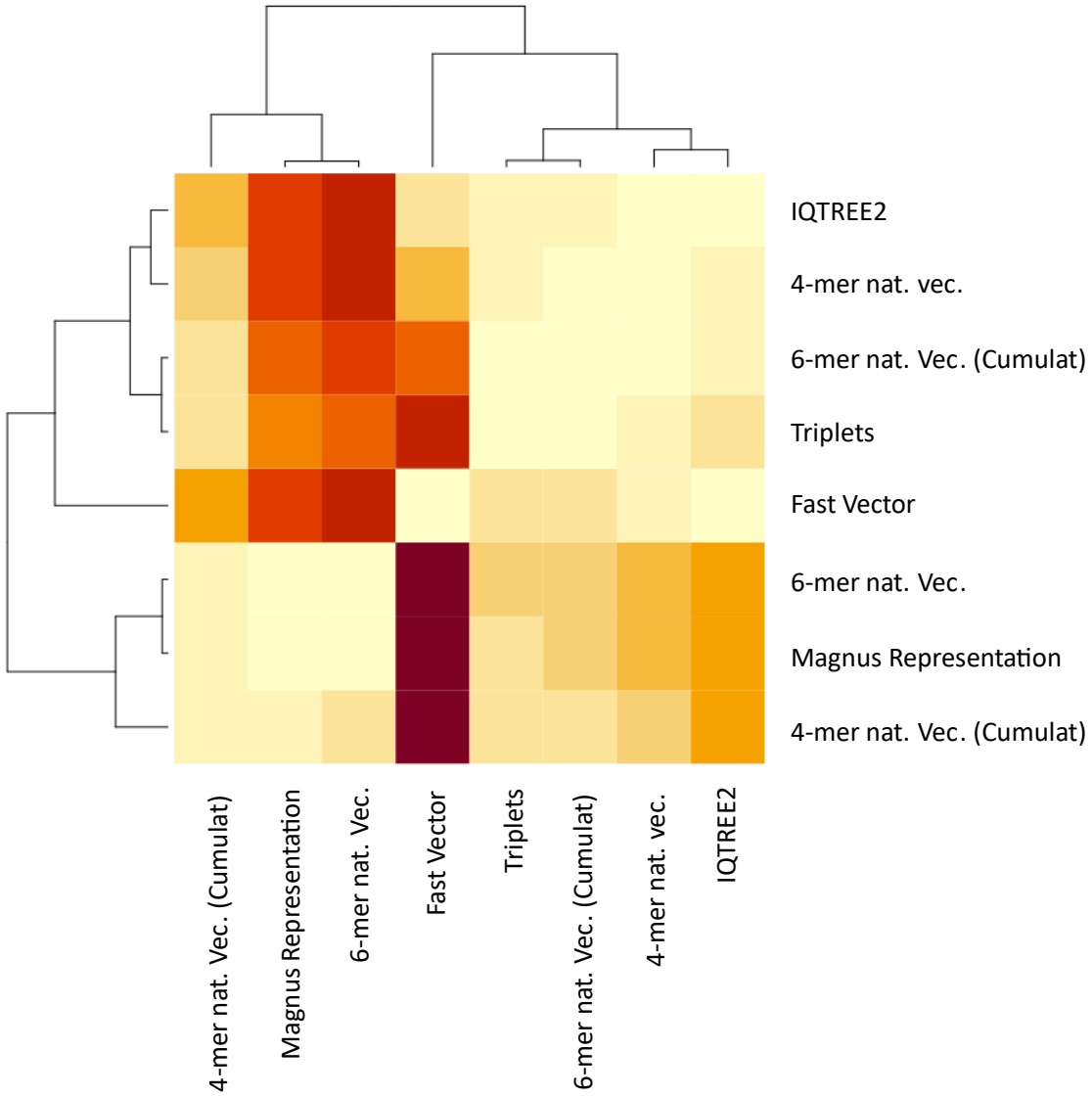
ZIELEZINSKI, A. et al. Benchmarking of alignment-free sequence comparison methods. **Genome Biology**, v. 20, n. 1, p. 144, 25 Dez. 2019. Disponível em: <https://genomebiology.biomedcentral.com/articles/10.1186/s13059-019-1755-7>.

ZIMERMAN, R. A. et al. Comparative Genomics and Characterization of SARS-CoV-2 P.1 (Gamma) Variant of Concern From Amazonas, Brazil. **Frontiers in Medicine**, v. 9, 15 Fev. 2022. Disponível em: <https://www.frontiersin.org/articles/10.3389/fmed.2022.806611/full>.

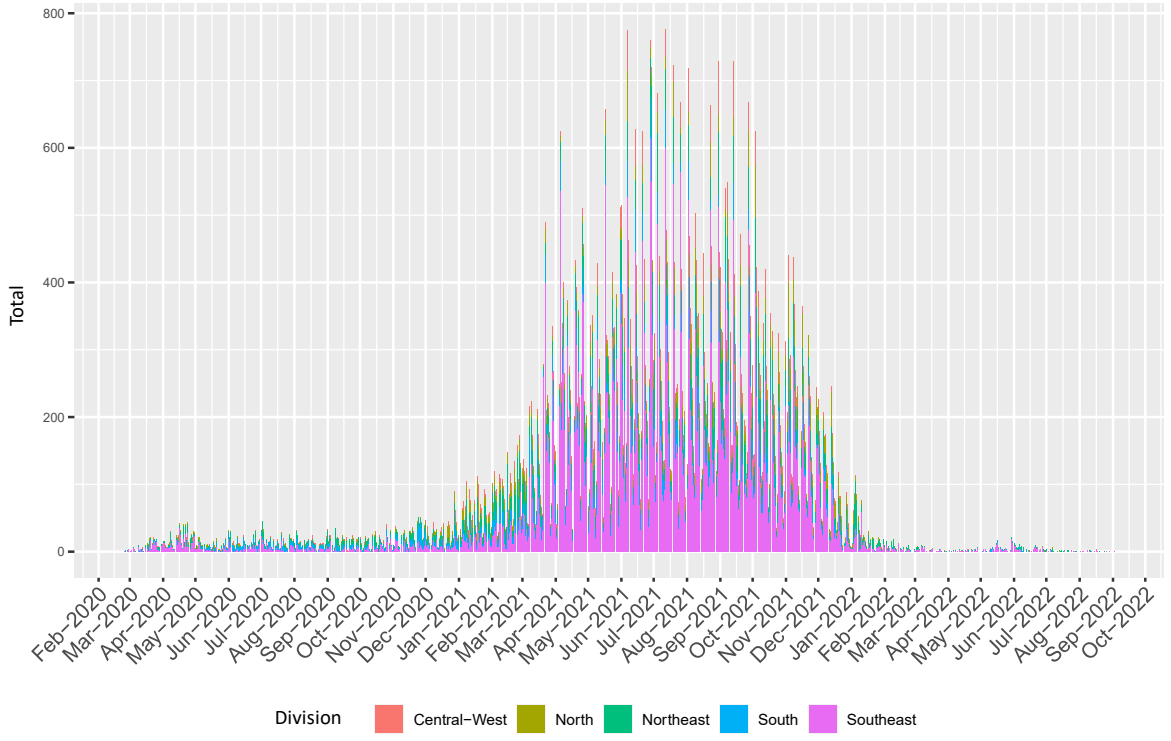
**Apêndice I -** Gráfico de violino com a distribuição dos valores de Bootstrap dos ramos internos das árvores geradas a partir de diferentes representações numéricas dos genomas da subfamília orthocoronavirinae.



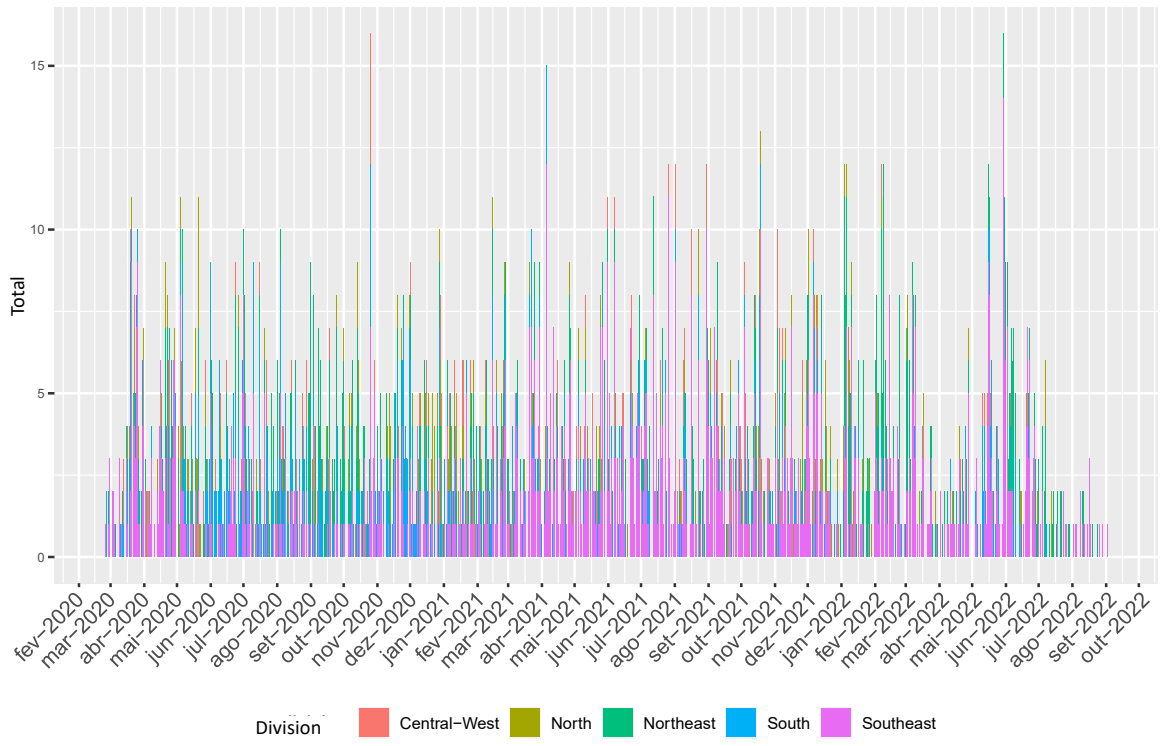
**Apêndice II -** Heatmap ilustrando a distância do cosseno entre as medidas estatísticas de morfologia das árvores filogenéticas geradas para as diferentes árvores criadas com as diferentes metodologias. O grupamento hierárquico ao lado mostra os padrões de proximidade da morfologia de cada árvore entre si.



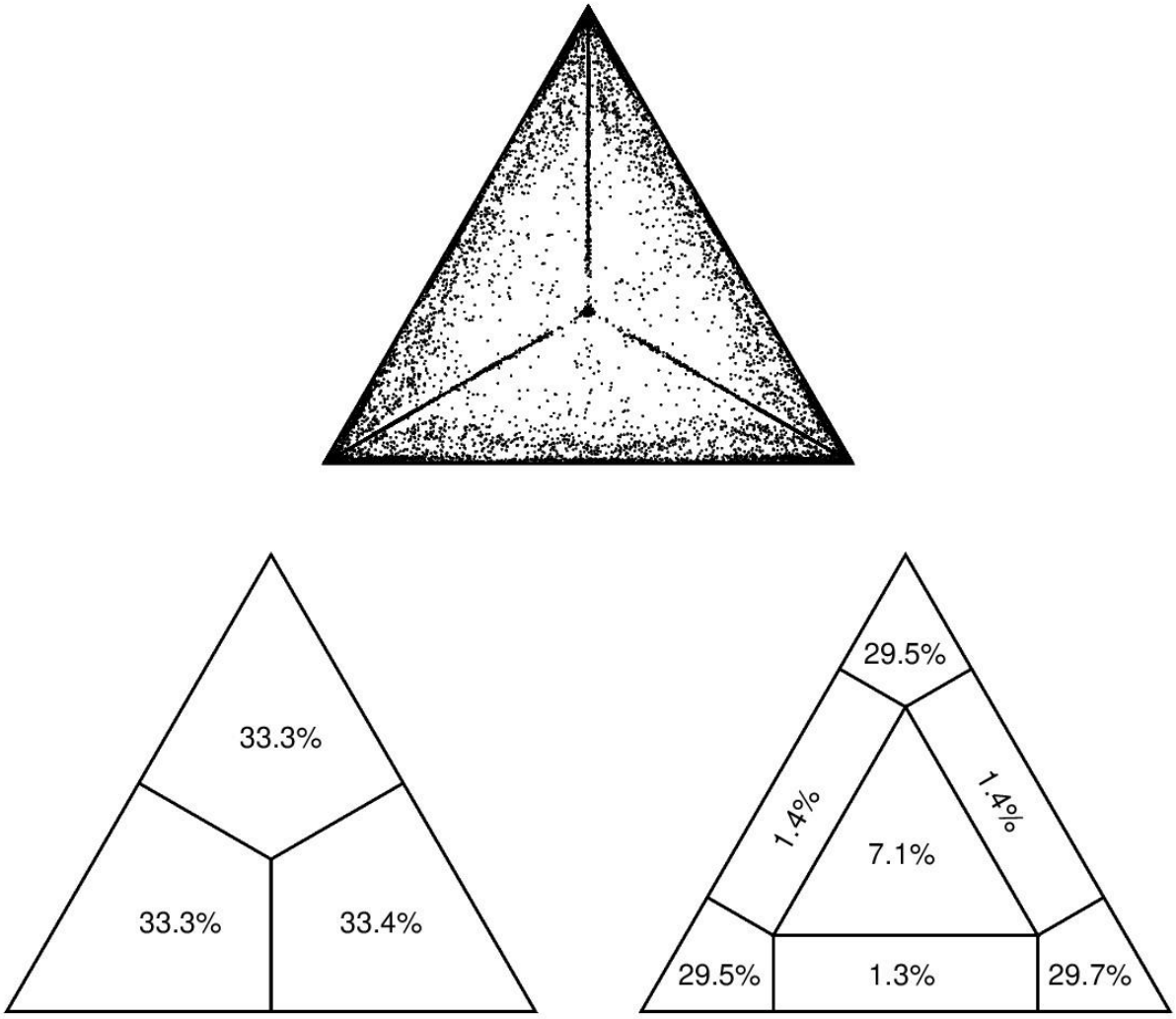
**Apêndice III -** Distribuição temporal das aproximadamente 86.000 sequências genômicas de SARS-CoV-2 sequenciadas no Brasil desde o começo da pandemia até outubro de 2022. As cores representam as regiões do Brasil no qual os genomas foram sequenciados. No eixo y temos a contagem absoluta das sequências geradas.



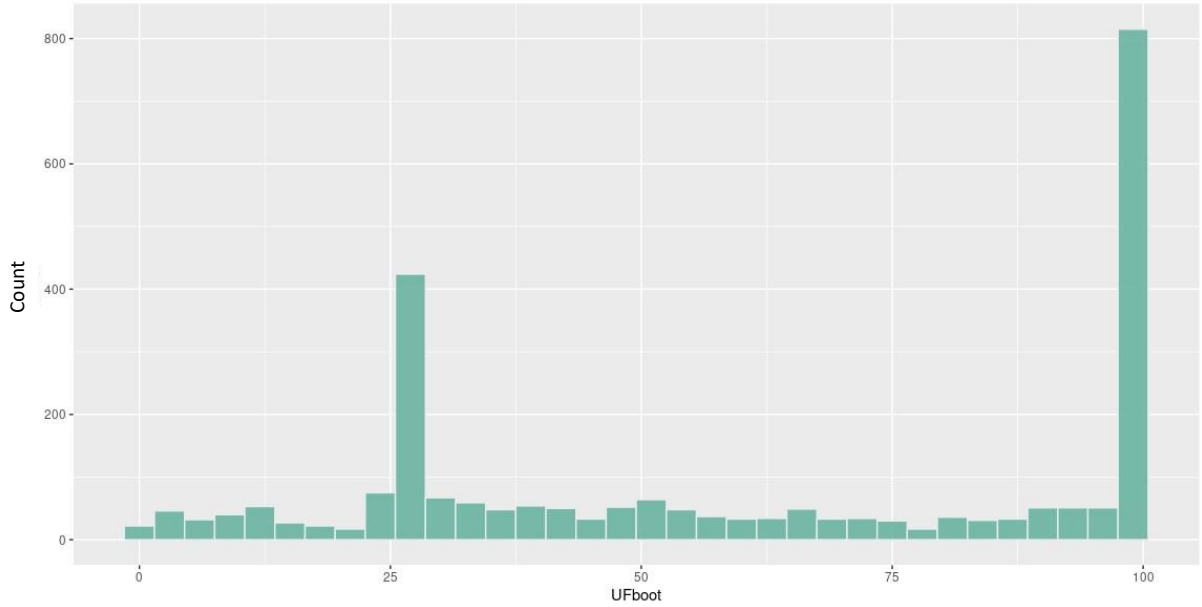
**Apêndice IV -** Distribuição temporal das aproximadamente 3.000 sequências genômicas de SARS-CoV-2 subamostradas dos 86.000 totais. As cores representam as regiões do Brasil no qual os genomas foram sequenciados. No eixo y temos a contagem absoluta das sequências selecionadas.



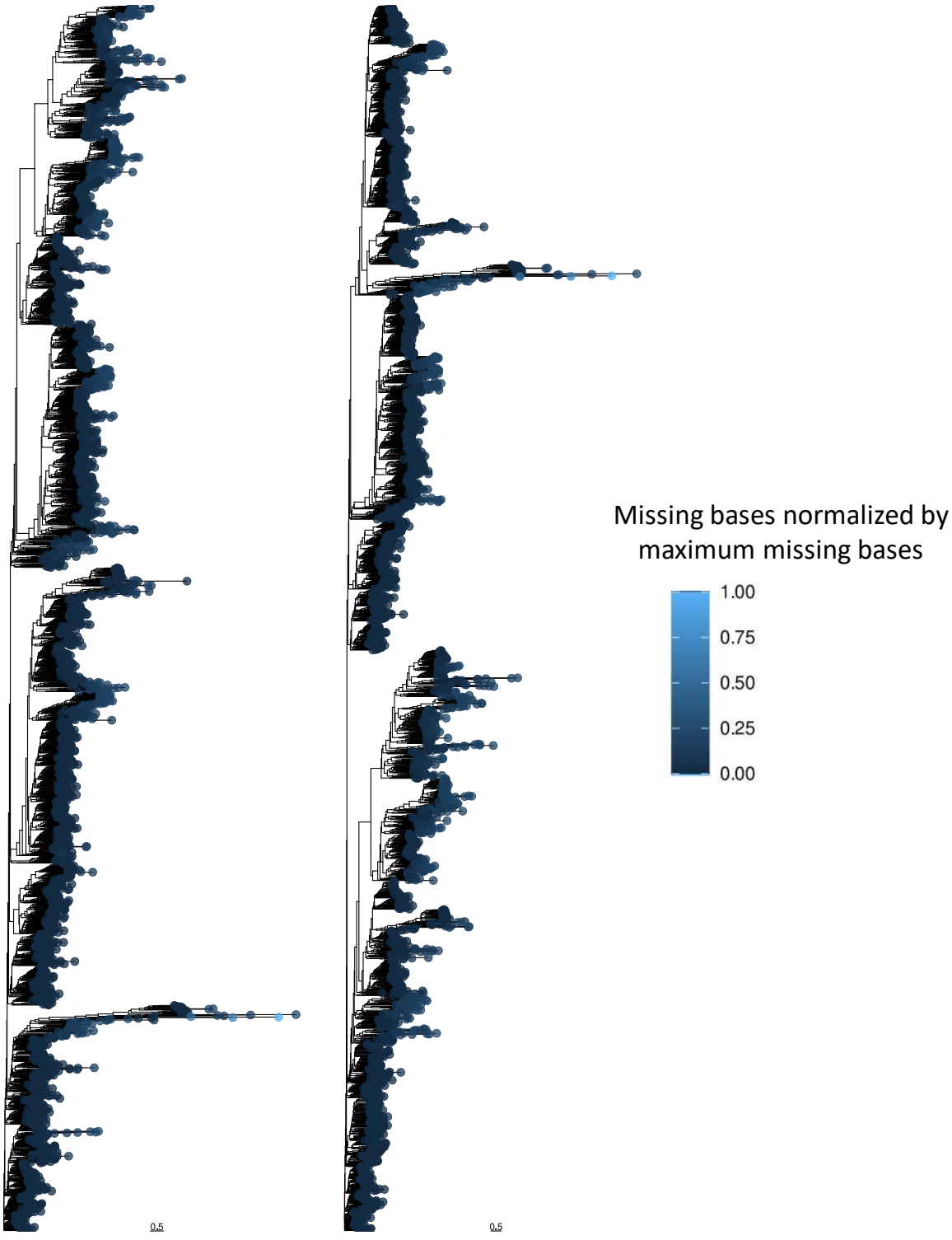
**Apêndice V -** Likelihood-mapping feito pela ferramenta IQTREE2 (modelo de substituição GTR+I+G) para as sequências subamostradas do Brasil. No triângulo superior, observamos as plotagens de verossimilhança das árvores dos quadruplets amostrados, à esquerda observamos a porcentagem distribuída nos 3 cantos do triângulo total e à direita temos a porcentagem dos pontos que ficaram em cada uma das regiões relacionadas a amostras com -sinais filogenéticos (88,7%), amostras com sinal net-like (4,1%) e amostras com sinais star-like (7,1%).



**Apêndice VI -** Distribuição dos valores de Ultrafast bootstrap para nossa árvore de referência de SARS-CoV-2 brasileiros gerada por máxima-verossimilhança.



**Apêndice VII –** Árvores geradas pelo algoritmo NJ a partir dos 4-mer natural vector (à esquerda) e dos cumulative 4-mer natural vector (à direita). As cores representam as quantidades de bases faltando, normalizadas pela quantidade máxima de bases faltantes.



**Apêndice VIII -** Gráfico de violino com a distribuição dos valores de bootstrap dos ramos internos das árvores geradas a partir de diferentes representações numéricas dos 3000 genomas de SARS-CoV-2 brasileiros.