VIGILÂNCIA EM SAÚDE: A IMPORTÂNCIA DA PADRONIZAÇÃO DE DADOS PARA CONTROLE DA PANDEMIA DA COVID-19

Health Surveillance: the importance of data standardization to control the COVID-19 pandemic.

Manuela Leal da Silva[22]
Maria Fernanda Ribeiro Dias[23]
Paula Lopes Cascabulho[24]
Isadora Nogueira Camelo[25]
Davi Ventura da Silva[26]
Diego Henrique Silvestre[27]
Evenilton Pessoa Costa[28]

Resumo: Neste trabalho, avaliamos a consistência na padronização dos dados genômicos depositados no banco de dados GISAID, considerando os filtros disponíveis pela própria plataforma. Os dados genômicos filtrados passaram por padronização dos rótulos das variáveis: "status do paciente", "sexo" e "idade". Rótulos redundantes foram transformados e categorizados, conforme o objetivo deste trabalho. Foram criados quatro rótulos (Vivo, Hospitalizado, Morto e Desconhecido) para a variável "status do paciente"; três rótulos (Feminino, Masculino e Desconhecido) para a variável "sexo"; e dois rótulos (apenas números reais e desconhecido) para a variável idade. Em seguida, analisamos o perfil global de genomas de SARS-CoV-2 sequenciados e depositados no GISAID, separando-os por continente. Analisamos também a distribuição de variantes do SARS-CoV-2 exclusivamente na América do Sul, bem como a distribuição das VOC e VOI conforme a idade e sexo dos pacientes. Foi feita uma análise quantitativa de sequências genômicas após aplicação dos filtros "localização geográfica", "genoma completo", "hospedeiro" e "status do paciente". Construímos um mapa de distribuição das variantes de SARS-CoV-2 pela América do Sul, destacando a participação de cada variante por todo o continente. Finalmente, mostramos a importância da vigilância genômica na descoberta de novas variantes de SARS-CoV-2. Além disso, observamos que a ausência de padronização no depósito de dados genômicos no GISAID levam a perda de uma quantidade valiosa

[22] Doutora em Ciências, realizou estágio de pós-doutoramento na École Normale Supérieure de Cachan junto ao Laboratoire de Biologie et Pharmacologie Appliquée (LBPA) (França) e estágios de pós-doutoramento junto ao Instituto de Biofísica Carlos Chagas Filho - UFRJ e ao Instituto Nacional de Metrologia, Qualidade e Tecnologia. É Professora Adjunta do Instituto de Biodiversidade e Sustentabilidade (NUPEM/UFRJ) da Universidade Federal do Rio de Janeiro - UFRJ, Docente Permanente e Coordenadora local do Programa de Pós-Graduação Multicêntrico em Ciências Fisiológicas da UFRJ (PMPGCF), Docente Permanente e Coordenadora Adjunta do Programa de Pós-Graduação em Biotecnologia do Inmetro (PPGBiotec) e Docente colaboradora do Programa de Pós-graduação em Biologia Computacional e Sistemas do Instituto Oswaldo Cruz (PGBCS/IOC/Fiocruz).
[23]Doutora em Ciências. Professora na Secretaria Estadual de Educação do Espírito Santo (SEDU/ES). marfedias@gmail.com. Orcid: 0000-0002-0069-3199
[24]Graduanda em Biomedicina pela Universidade Católica de Petrópolis. Iniciação científica. lopescascabulho@gmail.com. Orcid: 0000-0003-4308-0410.
[25]Graduanda em Ciências Biológicas do Instituto de Biodiversidade e Sustentabilidade (NUPEM/UFRJ) da Universidade Federal do Rio de Janeiro. Iniciação científica. E-mail isadoranogueiraok@gmail.com. Orcid: 0000-0001-6207-437X.
[26]Graduando em Ciências Biológicas do Instituto de Biodiversidade e Sustentabilidade (NUPEM/UFRJ) da Universidade Federal do Rio de Janeiro. Iniciação científica. davivilla09@gmail.com. Orcid 0000-0003-0544-4762.
[27]Mestre em Ciências Médicas. Doutorando do Programa de Pós-graduação em Nutrição da Universidade Federal do Rio de Janeiro. diegosilvestre@ufrj.br.Orcid: 0000-0003-3330-0256.
[28]Doutor em Biociências e Biotecnologia. Pós-doutorando do Programa de Pós-graduação Multicêntrico em Ciências Fisiológicas da Universidade Federal do Rio de Janeiro. eveniltonpessoa@gmail.com. Orcid: 0000-0003-3316-2834.

de dados. Em suma, esperamos que esta compilação de dados seja útil para as políticas públicas de saúde no controle da pandemia da COVID-19 e para a sociedade acadêmica como um todo.

Palavras-chave: Vigilância genômica; SARS-CoV-2; Banco de dados; Bioinformática; Mineração de dados.

Abstract: In this work, we evaluated consistency in the standardization of the genomic data deposited in GISAID database, considering the filters available by the platform itself. The filtered genomic data underwent standardization of variable labels: patient status, sex, and age. We transformed and categorized redundant tags according to the aim of this work. We Created four tags (Alive, Hospitalized, Dead, andUnknown) for the variable patient status, three for the variable sex (Female, Male andUnknown), and two for the age variable (only real and unknown numbers). Next, we analyzed the global profile of SARS-CoV-2 genomes sequenced and deposited in GISAID, separating them by continent. Then, we analyzed the distribution of SARS-CoV-2 variants in South America And The VOC and VOI distribution, according to the age and sex of patients. Next, we performed a quantitativeanalysisofgenomicsequencesafterapplyingthefilters' geographicallocation, complete genome, host, and patient status. We Constructed a distribution of SARS-CoV-2 variants across South America, highlighting the participation of each variant across the continent. Finally, we show the importance of genomic surveillance findings of the new SARS-CoV-2 variant and lack of standardization in the genomic data deposited in GISAID, leading the loss of valuable data. Together, we hope that this compilation of data could help public health policies handle the COVID-19 pandemic and for academic society as a whole.

Keywords: Genomic Surveillance; SARS-CoV-2; Database; Bioinformatics; Data mining.

1. Introdução

A Organização Mundial da Saúde (OMS) declarou, em 30 de janeiro de 2020, a COVID-19 como uma Emergência de Saúde Pública de Importância Internacional (ESPIN), posteriormente atingindo o status de pandemia. No Brasil, o governo intensificou medidas e recursos direcionados ao Centro de Operações de Emergência (COE) do Ministério da Saúde (MS) para monitorar a disseminação do vírus SARS-CoV-2 (OMS, 2020; Ministério da Saúde, 2020). O SARS-CoV-2 apresenta um genoma constituído de RNA fita simples, polaridade positiva (RNAfs+), não segmentado e com 27 a 32 quilobases (kb) de tamanho. Essas propriedades influenciam o surgimento de uma a duas mutações de nucleotídeos únicos por mês(1).

Dois anos se passaram, a pandemia continua em andamento, e o ano de 2022 se iniciou com centenas de variantes de SARS-CoV-2 mapeadas e circulando globalmente (2), e outras emergindo de países com baixas taxas de vacinação (3). O surgimento de novas variantes de SARS-CoV-2 levou as autoridades em saúde pública, tais como *Centers for DiseaseControlandPrevention* (CDC) e a OMS, a adotarem medidas para facilitar o controle da pandemia, como um sistema de classificação para as principais variantes, dividindo-as em grupos: variante de preocupação (*VariantofConcern*, VOC); variante de interesse (*VariantofInterest*, VOI) e variante sob monitoramento (*Variantsundermonitoring*, VUM) (4–6). Além disso, a OMS disponibilizou em agosto de 2021, um guia com orientações para vigilância de variantes de SARS-CoV-2, salientando a importância do sequenciamento genômico e o compartilhamento imediato das sequências, para entendimento e controle da pandemia (7).

A identificação de novas linhagens virais, enquanto elas estão em fase inicial de infecção, é uma das chaves para o controle de uma pandemia (8). A quantidade de genomas sequenciados e a inserção destes em bancos de dados mundiais, depende de sistemas de excelência no controle de vigilância epidemiológica. Devido à emergência de saúde global, mais de 10 milhões de genomas de SARS-CoV-2 já foram disponibilizados em bancos de dados públicos (2) como o*Global InitiativeonSharingAvian Influenza Data* (GISAID) e o *National Center for BiotechnologyInformation* (NCBI), que são plataformas de acesso aberto e acessíveis a pesquisadores em todo o mundo (9). O fornecimento desses dados evidencia os efeitos positivos de investimento em biologia computacional.

A divulgação e estudo em escala global do sequenciamento do material genético do SARS-CoV-2 facilitou o diagnóstico de novos casos de COVID-19 e identificou novas variantes, além de contribuir para que pesquisadores desenvolvessem vacinas eficazes contra

a doença (10). Embora a capacidade laboratorial de gerar dados referentes a sequenciamentos genéticos tenha aumentado muito, a capacidade de montar, analisar e interpretar esses dados genômicos têm sido mais difíceis de desenvolver. Para aumentar a capacidade analítica dos dados, os pesquisadores trabalham em duas extremidades: tornar a bioinformática e a análise genômica mais acessíveis aos não especialistas, e construir uma força de trabalho maior com experiência em bioinformática e epidemiologia genômica dentro da saúde pública. A melhora de desempenho dessas metodologias é utilizada para compor conjuntos de dados em métodos de Inteligência Artificial (IA), ferramentas que estão expandindo a capacidade da bioinformática, através de sondagem inteligente de genoma para classificação de características virais precisas (11).

No Brasil, o apagão de dados e a ausência de indicadores ampliam a vulnerabilidade frente à emergência sanitária atual e alertam para o aumento de ocupação de leitos de UTI (12). Esse cenário se intensifica devido a um surto de gripe, gerado pelo vírus influenza A (H3N2) (12), associado ao surgimento da nova variante da COVID-19, chamada Ômicron (13). O aumento exponencial dos casos exige contramedidas nos campos da vigilância epidemiológica, laboratorial e genômica (14).

Neste trabalho, abordaremos a importância da padronização de banco de dados de genomas voltado exclusivamente para a vigilância epidemiológica, e as dificuldades e limitações frente ao baixo investimento em Ciência e Tecnologia para com os centros de pesquisa dedicados ao sequenciamento de amostras de SARS-CoV-2. Os dados apresentados mostram que não só o Brasil, mas toda a América do Sul está aquém no quesito sequenciamento de amostras de SARS-CoV-2, quando comparado aos países desenvolvidos. Finalmente, apresentaremos neste trabalho uma forma eficaz e simplificada de montar, analisar e interpretar dados genômicos disponíveis, tornando a bioinformática e a análise genômica mais acessíveis aos não especialistas.

2. Materiais e Métodos

2.1 Extração e filtragem dos dados genômicos presentes na plataforma GISAID

Os dados genômicos utilizados neste trabalho foram obtidos através do banco de dados EpiCoV hospedado na plataforma GISAID (https://gisaid.org). Para a extração dos dados brutos iniciais foram contabilizadas todas as sequências de SARS-CoV-2, depositadas entre janeiro de 2020 até 10 de janeiro de 2022, e que respeitavam os seguintes parâmetros de interesse. Sendo eles:

i) Localização geográfica (países da América do Sul);

ii) Sequências completas (genoma do SARS-CoV-2);

iii) Hospedeiro (humano);

iv) Status do paciente (informações referentes ao estado clínico).

Após a seleção dos parâmetros mencionados acima, com exceção do status do paciente, foi realizado o download das sequências, com dois arquivos distintos para cada sequência. Um arquivo em formato FASTA, referente às sequências genômicas e um arquivo em formato TSV, em forma de tabela com informações sobre o vírus e o paciente infectado. As informações disponibilizadas são: idade, sexo, localização geográfica, comprimento da sequência genômica, data de coleta e data de submissão. O parâmetro referente ao status do paciente não vem inserido nas tabelas, e por esse motivo, foi necessário fazer um ajuste, com a adição de uma coluna contendo esse parâmetro, adicionado de forma manual.

A plataforma GISAID só permite o download de no máximo 5 mil sequências por vez. Por esse motivo a extração foi realizada de forma gradativa e posteriormente os arquivos em formato TSV foram unificados em apenas uma planilha.

2.2 Padronização e análise dos dados extraídos

A planilha final obtida da plataforma GISAID, de acordo com os parâmetros estabelecidos no item 2.1 foi padronizada, sendo excluídas as linhas em que os dados estavam parcialmente estruturados, inconstantes e/ou inexistentes.

A padronização e análise dos dados foram realizadas utilizando bibliotecas implementadas em linguagem *Python v3.9*, através da interface *Jupyter Notebook, software* livre com padrões abertos e serviços web para computação interativa em todas as linguagens de programação. As bibliotecas empregadas foram as seguintes: *Pandas v1.3.5*; *Numpy v1.19.3*; *Seaborn v0.11.2*; *Matplotlib v3.1.3*; *Plotly.express v0.4. 1*e*Scikit-learn v1.0.2*.

A padronização dos dados exigiu o tratamento de situações específicas:

- **Dados ausentes (*missing*):** Na checagem de valores faltantes na base de dados utilizamos o método *isnull*(), proveniente da biblioteca pandas. Esse método retorna *True* caso encontre valores do tipo NaN ou Nulo. Alguns dados faltantes foram tratados e mantidos, enquanto outros foram excluídos, havendo a eliminação de linhas, culminando na diminuição do conjunto de dados iniciais.
- **Padronização dos dados:** A padronização foi feita pelo pacote *preprocessing* da

biblioteca *Scikit-Learn*. Além disso, padronizamos as escalas dos dados utilizando a função *Standard Scaler()*.

2.3 Gráficos e mapas

Os gráficos e a análise descritiva dos dados foram produzidos usando os programas Jamovi v2.2.5 (https://www.jamovi.org/). Os mapas foram produzidos utilizando o QGIS v3.16 (https://www.qgis.org/pt_BR/site/index.html), o que permitiu a visualização, edição e a análise de dados georreferenciados. Os arquivos necessários foram obtidos através do Instituto Brasileiro de Geografia e Estatística (IBGE) no formato *shapefiles*.

3.Resultados

O GISAID compartilhava sequências do vírus da gripe desde 2006, e em 2020 começou a receber informações de sequências genômicas do SARS-CoV-2 e armazená-las no banco de dados EpiCoV em sua plataforma. Em 10 de janeiro de 2022, o banco de dados disponibiliza mais de 6 milhões de sequências de SARS-CoV-2, tornando-se uma referência mundial.

A parametrização dos dados foi realizada levando em consideração todos os depósitos feitos na América do Sul, os filtros disponíveis no GISAID e alguns desenvolvidos internamente. A padronização e concatenação dos dados foi feita de acordo com as variáveis internas do GISAID: “status do paciente”, “sexo” e “idade”. Cada variável apresentava diferentes rótulos para um mesmo parâmetro. Este fator resultou na necessidade de um fluxo de trabalho para parametrização dos dados obtidos **(Fig. 1)**.

A etapa de processamento dos dados resultou na divisão dos rótulos encontrados no GISAID nas seguintes categorias: Status do paciente (vivo, hospitalizado, morto ou desconhecido); Sexo (feminino, masculino e desconhecido); Idade (números reais e desconhecido) **(Fig. 1)**.

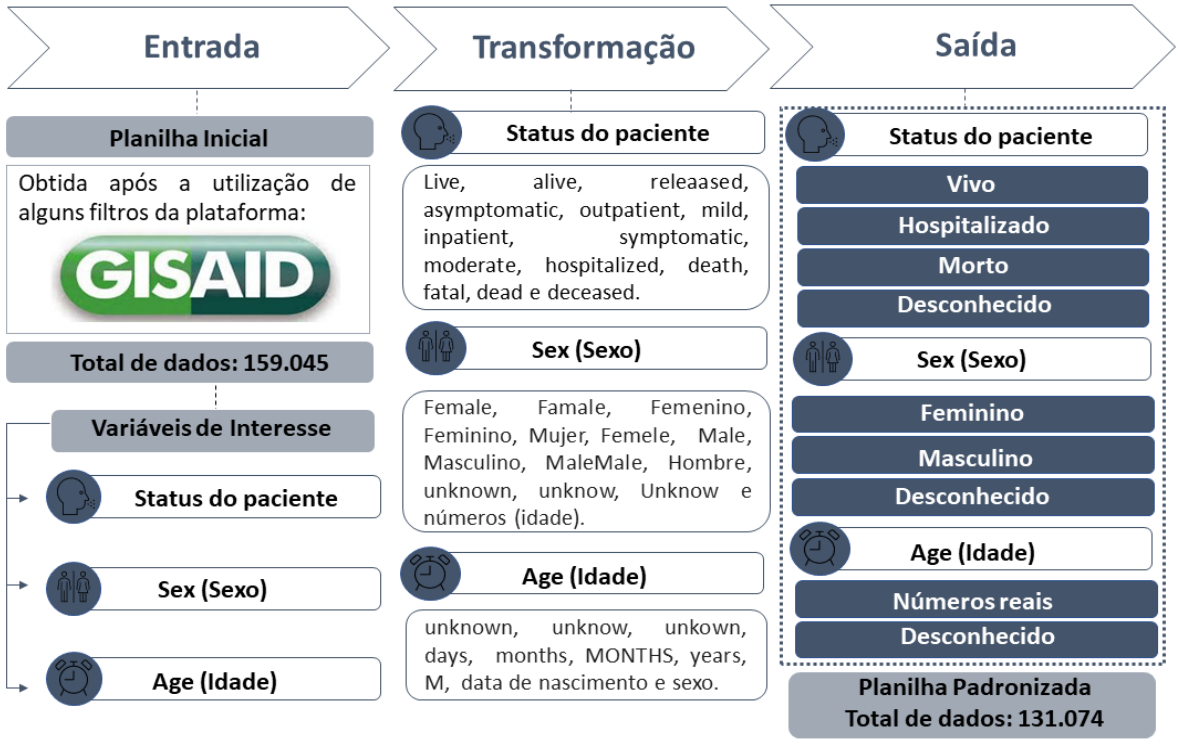

Figura 1. Representação esquemática do fluxo de trabalho adotado neste estudo. Todas as sequências do SARS-CoV-2 foram obtidas exclusivamente do GISAID. Os dados genômicos filtrados passaram por padronização dos rótulos das variáveis: "status do paciente", "sexo" e "idade". Rótulos redundantes foram transformados e categorizados, de acordo com o objetivo deste trabalho. Foram criados quatro rótulos (Vivo, Hospitalizado, Morto e desconhecido) para a variável "status do paciente"; três rótulos (Feminino, Masculino e Desconhecido) para a variável "sexo"; e dois rótulos (apenas números reais e desconhecido) para a variável idade.

3.1 Extração e filtragem dos dados genômicos presentes na plataforma GISAID

O depósito de genomas no banco de dados cresce a cada dia e até o momento em que foi realizada a última coleta de dados, o GISAID contava com 6.947.826 genomas depositados **(Fig. 2A)**. Após a aplicação dos filtros "Localização geográfica", "Genoma completo" e "Hospedeiro", os dados genômicos diminuíram para 159.045 sequências, expressando um corte de 97,71% sobre o total de sequências depositadas globalmente. Das 159.045 sequências baixadas, um subconjunto de 131.074 amostras foram utilizadas para as análises, visto que houve a necessidade de deletar sequências de conteúdos incoerentes ou ausentes (**Fig. 2B**).

Após a filtragem dos dados, observa-se que as VOCs*Gamma* e *Delta*, em conjunto, representam 87,35% das variantes de SARS-CoV-2 que predominaram na América do Sul (**Fig. 2**), sendo a variante *Gamma* a mais prevalente em todo o continente, seguida pela variante *Delta*. As VOCs*Alpha*, *Ômicron* e *Beta*, e as VOIs*Lambda* e *Mu*, representam 12,65% de todos os demais casos, tendo a variante *Beta*, a menor frequência dentro da América do Sul.

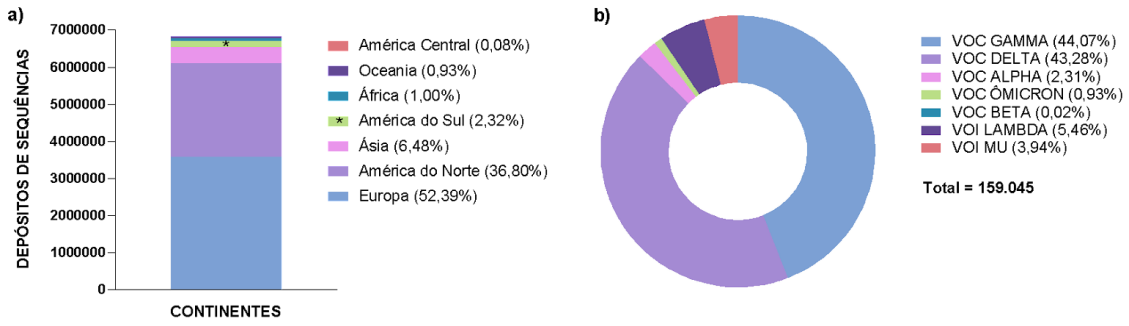

Figura 2. Perfil global de genomas de SARS-CoV-2 sequenciados e depositados no GISAID. **a)** Número de sequências genômicas por continente. **b)** Distribuição de variantes do SARS-CoV-2 na América do Sul. Os dados apresentados foram coletados entre 01/2020 até 01/2022 e levam em consideração os filtros disponíveis no GISAID (localização geográfica, sequências completas e hospedeiro).

A figura 3 apresenta os dados dos pacientes segundo a idade e o sexo dos infectados pelas variantes estudadas, apresentados em *boxplot*. Segundo os dados podemos observar que a mediana da idade dos pacientes varia entre 20 e 60 anos **(Fig. 3)**. Os elementos da caixa são definidos como: linha central (mediana ou segundo quartil), linha inferior (valor mínimo) e linha superior (valor máximo); a variação dos desvios padrão é determinada pelos pontos de dados mais discrepantes da mediana, denominados como *outlier*.

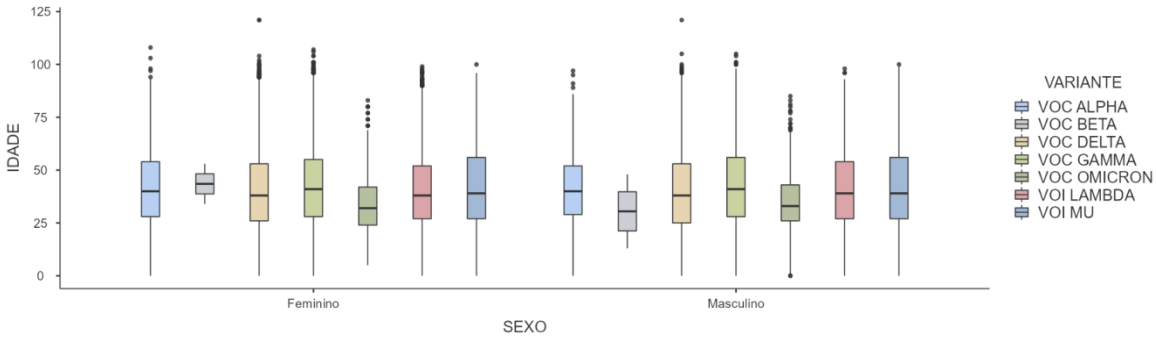

Figura 3. Análise da distribuição das VOC e VOI, de acordo com a idade e sexo dos pacientes.

3.2 Padronização e análise dos dados extraídos

Durante a extração de dados da plataforma GISAID, aplicamos um total de quatro filtros da própria plataforma para especificar os dados genômicos requeridos de acordo com os critérios iniciais estabelecidos para esse projeto. Estes filtros foram responsáveis pela maior redução de dados disponíveis, que contemplavam 6.947.826 de sequências depositadas no GISAID até o dia 10 de janeiro de 2022. Os filtros utilizados foram: Localização

geográfica (164.546), Genoma completo (159.069) e Hospedeiro (159.045) (**Fig. 4**). A exclusão dos dados faltantes resultou em 131.074 sequências genômicas, que compõem a planilha inicial.

O filtro "*Status* do paciente" resultou num total de 21.311 sequências genômicas, valor aquém do que se espera quando comparado à quantidade de dados depositados. Por essa razão, esse filtro não foi utilizado para eliminação dos dados, sendo atribuído uma nova coluna ao *dataset*, onde na linha referente a cada sequência que não possui essa informação, foi utilizada a categoria "Desconhecido" **(Fig. 4)**.

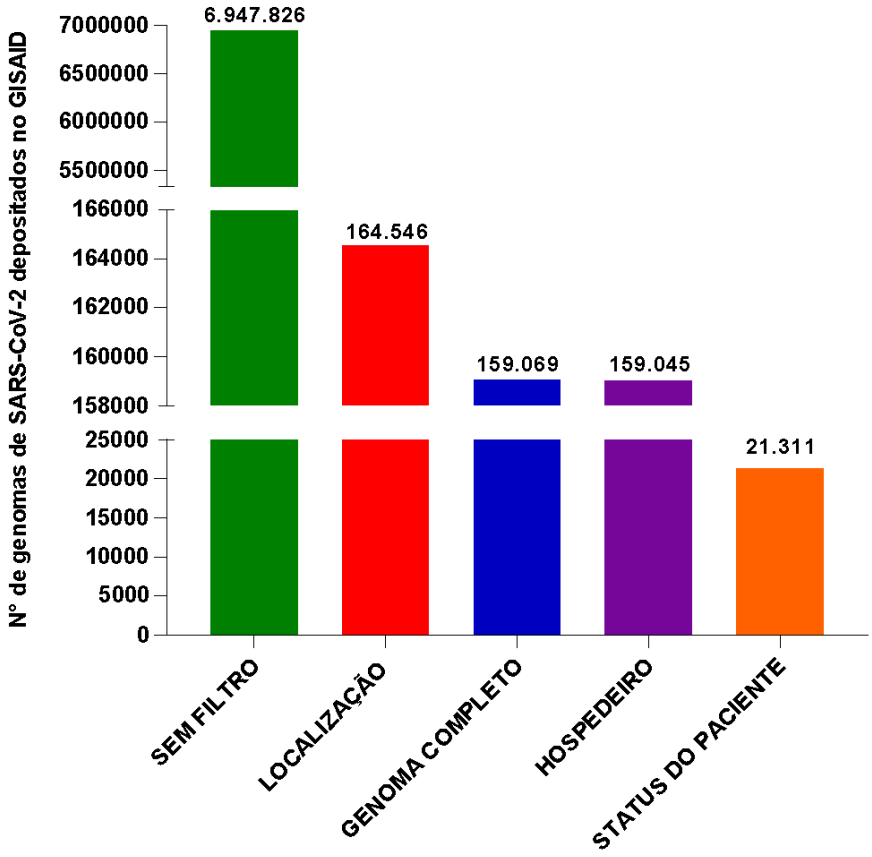

Figura 4. Análise quantitativa de sequências genômicas depositadas no GISAID após aplicação dos filtros "localização geográfica", "genoma completo", "hospedeiro" e"*Status* do paciente".O filtro "*Status* do paciente" foi utilizado de forma isolada, não sendo utilizado para exclusão de dados.

Após a aplicação dos filtros, foi possível traçar o perfil de disseminação das variantes em cada país da América do Sul. Essa análise nos dá um panorama das variantes mais prevalentes na pandemia em cada país da América do Sul (**Fig. 5**).

No mapa de distribuição de variantes de SARS-CoV-2 na América do Sul (**Fig. 5**) observa-se que as variantes de preocupação *Gamma* (VOC GAMMA) e Delta (VOC DELTA) tiveram um maior êxito em sua disseminação por esse continente, seguida de forma mais discreta pela variante *Alpha* (VOC ALPHA). A variante *Beta* (VOC BETA) foi a que

apresentou o menor número de casos em toda a região de estudo e por consequência, baixos índices de disseminação, tendo alguns registros oficiais apenas no Brasil (10), Suriname (4) e Guiana Francesa (2).

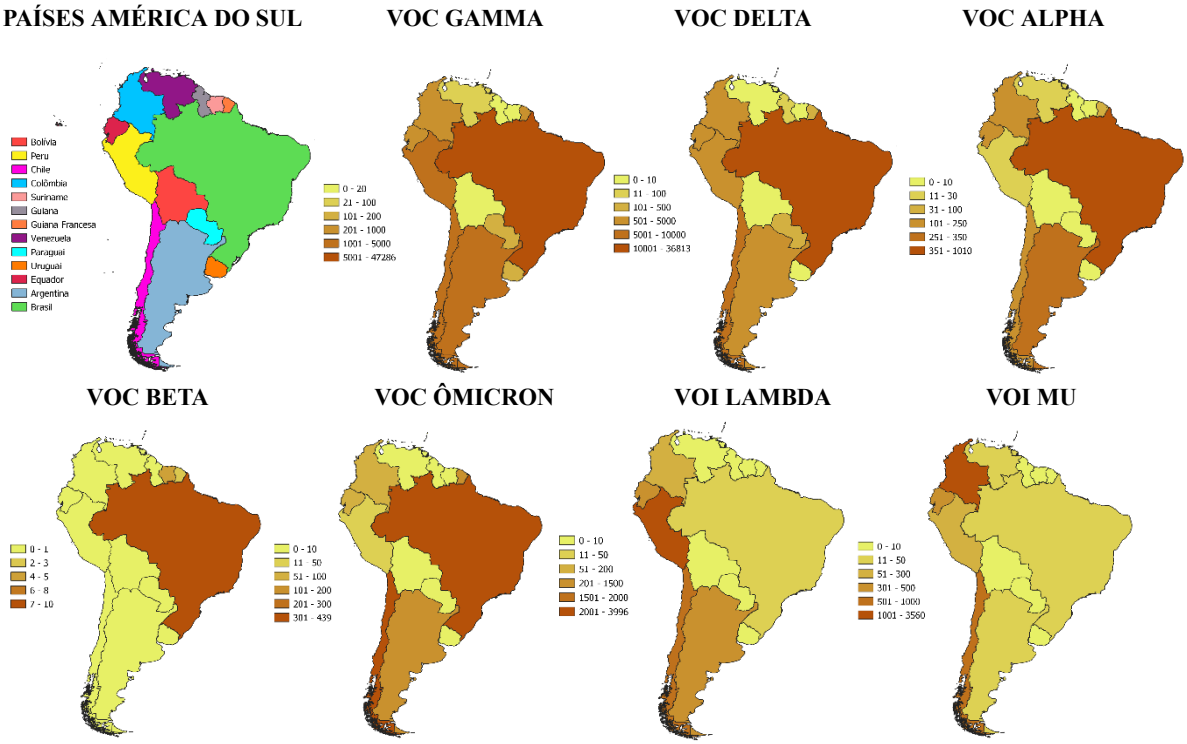

Figura 5. Mapa de distribuição das variantes de SARS-CoV-2 na América do Sul. Baseado nos dados obtidos no GISAID até o dia 10 de janeiro de 2022.

As variantes de interesse (VOI) *Lambda* e *Mu* expõem uma considerável diferença de distribuição pela América do Sul, ambas se disseminaram de forma majoritariamente desigual pelos países analisados. A variante Lambda apresenta maior grau de disseminação no Peru (3.996), país que a identificou inicialmente, e apresenta sucesso em seu espalhamento pelo Chile (1.800) e Argentina (1.093). Em contrapartida, a variante *Mu* possui um grande predomínio na Colômbia (3.560), local onde foi identificada pela primeira vez e apresenta um relativo aumento de ocorrências no Chile (839) e Equador (351), e baixos índices nos demais países sul-americanos (**Fig. 5**).

A recente variante *Ômicron* (VOC ÔMICRON), até o último dia de coleta de dados que compuseram o presente trabalho, tem apresentado maior prevalência no Brasil (439), seguido pelo Chile (349), Guiana Francesa (135) e Argentina (121) **(Fig. 5)**.

4.Discussão

Diante da pandemia de COVID-19 (15) e do aumento de variantes do SARS-CoV-2 (16), o crescimento exponencial da quantidade de dados biológicos, depositados em bancos de dados e oriundos do sequenciamento de genomas levanta alguns desafios, como a necessidade de armazenamento e o gerenciamento eficiente de grandes volumes de dados. Dentro da biologia computacional e, especificamente, na vigilância genômica, a extração de informações úteis a partir de dados biológicos demanda o desenvolvimento de ferramentas e métodos capazes de sincronizar e transformar os dados heterogêneos, em conhecimento biológico sobre o mecanismo subjacente estudado (17).

Desenvolvemos um *workflow* específico para este trabalho, focado na extração de informações da plataforma GISAID **(Fig. 1)**. Devido às dificuldades encontradas para a utilização dos filtros estipulados de forma padronizada, associadas principalmente à diminuição dos dados, foram utilizadas diferentes metodologias que auxiliaram na criação de um *dataset* estruturado. A utilização de fluxos de trabalho, como esse, é explorado na biologia computacional na etapa de pré-processamento dos dados e é um passo essencial para o sucesso das metodologias que irão utilizar esses dados, como é o caso de sistemas de aprendizagem de máquina (18). Na padronização dos dados extraídos do GISAID, os rótulos das variáveis '*Status* do paciente', 'Idade' e 'Sexo', possuíam vários sinônimos. Neste trabalho, adotou-se uma estratégia que auxiliou na aquisição das informações, além de uma mera descrição dos dados, fator que contribui efetivamente na vigilância epidemiológica, otimizando custos e tempo.

A obtenção de dados genômicos passa por um longo processo, desde coleta da amostra a ser sequenciada no hospedeiro, até o sequenciamento propriamente dito, utilizando metodologias confiáveis e, finalmente, chegando ao depósito das sequências em bancos de dados públicos(19). Esse processo demanda verbas públicas, materiais específicos e mão-de-obra especializada, sendo exatamente nesses pontos que se observa as imensas discrepâncias na amostragem populacional entre países desenvolvidos e os países em desenvolvimento, expressas na quantidade de amostras de sequências genômicas, depositadas no GISAID ou no NCBI, nos anos de 2020 e 2021, como mostrado na **Fig. 2**.

É importante ressaltar que o número de depósitos no GISAID começou a aumentar no final do ano de 2020 (20), e em fevereiro de 2021, o Reino Unido havia sequenciado mais da metade das 400 sequências genômicas de SARS-CoV-2 disponíveis na plataforma (21).Kalia e Sharma (2021) afirmam que pesquisadores do Reino Unido são os mais rápidos na inserção

de sequências nas plataformas públicas. O tempo médio de depósito é de 16 dias, sendo 5 vezes mais rápidos que países como Japão e Canadá (22). A organização e agilidade desse processo interfere na quantidade dos dados depositados.

A América do Norte e a Europa respondem hoje, por 89,19% de todos os genomas depositados no GISAID **(Fig. 2)**. Por outro lado, e demonstrando um expressivo contraste, a América do Sul e a África respondem 3,32%, uma diferença de 85,87%. Este dado evidencia a diferença de gastos e verbas públicas - destinadas à vigilância genômica e epidemiológica - quando comparamos países desenvolvidos e países em desenvolvimento.

Até o último dia de coleta de dados para o presente estudo, apenas 164.546 sequências correspondiam ao total das submissões dos 13 países constituintes da América do Sul, representando 2,37% do total de sequências depositadas na plataforma **(Fig. 2)**. Essa falta de acompanhamento da evolução viral prejudica a vigilância genômica e provoca uma resposta tardia dos órgãos de saúde pública.

O SARS-CoV-2 apresenta uma alta taxa de recombinação do seu DNA viral e por isso ele possui tantas variantes espalhadas por todo o globo. Mutações já foram descritas na proteína *spike*, local de clivagem da furina (entre os domínios S1 e S2), proteína do nucleocapsídeo, genes ORF, NSP3, NSP12, NSP14 e no gene M(23). Tais alterações podem acarretar numa diminuição da eficácia das vacinas, uma vez que as vacinas atuais utilizadas nas campanhas de vacinação são baseadas na sequência original da SARS-CoV-2 (24,25). Fica evidente, portanto, a importância do acompanhamento de novas variantes, por meio da vigilância genômica.

A figura 3 nos mostra que as variantes distribuídas na América do Sul têm uma taxa de infecção distribuída com uma idade mediana de 39 anos em ambos os sexos **(Fig. 3)**. Não foi possível traçar a relação entre níveis de infecção e taxa de mortalidade, sendo importante relatar o fato de que outros fatores adjacentes podem aumentar os perfis de óbito por COVID-19, como: obesidade, problemas cardíacos, doença renal crônica, ser fumante, etc.(26). Levin e colaboradores (2020), produziram um estudo baseado em metanálise e sugerem que o índice de fatalidade por COVID-19 aumenta progressivamente com a idade, em pacientes acima de 65 anos, estando com uma taxa de 0,4% aos 55 anos, 1,4% aos 65 anos, 4,6% aos 75 anos e 15% aos 85 anos (27). Contudo não podemos traçar este paralelo em nosso estudo devido a escassez de dados e ao enfoque do mesmo.

Tão importante quanto traçar o perfil de disseminação das variantes e depositar as sequências genômicas nas plataformas, está o processo de submissão dos dados, que necessita de informações pertinentes para complementar o dado de sequenciamento. Um processo de

submissão incompleto resulta em dados incoerentes e na necessidade de acréscimo de etapas no processo de vigilância genômica. No processo de ferramentas de inteligência artificial, a utilização de filtros para montagem de *dataset* constitui um passo importante e conta com a eficácia do pré-processamento. Quanto mais robustos os dados, melhor o desempenho do algoritmo e maior o ganho de tempo no processo de filtragem.

Este artigo aborda o pontapé inicial desse processo, com aplicação de filtros inerentes ao GISAID e outros, requeridos de acordo com o *workflow* **(Fig. 1)**.

A aplicação dos filtros, nos depósitos feitos na América do Sul, resultou na diminuição significativa do conjunto de dados de sequências genômicas, que se referem a 2,29% dos dados disponibilizados no GISAID **(Fig. 4)**. Isso representa a perda de 97,71% dos dados totais após a aplicação dos filtros: Localização geográfica (filtro que culminou na maior perda de dados), genoma completo e hospedeiro, fornecidos pela própria plataforma do GISAID **(Fig. 4)**. Esse fator alerta para a defasagem quantitativa de sequências depositadas por alguns países na plataforma e posterior dificuldade no processo de vigilância genômica e epidemiológica, por esses países. Alguns motivos são citados para justificar o processo tardio de sequências genômicas no GISAID. O primeiro deles é a incompreensão e não valorização da vigilância genômica e o segundo é a preocupação em reter informações, publicar ou patentear metodologias específicas (22).

Alguns trabalhos apresentam a disseminação de variantes em países como EUA e Índia (23,28), mas ainda não há descritos bibliográficos da disseminação de SARS-CoV-2 na América do Sul. Assim, damos o pontapé inicial para análises dessa região. Na figura 5, observamos a dispersão das variantes VOC e VOI nos países da América do Sul, sendo possível observar a alta concentração de dispersão das variantes *Delta, Alpha, Gamma, Lambda, Mu* e *Ômicron***(Fig. 5)**.

No Brasil, a variante *Gama* foi responsável pelo adoecimento em massa da população do Amazonas, uma vez que ela era mais transmissível, com um alto índice de internações hospitalares e de mortalidade (29), quando comparada com as linhagens dominantes da primeira onda. A variante *Gama* se espalhou pelo país, sendo responsável pela maior parte das infecções na segunda onda de COVID-19, entre janeiro e maio de 2021. Após um ano, a variante *Gama* deu espaço para as variantes *Delta* e *Ômicron*. Esta última apresenta curva crescente nos quadros de infecção registrados no país (12).

5.Considerações finais

Durante a pandemia de COVID-19, variantes de SARS-CoV-2 têm surgido e circulado

em todo o mundo, e autoridades em saúde pública, tais como Organização das Nações Unidas (ONU), *Centers for DiseaseControlandPrevention* (CDC) e Organização Mundial da Saúde (OMS), vem traçando estratégias para diminuir os danos causados. Nesse cenário, a vigilância genômica, associada a metodologias computacionais específicas - como métodos de IA- vêm sendo explorada, auxiliando na compreensão de padrões moleculares e predição de informações.

A aplicação de IA necessita de uma grande quantidade de dados pré-processados. O processo de obtenção desses dados esbarra em diversas dificuldades, desde a ausência de dados confiáveis até a falta de padronização nos bancos de dados.

A quantidade de sequências genômicas depositadas no GISAID é enorme, porém o rápido aumento do índice de sequenciamento gerou a submissão em larga escala desses dados - sem a preocupação com a padronização. Esse fator limita a utilização desses dados na aplicação de ferramentas de inteligência artificial, que necessitam de dados padronizados para o treinamento de um sistema. A observação desses fatores leva vários grupos de pesquisas a se debruçar sobre os bancos de dados existentes, observando e extraindo informações com foco em sua padronização (22).

Um sistema de vigilância genômica eficaz tende a permitir ao país portar um melhor controle no surgimento de potenciais variantes de SARS-CoV-2 e o andamento da pandemia. A padronização dos dados e um detalhamento de informações relevantes dos pacientes como sexo, idade, etnia, sintomas e os genomas sequenciados é útil para alimentar um sistema de aprendizagem de máquina como um aplicativo, capaz de predizer a possibilidade de disseminação viral em uma região ou auxiliar os profissionais de assistência à saúde a direcionarem tratamentos, por exemplo.

Diante disso, os dados de sequenciamento tornam-se um ponto chave, principalmente quando falamos de um banco de dados público e aberto como o GISAID, que teve um papel de fundamental importância para o compartilhamento de sequências e dados de SARS-CoV-2 ao longo da pandemia. Pontos assim só ressaltam o valor de domínios públicos de sequência que contribuem globalmente, não só para o combate de epidemias e pandemias, mas para prevenir ou aperfeiçoar as medidas de vigilância em saúde.

O uso de dados genômicos e a aplicação dos mesmos em sistemas de inteligência artificial, possibilita medir o grau de impacto, nacional e mundial, além de reforçar os cuidados e direcionamento para nações que necessitem de insumos para produção de vacinas e medidas de biossegurança, tornando as medidas de políticas e saúde públicas cada vez mais eficazes.

Referências:

1. Butera Y, Mukantwari E, Artesi M, Umuringa J d'arc, O'Toole ÁN, Hill V, et al. Genomicsequencingof SARS-CoV-2 in Rwandarevealstheimportanceofincomingtravelersonlineagediversity. Nat Commun [Internet]. 2021;12(1):1–11. Availablefrom: http://dx.doi.org/10.1038/s41467-021-25985-7
2. Romano CM, Melo FL. Genomicsurveillanceof SARS-CoV-2: A raceagainst time. Lancet RegHeal - Am [Internet]. 2021;1:100029. Availablefrom: https://doi.org/10.1016/j.lana.2021.100029
3. Kandeel M, Mohamed MEM, Abd El-Lateef HM, Venugopala KN, El-Beltagi HS. Omicronvariantgenomeevolutionandphylogenetics. J Med Virol. 2021;
4. WHO. Tracking SARS-CoV-2 variants [Internet]. WHO. 2022 [cited 2022 Feb 9]. Availablefrom: https://www.who.int/en/activities/tracking-SARS-CoV-2-variants/
5. Choi JY, Smith DM. SARS-CoV-2 VariantsofConcern. Yonsei Med J [Internet]. 2021 Nov 1 [cited 2022 Feb 10];62(11):961. Availablefrom: /pmc/articles/PMC8542474/
6. Instituto Butantan. OMS reclassifica gravidade e altera o grupo de variantes do SARS-CoV-2 - Instituto Butantan [Internet]. Instituto Butantan. 2021 [cited 2022 Feb 9]. Availablefrom: https://butantan.gov.br/noticias/oms-reclassifica-gravidade-e-altera-o-grupo-de-variantes-do-sars-cov-2
7. WHO. OMS declara emergência de saúde pública de importância internacional por surto de novo coronavírus - OPAS/OMS | Organização Pan-Americana da Saúde [Internet]. WHO. 2020 [cited 2022 Feb 7]. Availablefrom: https://www.paho.org/pt/news/30-1-2020-who-declares-public-health-emergency-novel-coronavirus
8. NanevSlavov S, Salvatore LeisterPatané J, dos Santos Bezerra R, Giovanetti M, Fonseca V, Jorge Martins A, et al. Genomicmonitoringunveiltheearlydetectionofthe SARS-CoV-2 B.1.351 lineage (20H/501Y.V2) in Brazil. medRxiv [Internet]. 2021 Apr 4 [cited 2022 Feb 7];2021.03.30.21254591. Availablefrom: https://www.medrxiv.org/content/10.1101/2021.03.30.21254591v1
9. Shu Y, McCauley J. GISAID: Global initiativeonsharingall influenza data – fromvisionto reality. Eurosurveillance [Internet]. 2017 Mar 30 [cited 2022 Feb 7];22(13):1. Availablefrom: /pmc/articles/PMC5388101/
10. The Lancet. Genomicsequencing in pandemics. Lancet (London, England) [Internet].

2021 Feb 6 [cited 2022 Feb 7];397(10273):445. Availablefrom: /pmc/articles/PMC7906659/

11. Ardakani AA, Kanafi AR, Acharya UR, Khadem N, Mohammadi A. Applicationofdeeplearningtechniquetomanage COVID-19 in routineclinicalpracticeusing CT images: Resultsof 10 convolutionalneural networks. ComputBiol Med [Internet]. 2020;121(March):103795. Availablefrom: https://doi.org/10.1016/j.compbiomed.2020.103795
12. FIOCRUZ. Observatório Covid-19 divulga boletim apenas com indicadores de leitos no SUS [Internet]. Rio de Janeiro; 2022 Jan [cited 2022 Feb 7]. Availablefrom: https://portal.fiocruz.br/noticia/observatorio-covid-19-divulga-boletim-apenas-com-indicadores-de-leitos-no-sus
13. WHO. Genomicsequencingof SARS-CoV-2: a guidetoimplementation for maximumimpactonpublichealth [Internet]. 2021 Jan [cited 2022 Feb 7]. Availablefrom: https://www.who.int/publications/i/item/9789240018440
14. Sallas J, Elidio GA, Rohlfs DB, De Medeiros AC, Guilhem DB. A vigilância genômica do SARS-CoV-2 no Brasil na resposta à pandemia da COVID-19. Rev Panam Salud Pública [Internet]. 2021 [cited 2022 Feb 7];45. Availablefrom: /pmc/articles/PMC8147732/
15. Holmes EC, Goldstein SA, Rasmussen AL, Robertson DL, Crits-Christoph A, Wertheim JO, et al. The originsof SARS-CoV-2: A critical review. Cell. 2021 Sep 16;184(19):4848–56.
16. Hirabara SM, Serdan TDA, Gorjao R, Masi LN, Pithon-Curi TC, Covas DT, et al. SARS-COV-2 Variants: DifferencesandPotentialofImmuneEvasion. Front CellInfect Microbiol. 2022 Jan 18;0:1401.
17. Larrañaga P, Calvo B, Santana R, Bielza C, Galdiano J, Inza I, et al. Machinelearning in bioinformatics. BriefBioinform [Internet]. 2006 Mar 1 [cited 2022 Feb 7];7(1):86–112. Availablefrom: https://academic.oup.com/bib/article/7/1/86/264025
18. Ekpenyong ME, Edoho ME, Inyang UG, Uzoka FM, Ekaidem IS, Moses AE, et al. A hybridcomputational framework for intelligentinter-continent SARS-CoV-2 sub-strainscharacterizationandprediction. Sci Rep [Internet]. 2021;11(1):1–25. Availablefrom: https://doi.org/10.1038/s41598-021-93757-w
19. Chiara M, D'Erchia AM, Gissi C, Manzari C, Parisi A, Resta N, et al. Next generationsequencingof SARS-CoV-2 genomes: Challenges, applicationsandopportunities. BriefBioinform. 2021;22(2):616–30.

20. Maxmen A. Onemillioncoronavirussequences: popular genome site hits mega milestone. Nature [Internet]. 2021 May 1 [cited 2022 Feb 7];593(7857):21. Availablefrom: https://pubmed.ncbi.nlm.nih.gov/33893460/
21. Burki T. Understandingvariantsof SARS-CoV-2. Lancet [Internet]. 2021 Feb 6 [cited 2022 Feb 7];397(10273):462. Availablefrom: http://www.thelancet.com/article/S0140673621002981/fulltext
22. Kalia K, Saberwal G, Sharma G. The lag in SARS-CoV-2 genomesubmissionsto GISAID. Nat Biotechnol [Internet]. 2021 Sep 1 [cited 2022 Feb 7];39(9):1058–60. Availablefrom: https://pubmed.ncbi.nlm.nih.gov/34376850/
23. Prajapat M, Handa V, Sarma P, Prakash A, Kaur H, Sharma S, et al. Update ongeographicalvariationanddistributionof SARS-nCoV-2: A systematic review. Indian J Pharmacol [Internet]. 2021 Jul 1 [cited 2022 Feb 7];53(4):310. Availablefrom: /pmc/articles/PMC8411960/
24. Wu F, Zhao S, Yu B, Chen YM, Wang W, Song ZG, et al. A new coronavirusassociatedwithhumanrespiratorydisease in China. Nature [Internet]. 2020 Mar 12 [cited 2022 Feb 10];579(7798):265–9. Availablefrom: https://pubmed.ncbi.nlm.nih.gov/32015508/
25. Dong Y, Dai T, Wei Y, Zhang L, Zheng M, Zhou F. A systematic review of SARS-CoV-2 vaccine candidates. SignalTransduct Target Ther [Internet]. 2020 Dec 1 [cited 2022 Feb 10];5(1). Availablefrom: https://pubmed.ncbi.nlm.nih.gov/33051445/
26. Voss JD, Skarzynski M, McAuley EM, Maier EJ, Gibbons T, Fries AC, et al. Variants in SARS-CoV-2 associatedwithmildorsevereoutcome. Evol Med PublicHeal [Internet]. 2021 Feb 26 [cited 2022 Feb 7];9(1):267. Availablefrom: /pmc/articles/PMC8385248/
27. Levin AT, Hanage WP, Owusu-Boaitey N, Cochran KB, Walsh SP, Meyerowitz-Katz G. Assessingthe Age SpecificityofInfection Fatality Rates for COVID-19: Systematic Review, Meta-Analysis, andPublicPolicyImplications. medRxiv [Internet]. 2020 Oct 31 [cited 2022 Feb 7];2020.07.23.20160895. Availablefrom: https://www.medrxiv.org/content/10.1101/2020.07.23.20160895v7
28. Jha N, Hall D, Kanakan A, Mehta P, Maurya R, Mir Q, et al. GeographicalLandscapeandTransmission Dynamics of SARS-CoV-2 VariantsAcrossIndia: A Longitudinal Perspective. Front Genet [Internet]. 2021 Dec 17 [cited 2022 Feb 7];12. Availablefrom: /pmc/articles/PMC8719586/
29. Freitas ARR, Beckedorff OA, Cavalcanti LP de G, Siqueira AM, Castro DB de, Costa CF da, et al. The emergenceof novel SARS-CoV-2 variant P.1 in Amazonas (Brazil)

wastemporallyassociatedwith a change in the age and sex profile of COVID-19 mortality: A populationbasedecologicalstudy. Lancet RegHeal - Am [Internet]. 2021 Sep 1 [cited 2022 Feb 7];1:100021. Availablefrom: http://www.thelancet.com/article/S2667193X21000132/fulltext