UNIVERSIDADE FEDERAL DE MINAS GERAIS

Instituto de Ciências Biológicas

Programa Interunidades de Pós-Graduação em Bioinformática

Joicymara Santos Xavier

THERMOMUTDB e SARS-CoV-2 AFRICA DASHBOARD: abordagens de ciência de dados para integração, análise e vigilância de dados biológicos

Belo Horizonte

2022

Joicymara Santos Xavier

THERMOMUTDB e SARS-CoV-2 AFRICA DASHBOARD: abordagens de ciência de dados para integração, análise e vigilância de dados biológicos

Versão Final

Tese apresentada ao Programa Interunidades de Pós-graduação em Bioinformática da Universidade Federal de Minas Gerais como requisito para obtenção do grau de “Doutora em Bioinformática”

Orientador: Dr. Douglas Eduardo Valente Pires
Co-orientadores: Dr. David Ascher e Dr. Tulio de Oliveira

Belo Horizonte
2022

043 Xavier, Joicymara Santos.
ThermoMutDB e SARS-COV-2 Africa Dashboard: abordagens de ciência de dados para integração, análise e vigilância de dados biológicos [manuscrito] / Joicymara Santos Xavier. – 2022.
167 f. : il. ; 29,5 cm.

Orientador: Dr. Douglas Eduardo Valente Pires. Co-orientadores: Dr. David Ascher e Dr. Tulio de Oliveira.
Tese (doutorado) – Universidade Federal de Minas Gerais, Instituto de Ciências Biológicas. Programa Interunidades de Pós-Graduação em Bioinformática.

1. Bioinformática. 2. Mutação de Sentido Incorreto. 3. Termodinâmica. 4. Betacoronavirus. 5. Ciência de Dados. 6. Base de Dados. 7. Proteômica. I. Pires, Douglas Eduardo Valente. II. Ascher, David Benjamin. III. Oliveira, Tulio de. IV. Universidade Federal de Minas Gerais. Instituto de Ciências Biológicas. V. Título.

CDU: 573:004

Ficha catalográfica elaborada pela bibliotecária Fabiane C M Reis – CRB 6 - 2680

UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS BIOLÓGICAS
PROGRAMA DE PÓS-GRADUAÇÃO EM BIOINFORMÁTICA

ATA DA DEFESA DE TESE

JOICYMARA SANTOS XAVIER

Às dezessete horas e trinta minutos do dia **27 de outubro de 2022**, reuniu-se, no aplicativo Zoom, a Comissão Examinadora de Tese, indicada pelo Colegiado do Programa, para julgar, em exame final, o trabalho intitulado: **"ThermoMutDB e SARS-CoV-2 Africa Dashboard: abordagens de ciência de dados para integração, análise e vigilância de dados biológicos"**, requisito para obtenção do grau de Doutora em **Bioinformática.** Abrindo a sessão, o Presidente da Comissão, **Dr. Douglas Eduardo Valente Pires**, após dar a conhecer aos presentes o teor das Normas Regulamentares do Trabalho Final, passou a palavra à candidata, para apresentação de seu trabalho. Seguiu-se a arguição pelos Examinadores, com a respectiva defesa da candidata. Logo após, a Comissão se reuniu, sem a presença da candidata e do público, para julgamento e expedição de resultado final. Foram atribuídas as seguintes indicações:

Professor(a)/Pesquisador(a)	**Instituição**	**Indicação**
Dr. Douglas Eduardo Valente Pires - Orientador	Fundação Oswaldo Cruz	Aprovada
Dr. David Ascher - Coorientador	The University of Queensland	Aprovada
Dr. Aristóteles Góes Neto	Universidade Federal de Minas Gerais	Aprovada
Dra. Marta Giovanetti	Fundação Oswaldo Cruz	Aprovada
Dr. Sandro Carvalho Izidoro	Universidade Federal de Itajubá	Aprovada
Dra. Lucianna Helene Silva dos Santos	Fundação Oswaldo Cruz	Aprovada
Dra. Valdete Maria Gonçalves de Almeida	Instituto Federal de Educação Ciência e Tecnologia do Norte	Aprovada

Pelas indicações, a candidata foi considerada: **Aprovada**

O resultado final foi comunicado publicamente à candidata pelo Presidente da Comissão. Nada mais havendo a tratar, o Presidente encerrou a reunião e lavrou a presente ATA, que será assinada por todos os membros participantes da Comissão Examinadora.

Belo Horizonte, 27 de outubro de 2022.

Documento assinado eletronicamente por **Marta Giovanetti**, **Usuária Externa**, em 27/10/2022, às 19:28, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Douglas Eduardo Valente Pires**, **Usuário Externo**, em 27/10/2022, às 20:05, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Lucianna Helene Silva dos Santos**, **Usuário Externo**, em 27/10/2022, às 20:06, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Aristoteles Goes Neto**, **Coordenador(a)**, em 27/10/2022, às 20:24, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Sandro Carvalho Izidoro**, **Usuário Externo**, em 28/10/2022, às 12:28, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Valdete Maria Gonçalves de Almeida**, **Usuária Externa**, em 31/10/2022, às 19:50, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **David Benjamin Ascher**, **Usuário Externo**, em 04/11/2022, às 15:49, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

A autenticidade deste documento pode ser conferida no site https://sei.ufmg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **1846463** e o código CRC **2027F96E**.

Referência: Processo nº 23072.262757/2022-61 SEI nº 1846463

Dedico este trabalho à minha mãe, que é e sempre será minha fonte de inspiração e força, e a meu pai, de quem herdei a calma e serenidade que balanceou a minha jornada até aqui.

Só enquanto eu respirar, vou me lembrar de você(s) - O Teatro Mágico

AGRADECIMENTOS

Agradeço imensamente a todas as pessoas que de alguma forma participaram da minha jornada até aqui. Nenhuma conquista é possível sozinho e eu sou privilegiada por ter as melhores (pessoas, instituições e financiadores) me apoiando.

Agradeço aos meus pais, Marivanda e Dalton (*in memoriam*), que sempre incentivaram meus estudos e, apesar de não mais estarem presentes fisicamente aqui, me acompanham diariamente na busca dos meus objetivos.

Aos meus irmãos, Geysa e Danilton, pelo suporte, companheirismo e por abraçarem meus sonhos junto comigo, inclusive ajudando com figuras, revisões e tudo que podem.

Ao Programa de Pós-Graduação em Bioinformática da UFMG por abrirem as portas para o desenvolvimento dessa pesquisa, em especial à Sheila e ao Tiago, que assim como toda a coordenação, não medem esforços para auxiliar os pós-graduandos.

Ao meu orientador Douglas, pela confiança e excelente condução do meu trabalho e além de tudo, suporte emocional e inspiração profissional. Ao meu co-orientador David, pela condução humanizada, paciência e sabedoria. Ao meu co-orientador Tulio, por ter abrido as portas para mim em um momento onde todos estávamos tão frágeis e ter me mostrado que fazer ciência vale a pena e pode ser divertido. Tenho muita sorte de ter sido supervisionada por vocês três.

À UFVJM por dar suporte ao meu afastamento para estudos. À FAPEMIG e à CAPES por fomentarem essa pesquisa através da bolsa de estudos e de doutorados sanduíche, respectivamente.

Aos colegas da Plataforma de Bioinformática da Fiocruz, em especial à Pâmela, Juliana, João, Amanda, Franscislon, Fausto, Gabriel e Fabiano pelo companheirismo, pelas risadas e por terem aceitado minha falta de habilidade com os *games*. Em especial à Pâmela, que foi o maior presente que o doutorado poderia me dar. Amiga, obrigada por todo suporte, por me ajudar a ser um ser humano melhor, por me incluir na sua família e por ser a melhor parceria que eu poderia ter.

A todas as mentoras do Código X por todos os momentos de troca e motivação.

À família CERI, nas pessoas da Cheryl, Monika, Houriiyah, Jennica, Marije, Nikita, Yajna, Arisha, Eduan, Yeshenee, Stephanie, San, Marta, Vagner, Zethu, Suzete e Tulio por acreditarem em mim e apoiarem minhas ideias.

À toda a minha família pelo carinho e apoio.

A todos os meus amigos pela torcida, orações e todo o apoio, principalmente Talissa, Rodrigo, Alicya, Pâmela, Danilo e Gael que também são como família pra mim.

À Deus e à Nossa Senhora pela proteção e alento nos momentos difíceis.

RESUMO

O crescimento exponencial na geração e disponibilização de dados biológicos, incluindo dados experimentais e genomas dos mais diversos organismos, impulsionou, nas últimas décadas, o surgimento de ferramentas computacionais que buscam predizer e entender uma variedade de fenômenos biológicos. Além disso, algumas iniciativas buscam informar e auxiliar na tomada de decisões sanitárias e de saúde. Para que seja possível o desenvolvimento desses e outros tantos trabalhos e análises relevantes, diversos bancos de dados têm sido mantidos como um recurso para orientar a inovação e a geração de novos *insights* biológicos. Nesse contexto, a mudança do paradigma de quantidade para qualidade de dados e informações têm sido cada vez mais necessária e crucial. No entanto, diferentes desafios são encontrados de acordo com a política de manutenção desses dados. As bases de dados utilizadas para pesquisas em Bioinformática, e também em outras áreas, podem ser abertas, onde a comunidade é responsável pela manutenção e verificação, ou mantidas por instituições que regulamentam a utilização desses dados. Neste trabalho, propusemos avaliar dois problemas relevantes em diferentes áreas de atuação da Bioinformática para entender como dados biológicos podem ser anotados e integrados de forma sustentável e otimizada a responder questões científicas relevantes e também informar a população. Como resultado, apresentamos duas abordagens que lidam com dois contextos diferentes, ThermoMutDB e SARS-CoV-2 Africa dashboard, tendo a qualidade de dados e entrega de informação científica como foco central. O ThermoMutDB é uma base de dados pública, manualmente curada, com dados termodinâmicos de proteínas. O SARS-CoV-2 Africa dashboard é uma ferramenta interativa para visualização e análises de dados genômicos de COVID-19 do continente africano. O ThermoMutDB propõe um paradigma colaborativo para verificação de dados para construção de bases de dados curadas com dados da literatura biomédica. O dashboard utiliza dados do GISAID (iniciativa que mantém e regula os dados genômicos dessa doença no mundo todo) através de um acordo entre as instituições e possibilita o acesso do público em geral a dados em tempo real que permitem a tomada de decisões em resposta à pandemia vigente. Resultados mostram que as ferramentas têm sido largamente utilizadas e têm potencial para impactar pesquisas futuras nas áreas de engenharia de proteínas e vigilância genômica, além da possibilidade de serem replicadas para outros contextos.

Palavras-chaves: bioinformática, mutações missense, termodinâmica, SARS-CoV-2, ciência de dados, bancos de dados, proteômica.

ABSTRACT

The exponential growth in the generation and availability of biological data, including experimental data and genome sequences of diverse organisms, has boosted, in recent decades, the emergence of computational tools that seek to predict and understand biological phenomena. In addition, some initiatives seek to inform and assist general public health and sanitary decision-making. In order to make it possible to develop these and many other relevant works and analyses, several databases have been maintained as a resource to guide innovation and the generation of new biological insights. In this context, the paradigm shift from quantity to quality of data and information has been proven increasingly necessary and crucial. However, different challenges are encountered according to the data maintenance policy. Databases used for research in Bioinformatics, and also in other areas, can be open, where the community is responsible for maintenance and verification, or maintained by institutions that regulate the use of this data. In this work, we proposed to evaluate two relevant problems in different areas of activity of Bioinformatics to understand how biological data can be annotated and integrated in a sustainable and optimized way to answer relevant scientific questions and also inform the population. As a result, we present two approaches that deal with two different contexts, ThermoMutDB and SARS-CoV-2 Africa dashboard, with data quality and scientific information delivery as a central focus. ThermoMutDB is a public, manually curated database of protein thermodynamic data. The SARS-CoV-2 Africa dashboard is an interactive tool for visualizing and analyzing COVID-19 genomic data from the African continent. ThermoMutDB proposes a collaborative data verification paradigm for building curated databases with data from the biomedical literature. The dashboard uses data from GISAID (an initiative that maintains and regulates the genomic data of this disease worldwide) through an agreement between the institutions. It allows the general public access to real-time data in order to guide decision-making in response to the current pandemic. Results show that the tools have been widely used and have the potential to impact future research in protein engineering and genomic surveillance, in addition to the possibility of being replicated in other contexts.

Keywords: bioinformatics, missense mutations, thermodynamics, SARS-CoV-2, data science, databases, proteomics.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SIGLAS

NAR	*Nucleic Acids Research*
IA	Inteligência Artificial
NGS	*Next Generation Sequence*
RMSE	*Root Mean Squared Error*
ML	*Machine Learning*
MLOPs	*Machine Learning Operations*
SARS-CoV-2	*Severe Acute Respiratory Syndrome Coronavirus 2*
GISAID	*Global Initiative on Sharing Avian Influenza Data*
CD	*Circular Dichroism*
DSC	*Differential Scanning Calorimetry*
Abs	Absorvância
Fl	Fluorescência
NMR	*Nuclear Magnetic Resonance*
mCSM	*mutation Cutoff Scanning Matrix*
COVID-19	Doença do coronavírus 2019
OMS	Organização Mundial da Saúde
CDC	Centro para Controle de Doenças e Prevenção
VOI	Variante de Interesse
VOC	*Variant of Concern*
VEME	*Bioinformatics Workshop on Virus Evolution and Molecular Epidemiology*

BIA	Laboratório de Bioinformática e Inteligência Artificial
API	*Application Programming Interface*
NLP	*Natural Language Process*
PMID	*Pubmed Identifier*
NIH	*National Institutes of Health*

SUMÁRIO

1. INTRODUÇÃO E JUSTIFICATIVA

Não é incomum encontrarmos atualmente a utilização de modelos de Aprendizado de Máquina para tentar responder às mais diversas questões biológicas existentes, como por exemplo: Quais mutações causam determinada doença? Como determinada mutação afeta a estabilidade de uma proteína? Quando a humanidade terá de enfrentar uma nova pandemia? [1], entre outras. Com o crescimento cada vez mais acelerado das técnicas de obtenção e interpretação de dados a partir de experimentos biológicos, responder computacionalmente essas questões têm ficado cada vez mais viável. Qualquer um, desde laboratórios individuais até instituições de larga escala podem, ou poderão em breve, gerar enormes quantidades de dados [1].

Diante da necessidade de guardar, catalogar, disponibilizar e garantir a confiabilidade e segurança dos dados, os bancos de dados biológicos são vistos como parte importante e essencial para a Bioinformática, Ciências da Saúde, Agrárias, entre outras. Inúmeras bases, que são fonte de dados para os mais variados métodos computacionais e principalmente de predição, estão disponíveis para as mais diversas áreas alvo da Bioinformática [2–4]. Fomentar e divulgar fontes de dados é, de fato, tão relevante para a Bioinformática, que a revista *Nucleic Acids Research* (NAR), desde 1993, dedica uma seção especial todo ano para publicações de bases de dados [5]. Até a escrita deste texto são 1.645 versões [6] diferentes de bases de dados biológicas publicados no NAR ao longo de 31 anos, incluindo publicações anteriores à criação da edição especial.

Por outro lado, apesar do esforço empregado no desenvolvimento, curadoria e divulgação de dados de qualidade, garantir que essas bases sejam compreensivas e livres de erros não é trivial. Depois de uma grande corrida ao longo dos anos em razão da maior disponibilidade de recursos para se gerar dados, impulsionados pelo avanço das técnicas de sequenciamento de nova geração, do inglês, *Next Generation Sequence* (NGS) [7,8], vivenciamos hoje novos desafios. O surgimento de ferramentas sofisticadas de Inteligência Artificial (IA) e o aumento do poder computacional, possibilitam análises e predição de comportamentos biológicos baseados em dados. No entanto, a quantidade de dados, outrora buscado para melhorar a precisão dos modelos, passa a não ser uma preocupação mais, dando espaço para o esforço relacionado à qualidade dos dados, em detrimento da quantidade.

A qualidade dos dados não só afeta as descobertas científicas como um todo, como também enviesa e limita ferramentas que são desenvolvidas para apoiá-las. [9] por exemplo,

analisa a performance de 16 preditores de estabilidade publicados ao longo dos anos e constatou que por mais de 15 anos, os esforços nessa área têm se estagnado em um RMSE (*root mean squared error*) médio de 1 kcal/mol. Embora ao longo desses anos, técnicas estatísticas e de IA das mais diversas tenham sido aplicadas, os problemas relativos aos dados permanecem os mesmos, o que impossibilita o avanço da área [9,10]. Além disso, erros em bases de dados podem agir em um efeito bola de neve, fazendo com que esses erros sejam propagados.

Uma iniciativa recente, chamada *Data Centric AI*, tem olhado para essas questões de qualidade de dados e ganhado bastante atenção por parte de pesquisadores e profissionais da área de Ciência de Dados em geral [11–14]. O movimento adverte para a mudança de foco, até então centrado nos ajustes de modelos e algoritmos de Aprendizado de Máquina (do inglês, *Machine Learning* - ML), para a sistematização e engenharia dos dados utilizados para treinar esses modelos.

Diante dessas mudanças de paradigma que estão em curso, a adoção das práticas que vêm sendo discutidas e sugeridas por movimentos como o *Data Centric AI*, pode ser crucial para se alcançar o próximo nível de maturidade para as ferramentas de Bioinformática, além de tirar vantagem da qualidade dos valiosos e numerosos dados biológicos que são gerados todos os dias, sendo eles, dados abertos, coletáveis ou mantidos por organizações. Além disso, o envolvimento da comunidade na melhoria desses dados, denominado *crowd reviewing*, pode ser crucial para a manutenção de repositórios de dados.

Os desafios relacionados a dados para Bioinformática, aqui focados no desenvolvimento de ferramentas de predição e análise, são diferentes dependendo, principalmente mas não restrito, dos mantenedores dos dados de interesse. Em geral, observa-se que bases de dados biológicas utilizadas para tarefas de predição e análise são bases de dados especializadas. Bases de dados especializadas são aquelas focadas em um organismo particular ou um tipo de dado específico, como por exemplo, bases de dados de sequência de um determinado patógeno [15–17], de propriedades de nucleotídeos [18], de efeitos de mutações [19], entre outras.

Bases de dados biológicos são em geral mantidos por grupos de pesquisa, consórcios ou iniciativas público-privadas. Quando não se tem uma organização financiando a manutenção daqueles dados, a dificuldade de atualização dos dados tende a ser maior. Por outro lado, bases de dados com dados sensíveis (de pacientes por exemplo) e que geralmente são atualizadas em tempo real, normalmente têm seus dados regulados e, portanto, acesso restrito. Enquanto essas bases de dados têm um fluxo bem estabelecido para atualização e

verificação, outras, dependem da anotação de usuários que depositam esses dados ou são curadas manualmente com dados advindos da literatura.

Para além das questões de curadoria e manutenção de bases de dados especializadas, observa-se que outra preocupação emergente tem sido em como transferir o conhecimento científico, principalmente em situações de interesse coletivo, para a sociedade em geral. Como em todas as áreas da ciência, a aplicação de paradigmas e ferramentas deve ser pensada no sentido de apoiar a descoberta científica, então, tendo isso em mente, essa tese de doutorado buscou compreender dois problemas relevantes relacionado a dados em Bioinformática e então propõe abordagens de Ciência de Dados que contribuem para o avanço das áreas relacionadas.

No primeiro caso, desejava-se desenvolver uma nova plataforma computacional para a identificação e predição de mutações compensatórias em proteínas. Mutações compensatórias são alterações que compensam o efeito de outra mutação, ou ainda, uma mutação que pode ser deletéria quando surge sozinha, mas é neutra quando combinada com outra mutação [20–22]. Mutações compensatórias podem desempenhar um papel relevante na resistência a antibióticos [20], na genética da conservação [23–25], além de terem sido associadas na formação de incompatibilidades de *Dobzhansky–Muller* [26], que diz respeito à evolução distinta de espécies. Todas essas evidências tornam o estudo de mutações compensatórias de interesse geral para a biologia evolutiva [27].

Para entender o efeito de uma determinada mutação em uma proteína, é preciso avaliar sua estabilidade na presença daquela mutação. Dessa forma, dados termodinâmicos como a temperatura em que a proteína perde a sua conformação original, a diferença de energia do estado enovelado para o desenovelado de uma proteína, entre outros, são essenciais para estudos que envolvem estabilidade. Como toda fase inicial de um projeto de Ciência de Dados, o primeiro objetivo envolvia definir um conjunto de dados de termodinâmica de proteínas, representativo e confiável, para treinar e testar os modelos de ML a serem desenvolvidos. Embora entender termodinâmica e consequentemente a estabilidade de uma proteína seja essencial para o entendimento dos efeitos causados por mutações, observou-se que a base de dados referência naquele momento apresentava problemas que comprometeriam a qualidade e confiabilidade dos estudos. Percebeu-se então que informações defasadas poderiam comprometer o estudo, assim como, a complexidade envolvida no processo de aquisição e armazenamento de dados. Nesse sentido, o ThermoMutDB, é proposto visando solucionar os problemas encontrados e também propõe uma abordagem colaborativa (*crowd reviewing*) para verificação e atualização dos dados.

A segunda abordagem surgiu da necessidade de esforços para conter, informar e tomar decisões frente à pandemia causada pelo vírus responsável pela síndrome respiratória aguda grave do coronavírus 2 (SARS-CoV-2, do inglês *Severe Acute Respiratory Syndrome Coronavirus 2*), identificado primeiramente na cidade de Wuhan na China em 2019 [28]. Logo após ser identificado, o vírus, se espalhou rapidamente, dando início a uma pandemia. Nesse sentido, o sequenciamento de genomas e vigilância têm sido fatores chave para o rastreamento e mitigação do desenvolvimento da pandemia causada pelo vírus SARS-CoV-2. Embora o continente africano tenha registrado uma proporção pequena de casos reportados globalmente e também número de mortes, alguns países do continente desempenharam papéis fundamentais em resposta à pandemia, através dos seus eforços de sequenciamento genômico. Assim como acordado com países do mundo todo, os dados genômicos do SARS-CoV-2 são depositados em uma base de dados única, o GISAID (do inglês, *Global Initiative on Sharing Avian Influenza Data*, www.gisaid.org). O GISAID é uma iniciativa público-privada global que provê acesso aberto a dados genômicos dos vírus influenza, coronavírus e monkepox [29,30]. Usuários que desejam utilizar os dados do GISAID devem aceitar as políticas de acesso e compartilhamento, assim como se identificar e informar como os dados serão utilizados. Uma vez que os dados disponibilizados pelo GISAID são de interesse não só científico mas também utilizados para informar a população e governo, mecanismos para que esses dados cheguem ao público alvo passam a se fazer cada vez mais necessários, dando espaço a ferramentas de Análise de Dados, como os *dashboards*.

Embora inúmeros *dashboards* tenham surgido ao longo da pandemia de COVID-19 apresentando vários dados globais e regionais, como número de casos, mortes reportadas e taxas de vacinação, nenhum focado no continente africano ainda havia sido desenvolvido. Levando-se em conta a enorme e valiosa quantidade de dados genômicos produzidos pelo continente africado e sua importância para entender a pandemia e ainda, por entender as limitações no uso e entendimento dos dados pelo público geral, existia uma notável necessidade de uma ferramenta analítica que pudesse prover a visualização desses dados em tempo real. Sendo assim, o SARS-CoV-2 Africa dashboard foi proposto com o objetivo de ser uma abordagem facilmente replicável para contextos semelhantes. A aplicação também foi desenvolvida visando ser uma ferramenta que permita que usuários interajam em tempo real com os dados gerados a partir das sequências genômicas depositadas no GISAID, ao mesmo tempo que não infringe seus termos de uso.

Em resumo, nesse estudo foram desenvolvidas duas abordagens de ciência de dados: o ThermoMutDB, uma base de dados termodinâmico de proteínas e o SARS-CoV-2 Africa

dashboard, uma ferramenta interativa para visualização de dados genômicos do vírus SARS-CoV-2.

Esta tese está organizada em formato de artigos, sendo que, no capítulo 2 são apresentados os objetivos do trabalho, os capítulos 3 e 4 se dedicam a embasar e descrever os dois estudos de caso realizados, através de uma introdução e a apresentação dos respectivos artigos na íntegra. Os capítulos seguintes se dedicam, respectivamente, à Discussão, Perspectivas Futuras e Conclusões.

2. OBJETIVOS

2.1. OBJETIVO GERAL

Avaliar problemas relevantes em diferentes áreas de atuação da Bioinformática e propor novos métodos de anotação e integração de dados biológicos que possam dar suporte à investigação de questões científicas relevantes de forma sustentável e otimizada. Além disso, as abordagens também devem ser capazes de informar a população utilizando dados atualizados e de qualidade.

2.2. OBJETIVOS ESPECÍFICOS

- Identificar pelo menos dois problemas relevantes em diferentes áreas de atuação da Bioinformática;
- Avaliar estado da arte em relação aos dados disponíveis para contorno dos problemas em cada área, indenficando lacunas que possam estar bloqueando os avanços das mesmas;
- Investigar e propor abordagens que permitam superar os desafios encontrados de forma otimizada e sustentável;
- Implementar abordagens propostas de forma que sejam reprodutíveis e replicáveis;
- Monitorar e manter as abordagens propostas.

3. TERMODINÂMICA DE PROTEÍNAS

Neste capítulo é apresentado o embasamento teórico necessário para o desenvolvimento do primeiro problema que foi estudado para essa pesquisa. Na primeira seção é discutido e definidos os conceitos acerca do efeito de mutações na estabilidade de proteínas. Na segunda seção, é descrito e apresentado como o efeito de mutações em proteínas podem ser analisados através de ferramentas computacionais.

3.1. O EFEITO DE MUTAÇÕES NA ESTABILIDADE DE PROTEÍNAS

Mutações são mecanismos eficientes utilizados pela natureza para introduzir diversidade em genomas, de modo a guiar a evolução. Mutações que alteram a sequência protéica e, portanto sua estrutura e função, chamadas mutações *missense*, são de interesse particular, dado seu papel em vários fenômenos biológicos relevantes. Esses fenômenos incluem predisposição a certos tipos de câncer [31,32], doenças hereditárias [33] e a emergência de resistência a medicamentos [34,35]. Os dados de pesquisas acerca desses fenômenos estão se acumulando rapidamente e os avanços em tecnologias emergentes devem ajudar a criar um ambiente de conhecimento favorável à obtenção de informações sobre genótipo-fenótipo. No entanto, elucidar as relações entre genótipo e fenótipo, continua sendo um problema complexo e interessante [36].

A sequência de aminoácidos de uma proteína determina sua estrutura tridimensional, função e estabilidade [37]. Entender e predizer o efeito de mutações na estrutura da proteína, função e no *fitness* do organismo é um grande desafio na biologia. Seu entendimento pode viabilizar prognósticos chave para se esclarecer as relações entre genótipo e fenótipo.

A estabilidade de uma proteína é medida pelo saldo líquido de energia que determina se ela está na conformação enovelada nativa ou em um estado desnaturado (desenovelado ou estendido). O estado nativo enovelado das estruturas proteicas é estabilizado por várias interações atômicas e de grupos, como hidrofóbica, eletrostática, ligação de hidrogênio, van der Waals e dissulfeto. O estado desenovelado é dominado por energias livres entrópicas e não-entrópicas. Dada a estrutura tri-dimensional aparentemente complicada de proteínas e o rápido enovelamento espontâneo, é uma espécie de paradoxo observar que sua estabilidade líquida seja tão pequena, tipicamente de 5-10 kcal/mol [38].

A determinação da estabilidade de uma proteína é feita através de vários experimentos, tais como, dicroísmo circular do inglês, *Circular Dichroism* (CD), *Differential*

Scanning Calorimetry (DSC), Absorvância (Abs), Fluorescência (Fl), Ressonância Magnética Nuclear, do inglês, *Nuclear Magnetic Resonance* (NMR), filtragem de gel, calorimetria isotermal e espalhamento de luz. Dentre esses métodos, CD, DSC, Fl e Abs são os mais utilizados para medir estabilidade [39].

A estabilidade de um estado nativo (N) é a mudança de energia livre (G) da transição de enovelamento (D), que é dado por:

$$\Delta G_{D \to N} = G(N) - G(D)$$

Onde, um estado nativo estável corresponde a um valor negativo de $\Delta G_{D \to N}$. Muitas mutações irão desestabilizar proteínas, aumentando a energia livre de Gibbs do estado nativo $G(N)$ e deixando a energia livre de Gibbs do estado desnaturado $G(D)$ relativamente inalterada. Isso torna $G(N) - G(D)$ menos negativo, deslocando o equilíbrio para longe do estado nativo [37].

A relativamente baixa estabilidade de estados nativos de proteínas naturais é medido quantitativamente por valores de

$$\Delta G = \Delta H - T\Delta S$$

Onde, H se refere à entalpia, S à entropia e T à temperatura. Estabilidades proteicas típicas a 25 ºC estão em um intervalo de -4,8 a -14,3 kcal mol^{-1} [37].

Estudos com mutantes naturais e projetados mostram os efeitos que modificações na sequência de aminoácidos podem causar na estabilidade de proteínas. Mutações diferentes, em diferentes posições, variam em importância. Uma pequena fração de resíduos parece ser absolutamente essencial para a estabilidade, sendo resistentes à substituição. Outros, dão contribuições significativas para a estabilidade. Nesses casos, a substituição diminui, mas não destrói completamente a estabilidade [37]. O efeito de uma mutação na estabilidade da proteína é dada através da diferença entre o ΔG do mutante e o ΔG da proteína selvagem, ou seja:

$$\Delta\Delta G = \Delta G_{D \to N}(selvagem) - \Delta G_{D \to N}(mutante)$$

Onde, valores negativos de $\Delta\Delta G$ implicam que o mutante desestabilizou a estrutura.

Como a maioria das mutações afeta a estabilidade e não a função [40], os níveis ou *fitness* das proteínas podem ser correlacionadas com os efeitos de estabilidade de mutações por um modelo de limiar simples, representado por uma sigmoide [41]. O *fitness* da proteína permanece inalterado enquanto a estabilidade permanecer acima de uma certo limiar (ΔG_t), como pode ser visto na Figura 1. No entanto, à medida que mais mutações vão se acumulando, o ΔG vai ficando menor que o ΔG_t e, portanto, a estabilidade diminui.

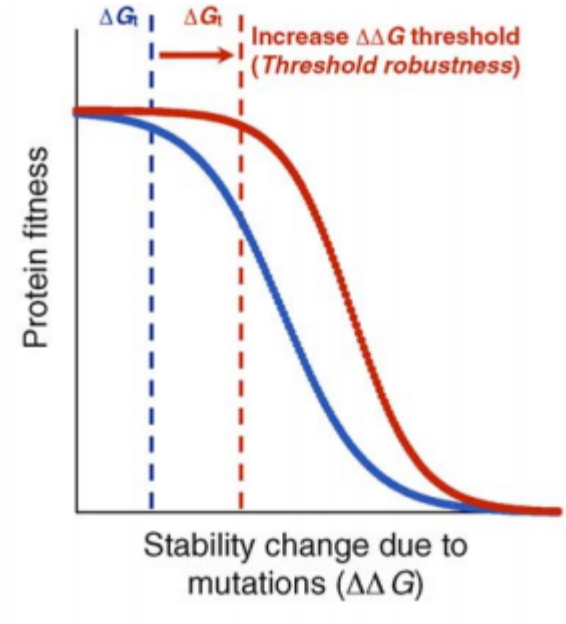


Figura 1. Modelo de limiar de robustez proposto por [42]. A natureza da sigmoidal indica que o *fitness* permanece praticamente inalterado desde que a estabilidade da proteína (ΔG) permaneça acima de um certo limiar (ΔG_t; linha azul). Aumentar o limiar de estabilidade resulta em maior tolerância a mutações ou à neutralidade (linha vermelha) (Adaptado de [42]).

3.2. ANÁLISE COMPUTACIONAL DO EFEITO DE MUTAÇÕES EM PROTEÍNAS

Elucidar experimentalmente os efeitos biofísicos de mutações é uma tarefa cara e demorada, usualmente limitada a um número pequeno de variantes com ensaios favoráveis. Com o passar dos anos, o acúmulo de informações caracterizadas experimentalmente permitiu o desenvolvimento e melhoria de ferramentas computacionais para análise de mutações. Essas ferramentas computacionais têm se mostrado indispensáveis para decifrar

correlações genótipo-fenótipo no câncer [43], doenças hereditárias [33] e detecção de resistência antimicrobiana [44], ajudando a guiar decisões clínicas e pesquisas futuras. Sendo assim, essa seção irá apresentar, de forma geral, uma plataforma de ferramentas, que é altamente aceita e utilizada pela comunidade científica para análise e predição de efeitos de mutações. Duas revisões acerca da plataforma foram publicadas como capítulo de livro e são intitulados "*A Comprehensive Computational Platform to Guide Drug Development Using Graph-Based Signature Methods*" e "*A Comprehensive Computational Platform to Guide Drug Development Using Graph-Based Signature Methods*", ambos publicados no *Methods in Molecular Biology [45,46]*. Os capítulos são anexados na íntegra após a breve discussão que é feita a seguir.

A plataforma *mutation Cutoff Scanning Matrix* (mCSM) [47] é uma extensa coleção de ferramentas *in silico* para predizer quantitativamente os efeitos de mutações missense no enovelamento, estrutura, dinâmica e interações em proteínas. Esta plataforma inclui ferramentas que lidam com mudanças na estabilidade proteica (mCSM-Stability, SDM e DUET), dinâmica e flexibilidade (DynaMut), iterações de proteínas com outras proteínas (mCSM-PPI e mCSM-PPI2), ácidos nucleicos (mCSM-DNA e mCSM-NA) e, ligantes e pequenas moléculas (mCSM-lig e CSM-lig). Um resumo desses métodos, bem como os links de acesso a cada um é mostrado na Tabela 1.

Tabela 1 : Ferramentas computacionais disponíveis na plataforma mCSM. (Adaptado e atualizado de [46]).

mCSM Tool	Função	URL
mCSM-Lig [48]	Prevê os efeitos de mutações de ponto único na estabilidade de um complexo proteína-ligante.	http://biosig.unimelb.edu.au/mcsm_lig/
mCSM-Stability [47]	Prevê os efeitos de uma mutação na estabilidade geral da proteína.	http://biosig.unimelb.edu.au/mcsm/stability
DUET [49]	Usa o mCSM-stability e o SDM2 para criar um consenso de predição sobre os efeitos de uma mutação na estabilidade	http://biosig.unimelb.edu.au/duet/

	da proteína.	
DynaMut [50]	Prevê os efeitos de uma mutação na estabilidade, flexibilidade e dinâmica proteica.	http://biosig.unimelb.edu.au/dynamut/
mCSM-PPI [47]	Prevê os efeitos de uma mutação dentro de uma proteína específica sobre seus impactos com as interações proteína-proteína no geral.	http://biosig.unimelb.edu.au/mcsm/protein_protein
mCSM-PPI2 [51]	Cria uma predição similar ao PPI, mas incorpora os efeitos de mutações em redes de interações inter-resíduos não covalentes usando kernels de grafos, informação evolucionária, métricas de redes complexas e termos energéticos.	http://biosig.unimelb.edu.au/mcsm_ppi2/
mCSM-DNA [47]	Prevê o impacto de mutações na interação de proteínas com DNA.	http://biosig.unimelb.edu.au/mcsm/protein_dna
mCSM-NA [52]	Prevê o impacto de mutações na interação de proteínas com ácidos nucleicos e usa farmacóforos e informação sobre propriedades dos ácidos nucleicos.	http://biosig.unimelb.edu.au/mcsm_na/
mCSM-AB [53]	Predição quantitativa dos efeitos de mutações missense na afinidade de ligação anticorpo-antígeno para guiar a engenharia racional de anticorpos.	http://biosig.unimelb.edu.au/mcsm_ab/
mCSM-AB2 [54]	Predições otimizadas dos efeitos de mutações na afinidade de ligação anticorpo-antígeno.	http://biosig.unimelb.edu.au/mcsm_ab2/

mmCSM-NA [55]	Prediz o efeito de múltiplas mutações na afinidade de ligação proteína-ácido nucleico	http://biosig.unimelb.edu.au/mmcsm_na/
cardioToxCSM [56]	Prediz seis tipos de resultados de toxicidade cardíaca.	https://biosig.lab.uq.edu.au/cardiotoxcsm
GRaSP-web [57]	Prevê resíduos do sítio de ligação do ligante putativo.	https://grasp.ufv.br
GASS-Metal [58]	Predição do sítio de ligação ao metal da proteína	https://gassmetal.unifei.edu.br/
cropCSM [59]	Plataforma computacional para ajudar a identificar herbicidas novos, potentes, não tóxicos e ambientalmente seguros.	http://biosig.unimelb.edu.au/crop_csm
CSM-carbohydrate [60]	Prevê a afinidade de ligação de complexos proteína-carboidratos.	http://biosig.unimelb.edu.au/csm_carbohydrate/
epitope3D [61]	Predição de epítopo de célula B conformacional	http://biosig.unimelb.edu.au/epitope3D

Essas ferramentas foram desenvolvidas utilizando o conceito de assinaturas baseadas em grafos [62], o qual representa a geometria e propriedades físico-químicas do ambiente estrutural de uma proteína selvagem como uma rede ou grafo. Essa representação é composta de uma série de nós, que descrevem o ambiente da mutação local, e arestas, que descrevem as distâncias entre camadas de interação entre os resíduos. A informação da mutação é capturada utilizando a mudança de farmacóforos entre o resíduo selvagem e o mutante, incluindo se doadores/aceptores de hidrogênio foram ganhos ou perdidos. Então, o conceito de assinaturas baseada em grafos pode ser definido, sem perda de generalidade, como um conjunto de características estruturais que, evidentemente, identifica objetos similares (por exemplo, mutações com efeitos semelhantes) que são, então, utilizadas como evidência para treinar e testar modelos preditivos.

A plataforma permite previsões biofísicas precisas, que, quando complementadas com outras ferramentas analíticas proteicas, podem prover uma visão detalhada do efeito específico de uma mutação em uma proteína. As ferramentas são implementadas em um *pipeline* analítico e preditivo, utilizando aprendizado de máquina supervisionado, para permitir a fácil e rápida caracterização de novas mutações e seus prováveis fenótipos clínicos. A Figura 2 ilustra, de forma geral, o funcionamento do *pipeline* utilizado pelas ferramentas da plataforma. O processo se inicia com a coleta e curadoria dos dados, depois são feitas as predições *in silico* que darão suporte para o estudo das consequências moleculares das mutações. Em seguida, as predições são correlacionadas com os dados biológicos utilizando-se abordagens de aprendizado de máquina, que inclui: validação de modelos, avaliação de métricas, teste cego, análise de performance, entre outras tarefas. Por fim é feita a aplicação e extrapolação do modelo.

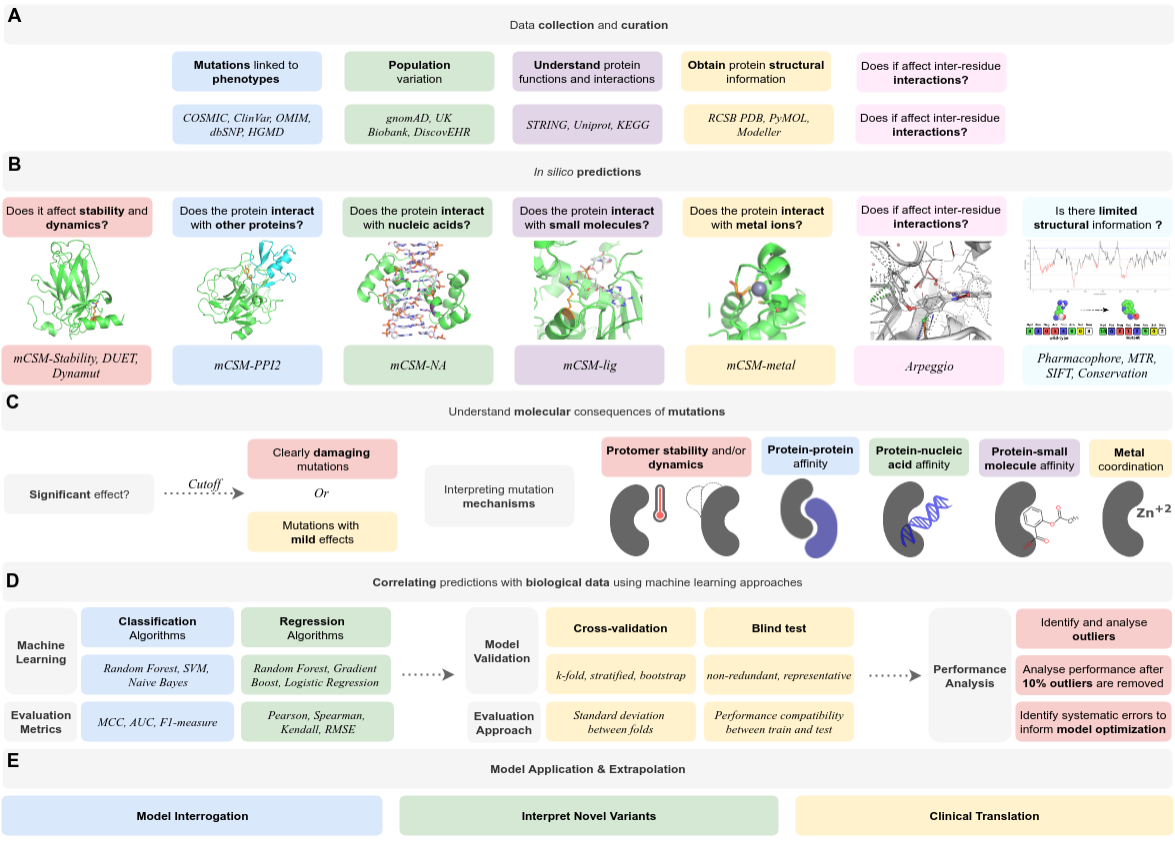


Figura 2: Visão geral do mecanismo de caracterização de mutações para entender suas consequências biológicas e guiar o desenvolvimento de ferramentas para predizer os resultados fenotípicos. Fonte: [46].

Foi demonstrado que essa abordagem tem grandes implicações no diagnóstico e na medicina personalizada na era pós-genômica [62–64]. Entretanto, ferramentas que lidam com os efeitos de mutações em proteínas baseadas em sequência ainda não estão disponíveis. Porém, para que ferramentas computacionais possam avançar nessa área, havia uma necessidade urgente de dados curados e atualizados.

Até o ano de 2020 o estado da arte em dados termodinâmicos de proteínas era a base de dados ProTherm [18], responsável pelos dados de treino de inúmeras ferramentas de predição, incluindo algumas das citadas na Tabela 1. No entanto, naquele momento, percebeu-se que o ProTherm não vinha mais sendo atualizado por 6 anos. Além disso, muitos dados depositados possuíam erros e algumas medidas estavam sem padronização. Sendo assim, a necessidade de curadoria dos dados e de se adquirir medições mais recentes se tornava bastante evidente. Para além da curadoria e aquisição de novos dados, uma abordagem que permitisse a contínua atualização e verificação dos dados também se fazia necessária. Dessa forma, o ThermoMutDB foi idealizado para ser uma nova base de dados termodinâmicos para mutações missense.

Chapter 7

A Comprehensive Computational Platform to Guide Drug Development Using Graph-Based Signature Methods

Douglas E. V. Pires, Stephanie Portelli, Pâmela M. Rezende, Wandré N. P. Veloso, Joicymara S. Xavier, Malancha Karmakar, Yoochan Myung, João P. V. Linhares, Carlos H. M. Rodrigues, Michael Silk, and David B. Ascher

Abstract

High-throughput computational techniques have become invaluable tools to help increase the overall success, process efficiency, and associated costs of drug development. By designing ligands tailored to specific protein structures in a disease of interest, an understanding of molecular interactions and ways to optimize them can be achieved prior to chemical synthesis. This understanding can help direct crucial chemical and biological experiments by maximizing available resources on higher quality leads. Moreover, predicting molecular binding affinity within specific biological contexts, as well as ligand pharmacokinetics and toxicities, can aid in filtering out redundant leads early on within the process. We describe a set of computational tools which can aid in drug discovery at different stages, from hit identification (EasyVS) to lead optimization and candidate selection (CSM-lig, mCSM-lig, Arpeggio, pkCSM). Incorporating these tools along the drug development process can help ensure that candidate leads are chemically and biologically feasible to become successful and tractable drugs.

Key words Graph-based signatures, mCSM, Mutation, Protein-ligand, Interatomic interactions, Docking, Drug development

1 Introduction

Structure-guided drug development uses knowledge of the three-dimensional structure of the biological target to more efficiently guide the design of small molecule binders. While it has become an integral strategy for both lead generation and optimization, the application of computational tools to take advantage of the explosion in structural information has often required specialist knowledge and resources and in some cases has been limited to commercial software.

Zoltán Gáspári (ed.), *Structural Bioinformatics: Methods and Protocols*, Methods in Molecular Biology, vol. 2112,
https://doi.org/10.1007/978-1-0716-0270-6_7,

Using the concept of graph-based signatures, we have developed a robust, user-friendly, and freely accessible platform to analyze protein structures and interactions [1–12] and guide disease characterization [13–28] and drug development [29–32]. These include methods to perform virtual screening (EasyVS), score protein-small molecule docking solutions (CSM-lig [3]), look at all the molecular interactions being made (Arpeggio [7]), identify mutations that are likely to affect compound binding (mCSM-lig [5]), and characterize the pharmacokinetic and toxicity properties of the proposed molecules (pkCSM [33, 34]). These have been successfully employed in a number of drug development projects [30–32, 35–37] and together comprise a powerful platform that allows users to enhance their structure-guided drug development efforts (Fig. 1). Here we discuss how this platform can be leveraged to guide drug development.

2 Materials

Here we present four structure-based tools to help guide drug development. For each method, users are required to provide:

1. **Wild-type protein structure in PDB format**: For all methods, a wild-type structure in the Protein Data Bank [38] format must be provided to perform the analysis. This can be an experimentally solved structure previously deposited into the Protein Data Bank (www.rcsb.org or http://www.ebi.ac.uk/pdbe/) or a model, for instance, obtained by comparative homology modeling. We have previously shown that homology models built using templates down to 25% sequence identity do not significantly affect the accuracy of the methods [9, 10]. For Arpeggio, CSM-lig, and mCSM-lig, the protein structure file needs to include the ligand of interest, either already present in the experimental structure or computationally docked into the binding site. PDB structures are required to have a valid chain identifier (*see* **Note 1**), a single conformation (multiple occupancies need to be filtered out; *see* **Note 2**), and a single model, in case of NMR structures (*see* **Note 3**).
2. **Three-letter code of the ligand of interest**: When a structure of a protein-ligand complex is provided to the predictive web servers (CSM-lig and mCSM-lig), users will be asked to provide a three-letter code that identifies the residue ID for that ligand within the PDB file, according to the PDB format standards. In addition to the three-letter code, CSM-lig also requires the canonical SMILES of the compound of interest for additional property calculations. Several tools are available to aid users to convert between small molecule formats. These include stand-alone packages such as OpenBabel [39] and Avogadro [40].

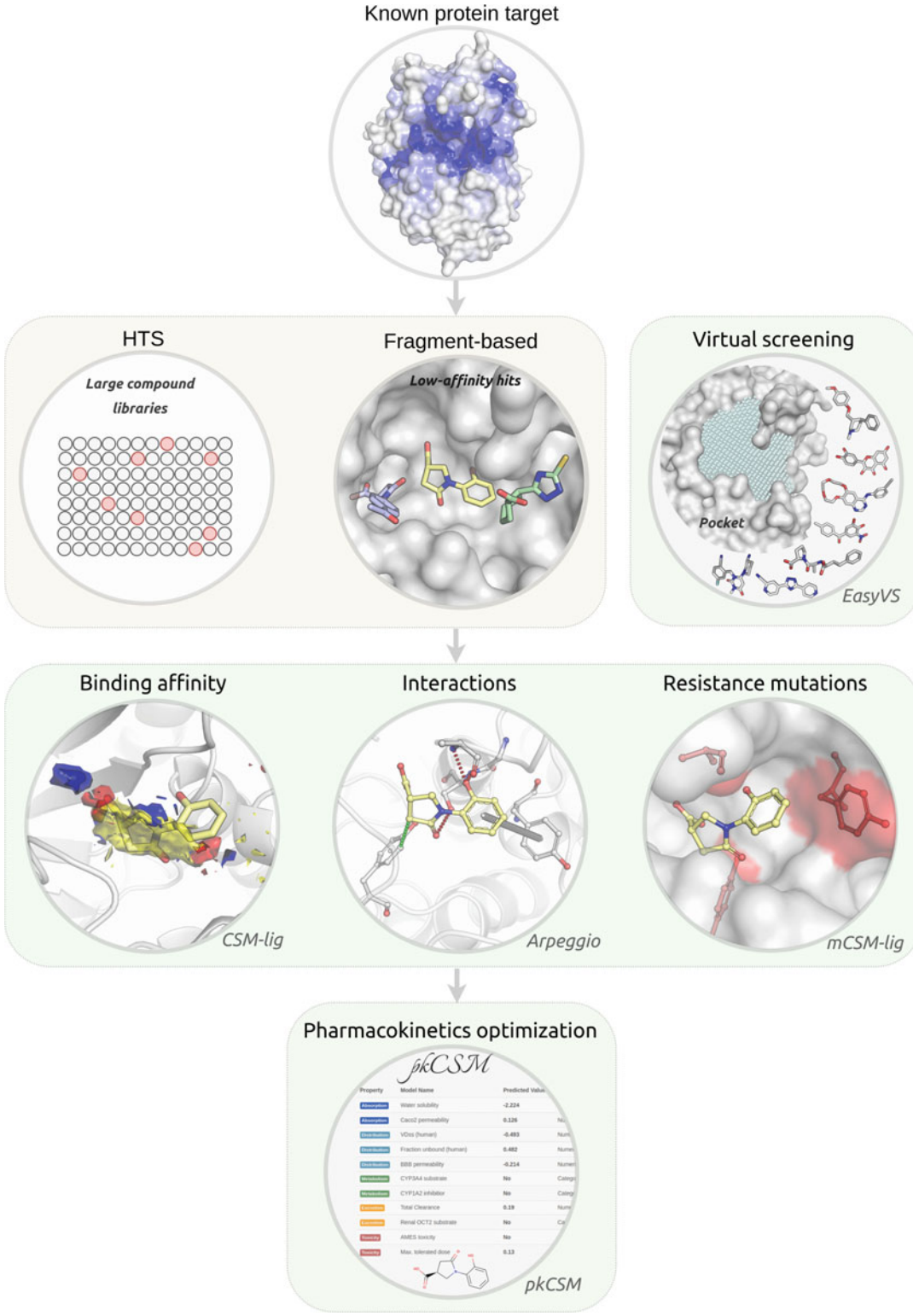


Fig. 1 A structure-based computational platform to guide drug development. To complement and support traditional experimental approaches, including high-throughput screening (HTS) and fragment-based drug discovery, this in silico platform supports hit identification via virtual screening, methods to better understand protein–small molecule interactions, affinity and effects of mutations, as well as the optimization of pharmacokinetic properties

3 Methods

3.1 Performing Automated Docking with EasyVS

1. Virtual screening is a powerful, high-throughput technique for computationally screening large libraries of small molecules (often in the order of millions) in order to identify those ligands which are most likely to bind to a drug target protein. When compared to traditional screening methods, this leads to significantly higher hit rates that can proceed to lead optimization [41, 42]. It can, however, be computationally intensive and usually requires specialist knowledge. EasyVS provides an easy-to-use web interface at http://biosig.unimelb.edu.au/easyvs/, allowing users to rapidly set up and analyze their virtual screening results.
2. Users can upload the structure of the protein target of interest as either a PDB file or by providing the PDB ID of a previously solved experimental structure. Any ligands, ions, or water molecules already bound to the provided structure will be disregarded.
3. On the following step, the provided PDB file or identifier will be processed, and pockets will be automatically detected using Ghecom [43] (Fig. 2a-1). Users can either select one of the identified pockets to determine the docking grid (the three-dimensional space where the ligands will be docked into) or provide specific grid coordinates and size (Fig. 2a-2).
4. Users then need to select the ligand library they want to screen, which includes libraries of purchasable compounds, natural products, or FDA-approved drugs (Fig. 2b). These can be further filtered based upon their molecular properties (e.g., Lipinski's rule of five [44] or the rule of three) or grouped by similarity.
5. The selected molecules will then be docked into the selected docking grid (Fig. 2c-1), and the top 20 poses per ligand can be downloaded. The server also provides an interactive visualization tool to compare ligand docking poses (Fig. 2c-2). The example on this figure shows the docking poses for ligands docked to the Ribosome-Inactivating Protein Ricin A (PDB ID: 1BR5). While poses are sorted by predicted affinity (kcal/mol) using autodock's scoring function, users can evaluate docking poses with alternative approaches, such as CSM-lig [3].

3.2 Predicting Protein-Small Molecule Affinity with CSM-lig

1. Following virtual screening or docking, the affinity of the top docked ligand poses can be quantified using CSM-lig. This is a machine learning-based tool which acts as a scoring function and enables the numerical affinity comparison between poses. It is implemented via an easy-to-use web interface at http://biosig.unimelb.edu.au/csm_lig, which is compatible with most operating systems and browsers.

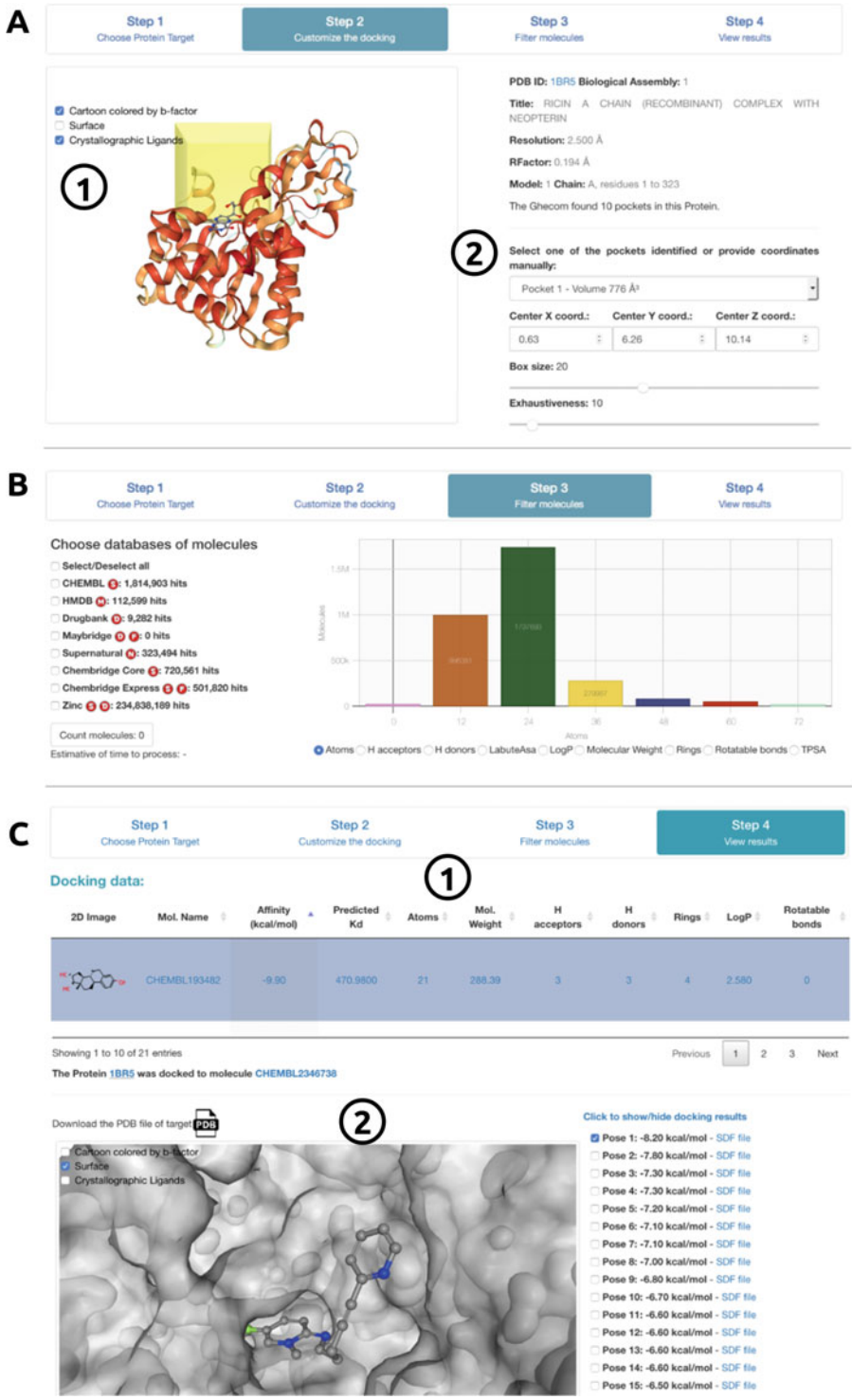


Fig. 2 Automated docking with EasyVS. After choosing a target of interest, EasyVS will automatically identify pockets (**a-1**) and allow user to further customize the docking protocol (**a-2**). A range of ligand libraries can be selected for docking (**b**), including FDA-approved drugs, purchasable compounds, and natural products, which can be further filtered based on physicochemical properties. Docking results are shown in tabular format (**c-1**), depicting ligands, their properties, and docking scores. An interactive viewer allows users to inspect the best poses for each ligand (**c-2**)

2. By selecting the "Predict" tab, users are presented with two job options, "Single Structure" and "Multiple Structures."
3. For "Single Structure" prediction, provide (Fig. 3a-1) the protein-small molecule complex you would like to evaluate the pose of in PDB format (Fig. 3a-2), the three-letter code for the small molecule (as in the provided PDB file) and (Fig. 3a-3) and the SMILES string of the small molecule.
4. Alternatively, for "Multiple Structures," provide two files. The first file (Fig. 3a-4) is a compressed zip file with all protein-small molecule PDB files you would like to evaluate. These could be, for instance, different poses or conformations for a given protein-ligand complex or multiple different complexes. The second (Fig. 3a-5) is a tab-separated file with the following information for each uploaded complex in the .zip file: (a) structure file name (file in PDB format), (b) three-letter code for the small molecule (as in the structure file), and (c) canonical SMILES for the small molecule.
5. The output prediction page for the "Single Structure" jobs depicted in Fig. 1b presents (Fig. 3b-1) the predicted affinity (as $-\log_{10}(\textit{affinity})$ in molar, meaning a compound with an affinity predicted as 1 nM would have a predicted value of 9). The example presented in the figure and the web server shows the affinity prediction for the ligand Zanamivir bound to human sialidase-2 (PDB ID: 2F0Z). For this complex, CSM-lig generates a score of 12.6, denoting very high affinity (larger numbers denote higher affinity). A depiction figure of the small molecule is shown, together with calculated properties, including molecular weight (in Da) and partition coefficient (log *P*), among others (Fig. 3b-2). An interactive visualization of the protein-small molecule complex is also exhibited (Fig. 3b-3). The interatomic non-covalent interactions between protein and small molecule are also calculated and are available as a downloadable Pymol [45] session (Fig. 3b-4). Pharmacokinetics and toxicity predictions by pkCSM for the provided small molecule are also available by clicking on the red button at the bottom-left corner of the results page.
6. The output for "Multiple Structures" jobs are shown in tabular format (Fig. 3c-1), depicting predicted affinity values, SMILES identifying the molecules and their calculated molecular properties. These results are available as a tabular file and can be downloaded (Fig. 3c-2).

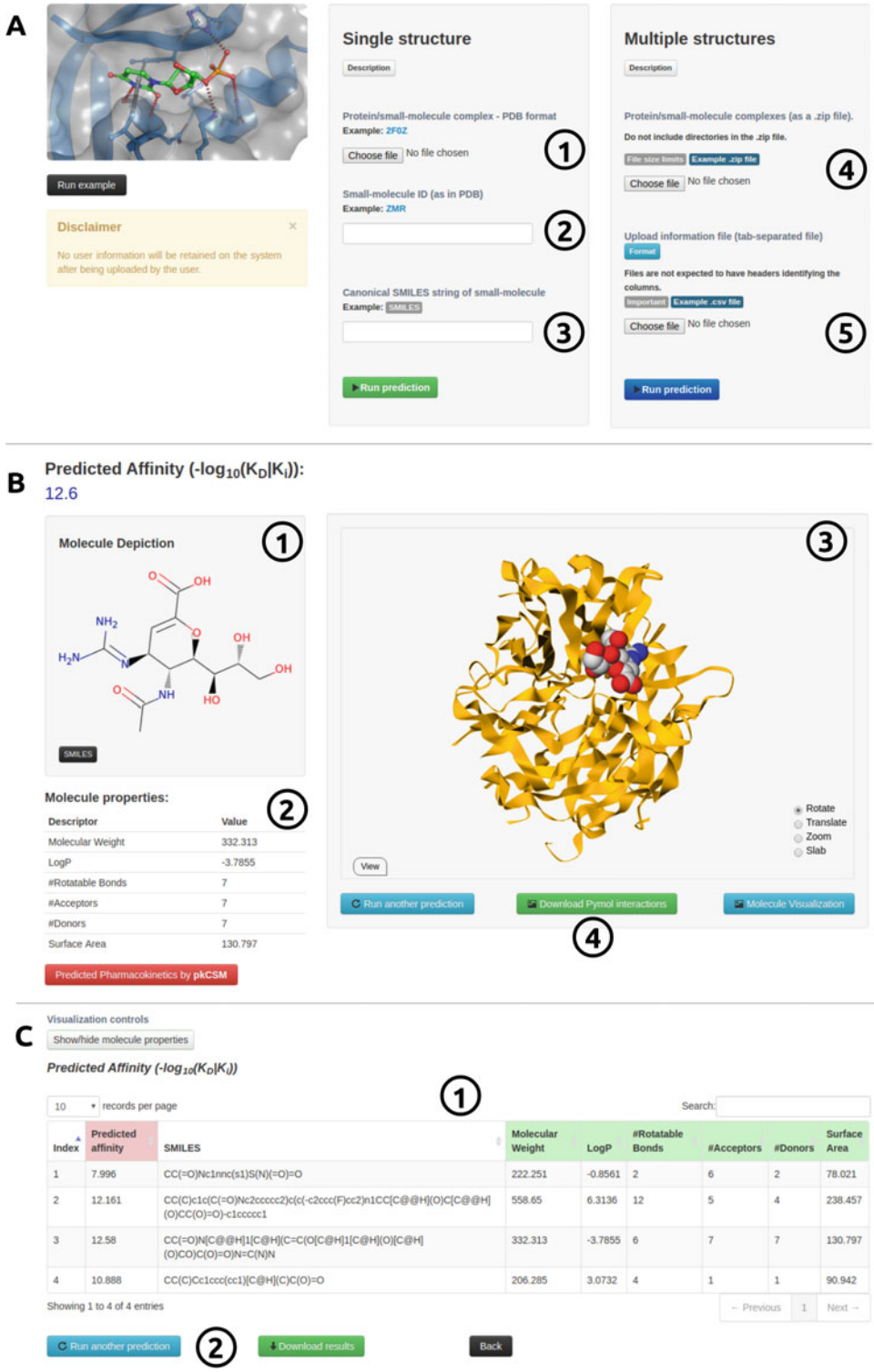


Fig. 3 CSM-lig submission and results web interface. The submission page (**a**) allows users to provide either single or multiple protein-ligand complexes for evaluation. The results page for single complex/pose assessment (**b**) provides the calculated affinity, ligand properties and depiction, as well as an interactive visualization of the complex. For multiple poses, CSM-lig provides the predicted affinities in a downloadable tabular format, together with ligand properties (**c**)

3.3 Depicting and Analyzing Protein-Small Molecule Interactions with Arpeggio

1. Once a structure of the target protein with the candidate molecule is available, either through experimental determination or docking or other alternative approach (for instance, those combining blind docking with molecular dynamics like the Wrap 'n' Shake method [46]), Arpeggio enables the visualization of intermolecular interactions occurring between the lead and its target. During lead optimization, Arpeggio can therefore be used to understand the mechanism of binding and guide medicinal chemistry efforts.
2. Arpeggio is freely available as a user-friendly web interface and is compatible with multiple operating systems and browsers. Open up the prediction server, http://biosig.unimelb.edu.au/arpeggioweb/, on a web browser of your preference.
3. Provide the complexed protein structure of interest by either uploading it as a PDB file or providing the PDB ID of the experimentally solved structure in complex with the ligand of interest (Fig. 4a-1).
4. Select the ligand or ligands of interest under the "Heteroatom" selection heading to calculate all molecular interactions being made by that ligand (Fig. 4b-1; *see* **Note 4**).
5. The results page will show an interactive image of all the molecular interactions made by the ligand(s) selected (Fig. 5a) and a table with a count of the total number of specific molecular interactions being made, including hydrophobic interactions, hydrogen bonds, pi-interactions, and ionic interactions (Fig. 4c).
6. A Pymol session file (PSE file) containing the submitted PDB file and all of the calculated interactions can be downloaded and opened in Pymol to enable visualization of the interaction network in 3D and to facilitate high-quality image generation for manuscripts (Fig. 5b).

3.4 Predicting the Effects of Mutations on Small Molecule Affinity with mCSM-lig

1. During lead optimization, it is important to consider how genetic diversity might affect the binding of candidate molecules and, in particular, if resistance is likely to arise. mCSM-lig uses graph-based signatures to calculate the change upon mutation in small molecule binding affinity. In order to run a prediction, open up the mCSM-lig server at http://biosig.unimelb.edu.au/mcsm_lig/ on a web browser of your preference (the web server is compatible with the most common operating systems and browsers).
2. Users are required to provide the protein structure in complex with the ligand of interest by either uploading a PDB file or supplying a valid four-letter code PDB accession code of a deposited experimental structure (Fig. 6a-1). Users also need to provide the mutation information, the mutation chain, the

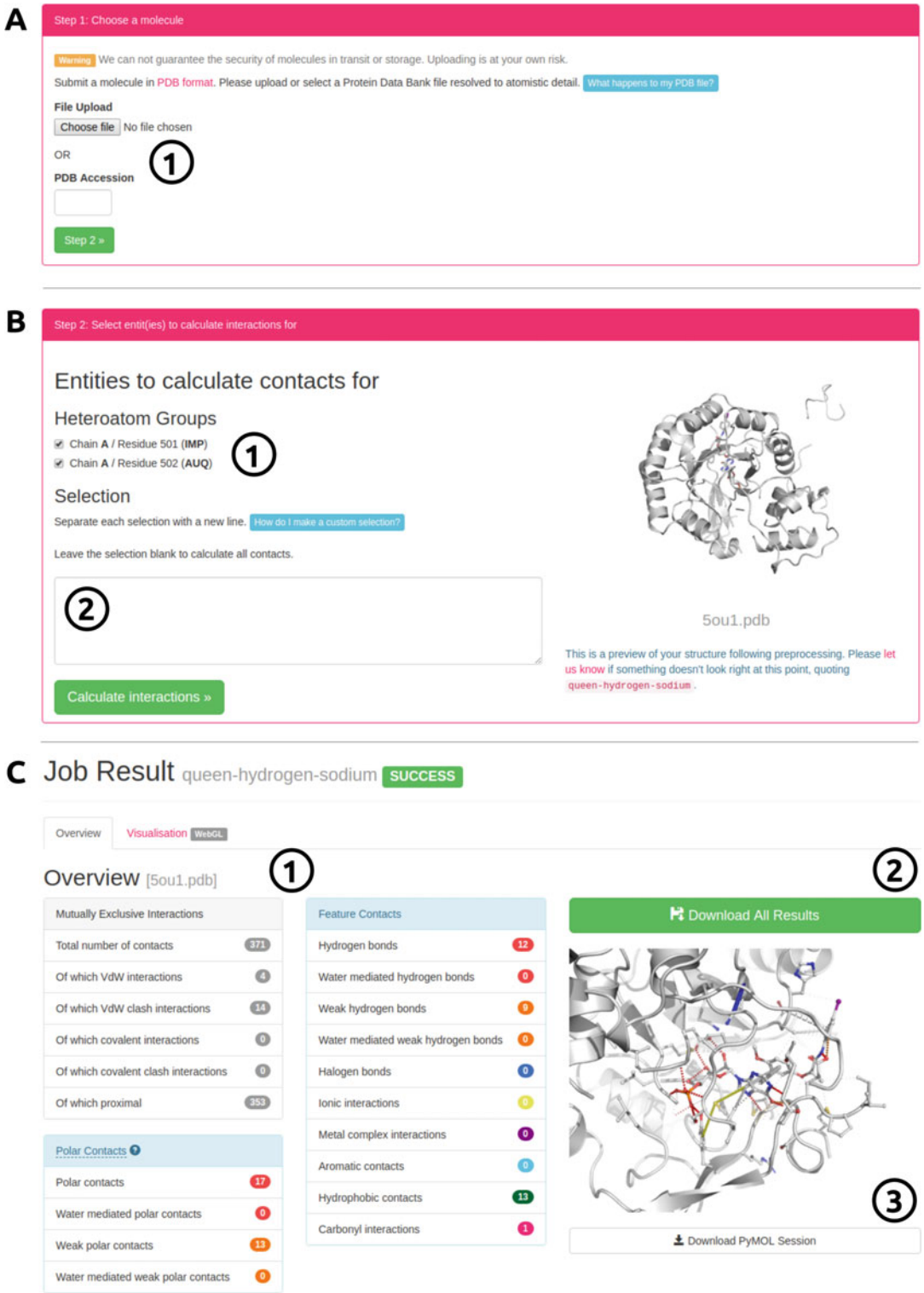


Fig. 4 Arpeggio submission and results web interface. (**a**) The submission page allows users to either provide their own PDB file or an accession code of a deposited experimental structure of the protein of interest. By selecting the molecule of interest (**b**), all molecular interactions will be calculated and displayed (**c**)

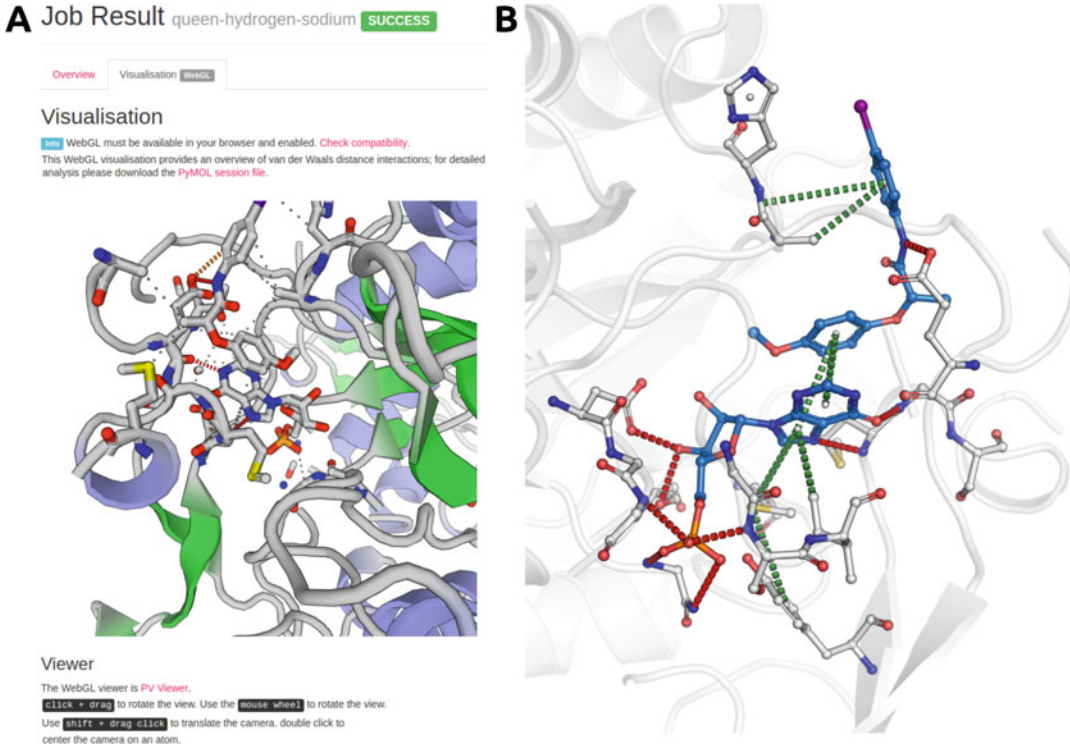


Fig. 5 Molecular interaction visualization using Arpeggio. The molecular interactions calculated by Arpeggio can be visualized either online (**a**) or by downloading the PSE file for visualization in Pymol (**b**)

three-letter code of the ligand of interest in the PDB file, and the approximate binding affinity (in nM) (Fig. 6a-2). If the binding affinity is not available, this can be approximated using CSM-lig. The mCSM-lig values do not vary significantly across most biologically relevant binding affinities.

3. After processing, the results page is shown (Fig. 6b-1), which includes information about the mutation and the predicted effects on the ligand binding affinity. An interactive molecular visualization is shown, allowing users to inspect the wild-type residue environment (Fig. 6b-2).
4. Predicted effects are outputted as the log fold change in binding affinity, in which negative values denote destabilizing mutations and positive values, stabilizing ones. The example shown in Fig. 6 and the web server depicts the prediction for a mutation on the HIV-1 protease bound to an inhibitor. Mutation from Aspartic Acid to Asparagine on residue position 30 is predicted to considerably reduce protein-ligand affinity. While users should interpret the values in the context of the protein system being studied, for competitive binding inhibitors, it is often important to consider the relative effect of a mutation on not only inhibitor binding but also the competitive ligand. This

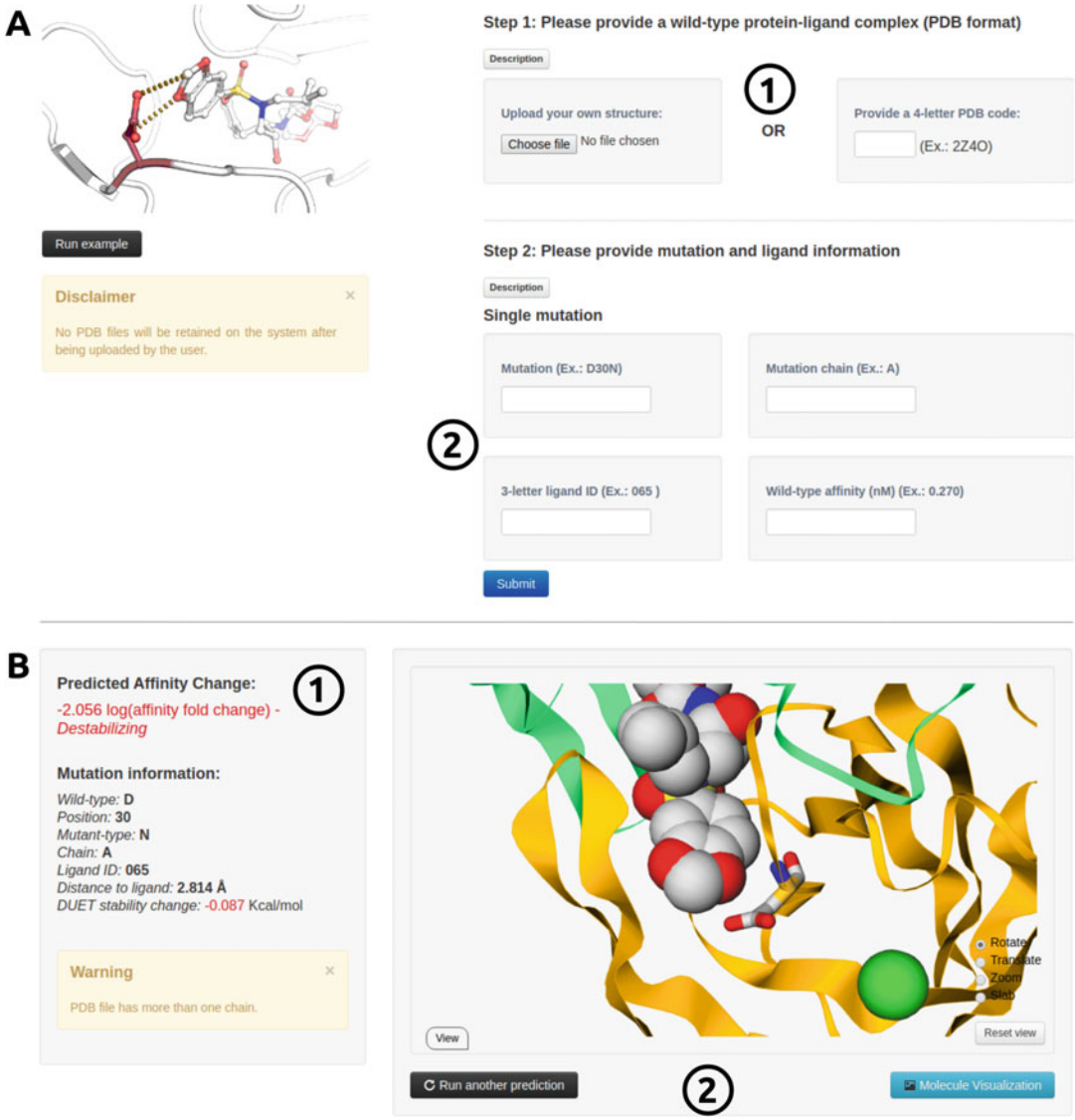


Fig. 6 mCSM-lig submission and results web interface. To predict the effects of a mutation on protein-ligand affinity, users need to provide a protein-ligand structure of interest (**a-1**) as well as mutation and ligand information (**a-2**). Once the calculations have finished, the results page will show the predicted change in ligand binding affinity (**b-1**) as well as an interactive visualization of the mutated residue within its molecular environment (**b-2**)

can be done by submitting a structure of the protein containing the ligand. Resistance mutations are more likely to affect, or have a larger effect, on inhibitor binding affinity than the natural ligand. This has been used to successfully preemptively guide detection of likely resistance variants [29–31, 47–53].

4 Notes

1. The chain ID for the provided PDB file is a mandatory field for CSM-Lig and mCSM-Lig, and blank characters are not allowed. It is possible that homology modeling tools might not automatically add a chain ID. If this is the case, the user will need to modify the PDB file prior to submission to the servers. There are several tools available to perform this task.
2. Another source of error comes from multiple occupancies, common in high-resolution experimental X-ray crystal structures. Multiple occupancies should first be filtered out, with the highest occupancy conformation normally selected.
3. NMR experimental structures often contain multiple models. It is an important practice to filter NMR structures, selecting a single model. The predictive tool will show a warning message in case multiple models are identified.
4. Arpeggio will sometimes fail if the PDB file contains an element with upper and lower case letters (e.g., Fe as opposed to FE). These can be altered using a text editor.

Acknowledgments

This work was supported by the Australian Government Research Training Program Scholarships [to S.P., M.K., Y.M., C.H.M.R.]; the Jack Brockhoff Foundation [JBF 4186, 2016 to D.B.A.]; a Newton Fund RCUK-CONFAP Grant awarded by The Medical Research Council (MRC) and Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG) [MR/M026302/1 to D.B.A. and D.E.V.P.]; the National Health and Medical Research Council of Australia [APP1072476 to D.B.A.]; the Instituto René Rachou (IRR/FIOCRUZ Minas), Brazil, and Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) [to D.E.V.P., P.M.R.]; and the Department of Biochemistry and Molecular Biology, University of Melbourne [to D.B.A.].

References

1. Pires DE, Ascher DB, Blundell TL (2014) mCSM: predicting the effects of mutations in proteins using graph-based signatures. Bioinformatics 30(3):335–342. https://doi.org/10.1093/bioinformatics/btt691
2. Pires DE, Ascher DB, Blundell TL (2014) DUET: a server for predicting effects of mutations on protein stability using an integrated computational approach. Nucleic Acids Res 42 (Web Server issue):W314–W319. https://doi.org/10.1093/nar/gku411
3. Pires DE, Ascher DB (2016) CSM-lig: a web server for assessing and comparing protein-small molecule affinities. Nucleic Acids Res 44 (W1):W557–W561. https://doi.org/10.1093/nar/gkw390
4. Pires DE, Ascher DB (2016) mCSM-AB: a web server for predicting antibody-antigen affinity changes upon mutation with graph-based

signatures. Nucleic Acids Res 44(W1): W469–W473. https://doi.org/10.1093/nar/gkw458

5. Pires DE, Blundell TL, Ascher DB (2016) mCSM-lig: quantifying the effects of mutations on protein-small molecule affinity in genetic disease and emergence of drug resistance. Sci Rep 6:29575. https://doi.org/10.1038/srep29575
6. Pires DE, Chen J, Blundell TL, Ascher DB (2016) In silico functional dissection of saturation mutagenesis: interpreting the relationship between phenotypes and changes in protein stability, interactions and activity. Sci Rep 6:19848. https://doi.org/10.1038/srep19848
7. Jubb HC, Higueruelo AP, Ochoa-Montano B, Pitt WR, Ascher DB, Blundell TL (2017) Arpeggio: a web server for calculating and visualising interatomic interactions in protein structures. J Mol Biol 429(3):365–371. https://doi.org/10.1016/j.jmb.2016.12.004
8. Pandurangan AP, Ochoa-Montano B, Ascher DB, Blundell TL (2017) SDM: a server for predicting effects of mutations on protein stability. Nucleic Acids Res 45(W1): W229–W235. https://doi.org/10.1093/nar/gkx439
9. Rodrigues CH, Ascher DB, Pires DE (2018) Kinact: a computational approach for predicting activating missense mutations in protein kinases. Nucleic Acids Res 46(W1): W127–W132. https://doi.org/10.1093/nar/gky375
10. Rodrigues CH, Pires DE, Ascher DB (2018) DynaMut: predicting the impact of mutations on protein conformation, flexibility and stability. Nucleic Acids Res 46(W1):W350–W355. https://doi.org/10.1093/nar/gky300
11. Pires DE, Blundell TL, Ascher DB (2015) Platinum: a database of experimentally measured effects of mutations on structurally defined protein-ligand complexes. Nucleic Acids Res 43(Database issue):D387–D391. https://doi.org/10.1093/nar/gku966
12. Pires DEV, Ascher DB (2017) mCSM-NA: predicting the effects of mutations on protein-nucleic acids interactions. Nucleic Acids Res 45 (W1):W241–W246. https://doi.org/10.1093/nar/gkx236
13. Jafri M, Wake NC, Ascher DB, Pires DE, Gentle D, Morris MR, Rattenberry E, Simpson MA, Trembath RC, Weber A, Woodward ER, Donaldson A, Blundell TL, Latif F, Maher ER (2015) Germline mutations in the CDKN2B tumor suppressor gene predispose to renal cell carcinoma. Cancer Discov 5(7):723–729. https://doi.org/10.1158/2159-8290.CD-14-1096
14. Jubb H, Blundell TL, Ascher DB (2015) Flexibility and small pockets at protein-protein interfaces: new insights into druggability. Prog Biophys Mol Biol 119(1):2–9. https://doi.org/10.1016/j.pbiomolbio.2015.01.009
15. Usher JL, Ascher DB, Pires DE, Milan AM, Blundell TL, Ranganath LR (2015) Analysis of HGD gene mutations in patients with alkaptonuria from the United Kingdom: identification of novel mutations. JIMD Rep 24:3–11. https://doi.org/10.1007/8904_2014_380
16. Coelho MB, Ascher DB, Gooding C, Lang E, Maude H, Turner D, Llorian M, Pires DE, Attig J, Smith CW (2016) Functional interactions between polypyrimidine tract binding protein and PRI peptide ligand containing proteins. Biochem Soc Trans 44(4):1058–1065. https://doi.org/10.1042/BST20160080
17. Kano FS, Souza-Silva FA, Torres LM, Lima BA, Sousa TN, Alves JR, Rocha RS, Fontes CJ, Sanchez BA, Adams JH, Brito CF, Pires DE, Ascher DB, Sell AM, Carvalho LH (2016) The presence, persistence and functional properties of Plasmodium vivax Duffy binding protein II antibodies are influenced by HLA class II allelic variants. PLoS Negl Trop Dis 10(12):e0005177. https://doi.org/10.1371/journal.pntd.0005177
18. Nemethova M, Radvanszky J, Kadasi L, Ascher DB, Pires DE, Blundell TL, Porfirio B, Mannoni A, Santucci A, Milucci L, Sestini S, Biolcati G, Sorge F, Aurizi C, Aquaron R, Alsbou M, Lourenco CM, Ramadevi K, Ranganath LR, Gallagher JA, van Kan C, Hall AK, Olsson B, Sireau N, Ayoob H, Timmis OG, Sang KH, Genovese F, Imrich R, Rovensky J, Srinivasaraghavan R, Bharadwaj SK, Spiegel R, Zatkova A (2016) Twelve novel HGD gene variants identified in 99 alkaptonuria patients: focus on 'black bone disease' in Italy. Eur J Hum Genet 24(1):66–72. https://doi.org/10.1038/ejhg.2015.60
19. Silvino AC, Costa GL, Araujo FC, Ascher DB, Pires DE, Fontes CJ, Carvalho LH, Brito CF, Sousa TN (2016) Variation in human cytochrome P-450 drug-metabolism genes: a gateway to the understanding of Plasmodium vivax relapses. PLoS One 11(7):e0160172. https://doi.org/10.1371/journal.pone.0160172
20. White RR, Ponsford AH, Weekes MP, Rodrigues RB, Ascher DB, Mol M, Selkirk ME, Gygi SP, Sanderson CM, Artavanis-Tsakonas K (2016) Ubiquitin-dependent modification of skeletal muscle by the parasitic nematode, Trichinella spiralis. PLoS Pathog 12(11):

e1005977. https://doi.org/10.1371/journal.ppat.1005977

21. Casey RT, Ascher DB, Rattenberry E, Izatt L, Andrews KA, Simpson HL, Challis B, Park SM, Bulusu VR, Lalloo F, Pires DEV, West H, Clark GR, Smith PS, Whitworth J, Papathomas TG, Taniere P, Savisaar R, Hurst LD, Woodward ER, Maher ER (2017) SDHA related tumorigenesis: a new case series and literature review for variant interpretation and pathogenicity. Mol Genet Genomic Med 5(3):237–250. https://doi.org/10.1002/mgg3.279
22. Jubb HC, Pandurangan AP, Turner MA, Ochoa-Montano B, Blundell TL, Ascher DB (2017) Mutations at protein-protein interfaces: small changes over big surfaces have large impacts on human health. Prog Biophys Mol Biol 128:3–13. https://doi.org/10.1016/j.pbiomolbio.2016.10.002
23. Ramdzan YM, Trubetskov MM, Ormsby AR, Newcombe EA, Sui X, Tobin MJ, Bongiovanni MN, Gras SL, Dewson G, Miller JML, Finkbeiner S, Moily NS, Niclis J, Parish CL, Purcell AW, Baker MJ, Wilce JA, Waris S, Stojanovski D, Bocking T, Ang CS, Ascher DB, Reid GE, Hatters DM (2017) Huntingtin inclusions trigger cellular quiescence, deactivate apoptosis, and lead to delayed necrosis. Cell Rep 19(5):919–927. https://doi.org/10.1016/j.celrep.2017.04.029
24. Soardi FC, Machado-Silva A, Linhares ND, Zheng G, Qu Q, Pena HB, Martins TMM, Vieira HGS, Pereira NB, Melo-Minardi RC, Gomes CC, Gomez RS, Gomes DA, Pires DEV, Ascher DB, Yu H, Pena SDJ (2017) Familial STAG2 germline mutation defines a new human cohesinopathy. NPJ Genom Med 2(1):7. https://doi.org/10.1038/s41525-017-0009-4
25. Traynelis J, Silk M, Wang Q, Berkovic SF, Liu L, Ascher DB, Balding DJ, Petrovski S (2017) Optimizing genomic medicine in epilepsy through a gene-customized approach to missense variant interpretation. Genome Res 27(10):1715–1729. https://doi.org/10.1101/gr.226589.117
26. Trezza A, Bernini A, Langella A, Ascher DB, Pires DEV, Sodi A, Passerini I, Pelo E, Rizzo S, Niccolai N, Spiga O (2017) A computational approach from gene to structure analysis of the human ABCA4 transporter involved in genetic retinal diseases. Invest Ophthalmol Vis Sci 58(12):5320–5328. https://doi.org/10.1167/iovs.17-22158
27. Andrews KA, Ascher DB, Pires DEV, Barnes DR, Vialard L, Casey RT, Bradshaw N, Adlard J, Aylwin S, Brennan P, Brewer C, Cole T, Cook JA, Davidson R, Donaldson A, Fryer A, Greenhalgh L, Hodgson SV, Irving R, Lalloo F, McConachie M, McConnell VPM, Morrison PJ, Murday V, Park SM, Simpson HL, Snape K, Stewart S, Tomkins SE, Wallis Y, Izatt L, Goudie D, Lindsay RS, Perry CG, Woodward ER, Antoniou AC, Maher ER (2018) Tumour risks and genotype-phenotype correlations associated with germline variants in succinate dehydrogenase subunit genes SDHB, SDHC and SDHD. J Med Genet 55(6):384–394. https://doi.org/10.1136/jmedgenet-2017-105127
28. Hnizda A, Fabry M, Moriyama T, Pachl P, Kugler M, Brinsa V, Ascher DB, Carroll WL, Novak P, Zaliova M, Trka J, Rezacova P, Yang JJ, Veverka V (2018) Relapsed acute lymphoblastic leukemia-specific mutations in NT5C2 cluster into hotspots driving intersubunit stimulation. Leukemia 32(6):1393–1403. https://doi.org/10.1038/s41375-018-0073-5
29. Albanaz ATS, Rodrigues CHM, Pires DEV, Ascher DB (2017) Combating mutations in genetic disease and drug resistance: understanding molecular mechanisms to guide drug design. Expert Opin Drug Discov 12(6):553–563. https://doi.org/10.1080/17460441.2017.1322579
30. Park Y, Pacitto A, Bayliss T, Cleghorn LA, Wang Z, Hartman T, Arora K, Ioerger TR, Sacchettini J, Rizzi M, Donini S, Blundell TL, Ascher DB, Rhee K, Breda A, Zhou N, Dartois V, Jonnala SR, Via LE, Mizrahi V, Epemolu O, Stojanovski L, Simeons F, Osuna-Cabello M, Ellis L, MacKenzie CJ, Smith AR, Davis SH, Murugesan D, Buchanan KI, Turner PA, Huggett M, Zuccotto F, Rebollo-Lopez MJ, Lafuente-Monasterio MJ, Sanz O, Diaz GS, Lelievre J, Ballell L, Selenski C, Axtman M, Ghidelli-Disse S, Pflaumer H, Bosche M, Drewes G, Freiberg GM, Kurnick MD, Srikumaran M, Kempf DJ, Green SR, Ray PC, Read K, Wyatt P, Barry CE 3rd, Boshoff HI (2017) Essential but not vulnerable: indazole sulfonamides targeting inosine monophosphate dehydrogenase as potential leads against Mycobacterium tuberculosis. ACS Infect Dis 3(1):18–33. https://doi.org/10.1021/acsinfecdis.6b00103
31. Singh V, Donini S, Pacitto A, Sala C, Hartkoorn RC, Dhar N, Keri G, Ascher DB, Mondesert G, Vocat A, Lupien A, Sommer R, Vermet H, Lagrange S, Buechler J, Warner DF, McKinney JD, Pato J, Cole ST, Blundell TL, Rizzi M, Mizrahi V (2017) The inosine monophosphate dehydrogenase, GuaB2, is a vulnerable new bactericidal drug target for tuberculosis. ACS Infect Dis 3(1):5–17. https://doi.org/10.1021/acsinfecdis.6b00102

32. Trapero A, Pacitto A, Singh V, Sabbah M, Coyne AG, Mizrahi V, Blundell TL, Ascher DB, Abell C (2018) Fragment-based approach to targeting inosine-5′-monophosphate dehydrogenase (IMPDH) from Mycobacterium tuberculosis. J Med Chem 61(7):2806–2822. https://doi.org/10.1021/acs.jmedchem.7b01622
33. Pires DE, Blundell TL, Ascher DB (2015) pkCSM: predicting small-molecule pharmacokinetic and toxicity properties using graph-based signatures. J Med Chem 58 (9):4066–4072. https://doi.org/10.1021/acs.jmedchem.5b00104
34. Pires DEV, Kaminskas LM, Ascher DB (2018) Prediction and optimization of pharmacokinetic and toxicity properties of the ligand. Methods Mol Biol 1762:271–284. https://doi.org/10.1007/978-1-4939-7756-7_14
35. Sigurdardottir AG, Winter A, Sobkowicz A, Fragai M, Chirgadze D, Ascher DB, Blundell TL, Gherardi E (2015) Exploring the chemical space of the lysine-binding pocket of the first kringle domain of hepatocyte growth factor/scatter factor (HGF/SF) yields a new class of inhibitors of HGF/SF-MET binding. Chem Sci 6(11):6147–6157. https://doi.org/10.1039/c5sc02155c
36. Ascher DB, Jubb HC, Pires DE, Ochi T, Higueruelo A, Blundell TL (2015) Protein-protein interactions: structures and druggability. In: Scapin G, Patel D, Arnold E (eds) Multifaceted roles of crystallography in modern drug discovery. NATO science for peace and security series a: chemistry and biology. Springer, Netherlands, pp 141–163. https://doi.org/10.1007/978-94-017-9719-1_12
37. Pandurangan AP, Ascher DB, Thomas SE, Blundell TL (2017) Genomes, structural biology and drug discovery: combating the impacts of mutations in genetic disease and antibiotic resistance. Biochem Soc Trans 45(2):303–311. https://doi.org/10.1042/BST20160422
38. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) The Protein Data Bank. Nucleic Acids Res 28(1):235–242
39. O'Boyle NM, Banck M, James CA, Morley C, Vandermeersch T, Hutchison GR (2011) Open babel: an open chemical toolbox. J Cheminform 3:33. https://doi.org/10.1186/1758-2946-3-33
40. Hanwell MD, Curtis DE, Lonie DC, Vandermeersch T, Zurek E, Hutchison GR (2012) Avogadro: an advanced semantic chemical editor, visualization, and analysis platform. J Cheminform 4(1):17. https://doi.org/10.1186/1758-2946-4-17
41. Ascher DB, Crespi GA, Ng HL, Morton CJ, Parker MW (2008) Novel therapeutic approaches to treat Alzheimer's disease and memory disorders. J Proteomics Bioinform 1:464–476
42. Chai SY, Yeatman HR, Parker MW, Ascher DB, Thompson PE, Mulvey HT, Albiston AL (2008) Development of cognitive enhancers based on inhibition of insulin-regulated aminopeptidase. BMC Neurosci 9(Suppl 2):S14. https://doi.org/10.1186/1471-2202-9-S2-S14
43. Kawabata T (2010) Detection of multiscale pockets on protein surfaces using mathematical morphology. Proteins 78(5):1195–1211. https://doi.org/10.1002/prot.22639
44. Lipinski CA, Lombardo F, Dominy BW, Feeney PJ (2001) Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. Adv Drug Deliv Rev 46 (1–3):3–26
45. Schrodinger, LLC (2015) The PyMOL molecular graphics system, version 1.8
46. Balint M, Jeszenoi N, Horvath I, van der Spoel D, Hetenyi C (2017) Systematic exploration of multiple drug binding sites. J Cheminform 9(1):65. https://doi.org/10.1186/s13321-017-0255-6
47. Ascher DB, Wielens J, Nero TL, Doughty L, Morton CJ, Parker MW (2014) Potent hepatitis C inhibitors bind directly to NS5A and reduce its affinity for RNA. Sci Rep 4:4765. https://doi.org/10.1038/srep04765
48. Phelan J, Coll F, McNerney R, Ascher DB, Pires DE, Furnham N, Coeck N, Hill-Cawthorne GA, Nair MB, Mallard K, Ramsay A, Campino S, Hibberd ML, Pain A, Rigouts L, Clark TG (2016) Mycobacterium tuberculosis whole genome sequencing and protein structure modelling provides insights into anti-tuberculosis drug resistance. BMC Med 14(1):31. https://doi.org/10.1186/s12916-016-0575-9
49. Hawkey J, Ascher DB, Judd LM, Wick RR, Kostoulias X, Cleland H, Spelman DW, Padiglione A, Peleg AY, Holt KE (2018) Evolution of carbapenem resistance in Acinetobacter baumannii during a prolonged infection. Microb Genom 4. https://doi.org/10.1099/mgen.0.000165
50. Holt KE, McAdam P, Thai PVK, Thuong NTT, Ha DTM, Lan NN, Lan NH, Nhu NTQ, Hai HT, Ha VTN, Thwaites G, Edwards DJ, Nath AP, Pham K, Ascher DB, Farrar J, Khor CC, Teo YY, Inouye M, Caws M, Dunstan SJ (2018) Frequent transmission of the Mycobacterium tuberculosis Beijing lineage

and positive selection for the EsxW Beijing variant in Vietnam. Nat Genet 50 (6):849–856. https://doi.org/10.1038/s41588-018-0117-9

51. Karmakar M, Globan M, Fyfe JAM, Stinear TP, Johnson PDR, Holmes NE, Denholm JT, Ascher DB (2018) Analysis of a novel pncA mutation for susceptibility to pyrazinamide therapy. Am J Respir Crit Care Med 198 (4):541–544. https://doi.org/10.1164/rccm.201712-2572LE
52. Portelli S, Phelan JE, Ascher DB, Clark TG, Furnham N (2018) Understanding molecular consequences of putative drug resistant mutations in Mycobacterium tuberculosis. Sci Rep 8 (1):15356. https://doi.org/10.1038/s41598-018-33370-6
53. Vedithi SC, Malhotra S, Das M, Daniel S, Kishore N, George A, Arumugam S, Rajan L, Ebenezer M, Ascher DB, Arnold E, Blundell TL (2018) Structural implications of mutations conferring rifampin resistance in Mycobacterium leprae. Sci Rep 8(1):5016. https://doi.org/10.1038/s41598-018-23423-1

Chapter 1

Identifying Genotype–Phenotype Correlations via Integrative Mutation Analysis

Edward Airey, Stephanie Portelli, Joicymara S. Xavier, Yoo Chan Myung, Michael Silk, Malancha Karmakar, João P. L. Velloso, Carlos H. M. Rodrigues, Hardik H. Parate, Anjali Garg, Raghad Al-Jarf, Lucy Barr, Juliana A. Geraldo, Pâmela M. Rezende, Douglas E. V. Pires, and David B. Ascher

Abstract

Mutations in protein-coding regions can lead to large biological changes and are associated with genetic conditions, including cancers and Mendelian diseases, as well as drug resistance. Although whole genome and exome sequencing help to elucidate potential genotype–phenotype correlations, there is a large gap between the identification of new variants and deciphering their molecular consequences. A comprehensive understanding of these mechanistic consequences is crucial to better understand and treat diseases in a more personalized and effective way. This is particularly relevant considering estimates that over 80% of mutations associated with a disease are incorrectly assumed to be causative. A thorough analysis of potential effects of mutations is required to correctly identify the molecular mechanisms of disease and enable the distinction between disease-causing and non–disease-causing variation within a gene. Here we present an overview of our integrative mutation analysis platform, which focuses on refining the current genotype–phenotype correlation methods by using the wealth of protein structural information.

Key words Genotype–phenotype correlations, Graph-based signatures, mCSM, Mutation, Protein structure, Protein interactions

1 Introduction

Proteins are versatile molecules, responsible for orchestrating a wide range of biological processes. They comprise a single polypeptide chain of amino acids, which folds in 3D space into dynamic structures. How a protein folds is important for determining its functions, including activities and interactions with other molecules. These structures are highly coordinated and conserved across evolution, and small perturbations in the amino acid sequence can disrupt these shapes, functions, and interactions [1, 2]. While

Hugh Cartwright (ed.), *Artificial Neural Networks*, Methods in Molecular Biology, vol. 2190,
https://doi.org/10.1007/978-1-0716-0826-5_1,

missense mutations, causing a change to a single amino acid, are generally less structurally disruptive than nonsense mutations, their effects are highly variable and can be wide-ranging, making their molecular consequences harder to determine. Despite their subtle effects, missense substitutions are related with many different genetic conditions, including cancer, Mendelian diseases, and the emergence of drug resistance.

The introduction of a missense mutation can have many molecular effects, including altering how the protein folds, its dynamics, posttranslational modifications, half-life, localization, activity, and molecular interactions [3]. When analyzing a new mutation, an integrative approach is therefore important to consider the effects it might have on all of these aspects. This enables the identification of specific functional, and structural changes imparted by the mutations, which is essential for a molecular understanding. It can also explain why mutations in the same protein might lead to different diseases, why mutations might cluster in 3D space and how those genetic changes present phenotypically.

Although many assume that an unfavorable phenotype (e.g., pathogenic, drug-resistant) is the result of large, overall destabilizing mutations, mutations with milder effects are often more prevalent in a population, as they are generally under less selective pressure [4, 5]. For example, by assessing mutations in three different tuberculosis proteins that lead to resistance, we have shown that the most frequent resistant mutations were more likely to be associated with overall mild functional effects, and associated reduced fitness cost, allowing for increased prevalence within the bacterial population [4].

Experimentally elucidating the biophysical effects of mutations is an expensive and time-consuming task, usually limited to a few variants in proteins with amenable assays. Over the years, the accumulation of information of experimentally characterized mutations has enabled the development and improvement of computational mutational analysis tools [6]. These computational platforms have shown to be invaluable assets to decipher genotype–phenotype correlations in cancer [7–19], Mendelian diseases [20–26], and detection of antimicrobial resistance [4, 15, 27–35], guiding clinical decisions and driving further research. Here, we introduce a general computational pipeline that uses in silico biophysical predictions and machine learning approaches to harness the wealth of available biological and protein structural information and give insights into genotype–phenotype correlation for clinical use [10].

The mutation cutoff scanning matrix (mCSM) platform is the only comprehensive collection of in silico tools for quantitatively predicting the effects of missense mutations on protein folding, structure, dynamics, and interactions. It includes tools which calculate all possible molecular interactions (Arpeggio [36]), account for changes in protein stability (mCSM-Stability [37], SDM [38],

DUET [39], mCSM-membrane [40], dynamics (DynaMut [41]), protein interactions with other proteins (mCSM-PPI [37], mCSM-PPI2 [42], mCSM-AB [43], mCSM-AB2 [44], mmCSM-AB [45], nucleic acids (mCSM-DNA [37], mCSM-NA [46]), and small molecule ligands (mCSM-lig [47], CSM-lig [48]).

These tools were built using the concept of graph-based signatures [49, 50], which represent the geometry and physicochemical properties of the wild-type protein structure environment as a network or graph, composed of a series of nodes, describing the local mutation environment, and edges, describing the distances between interacting "layers" of surrounding residues. Information on the mutation is captured using the pharmacophore change between the wild-type and the mutant residue, including whether charges or hydrogen donors/acceptors have been gained or lost [37].

This platform allows for accurate biophysical predictions, which, when complemented with other protein analytical tools, can provide a detailed landscape on the specific mutational effects on a protein. We have implemented these within an analytical and supervised machine learning predictive pipeline (Fig. 1), to enable easy and fast characterization of novel mutations and their likely

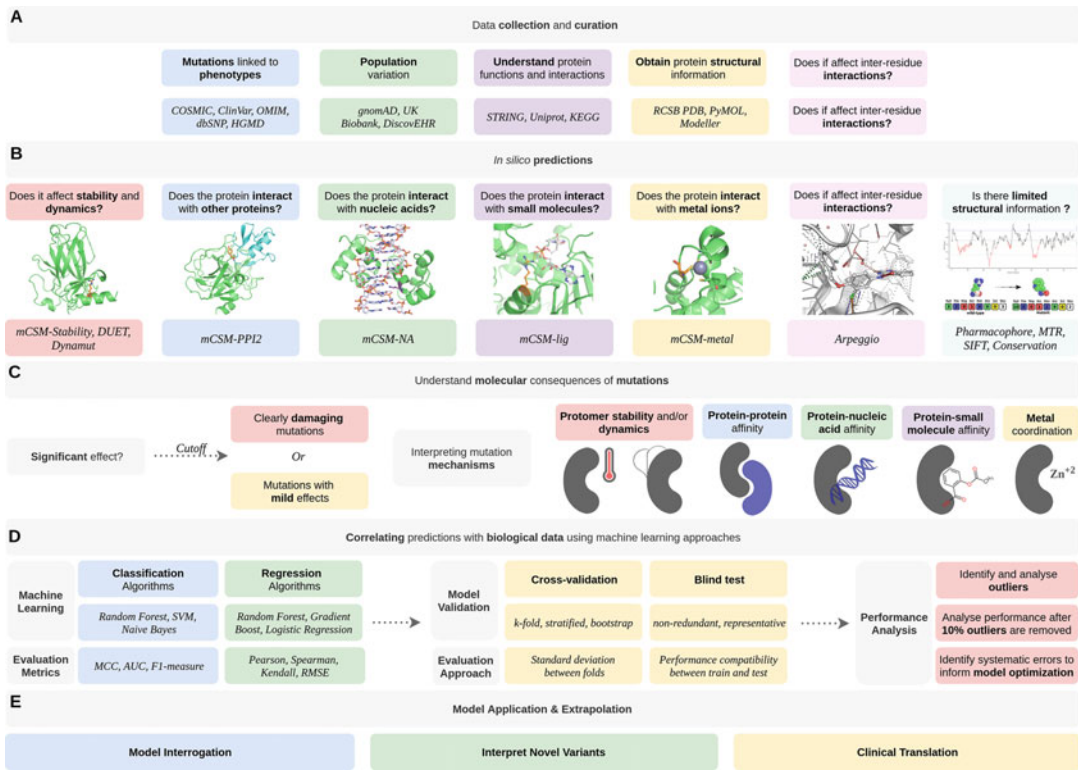


Fig. 1 An overview of the mechanistic characterization of mutations and their biological consequences, to guide the development of tools to predict phenotypic outcomes

clinical phenotypes. This approach has been shown to have big implications in diagnostic and personalized medicine in the post-genomic era.

2 Materials

2.1 Data Curation

2.1.1 Mutation Curation

The foremost requirement for training a machine learning model is appropriate high-quality experimental/clinical data, with suitable representation of the classes under comparison. For human disease, a wealth of freely accessible collections of curated data exist. Previously reported mutations through publications and functional studies are available from dbSNP [51], the largest freely available repository of genetic variation. Variants with evidence of pathogenicity can be viewed from the Human Gene Mutation Database (HGMD) [52] and ClinVar [53], and from disease-specific datasets such as the Catalogue of Somatic Mutations in Cancer (COSMIC). Standing variation is available from genomic sequencing efforts of healthy populations, including over 140,000 healthy humans in gnomAD [54] and 50,000 whole exomes currently available in UK Biobank [55].

When combining data from multiple sources, it is important that all datapoints are comparable. If using genetic coordinates, they should be found on the same assembly of the genome (e.g., GRCh38 vs GRCh37). The mutations themselves (whether reported as genetic or amino acid changes) must be reported on the same transcript, as most genes have multiple reported coding sequences.

2.1.2 Protein Structure Curation

The sequence and functional information for a specific protein can be obtained from Uniprot (https://www.uniprot.org/) [56]. To run the mCSM tools we need crystallographic structures, which can be downloaded from the Protein Data Bank (PDB;http://www.rcsb.org/) [57] or generated via homology modeling or molecular docking (to run mCSM-PPI, mCSM-Lig, or mCSM-NA). Once we have the variant information collected from the resources in Subheading 2.1.1, we map these variants on to the identified protein structures to help visualize the spread and identify potential hotspots, which is easily done using visualization software such as PyMol, as it enables selection of residues being mutated in a 3D manner.

2.2 An Overview of Computational Tools to Analyze Missense Mutations

Over the past two decades there has been an unprecedented growth in both computational power and the amount of biological data available. This has facilitated the development of numerous sequence (Table 1) and structural (Table 2) based computational tools to guide mutation characterization.

Table 1
Available sequence-based predictive tools for mutation analysis

Protein stability and dynamics	
Method	**Corr.[a]**
I-Mutant 2.0	0.62
Auto-Mute	0.64[a]
MUpro	0.75
DynaMine	0.63
DDGun	0.49
INPS-MD/3D	0.58
iStable	0.56[b]
iPTREEE - STAB	0.70
ProMaya	0.79

[a]Pearson's correlation
[b]MCC

Table 2
Available structure-based predictive tools for mutation analysis

Protein stability and dynamics		Protein–protein affinity		Protein–nucleic acid affinity		Protein–small molecule affinity	
Method	**Corr.[a]**	**Method**	**Corr.[b]**	**Method**	**Corr.[c]**	**Method**	**Corr.[d]**
mCSM-Stability	0.69	mCSM-PPI	0.16	mCSM-NA	0.70	mCSM-lig	0.63
DUET	0.68	mCSM-PPI2	0.42				
DynaMut	0.70	BeAtMuSiC	0.28				
SDM2	0.61	MutaBind	0.41				
STRUM	0.79	FoldX	0.12				
PopMuSiC 2.1	0.63	MMPBSA	0.19				
CUPSAT	0.78						
Eris	0.75						
INPS-MD/3D	0.72						

[a]Pearson's correlation when evaluated on blind-test sets derived from the ProTherm database
[b]Kendall rank correlation coefficient on 1007 single-point mutations from CAPRI (T55)
[c]Pearson's correlation on 331 single-point mutations from 38 protein–nucleic acid complexes
[d]Pearson's correlation on 763 single-point mutations from 200 protein–ligand complexes

The mCSM platform is the only available approach to consider all possible molecular effects and has therefore formed the central component of our mutational analysis pipeline. All mCSM

Platform tools are available freely as websites compatible with most web-browsers, but Google Chrome is recommended. A summary of these methods and links to access them is described in Table 3.

Table 3
Computational tools available in the mCSM platform

mCSM tool	Type	Function
Arpeggio[a]	Protein interaction	Calculates 13 different types of interactions between atoms including hydrogen bonds, halogen bonds, carbonyl interactions, and others.
MTR-Viewer[b]	Missense tolerance	A measure of a gene's regional tolerance to missense variation.
mCSM-Stability[c]	Stability	Predict the effects of a mutation on the overall protein stability
SDM2[d]	Stability	Predicts the change in protein stability due to a single mutation using conformationally constrained environment-dependent amino acid substitution tables.
DUET[e]	Stability	Uses mCSM-Stability and SDM2 in order to create a consensus prediction the effects of a mutation on protein stability
DynaMut[f]	Flexibility	Looks to predict the effects of a mutation on protein stability, flexibility, and dynamics
mCSM-PPI[g]	Protein interaction	Predicts the effects of a mutation within a specified protein on its impact with overall protein–protein interactions.
mCSM-PPI2[h]	Protein interaction	Creates a similar prediction to PPI but incorporates the effects of mutations on interresidue noncovalent interaction network using graph kernels, evolutionary information, complex network metrics, and energetic terms.
mCSM-DNA[i]	Protein interaction	Predicts the impact of mutations on the protein interaction with DNA.
mCSM-NA[j]	Protein interaction	Predicts the impact of mutations on the protein interaction with nucleic acids, and uses pharmacophore and information about nucleic acid properties.
mCSM-Lig[k]	Protein interaction	Predicts the effects of single-point mutations on the stability of a protein–ligand complex.

[a]http://biosig.unimelb.edu.au/arpeggioweb/
[b]http://biosig.unimelb.edu.au/mtr-viewer/
[c]http://biosig.unimelb.edu.au/mcsm/stability
[d]http://marid.bioc.cam.ac.uk/sdm2
[e]http://biosig.unimelb.edu.au/duet/
[f]http://biosig.unimelb.edu.au/dynamut/
[g]http://biosig.unimelb.edu.au/mcsm/protein_protein
[h]http://biosig.unimelb.edu.au/mcsm_ppi2/
[i]http://biosig.unimelb.edu.au/mcsm/protein_dna
[j]http://biosig.unimelb.edu.au/mcsm_na/
[k]http://biosig.unimelb.edu.au/mcsm_lig/

3 Methods

3.1 Predicting and Analyzing Structural and Biophysical Effects of Mutations Using the mCSM Platform

The mCSM methods can be categorized by purpose. As shown in Fig. 1, methods are chosen depending on interactions made, and what structural information is available. Below we discuss how each type of predictor can be used and interpreted.

- The user should choose the appropriate tools based on what information is available on their protein of interest (Fig. 1).
- In general, each mCSM tool requires a wild-type protein file, in the PDB format, and the single-point mutation or a list of mutations. Some tools may require additional specific information; Table 4 shows the inputs required for each tool. **Notes 1** and **2** highlight some common issues with the submission inputs.

3.2 mCSM Platform Output

The results of Arpeggio are shown in Fig. 2.

3.2.1 Arpeggio

- After submitting a job, an overview of the type and number of atomic interactions within the protein is shown (Fig. 2a). Arpeggio calculates all types of molecular interactions (Table 5), which are displayed and downloadable along with a visual representation of the atomic contacts overlaid on the protein structure (Fig. 2b).
- The number of each interaction/contact and PyMOL session files can be downloaded for a more detailed analysis.

3.2.2 MTR-Viewer

Gene Viewer

- The MTR gene viewer [5] results page (Fig. 3) shows predicted MTR scores in an interactive line graph with a control panel which allows users to adjust the window size and the ethnicity for MTR estimates. A line graph (Fig. 3a) displays regions that have high variation, low-MTR scored; those in red are most likely to be pathogenic. Any ethnicity-specific MTR scores are shown in blue on the line graph.
- The first lollipop plot (Fig. 3b) shows observed missense (yellow) and synonymous (green) variations based on gnomeAD.
- If the gene of interest is a ClinVar pathogenic gene, their pathogenic (red) and benign (blue) missense variants are displayed under the gnomeAD lollipop plot (Fig. 3c).
- Users can browse results of alternative-transcript (Fig. 3d) of the given query if available.

Variant Query

- The variant query result page (Fig. 4) shows MTR scores for each user-supplied missense variant, providing the estimated regional intolerance. Low MTR scores indicate stronger purifying selection within the population. Users can also press "view" next to a variant to show its position within its gene transcript.

Table 4
Information required to run each mCSM program

		Inputs	
mCSM tool	**Task**	**Step 1**	**Step 2**
Arpeggio	Calculate	Molecule in PDB format or PDB accession code.	Select desired interaction calculation. You can select any (including multiple) part of the PDB file using the syntax: /1/2/3 Where: 1. Chain ID. 2. Residue number. 3. Atom name.
MTR-Viewer	Gene Viewer Variant Queries	Gene, ensembl ID, or Refseq ID Variants as GrCh37 genomic coordinates.	Select window size and overlay sub-population
mCSM-Stability, mCSM-PPI, mCSM-DNA	Prediction	Wild-type protein file in PDB format. For mCSM-PPI and mCSM-DNA, the structure of the complex in PDB format is required.	Single mutation (code and mutation chain), file with a list of mutations and its respective chains or code of residue and the mutation chain.
SDM2	Prediction	Wild-type protein structure in a PDB format or PDB accession code.	Single mutation (code and mutation chain) or residue/position code and the mutation chain.
DUET	Prediction	Wild-type protein structure in a PDB format or PDB accession code.	Single mutation (code and mutation chain)
DynaMut	Analysis Prediction	Wild-type protein structure in a PDB format or PDB accession code. Wild-type protein structure in a PDB format or PDB accession code.	The selection of a Force Field and email (optional field). Single mutation (code and mutation chain) or file with a list of mutations and its respective chains, and email (optional field).
mCSM-PPI2	Prediction Analysis	The structure of the complex in PDB format or corresponding PDB accession code. The structure of the complex in PDB format or corresponding PDB accession code.	Single mutation (code and mutation chain) or file with a list of mutations and its respective chains, and email (optional field). Mutation details (alanine scanning or saturation mutagenesis) and email (optional field).
mCSM-NA	Prediction	The structure of the complex in PDB format or corresponding PDB accession code.	Single mutation (code and mutation chain) or file with a list of mutations and its respective chains, and the selection of the Nucleic Acid Type.
mCSM-Lig	Prediction	The structure of the complex in PDB format or corresponding PDB accession code.	Single mutation (code and mutation chain) and ligand information (three-letter ligand ID and estimated wild-type affinity).

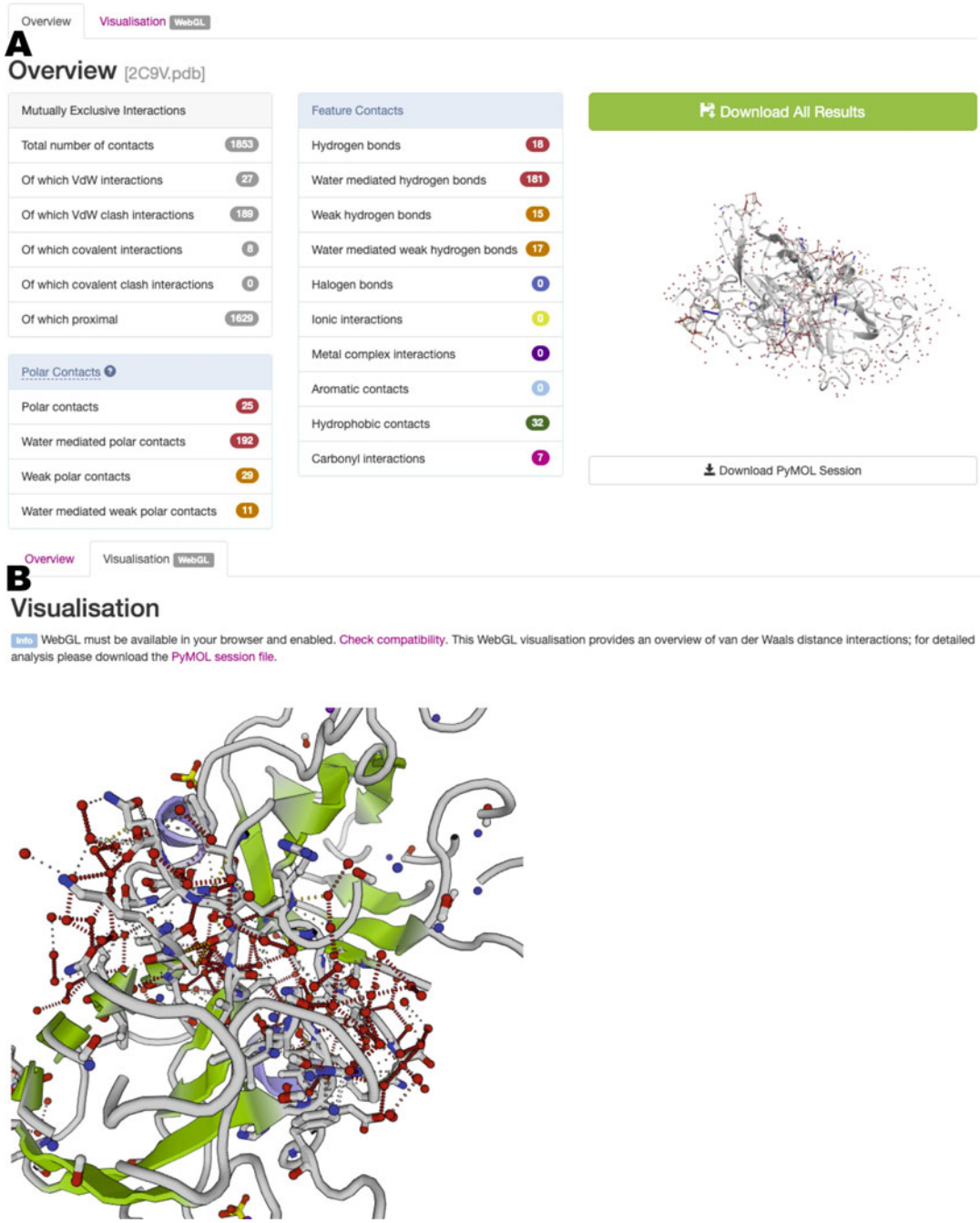


Fig. 2 Output of the Arpeggio tool. (**a**) Overview of the output for the inputted protein including the different types of interactions. (**b**) Visualization of the interactions shown on a protein structure

Table 5
Atomic interactions calculated by Arpeggio

Atomic interaction	Description	Arpeggio class	Bond energy (kJ/mol)
Van der Waals (dipole)	Permanent, induced and instantaneous dipoles	VWD	1–9
Hydrophobic	Between aliphatic and aromatic atoms	Hydrophobic	4–12
Hydrogen bond	Between carboxyl, amide, imidazole, guanidine, amino, hydroxyl and phenolic groups	Hydrogen bonds, weak hydrogen bond, polar contacts, halogen bonds, carbonyl interactions	8–40
Pi interactions	From/to rings	Aromatic contacts	6–70
Electrostatic	Between carboxyl and amino groups	Ionic interactions, metal complex	42–84

3.2.3 mCSM-Stability/PPI/DNA

The impact of mutations on protein stability, protein–protein binding affinity, and protein–DNA affinity can be predicted by mCSM-Stability, mCSM-PPI, mCSM-DNA with three types of prediction; single, multiple and systematic mutation.

Single Mutation

- If the single mutation option is selected in one of the tools within the mCSM platform, it will be shown on a results page after processing. This information includes the predicted value changes (protein stability, protein–protein interaction, protein–DNA interaction) as measured by the change in Gibbs Free Energy $\Delta\Delta G$ kcal/mol (Fig. 5), which is classified as highly destabilizing ($\Delta\Delta G \leq -2$ kcal/mol), destabilizing (-2 kcal/mol $< \Delta\Delta G < 0$ kcal/mol), stabilizing (0 kcal/mol $\leq \Delta\Delta G < 2$ kcal/mol), or highly stabilizing ($\Delta\Delta G \geq 2$ kcal/mol).
- If the structure of a complex is submitted to mCSM-Stability, it will calculate the predicted change in stability of the entire complex. It is therefore often advisable to also run predictions on a PDB file containing the protomer chain alone.
- For mCSM-PPI and mCSM-DNA, for mutations further than 12 Å from the interaction, the mCSM predictions are not considered, and are set to 0, as the graph-based signatures capture a smaller radius of environmental data, and there are fewer mutations located further away than 12 Å in the datasets used to train the methods.
- Also shown is an interactive 3D visual representation of the uploaded PDB file (Fig. 5a, right).

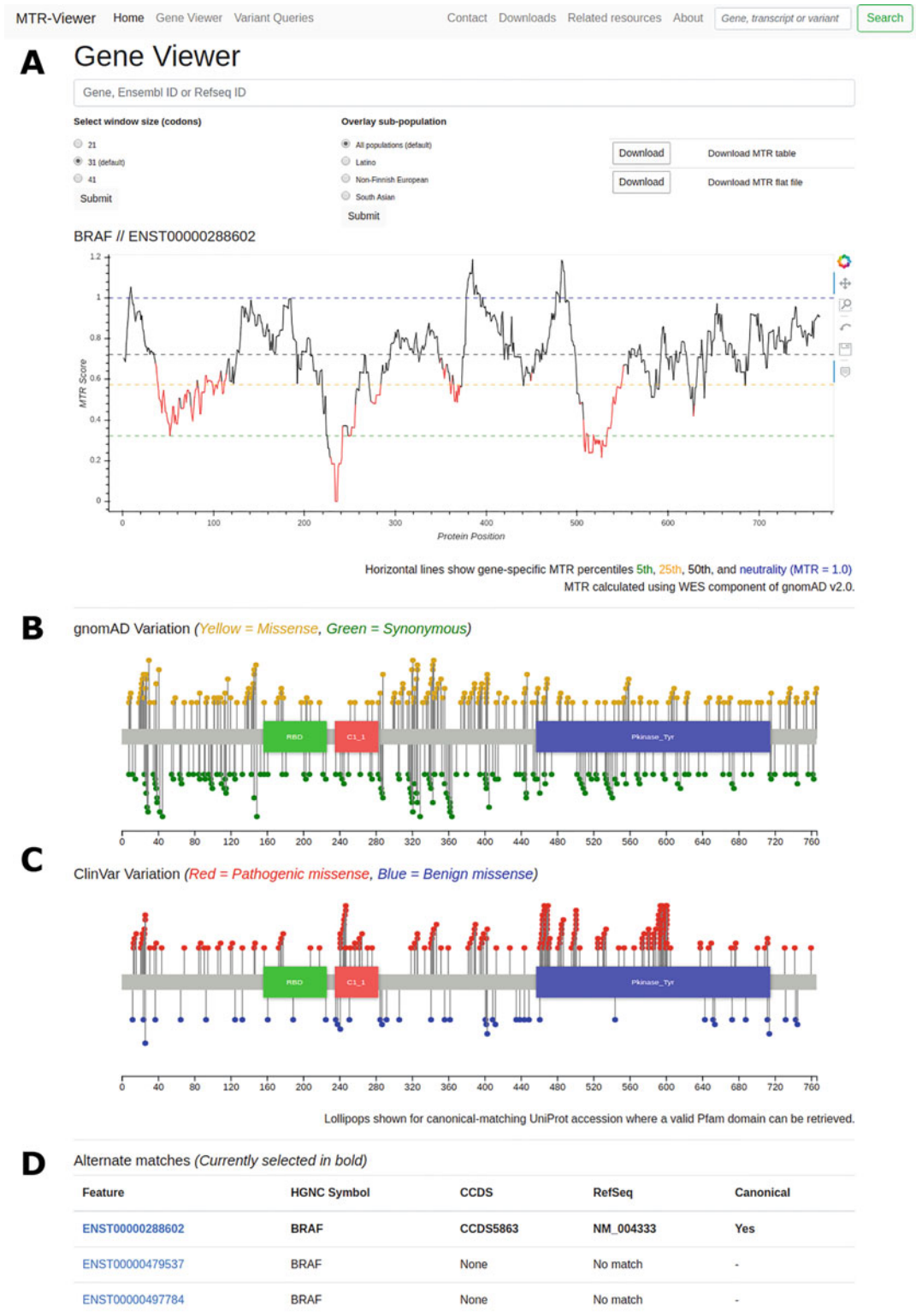
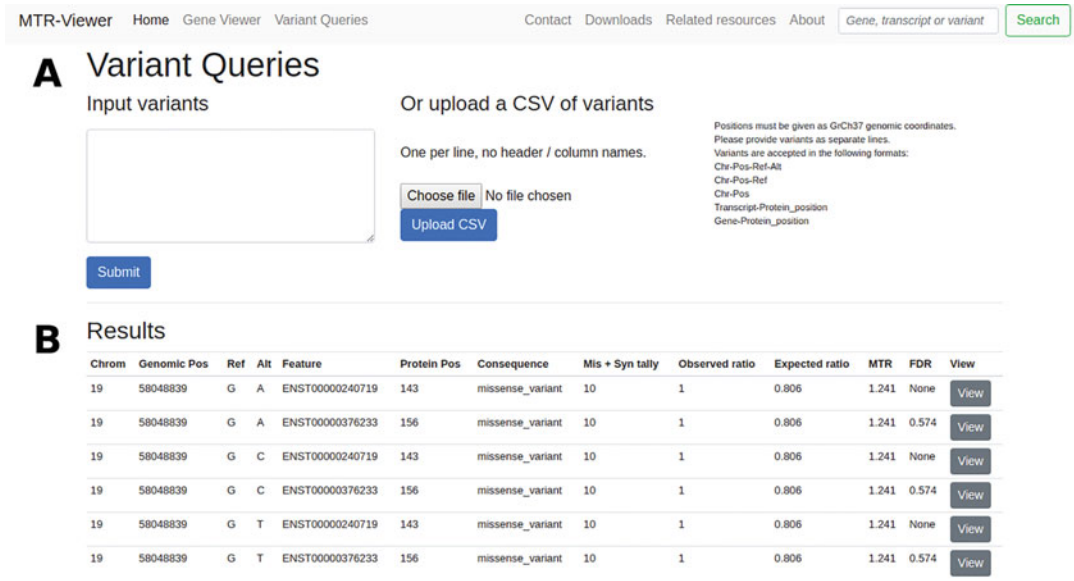


Fig. 3 The MTR Gene Viewer result page. (**a**) The line graph shows MTR scores in red for variations distant from neutrality across the transcript according to selected window size (codons) and subpopulation option. (**b**) The lollipop plot shows observed gnomeAD variation in yellow and green for missense and synonymous variation. (**c**) The second lollipop plot displays pathogenic (red) and benign (blue) missense variants based on ClinVar annotation. (**d**) The alternate transcripts can be shown in a table with RefSeq ID



Chrom	Genomic Pos	Ref	Alt	Feature	Protein Pos	Consequence	Mis + Syn tally	Observed ratio	Expected ratio	MTR	FDR	View
19	58048839	G	A	ENST00000240719	143	missense_variant	10	1	0.806	1.241	None	View
19	58048839	G	A	ENST00000376233	156	missense_variant	10	1	0.806	1.241	0.574	View
19	58048839	G	C	ENST00000240719	143	missense_variant	10	1	0.806	1.241	None	View
19	58048839	G	C	ENST00000376233	156	missense_variant	10	1	0.806	1.241	0.574	View
19	58048839	G	T	ENST00000240719	143	missense_variant	10	1	0.806	1.241	None	View
19	58048839	G	T	ENST00000376233	156	missense_variant	10	1	0.806	1.241	0.574	View

Fig. 4 MTR Variant Queries result page. Calculated results and information for the given input variants (or a CSV). User can check the details through MTR Gene Viewer by clicking on the view button

Multiple or Systematic

- If the option for inputting a list of mutations or systematic was used to analyze the PDB file, then after processing, results will be shown in tabulated form (Fig. 5b), including mutation specific information such as the residue solvent accessibility (RSA), as well as the predicted ΔΔG.
- Each result is also classified, using the predicted ΔΔG value, as highly destabilizing, destabilizing, stabilizing, or highly stabilizing.
- Users can search the result table or download results into a tab-separated text file.

3.2.4 SDM

SDM uses environment-specific amino acid substitution tables [38] and structural features including residue depth [15] and packing density to predict the impact of mutations on protein stability. The result page of single and list mutation is as follows.

Single Mutation

- The single mutation result page (Fig. 6a) provides predicted protein stability changes (ΔΔG), in addition to structural information implemented in SDM including secondary structure, RSA, residue depth and residue occluded packing density (OSP), sidechain–sidechain hydrogen bond (HBOND_SS), sidechain–main chain amide hydrogen bond (HBOND_SN), and sidechain–main chain carbonyl hydrogen bond (HBOND_SO). The integrated 3D viewer also shows the

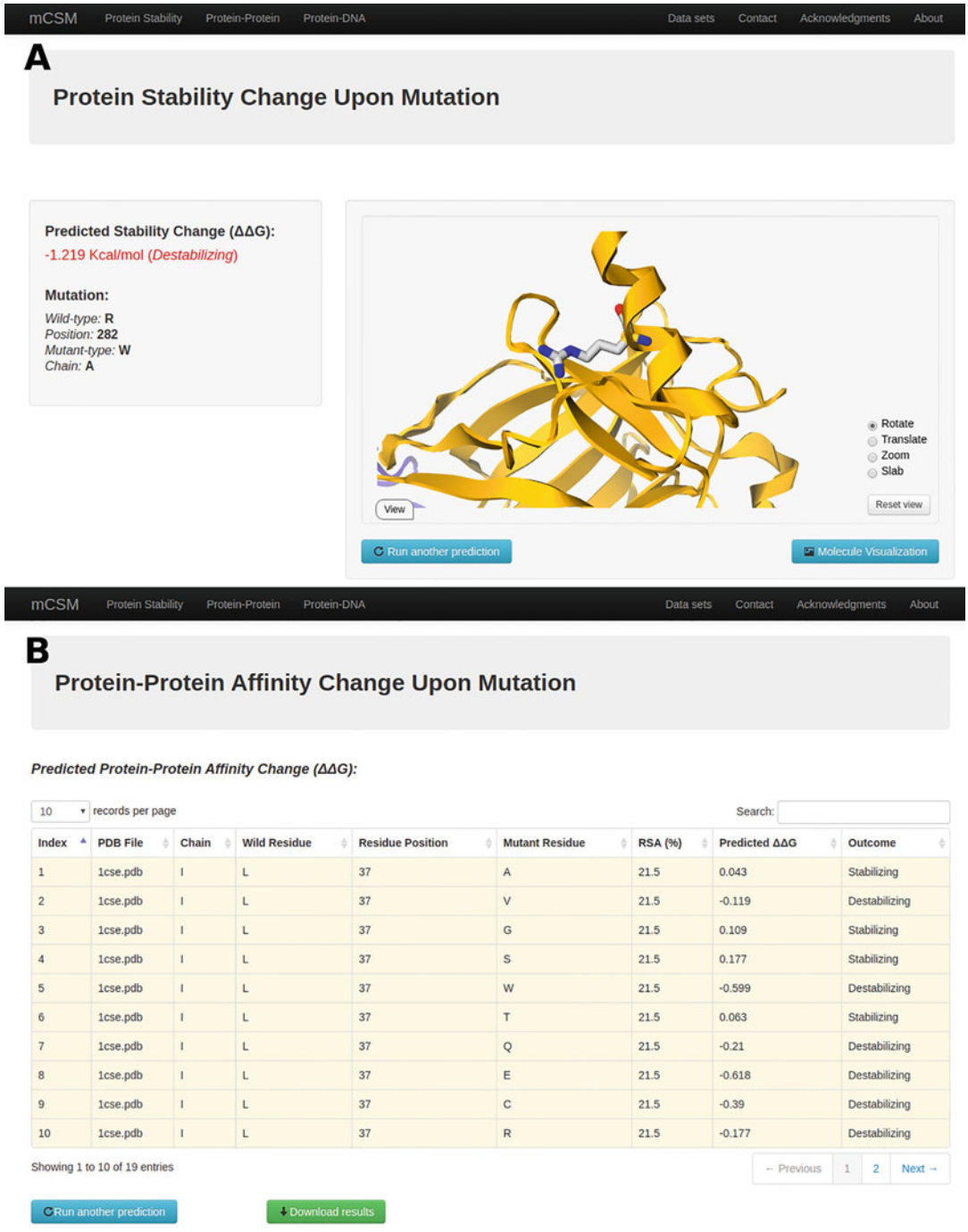


Fig. 5 Result pages for mCSM-Stability, mCSM-PPI and mCSM-DNA. (**a**) mCSM-Stability (single mutation) and (**b**) mCSM-PPI (multiple/systematic mutation). (**a**) The single prediction for example mCSM-Stability page supports 3D interactive viewer for structural analysis. (**b**) The results and information from multiple/systematic prediction for example mCSM-PPI are shown in a table

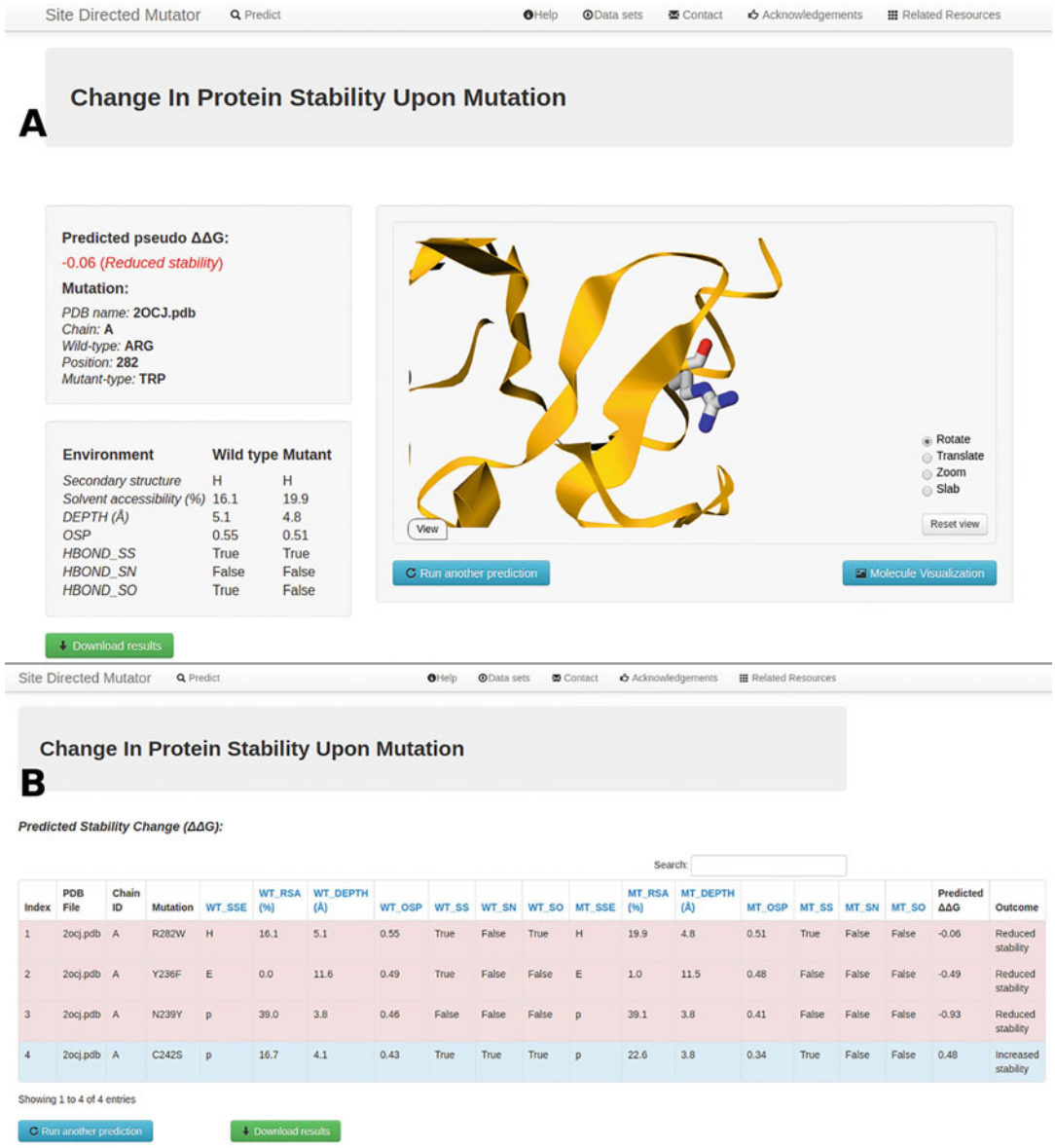


Fig. 6 SDM prediction results for single and list prediction. (**a**) The single prediction displays the predicted ΔΔG with information used on the left panel and 3D structure in a ribbon (protein) and a stick (wild-type amino acid) representation. (**b**) The list prediction gives detailed structural information and predicted ΔΔG in a tabulated form highlighted according to stabilizing (blue) and destabilizing (red) mutation

structure and its wild-type amino acids in ribbon and stick representation.

- Stability changes (ΔΔG) are shown in red with a negative sign if the mutation is predicted to be destabilizing, and in blue with a positive sign if the mutation is predicted to be stabilizing.

Multiple Mutations

- The predicted SDM ΔΔG for a given mutation list is displayed in a tabulated format (Fig. 6b) with their structural features. Users can download all mutant PDB structures and their predicted values in individual files.

3.2.5 DUET

Single Mutation

- The DUET result page (Fig. 7a) provides the predicted stability changes (ΔΔG) with integrated features such as secondary structure and stability changes from mCSM and SDM. While DUET refers to both mCSM and SDM scores, the prediction result can vary between the two methods.
- In the structure viewer (Fig. 7a right), the wild-type amino acid is shown in stick form and users can download the corresponding mutant structure file in PDB format.

Systematic Mutations

- With the systematic prediction (Fig. 7b), users can examine the predicted changes in protein stability using DUET, mCSM, and SDM for all nineteen possible mutations at a given residue position.
- The predictions and the structural information used to calculate the DUET scores are displayed in a downloadable table.

3.2.6 DynaMut

Users can use DynaMut to assess the impact of mutations on protein dynamics and stability with single and list mutation prediction.

Single Mutation

- The results of mutational effects on protein dynamics and stability are shown in Fig. 8a: ΔΔG predictions, interatomic interactions, deformation and fluctuation analysis.
- The ΔΔG prediction page provides predicted values from normal mode analysis (NMA)-based prediction (ΔΔG ENCoM), vibrational entropy energy changes ($\Delta\Delta S_{Vib}$ ENCoM), and other structure-based stability predictions (ΔΔG mCSM, ΔΔG SDM, ΔΔG DUET). Users can visually assess mutational effects on protein flexibility which is colored on the protein structure by vibrational entropy (Fig. 8b) for the region gaining (red) or losing (blue) flexibility. This 3D representation can be downloaded into a Pymol session, high resolution image and CSV file.
- Through the interatomic interactions tab, users can compare molecular interactions between wild-type and mutant structures. The PDB structure with interatomic interactions can be retrieved as a Pymol session file.
- The mutational effects on protein dynamics are shown in the deformation and fluctuation tab. Users can evaluate changes in the amount of local flexibility and atomic fluctuation upon mutation in 3D visual representation; results are downloadable as a CSV file and a Pymol session file.

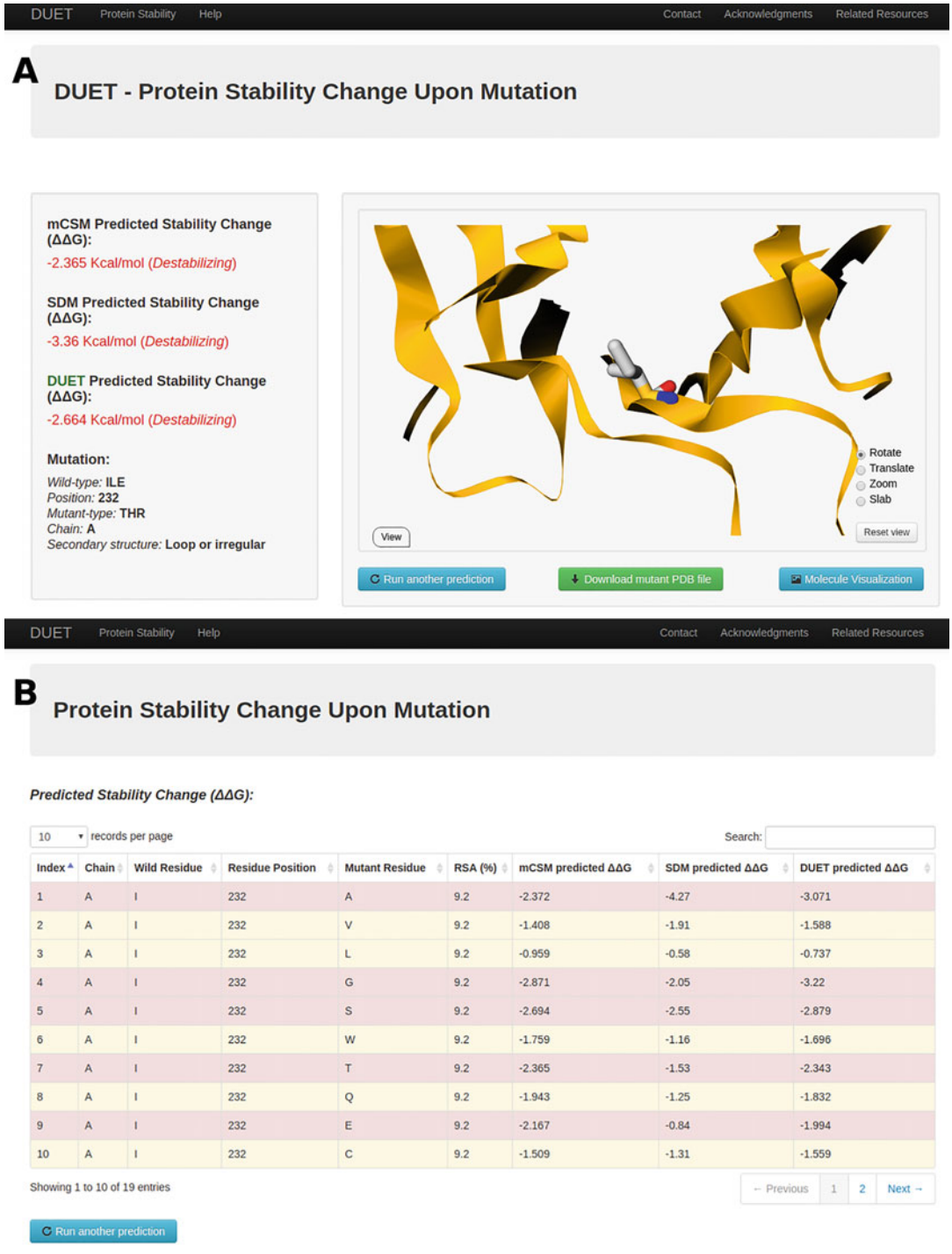


Fig. 7 DUET result pages for single and systematic prediction. (**a**) The single prediction result of DUET shows predicted ΔΔG across SDM and mCSM-Stability with mutation details. (**b**) Systematic prediction results including ΔΔG from DUET, SDM and mCSM-Stability and relative solvent accessible area of wild-type structure

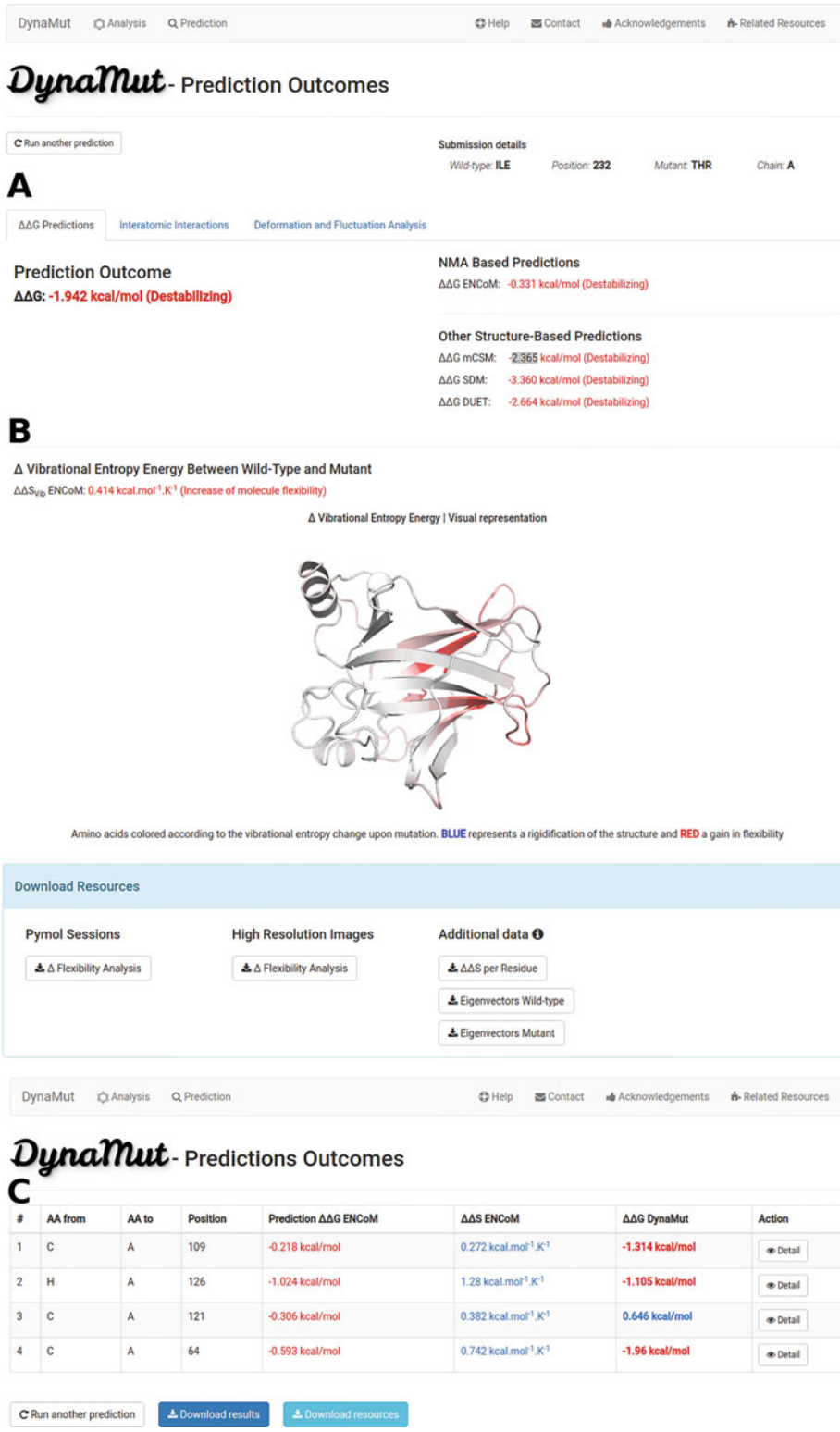


Fig. 8 DynaMut result pages. The single prediction shows predicted DynaMut ΔΔG (**a**, left) and predicted protein stability (ΔΔG) from mCSM-Stability, SDM and DUET and flexibility changes (ΔΔG ENCoM). Users can check vibrational energy changes upon mutation in the panel B. For a multiple mutation, (**b**) list prediction result page shows predicted DynaMut ΔΔG and links to access the corresponding single prediction in table

Multiple Mutations

- For a given mutation list, DynaMut gives all predicted values, including $\Delta\Delta G_{Stability}$ENCoM, $\Delta\Delta S_{Vib}$ ENCoM, and $\Delta\Delta G_{Stability}$ DynaMut, in table format (Fig. 8c). A more detailed analysis is available through the single prediction page of each mutation by clicking on the "Detail" button.

3.2.7 *mCSM-PPI2*

mCSM-PPI2 supports two types of protein–protein affinity prediction: mutation prediction and binding analysis. Mutation prediction gives predicted protein–protein affinity changes based on a given protein–protein complex and the mutation information. Binding analysis considers interface residues within 5 Å from different chains in the complex structure for alanine scanning and saturation mutagenesis.

Single Mutation

- mCSM-PPI2 displays predicted binding affinity changes (ΔΔG) upon mutation in two classes, destabilizing and stabilizing. Mutation details such as the distance to the interface from the given mutation position are also shown (Fig. 9).
- For mutations further than 12 Å from the interaction, the mCSM predictions are not considered, and are set to 0, as the graph-based signatures capture a smaller radius of environmental data, and there were fewer mutations located further away than 12 Å in the datasets used to train the methods.
- Users can assess the mutational impact in atomic/residue level through a 3D interactive viewer and a 2D graph. The molecular viewer provides Arpeggio inter/intra interactions for wild-type and mutant structures and the interaction changes between wild-type and mutant allows for investigation of the relationship between nonbonded interaction and protein–protein affinity. For residue-level analysis, the 2D graph can be used to study interresidue interactions of wild-type and mutant in a simple and user-friendly representation.

List Mutation

- For multiple mutation analysis, the result page tabulates predicted ΔΔG with mutation details. Users can access detailed results of each mutation through the single mutation result page and download all entries as a CSV file.

Alanine Scanning

- To identify residues with a greater contribution to the energy of binding (hot-spot) at the interface of interaction, alanine scanning can be used by predicting protein–protein binding affinity changes upon mutations to alanine across all identified interface residues. The predicted ΔΔG values are displayed in table, bar chart, and 3D viewer (Fig. 10a).
- Users can assess the effects of alanine mutation on the interface residues through a bar graph and 3D viewer colored in red and blue for destabilizing and stabilizing mutations, respectively.

Fig. 9 mCSM-PPI2 single prediction result page. The predicted ΔΔG is shown along with two interaction viewers: 3D interactive molecule viewer for atomic interaction analysis and 2D diagram for residue-level interaction analysis

Fig. 10 mCSM-PPI2 interface scanning result pages. The result pages of (**a**) alanine scanning and (**b**) saturation mutagenesis provide a bar chart and a heatmap colored by predicted ΔΔG and average predicted ΔΔG from the nineteen possible mutations, respectively

Saturation Mutagenesis

- The saturation mutagenesis provides the most exhaustive prediction, showing predicted ΔΔG for all identified interface residues when they are changed into nineteen different amino acids. The results are shown in table, heatmap, and 3D molecule viewer, and the interface residues of the 3D viewer are colored by the average ΔΔG of all mutations for each residue.

3.2.8 mCSM-NA

Single Prediction

- The predicted protein–nucleic acid affinity changes on a given structure are shown (Fig. 11a) with other properties such as the type of nucleic acid, solvent accessibility of wild-type protein, and predicted mutational effects from mCSM-Stability.
- For mutations further than 12 Å from the interaction, the mCSM predictions are not considered, and are set to 0, as the graph-based signatures capture a smaller radius of environmental data, and there were fewer mutations located further away than 12 Å in the datasets used to train the methods.
- The molecule visualization panel shows the protein–nucleic acid complex with the wild-type amino acid, and the mutation as a stick representation. mCSM-NA allows users to further investigate inter/intraresidue interactions by downloading Pymol session file.

List Mutation

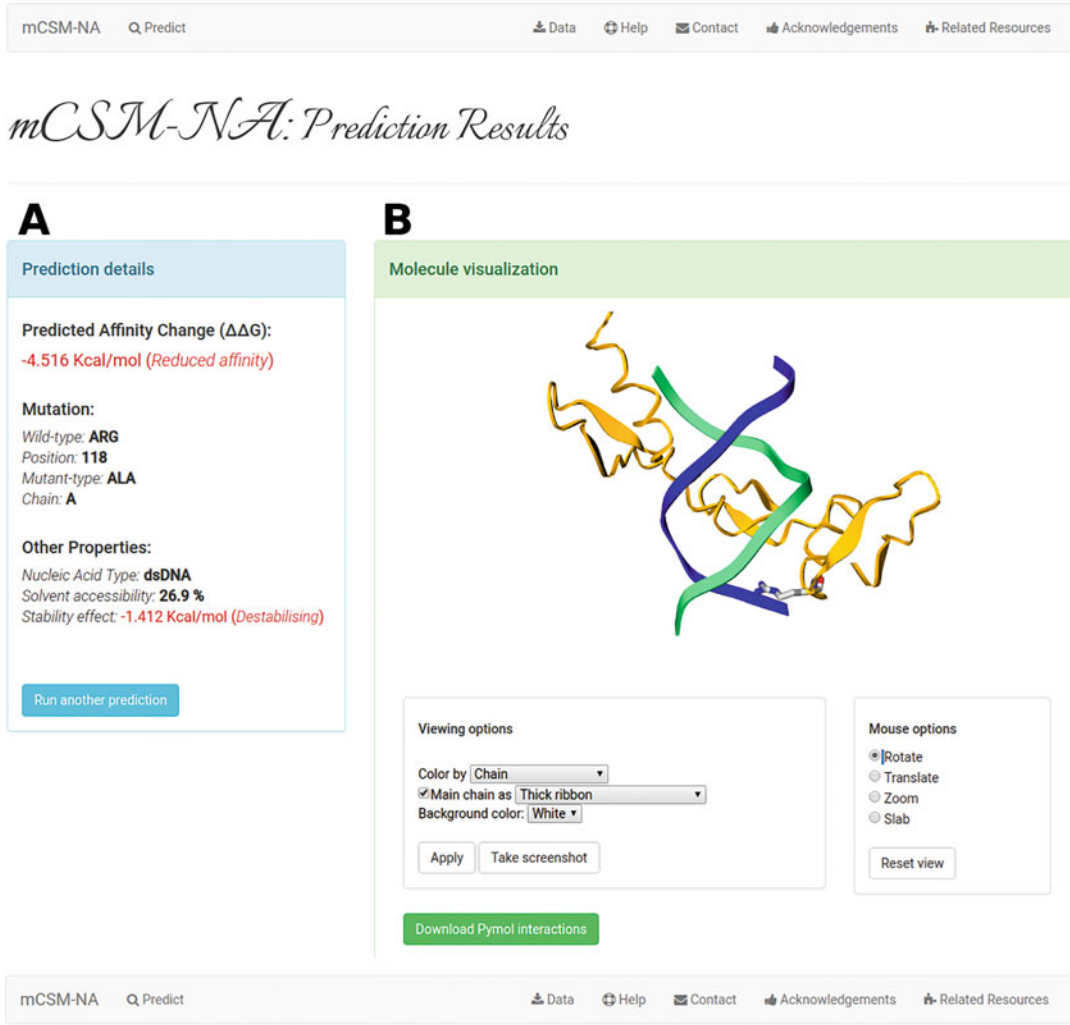
- mCSN-NA provides predicted protein–nucleic acid affinity changes, wild-type RSA, and mutation information for a given list of mutations in a table which is also downloadable in TSV format.

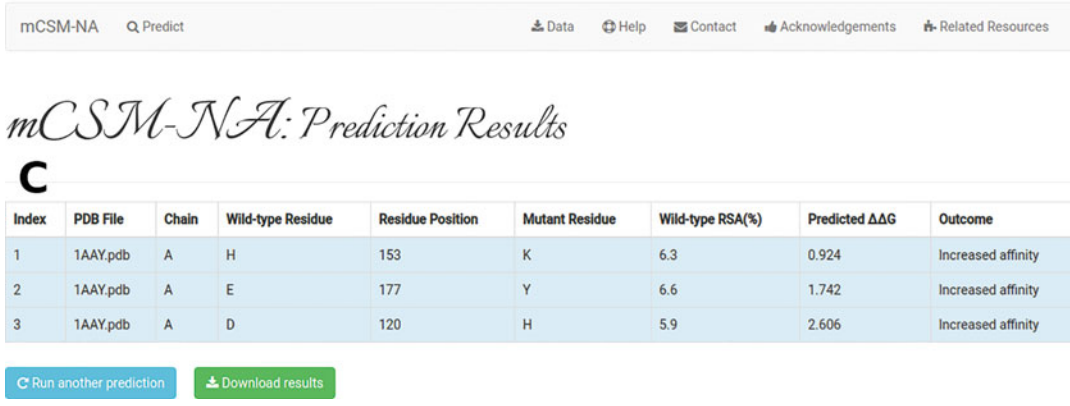
3.2.9 mCSM-lig

- mCSM-lig predicts affinity changes (log affinity fold) between a protein and its ligand upon mutation (Fig. 12a) using additional information such as the closest distance between wild-type residue and ligand and the protein stability change (Kcal/mol) from DUET. The stabilizing and destabilizing mutations are shown in positive and negative values respectively.
- For mutations further than 12 Å from the interaction, the mCSM predictions are not considered, and are set to 0, as the graph-based signatures capture a smaller radius of environmental data, and there were fewer mutations located further away than 12 Å in the datasets used to train the methods.
- The wild-type amino acid and ligand are shown in stick and sphere representations in 3D molecule viewer, respectively.

3.3 Identification of Driving Molecular Consequences

The outputs of the predictive tools described above provide the basis for an initial heuristic examination. When trying to interpret the molecular consequences of a specific variant, it is important to remember that phenotypic outcomes are often the result of the





Index	PDB File	Chain	Wild-type Residue	Residue Position	Mutant Residue	Wild-type RSA(%)	Predicted ΔΔG	Outcome
1	1AAY.pdb	A	H	153	K	6.3	0.924	Increased affinity
2	1AAY.pdb	A	E	177	Y	6.6	1.742	Increased affinity
3	1AAY.pdb	A	D	120	H	5.9	2.606	Increased affinity

Fig. 11 mCSM-NA result pages for single and list mutation prediction. In the single prediction result page, predicted protein–DNA affinity changes and mutation information are displayed in the prediction details (**a**) and the 3D viewer shows protein–DNA complex and wild-type amino acid in a ribbon and stick representation (**b**). The results of list prediction are shown in a tabulated form (**c**) and users can save the results in a TSV format

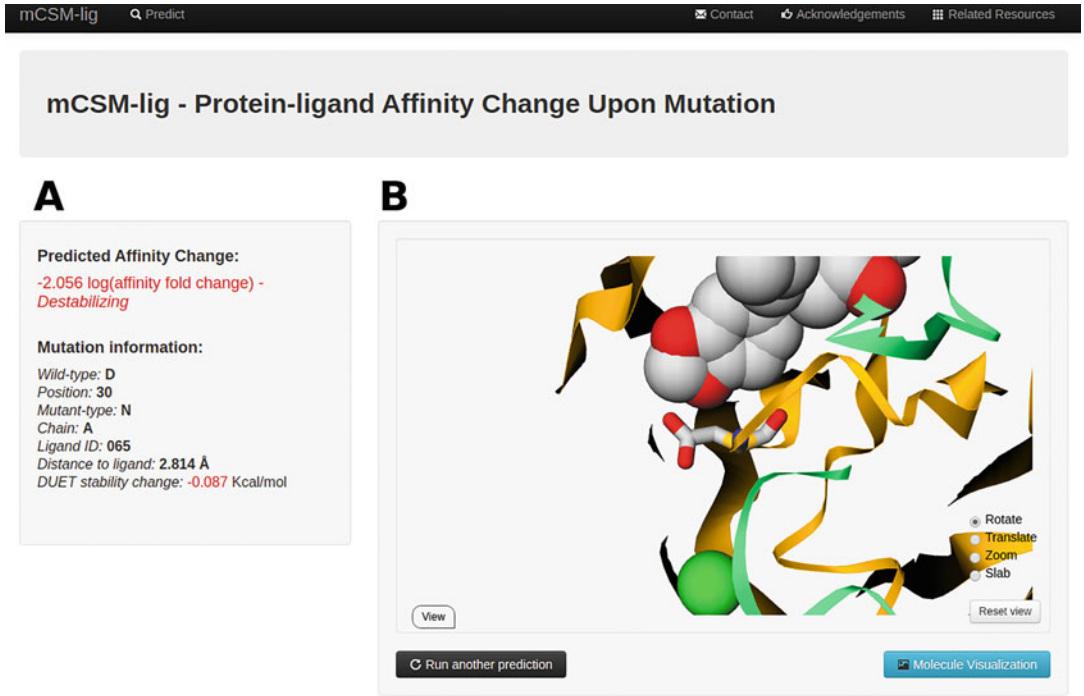


Fig. 12 mCSM-lig result page. (**a**) The predicted affinity change between protein and ligand upon mutation is shown in logarithm scale. (**b**) The protein and ligand are displayed in 3D viewer with a ribbon (for protein), a stick (for wild-type amino acid), and a sphere (for ligand) representation

combination of multiple molecular changes. For coding mutations, we initially ask ourselves three questions:

1. Is the mutation within 5 Å of an interface? If so, is the mutation more likely to disrupt the interaction ($\Delta\Delta G < \pm 0.5$ kcal/mol) based on the corresponding mCSM output (e.g., mCSM-PPI, mCSM-DNA, mCSM-NA, mCSM-Lig)? If the mutation is further than 12 Å away, it is less likely to disrupt the interaction directly, so the mCSM predictions are less reliable.
2. Is the mutation likely to disrupt protein folding and stability? mCSM-Stability, SDM, DUET, and DynaMut provide insight into this, with mutations leading to $\Delta\Delta G < \pm 0.5$ kcal/mol more likely to have a significant biological effect. Mutations at buried residues are more likely to have a larger effect on protein stability.
3. Is the mutation a special case that is more likely to lead to disruption of the protein due to unique geometry restraints of the residues (*see* **Notes 3** and **4**)?

To more exhaustively explore how mutations in a protein lead to a phenotype, and to identify those molecular features that best

capture the driving of the molecular mechanisms, an investigation into the performance of each inputted feature should be conducted in order to construct the highest performing predictive model.

A more robust method for selecting which features are most informative can be performed using feature selection in R, a statistical programming language. While R is powerful enough itself to create classification models, we can also use it to measure which features from our predictive tools' output are most effective in stratifying mutations. Two effective approaches are:

1. A random forest classification algorithm to measure feature importance using a set of mutations with known class labels (e.g., pathogenic/nonpathogenic, deleterious/nondeleterious).
2. The Boruta Algorithm performs permutations of the data to statistically compare each feature's importance with that attainable at random, and uses this to eliminate uninformative features. The package in R provides a graphical output using boxplots.

Features that score highly provide evidence that the molecular consequence that they measure is relevant to how mutations lead to the phenotype of interest. The algorithm can also highlight correlation between features. When two or more features are highly correlated and are likely measuring the same information, only one should be used in subsequent predictive model development to remove redundancy, minimize noise and avoid bias from weighting a model in favor of a particular attribute. The model should also have the fewest possible features that perform best. Using too many features may generate a model that performs accurately on training data but cannot be generalized to real-world data.

3.4 Machine Learning Phenotypes: Building a Predictive Classifier

An initial understanding of molecular mechanisms imparted by disease-causing mutations is a crucial step toward establishing a genotype–phenotype correlation. However, manual analysis of different results can often miss underlying, statistically significant relationships among different mutational measurements, which can help relate them to the phenotype. Machine learning, and in particular supervised learning, addresses this issue by providing a set of tools for the efficient analysis of labeled data (e.g., experimentally characterized mutations) in order to derive a model that describes a phenomenon, aiming for generalization (applying it to unseen data). The identification of patterns and associations within the data will further help the predictive model establish a distinction between mutations within the same gene leading to different phenotypes, and hence the development of an effective predictive tool that can be used to interpret novel clinical variants.

Here, our goal is to build a machine learning classifier to distinguish between pathogenic vs. nonpathogenic mutations in a

given gene. Multiple steps are required to obtain a nonbiased, accurate predictor:

1. Dataset curation: Machine learning algorithms require a well-curated dataset. In a supervised machine learning approach, all data labels (here, pathogenic or nonpathogenic for each mutation) must be known in order to enable correlations to be assessed between labels (e.g., phenotypes) and features/properties used as evidence to represent each data point (e.g., mutations). The quality of a classifier directly depends on the quality of the data used to build it, so accurate clinical sources are required to justify labeling mutations as pathogenic or nonpathogenic. In this case, generally, nonpathogenic variants can be curated from population variant databases such as GnomAD, usually taking into account frequent mutations. Even common variants, however, may still be linked to a disease, especially if it is a weakly penetrative mutation or recessive condition, which would add noise to the data set and thus complicate the task of building a general predictive model. In situations where other biologically relevant information is present, such as cellular fitness cost, it is essential that this type of information is present for every mutation in a dataset, as a supervised algorithm cannot handle missing data labels. The initial dataset should contain a representative set of mutations within all phenotype classes (pathogenic and nonpathogenic), and ideally, present a balanced number of instances between classes, to minimize biases toward overrepresented classes in the resultant model. More details on metrics used to evaluate the performance of predictive models on an imbalanced dataset are discussed below.
2. Feature generation: The feature generation stage is crucial as it provides descriptive information about each mutation, to be used by the learning algorithm to finally classify the phenotype of a mutation. As described above, features can encompass a diverse range of mutational information:
   (a) Protein stability and dynamics (mCSM-Stability, DUET, SDM, Dynamut).
   (b) Protein functional changes such as changes in affinity for other proteins (mCSM-PPI2), nucleic acids (mCSM-NA), and ligands (mCSM-lig).
   (c) At the residue level, changes in protein pharmacophore and local residue environment such as changes in interatomic interactions (Arpeggio) are also important, as some mutations at the same locus can have different phenotypes.
   (d) Sequence-level predictors (SIFT, Polyphen, SNAP2).

(e) Evolutionary-based predictors (ConSurf), population based mutational tolerance (MTR-Viewer), as well as amino acid substitution matrices (e.g., PAM30, BLOSUM62, PSSM) offer added information on the likelihood of one mutation to change into another.

Feature generation is directly dependent on the wild-type biological functions of the protein, which is why an understanding of the biological relevance is important at the very beginning of this process.

3. Training and Testing sets: The data collected must be divided into training and testing sets to assess the generalization power of a classifier, that is, its ability to correctly predict on new data, and to ensure that it has not been over- or undertrained. Data used to train the model should be different, nonredundant, from the data used to test the model. It is common practice to divide the original dataset into Training and Test sets at the start of learning. For small datasets, a large proportion of the data may need to be segregated into the Test set to provide sufficient data to accurately measure performance of the trained model. This can be done in a bootstrapping procedure or through cross-validation, when the original data set is divided into *k*-folds and each is taken iteratively as the test set while remaining data are used in training (*k*-fold cross-validation).
4. Feature selection: The features selected for training can strongly influence accuracy, so it is important to select only informative features, and eliminate irrelevant or nondiscriminative ones, which are a common source of noise. Feature selection can also help reduce overfitting and reduce training time, as it aims to generate simpler, more concise models. Feature selection methods provided in the Python machine learning library, Scikit-Learn [58], include univariate selection, feature importance, correlation matrix, and recursive feature elimination or addition. Alternatively, forward stepwise selection can be performed as a greedy heuristic in which features are included iteratively, one at a time, based on their individual performance contributions.
5. Machine learning platforms: Different tools have been developed for implementing machine learning. Some offer a graphical user interface (GUI), such as Weka [59], while some run as python packages through the command line, such as Scikit-Learn. Different packages for different programming languages offer similar algorithms and options to adjust the algorithm parameters according to specific tasks. The major classification algorithms we test are Naive Bayes, Decision Trees, K-Nearest Neighbor, Support Vector Machines, and Ensemble Classifiers. It is good practice to compare

representative algorithms of each class, provided that the algorithm is compatible with the dataset type. Within weka, this can be done automatically using the auto-weka function. In cases where the training set is unbalanced, oversampling or undersampling of the training data can be used to achieve a better representation of classes within the classification model-building stage, preventing model bias in always detecting the predominant class and achieving a false high performance.

6. Model validation: The primary tool in the validation of a model is the use of a nonredundant independent test set, also called blind test.

   Validation can be furthered using internal data testing such as k-fold cross validation, in which the dataset is divided into k subsets. One subset is used as a test set, while the remaining $(k - 1)$ subsets are used to train a model. The process is repeated k times, until all the data have been used in both training and test sets. The final model performance is calculated as the average of the performances of all k iterations. We will often vary k based on the size of the dataset. When the training set is small (e.g., ~200 data points), we may use leave-one-out validation, where k is equal to the size of the dataset. An important aspect when selecting predictive models is consistency in performance between the training and test sets. This usually indicates a robust model, within which discrepancies (e.g., a much higher performance on training than with the test set) might indicate overfitting.

7. Model evaluation: Several different evaluation metrics may be used for classification tasks. These are generally calculated on values obtained from a confusion matrix, which is a summary of the data points, and their actual and predicted phenotypes (Table 6).

8. From the distributions of data points within the matrix, descriptive metrics can be calculated:

   (a) accuracy (number of correct predictions: [(TP + TN)/TOTAL]),

   (b) precision (rate of correctly predicted positive instances from all assigned as positives: [TP/(TP + FP)],

Table 6
Description of a confusion matrix

Predicted value	Actual value Positive	Negative
Positive	True positive	False positive
Negative	False negative	True negative

(c) recall (rate of correctly predicted positive instances from all real positive instances: [TP/(TP + FN)],

(d) *f*-score (a weighted average of recall and precision), and,

(e) Matthews correlation coefficient (MCC) a balanced measure between true positives and true negatives

$$\left[(\mathrm{TP} \times \mathrm{TN}) - (\mathrm{FP} \times \mathrm{FN}) / \sqrt{(\mathrm{TP} + \mathrm{FP})(\mathrm{TP} + \mathrm{FN})(\mathrm{TN} + \mathrm{FP})(\mathrm{TN} + \mathrm{FN})}\right]$$

where TP = True positive; TN = True negative; FP = False positive; and FN = False negative.

Classifier performance can also be described graphically using a Receiver Operating Characteristic curve, which compares the TP Rate and TN Rate. The closer the area under the curve is to 1, the better the classifier performance.

These metrics should be used in a combinatorial fashion across all elements of training, test, and cross-validation stages to compare model performance during different stages of classifier optimization. When the dataset is imbalanced, balanced measures such as MCC should be prioritized, as other measures might bias for an overtrained model on the dominant dataset.

4 Notes

1. Often following curation, the distribution of number of pathogenic and benign mutations is unbalanced, which can affect efforts to build predictive tools using machine learning. Two approaches that can help include oversampling of the underrepresented class, or undersampling of the overrepresented class. Evaluation metrics that are less biased toward unbalanced classes, such as the Matthew's correlation coefficient, precision-recall curves, and Kendall correlations, should also be preferentially used.
2. The chain ID for the provided PDB file is a mandatory field for all the structure-based methods; blank characters are not allowed. It is possible that homology modeling tools might not automatically add a chain ID. If this is the case, the user will need to modify the PDB file prior to submission to the servers. Several tools exist to perform this task (e.g., http://www.canoz.com/sdh/renamepdbchain.pl).
3. Special cases: Mutations to and from prolines. Prolines are the only amino acid whose amino group is connected to the side-chain, which in the context of the peptide bond greatly limits torsional angles. The nature of this residue therefore needs to be taken into account while analyzing mutation effects. For instance, (1) mutations to prolines in the middle of alpha-

helices can introduce kinks, affecting local structure and (2) since prolines are commonly found in turns and loops, their substitution might interfere with the formation of secondary structures such as hairpins.

4. Special cases: mutations of positive-phi glycines. Similarly to prolines, positive phi glycines, while rare in experimental structures, deserve special consideration due to their torsional angles. Glycines are the only residues capable of adopting positive-phi angles. These glycines are usually conserved across evolution, meaning that mutations on positive-phi glycines, especially on loops and hairpins, tend to be destabilizing.

Acknowledgments

This work was supported by Australian Government Research Training Program Scholarships [to S.P., M.K., Y.M., C.H.M.R.]; the Jack Brockhoff Foundation [JBF 4186, 2016 to D.B.A.]; a Newton Fund RCUK-CONFAP Grant awarded by The Medical Research Council (MRC) and Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG) [MR/M026302/1 to D.B.A. and D.E.V.P.]; and the National Health and Medical Research Council of Australia [APP1072476 to D.B.A.].

References

1. Jatana N, Ascher DB, Pires DEV et al (2019) Human LC3 and GABARAP subfamily members achieve functional specificity via specific structural modulations. Autophagy:1–17. https://doi.org/10.1080/15548627.2019.1606636
2. Abayakoon P, Jin Y, Lingford JP et al (2018) Structural and biochemical insights into the function and evolution of sulfoquinovosidases. ACS Cent Sci 4(9):1266–1273. https://doi.org/10.1021/acscentsci.8b00453
3. Ascher DB, Cromer BA, Morton CJ et al (2011) Regulation of insulin-regulated membrane aminopeptidase activity by its C-terminal domain. Biochemistry 50(13):2611–2622. https://doi.org/10.1021/bi101893w
4. Portelli S, Phelan JE, Ascher DB et al (2018) Understanding molecular consequences of putative drug resistant mutations in Mycobacterium tuberculosis. Sci Rep 8(1):15356. https://doi.org/10.1038/s41598-018-33370-6
5. Silk M, Petrovski S, Ascher DB (2019) MTR-Viewer: identifying regions within genes under purifying selection. Nucleic Acids Res 47(W1):W121–W126. https://doi.org/10.1093/nar/gkz457
6. Pires DE, Blundell TL, Ascher DB (2015) Platinum: a database of experimentally measured effects of mutations on structurally defined protein-ligand complexes. Nucleic Acids Res 43(Database issue):D387–D391. https://doi.org/10.1093/nar/gku966
7. Lucy G, Douglas EVP, Álvaro O-N et al (2014) An integrated computational approach can classify VHL missense mutations according to risk of clear cell renal carcinoma. Human Molecular Genetics, 23(22):5976–5988. https://doi.org/10.1093/hmg/ddu321
8. Blaszczyk M, Harmer NJ, Chirgadze DY et al (2015) Achieving high signal-to-noise in cell regulatory systems: spatial organization of multiprotein transmembrane assemblies of FGFR and MET receptors. Prog Biophys Mol Biol 118(3):103–111. https://doi.org/10.1016/j.pbiomolbio.2015.04.007
9. Jafri M, Wake NC, Ascher DB et al (2015) Germline mutations in the CDKN2B tumor suppressor gene predispose to renal cell carcinoma. Cancer Discov 5(7):723–729. https://doi.org/10.1158/2159-8290.CD-14-1096

10. Pacitto A, Ascher DB, Wong LH et al (2015) Lst4, the yeast Fnip1/2 orthologue, is a DENN-family protein. Open Biol 5 (12):150174. https://doi.org/10.1098/rsob.150174
11. Pires DE, Chen J, Blundell TL et al (2016) In silico functional dissection of saturation mutagenesis: interpreting the relationship between phenotypes and changes in protein stability, interactions and activity. Sci Rep 6:19848. https://doi.org/10.1038/srep19848
12. Albanaz ATS, Rodrigues CHM, Pires DEV et al (2017) Combating mutations in genetic disease and drug resistance: understanding molecular mechanisms to guide drug design. Expert Opin Drug Discov 12(6):553–563. https://doi.org/10.1080/17460441.2017.1322579
13. Casey RT, Ascher DB, Rattenberry E et al (2017) SDHA related tumorigenesis: a new case series and literature review for variant interpretation and pathogenicity. Mol Genet Genomic Med 5(3):237–250. https://doi.org/10.1002/mgg3.279
14. Jubb HC, Pandurangan AP, Turner MA et al (2017) Mutations at protein-protein interfaces: small changes over big surfaces have large impacts on human health. Prog Biophys Mol Biol 128:3–13. https://doi.org/10.1016/j.pbiomolbio.2016.10.002
15. Pandurangan AP, Ascher DB, Thomas SE et al (2017) Genomes, structural biology and drug discovery: combating the impacts of mutations in genetic disease and antibiotic resistance. Biochem Soc Trans 45(2):303–311. https://doi.org/10.1042/BST20160422
16. Sibanda BL, Chirgadze DY, Ascher DB et al (2017) DNA-PKcs structure suggests an allosteric mechanism modulating DNA double-strand break repair. Science 355 (6324):520–524. https://doi.org/10.1126/science.aak9654
17. Rodrigues CH, Ascher DB, Pires DE (2018) Kinact: a computational approach for predicting activating missense mutations in protein kinases. Nucleic Acids Res 46(W1): W127–W132. https://doi.org/10.1093/nar/gky375
18. Hnizda A, Fabry M, Moriyama T et al (2018) Relapsed acute lymphoblastic leukemia-specific mutations in NT5C2 cluster into hotspots driving intersubunit stimulation. Leukemia 32 (6):1393–1403. https://doi.org/10.1038/s41375-018-0073-5
19. Andrews KA, Ascher DB, Pires DEV et al (2018) Tumour risks and genotype-phenotype correlations associated with germline variants in succinate dehydrogenase subunit genes SDHB, SDHC and SDHD. J Med Genet 55 (6):384–394. https://doi.org/10.1136/jmedgenet-2017-105127
20. Usher JL, Ascher DB, Pires DE et al (2015) Analysis of HGD gene mutations in patients with alkaptonuria from the United Kingdom: identification of novel mutations. JIMD Rep 24:3–11. https://doi.org/10.1007/8904_2014_380
21. Nemethova M, Radvanszky J, Kadasi L et al (2016) Twelve novel HGD gene variants identified in 99 alkaptonuria patients: focus on 'black bone disease' in Italy. Eur J Hum Genet 24(1):66–72. https://doi.org/10.1038/ejhg.2015.60
22. Ramdzan YM, Trubetskov MM, Ormsby AR et al (2017) Huntingtin inclusions trigger cellular quiescence, deactivate apoptosis, and lead to delayed necrosis. Cell Rep 19(5):919–927. https://doi.org/10.1016/j.celrep.2017.04.029
23. Traynelis J, Silk M, Wang Q et al (2017) Optimizing genomic medicine in epilepsy through a gene-customized approach to missense variant interpretation. Genome Res 27 (10):1715–1729. https://doi.org/10.1101/gr.226589.117
24. Trezza A, Bernini A, Langella A et al (2017) A computational approach from gene to structure analysis of the human ABCA4 transporter involved in genetic retinal diseases. Invest Ophthalmol Vis Sci 58(12):5320–5328. https://doi.org/10.1167/iovs.17-22158
25. Ascher DB, Spiga O, Sekelska M et al (2019) Homogentisate 1,2-dioxygenase (HGD) gene variants, their analysis and genotype-phenotype correlations in the largest cohort of patients with AKU. Eur J Hum Genet 27(6):888–902. https://doi.org/10.1038/s41431-019-0354-0
26. Soardi FC, Machado-Silva A, Linhares ND et al (2017) Familial STAG2 germline mutation defines a new human cohesinopathy. NPJ Genom Med 2:7. https://doi.org/10.1038/s41525-017-0009-4
27. Phelan J, Coll F, McNerney R et al (2016) Mycobacterium tuberculosis whole genome sequencing and protein structure modelling provides insights into anti-tuberculosis drug resistance. BMC Med 14:31. https://doi.org/10.1186/s12916-016-0575-9
28. Silvino AC, Costa GL, Araujo FC et al (2016) Variation in human cytochrome P-450 drug-metabolism genes: a gateway to the understanding of Plasmodium vivax relapses. PLoS One 11(7):e0160172. https://doi.org/10.1371/journal.pone.0160172

29. White RR, Ponsford AH, Weekes MP et al (2016) Ubiquitin-dependent modification of skeletal muscle by the parasitic nematode, Trichinella spiralis. PLoS Pathog 12(11): e1005977. https://doi.org/10.1371/journal.ppat.1005977
30. Hawkey J, Ascher DB, Judd LM et al (2018) Evolution of carbapenem resistance in Acinetobacter baumannii during a prolonged infection. Microb Genom 4(3). https://doi.org/10.1099/mgen.0.000165
31. Holt KE, McAdam P, Thai PVK et al (2018) Frequent transmission of the Mycobacterium tuberculosis Beijing lineage and positive selection for the EsxW Beijing variant in Vietnam. Nat Genet 50(6):849–856. https://doi.org/10.1038/s41588-018-0117-9
32. Karmakar M, Globan M, Fyfe JAM et al (2018) Analysis of a Novel pncA mutation for susceptibility to pyrazinamide therapy. Am J Respir Crit Care Med 198(4):541–544. https://doi.org/10.1164/rccm.201712-2572LE
33. Vedithi SC, Malhotra S, Das M et al (2018) Structural implications of mutations conferring rifampin resistance in Mycobacterium leprae. Sci Rep 8(1):5016. https://doi.org/10.1038/s41598-018-23423-1
34. Karmakar M, Rodrigues CHM, Holt KE et al (2019) Empirical ways to identify novel Bedaquiline resistance mutations in AtpE. PLoS One 14(5):e0217169. https://doi.org/10.1371/journal.pone.0217169
35. Ascher DB, Wielens J, Nero TL et al (2014) Potent hepatitis C inhibitors bind directly to NS5A and reduce its affinity for RNA. Sci Rep 4:4765. https://doi.org/10.1038/srep04765
36. Jubb HC, Higueruelo AP, Ochoa-Montano B et al (2017) Arpeggio: a web server for calculating and visualising interatomic interactions in protein structures. J Mol Biol 429(3):365–371. https://doi.org/10.1016/j.jmb.2016.12.004
37. Pires DE, Ascher DB, Blundell TL (2014) mCSM: predicting the effects of mutations in proteins using graph-based signatures. Bioinformatics 30(3):335–342. https://doi.org/10.1093/bioinformatics/btt691
38. Pandurangan AP, Ochoa-Montano B, Ascher DB et al (2017) SDM: a server for predicting effects of mutations on protein stability. Nucleic Acids Res 45(W1):W229–W235. https://doi.org/10.1093/nar/gkx439
39. Pires DE, Ascher DB, Blundell TL (2014) DUET: a server for predicting effects of mutations on protein stability using an integrated computational approach. Nucleic Acids Res 42(Web Server issue):W314–W319. https://doi.org/10.1093/nar/gku411
40. Douglas EVP, Carlos HMR, David BA et al (2020) mCSM-membrane: predicting the effects of mutations on transmembrane proteins, Nucleic Acids Research, gkaa416. https://doi.org/10.1093/nar/gkaa416
41. Rodrigues CH, Pires DE, Ascher DB (2018) DynaMut: predicting the impact of mutations on protein conformation, flexibility and stability. Nucleic Acids Res 46(W1):W350–W355. https://doi.org/10.1093/nar/gky300
42. Rodrigues CHM, Myung Y, Pires DEV et al (2019) mCSM-PPI2: predicting the effects of mutations on protein-protein interactions. Nucleic Acids Res 47(W1):W338–W344. https://doi.org/10.1093/nar/gkz383
43. Pires DE, Ascher DB (2016) mCSM-AB: a web server for predicting antibody-antigen affinity changes upon mutation with graph-based signatures. Nucleic Acids Res 44(W1):W469–W473. https://doi.org/10.1093/nar/gkw458
44. Yoochan M, Carlos HMR, David BA, Douglas EVP et al (2020) mCSM-AB2: guiding rational antibody design using graphbased signatures, Bioinformatics. 36(5):1453–1459. https://doi.org/10.1093/bioinformatics/btz779
45. Yoochan M, Douglas EVP, David BA et al. mmCSM-AB: guiding rational antibody engineering through multiple point mutations, Nucleic Acids Research, gkaa389. https://doi.org/10.1093/nar/gkaa389
46. Pires DEV, Ascher DB (2017) mCSM-NA: predicting the effects of mutations on protein-nucleic acids interactions. Nucleic Acids Res 45(W1):W241–W246. https://doi.org/10.1093/nar/gkx236
47. Pires DE, Blundell TL, Ascher DB (2016) mCSM-lig: quantifying the effects of mutations on protein-small molecule affinity in genetic disease and emergence of drug resistance. Sci Rep 6:29575. https://doi.org/10.1038/srep29575
48. Pires DE, Ascher DB (2016) CSM-lig: a web server for assessing and comparing protein-small molecule affinities. Nucleic Acids Res 44(W1):W557–W561. https://doi.org/10.1093/nar/gkw390
49. Douglas EVP et al (2011) Cutoff Scanning Matrix (CSM): structural classification and function prediction by protein inter-residue distance patterns. BMC genomics (12) No. S4. BioMed Central
50. Douglas EVP, Raquel CM-M, Carlos HS, Frederico FC, Wagner M Jr (2013) aCSM: noise-free graphbased signatures to large-scale receptor-based ligand prediction, Bioinformatics 29(7):855–861. https://doi.org/10.1093/bioinformatics/btt058

51. Sherry ST, Ward MH, Kholodov M et al (2001) dbSNP: the NCBI database of genetic variation. Nucleic Acids Res 29(1):308–311. https://doi.org/10.1093/nar/29.1.308
52. Stenson PD, Mort M, Ball EV et al (2017) The Human Gene Mutation Database: towards a comprehensive repository of inherited mutation data for medical research, genetic diagnosis and next-generation sequencing studies. Hum Genet 136(6):665–677. https://doi.org/10.1007/s00439-017-1779-6
53. Landrum MJ, Lee JM, Benson M et al (2018) ClinVar: improving access to variant interpretations and supporting evidence. Nucleic Acids Res 46(D1):D1062–D1067. https://doi.org/10.1093/nar/gkx1153
54. Karczewski KJ, Francioli LC, Tiao G et al (2019) Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. bioRxiv:531210. https://doi.org/10.1101/531210
55. Sudlow C, Gallacher J, Allen N et al (2015) UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS Med 12(3): e1001779. https://doi.org/10.1371/journal.pmed.1001779
56. UniProt Consortium T (2018) UniProt: the universal protein knowledgebase. Nucleic Acids Res 46(5):2699. https://doi.org/10.1093/nar/gky092
57. Rose PW, Prlic A, Altunkaya A et al (2017) The RCSB protein data bank: integrative view of protein, gene and 3D structural information. Nucleic Acids Res 45(D1):D271–D281. https://doi.org/10.1093/nar/gkw1000
58. Pedregosa F, Varoquaux G, Gramfort A et al (2011) Scikit-learn: machine learning in python. J Mach Learn Res 12:2825–2830
59. Witten IH, Frank E, Hall MA et al (2016) Data mining, fourth edition: practical machine learning tools and techniques. Morgan Kaufmann, Burlington

3.3. THERMOMUTDB: A THERMODYNAMIC DATABASE FOR MISSENSE MUTATIONS

Esta seção apresenta o desenvolvimento do ThermoMutDB, uma base de dados termodinâmicos para mutações missense. ThermoMutDB é uma base de dados manualmente curada e com medidas padronizadas. O desenvolvimento envolveu quatro etapas: (1) A dupla checagem da base de dados Protherm [18], que, no momento do desenvolvimento do trabalho era a base de dados referência em dados termodinâmicos para mutações missense, porém continha erros e não vinha sendo atualizada por vários anos; (2) Inserção de eventuais dados contidos nas referências do Protherm e que não foram inseridos; (3) coleta de novos dados; e (4) desenvolvimento do servidor web.

Esta seção é apresentada em formato de artigo, com informações adicionais à publicação nos anexos A e B. O artigo apresentado [65] foi publicado na edição de Banco de Dados da revista *Nucleic Acids Research* (fator de impacto 19,6) em Janeiro de 2021. Alguns desdobramentos pós-publicação e também análise da utilização da ferramenta pela comunidade são apresentados no capítulo de Discussão.

Published online 23 October 2020 *Nucleic Acids Research, 2021, Vol. 49, Database issue* ***D475–D479***
doi: 10.1093/nar/gkaa925

ThermoMutDB: a thermodynamic database for missense mutations

Joicymara S. Xavier[1,2], Thanh-Binh Nguyen[3], Malancha Karmarkar[3,4], Stephanie Portelli[3,4], Pâmela M. Rezende[2], João P.L. Velloso[2], David B. Ascher[3,4,5,*] and Douglas E.V. Pires[3,4,6,*]

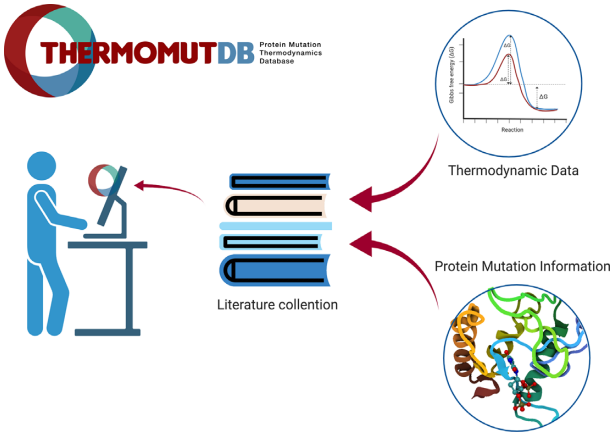
[1]Institute of Agricultural Sciences, Universidade Federal dos Vales do Jequitinhonha e Mucuri, [2]Instituto René Rachou, Fundação Oswaldo Cruz, [3]Bio 21 Institute, University of Melbourne, [4]Computational Biology and Clinical Informatics, Baker Heart and Diabetes Institute, [5]Department of Biochemistry, University of Cambridge and [6]School of Computing and Information Systems, University of Melbourne

Received August 15, 2020; Revised September 21, 2020; Editorial Decision October 05, 2020; Accepted October 12, 2020

ABSTRACT

Proteins are intricate, dynamic structures, and small changes in their amino acid sequences can lead to large effects on their folding, stability and dynamics. To facilitate the further development and evaluation of methods to predict these changes, we have developed ThermoMutDB, a manually curated database containing >14,669 experimental data of thermodynamic parameters for wild type and mutant proteins. This represents an increase of 83% in unique mutations over previous databases and includes thermodynamic information on 204 new proteins. During manual curation we have also corrected annotation errors in previously curated entries. Associated with each entry, we have included information on the unfolding Gibbs free energy and melting temperature change, and have associated entries with available experimental structural information. ThermoMutDB supports users to contribute to new data points and programmatic access to the database via a RESTful API. ThermoMutDB is freely available at: http://biosig.unimelb.edu.au/thermomutdb.

GRAPHICAL ABSTRACT



INTRODUCTION

Protein thermodynamic stability is a fundamental property of proteins that significantly influences their structure, function, expression, and solubility. Changes in protein stability have been shown to be a main driving molecular mechanism of genetic diseases (1–8) and even drug resistance (9–18). Small changes in the protein sequence can have significant consequences on their intricate structures, reflected in changes in their stability and ability to correctly fold (19). This is often a significant consideration whenever considering a new mutation, whether in the context of protein engineering or variant characterisation (20,21).

The accurate prediction of the effects of mutations on protein stability remains a complex and challenging problem. The development of computational approaches to tackle this have required large mutational datasets, however in turn have been limited by the quantity and quality of data available.

*To whom correspondence should be addressed. Tel: +61 3 8344 8185; Email: douglas.pires@unimelb.edu.au
Correspondence may also be addressed to David B. Ascher. Email: david.ascher@unimelb.edu.au

One of the first databases to collect information on the effects of mutations on protein stability, ProTherm (22), led to the exploration and rapid development of new computational approaches (23–28). However, this database has not been updated for 7 years and many errors have been identified previously (29,30), limiting both previous methods and future developments.

To overcome this, we have developed a new comprehensive and user-friendly resource for thermodynamic data from protein mutations, ThemoMutDB. Figure 1 depicts the database development workflow, which is divided into three main stages: (i) data acquisition and curation, (ii) mutation annotation and (iii) web-server development. By using a rigorous and careful data curation approach, ThemoMutDB represents a significant improvement in both the quantity and quality of data. This will not only enable the development of a new generation of methods but also an unbiased assessment of previously proposed ones.

MATERIALS AND METHODS

Data acquisition and curation

Data acquisition for ThermoMutDB was divided into two steps: manual checking of previously mined data available in other resources (Figure 1A) and manual literature curation of new thermodynamic data (Figure 1B). Within ThermoMutDB we captured thermodynamic information, experimental conditions, and literature citations. We also standardized measurements and calculations across the data entries, including temperature in Kelvin, energy in kcal/mol, and Gibbs free energy ($\Delta\Delta G$) as in the formula:

$$\Delta\Delta G = \Delta G\,(\text{wild-type}) - \Delta G\,(\text{mutant})$$

where negative $\Delta\Delta G$ values indicate that the mutation has destabilized the protein and positive $\Delta\Delta G$ values that the mutant protein is more stable.

On the first data acquisition stage, all 1,902 references in ProTherm were manually checked and validated. References that did not contain data about missense mutations were removed, leaving 829 papers. During this process, errors in data fields were corrected, duplicate entries were removed, and 329 new data-points not previously captured, but present in the original papers, were included.

New data were identified through manual literature curation. Optimized search terms (Supplementary Figure S1) were used to identify an initial pool of over 34,000 manuscripts available on PubMed. These were further narrowed down to those that contained experimental thermodynamic results for missense mutations. In total, 393 papers were analyzed and 5,654 new data points obtained, which were confirmed by at least two independent curators. Supplementary Figure S2 shows the distribution of unique mutations collected per year.

Mutation annotation

Collected mutations were mapped to protein structures available at the Protein Data Bank using (31). Different characteristics of the wild-type residue environment were calculated, including secondary structure, torsional angles, relative solvent accessibility (32) and residue depth (33). Additional residue-level information used to annotate the mutations included different substitution matrix scores. Mutation annotations were calculated using the Biopython (34). Mutation effects are also depicted via pharmacophore modeling (23). Pharmacophore modeling has been introduced in the context of mutation analysis in a previous work (23) to characterise the effect of mutations based on the differences in atom counts per pharmacophore type. Mutations that do not map to any available experimental structures are still listed but without any structure-based features calculated.

Database and web interface implementation

The database architecture was developed using SQLAlchemy, a database toolkit for Python (version 2.7). All data is stored in an SQLite database and available to download at http://biosig.unimelb.edu.au/thermomutdb/downloads. The backend system was developed using the Flask Python module (version 1.0.2) and the REStful API uses RestX extension for Flask (version 0.2.0). The web interface was implemented using the Bootstrap (version 4) framework. It also uses HTML5, CSS, JavaScript, and JQuery. JINJA2 templating language for Python was used to dynamically generate HTML templates.

RESULTS

Web interface and usage

ThermoMutDB contains information of the protein, mutational information, experimental methods and conditions, thermodynamic parameters, derived data, and literature information (details are available in Supplementary Table S1 and Figure S2). The database provides a user-friendly web interface that contains five modules: *Explore and Browse*, *Contribute*, *Downloads*, *API* and a detailed tutorial.

Explore and browse. In order to access the data, a search can be performed. This can be done either by selecting the 'Browse' page from the navigation bar or by writing the desired words on search input available on the 'Home' page. In both cases, users can use different filter combinations (Figure 2A), include or exclude columns, and download selected results in several formats (JSON, XML, CSV, TXT, SQL, MS-Excel and PDF).

The search results are shown in an interactive table, with columns providing experimental information recovered from literature and also derived properties (Figure 2B). Aiming to improve user experience, it is possible to visualize a summary for each entry by clicking on the '+' icon. This option can lead to a 'Details' page that shows all information about the mutation and provides related files to download (Supplementary Figure S3).

User contributions. To facilitate a continuous database update, we have implemented a user's contribution section (Supplementary Figure S4), which allows the scientific community to share new data or identify potential errors that will be manually checked by our team. To submit contributions it is just required to fill the form with mutation and

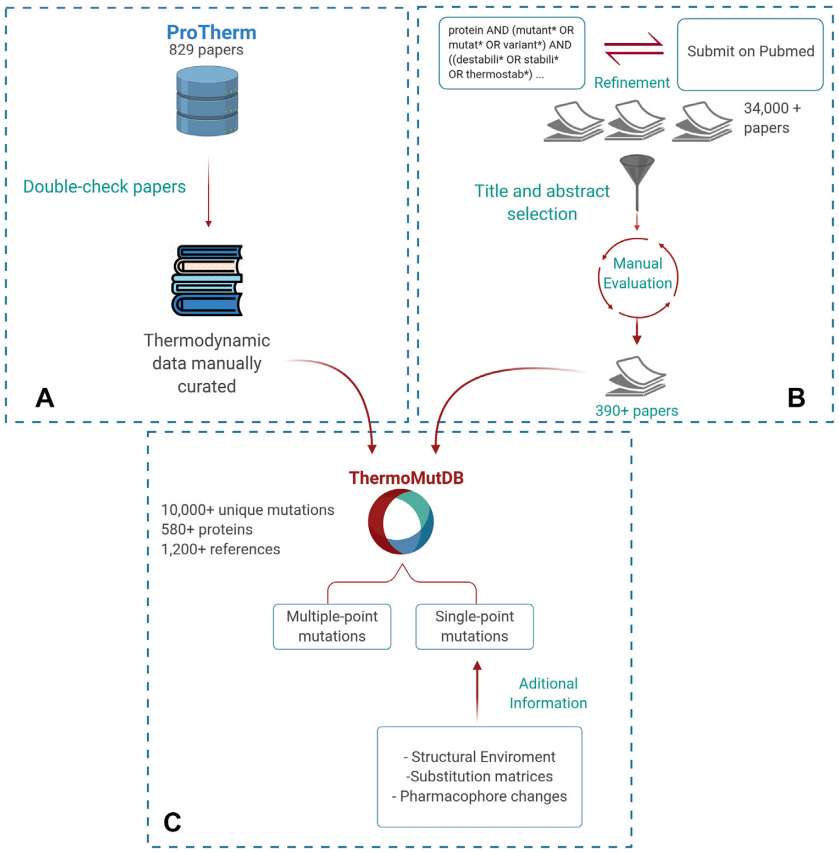


Figure 1. ThermoMutDB workflow for data acquisition and processing. The development workflow is divided into three steps: (**A**) verification of previously available mutation thermodynamics information (**B**) collection and manual curation of new data and (**C**) data aggregation and mutation annotation.

thermodynamics data, to inform a contact email and a reference (paper published, accepted, or pre-print). Although significant effort has been devoted to ensure high quality data curation, users have the option to report any issues with the data to our team. These are important efforts to further expand and improve the database.

Downloads. All data in the database can be downloaded from the 'Download' page in CSV or JSON formats. It is also possible to download the protein structure files related to data available.

Programmatic access via an API. ThermoMutDB supports programmatic access via a RESTful API to allow other services to harness our data easily. The 'API' page provides documentation of all endpoints available and allows users to execute queries using provided fields. Other queries can be performed by passing parameters through the URL (Supplementary Figure S5).

Data statistics

Examining the distribution of mutations in the ThermoMutDB reveals a number of natural biases that need to be taken into consideration when developing, or evaluating, new predictive tools. ThermoMutDB contains thermodynamic information on 14,669 mutations across 588 proteins. This represents a significant increase over ProTherm, with a 83% increase in unique mutations and over 300 new proteins. Supplementary Figure S6 shows the distribution of unique mutations collected per year. The majority of these are single-point mutations (82.8%), with mutations to alanine being over-represented (Figure 3D). This becomes evident when we look at the distribution of wild-type and mutant amino acid residues within the database (Supplementary Figure S7). The most frequent mutations were from Leucine and Valine to Alanine, while 10 mutations were not present in the dataset, including W→G, W→P and C→K among others, which seem to denote large changes in residue physicochemical properties.

As would be expected by chance, two thirds of mutations within the database are destabilising (Supplementary Figure S8). This natural bias creates an extra challenge for computational methods built using this information, in particular those based on machine learning approaches, regarding the prediction of stabilising mutations, which are less well represented. It is important to note, however, that the data on ThermoMutDB represents an increase of over 100% in stabilising mutations in comparison with previous resources. No apparent correlation was identified between the mutation effects and their location within protein structures, with mutations leading to increased and decreased stability similarly distributed across protein structures when looking at residue depth (Supplementary Figure S9). Muta-

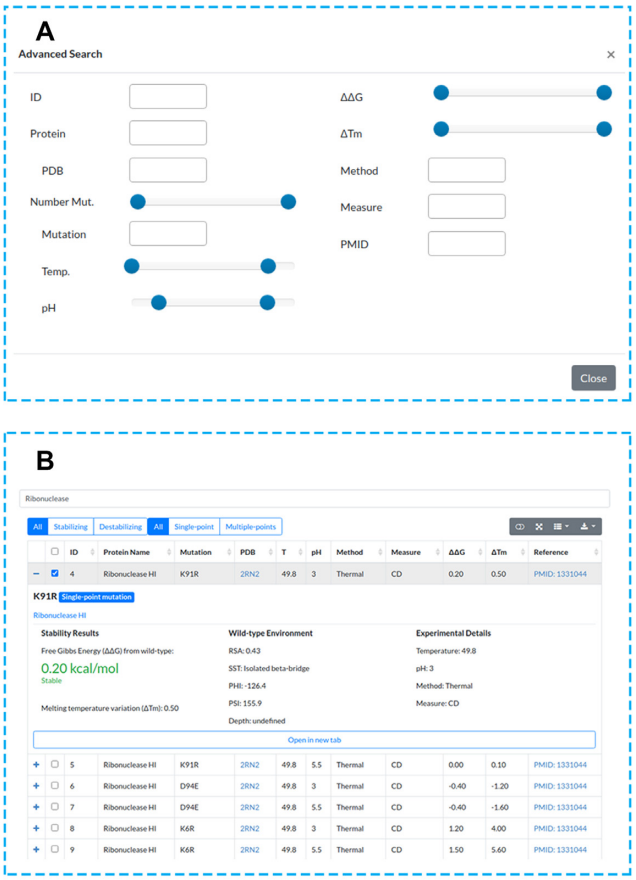


Figure 2. ThermoMutDB web interface search and results pages. (**A**) ThermoMutDB offers 12 query modes, with detailed information available about each query type through the 'Help' page at the top navigation bar and through on-page help in the form of question mark tooltips. (**B**) The general layout of the result page, showing a summary of information for each entry as well as detailed view.

Figure 3. Composition of ThermoMutDB entries. (**A**) depicts the distribution of phylogenetic kingdoms of proteins in the database. (**B**) highlights the distribution of thermodynamic effects of mutation in the database, given as the variation in Gibbs Free Energy ($\Delta\Delta G$). (**C**) Experimental variability of mutation assessed under different conditions and groups. (**D**) Distribution of mutations in ThermoMutDB based on type (mutation to alanine/non-alanine), their location and residue environment.

tions in ThermoMutDB are spread across different protein classes (Supplementary Figure S10) and diverse in terms of secondary structure (Supplementary Figure S11).

Within ThermoMutDB, we identified mutations that had been experimentally measured at least twice and, by comparing the variance between these replicate results (Figure 3C), we identified a Pearson's correlation of 0.9. This provides a measure of the intrinsic noise in the data, and suggests a theoretical maximum performance that should be expected for predictive stability tools built using this data.

DISCUSSION

ThermoMutDB represents a significant increase in availability, reliability and diversity of thermodynamics data linking effects of mutations to protein stability. We believe this resource will have a significant impact on understanding the effects of mutations on protein structure and stability. It will enable experimental scientists to identify previously characterised mutations in proteins of interest, and provide computational scientists with a comprehensive and refined set of experimental data to query the relationship between changes in protein sequence and stability, facilitating the development of new computational tools to analyse these relationships and develop prediction algorithms.

New mutation thermodynamics data collected and compiled in ThermoMutDB will also allow for more robust, comprehensive and independent validation of currently available computational predictors. The database will be continuously maintained and updated, enabling submission of user contributions and data access through an intuitive web-based interface (http://biosig.unimelb.edu.au/thermomutdb) as well as programmatic access through an API.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

D.B.A. and D.E.V.P. were funded by a Newton Fund RCUK-CONFAP Grant awarded by the Medical Research Council (MRC) and Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG) [MR/M026302/1]; Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq); Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES); Jack Brockhoff Foundation [JBF 4186, 2016]; Wellcome Trust [200814/Z/16/Z]; Investigator Grant from the National Health and Medical Research Council (NHMRC) of Australia [GNT1174405]. Funding for open access charge: Wellcome Trust.
Conflict of interest statement. None declared.

REFERENCES

1. Jafri,M., Wake,N.C., Ascher,D.B., Pires,D.E., Gentle,D., Morris,M.R., Rattenberry,E., Simpson,M.A., Trembath,R.C., Weber,A. *et al.* (2015) Germline mutations in the CDKN2B tumor suppressor gene predispose to renal cell carcinoma. *Cancer Discov.*, **5**, 723–729.
2. Ramdzan,Y.M., Trubetskov,M.M., Ormsby,A.R., Newcombe,E.A., Sui,X., Tobin,M.J., Bongiovanni,M.N., Gras,S.L., Dewson,G., Miller,J.M.L. *et al.* (2017) Huntingtin inclusions trigger cellular quiescence, deactivate apoptosis, and lead to delayed necrosis. *Cell Rep.*, **19**, 919–927.
3. Soardi,F.C., Machado-Silva,A., Linhares,N.D., Zheng,G., Qu,Q., Pena,H.B., Martins,T.M.M., Vieira,H.G.S., Pereira,N.B., Melo-Minardi,R.C. *et al.* (2017) Familial STAG2 germline mutation defines a new human cohesinopathy. *NPJ Genom Med*, **2**, 7.
4. Andrews,K.A., Ascher,D.B., Pires,D.E.V., Barnes,D.R., Vialard,L., Casey,R.T., Bradshaw,N., Adlard,J., Aylwin,S., Brennan,P. *et al.* (2018) Tumour risks and genotype-phenotype correlations associated with germline variants in succinate dehydrogenase subunit genes SDHB, SDHC and SDHD. *J. Med. Genet.*, **55**, 384–394.
5. Ascher,D.B., Spiga,O., Sekelska,M., Pires,D.E.V., Bernini,A., Tiezzi,M., Kralovicova,J., Borovska,I., Soltysova,A., Olsson,B. *et al.* (2019) Homogentisate 1,2-dioxygenase (HGD) gene variants, their analysis and genotype-phenotype correlations in the largest cohort of patients with AKU. *Eur. J. Hum. Genet.*, **27**, 888–902.
6. Hildebrand,J.M., Kauppi,M., Majewski,I.J., Liu,Z., Cox,A.J., Miyake,S., Petrie,E.J., Silk,M.A., Li,Z., Tanzer,M.C. *et al.* (2020) A missense mutation in the MLKL brace region promotes lethal neonatal inflammation and hematopoietic dysfunction. *Nat. Commun.*, **11**, 3150.
7. Pires,D.E.V., Rodrigues,C.H.M. and Ascher,D.B. (2020) mCSM-membrane: predicting the effects of mutations on transmembrane proteins. *Nucleic Acids Res.*, **48**, W147–W153.
8. Trezza,A., Bernini,A., Langella,A., Ascher,D.B., Pires,D.E.V., Sodi,A., Passerini,I., Pelo,E., Rizzo,S., Niccolai,N. *et al.* (2017) A computational approach from gene to structure analysis of the human ABCA4 transporter involved in genetic retinal diseases. *Invest. Ophthalmol. Vis. Sci.*, **58**, 5320–5328.
9. Ascher,D.B., Wielens,J., Nero,T.L., Doughty,L., Morton,C.J. and Parker,M.W. (2014) Potent hepatitis C inhibitors bind directly to NS5A and reduce its affinity for RNA. *Sci. Rep.*, **4**, 4765.
10. Phelan,J., Coll,F., McNerney,R., Ascher,D.B., Pires,D.E., Furnham,N., Coeck,N., Hill-Cawthorne,G.A., Nair,M.B., Mallard,K. *et al.* (2016) Mycobacterium tuberculosis whole genome sequencing and protein structure modelling provides insights into anti-tuberculosis drug resistance. *BMC Med.*, **14**, 31.
11. Karmakar,M., Globan,M., Fyfe,J.A.M., Stinear,T.P., Johnson,P.D.R., Holmes,N.E., Denholm,J.T. and Ascher,D.B. (2018) Analysis of a novel pncA mutation for susceptibility to pyrazinamide therapy. *Am. J. Respir. Crit. Care Med.*, **198**, 541–544.
12. Portelli,S., Phelan,J.E., Ascher,D.B., Clark,T.G. and Furnham,N. (2018) Understanding molecular consequences of putative drug resistant mutations in Mycobacterium tuberculosis. *Sci. Rep.*, **8**, 15356.
13. Karmakar,M., Rodrigues,C.H.M., Holt,K.E., Dunstan,S.J., Denholm,J. and Ascher,D.B. (2019) Empirical ways to identify novel Bedaquiline resistance mutations in AtpE. *PLoS One*, **14**, e0217169.
14. Hawkey,J., Ascher,D.B., Judd,L.M., Wick,R.R., Kostoulias,X., Cleland,H., Spelman,D.W., Padiglione,A., Peleg,A.Y. and Holt,K.E. (2018) Evolution of carbapenem resistance in Acinetobacter baumannii during a prolonged infection. *Microbial Genomics*, **4**, e000165.
15. Holt,K.E., McAdam,P., Thai,P.V.K., Thuong,N.T.T., Ha,D.T.M., Lan,N.N., Lan,N.H., Nhu,N.T.Q., Hai,H.T., Ha,V.T.N. *et al.* (2018) Frequent transmission of the Mycobacterium tuberculosis Beijing lineage and positive selection for the EsxW Beijing variant in Vietnam. *Nat. Genet.*, **50**, 849–856.
16. Vedithi,S.C., Malhotra,S., Das,M., Daniel,S., Kishore,N., George,A., Arumugam,S., Rajan,L., Ebenezer,M., Ascher,D.B. *et al.* (2018) Structural implications of mutations conferring rifampin resistance in Mycobacterium leprae. *Sci. Rep.*, **8**, 5016.
17. Karmakar,M., Rodrigues,C.H.M., Horan,K., Denholm,J.T. and Ascher,D.B. (2020) Structure guided prediction of Pyrazinamide resistance mutations in pncA. *Sci. Rep.*,**10**, 1875.
18. Portelli,S., Olshansky,M., Rodrigues,C.H.M., D'Souza,E.N., Myung,Y., Silk,M., Alavi,A., Pires,D.E.V. and Ascher,D.B. (2020) Exploring the structural distribution of genetic variation in SARS-CoV-2 with the COVID-3D online resource. *Nat. Genet.*, **52**, 999–1001.
19. Pandurangan,A.P., Ascher,D.B., Thomas,S.E. and Blundell,T.L. (2017) Genomes, structural biology and drug discovery: combating the impacts of mutations in genetic disease and antibiotic resistance. *Biochem. Soc. Trans.*, **45**, 303–311.
20. Wijma,H.J., Floor,R.J. and Janssen,D.B. (2013) Structure- and sequence-analysis inspired engineering of proteins for enhanced thermostability. *Curr. Opin. Struct. Biol.*, **23**, 588–594.
21. Pires,D.E., Chen,J., Blundell,T.L. and Ascher,D.B. (2016) In silico functional dissection of saturation mutagenesis: interpreting the relationship between phenotypes and changes in protein stability, interactions and activity. *Sci. Rep.*, **6**, 19848.
22. Gromiha,M.M., An,J., Kono,H., Oobatake,M., Uedaira,H. and Sarai,A. (1999) ProTherm: thermodynamic database for proteins and mutants. *Nucleic Acids Res.*, **27**, 286–288.
23. Pires,D.E., Ascher,D.B. and Blundell,T.L. (2014) mCSM: predicting the effects of mutations in proteins using graph-based signatures. *Bioinformatics*, **30**, 335–342.
24. Pires,D.E., Ascher,D.B. and Blundell,T.L. (2014) DUET: a server for predicting effects of mutations on protein stability using an integrated computational approach. *Nucleic Acids Res.*, **42**, W314–W319.
25. Pandurangan,A.P., Ochoa-Montano,B., Ascher,D.B. and Blundell,T.L. (2017) SDM: a server for predicting effects of mutations on protein stability. *Nucleic Acids Res.*, **45**, W229–W235.
26. Rodrigues,C.H., Pires,D.E. and Ascher,D.B. (2018) DynaMut: predicting the impact of mutations on protein conformation, flexibility and stability. *Nucleic Acids Res.*, **46**, W350–W355.
27. Laimer,J., Hiebl-Flach,J., Lengauer,D. and Lackner,P. (2016) MAESTROweb: a web server for structure-based protein stability prediction. *Bioinformatics*, **32**, 1414–1416.
28. Quan,L., Lv,Q. and Zhang,Y. (2016) STRUM: structure-based prediction of protein stability changes upon single-point mutation. *Bioinformatics*, **32**, 2936–2946.
29. Yang,Y., Urolagin,S., Niroula,A., Ding,X., Shen,B. and Vihinen,M. (2018) PON-tstab: protein variant stability predictor. Importance of training data quality. *Int. J. Mol. Sci.*, **19**, 1009.
30. Fang,J. (2020) A critical review of five machine learning-based algorithms for predicting protein stability changes upon mutation. *Brief. Bioinform.*, **21**, 1285–1292.
31. Martin,A.C. (2005) Mapping PDB chains to UniProtKB entries. *Bioinformatics*, **21**, 4297–4301.
32. Gilis,D. and Rooman,M. (1997) Predicting protein stability changes upon mutation using database-derived potentials: solvent accessibility determines the importance of local versus non-local interactions along the sequence. *J. Mol. Biol.*, **272**, 276–290.
33. Chakravarty,S. and Varadarajan,R. (1999) Residue depth: a novel parameter for the analysis of protein structure and stability. *Structure*, **7**, 723–732.
34. Cock,P.J., Antao,T., Chang,J.T., Chapman,B.A., Cox,C.J., Dalke,A., Friedberg,I., Hamelryck,T., Kauff,F., Wilczynski,B. *et al.* (2009) Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, **25**, 1422–1423.

ThermoMutDB: a thermodynamic database for missense mutations

Joicymara S. Xavier[1,2], Thanh-Binh Nguyen[3], Malancha Karmarkar[3,4],Stephanie Portelli [3,4], Pâmela M. Rezende[2], João P. L. Velloso[2], David B. Ascher[3,4,5*], Douglas E. V. Pires[3,4,6*]

[1]Institute of Agricultural Sciences, Universidade Federal dos Vales do Jequitinhonha e Mucuri;
[2]Instituto René Rachou, Fundação Oswaldo Cruz;
[3]Computational Biology and Clinical Informatics, Baker Heart and Diabetes Institute;
[4]Bio 21 Institute, University of Melbourne;
[5]Department of Biochemistry, University of Cambridge;
[6]School of Computing and Information Systems, University of Melbourne

*To whom correspondence should be addressed. Tel: +61 3 8344 8185; Email: douglas.pires@unimelb.edu.au. Correspondence may also be addressed to David B. Ascher. Email: david.ascher@unimelb.edu.au.

protein AND (mutant* OR mutat* OR variant*) AND ((destabili* OR stabili* OR thermostab*) OR (kcal/mol OR kj/mol OR kcal mol OR kj mol) OR (“free energy” OR “gibb* free energy" OR "melting temperature")) NOT(review[Publication Type] OR "molecular dynamics" OR predict*)

Figure S1: Search query used to identify and collect mutation thermodynamics data from publications available on PubMed. This query was designed to encompass mutations in proteins and their experimentally measured effects while excluding works involving their silico characterization.

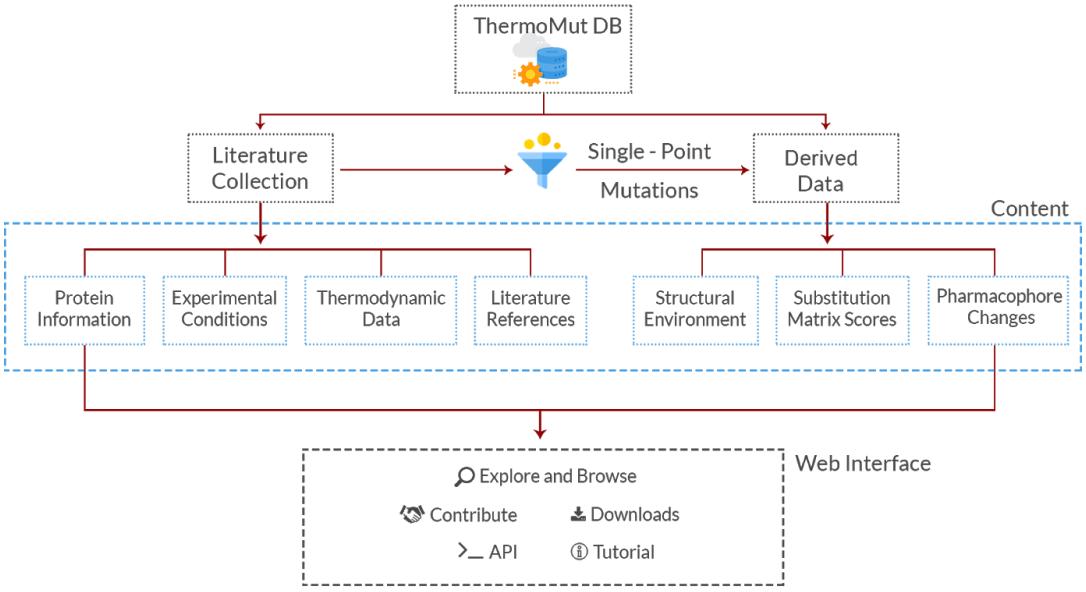


Figure S2: Schematic workflow of ThermoMutDB data organization.

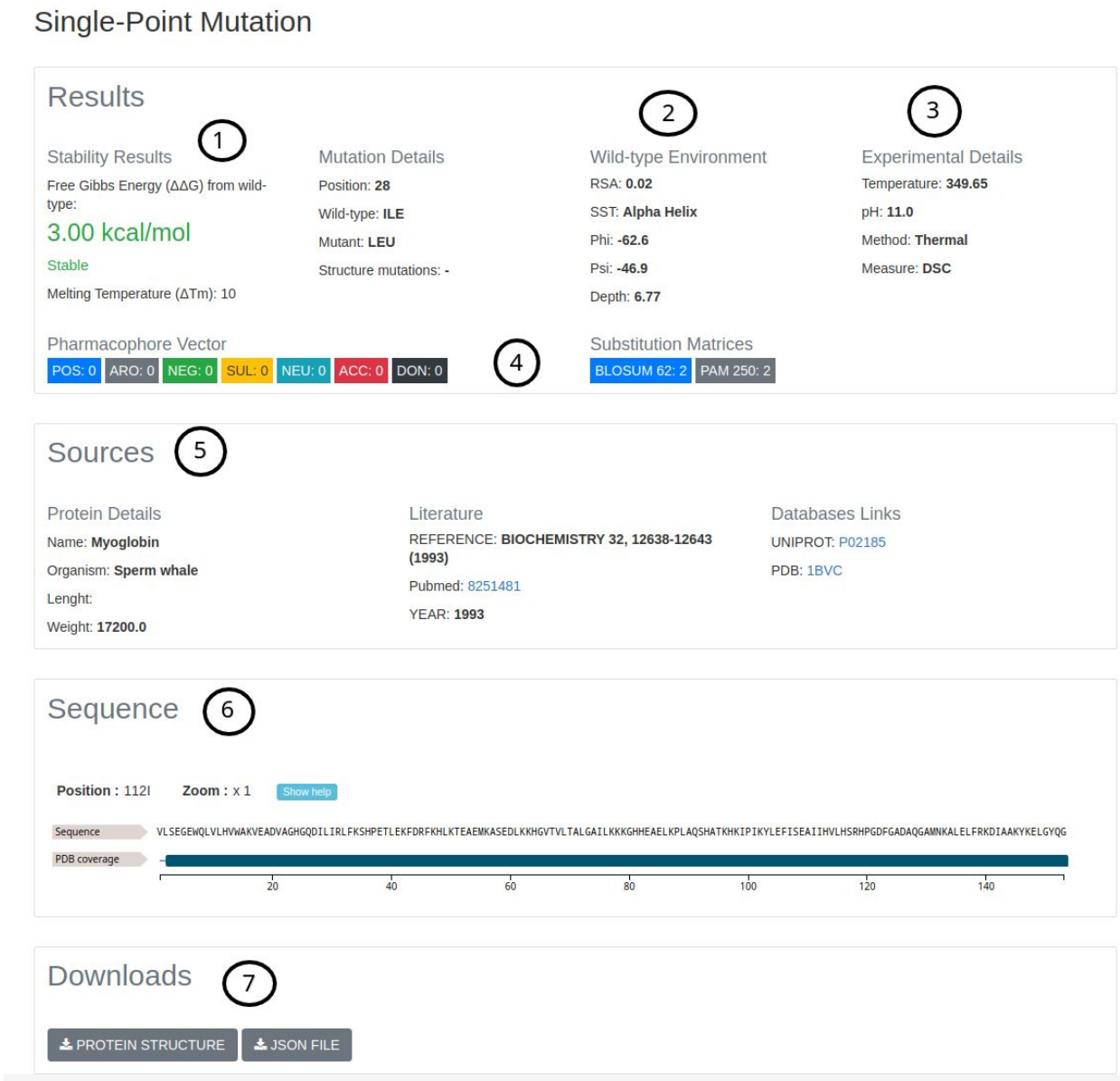


Figure S3: ThermoMutDB mutation details page. The mutation page provides information about the mutation and its thermodynamic effects (1), residue environment properties (2), experimental conditions (3), as well as substitution matrix scores for the mutation and their effects in terms of pharmacophore changes (4). In this page, the literature information is provided, with links to external databases (5), followed by the alignment of PDB sequence against the Uniprot sequence (6), as well as the option to download the entry and protein structure (7).

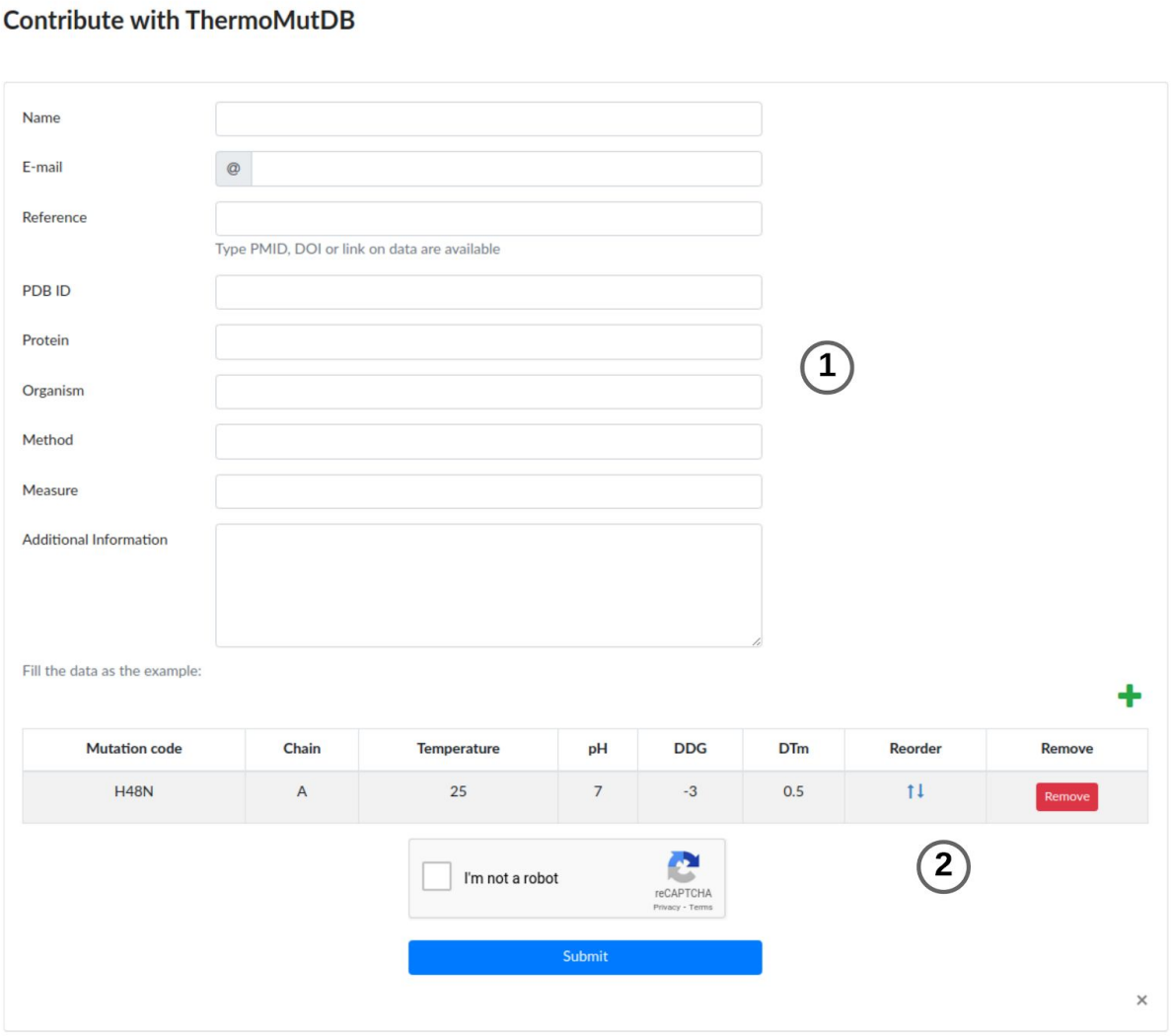


Figure S4: User contribution page. ThermoMutDB allows users to submit their own contributions to be manually curated and incorporated into the database. Users are requested to provide information on the publication, protein and experimental protocol (1) and can include thermodynamic information for one or more mutations (2).

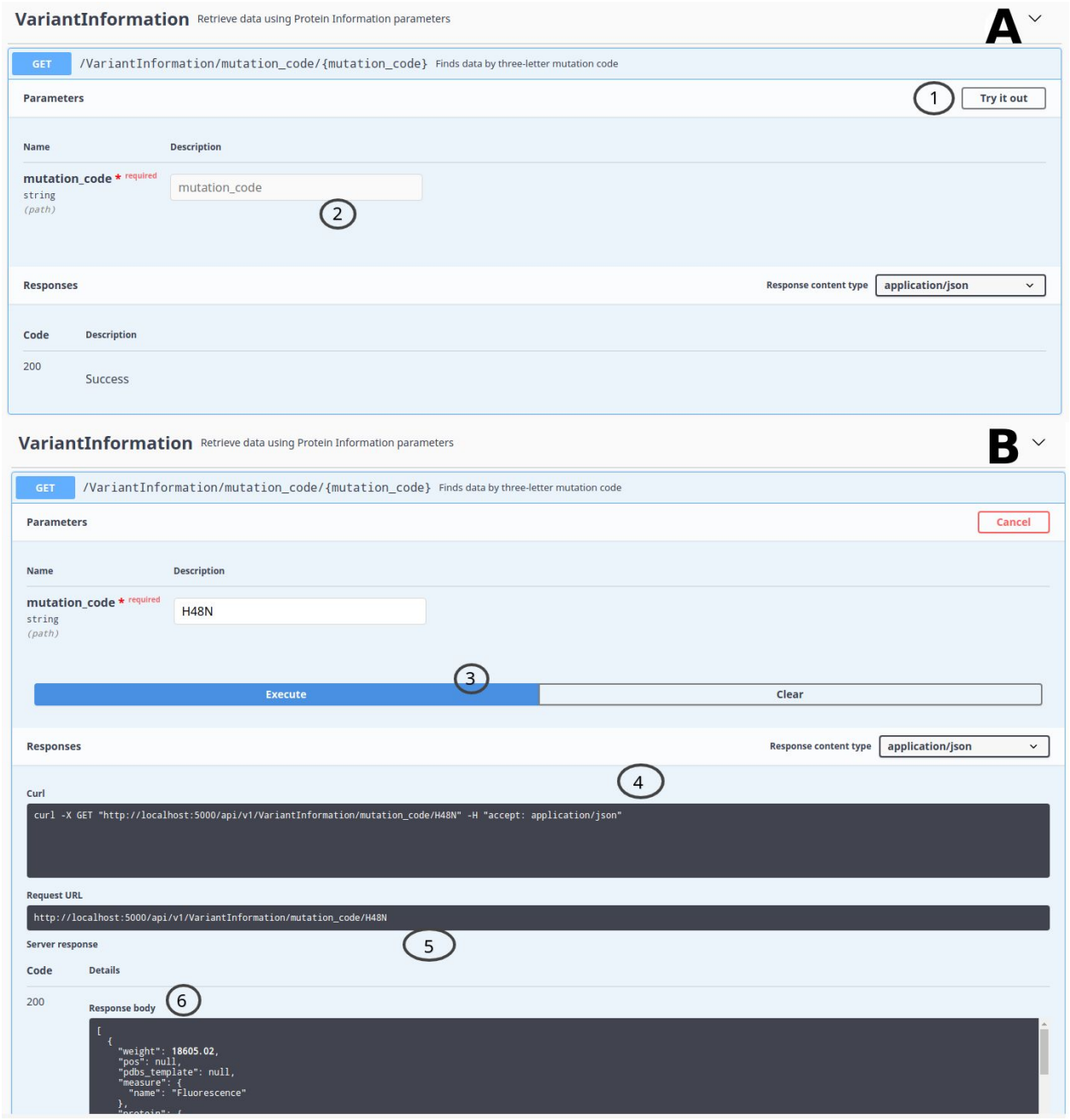


Figure S5: Programmatic Access Via an API. The figure depicts advanced search options for ThermoMutDB and respective URL parameters. (1) To start, click on “Try it out”, (2) type the desired parameters, and (3) click on “Execute” button. The response shows (4) Curl command, (5) URL request, and (6) the response body.

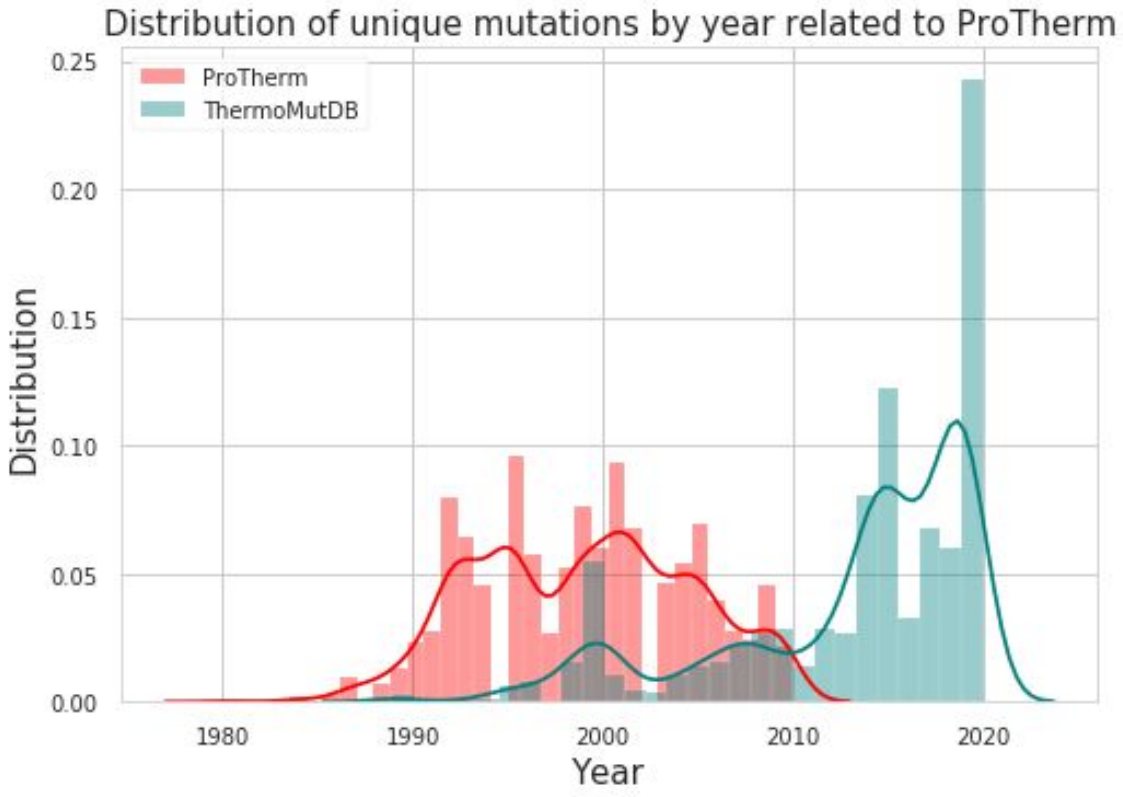


Figure S6: Distribution of entries on ThermoMutDB over the years, highlighted by origin.

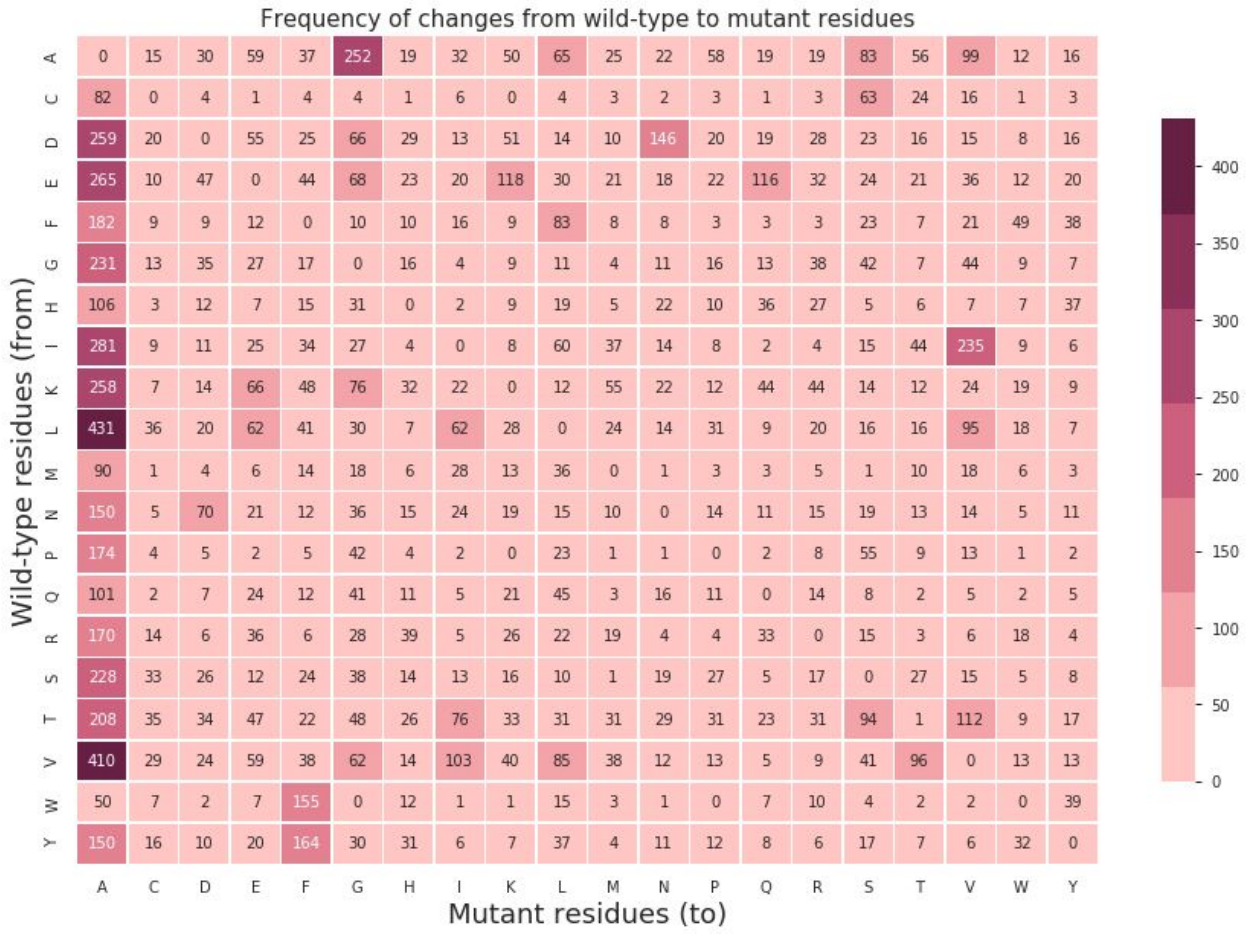


Figure S7: Frequency of changes from wild-type to mutant residues within the ThermoMutDB. The majority of mutations are to alanine, characterising alanine-scanning experiments.

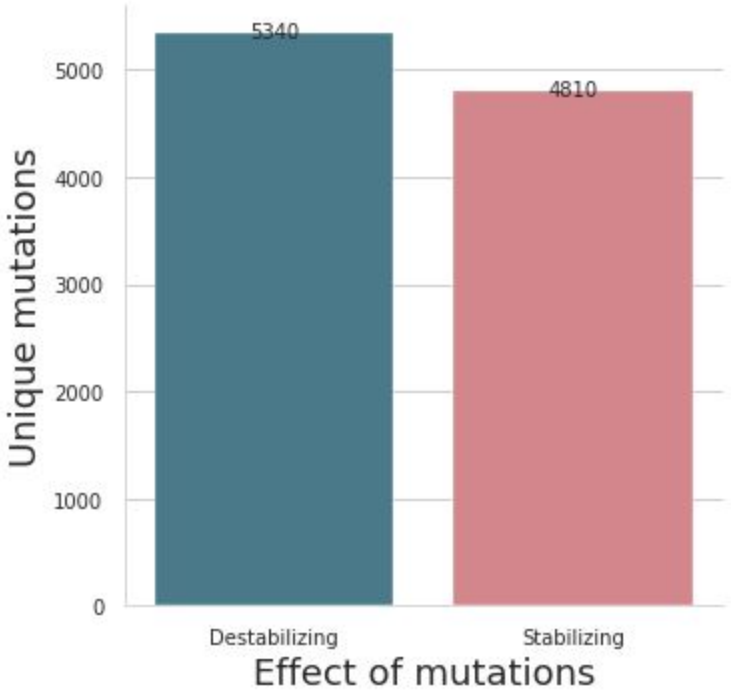


Figure S8: ThermoMutDB entries divided based on mutation effects.

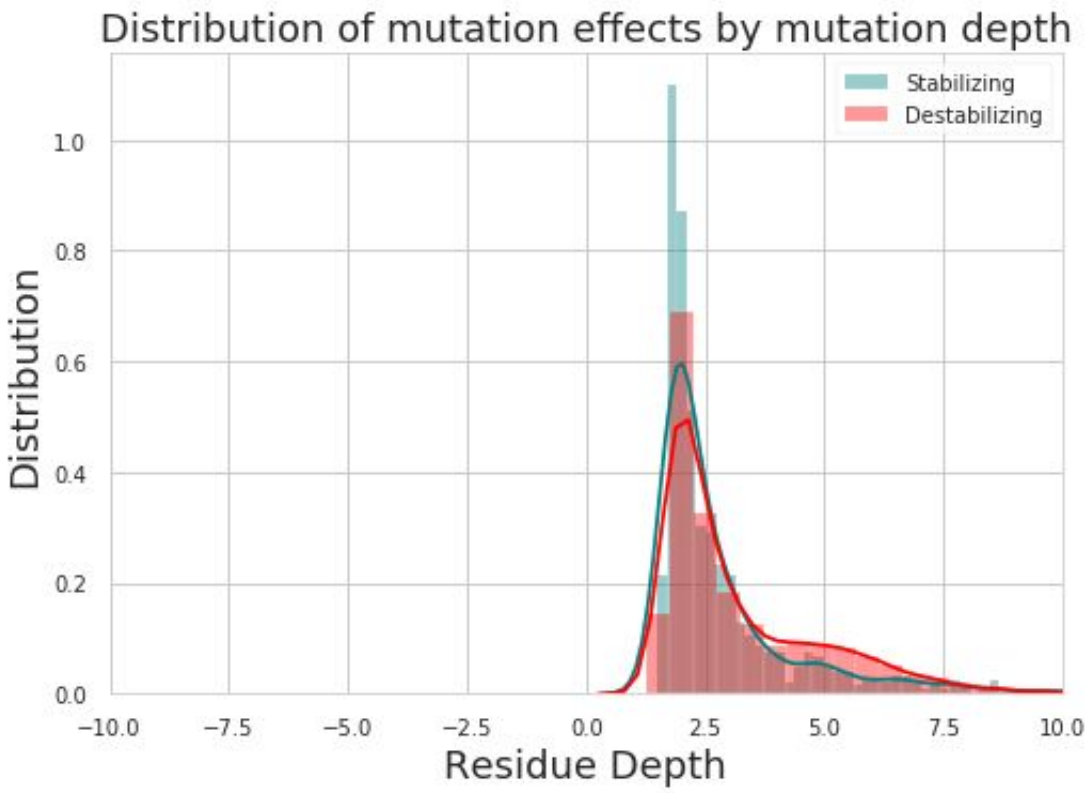


Figure S9: Distribution of residue depth for stabilising and destabilising mutations in ThermoMutDB.

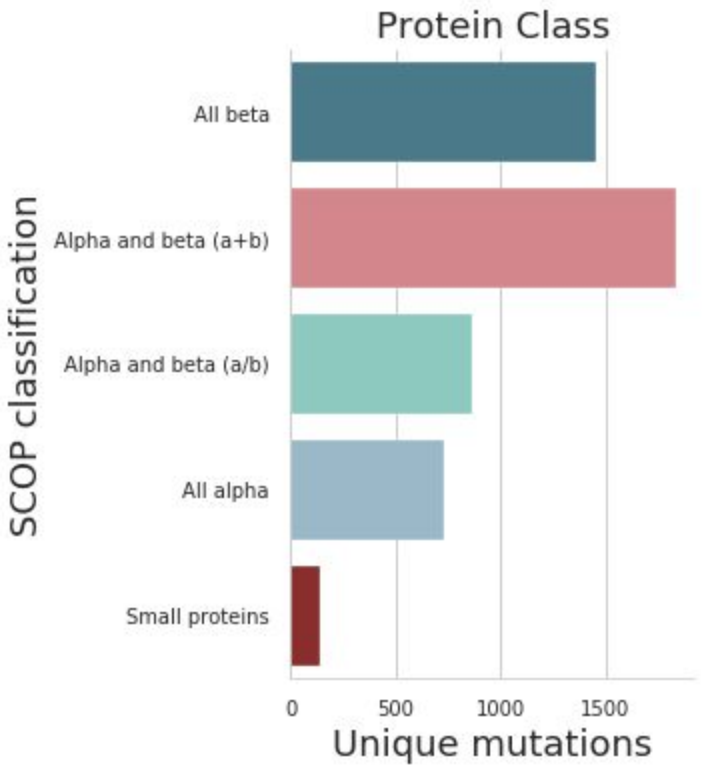


Figure S10: Distribution of mutation on ThermoMutDB based on protein classification according to SCOP.

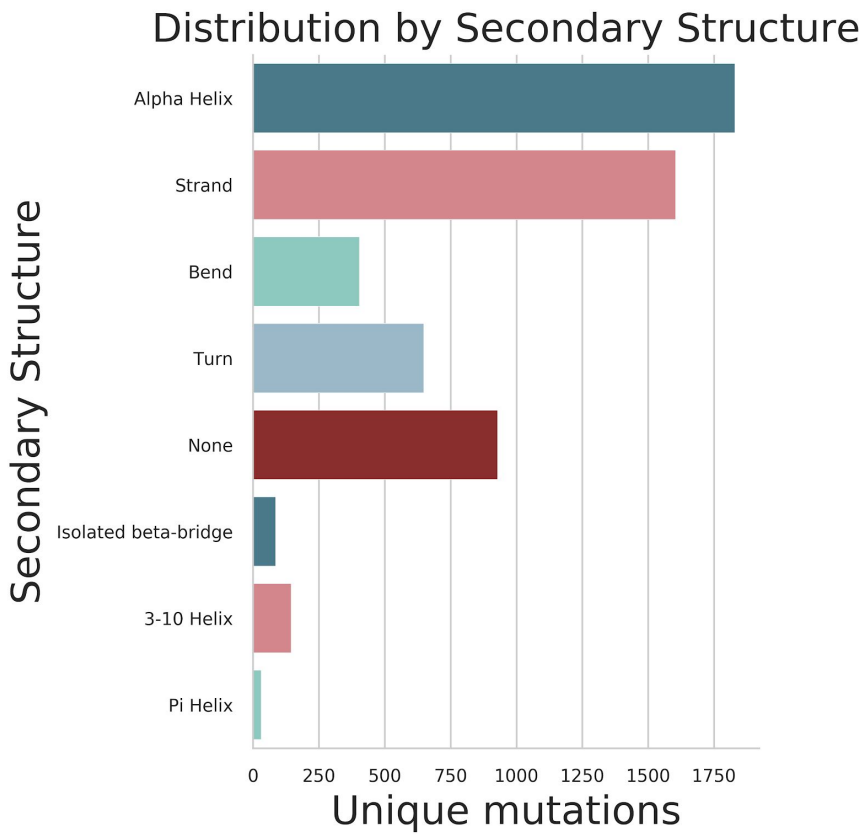


Figure S11: Distribution of mutation on ThermoMutDB based on secondary structure.

Table S1: Information content of ThermoMutDB

Content	Description	Measure (if there is)	single	multiple
Protein Information				
Protein name	Protein Name		X	X
Source	Protein Organism		X	X
Uniprot	Uniprot Code		X	X
PDB wild	Protein Data Bank code		X	X
PDB mutant	PDB code for mutant (when is available)		X	X
PDBs template	PDBs used as a template to model wild_type (when is available and necessary)		X	X
Length	Length of sequence		X	X
Weight	Molecular weight		X	X
PIR ID	Protein Information Resource		X	X
SWISSPROT ID	Code of Swiss-Prot (Revised Entries)		X	X
Mutation	Three-digits mutation code		X	X
Mutated chain	Mutated chain		X	X
Structure mutations	Number of mutations in the structure		X	X
Structure coverage	Structure coverage		X	X
Mutation count	Number of mutations		X	X
Experimental Conditions				
Temperature	Experimental temperature	Kelvin (K)	X	X
pH	Experimental pH		X	X
Measure	Experimental techniques for studying protein folding.		X	X
Method	Techniques to denature a protein		X	X

Thermodynamic data				
ΔΔG	Variation of Free Gibbs Energy on the experiment	$kcal/mol^{-1}$	X	X
ΔTm	Variation of Melting Temperature on the experiment	Kelvin (K)	X	X
Structural environment				
SST	Secondary Structure classification		X	
RSA	Relative accessible surface area		X	
PHI	Phi angle value		X	
PSI	Psi angle value		X	
Residue Depth	The average distance of atoms of wild-type residue from the solvent accessible area		X	
CA Depth	The average distance of atoms of CA from the solvent accessible area		X	
Relative B Factor	Temperature factor		X	
Substitution matrices scores				
Blosum 62	BLOSUM 62 matrix score		X	
Pam 250	PAM 250 matrix score		X	
Pharmacophore changes				
POS	Positive		X	
NEG	Negative		X	
ACC	Hydrogen bond acceptors		X	
DON	Hydrogen bond donors		X	
ARO	Aromatic rings		X	
SUL	Sulfuric acid		X	
NEU	Neutral		X	
Literature information				
Reference	Publication reference		X	X
PMID	Pubmed code of publication		X	X

DOI	Digital Object Identifier of publication		X	X
YEAR	Year of publication		X	X

4. VIGILÂNCIA GENÔMICA NA PANDEMIA DE COVID-19

A doença do coronavírus 2019 (COVID-19) é a expressão clínica da nova infecção pela síndrome respiratória aguda grave do coronavírus 2 (SARS-CoV-2) que teve início no ano de 2019 em Wuhan, na China [28]. No ano seguinte, em Março de 2020, a Organização Mundial da Saúde (OMS) classificou a doença como Emergência de Saúde Pública de Importância Internacional e posteriormente como Pandemia. Muitas preocupações então começaram a surgir, principalmente devido à rápida disseminação do vírus e à incapacidade da maioria dos hospitais atenderem à alta demanda por hospitalizações [66–68].

O SARS-CoV-2 pertence a uma grande família de vírus, a Coronavírus. Os coronavírus são vírus de RNA de cadeia simples positiva (+ssRNA) distribuídos largamente entre humanos, outros mamíferos e pássaros, causando doenças respiratórias, entéricas e neurológicas [69]. Assim como grande parte das doenças virais, a COVID-19 pode ser contraída principalmente através de gotículas de saliva transmitidas pelo ar, através do contato direto com uma pessoa infectada, ou ainda, indiretamente, através de superfícies contaminadas com o vírus. Devido à sua alta capacidade de transmissão, infecção e adaptação, medidas para contenção do vírus passaram a ser recomendadas pela OMS e implementadas no mundo todo. Para além das medidas de restrição e isolamento, a vigilância genômica tem sido consistentemente utilizada para monitorar a evolução e transmissão do vírus e também identificar novas variantes de preocupação que frequentemente emergem, em razão da sua alta capacidade de mutação [70].

A vigilância genômica é o processo de monitoramento constante de patógenos e análise de suas semelhanças e diferenças genéticas. Ela ajuda pesquisadores, epidemiologistas e autoridades de saúde pública a monitorar a evolução de agentes de doenças infecciosas, alertar sobre a disseminação de patógenos, além de adaptar intervenções e recomendações para o público. Além disso, a vigilância genômica é essencial para se desenvolver e adaptar contramedidas como vacinas para mitigar ou acabar com a propagação de doenças [70].

Durante a pandemia, o Comitê de Emergência do Regulamento Sanitário Internacional para COVID-19 recomendou repetidamente que os estados fortalecessem as estratégias de vigilância genômica e também que fornecessem financiamento regular para bens globais específicos, incluindo sequenciamento genômico [71,72]. A vigilância genômica combina dados genômicos e epidemiológicos com ferramentas de bioinformática, gerando informações essenciais para a compreensão da origem e futuro do vírus. Um sistema contínuo

e estruturado de genômica viral, epidemiologia e bioinformática, integrado com dados de vigilância, pode fornecer dados valiosos para a busca de respostas adequadas à emergência e reemergência de vírus [73].

A vigilância eficaz requer a coleta de dados de sequência suficientes de populações representativas para detectar novas variantes e monitorar as tendências nas variantes circulantes. Todos os vírus sofrem mutação à medida que se replicam e se espalham em uma população, dessa forma, toda vez que o SARS-CoV-2 se replica, há uma oportunidade para modificação do vírus. Quando uma dessas mudanças afeta a capacidade do vírus de se espalhar ou causar doenças, ele pode adquirir uma vantagem competitiva sobre as outras linhagens de SARS-CoV-2. Com o tempo, certas linhagens com essas vantagens tornam-se mais prevalentes e circulam em uma população. Quando uma linhagem ou grupo de linhagens possui características que impactam a saúde pública, o Centro para Controle de Doenças e Prevenção (CDC) pode classificá-las como uma Variante de Interesse (VOI) ou Variante de Preocupação, do inglês, *Variant of Concern* (VOC) [74].

4.1. VIGILÂNCIA GENÔMICA NO CONTINENTE AFRICANO

Embora o continente Africano conte com uma pequena proporção global de casos e mortes reportados, alguns países africanos têm desempenhado um papel fundamental na resposta à pandemia através dos seus esforços de sequenciamento genômico. Duas das cinco VOCs de importância global, Beta e Omicron, por exemplo, foram identificadas na África, além de duas subvariantes da Omicron (BA.4 e BA.5). Essas contribuições foram possíveis graças ao dedicado sistema de vigilância genômica, sequenciamento em tempo real e esforços para liberação dos dados que vêm sendo fomentados no continente [75–77].

O artigo *“The evolving SARS-CoV-2 epidemic in Africa: Insights from rapidly expanding genomic surveillance”* (Apêndice A) discute como alguns países da África aumentaram a capacidade de vigilância genômica no continente e apresenta os resultados obtidos depois de mais de dois anos de esforços que foram capazes de sequenciar mais de 100.000 genomas de SARS-CoV-2.

4.2. SARS-COV-2 AFRICA DASHBOARD: AN INTERACTIVE TOOL FOR VISUALIZING COVID-19 GENOMICS DATA

Esta seção apresenta o desenvolvimento do SARS-CoV-2 Africa Dashboard, uma ferramenta interativa de visualização de dados genômicos para o continente africano em tempo real. A principal motivação do desenvolvimento da aplicação foi permitir que tomadores de decisão e a população em geral tenham acesso às informações provindas dos dados genômicos de COVID-19 produzidos pelo continente africano, um dos líderes no combate à COVID através de vigilância genômica no sul global [78].

Esta seção é apresentada em formato de artigo com seu respectivo material suplementar. O artigo apresentado foi publicado na Nature Microbiology (fator de impacto 30,964) em outubro de 2022. Este trabalho também ganhou segundo lugar entre os melhores posters do módulo NGS no 26º *Bioinformatics Workshop on Virus Evolution and Molecular Epidemiology* (VEME) (Anexo D).

https://doi.org/10.1038/s41564-022-01276-9

SARS-CoV-2 Africa dashboard for real-time COVID-19 information

Joicymara S. Xavier, Monika Moir, Houriiyah Tegally, Nikita Sitharam, Wasim Abdool Karim, James E. San, Joana Linhares, Eduan Wilkinson, David B. Ascher, Cheryl Baxter, Douglas E. V. Pires & Tulio de Oliveira

The SARS-CoV-2 Africa dashboard is an interactive tool that enables visualization of SARS-CoV-2 genomic information in African countries. The customizable app allows users to visualize the number of sequences deposited in each country, and the variants circulating over time. Our dashboard enables near real-time exploration of public data that can inform policymakers, healthcare professionals and the public about the ongoing pandemic.

COVID-19 is the clinical manifestation of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) infection[1]. The COVID-19 pandemic has been ongoing for more than two years, with the first case of COVID-19 in Africa reported in Egypt in mid-February 2020 (ref. [2]). The first SARS-CoV-2 genome sequenced in Africa was reported in March 2020 (ref. [3]). Genomic sequencing and surveillance have played a crucial role in monitoring and mitigating the COVID-19 pandemic. There have been approximately 12 million cases and more than 256,000 deaths reported to date in Africa[4], and African countries have contributed substantial amounts of genomic sequencing data to global agencies. For example, two of the five variants of concern (Beta and Omicron) were first identified in Africa through genomic surveillance systems and real-time sequencing and data release[5,6].

During the first year of the pandemic, SARS-CoV-2 genomes from Africa were mainly produced for a small number of countries with genomes available from 38 of the 54 African countries[7]. Subsequently, the Africa Centres for Disease Control and Prevention (Africa CDC) and the World Health Organization Regional Office for Africa (WHO AFRO) invested in capacity building and provided resources to equip more African countries to produce genomes locally[8]. For example, the African Union Commission and Africa CDC launched the Africa Pathogen Genomics Initiative (Africa PGI) with an initial investment of US$100 million. Currently, more than 100,000 genomes, originating from 51 African countries and 4 independent overseas territories, are publicly available from Global Initiative on Sharing Avian Influenza Data (GISAID)[9].

Dashboards for live COVID-19 information

Online dashboards presenting global and regional COVID-19 data, including case numbers, reported deaths and vaccination rates, have proliferated since the onset of the pandemic[10–12]. These dashboards have a vital role in guiding the public health response and decision-making by policymakers, public health officials and scientists[13]. Data visualization in dashboards also keeps the public abreast of the state of the pandemic. Examples of genomic dashboards include the Welcome Sanger Institute's COVID-19 Genomic Surveillance dashboard (https://go.nature.com/3U9wS8R) and the COVID-19 Genomics UK Consortium dashboard (https://go.nature.com/3Fw32r2). These dashboards include the number of genomes sequenced and the proportion of variants identified in the sequenced genomes, as well as information on the mutations in the lineages of interest. Although these dashboards display important genomic information about England, there was initially no genomics dashboard for the African continent. We therefore set out to devise a dashboard that provides real-time analytical tools for visualization of a genomics-oriented understanding of the state of the pandemic on the African continent.

Data inputs for the SARS-CoV-2 Africa dashboard

The SARS-CoV-2 Africa dashboard is an open-source web-based graphical user interface for presentation of the data produced by genomic surveillance of COVID-19 on the continent, and for provision of details of variants that are currently circulating. The dashboard is supported by the main commercially available web browsers, including Google Chrome, Mozilla Firefox, Microsoft Edge and Safari. The dashboard collates all sequencing data available in GISAID, with metadata linking data to a specific country in Africa, and uses these metadata to display temporal and spatial trends in SARS-CoV-2 evolution in Africa. Genomics data are incorporated into the dashboard using an application programming interface (API) via an agreement with GISAID. The web application processes it, and includes a data quality assessment that can eliminate poor quality registers — for example, sequences assigned to a variant that was submitted before the variant was identified (Fig. 1a).

Data processing in the SARS-CoV-2 Africa dashboard

SARS-CoV-2 genomes are accessioned on GISAID with contextual metadata (such as patient details, collection and sampling strategies, and sequencing and assembly methods) that are subjected to curation by GISAID before release. GISAID data can be freely accessed and downloaded by users after registration. The data acquisition and processing pipelines use Python 3.6 and the web interface is implemented using Streamlit (https://go.nature.com/3DqDE3o), with charts created using Plotly[14]. The code can be locally installed for customization in a Conda environment[15]. Code and dependencies can be installed by cloning the Github repository, available at: https://go.nature.com/3WjtMRw.

Performance of the SARS-CoV-2 Africa dashboard

To evaluate dashboard performance, we implemented an experiment using ApacheBench version 2.3 (https://go.nature.com/3WjtZEi) and varying for different levels of concurrency (10, 100, 500 and 1,000 simultaneous access). For each level of concurrence, we performed

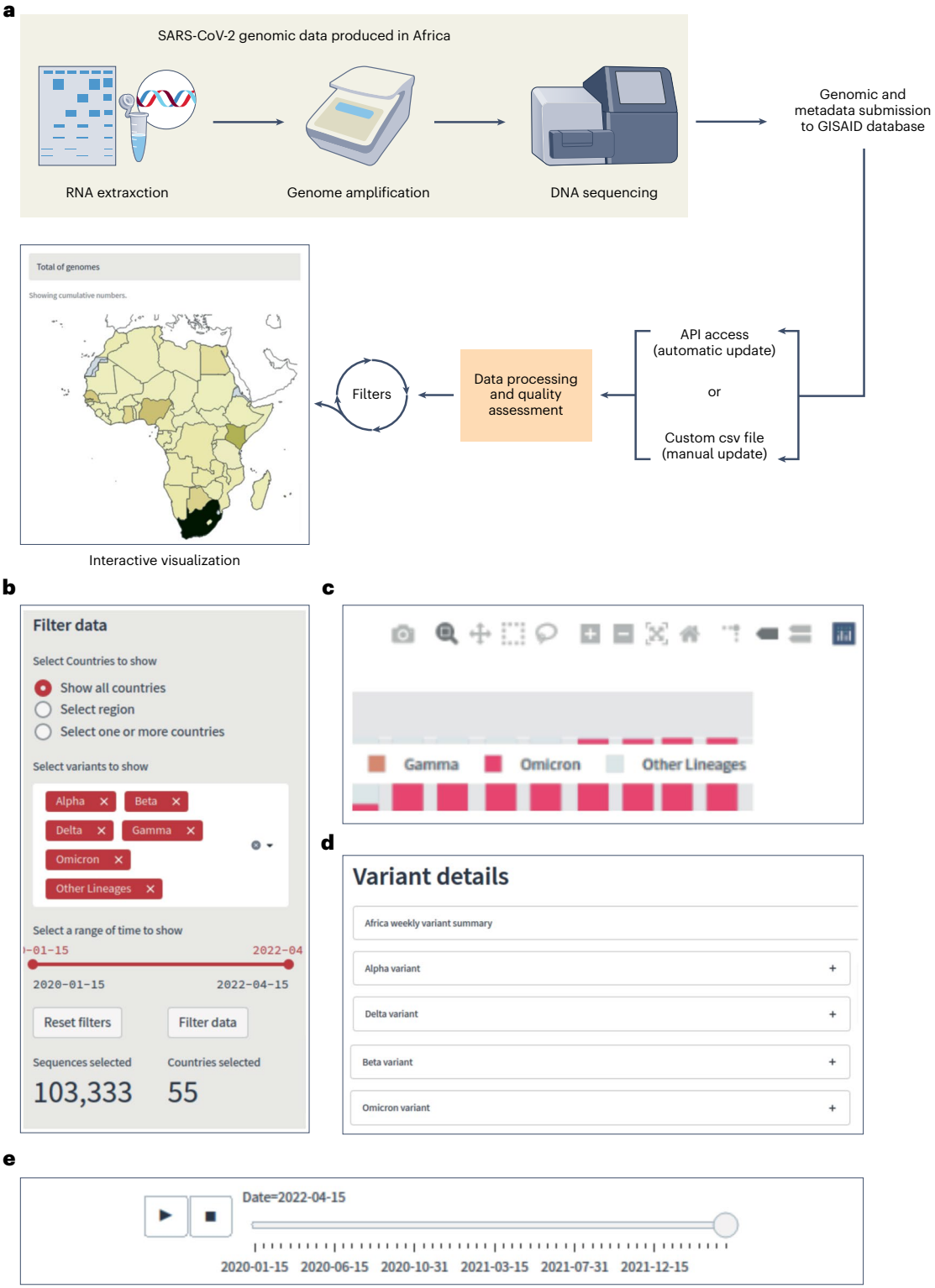


Fig. 1 | Incorporation of SARS-CoV-2 data into the SARS-CoV-2 Africa dashboard. a, An overview of the main features of the interface. **b**, General filters that allow users to select the data of interest. **c**, Figure controls that allow users to enable and disable legend elements and labels, select a part of the figure, zoom in and out, and download the plot. **d**, A tabulated description and mutation map that is provided for variants of interest and/or importance at the time of access. **e**, A timeline player that displays the mapped progression of the pandemic over time, based on the filter selection (**b**).

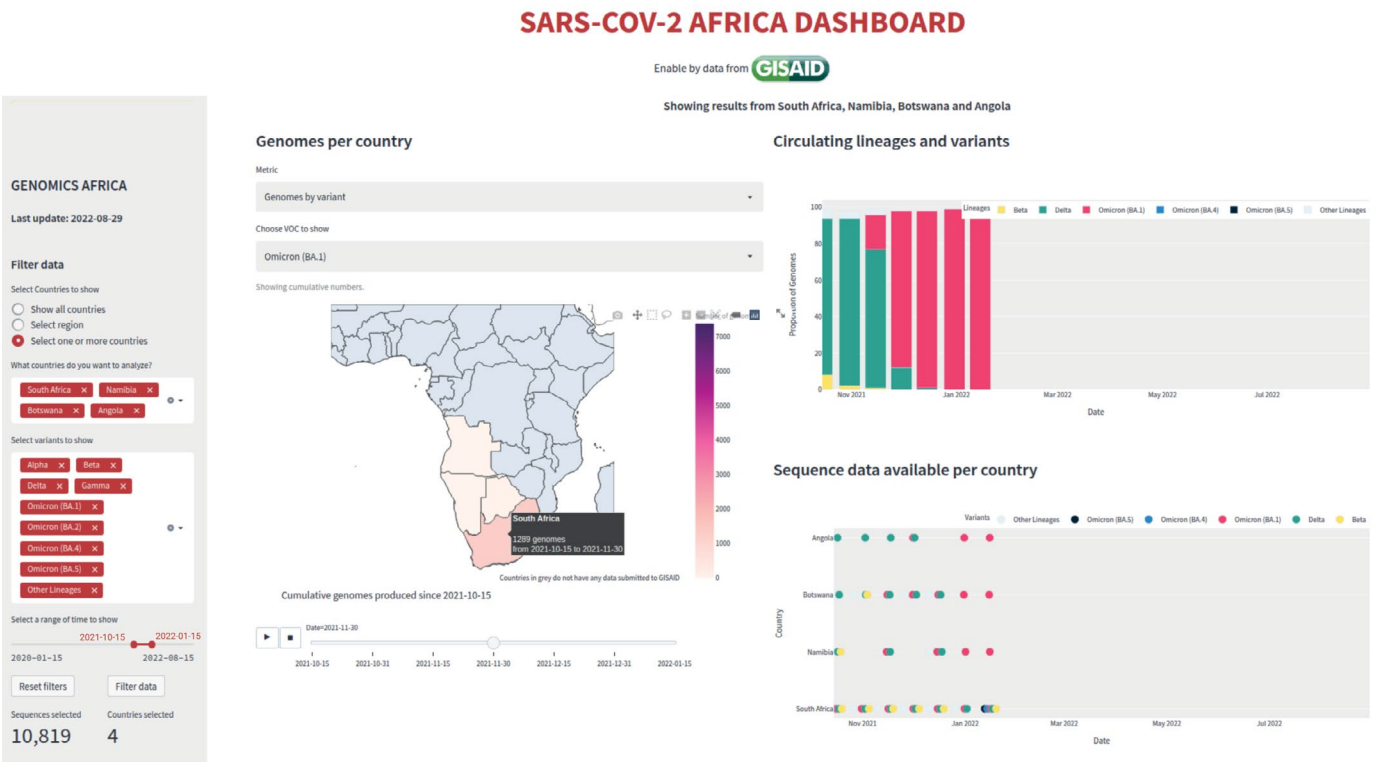


Fig. 2 | A case study of Omicron spread. The case study scenario screenshot shows how to investigate Omicron spread in Namibia and neighbouring countries using the filters and charts provided by the SARS-CoV-2 Africa dashboard interface. Logo courtesy of the GISAID Initiative.

5,000 requisitions, which showed that the dashboard performed well for simultaneous access. For example, in the last level (1,000 requests), only 0.08% of the total requisitions were not answered. All requests for 0–1,000 were completed within an average 1.862 to 25.841 ms (Supplementary Table 1). The dashboard provides interactive visualizations of the temporal and spatial distributions of SARS-CoV-2 variants and their prevalence across different African regions and countries. Several filters are provided to customize the visualizations according to user needs. The main features of the interface are four modules (Fig. 1b–e): general filters allow users to select the data of interest, figure controls allow users to customize the display and snapshot a desired plot, a tabulated weekly summary of variant details is provided, and drop-down mutation maps for variants that are of interest can be used. We also included a timeline player that displays the progression of the pandemic over time, based on user-defined filter selections.

Case study for Omicron spread

A hypothetical example of the application of this dashboard is shown in Fig. 2. The scenario is that the Minister of Health in Namibia wants to understand the spread of the Omicron lineage in neighbouring countries after reports of Omicron in South Africa and Botswana, to better understand how Namibia may be affected. The users would use filters to display the number of Omicron lineage genomes in each neighbouring country. In Fig. 2, on the left-hand side panel of the dashboard, there are filters that allow data for specific countries, specific regions or all countries to be shown. In this example, Namibia and its neighbouring countries (South Africa, Botswana and Angola) have been selected. On the interactive map titled 'Genomes per country', the metric 'Genomes by variant' has been selected. In this case, the Omicron variant was selected. As seen in Fig. 2, if the cursor is hovering over a country on the map, the name of the country, the number of genomes produced by that country and the date are displayed.

When studying the figures on the dashboard with these filters applied, one can see that the proportion of genomes deposited in GISAID at the end of October 2021 is dominated by the Delta lineage, with few remnant Beta genomes. Within the first two weeks of November 2021, the Omicron BA.1 lineage rapidly increased to comprise 19% of all genomes. Watching the sliding scale animation for Omicron lineages on the map displays the early detection of the lineage in South Africa, with swift progression from a low (light pink) to high (dark purple) number of cumulative genomes. From these visualizations, the minister would be aware of the rapid spread of Omicron and its growth advantage over Delta, and would be able to see that, at that time, Omicron had the potential to be the dominant variant in southern Africa. The minister would be empowered with the information needed to enable consultation with local researchers, public health officials and clinicians for the provision of local and regional public health responses to mitigate the effects of Omicron on the population.

Outlook

Numerous dashboards for global and regional COVID-19 data, such as case numbers, reported deaths and vaccination rates, have proliferated since the onset of the COVID-19 pandemic. These dashboards have been vital in guiding the public health response and decision-making by policymakers, public health officials and scientists[13]. Data visualizations produced by these dashboards have also been useful for keeping

the public informed. Genomic surveillance of SARS-CoV-2 has been crucial in monitoring the progression of the pandemic, particularly in the low-vaccination landscape of Africa, where globally important variants have emerged and are likely to continue to appear.

Africa has generated a wealth of genomic surveillance data, with more than 129,000 SARS-CoV-2 genomes currently available on GISAID. The SARS-CoV-2 Africa dashboard is the first detailed online, real-time and interactive tool produced for the Global South. It provides simple and clear graphics that are easy to interpret and equips developers to analyse and visualize the data themselves, by allowing manual input of data via custom csv files, formatted as per the provided template (Supplementary Information). Our dashboard makes often intimidating and complex genomic data accessible to all users, and can be used to inform policy and guide the public health response in Africa and for Africa. All datasets used in our dashboard are in publicly accessible repositories. Genomic data are available from the GISAID database (https://www.gisaid.org/). The SARS-CoV-2 Africa dashboard is freely available at https://climade.health/dashboard/covid-africa/. Source code is available at https://github.com/CERI-KRISP/SARS-Cov-2-Africa-dashboard. Supplementary methods are available at https://doi.org/10.25413/sun.19722025.

Joicymara S. Xavier[1,2,3] ✉, **Monika Moir**[1], **Houriiyah Tegally**[1,4], **Nikita Sitharam**[1], **Wasim Abdool Karim**[1], **James E. San**[1,4], **Joana Linhares**[1], **Eduan Wilkinson**[1], **David B. Ascher**[5,6,7], **Cheryl Baxter**[1,8] ✉, **Douglas E. V. Pires**[5,6,9] ✉ & **Tulio de Oliveira**[1,4,8,10] ✉

[1]Centre for Epidemic Response and Innovation (CERI), School of Data Science and Computational Thinking, Stellenbosch University, Stellenbosch, South Africa. [2]Institute of Agricultural Sciences, Universidade Federal dos Vales do Jequitinhonha e Mucuri, Unaí, Brazil. [3]Instituto René Rachou, Fundação Oswaldo Cruz, Belo Horizonte, Brazil. [4]KwaZulu-Natal Research Innovation and Sequencing Platform (KRISP), Nelson R. Mandela School of Medicine, University of KwaZulu-Natal, Durban, South Africa. [5]Systems and Computational Biology, Bio 21 Institute, University of Melbourne, Melbourne, Victoria, Australia. [6]Computational Biology and Clinical Informatics, Baker Heart and Diabetes Institute, Melbourne, Victoria, Australia. [7]School of Chemistry and Molecular Biosciences, University of Queensland, Brisbane, Queensland, Australia. [8]Centre for the AIDS Programme of Research in South Africa (CAPRISA), Durban, South Africa. [9]School of Computing and Information Systems, University of Melbourne, Melbourne, Victoria, Australia. [10]Department of Global Health, University of Washington, Seattle, WA, USA.

✉e-mail: joicy.xavier@ufvjm.edu.br; cbaxter@sun.ac.za; douglas.pires@unimelb.edu.au; tulio@sun.ac.za

Published online: 22 December 2022

References

1. Li, Q. et al. *N. Engl. J. Med.* **382**, 1199–1207 (2020).
2. Medhat, M. A. & El Kassas, M. *J. Glob. Health* **10**, 010368 (2020).
3. Oluniyi, P. *First African SARS-CoV-2 Genome Sequence from Nigerian COVID-19 Case* (Virological, 2022).
4. Ritchie, H. et al. *Coronavirus Pandemic (COVID-19)* (Our World in Data, October 2022).
5. Viana, R. et al. *Nature* **603**, 679–686 (2022).
6. Tegally, H. et al. Preprint at *medRxiv* https://doi.org/10.1101/2021.09.23.21264018 (2021).
7. Wilkinson, E. et al. *Science* **374**, 423–431 (2021).
8. Tegally, H. et al. *Science* https://doi.org/10.1126/science.abq5358 (2022).
9. Khare, S. et al. *China CDC Wkly* **3**, 1049–1051 (2021).
10. Wright, D. W. et al. *Virus Evol.* **8**, veac023 (2022).
11. *COVID-19 Genomic Surveillance* (Broad Institute, October 2022); https://go.nature.com/3NsMXVi
12. *COVID Data Tracker* (CDC, October 2022); https://go.nature.com/3jcxeZx
13. Peeples, L. *Nature* **603**, 564–567 (2022).
14. *Collaborative Data Science* (Plotly Technologies Inc., 2015).
15. *Anaconda Distribution* (Anaconda Software Distribution, 2016); https://www.anaconda.com/

Acknowledgements

We thank GISAID for providing real-time access to metadata available within the database. SARS-CoV-2 sequencing at CERI is supported, in part, by grants from the South African Medical Research Council (SAMRC), World Health Organization, the Rockefeller Foundation, the Abbott Pandemic Defense Coalition (APDC), the National Institute of Health USA (U01 AI151698) for the United World Antivirus Research Network (UWARN) and the INFORM Africa project through IHVN (U54 TW012041), the South African Department of Science and Innovation (DSI) and the SAMRC under the BRICS JAF no. 2020/049. J.S.X. was supported by Coordenação de Aperfeiçoamento de Pessoal de Nível Superior — Brasil (Capes) — finance code 001. D.E.V.P. received funding from an Oracle for Research Grant. Our research included local African researchers from the Centre for Epidemic Response and Innovation (CERI) and the KwaZulu-Natal Research Innovation and Sequencing Platform (KRISP). Local researchers participated in all steps from study design to implementation and authorship (C.B., E.W., H.T., J.E.S., M.M., N.S., T.d.O. and W.A.K.). Participating authors from CERI and KRISP helped to guide the development of the dashboard so that it is locally relevant and useful in Africa. Roles and responsibilities were agreed among collaborators before the start of the project. Guidance on our dashboard has been disseminated to local African researchers via an instructional webinar, as part of the Africa COVID-19 Genomics Training webinar series, hosted by the African Union and Africa CDC. Our research is not restricted to researchers only, and does not pose any health, safety, security or other risks.

Author contributions

J.S.X., H.T., T.d.O., J.E.S. and E.W. devised the project. J.S.X. developed the SARS-CoV-2 Africa dashboard, with contributions from J.E.S., J.L. and W.A.K. H.T., T.d.O., E.W., W.A.K., D.E.V.P., J.E.S., N.S., C.B. and D.B.A. helped with methodology, tests and improvements of the SARS-CoV-2 Africa dashboard and this manuscript. T.d.O. and C.B. managed the project and funding. J.S.X. and M.M. wrote the manuscript, and all authors read and reviewed it.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41564-022-01276-9.

Peer review information *Nature Microbiology* thanks Amir Bahmani and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Comment

https://doi.org/10.1038/s41564-022-01276-9

SARS-CoV-2 Africa dashboard for real-time COVID-19 information

In the format provided by the authors and unedited

Supplementary table 1. Benchmark for SARS-CoV-2 Africa Dashboard concurrent requests:*

Concurrence level	**Complete requests**	**Failed requests**	**Requests per second (mean)**	**Time per request ** [ms] (mean)**
10	5000	0	38.7	25.841
100	5000	0	225.45	4.436
500	5000	0	536.93	1.862
1000	5000	4	373.66	2.676

* All the experiments were performed using the same wifi network

**Across all concurrent requests.

SUPPLEMENTARY INFORMATION

SARS-CoV-2 Africa Dashboard: An interactive tool for visualizing COVID-19 genomics data

Joicymara S. Xavier[1,2,3*]Monika Moir[1], Houriiyah Tegally[1,7,], Nikita Sitharam[1], Wasim Abdool Karim[1], James E. San[1,7], Joana Linhares[1], Eduan Wilkinson[1], David B. Ascher [4,5,10], Cheryl Baxter[1,8], Douglas E. V. Pires[4,5,6*], Tulio de Oliveira[1, 7, 8, 9*]

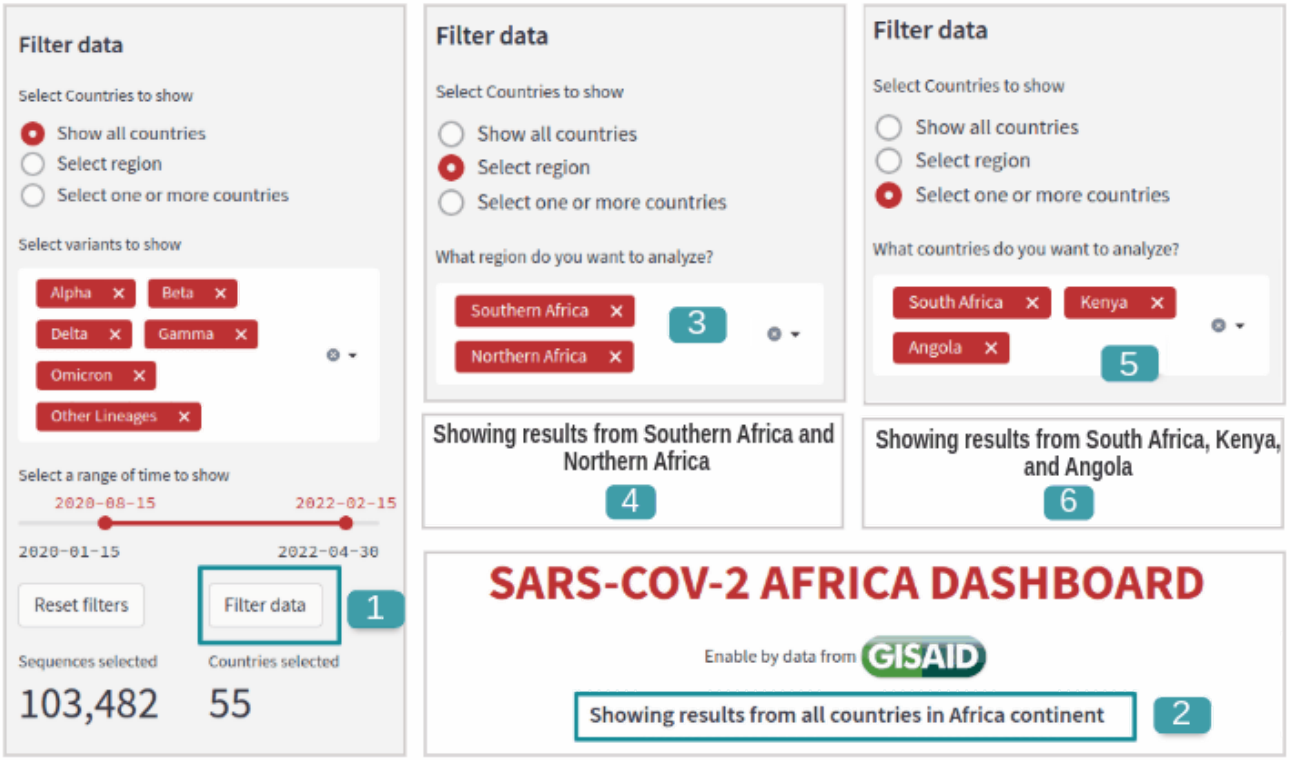
[1]Centre for Epidemic Response and Innovation (CERI), School of Data Science and Computational Thinking, Stellenbosch University, Stellenbosch, South Africa.
[2]Institute of Agricultural Sciences, Universidade Federal dos Vales do Jequitinhonha e Mucuri, Unaí, Brazil
[3]Instituto René Rachou, Fundação Oswaldo Cruz, Belo Horizonte, Brazil
[4]Systems and Computational Biology, Bio 21 Institute, University of Melbourne, Melbourne, Australia
[5]Computational Biology and Clinical Informatics, Baker Heart and Diabetes Institute, Melbourne, Australia
[6]School of Computing and Information Systems, University of Melbourne, Melbourne, Australia
[7]KwaZulu-Natal Research Innovation and Sequencing Platform (KRISP), Nelson R Mandela School of Medicine, University of KwaZulu-Natal, Durban, South Africa.
[8]Centre for the AIDS Programme of Research in South Africa (CAPRISA), Durban, South Africa
[9]Department of Global Health, University of Washington; Seattle, USA.
[10]School of Chemistry and Molecular Biosciences, University of Queensland

*Corresponding authors: TdO, DEVP and JSX., email: tulio@sun.ac.za; douglas.pires@unimelb.edu.au; joicy.xavier@ufvjm.edu.br, tel: +27 82 962 4219; +61 3 8344 8185; +55 38 991717950.

HOW TO USE SARS-CoV-2 AFRICA DASHBOARD

SARS-CoV-2 Africa dashboard requires input data in two ways: by using a GISAID API or a custom tsv file formatted as per the provided template. To use an API, the developer is required to enter into a Data Provision Agreement with GISAID. The use of an API allows for automatic updating of the dashboard. Here we show two tutorials on how to use the SARS-CoV-2 Africa dashboard online and also how to reproduce the code locally.

Tutorial 1. Using SARS-CoV-2 Africa Dashboard online: an example use case

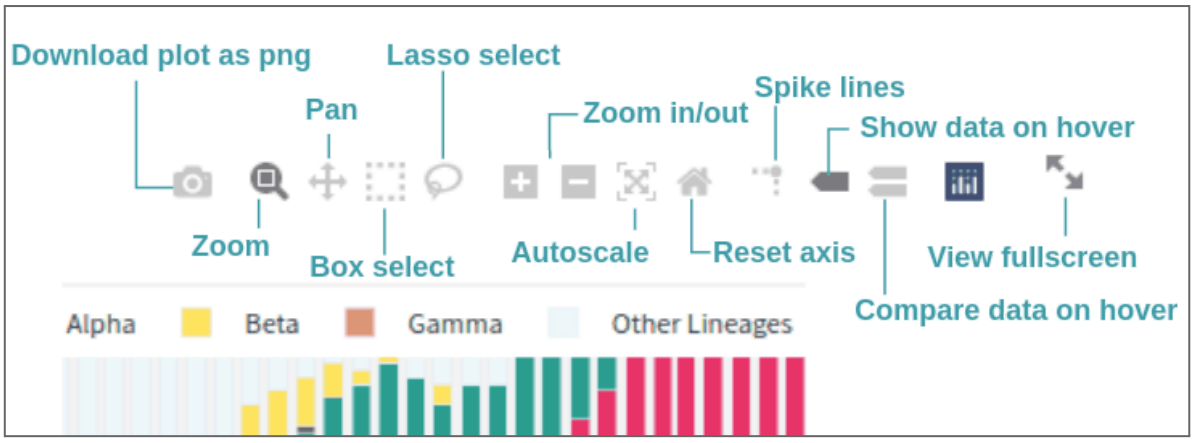


Supplementary figure 1: *Use case example of filtering data selections to customize the SARS-CoV-2 variant, geographic and temporal data display on the SARS-CoV-2 Africa Dashboard.*

When accessing the dashboard online, the default settings are to display the data for all African countries (Show all countries), for all SARS-CoV-2 variants (Select variants to show) with the full-time range of available data (Select a range of time to show). Filters may be applied with radio buttons and drop-down menus to customize the geographic distribution of data on the dashboard display. Similarly, a drop-down menu allows for specific variants to be chosen. A simple sliding time window may be used to adjust the temporal range of the display. These filters are applied with the

'Filter data' button as shown in panel [1]. When filters are implemented the number of sequences and countries selected, shown below the Reset filters and Filter data buttons, are updated. The dashboard display will also update to show which region or countries have been selected for data visualization [panel 2].

The 'Select countries to show' filter has three radio buttons to show all countries, select regions, and select one or more countries. When 'select region' is chosen, a drop-down menu becomes available with six African regions for selection. Multiple regions may be chosen [3] which will display in text on the dashboard [4]. When the 'Select one or more countries' radio button is selected, a drop-down menu with the list of African countries is displayed [5 and 6]. Any applied filters of region, countries, and variants may be removed by clicking on the exit button (X) of the particular filter or by pressing 'Reset filters' [1].



***Supplementary* figure 2:** *Several figure controls are available to download and modify the selected plot.*

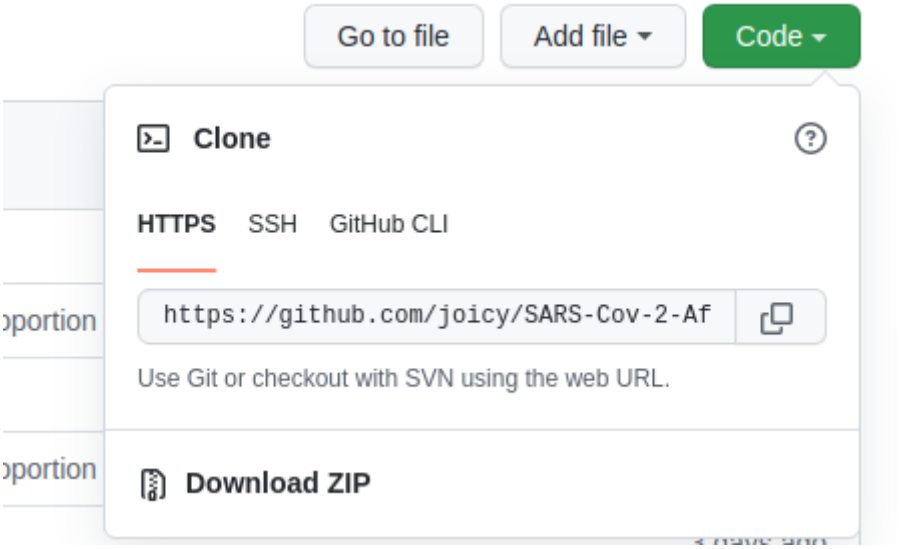
Figure controls are available on the top-right corner of each plot, their functions are as follows:

- Download the plot in png format
- The zoom control magnifies the area of the figure beneath the mouse cursor
- Pan around the figure
- Box select and lasso select controls may be used to highlight a particular area of the figure while the remaining portions of the figure are grayed out
- Zoom in (+) and zoom out (-)
- Autoscale automatically resizes the figure for a better visualization
- Reset axis resets the axes and figure size to the original display size
- Spike lines display a vertical and horizontal line from the mouse cursor location to the spike base on the x- and y- axes of the plot

- Show data on hover control enables the display of a single data label for the point directly beneath the mouse cursor
- Compare data on hover control allows for the display of multiple data labels for the point directly beneath the mouse cursor
- View fullscreen switches the active window to a full-screen view of the specific figure

Tutorial 2. Using SARS-CoV-2 Africa Dashboard offline

1. Download the code at https://github.com/joicy/SARS-Cov-2-Africa-dashboard



2. Install the dependencies using Conda environment:
   Start by opening your terminal and going into the project folder

```
(base) joicy@joicy-pc:~$ cd SARS-Cov-2-Africa-dashboard/
(base) joicy@joicy-pc:~/SARS-Cov-2-Africa-dashboard$ conda env create -f requirements.yml
```

   Type: conda env create -f requirements.yml

   Activate the conda environment:

```
(base) joicy@joicy-pc:~/SARS-Cov-2-Africa-dashboard$ conda activate SARS-Cov-2-Africa-dashboard
(SARS-Cov-2-Africa-dashboard) joicy@joicy-pc:~/SARS-Cov-2-Africa-dashboard$
```

   Type: conda (or source) activate SARS-Cov-2-Africa-dashboard, the name of the environment will appear at the beginning of the line.

3. Now, you can run the dashboard using your own metadata or via setting up a GISAID API.

Using metadata:

- Edit config.py file and set *data_source* variable with your option: *data_source="metadata"*

```
#Fill data_source variable with 'GISAID_API' if you are going to use GISAID feed data or 'metadata' to use a metadata file
data_source = "metadata"
```

- Create your metadata based on *data/template_metadata.csv* (required columns shown below) and save it in *data/metadata.csv*

	A	B	C	D	E	F
1	**lineage**	**collection_date**	**subm_date**	**region**	**country**	**province**
2	B.1.351	2021-03-18	2022-02-21	Africa	Kenya	Nairobi
3	B.1.612	2020-12-01	2021-06-24	Africa	Gabon	
4	B.1.351	2020-12-01	2021-06-24	Africa	Gabon	
5	B.1.612	2021-01-01	2021-06-24	Africa	Gabon	
6	B.1.351	2021-01-01	2021-06-24	Africa	Gabon	
7						
8						

Using GISAID API:

- Edit config.py and set *data_source* variable with your option: *data_source="GISAID_API"*

- Work with GISAID to get a Data Provision Agreement.

- Define the following environment variables in config.py:
  - GISAID_URL
  - GISAID_USERNAME
  - GISAID_PASSWORD

```
#Fill data_source variable with 'GISAID_API' if you are going to use GISAID feed data or 'metadata' to use a metadata file
data_source = "GISAID_API"

# Setup GISAID variables if you are using GISAID feed data
GISAID_URL = 'https://www.                                    'provision.json.xz'
GISAID_USERNAME =
GISAID_PASSWORD =
```

- Perform the edits required to customize your data if your needs differ from the standard in *source/data_process.py*

- If your data is not from the African continent, you must replace the geojson file in data/africa.geojson. We recommend this repository for accessing geojson files.

5. DISCUSSÃO

É evidente que as ferramentas de Ciência de Dados têm sido essenciais para o avanço das pesquisas em todas as áreas do conhecimento e para a Bioinformática não é diferente. Com vacinas sendo desenvolvidas em tempo recorde, como aconteceu recentemente com as vacinas para COVID-19 [79], empresas líderes no sequenciamento de genomas desenvolvendo soluções cada vez mais rápidas e baratas [80], podemos ter uma clara noção da importância e do quanto a velocidade da produção de dados biológicos pode ainda aumentar.

Atentos à necessidade de dados cada vez mais acurados, significativos e informativos, iniciativas têm surgido no sentido de buscar soluções para o compartilhamento e reutilização de dados provenientes de experimentos científicos. Essas iniciativas podem ter o intuito tanto de criar novos métodos quanto de informar a população acerca do desenvolvimento de uma nova doença, por exemplo. Nesta tese, apresentou-se dois projetos que propõem soluções, respectivamente, para esses dois cenários. No primeiro, deparou-se com um cenário de uma importante área de estudo, a termodinâmica de proteínas, que enfrentava há 15 anos uma barreira em relação à obtenção e curadoria de dados. No segundo, foi identificado uma lacuna entre a enorme quantidade de dados gerados a partir do trabalho de vigilância genômica realizado no continente africano e a efetiva comunicação desses dados como ferramenta de tomada de decisão pela comunidade e gestores públicos.

As soluções propostas para os dois casos, envolveu a engenharia de dados em diferentes frentes. No primeiro caso, é proposto uma abordagem de bases de dados manualmente curada a partir de dados da literatura que tem por objetivo continuar sendo mantida através da colaboração das partes interessadas. No segundo caso, foi desenvolvido um dashboard interativo que facilita o entendimento dos dados através de visualizações que permitem análises de formas diversas.

O ThermoMutDB, uma base de dados termodinâmicos de mutações missense de proteínas dos mais diversos organismos, foi desenvolvida através da dupla checagem da base de dados Protherm [18] e aquisição de novas entradas através de busca na literatura. Atualmente, os dados do ThermoMutDB se encontram em sua 3ª versão, totalizando 13.337 entradas manualmente curadas. A confirmação da relevância do trabalho, amplamente discutida neste documento e também no artigo publicado, se dá já no processo de submissão do artigo. Na mesma edição em que o ThermoMutDB foi publicado (*2021 Database Issue* do

NAR) outros dois artigos para bases de dados thermodinâmicos também foram aceitos [81,82], incluindo o próprio Protherm, que voltou a ser atualizado depois de 15 anos.

Essa retomada do interesse em dados termodinâmicos indica uma necessidade que vinha sendo criada por esse hiato deixado pelos autores do Protherm, paralelamente à necessidade do desenvolvimento de ferramentas que utilizam esses dados. Em pouco mais de um ano da sua publicação, o ThermoMutDB tem, até a data da escrita, 25 citações, sendo que 7 delas são referentes a trabalhos que propõem novos métodos e/ou ferramentas. Além disso, como detalhado no ANEXO B, a base de dados tem sido largamente utilizada no mundo todo, contando com mais de 1.800 usuários únicos adquiridos nesses dois anos de publicação. Para além dos acessos, usuários têm contribuído com a contínua curadoria dos dados, reportando erros quando são encontrados, o que permite com que os dados sejam continuamente revisados e atualizados. Atualmente, uma nova ferramenta, chamada LitCurate, vem sendo desenvolvida para apoiar o processo de curadoria não só do ThermoMutDB quanto de outras bases curadas a partir de dados da Literatura. O LitCurate permitirá que a criação de comunidade para curadoria dos dados proposta pelo ThermoMutDB seja ainda mais eficiente. Mais detalhes a respeito dessa nova ferramenta serão discutidas na seção Perspectivas Futuras.

Por sua vez, o SARS-CoV-2 Africa Dashboard, uma ferramenta web para visualização interativa de informações a partir de dados genômicos de COVID-19 do continente Africano depositados no GISAID. O GISAID é uma iniciativa público-privada para armazenamento e compartilhamento de dados de COVID-19 e outros vírus. Em razão dos termos de uso praticados por esta iniciativa e pela enorme quantidade de dados em questão, o acesso a essas informações não é trivial para a população em geral e tomadores de decisão, como governantes, dirigentes de hospitais, entre outros. Adicionalmente, alguns países do continente africano têm liderado os esforços de vigilância genômica de COVID-19 desde o início da pandemia e portanto, também é importante para o mundo observar o que acontece neste continente. O desenvolvimento do dashboard passa a propiciar, para uma enorme quantidade de pessoas dos mais diferentes níveis de entendimento genômico, uma ferramenta que lida, de forma simples e interativa, com os valiosos dados produzidos pelo continente.

Além de fazer análises simples, o dashboard também tem a proposta de auxiliar outros grupos no desenvolvimento de novos dashboards regionais através da reutilização do seu código. O código é disponibilizado no Github (https://github.com/CERI-KRISP/SARS-Cov-2-Africa-dashboard) e pode ser replicado livremente. As tecnologias de desenvolvimento utilizadas também foram planejadas para

serem de simples utilização por qualquer pessoa com nível de programação básico, fomentando e facilitando assim sua disseminação. Embora o dashboard tenha sido disponibilizado recentemente e ainda não tenha sido oficialmente publicado, já conta com um considerável número de visitas, 56 acessos únicos, confirmando o interesse dos usuários no seu conteúdo. Muito ainda se pretende fazer e aperfeiçoar, como será discutido nas perspectivas futuras, incluindo a replicação do dashboard para outros patógenos e a implementação de ferramentas que permitam a análise dos dados genômicos em si através do upload por parte do usuário de arquivos FASTQ, por exemplo.

Ambas as soluções se mostraram eficazes e promissoras para dar suporte ao avanço das pesquisas em suas áreas de atuação. Além do mais, ambas são replicáveis e servem como trabalho inicial para o desenvolvimento e aplicação de novos paradigmas relacionados a dados biológicos, de saúde e de bioinformática que se mostram tão proeminentes para um futuro próximo.

6. PERSPECTIVAS FUTURAS

Durante o desenvolvimento dessa pesquisa, aprendizados transversais e soluções para problemas no entorno do desenvolvimento das ferramentas, foram gerando novos trabalhos que culminaram na criação de um novo laboratório de pesquisa focado principalmente na Curadoria de Dados Biológicos, Desenvolvimento de Ferramentas de Predição e Disponibilização de Dados. O Laboratório de Bioinformática e Inteligência Artificial (BIA) já possui colaborações estabelecidas e alguns trabalhos em andamento que podem ser divididos em duas infraestruturas gerais: Um ecossistema de curadoria de dados e o desenvolvimento de uma plataforma de Data Lake e MLOps (*Machine Learning Operations)*, ilustrados na Figura 3.

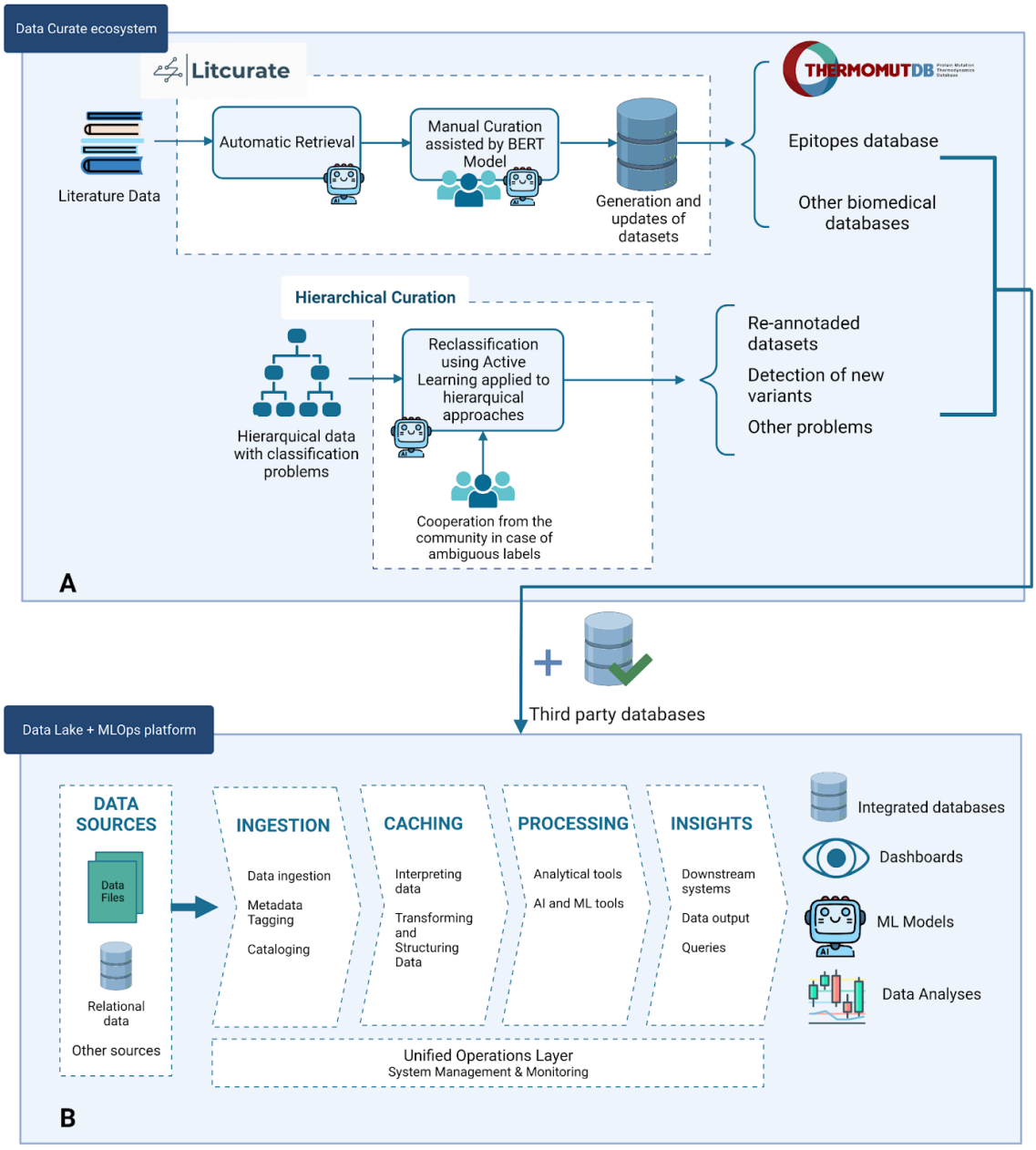


Figura 3 : Representação da integração entre as propostas de trabalhos futuros. **A**. Ecossistema de curadoria de dados: Dois projetos compõem a curadoria de dados, sendo o primeiro, Litcurate uma plataforma para curadoria de dados da literatura envolvendo interação humana suportada por Aprendizado de Máquina. O segundo projeto tem foco na curadoria de bases de dados hierárquicas utilizando Aprendizado de Máquina Ativo, que também envolve colaboração humana. **B**. Proposta de infraestrutura de Data Lake e ferramentas de MLOps para integração dos dados gerados pelo ecossistema de curadoria com bases de dados de terceiro com o objetivo de gerar bases de dados integradas, dashboards interativos, modelos de Machine Learning e análises de dados que possam ser atualizadas em tempo real, se necessário.

6.1. ECOSSISTEMA PARA CURADORIA DE DADOS

A primeira infraestrutura, o ecossistema para curadoria de dados (Figura 3A), nasce das lições aprendidas com a curadoria do ThermoMutDB e também da colaboração para avaliação de abordagens hierárquicas de aprendizado de máquina para classificar bancos de dados biológicos [83,84]. Em ambos os casos, pretende-se criar uma comunidade ou pequenos consórcios de pessoas interessadas nos dados que estão sendo tratados/verificados para que essas possam trabalhar cooperativamente com modelos de ML na curadoria dos dados.

No primeiro caso, apresentamos o LitCurate, um projeto que já está em andamento, cujo foco principal é apoiar a curadoria manual e, no futuro, automatizar todo o processo de recuperação de dados. O LitCurate tem foco nos artigos indexados pelo Pubmed[1], que compreende a maior parte da literatura biomédica publicada em revistas relevantes. Utilizamos os PMIDs (*Pubmed Identifiers*), selecionados para compor o ThermoMutDB e também aqueles que foram excluídos durante a curadoria, para treinar e avaliar um modelo de ML para recuperação de artigos do Pubmed, utilizando a ferramenta LitSuggest [85]. O LitSuggest é um sistema que utiliza aprendizado de máquina para recomendação e curadoria de literatura do Pubmed. Os PMIDs de interesse (positivos) e de não interesse (negativos) são passados como entrada para o LitSuggest, que treina um modelo para aquele projeto específico e a partir daí pode sugerir semanalmente novos artigos com suas respectivas probabilidades de pertencerem à busca desejada.

Uma API (do inglês, *Application Programming Interface*) para ingestão da compilação dos dados do LitSuggest para dados termodinâmicos de mutações, assim como os resultados dos dados recuperados estão sendo validados.

Uma vez recuperados os artigos, o nosso sistema, o LitCurate, que irá segmentar os usuários para cada base de dados de interesse, e ainda diferenciá-los entre Curador (usuário iniciante que faz a anotação dos dados) e Especialista (usuário experiente que pode, além de fazer a anotação, verificar artigos que foram anotados). Na primeira versão do sistema, essa anotação será feita através do preenchimento dos dados em formulário específico para cada base de dados. Em trabalhos futuros, pretende-se investigar modelos de atenção baseados em NLP (do inglês, *Natural Language Process*) [86] que devem ser capazes de indicar no artigo onde possivelmente a informação buscada está. Com essa funcionalidade será possível

[1] https://pubmed.ncbi.nlm.nih.gov/

reduzir consideravelmente o tempo de trabalho do curador. Uma vez que estes modelos forem avaliados e validados, poderá ser possível eliminar o papel do curador, através da recuperação automática dos dados nos artigos, restando apenas os especialistas, responsáveis pela dupla checagem das anotações.

Por fim, os dados recuperados são então utilizados para geração e atualização de bases de dados biomédicas. Além do ThermoMutDB que será mantido por essa infraestrutura, pretende-se dar início em breve, a uma nova base de dados de epítopos, que são utilizados para o desenvolvimento de ferramentas para apoiar o desenvolvimento de vacinas e imunoterapias.

O segundo projeto que integra o ecossistema de curadoria de dados é relacionado a bases de dados hierárquicas. Grande parte dos conjuntos de dados biológicos têm uma natureza hierárquica, como taxonomia do organismo [87,88], domínios estruturais de proteínas [89,90], vias metabólicas [91], entre outros. Assim como as demais bases de dados discutidas aqui, essas bases também estão passíveis de erros e apresentam desafios particulares de classificação, devido a hierarquia. [92] propôs um framework para avaliação e classificação de bases hierárquicas utilizando aprendizado de máquina.

Para esse projeto, pretendemos utilizar o framework proposto por [92], associado ao Aprendizado Ativo [93,94]. O Aprendizado Ativo, em linhas gerais, é uma técnica de aprendizado de máquina semi-supervisionado com que é possível treinar um modelo a partir de um pequeno conjunto de dados, e, à medida que amostras são treinadas, o algoritmo usa esse conjunto para treinar as demais. O aprendizado ativo utiliza-se também de um oráculo, que será consultado toda vez que surgir um rótulo ambíguo no processo. Os usuários do ecossistema poderão colaborar com o processo, sendo esse oráculo.

Uma aplicação possível para esse tipo de anotação, para além da reclassificação de bases de dados, é a detecção de novas variantes, por exemplo. Usualmente, os algoritmos de ML utilizados para anotação de sequências genômicas tentam classificar uma nova sequência a uma nomenclatura já conhecida. O aprendizado ativo pode auxiliar a identificar que aquela sequência pode ainda não ter uma nomenclatura conhecida, no momento que ela chega como entrada para o modelo. Dessa forma, ao invés de continuar tentando classificar o que é desconhecido, o modelo deverá consultar o oráculo, que terá a tarefa de fazer essa verificação.

6.2. DATA LAKE E PLATAFORMA DE MLOPs

Levando em consideração a enorme quantidade de dados e a diversidade de tarefas possíveis para serem realizadas com esses dados, como integração entre bases de dados, incluindo bases de dados de terceiros, e a complexidade envolvida no gerenciamento, versionamento, transformação e segurança desses dados, uma infraestrutura de *Data Lake [95]* está sendo desenvolvida. Além da preocupação relacionada aos dados, surge também a necessidade de se acompanhar o ciclo de vida de um modelo de ML, área que é o foco do MLOps [96]. As ferramentas de MLOps visam reunir um conjunto de melhores práticas para implantar e manter modelos de ML em produção [96–99], o que permite o re-treinamento contínuo dos modelos.

Como mostrado na Figura 3B, a arquitetura é dividida em camadas, que vai da Ingestão de Dados, que pode ser de diferentes fontes e formatos, até a disponibilização desses dados para análises, sistemas de ML, dashboards, etc. Algumas das características da arquitetura de Data Lake está na capacidade de centralização de dados de diversas fontes em um só lugar, ao mesmo tempo que aceita dados estruturados, semi-estruturados e não-estruturados, tem baixo custo e também suporta regras de segurança e proteção de dados.

Uma prova de conceito foi feita, por um dos alunos do BIA, utilizando dados do ThermoMutDB e do SARS-CoV-2 dashboard para as duas primeiras camadas do Data Lake (ingestão e *caching*). Além disso, essa arquitetura está sendo planejada para atender também a necessidade do INFORM Africa *consortium*, um hub de Pesquisa em Ciência de Dados financiado pelo *National Institutes of Health* (NIH) para Descoberta e Inovação em Saúde na África (DS-I África). O INFORM Africa é um consórcio de 5 anos, composto por 3 projetos que buscam identificar intersecções entre pacientes de COVID-19 e HIV. Esses projetos são suportados pelo *Data Management and Analysis Core* (DMAC) e *Next Generation Sequencing* (NGS) *Core*, responsável pelo desenvolvimento da plataforma em questão. Nós propusemos [100] e estamos desenvolvendo, uma arquitetura que será capaz de capturar e fornecer suporte de análise para dados relevantes, oportunos, precisos e coerentes que possam ser interpretados e acessados por colaboradores em vários países africanos.

Um outro projeto que integrará a plataforma de Data Lake e MLOps é o preditor de mutações compensatórias, que deu início ao projeto do ThermoMutDB. Este preditor será o primeiro produto de todo o ecossistema, uma vez que utilizará os dados do ThermoMutDB, advindos do ecossistema de curadoria, através do LitCurate, que deverá ser ingerido, transformado e gerenciado pelo Data Lake. Uma vez desenvolvido, o modelo de ML poderá

ser o primeiro da nossa plataforma a ser retreinado conforme a base de dados for atualizada, utilizando os paradigmas e ferramentas do MLOps.

7. CONCLUSÕES

Nestre trabalho, exploramos dois tópicos relevantes para a Bioinformática, um dentro da Bioinformática Estrutural e o outro na Vigilância Genômica e em ambos, estratégias particulares relacionadas à engenharia dos dados apareceram como tema central. Propusemos duas abordagens, uma para curadoria e disponibilização dos dados e a outra para integração e disponibilização de informações através de visualizações interativas. As duas abordagens, conjuntas, perpassam por todo o pipeline de Ciência de Dados, saindo dos dados brutos e curadoria até a entrega de informação útil para o público alvo.

Durante o desenvolvimento das abordagens foi possível entender e identificar as particularidades associadas aos problemas biológicos e biomédicos de cada área e propor abordagens que, associados a tecnologias emergentes, podem ser aliadas na elucidação de questões biológicas e de saúde pública relevantes.

Propusemos uma abordagem para curadoria de bases de dados biomédicas baseada na literatura que independe do esforço de uma instituição ou grupo de pesquisa, deixando essa tarefa para a comunidade, ao mesmo tempo que controla os usuários colaboradores. A rotulagem de dados recrutando pessoas envolvidas nas comunidades é uma das práticas recomendadas pelo movimento Data Centric AI[2] e mostrou-se efetivo nesses primeiros anos do ThermoMutDB, mesmo que ainda não tenha sido completamente estruturado.

Desenvolvemos também um dashboard interativo que permite com que o público geral e tomadores de decisão tenham acesso fácil a informações advindas da enorme quantidade de genomas de SARS-CoV-2 gerados pelo continente africano, cabendo a nós o acordo com a instuição provedora dos dados. A ferrmenta foi desenvolvida com ferramentas open-source e pode ser reproduzida para atender outras regiões ou outros patógenos.

A proposta e desenvolvimento das abordagens puderam conferir uma visão geral das necessidades das duas áreas e propor soluções para integração e gerenciamento de dados através de uma plataforma robusta de Ciência de Dados. Essa plataforma possibilitará não só o desenvolvimento de ferramentas específicas para essas áreas em questão, como também responder outras questões biológicas relevantes.

Alguns dos projetos futuros para o uso das plataformas propostas incluem a manutenção e atualização do ThermoMutDB, a geração de outras bases de dados que são comuns aos preditores da plataforma mCSM, o desenvolvimento de um preditor para identificar mutações compensatórias, a reclassificação de bases de dados hieráquicas, como

[2] https://datacentricai.org/labeling-and-crowdsourcing/

bases de dados taxonômicas, o desenvolvimento de um preditor para identificar novas variantes de COVID-19, entre outros trabalhos.

Conclui-se portanto, que essa pesquisa pôde contribuir para o avanço das áreas relacionadas através de abordagens simples mas que têm demonstrado ser eficientes para os seus contextos. Observa-se também que ainda há muitas oportunidades relacionadas à engenharia e qualidade de dados que poderão ser explorados para dar suporte à análises biológicas mais confiáveis.

REFERÊNCIAS

1 Van der Auwera, G.A. and O'Connor, B.D. (2020) *Genomics In The Cloud*, O'reilly Media.

2 Paiva, V. de A. *et al.* (2022) Protein structural bioinformatics: An overview. *Comput. Biol. Med.* 147, 105695.

3 Chen, J. and Coppola, G. (2018) Bioinformatics and genomic databases. *Handb. Clin. Neurol.* 147, 75–92.

4 Chen, C. *et al.* (2017) Protein bioinformatics databases and resources. *Methods Mol. Biol.* 1558, 3–39.

5 Imker, H.J. (2018) 25 years of molecular biology databases: A study of proliferation, impact, and maintenance. *Front. Res. Metr. Anal.* 3.

6 Rigden, D.J. and Fernández, X.M. (2022) The 2022 Nucleic Acids Research database issue and the online molecular biology database collection. *Nucleic Acids Res.* 50, D1–D10.

7 Grishin, D. *et al.* (2018) Accelerating Genomic Data Generation and Facilitating Genomic Data Access Using Decentralization, Privacy-Preserving Technologies and Equitable Compensation.

8 Kahn, S.D. (2011) On the future of genomic data. *Science* 331, 728–729.

9 Pucci, F. *et al.* (2022) Artificial intelligence challenges for predicting the impact of mutations on protein stability. *Curr. Opin. Struct. Biol.* 72, 161–168.

10 National Research Council (US) Board on Biology (2000) *Bioinformatics: Converting Data to Knowledge: Workshop Summary*, National Academies Press (US).

11 Abedjan, Z. (2022) Enabling data-centric AI through data quality management and data literacy. *it - Information Technology* 64, 67–70.

12 Hamid, O.H. (2022) , From Model-Centric to Data-Centric AI: A Paradigm Shift or Rather a Complementary Approach? , in *2022 8th International Conference on Information Technology Trends (ITT)*, pp. 196–199.

13 Whang, S.E. *et al.* (2021) Data Collection and Quality Challenges in Deep Learning: A Data-Centric AI Perspective. *arXiv*.

14 Pan, I. *et al.* (2022) Data-centric Engineering: integrating simulation, machine learning and statistics. Challenges and opportunities. *Chem. Eng. Sci.* 249, 117271.

15 Kuiken, C. *et al.* (2003) HIV sequence databases. *AIDS Rev.* 5, 52–61.

16 Pleasance, E.D. *et al.* (2010) A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature* 463, 191–196.

17 Abbott, K.L. *et al.* (2015) The Candidate Cancer Gene Database: a database of cancer driver genes from forward genetic screens in mice. *Nucleic Acids Res.* 43, D844-8.

18 Kumar, M.D.S. *et al.* (2006) ProTherm and ProNIT: thermodynamic databases for proteins and

protein-nucleic acid interactions. *Nucleic Acids Res.* 34, D204-6.

19 Pires, D.E.V. *et al.* (2015) Platinum: a database of experimentally measured effects of mutations on structurally defined protein-ligand complexes. *Nucleic Acids Res.* 43, D387-91.

20 Maisnier-Patin, S. and Andersson, D.I. (2004) Adaptation to the deleterious effects of antimicrobial drug resistance mutations by compensatory evolution. *Res. Microbiol.* 155, 360–369.

21 Wright, S. (1982) The shifting balance theory and macroevolution. *Annu. Rev. Genet.* 16, 1–19

22 Kimura, M. (1990) Some models of neutral evolution, compensatory evolution, and the shifting balance process. *Theor. Popul. Biol.* 37, 150–158.

23 Whitlock, M.C. (2000) Fixation of new alleles and the extinction of small populations: drift load, beneficial alleles, and sexual selection. *Evolution* 54, 1855–1861.

24 Poon, A. and Otto, S.P. (2000) Compensating for our load of mutations: freezing the meltdown of small populations. *Evolution* 54, 1467–1479.

25 Whitlock, M.C. *et al.* (2003) , Compensating for the meltdown: the critical effective size of a population with deleterious and compensatory mutations. , in *Annales Zoologici Fennici*, 169–183.

26 Kondrashov, A.S. *et al.* (2002) Dobzhansky-Muller incompatibilities in protein evolution. *Proc Natl Acad Sci USA* 99, 14878–14883.

27 Davis, B.H. *et al.* (2009) Compensatory mutations are repeatable and clustered within proteins. *Proceedings of the Royal Society B: Biological Sciences* 276, 1823–1827.

28 Li, Q. *et al.* (2020) Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus-Infected Pneumonia. *N. Engl. J. Med.* 382, 1199–1207.

29 Korber, B. *et al.* (2020) Tracking Changes in SARS-CoV-2 Spike: Evidence that D614G Increases Infectivity of the COVID-19 Virus. *Cell* 182, 812-827.

30 Swaminathan, S. (2020) The WHO's chief scientist on a year of loss and learning. *Nature* 588, 583–585.

31 Gossage, L. *et al.* (2014) An integrated computational approach can classify VHL missense mutations according to risk of clear cell renal carcinoma. *Hum. Mol. Genet.* 23, 5976–5988.

32 Jafri, M. *et al.* (2015) Germline mutations in the CDKN2B tumor suppressor gene predispose to renal cell carcinoma. *Cancer Discov.* 5, 723–729.

33 Nemethova, M. *et al.* (2016) Twelve novel HGD gene variants identified in 99 alkaptonuria patients: focus on 'black bone disease'in Italy. *European Journal of Human Genetics* 24, 66.

34 Phelan, J. *et al.* (2016) Mycobacterium tuberculosis whole genome sequencing and protein structure modelling provides insights into anti-tuberculosis drug resistance. *BMC Med.* 14, 31.

35 Handel, A. *et al.* (2006) The role of compensatory mutations in the emergence of drug resistance. *PLoS Comput. Biol.* 2, e137.

36 Thorisson, G.A. *et al.* (2009) Genotype-phenotype databases: challenges and solutions for the post-genomic era. *Nat. Rev. Genet.* 10, 9–18.

37 Lesk, A. (2010) *Introduction to protein science: architecture, function, and genomics*, Oxford university press.

38 Pace, C.N. (1990) Measuring and increasing protein stability. *Trends Biotechnol.* 8, 93–98.

39 Gromiha, M.M. (2010) *Protein bioinformatics: from sequence to function*, Academic Press.

40 DePristo, M.A. *et al.* (2005) Missense meanderings in sequence space: a biophysical view of protein evolution. *Nat. Rev. Genet.* 6, 678–687.

41 Tokuriki, N. and Tawfik, D.S. (2009) Stability effects of mutations and protein evolvability. *Curr. Opin. Struct. Biol.* 19, 596–604.

42 Tokuriki, N. *et al.* (2009) Do viral proteins possess unique biophysical features? *Trends Biochem. Sci.* 34, 53–59.

43 Yi, S. *et al.* (2017) Functional variomics and network perturbation: connecting genotype to phenotype in cancer. *Nat. Rev. Genet.* 18, 395–410.

44 Eilbeck, K. *et al.* (2017) Settling the score: variant prioritization and Mendelian disease. *Nat. Rev. Genet.* 18, 599–612.

45 Pires, D.E.V. *et al.* (2020) A Comprehensive Computational Platform to Guide Drug Development Using Graph-Based Signature Methods. *Methods Mol. Biol.* 2112, 91–106.

46 Airey, E. *et al.* (2021) Identifying Genotype-Phenotype Correlations via Integrative Mutation Analysis. *Methods Mol. Biol.* 2190, 1–32.

47 Pires, D.E.V. *et al.* (2014) mCSM: predicting the effects of mutations in proteins using graph-based signatures. *Bioinformatics* 30, 335–342.

48 Pires, D.E.V. *et al.* (2016) mCSM-lig: quantifying the effects of mutations on protein-small molecule affinity in genetic disease and emergence of drug resistance. *Sci. Rep.* 6, 29575.

49 Pires, D.E.V. *et al.* (2014) DUET: a server for predicting effects of mutations on protein stability using an integrated computational approach. *Nucleic Acids Res.* 42, W314-9.

50 Rodrigues, C.H. *et al.* (2018) DynaMut: predicting the impact of mutations on protein conformation, flexibility and stability. *Nucleic Acids Res.* 46, W350–W355.

51 Rodrigues, C.H.M. *et al.* (2019) mCSM-PPI2: predicting the effects of mutations on protein-protein interactions. *Nucleic Acids Res.* 47, W338–W344.

52 Pires, D.E.V. and Ascher, D.B. (2017) mCSM-NA: predicting the effects of mutations on protein-nucleic acids interactions. *Nucleic Acids Res.* 45, W241–W246.

53 Pires, D.E.V. and Ascher, D.B. (2016) mCSM-AB: a web server for predicting antibody-antigen affinity changes upon mutation with graph-based signatures. *Nucleic Acids Res.* 44, W469-73.

54 Myung, Y. *et al.* (2020) mCSM-AB2: guiding rational antibody design using graph-based signatures. *Bioinformatics* 36, 1453–1459.

55 Nguyen, T.B. *et al.* (2021) mmCSM-NA: accurately predicting effects of single and multiple mutations on protein-nucleic acid binding affinity. *NAR Genom. Bioinform.* 3, lqab109.

56 Iftkhar, S. *et al.* (2022) cardioToxCSM: A Web Server for Predicting Cardiotoxicity of Small Molecules. *J. Chem. Inf. Model.*

57 Santana, C.A. *et al.* (2022) GRaSP-web: a machine learning strategy to predict binding sites based on residue neighborhood graphs. *Nucleic Acids Res.*

58 Paiva, V.A. *et al.* (2022) GASS-Metal: identifying metal-binding sites on protein structures using genetic algorithms. *Brief. Bioinformatics*.

59 Pires, D.E.V. *et al.* (2022) cropCSM: designing safe and potent herbicides with graph-based signatures. *Brief. Bioinformatics*.

60 Nguyen, T.B. *et al.* (2022) CSM-carbohydrate: protein-carbohydrate binding affinity prediction and docking scoring function. *Brief. Bioinformatics*.

61 da Silva, B.M. *et al.* (2022) epitope3D: a machine learning method for conformational B-cell epitope prediction. *Brief. Bioinformatics*.

62 Pires, D.E.V. *et al.* (2011) Cutoff Scanning Matrix (CSM): structural classification and function prediction by protein inter-residue distance patterns. *BMC Genomics* 12 Suppl 4, S12.

63 Abayakoon, P. *et al.* (2018) Structural and Biochemical Insights into the Function and Evolution of Sulfoquinovosidases. *ACS Cent. Sci.* 4, 1266–1273.

64 Ragoza, M. *et al.* (2017) Protein-Ligand Scoring with Convolutional Neural Networks. *J. Chem. Inf. Model.* 57, 942–957.

65 Xavier, J.S. *et al.* (2021) ThermoMutDB: a thermodynamic database for missense mutations. *Nucleic Acids Res.* 49, D475–D479.

66 Bedford, J. *et al.* (2020) COVID-19: towards controlling of a pandemic. *Lancet* 395, 1015–1018.

67 Chen, J. (2020) Pathogenicity and transmissibility of 2019-nCoV-A quick overview and comparison with other emerging viruses. *Microbes Infect.* 22, 69–71.

68 Johnson, H.C. *et al.* (2020) Potential scenarios for the progression of a COVID-19 epidemic in the European Union and the European Economic Area, March 2020. *Euro Surveill.* 25.

69 Zhu, N. *et al.* (2020) A Novel Coronavirus from Patients with Pneumonia in China, 2019. *N. Engl. J. Med.* 382, 727–733.

70 Carter, L.L. *et al.* (2022) Global genomic surveillance strategy for pathogens with pandemic and epidemic potential 2022-2032. *Bull. World Health Organ.* 100, 239-239A.

71 The Independent Panel for Pandemic Preparedness & Response, T. independent panel May-(2021) , COVID-19: make it the last pandemic. , *Geneva: The Independent Panel for*

Pandemic Preparedness and Response. [Online]. Available: https://theindependentpanel.org/wp-content/uploads/2021/05/COVID-19-Make-it-the-Last-Pandemic_final.pdf. [Accessed: 01-Oct-2022]

72 WHO, W.H.O. 19-Apr-(2021) , Statement on the seventh meeting of the International Health Regulations (2005) Emergency Committee regarding the coronavirus disease (COVID-19) pandemic.[Online].Available:https://www.who.int/news/item/19-04-2021-statement-on-the-seventh-meeting-of-the-international-health-regulations-(2005)-emergency-committee-regarding-the-coronavirus-disease-(covid-19)-pandemic. [Accessed: 02-Oct-2022]

73 PAHO, P. 08-Oct-(2020) , Nota Técnica: Caracterização genômica de SARS-CoV-2 e variantescirculantes na região das Américas. [Online]. Available: https://www.paho.org/pt/documentos/nota-tecnica-caracterizacaogenomica-sars-cov-2-e-variantes-circulantes-na-regiao-das. [Accessed: 04-Oct-2022]

74 CDC, C. for D. and C.P. 24-Jan-(2022) , What is Genomic Surveillance? . [Online]. Available: https://www.cdc.gov/coronavirus/2019-ncov/variants/genomic-surveillance.html. [Accessed: 04-Oct-2022]

75 Tegally, H. *et al.* (2021) Rapid replacement of the Beta variant by the Delta variant in South Africa. *medRxiv*.

76 Viana, R. *et al.* (2022) Rapid epidemic expansion of the SARS-CoV-2 Omicron variant in southern Africa. *Nature* 603, 679–686.

77 Tegally, H. *et al.* (2022) Emergence of SARS-CoV-2 Omicron lineages BA.4 and BA.5 in South Africa. *Nat. Med.* 28, 1785–1790

78 Wadman, M. (2022) An advocate for Africa. *Science* 378, 17–21.

79 Barrett, A.D.T. *et al.* (2022) The rapid progress in COVID vaccine development and implementation. *npj Vaccines* 7, 20.

80 WIRED, E.M. 29-Sep-(2022) , The Era of Fast, Cheap Genome Sequencing Is Here. , *Wired*. [Online]. Available: https://www.wired.com/story/the-era-of-fast-cheap-genome-sequencing-is-here/. [Accessed: 09-Oct-2022]

81 Nikam, R. *et al.* (2021) ProThermDB: thermodynamic database for proteins and mutants revisited after 15 years. *Nucleic Acids Res.* 49, D420–D424.

82 Stourac, J. *et al.* (2021) FireProtDB: database of manually curated protein stability data. *Nucleic Acids Res.* 49, D319–D324.

83 Rezende, P.M. *et al.* (2022) Evaluating hierarchical machine learning approaches to classify biological databases. *Brief. Bioinformatics*.

84 Marinho Rezende, P. and Santos Xavier, J. (2019) , Uso de aprendizado de máquina para criação de um Banco de Dados com informações taxonômicas de gene de rRNA 16s. , in *Anais do 14º Simpósio Brasileiro de Automação Inteligente*.

85 Allot, A. *et al.* (2021) LitSuggest: a web-based system for literature recommendation and

curation using machine learning. *Nucleic Acids Res.* 49, W352–W358.

86 Kamath, U. *et al.* (2022) Pre-trained and Application-Specific Transformers. In *Transformers for machine learning: A deep dive* pp. 155–186, Chapman and Hall/CRC.

87 Söhngen, C. *et al.* (2016) BacDive--The Bacterial Diversity Metadatabase in 2016. *Nucleic Acids Res.* 44, D581-5.

88 Quast, C. *et al.* (2013) The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.* 41, D590-6.

89 Murzin, A.G. and Bateman, A. (2001) CASP2 knowledge-based approach to distant homology recognition and fold prediction in CASP4. *Proteins* Suppl 5, 76–85.

90 Sandaruwan, P.D. and Wannige, C.T. (2021) An improved deep learning model for hierarchical classification of protein families. *PLoS ONE* 16, e0258625.

91 Kulmanov, M. and Hoehndorf, R. (2021) DeepGOPlus: improved protein function prediction from sequence. *Bioinformatics* DOI: 10.1093/bioinformatics/btaa763.

92 Rezende, P.M. *et al.* Evaluating hierarchical machine learning approachesto classify biological databases. *Brienfings in Bioinformatcs*.

93 Ertekin, S. *et al.* (2007) , Active learning for class imbalance problem. , 7, pp. 823–824

94 Settles, B. (2009) Active learning literature survey.

95 Giebler, C. *et al.* (2019) Leveraging the data lake: current state and challenges. In *Big data analytics and knowledge discovery: 21st international conference, dawak 2019, linz, austria, august 26–29, 2019, proceedings* 11708 (Ordonez, C. et al., eds), pp. 179–188, Springer International Publishing

96 Kreuzberger, D. *et al.* (2022) Machine Learning Operations (MLOps): Overview, Definition, and Architecture. *arXiv* DOI: 10.48550/arxiv.2205.02302

97 Alla, S. and Adari, S.K. (2021) What Is MLOps? In *Beginning MLOps with MLFlow: Deploy Models in AWS SageMaker, Google Cloud, and Microsoft Azure* pp. 79–124, Apress

98 Makinen, S. *et al.* (2021) , Who needs mlops: what data scientists seek to accomplish and how can mlops help? , in *2021 IEEE/ACM 1st Workshop on AI Engineering - Software Engineering for AI (WAIN)*, pp. 109–112

99 Renggli, C. *et al.* (2021) A Data Quality-Driven View of MLOps. *arXiv* DOI: 10.48550/arxiv.2102.07750

100 Poongavanan, J. *et al.* (2022) Managing and assembling population-scale data streams, tools and workflows to plan for future pandemics within the INFORM Africa Consortium Authors. *Unpublished* DOI: 10.13140/rg.2.2.32839.57766

A ARTIGO THE EVOLVING SARS-COV-2 EPIDEMIC IN AFRICA: INSIGHTS FROM RAPIDLY EXPANDING GENOMIC SURVEILLANCE

Cite as: H. Tegally *et al*., *Science* 10.1126/science.abq5358 (2022).

The evolving SARS-CoV-2 epidemic in Africa: Insights from rapidly expanding genomic surveillance

All authors and their affiliations appear at the end of this paper.

Investment in SARS-CoV-2 sequencing in Africa over the past year has led to a major increase in the number of sequences generated, now exceeding 100,000 genomes, used to track the pandemic on the continent. Our results show an increase in the number of African countries able to sequence domestically, and highlight that local sequencing enables faster turnaround time and more regular routine surveillance. Despite limitations of low testing proportions, findings from this genomic surveillance study underscores the heterogeneous nature of the pandemic and shed light on the distinct dispersal dynamics of Variants of Concern, particularly Alpha, Beta, Delta, and Omicron, on the continent. Sustained investment for diagnostics and genomic surveillance in Africa is needed as the virus continues to evolve, while the continent faces many emerging and re-emerging infectious disease threats. These investments are crucial for pandemic preparedness and response and will serve the health of the continent well into the 21st century.

What originally started as a small cluster of pneumonia cases in Wuhan, China over two years ago (*1*), quickly turned into a global pandemic. Coronavirus Disease 2019 (COVID-19) is the clinical manifestation of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) infection; and by March 2022 there had been over 437 million reported cases and over 5.9 million reported deaths (*2*). Though Africa accounts for the lowest number of reported cases and deaths thus far, with ~11.3 million reported cases and 245 000 reported deaths as of February 2022, the continent has played an important role in shaping the scientific response to the pandemic with the implementation of genomic surveillance and the identification of two of the five variants of concerns (VOCs) (*3*, *4*).

Since it emerged in 2019, SARS-CoV-2 has continued to evolve and adapt (*5*). This has led to the emergence of several viral lineages that carry mutations that confer some viral adaptive advantages that increase transmission and infection (*6*, *7*), or counter the effect of neutralizing antibodies from vaccination (*8*) or previous infections (*9–11*). The World Health Organization (WHO) classifies certain viral lineages as variants of concert (VOCs) or variants of interest (VOIs) based on the potential impact they may have on the pandemic, with VOCs regarded as the highest risk. To date, five VOCs have been classified by the WHO, two of which were first detected on the African continent (Beta and Omicron) (*3*, *4*, *12*), while two more (Alpha and Delta) (*12*, *13*) have spread extensively on the continent in successive waves. The remaining VOC, Gamma (*14*), originated in Brazil and had a limited influence in Africa with only four recorded sequenced cases.

For genomic surveillance to be useful for public health responses, sampling for sequencing needs to be both spatially and temporally representative. In the case of SARS-CoV-2 in Africa, this means extending the geographic coverage of sequencing capacity to capture the dynamic genomic epidemiology in as many locations as possible. In a meta-analysis of the first 10 000 SARS-CoV-2 sequences generated in 2020 from Africa (*15*) several blind spots were identified with regards to genomic surveillance on the continent. Since then, much investment has been devoted to building capacity for genomic surveillance in Africa, coordinated mostly by the Africa Centers for Disease Control (Africa CDC) and the regional office of the WHO in Africa (or WHO AFRO), but also provided by several national and international partners resulting in an additional 90 000 sequences shared over the past year (April 2021 - March 2022). This makes the sequencing effort for SARS-CoV-2 a phenomenal milestone. In comparison, only 12 000 whole genome influenza sequences (*16*) and only ~3 700 whole genome HIV sequences (*17*) from Africa have been shared publicly even though HIV has plagued the continent for decades.

Here we describe how the first 100 000 SARS-CoV-2 sequences from Africa have helped describe the pandemic on the continent, how this genomic surveillance in Africa has expanded, and how we adapted our sequencing methods to deal with an evolving virus. We also highlight the impact that genomic sequencing in Africa has had on the global public health response, particularly through the identification and early analysis of new variants. Finally, we also describe here for the first time how the Delta and Omicron variants have spread across the continent, and how their transmission dynamics were distinct from the Alpha and Beta variants that preceded them.

Results

Epidemic waves driven by variant dynamics and geography

Scaling up sequencing in Africa has provided a wealth of information on how the pandemic unfolded on the continent. The epidemic has largely been spatially heterogeneous across Africa, but most countries have experienced multiple waves of infection (*18–29*), with significant local and regional diversity in the first and to a lesser extent the second waves, followed by successive sweeps of the continent with Delta and Omicron (Fig. 1A). In all regions of the continent, different lineages and VOIs evolved and co-circulated with VOCs and in some cases, contributed considerably to epidemic waves.

In North Africa (Fig. 1B and fig. S1A), B.1 lineages and Alpha dominated in the first and second wave of the pandemic and were replaced by Delta and Omicron in the third and fourth waves, respectively. Interestingly, the C.*36* and C.*36.3* sub-lineage dominated the epidemic in Egypt (~40% of reported infections) before July 2021 when it was replaced by Delta (*30*). Similarly, in Tunisia the first and second waves were associated with the B.1.160 lineage and were replaced by Delta during the country's third wave of infections. In southern Africa (Fig. 1C and fig. S1C), we see a similar pandemic profile with B.1 dominating the first wave, but instead of Alpha, Beta was responsible for the second wave, followed by Delta and Omicron. Another lineage that was flagged for close monitoring in the region was C.1.2, due to its mutational profile and predicted capacity for immune escape (*31*). However, the C.1.2 lineage did not cause many infections in the region as it was circulating at a time when Delta was dominant. In West Africa (Fig. 1D and fig. S1B), the B.1.*525* lineage caused a large proportion of infections in the second and third waves where it shared the pandemic landscape with the Alpha variant. As with other regions on the continent, these variants were later replaced by the Delta and then Omicron VOCs in successive waves. In Central Africa (Fig. 1E and fig. S1D), the B.1.620 lineage caused most of the infections between January and June 2021 (*32*) before systematically being replaced by Delta and then Omicron. Lastly, in East Africa (Fig. 1F and fig. S1E) the A.23.1 lineage dominated the second wave of infections in Uganda (*33*) and much of East Africa. In all of these regions, minor lineages such as B.1.*525*, C.*36* and A.23.1 were eventually replaced by VOCs that emerged in later waves.

Finally, we directly compared the official recorded cases in Africa with the ongoing SARS-CoV-2 genomic surveillance (GISAID date of access 2022-03-31) for a crude estimation of variants' contribution to cases. We observe that Delta was responsible for an epidemic wave between May and October 2021 (Fig. 1A) and had the greatest impact on the continent with almost 34.2% of overall infections in Africa possibly attributed to it. Beta was responsible for an epidemic wave at the end of 2020 and beginning of 2021 (Fig. 1A), with 13.3% of infections overall attributed to it. Notably, Alpha, despite being predominant in other parts of the world at the beginning of 2021, had only minimal significance in Africa, accounting for just 4.3% of infections. At the time of writing, the Omicron VOC had contributed to 21.6% of overall sequenced infections. At this time the Omicron wave was still unfolding globally and in Africa with the expansion of several sub-lineages (*34*), such that its full impact is yet to be determined. However, due to increased population immunity (*35*), from SARS-CoV-2 infection and vaccination (fig. S2), the impact of Omicron on mortality has been less in comparison to the other VOCs, as can be observed by the relatively low death rate in South Africa during the Omicron wave (*36*). The findings from mapping epidemiological numbers onto genomic surveillance data are reliable as far as the proportional scaling of genomic sampling across Africa with the size and timing of epidemic waves (fig. S3; b = 0.011, SE = 0.001, $p < 2 \times 10^{-16}$).

This comes with the obvious caveats that testing and reporting practices have varied widely across the continent, along with genomic surveillance volumes throughout the pandemic. Countries in Africa with reported data have tested in proportions from as little as 0.1 daily tests per million population to more than 1 000 tests per million (fig. S4). Some countries have consistently tested at high proportions, for example South Africa, Botswana, Morocco and Tunisia. Incidentally, these countries have also generally reported more cases per million, providing an indication that recorded low incidence in other parts of the continent has been an underestimate due to low testing rates. However, even for these countries, epidemic numbers are certainly under represented and under detected, given that in several timeframes, test positivity rates were still on the higher end, approaching or exceeding 20% (fig. S4), and as concluded by seroprevalence surveys and estimates of true infection burdens in Africa (*37*, *38*). Findings of attributing case numbers of variants must therefore be interpreted in context of this limitation but can nevertheless provide a qualitative overview of the spatial and temporal dynamics of VOCs in relation to epidemic progression in Africa.

The African regional- (table S1) and country-specific (table S2) NextStrain builds also clearly support the changing nature of the pandemic over time. From these builds we observe a strong association of B.1-like viruses circulating on the continent during the first wave. These "ancestral" lineages were subsequently replaced by the Alpha and Beta variants which dominated the pandemic landscape during the second wave, and were later replaced by the Delta and Omicron variants during the third and fourth waves.

Optimizing surveillance coverage in Africa

By mapping and comparing the locations of specimen

sampling laboratories to the sequencing laboratories, a number of aspects regarding the expansion of genomic surveillance on the continent became clear. First, even though several countries in Africa started sequencing SARS-CoV-2 in the first months of the pandemic, local sequencing capacity was initially limited. However, local sequencing capabilities slowly expanded over time, particularly after the emergence of VOCs (Fig. 2A). The fact that almost half of all SARS-CoV-2 sequencing in Africa was performed using the Oxford Nanopore technology (ONT), which is relatively low-cost compared to other sequencing technologies and better adapted to modest laboratory infrastructures, illustrates one component of how this rapid scale-up of local sequencing was achieved (fig. S5). Yet, to rely only on local sequencing would have thwarted the continent's chance at a reliable genomic surveillance program. At the time of writing, there were 52/55 countries in Africa with SARS-CoV-2 genomes deposited in GISAID, however, there were still 16 countries with no reported local sequencing capacity (Fig. 2A) and undoubtedly many with limited capacity to meet demand during pandemic waves.

To tackle this, three centers of excellence and various regional sequencing hubs were established to maximize resources available in a few countries to assist in genomic surveillance across the continent. This sequencing is done either as the sole source of viral genomes for those countries (e.g., Angola, South Sudan and Namibia) or concurrently with local efforts to increase capacity during resurgences (Fig. 2B). Sequencing is further supplemented by a number of countries utilizing facilities outside of Africa. Ultimately, a mix of strategies from local sequencing, collaborative resource sharing among African countries and sequencing with academic collaborators outside the continent helped close surveillance blind spots (Fig. 2C). Countries in sub-Saharan Africa, particularly in Southern and East Africa, most benefited from the regional sequencing networks, while countries in West and North Africa often partnered with collaborators outside of Africa.

The success of pathogen genomic surveillance programs relies on how representative it is of the epidemic under investigation. For SARS-CoV-2, this is often measured in terms of the percentage of reported cases sequenced and the regularity of sampling. African countries were positioned across a range of different combinations of overall proportion and frequency of genomic sampling (Fig. 2D). While the ultimate goal would be to optimize both of these parameters, a lower proportion of sampling can also be useful if frequency of sampling is maintained as high as possible. For instance, South Africa and Nigeria, who have both sequenced ~1% of cases overall, can be considered to have successful genomic surveillance programs on the basis that sampling is representative over time, and has enabled the timely detection of variants (Beta, Eta, Omicron).

Additionally, for genomic surveillance to be most useful for rapid public health response during a pandemic, sequencing would ideally be done in real-time or in a framework as close as possible to that. We show a general trend of decreasing sequencing turnaround time in Africa (fig. S6), particularly from a mean of 182 days between October to December 2020 to a mean of 50 days over the same period a year later, although this does come with several caveats. First, we measure sequencing turnaround time in the most accessible manner, which is by comparing the date of sampling of a specimen to the date its sequence was deposited in GISAID. Generally, the genomic data potentially informs the public health response more rapidly than reflected here, particularly when it comes to local outbreak investigations or variant detection. This analysis is also confounded by various factors such as country-to-country variation in these trends (fig. S7), delays in data sharing, and potential retrospective sequencing, particularly by countries joining sequencing efforts at later stages of the pandemic. The most critical caveat is the fact that sequencing from the most recently collected samples (e.g., over the last six months) may still be ongoing. The shortening duration between sampling and genomic data sharing is nevertheless a positive takeaway, given that this data also feeds into continental and global genomic monitoring networks. Overall, the continental average delay from specimen collection to sequencing submission is 87 days with 10 countries having an average turnaround time of less than 60 days and Botswana of less than 30 days (fig. S8).

Most importantly in the context of optimizing genomic surveillance, we found that the route taken to sequencing impacts the speed of data generation. Local sequencing has significantly faster sequencing turn-around times of the three frameworks we investigated (median of 51 days), followed by sequencing within regional sequencing networks in Africa (median of 93 days) and finally outsourced sequencing to countries outside Africa (median of 113 days) (Fig. 2E). This finding strongly supports the investments in local genomic surveillance, to generate timely and regular data for local and regional decision making. Finally, we show that it is beneficial in several ways for countries to undertake genomic surveillance through several sequencing laboratories, rather than centralizing efforts. For instance, we estimate strong correlations between the numbers of sequencing laboratories per country with the total number of genomes produced by that country (method, correlation value), the total number of *epiweeks* for which sequencing data was produced (method, correlation value), and importantly, sequencing turnaround time (method, correlation value) (Fig. 2F).

With the increase in sequencing capacity on the continent, a decrease in the time taken to detect new variants was observed. For example, the Beta variant was identified in

December 2020 in South Africa (*4*), but sampling and molecular clock analyses suggest the variant originated in September 2020. This three-month lag in detection means that a new variant, like Beta, has ample time to spread over a large geographic region prior to its detection. However, by the end of 2021, the time to detect a new variant was substantially improved. Phylogenetic and molecular clock analyses suggest that the Omicron variant originated around 9 October 2021 (95% Highest posterior density or HPD: 30 September - 20 October 2021) and the variant was described on 23rd November 2021 (*3*). Thus, Omicron was detected within ~5 weeks from origin compared to the Beta variant (~16 weeks) and the Alpha variant, detected in the UK (~10 weeks). More importantly, the time from sequence deposition to the WHO declaring the new variant a VOC was substantially shortened to 72 hours for the Omicron variant.

To interpret insights from the described genomic surveillance in Africa, it is important to understand the context of epidemiological reporting and sampling strategies utilized for sequencing on the continent (table S3). Most countries provided daily reports of newly recorded cases, while a few provided weekly and monthly reports. For most countries, surveillance was mainly focused on the major cities, suggesting potential cryptic circulation in rural areas. We find that at the onset of the pandemic, surveillance was focused on identification of imported cases from incoming travelers or local residents returning from various countries. As community transmissions began to emerge, the focus shifted toward regular surveillance and outbreak investigations. Together, these three strategies account for the vast majority of samples generated on the continent and analyzed here. As the pandemic progressed and vaccines were made available, some countries on the continent began to explore other sampling strategies such as reinfections, environmental samples such as waste water samples, and vaccine breakthrough cases to gain new insights into the evolutionary dynamics of SARS-CoV-2. The utility of sequencing for viral evolution tracking and VOC detection in the way described above is obviously also dependent on sampling proportions, especially within sampling for regular surveillance.

The speed of SARS-CoV-2 evolution has complicated sequencing efforts. Common methods of RNA sequencing include reverse transcription followed by double stranded DNA amplification using sequence-specific primer sets (*39*). Ongoing SARS-CoV-2 evolution has necessitated the continual evaluation and updating of these primer sets to ensure their sustained utility during genomic surveillance efforts. Here, we examined the current set of genomes to determine aspects of the sequencing that might be improved in the future. Many of the primer sets used were designed using viral sequences from the start of the pandemic and may require updating to keep pace with evolution. Indeed, the ARTIC primer sets are currently in version 4.1 (*40*). The Entebbe primer set was designed mid-2020 well into the first year of the epidemic and used an algorithm and design that accommodates evolution (*41*).

The effects of viral evolution on sequencing patterns can be seen with low median unspecified nucleotide (N)-values (a consequence of primer dropout or low coverage at that site) observed for the first 12 months of the epidemic with an increase from October 2020 (Fig. 3A). Additional challenges appear (indicated by increasing median N values) as the virus further evolved into Delta and Omicron lineages from January 2021 onward (Fig. 3A). Examining the role of sequencing technology, it appears that the two major technologies used (Illumina and ONT) have similar gap profiles (as measured by mean N count per genome) while Ion Torrent, MGI and Sanger show reduced mean N count per genome (Fig. 3B). Likely factors for this pattern are the primers used in sequencing, with primer choice playing a key role in the quantity of gaps (Fig. 3C). The mean N count per genome varied with viral lineage (Fig. 4D). There was a modest difference in mean N count per genome across the lineages. Lineages that returned no classification with Pangolin ("None") showed the highest mean N count, suggesting that high mean N count per genome was probably the basis for failed classification. The more recent lineages Delta (e.g., AY.39, AY.75) and Omicron (BA.1.1, BA.2) also showed higher mean N count per genome consistent with virus evolution impairing primer function. This pattern is further explored in fig. S9 with position of gaps showing an enrichment in the genome regions after position 19 000 with frequent gaps disrupting the spike coding region.

Phylogenetic insights into the rise and spread of variants of concern in Africa

During the first wave of infections in 2020 in Africa, as was the case globally, the majority of corresponding genomes were classified as PANGO B.1 (n=2 456) or B.1.1 viruses (n=1 329). Toward the end of 2020, more distinct viral lineages started to appear. The most important of which that impacted the African continent are: B.1.525 (n=797), B.1.1.318 (n=398) (*42*), B.1.1.418 (n=395), A.23.1 (n=358) (*15*, *29*, *31*, *33*), C.1 (n=446) (*29*), C.1.2 (n=300) (*31*), C.36 (n=305) (*30*, *43*), B.1.1.54 (n=287) (*15*, *29*, *31*, *33*), B.1.416 (n=272), B.1.177 (n=203), B.1.620 (n=138), and B.1.160 (n=61), (*32*) (fig. S10, A and B). Our discrete state phylogeographic inference from phylogenetic reconstruction of non-VOC African sequences and an equal number of external references revealed that African countries were primarily seeded by multiple introductions of viral lineages from abroad (mainly Europe) at the beginning of the pandemic. The observed pattern of non-VOC viral lineage movement then consistently shifted toward more intercontinental exchanges (fig. S10C). Mapping out the

spatial routes of dissemination shows that various countries in all subregions of the continent acted as sources of these viral lineages at one point or another (fig. S10D). While uneven testing rates and proportions of samples sequenced on the continent may have influenced these inferences (discussed below), the results presented here are in line with the fact that these most predominant non-VOC lineages in Africa, except B.1.177, emerged and circulated widely in different sub-regions (Fig. 1).

Similar to the pandemic globally, VOCs became increasingly important in Africa toward the end of 2020. The Alpha, Beta, Delta and Omicron variants demonstrate many similarities as well as differences in the way they spread on the continent. For all these VOCs, we observe large regional monophyletic transmission clusters in each of their phylogenetic reconstructions in Africa (fig. S11). This suggests an important extent of continental dissemination within Africa. Alpha and Beta were epidemiologically important in distinct regions of the continent with Alpha primarily circulating in West, North and most of Central Africa, Beta in southern and most of East Africa, and only substantially co-circulated in a few countries such as Angola, Kenya, Comoros, Burundi and Ghana (Fig. 1 and fig. S12). However, we may not have enough resolution in the geospatial data to know how much they were truly co-circulating throughout these countries, or whether there were regional outbreaks of Alpha and Beta within these countries. In Kenya, for example, Beta was detected more in coastal regions, and Alpha more inland (*26*, *44*). In contrast, Delta and Omicron variants sequentially dominated the majority of infections on the entire continent shortly after their emergence (Fig. 4A and fig. S12).

The Alpha variant was first identified in December 2020 in the UK and has since spread globally. In Africa, Alpha was detected in 43 countries with evidence of community transmission, based on phylogenetic clustering, in many countries including Ghana, Nigeria, Kenya, Gabon and Angola (fig. S11). Discrete state maximum likelihood reconstruction from a globally case-sensitive genomic subsampling inferred at least 80 introductions (95% CI: 78 - 82) into Africa with the bulk of imports attributed to the US (>47%) and the UK (>25%) (Fig. 4B). Only 1% of imports into any particular African country were attributed to another African nation. Phylogeographic reconstruction enriched in African sequences revealed that of those, >85% of the intercontinental Alpha exchanges in Africa originated from West African countries (Fig. 4C). This occurred in spite of initial importations of the Alpha variant from Europe into all regions of the continent (fig. S13B), but is in line with Alpha having dominated circulation mostly in West Africa (fig. S12). In countries where Alpha was introduced but did not grow and cause an expansion of cases, this can be explained by competition with the already established Beta variant, which simultaneously circulated. The characteristics of multiple introductions of Alpha intro Africa and between African countries is similar to the spread of Alpha documented in the UK, Scotland and Ireland (*45–47*).

The second VOC, Beta, was identified in December 2020 in South Africa (*4*). However, sampling and molecular clock analyses suggest that the variant originated around September 2020 (fig. S11). At the end of 2020 and beginning of 2021, Beta was driving a second wave of infection in South Africa and quickly spread to other countries within the region. The concurrent introductions and spread of Alpha and other variants (Eta, A.23.1) in other regions of the continent may have reduced the Beta variant's initial growth, limiting its spread to largely southern Africa, and to a lesser extent the East Africa region. Beta spread to at least 114 countries globally, including 37 countries and territories in Africa. For this variant, viral circulation and geographical exchanges occurred predominantly within the continent. Indeed, phylogeographic reconstruction from a globally case-sensitive sampling revealed that of the 810 (95% CI: 803 - 818) inferred introductions of the Beta variant into African countries, only 110 (95% CI: 105 - 115; 13%) were attributed to sources outside the continent (fig. S13C), while more than half of introductions were attributed to South Africa (63%) (Fig. 4C). This is in line with expectations as the variant originated in South Africa. Beyond southern Africa, most of the introductions back into the continent were attributed to France and other EU countries into the French overseas territories, Mayotte and Reunion, and other Francophone African countries. Africa-focused phylogeographic analysis revealed a similar spatial pattern showing southern countries as substantial sources of the variant, followed in small numbers by countries in East Africa (Fig. 4C).

The fourth VOC observed was Delta (*13*), which rose to prominence in April 2021 in India, where it fuelled an explosive second wave. Since its emergence, Delta was detected in >170 countries, including 37 African countries and territories (fig. S11). Our global case-sensitive subsampled analysis infers at least 100 (95% CI: 93 - 106) introductions of the Delta variant into Africa, with the bulk attributed to India (~72%), mainland Europe (~8%), the UK (~5%), and the US (~2.5%). Viral introductions of Delta also occurred from one African country to others, in 7% of inferred introductions. From our Africa-focused phylogeographic inferences, we infer that viral dissemination of Delta within Africa was not restricted to or dominated by any particular region unlike Alpha and Beta, but rather spread across the entire continent (Fig. 4C). Following introductions from Asia in the middle of 2021, Delta rapidly replaced the other circulating variants (Fig. 4A). For example, in southern African countries, the Delta variant rapidly displaced Beta and by June-2021 was circulating at very high (>90%) frequencies (*48*).

The latest VOC, Omicron, was identified and characterized in November 2021, in southern Africa (*3*). At the time of writing, the variant has been detected and caused waves of infections in >160 countries including 39 African countries and two overseas territories (fig. S11). Due to the genetic distance between them and their sequential epidemic expansion globally (rather than simultaneous), phylogenies were reconstructed separately for Omicron BA.1 and BA.2. Our discrete ancestral state reconstruction from a global case-sensitive sampling for Omicron BA.1 infers at least 55 (95% CI: 47 - 62) viral exports of BA.1 out of various African countries, of which 31 (95% CI: 25 - 36) were toward Europe and 8 (95% CI: 6 - 10) toward North America (Fig. 4B). Following explosive expansion of Omicron around the world, we inferred even more reintroductions of the variant back into Africa, at least 69 (95% CI: 60 - 78) from Europe and 102 (95% CI: 92 - 112) from North America (Fig. 4B). From our Africa-focused phylogeographic reconstructions, we determine that, as with Delta, routes of dissemination of this variant involved all regions of the continent spatially (Fig. 4C). Yet, ~75% of all BA.1 viral movement volume in Africa happened between southern African countries, likely due to rapid epidemic expansion in the region soon after its detection (*3*). Omicron BA.2's reach in Africa was limited at the time of writing, with only 3 260 sequences from 19 countries attributed to BA.2 on GISAID (Date of access: 2022-03-31) (15% of all Omicron sequences from Africa). Our discrete ancestral state reconstruction from a global case-sensitive sampling for Omicron BA.2 infers at least 68 (95% CI: 53 - 84) viral exports out of African countries, of which the majority were toward Europe (~88%) (Fig. 4B). We also infer at least 99 (95% CI: 87 - 109) separate introduction or reintroduction events of BA.2 back into African countries, of which ~65% are from Europe and ~30% from Asia, primarily from India (Fig. 4B). This is consistent with India having experienced one of the earliest large BA.2 waves globally. In the context of global incidence of BA.2, this case-sensitive phylogeographic analysis revealed that only 0.01% of viral movements of this lineage globally happened from one African country to another. Our Africa-focused analysis inferred a similar pattern of BA.2 spatial diffusion within African to BA.1 (Fig. 4C). However, given that this accounted for such a small percentage of global BA.2 movements, BA.2 diffusion from one African country to another is unlikely to have had a significant impact on epidemiological expansion, compared to introductions from Asia, Europe or North America.

Globally, dissemination of the SARS-CoV-2 virus throughout the pandemic was intricately linked with human mobility patterns (*49–53*). To determine the validity of the VOC movement patterns that we infer into and within the Africa continent in this study, we compared viral import and export events to and from South Africa with travel to the country. In December 2020, the UK accounted for the 5th highest number of passengers entering South Africa, while other countries with the top 9 sources of travellers were all neighboring countries in southern Africa (fig. S14A). Considering that incidence of the Alpha variant was insignificant in the region, this supports our inference of the UK contributing 60% of Alpha introductions to South Africa (fig. S15A). In March 2021, the US, Germany, the UK and India were among the top 12 sources of travellers to South Africa behind 8 African countries (fig. S14B). During this time of Delta dissemination globally, we infer that ~90% of introductions of Delta into South Africa originated in the UK, the US and India (fig. S15B). At the end of 2021, most introductions or re-introductions of Omicron to the country came from the UK, the US or Botswana, corresponding to locations of both high Omicron incidence at the time, and high numbers of passengers to South Africa (figs. S14C and S15C). These travel patterns also fit the findings that ~89%,~70% and ~75% of Beta, Delta and Omicron exports respectively from South Africa to other African countries were directed to locations of southern Africa (figs. S14, D and E, and S15, D and E).

Discussion, limitations and conclusions

By April 2020, a total of 20 African countries were able to sequence the virus within their own borders. This was largely made possible by other pre-existing sequencing efforts on the continent focused on other human pathogens (e.g., HIV, TB, Ebola and H1N1). However, these efforts were quickly limited by global supply chain issues and in many countries sequencing efforts dramatically slowed down or stopped toward the end of 2020. In order to facilitate more sequencing on the continent over the course of the past year (April 2021 - March 2022) the Africa CDC and partners invested heavily to support genomic surveillance on the continent. This included the transfer of 24 new sequencing platforms (including MinIon, GridIon, MiSeq and NextSeq), the distribution of reagents and flow cells to support the sequencing of 100 000 positive samples, the training of >230 students and technicians in wet laboratory and bioinformatic techniques and additional grants to support 10 regional sequencing hubs. This investment has started bearing fruit and should be intensified as the virus continues to evolve, requiring the adaptation of methodologies locally on the continent to keep pace with the emergence of variants. The continued development of sequencing protocols in Africa is of crucial importance (*41*, *54*, *55*) given the number of variants and lineages that emerged in, and were introduced to, the continent. In Northern Africa, the SARS-CoV-2 pandemic was caused by waves of infections that were similar to those seen in Europe (first wave = B.1 descendants, second wave = Alpha, third wave = Delta and forth wave = Omicron), in southern Africa the pattern was similar but with a Beta wave instead of an Alpha one. In East Africa, the pandemic was more complex, involving both

Alpha and Beta as well as its own lineage A.23.1 before the arrival of Delta and Omicron. Central Africa experienced epidemic patterns sometimes mirroring East Africa and other times southern Africa. In West Africa, Eta made a significant contribution to both a second wave (together with alpha) and a third wave (together with Delta). The factors that resulted in these regional differences are not clear but could be due to differences in human mobility, founder effects, competition between lineages or the immunity induced by earlier waves in a region.

Public health benefits of such broadly inclusive genomic surveillance are manifold. The most prominent insight from this expanded genomic surveillance in Africa has been an early warning capacity for the world following the detection of new lineages and variants, most recently relevant in the detection of Omicron BA.1, BA.2, BA.3, BA.4 and BA.5 subvariants (*3*, *4*, *34*). Furthermore, the reporting of local SARS-CoV-2 sequences made the epidemic more immediate to the Ministries of Health from the reporting African countries. It became clear early on that the viral evolution is global and the transmission of the virus is extremely rapid which guided mitigation strategies. The generation and the availability of local sequences also validated local diagnostics and allowed investigators to determine if nucleic acid based diagnostics in use could still detect local variants. The detection of SARS-CoV-2 in returning travellers and truck drivers indicated routes that the virus might be using to enter a country and guided early efforts to slow the virus entry and gain time to establish vaccination plans. Later the difficulty of stopping the virus at borders combined with the data that the variants were already in community circulation allowed public health officials to focus efforts and limited resources on vaccination rather than on border controls. The detection and reporting of the more recent lineages with enhanced transmission (i.e., Omicron) and the ability to bypass existing immunity is important information and an early alert to the public health officials globally that the epidemic was still proceeding. As the pandemic progresses in an evolving global context, we provide evidence that with each new variant, transmission dynamics are changing and the use of sequencing with phylogenetics could potentially alter decisions of public health measures. For example, the demonstrated shift away from regional dynamics of Alpha and Beta toward more global patterns with Delta and Omicron can provide insights to public health officials as they anticipate epidemic developments locally. With Omicron it became clear that although the variant expanded first in Africa, the continent ultimately had a minimal role in global dissemination, and continental expansion beyond southern Africa was most influenced by external introductions, in contrast to the Beta variant. All of these public health benefits to sequencing SARS-CoV-2 is primarily amplified, as we show in this study, if the sequencing can be conducted locally within a country, which strongly supports the continued investment into pathogen sequencing on the continent.

In spite of the recent successful expansion of genomics surveillance in Africa, additional work remains necessary. Even with the Africa CDC - Africa PGI's and other investments, there are still 16 countries with no sequencing capacity within their own borders. These countries' only option is to send samples to continental sequencing hubs or to centers outside of the continent, which increases the turnaround times and limits the utility of genomic surveillance for public health decision making. Secondly, not all countries are willing to share data openly in a timely fashion for fear of being subject to travel bans or restrictions which could bring substantial economic harm. Such hesitancy has obvious potential ramifications for the future of genomic surveillance on the continent. Furthermore, with the expansion of sequencing on the continent there is a growing need for more bioinformatics support and knowledge to allow investigators to analyze and report their data in a reasonable timeframe that makes it useful for public health response. It is also clear the SARS-CoV-2 sequencing primers are not a static development and may require updating as the virus evolves. A number of research groups have been addressing the SARS-CoV-2 sequencing primer questions. Issues of gaps in the genomes due to missing amplicons have been discussed (*56*, *57*). The ARTIC primer set has gone through a number of revisions to accommodate virus evolution (*39*, *40*). Additional longer amplicon methods have been published (*58–60*) including methods to use a subset of ARTIC primers (*61*).

The patterns we describe here are of course limited to reported cases, and applies to both the phylogeographic as well as the epidemiology inferences. As such, the results need to be interpreted with these limitations in mind. Our primary phylogeographic inference relied on a sampling strategy considering all high quality African sequences and an equal number of external references. Though this strategy has the advantage of placing all African sequences in a phylogenetic context, it introduces a bias when applied to discrete ancestral state reconstruction as more internal nodes are inferred to be from Africa. To address this we performed an even sampling of global cases, based on reported case counts through time, to compare against our over sampled inference. The even sampling approach has the benefit that the discrete ancestral state reconstruction is not biased by uneven sampling. Comparing the two there are obvious differences, most notably that the number of inferred introductions into Africa is proportional to sampling proportions (fig. S16), as we no longer consider all African sequences but just a small subset against a global sample. However, inferences from the two approaches correspond well with one another. For example, considering Alpha we still observed the vast majority of

introductions into Africa to originate from Western Europe. Patterns of dissemination within Africa are more robustly comparable between the two, for instance that countries in West Africa were the biggest source of Alpha within the continent. High concordance between the two inference methods were also observed for other VOCs for dispersal routes within Africa which gives us confidence in the inferred patterns we observe here. Although we represent an inference based on over sampling and case sensitive sampling, it is currently not possible to explore how under sampling affects the phylogeographic reconstruction due to uneven testing rates. Additionally, the robustness of the phylogeographic inference can also be affected by the underlying methodology used. Broad consensus would favor the use of Bayesian methods for phylogeographic reconstruction, which is often considered to be the "gold standard" in the field. The main drawbacks of Bayesian methods are that they can only be applied to a relatively small number of sequences at a time (<1,000) and are extremely computationally and time intensive. Given the explosion of sequence data over the past two years, the scientific community will have to adapt or put forth new analytical methods to fully capitalize on the global sequencing efforts for SARS-CoV-2.

Despite our best attempts to consider and minimize genomic sampling bias, the accuracy of the resulting phylogenetic inferences is limited by the available epidemiological and genomic data, leading to unaccounted biases in the estimates of viral movements. This includes limited testing and subsequent sequencing in many African countries. Although the percentage of reported cases sequenced in African countries (0.01 - 10%, mean = 1.27%) is not far from global figures (0.01-16%, mean = 1.31%), testing rates and infection-to-detection ratios in Africa were some of the lowest globally (*38*, *62*). Together with estimates of excess mortality being as much as 20-fold more than the reported numbers in African countries (*63*), these are strong indications of undetected and underreported epidemic sizes in Africa, leading to undersampling of genomic data (*62*) and thus underestimates of viral exchange inferences in our study. Some countries with no publicly available SARS-CoV-2 sequences are by definition completely missing in our inference. This in turn means that inferred routes of viral transmission within Africa could be missing important intermediate locations, although this is potentially true around the world. Nevertheless, we believe that the viral movement inferences that we discuss in this study provide a likely qualitative description of the patterns of SARS-CoV-2 migration into, out of, and within Africa.

Finally, we should also mention uneven sequencing and reporting standards across the different laboratories on the continent - and globally, for that matter. Different groups use different measures for what constitutes a high quality sequence (e.g., 70% vs 80% sequence coverage) or using different sequencing depth coverage. This lack of standardization globally complicates the direct comparison of sequences that may have been submitted to GISIAD using different criteria further biasing any inference. Given the sheer size of SARS-CoV-2 sequencing, with ~10 million whole genome sequences shared on the GISAID database (31st March 2022), there is an urgent need for global standards with regards to sequence quality and associated metadata.

In conclusion, Africa needs to continue expanding genomic sequencing technologies on the continent in conjunction with diagnostics capabilities. This holds true not just for SARS-CoV-2 but for other emerging or re-emerging pathogens on the continent. For example, WHO announced in February 2022 the re-emergence of wild polio in Africa, while sporadic influenza H1N1, measles and Ebola outbreaks continue to occur on the continent. The Africa CDC has estimated that over 200 pathogen outbreaks are reported across the continent every year. Beyond the current pandemic, continued investment in diagnostic and sequencing capacity for these pathogens could serve the public health of the continent well into the 21st century.

Methods and methods

Ethics statement

This project relied on sequence data and associated metadata publicly shared by the GISAID data repository and adhere to the terms and conditions laid out by GISAID (*16*). The African samples processed in this study were obtained anonymously from material exceeding the routine diagnosis of SARS-CoV-2 in African public and private health laboratories. Individual institutional review board (IRB) references or material transfer agreements (MTAs) for countries are listed below.

Angola - (MTA - CON8260), Botswana - Genomic surveillance in Botswana was approved by the Health Research and Development Committee (Protocol HPDME 13/18/1), Egypt - Surveillance in Egypt was approved by the Research Ethics Committee of the National Research Centre (Egypt) (protocol number 14 155, dated March 22, 2020), Kenya - samples were collected under the Ministry of Health protocols as part of the national COVID-19 public health response. The whole genome sequencing study protocol was reviewed and approved by the Scientific and Ethics Review Committee (SERU) at Kenya Medical Research Institute (KEMRI), Nairobi, Kenya (SERU protocol #4035), Nigeria – (NHREC/01/01/2007), Mali - study of the sequence of SARS-CoV-2 isolates in Mali - Letter of Ethical Committee (N0-2020 /201/CE/FMPOS/FAPH of 09/17/2020), Mozambique - (MTA - CON7800), Malawi - (MTA - CON8265), South Africa - The use of South African samples for sequencing and genomic surveillance were approved by University of KwaZulu-Natal Biomedical Research Ethics Committee (ref. BREC/00001510/2020); the University of the Witwatersrand Human Research Ethics Committee

(HREC) (ref. M180832); Stellenbosch University HREC (ref. N20/04/008_COVID-19); the University of the Free State Research Ethics Committee (ref. UFS-HSD2020/1860/2710) and the University of Cape Town HREC (ref. 383/2020), Tunisia - for sequences derived from sampling in Tunisia, all patients provided their informed consent to use their samples for sequencing of the viral genomes. The ethical agreement was provided to the research project ADAGE (PRFCOVID19GP2) by the Committee of protection of persons (Tunisian Ministry of Health) under the reference (CPP SUD N 0265/2020), Uganda - The use of samples and sequences from Uganda were approved by the Uganda Virus Research Institute - Research and Ethics Committee UVRI-REC Federalwide Assurance [FWA] FWA No. 00001354, study reference - GC/127/20/04/771 and by the Uganda National Council for Science and Technology, reference number - HS936ES) and Zimbabwe (MTA - CON8271).

Epidemiological and genomic data dynamics

We analyzed trends in daily numbers of cases of SARS-CoV-2 in Africa up to 31st March 2022 from publicly released data provided by the Our World in Data repository for the continent of Africa (https://github.com/owid/covid-19-data/tree/master/public/data) as a whole and for individual countries (*2*). To provide a comparable view of epidemiological dynamics over time in various countries, the variable under primary consideration for Fig. 1 was 'new cases per million (smoothed)'. To calculate the genomic sampling proportion and frequency for each country for Fig. 2, the total number of recorded cases at 31st March was considered, as well as the total length of time for which each country has recorded cases of SARS-CoV-2.

Genomic metadata was downloaded for all African entries on GISAID for the same time period (date of access: 31st March 2022). From this, information extracted from all entries for this study included: date of sampling, country of sampling, viral lineage and clade, originating laboratory, sequencing laboratory, and date of submission to the GISAID database. The geographical locations of the originating and sequencing laboratories were manually curated. Sequences originating and sequenced in the same country were defined as locally sequenced, irrespective of specific laboratory or finer location. Sequences originating in one African country and sequenced in another were defined as sequenced within regional sequencing networks. Sequences sequenced in a location not within Africa were labeled as sequenced outside Africa. Sequencing turnaround time was defined as the number of days elapsed from specimen collection to sequence submission to GISAID. Sequencing technology information for all African entries was also downloaded from GISAID on 31st March 2022.

Primer choice and sequencing outcomes

All SARS-CoV-2 genomes from African countries were retrieved from GISAID (*16*) for submission dates from 1 December 2019 to 31st March 2022 yielding 100 470 entries. Associated metadata for the entries were also retrieved, including collection date, submission date, country, viral strain and sequencing technology. Data on the primers used for the sequencing were requested from investigators and yielded primer data for 13 973 of the entries (~13%). The total N (bases with low sequence depth) per genome were counted, results from which were then used for genome quality analysis and visualization. Gap locations in the genomes were mapped and visualized compared to the original Wuhan strain (*64*).

Phylogenetic investigation

All African sequences on the GISAID sequence database (*16*) were downloaded on the 31st of March 2022 (n=100 470). Of this, Alpha accounted for 3 851 sequences, Beta accounted for 14 548 sequences, Delta accounted for 35 027 sequences, Omicron for 21 708, while 25 336 sequences were classified as none-VOCs. Prior to any phylogenetic inference we performed some quality assessment on the sequences to exclude incomplete or problematic sequences as well as sequences lacking complete metadata. Briefly, all African sequences were passed through the NextClade analysis pipeline (*65*) in order to identify and exclude: (*i*) sequences missing >10% of the SARS-CoV-2 genome, (*ii*) sequences that deviate by >70 nucleotides from the Wuhan reference strain, (*iii*) sequences with >10 ambiguous bases, (*iv*) clustered mutations, and (*v*) sequences flagged with private mutations by NextClade. Additionally, Omicron variants were screened for traces of viral recombination with RDP5.23 (*66*) using default settings and a p-value of ≤0.05 as evidence of recombination. A large number of sequences were removed (n=57 421) with incomplete sequences (<90% genome coverage) being the biggest contributor. This produced a final African dataset of 43 049 high quality African sequences. Due to the sheer size of the dataset we opted to perform independent phylogenetic inferences on the main VOCs (Alpha, Beta, Delta and Omicron BA.1 and BA.2) that have spread on the African continent, as well as a separate inference for all non-VOC SARS-CoV-2 sequences.

In order to evaluate the spread of the virus on the African continent we aligned the African datasets against a large number of globally representative sequences from around the world. Due to the oversampling of some variants or lineages we performed a random down sampling while retaining the oldest two known variants from each country. Reference sequences were respectively aligned with their African counterparts independently with NextAlign (*65*). Each of the alignments were then used to infer maximum likelihood (ML) tree topologies in FastTree v 2.0 (*67*) using the General Time Reversible (GTR) model of nucleotide substitution and

a total of 100 bootstrap replicates (*68*). The resulting ML tree topologies were first inspected in TempEst (*69*) to identify any sequences that deviate more than 0.0001 from the residual mean. Following the removal of potential outliers in R with the ape package (*70*), the resulting ML-trees were then transformed into time calibrated phylogenies in TreeTime (*71*) by applying a rate of 8x10e-4 substitution per site per year (*72*) in order to transform the branches into units of calendar time. Time calibrated trees were then visualized along with associated metadata in R using ggtree (*73*) and other packages.

We performed a basic viral dispersal analysis for each of the VOCs (excluding Gamma), as well as for the non-VOC dataset. Briefly, a migration model was fitted to each of the time calibrated tree topologies in TreeTime, mapping the country location of sampled sequences to the external tips of the trees. The *mugration* model of TreeTime also infer the most likely location for internal nodes in the trees. Using a custom python script we could then count the number of state changes by iterating over each phylogeny from the root to the external tips. We count state changes when an internal node transitions from one country to a different country in the resulting child-node or tip(s). The timing of transition events is then recorded which serve as the estimated import or export event. To infer some confidence around these estimates, we performed ten replicates for each of the dataset by random selection from the 100 bootstrap trees. Due to the high uncertainty in the inferred locations for deep internal nodes in the trees we truncated state changes to the earliest date of sampling in each dataset. All data analytics were performed using custom python and R scripts and results visualized using the ggplot libraries (*74*). Such phylogeographic methods are always subject to uneven sampling through time (i.e., over the course of the pandemic) and through space (by sampling location). To address this we have performed a case sensitive analysis to investigate the effects of oversampling African locations on the inferred number of viral introductions. Furthermore, in a previous analysis (*15*) we performed a sensitivity analysis to address some of these issues and found no substantial variations in estimates.

Case sensitive phylogeographic inference

To address the potential over sampling of African sequences relative to global reference in the above mentioned analyses we performed another phylogeographic inference on subsamples based on global case counts to try and eliminate oversampling bias in our inference. To this end, we considered all high quality sequences for each of the VOCs (Alpha, Beta, Delta and Omicron BA.1 and BA.2) globally over the same sampling period (till 31st of March 2022). We used subsampler (https://github.com/andersonbrito/subsampler) to generate subsamples for each variant based on globally reported cases. In short, subsampler uses a case count matrix of daily cases, along with the fasta sequences and GISAID associated metadata to sample a user defined number of sequences. For each VOC and for BA.1 and BA.2 we performed 10 samplings using different number seeds in order to sample datasets of ~20 000. Once again, sampled sequences were screened for viral recombination as described above and sequences with signs of recombination were removed. Subsampler has the added advantage that it disregards poor quality sequences (e.g., <90% coverage) and sequences with missing metadata (e.g., exact date of sampling). Each dataset was then subjected to the same analytical pipeline as mentioned above to infer the viral transitions between Africa and the rest of the world.

Regional and country specific NextStrain builds

In order to investigate more granular changes in lineage dynamics within a specific country or region in Africa we utilized the NextStrain pipeline (https://github.com/nextstrain/ncov) to generate the regional and country-specific builds for African countries (*75*). First, all sequence data and metadata were retrieved from the GISAID sequence database and filtered for Africa based on the 'region' tab, for inclusion in regional- and country-specific African builds. For country-specific builds ~4 000 sequences from a given country were randomly selected and analyzed against ~1 000 randomly selected sequences from the Africa 'nextregions' records that do not match the focal country of interest. For region specific (e.g., West Africa), ~4 000 sequences from the focal region are selected at random and analyzed against ~1 000 randomly selected sequences from the Africa 'nextregions' records that do not match the focal region of interest. The methodological pipeline for NextStrain is well documented and performs all analyses within one workflow, including filtering of sequences, alignment, tree inference, molecular clock and ancestral state reconstruction. For more information please visit, https://docs.nextstrain.org/en/latest/index.html.

All region- and country-specific builds are regularly updated to keep track of the evolving pandemic on the continent. All builds are publicly available under the links provided in tables S1 and S2 as well as on the NextStrain webpage (https://nextstrain.org/sars-cov-2/#datasets).

REFERENCES AND NOTES

1. Q. Li, X. Guan, P. Wu, X. Wang, L. Zhou, Y. Tong, R. Ren, K. S. M. Leung, E. H. Y. Lau, J. Y. Wong, X. Xing, N. Xiang, Y. Wu, C. Li, Q. Chen, D. Li, T. Liu, J. Zhao, M. Liu, W. Tu, C. Chen, L. Jin, R. Yang, Q. Wang, S. Zhou, R. Wang, H. Liu, Y. Luo, Y. Liu, G. Shao, H. Li, Z. Tao, Y. Yang, Z. Deng, B. Liu, Z. Ma, Y. Zhang, G. Shi, T. T. Y. Lam, J. T. Wu, G. F. Gao, B. J. Cowling, B. Yang, G. M. Leung, Z. Feng, Early transmission dynamics in Wuhan, China, of novel coronavirus-infected pneumonia. *N. Engl. J. Med.* **382**, 1199–1207 (2020). doi:10.1056/NEJMoa2001316 Medline
2. J. Hasell, E. Mathieu, D. Beltekian, B. Macdonald, C. Giattino, E. Ortiz-Ospina, M. Roser, H. Ritchie, A cross-country database of COVID-19 testing. *Sci. Data* **7**, 345 (2020). doi:10.1038/s41597-020-00688-8 Medline
3. R. Viana, S. Moyo, D. G. Amoako, H. Tegally, C. Scheepers, C. L. Althaus, U. J.

Anyaneji, P. A. Bester, M. F. Boni, M. Chand, W. T. Choga, R. Colquhoun, M. Davids, K. Deforche, D. Doolabh, L. du Plessis, S. Engelbrecht, J. Everatt, J. Giandhari, M. Giovanetti, D. Hardie, V. Hill, N. Y. Hsiao, A. Iranzadeh, A. Ismail, C. Joseph, R. Joseph, L. Koopile, S. L. Kosakovsky Pond, M. U. G. Kraemer, L. Kuate-Lere, O. Laguda-Akingba, O. Lesetedi-Mafoko, R. J. Lessells, S. Lockman, A. G. Lucaci, A. Maharaj, B. Mahlangu, T. Maponga, K. Mahlakwane, Z. Makatini, G. Marais, D. Maruapula, K. Masupu, M. Matshaba, S. Mayaphi, N. Mbhele, M. B. Mbulawa, A. Mendes, K. Mlisana, A. Mnguni, T. Mohale, M. Moir, K. Moruisi, M. Mosepele, G. Motsatsi, M. S. Motswaledi, T. Mphoyakgosi, N. Msomi, P. N. Mwangi, Y. Naidoo, N. Ntuli, M. Nyaga, L. Olubayo, S. Pillay, B. Radibe, Y. Ramphal, U. Ramphal, J. E. San, L. Scott, R. Shapiro, L. Singh, P. Smith-Lawrence, W. Stevens, A. Strydom, K. Subramoney, N. Tebeila, D. Tshiabuila, J. Tsui, S. van Wyk, S. Weaver, C. K. Wibmer, E. Wilkinson, N. Wolter, A. E. Zarebski, B. Zuze, D. Goedhals, W. Preiser, F. Treurnicht, M. Venter, C. Williamson, O. G. Pybus, J. Bhiman, A. Glass, D. P. Martin, A. Rambaut, S. Gaseitsiwe, A. von Gottberg, T. de Oliveira, Rapid epidemic expansion of the SARS-CoV-2 Omicron variant in southern Africa. *Nature* **603**, 679–686 (2022). doi:10.1038/s41586-022-04411-y Medline

4. H. Tegally, E. Wilkinson, M. Giovanetti, A. Iranzadeh, V. Fonseca, J. Giandhari, D. Doolabh, S. Pillay, E. J. San, N. Msomi, K. Mlisana, A. von Gottberg, S. Walaza, M. Allam, A. Ismail, T. Mohale, A. J. Glass, S. Engelbrecht, G. Van Zyl, W. Preiser, F. Petruccione, A. Sigal, D. Hardie, G. Marais, N. Y. Hsiao, S. Korsman, M. A. Davies, L. Tyers, I. Mudau, D. York, C. Maslo, D. Goedhals, S. Abrahams, O. Laguda-Akingba, A. Alisoltani-Dehkordi, A. Godzik, C. K. Wibmer, B. T. Sewell, J. Lourenço, L. C. J. Alcantara, S. L. Kosakovsky Pond, S. Weaver, D. Martin, R. J. Lessells, J. N. Bhiman, C. Williamson, T. de Oliveira, Detection of a SARS-CoV-2 variant of concern in South Africa. *Nature* **592**, 438–443 (2021). doi:10.1038/s41586-021-03402-9 Medline
5. D. P. Martin, S. Weaver, H. Tegally, J. E. San, S. D. Shank, E. Wilkinson, A. G. Lucaci, J. Giandhari, S. Naidoo, Y. Pillay, L. Singh, R. J. Lessells, R. K. Gupta, J. O. Wertheim, A. Nekturenko, B. Murrell, G. W. Harkins, P. Lemey, O. A. MacLean, D. L. Robertson, T. de Oliveira, S. L. Kosakovsky Pond, NGS-SA, COVID-19 Genomics UK (COG-UK), The emergence and ongoing convergent evolution of the SARS-CoV-2 N501Y lineages. *Cell* **184**, 5189–5200.e7 (2021). doi:10.1016/j.cell.2021.09.003 Medline
6. F. Campbell, B. Archer, H. Laurenson-Schafer, Y. Jinnai, F. Konings, N. Batra, B. Pavlin, K. Vandemaele, M. D. Van Kerkhove, T. Jombart, O. Morgan, O. le Polain de Waroux, Increased transmissibility and global spread of SARS-CoV-2 variants of concern as at June 2021. *Euro Surveill.* **26**, (2021). doi:10.2807/1560-7917.ES.2021.26.24.2100509 Medline
7. B. Korber, W. M. Fischer, S. Gnanakaran, H. Yoon, J. Theiler, W. Abfalterer, N. Hengartner, E. E. Giorgi, T. Bhattacharya, B. Foley, K. M. Hastie, M. D. Parker, D. G. Partridge, C. M. Evans, T. M. Freeman, T. I. de Silva, C. McDanal, L. G. Perez, H. Tang, A. Moon-Walker, S. P. Whelan, C. C. LaBranche, E. O. Saphire, D. C. Montefiori, A. Angyal, R. L. Brown, L. Carrilero, L. R. Green, D. C. Groves, K. J. Johnson, A. J. Keeley, B. B. Lindsey, P. J. Parsons, M. Raza, S. Rowland-Jones, N. Smith, R. M. Tucker, D. Wang, M. D. Wyles, Sheffield COVID-19 Genomics Group, Tracking changes in SARS-CoV-2 spike: Evidence that D614G increases infectivity of the COVID-19 virus. *Cell* **182**, 812–827.e19 (2020). doi:10.1016/j.cell.2020.06.043 Medline
8. E. Hacisuleyman, C. Hale, Y. Saito, N. E. Blachere, M. Bergh, E. G. Conlon, D. J. Schaefer-Babajew, J. DaSilva, F. Muecksch, C. Gaebler, R. Lifton, M. C. Nussenzweig, T. Hatziioannou, P. D. Bieniasz, R. B. Darnell, Vaccine breakthrough infections with SARS-CoV-2 variants. *N. Engl. J. Med.* **384**, 2212–2218 (2021). doi:10.1056/NEJMoa2105000 Medline
9. D. Planas, D. Veyer, A. Baidaliuk, I. Staropoli, F. Guivel-Benhassine, M. M. Rajah, C. Planchais, F. Porrot, N. Robillard, J. Puech, M. Prot, F. Gallais, P. Gantner, A. Velay, J. Le Guen, N. Kassis-Chikhani, D. Edriss, L. Belec, A. Seve, L. Courtellemont, H. Péré, L. Hocqueloux, S. Fafi-Kremer, T. Prazuck, H. Mouquet, T. Bruel, E. Simon-Lorière, F. A. Rey, O. Schwartz, Reduced sensitivity of SARS-CoV-2 variant Delta to antibody neutralization. *Nature* **596**, 276–280 (2021). doi:10.1038/s41586-021-03777-9 Medline
10. S. Yue, Z. Li, Y. Lin, Y. Yang, M. Yuan, Z. Pan, L. Hu, L. Gao, J. Zhou, J. Tang, Y. Wang, Q. Tian, Y. Hao, J. Wang, Q. Huang, L. Xu, B. Zhu, P. Liu, K. Deng, L. Wang, L. Ye, X. Chen, Sensitivity of SARS-CoV-2 variants to neutralization by convalescent sera and a VH3-30 monoclonal antibody. *Front. Immunol.* **12**, 751584 (2021). doi:10.3389/fimmu.2021.751584 Medline
11. S. Cele, I. Gazy, L. Jackson, S.-H. Hwa, H. Tegally, G. Lustig, J. Giandhari, S. Pillay, E. Wilkinson, Y. Naidoo, F. Karim, Y. Ganga, K. Khan, M. Bernstein, A. B. Balazs, B. I. Gosnell, W. Hanekom, M. S. Moosa, R. J. Lessells, T. de Oliveira, A. Sigal, Network for Genomic Surveillance in South Africa, COMMIT-KZN Team, Escape of SARS-CoV-2 501Y.V2 from neutralization by convalescent plasma. *Nature* **593**, 142–146 (2021). doi:10.1038/s41586-021-03471-w Medline
12. B. Meng, S. A. Kemp, G. Papa, R. Datir, I. A. T. M. Ferreira, S. Marelli, W. T. Harvey, S. Lytras, A. Mohamed, G. Gallo, N. Thakur, D. A. Collier, P. Mlcochova, L. M. Duncan, A. M. Carabelli, J. C. Kenyon, A. M. Lever, A. De Marco, C. Saliba, K. Culap, E. Cameroni, N. J. Matheson, L. Piccoli, D. Corti, L. C. James, D. L. Robertson, D. Bailey, R. K. Gupta, COVID-19 Genomics UK (COG-UK) Consortium, Recurrent emergence of SARS-CoV-2 spike deletion H69/V70 and its role in the Alpha variant B.1.1.7. *Cell Rep.* **35**, 109292 (2021). doi:10.1016/j.celrep.2021.109292 Medline
13. P. Mlcochova, S. A. Kemp, M. S. Dhar, G. Papa, B. Meng, I. A. T. M. Ferreira, R. Datir, D. A. Collier, A. Albecka, S. Singh, R. Pandey, J. Brown, J. Zhou, N. Goonawardane, S. Mishra, C. Whittaker, T. Mellan, R. Marwal, M. Datta, S. Sengupta, K. Ponnusamy, V. S. Radhakrishnan, A. Abdullahi, O. Charles, P. Chattopadhyay, P. Devi, D. Caputo, T. Peacock, C. Wattal, N. Goel, A. Satwik, R. Vaishya, M. Agarwal, A. Mavousian, J. H. Lee, J. Bassi, C. Silacci-Fegni, C. Saliba, D. Pinto, T. Irie, I. Yoshida, W. L. Hamilton, K. Sato, S. Bhatt, S. Flaxman, L. C. James, D. Corti, L. Piccoli, W. S. Barclay, P. Rakshit, A. Agrawal, R. K. Gupta, Indian SARS-CoV-2 Genomics Consortium (INSACOG), Genotype to Phenotype Japan (G2P-Japan) Consortium, CITIID-NIHR BioResource COVID-19 Collaboration, SARS-CoV-2 B.1.617.2 Delta variant replication and immune evasion. *Nature* **599**, 114–119 (2021). doi:10.1038/s41586-021-03944-y Medline
14. N. R. Faria, T. A. Mellan, C. Whittaker, I. M. Claro, D. D. S. Candido, S. Mishra, M. A. E. Crispim, F. C. S. Sales, I. Hawryluk, J. T. McCrone, R. J. G. Hulswit, L. A. M. Franco, M. S. Ramundo, J. G. de Jesus, P. S. Andrade, T. M. Coletti, G. M. Ferreira, C. A. M. Silva, E. R. Manuli, R. H. M. Pereira, P. S. Peixoto, M. U. G. Kraemer, N. Gaburo Jr., C. D. C. Camilo, H. Hoeltgebaum, W. M. Souza, E. C. Rocha, L. M. de Souza, M. C. de Pinho, L. J. T. Araujo, F. S. V. Malta, A. B. de Lima, J. D. P. Silva, D. A. G. Zauli, A. C. S. Ferreira, R. P. Schnekenberg, D. J. Laydon, P. G. T. Walker, H. M. Schlüter, A. L. P. Dos Santos, M. S. Vidal, V. S. Del Caro, R. M. F. Filho, H. M. Dos Santos, R. S. Aguiar, J. L. Proença-Modena, B. Nelson, J. A. Hay, M. Monod, X. Miscouridou, H. Coupland, R. Sonabend, M. Vollmer, A. Gandy, C. A. Prete Jr., V. H. Nascimento, M. A. Suchard, T. A. Bowden, S. L. K. Pond, C.-H. Wu, O. Ratmann, N. M. Ferguson, C. Dye, N. J. Loman, P. Lemey, A. Rambaut, N. A. Fraiji, M. D. P. S. S. Carvalho, O. G. Pybus, S. Flaxman, S. Bhatt, E. C. Sabino, Genomics and epidemiology of the P.1 SARS-CoV-2 lineage in Manaus, Brazil. *Science* **372**, 815–821 (2021). doi:10.1126/science.abh2644 Medline
15. E. Wilkinson, M. Giovanetti, H. Tegally, J. E. San, R. Lessells, D. Cuadros, D. P. Martin, D. A. Rasmussen, A. N. Zekri, A. K. Sangare, A.-S. Ouedraogo, A. K. Sesay, A. Priscilla, A.-S. Kemi, A. M. Olubusuyi, A. O. O. Oluwapelumi, A. Hammami, A. A. Amuri, A. Sayed, A. E. O. Ouma, A. Elargoubi, N. A. Ajayi, A. F. Victoria, A. Kazeem, A. George, A. J. Trotter, A. A. Yahaya, A. K. Keita, A. Diallo, A. Kone, A. Souissi, A. Chtourou, A. V. Gutierrez, A. J. Page, A. Vinze, A. Iranzadeh, A. Lambisia, A. Ismail, A. Rosemary, A. Sylverken, A. Femi, A. Ibrahimi, B. Marycelin, B. S. Oderinde, B. Bolajoko, B. Dhaala, B. L. Herring, B.-M. Njanpop-Lafourcade, B. Kleinhans, B. McInnis, B. Tegomoh, C. Brook, C. B. Pratt, C. Scheepers, C. G. Akoua-Koffi, C. N. Agoti, C. Peyrefitte, C. Daubenberger, C. M. Morang'a, D. J. Nokes, D. G. Amoako, D. L. Bugembe, D. Park, D. Baker, D. Doolabh, D. Ssemwanga, D. Tshiabuila, D. Bassirou, D. S. Y. Amuzu, D. Goedhals, D. O. Omuoyo, D. Maruapula, E. Foster-Nyarko, E. K. Lusamaki, E. Simulundu, E. M. Ong'era, E. N. Ngabana, E. Shumba, E. El Fahime, E. Lokilo, E. Mukantwari, E. Philomena, E. Belarbi, E. Simon-Loriere, E. A. Anoh, F. Leendertz, F. Ajili, F. O. Enoch, F. Wasfi, F. Abdelmoula, F. S. Mosha, F. T. Takawira, F. Derrar, F. Bouzid, F. Onikepe, F. Adeola, F. M. Muyembe, F. Tanser, F. A. Dratibi, G. K. Mbunsu, G. Thilliez, G. L. Kay, G. Githinji, G. van Zyl, G. A. Awandare, G. Schubert, G. P. Maphalala, H. C. Ranaivoson, H. Lemriss, H. Anise, H. Abe, H. H. Karray, H. Nansumba, H. A. Elgahzaly, H. Gumbo, I. Smeti, I. B. Ayed, I. Odia, I. B. Ben Boubaker, I. Gaaloul, I. Gazy, I. Mudau, I. Ssewanyana, I. Konstantinus, J. B. Lekana-Douk, J.-C. C. Makangara, J. M. Tamfum, J.-M. Heraud, J. G. Shaffer, J. Giandhari, J. Li, J. Yasuda, J. Q. Mends, J. Kiconco, J. M. Morobe, J. O. Gyapong, J. C. Okolie, J. T. Kayiwa, J. A. Edwards, J. Gyamfi, J. Farah, J.

Nakaseegu, J. M. Ngoi, J. Namulondo, J. C. Andeko, J. J. Lutwama, J. O'Grady, K. Siddle, K. T. Adeyemi, K. A. Tumedi, K. M. Said, K. Hae-Young, K. O. Duedu, L. Belyamani, L. Fki-Berrajah, L. Singh, L. O. Martins, L. Tyers, M. Ramuth, M. Mastouri, M. Aouni, M. El Hefnawi, M. I. Matsheka, M. Kebabonye, M. Diop, M. Turki, M. Paye, M. M. Nyaga, M. Mareka, M.-M. Damaris, M. W. Mburu, M. Mpina, M. Nwando, M. Owusu, M. R. Wiley, M. T. Youtchou, M. O. Ayekaba, M. Abouelhoda, M. G. Seadawy, M. K. Khalifa, M. Sekhele, M. Ouadghiri, M. M. Diagne, M. Mwenda, M. Allam, M. V. T. Phan, N. Abid, N. Touil, N. Rujeni, N. Kharrat, N. Ismael, N. Dia, N. Mabunda, N. Y. Hsiao, N. B. Silochi, N. Nsenga, N. Gumede, N. Mulder, N. Ndodo, N. H. Razanajatovo, N. Iguosadolo, O. Judith, O. C. Kingsley, O. Sylvanus, O. Peter, O. Femi, O. Idowu, O. Testimony, O. E. Chukwuma, O. E. Ogah, C. K. Onwuamah, O. Cyril, O. Faye, O. Tomori, P. Ondoa, P. Combe, P. Semanda, P. E. Oluniyi, P. Arnaldo, P. K. Quashie, P. Dussart, P. A. Bester, P. K. Mbala, R. Ayivor-Djanie, R. Njouom, R. O. Phillips, R. Gorman, R. A. Kingsley, R. A. A. Carr, S. El Kabbaj, S. Gargouri, S. Masmoudi, S. Sankhe, S. B. Lawal, S. Kassim, S. Trabelsi, S. Metha, S. Kammoun, S. Lemriss, S. H. A. Agwa, S. Calvignac-Spencer, S. F. Schaffner, S. Doumbia, S. M. Mandanda, S. Aryeetey, S. S. Ahmed, S. Elhamoumi, S. Andriamandimby, S. Tope, S. Lekana-Douki, S. Prosolek, S. Ouangraoua, S. A. Mundeke, S. Rudder, S. Panji, S. Pillay, S. Engelbrecht, S. Nabadda, S. Behillil, S. L. Budiaki, S. van der Werf, T. Mashe, T. Aanniz, T. Mohale, T. Le-Viet, T. Schindler, U. J. Anyaneji, U. Chinedu, U. Ramphal, U. Jessica, U. George, V. Fonseca, V. Enouf, V. Gorova, W. H. Roshdy, W. K. Ampofo, W. Preiser, W. T. Choga, Y. Bediako, Y. Naidoo, Y. Butera, Z. R. de Laurent, A. A. Sall, A. Rebai, A. von Gottberg, B. Kouriba, C. Williamson, D. J. Bridges, I. Chikwe, J. N. Bhiman, M. Mine, M. Cotten, S. Moyo, S. Gaseitsiwe, N. Saasa, P. C. Sabeti, P. Kaleebu, Y. K. Tebeje, S. K. Tessema, C. Happi, J. Nkengasong, T. de Oliveira, A year of genomic surveillance reveals how the SARS-CoV-2 pandemic unfolded in Africa. *Science* **374**, 423–431 (2021). doi:10.1126/science.abj4336 Medline

16. Y. Shu, J. McCauley, GISAID: Global initiative on sharing all influenza data - from vision to reality. *Euro Surveill.* **22**, 30494 (2017). doi:10.2807/1560-7917.ES.2017.22.13.30494 Medline
17. C. Kuiken, B. Korber, R. W. Shafer, HIV sequence databases. *AIDS Rev.* **5**, 52–61 (2003). Medline
18. D. L. Bugembe, J. Kayiwa, M. V. T. Phan, P. Tushabe, S. Balinandi, B. Dhaala, J. Lexow, H. Mwebesa, J. Aceng, H. Kyobe, D. Ssemwanga, J. Lutwama, P. Kaleebu, M. Cotten, Main routes of entry and genomic diversity of SARS-CoV-2, Uganda. *Emerg. Infect. Dis.* **26**, 2411–2415 (2020). doi:10.3201/eid2610.202575 Medline
19. T. Mashe, F. T. Takawira, L. de Oliveira Martins, M. Gudza-Mugabe, J. Chirenda, M. Munyanyi, B. V. Chaibva, A. Tarupiwa, H. Gumbo, A. Juru, C. Nyagupe, V. Ruhanya, I. Phiri, P. Manangazira, A. Goredema, S. Danda, I. Chabata, J. Jonga, R. Munharira, K. Masunda, I. Mukeredzi, D. Mangwanya, A. Trotter, T. Le Viet, S. Rudder, G. Kay, D. Baker, G. Thilliez, A. V. Gutierrez, J. O'Grady, M. Hove, S. Mutapuri-Zinyowera, A. J. Page, R. A. Kingsley, G. Mhlanga, COVID-19 Genomics UK Consortium, SARS-CoV-2 Research Group, Genomic epidemiology and the role of international and regional travel in the SARS-CoV-2 epidemic in Zimbabwe: A retrospective study of routinely collected surveillance data. *Lancet Glob. Health* **9**, e1658–e1666 (2021). doi:10.1016/S2214-109X(21)00434-4 Medline
20. A. Chouikha, W. Fares, A. Laamari, S. Haddad-Boubaker, Z. Belaiba, K. Ghedira, W. Kammoun Rebai, K. Ayouni, M. Khedhiri, S. Ben Halima, H. Krichen, H. Touzi, I. Ben Dhifallah, F. Z. Guerfali, C. Atri, S. Azouz, O. Khamessi, M. Ardhaoui, M. Safer, N. Ben Alaya, I. Guizani, R. Kefi, M. Gdoura, H. Triki, Molecular epidemiology of SARS-CoV-2 in Tunisia (North Africa) through several successive waves of COVID-19. *Viruses* **14**, 624 (2022). doi:10.3390/v14030624 Medline
21. F. Ntoumi, C. C. Mfoutou Mapanguy, A. Tomazatos, S. R. Pallerla, L. T. K. Linh, N. Casadei, A. Angelov, M. Sonnabend, S. Peter, P. G. Kremsner, T. P. Velavan, Genomic surveillance of SARS-CoV-2 in the Republic of Congo. *Int. J. Infect. Dis.* **105**, 735–738 (2021). doi:10.1016/j.ijid.2021.03.036 Medline
22. Y. Butera, E. Mukantwari, M. Artesi, J. D'Arc Umuringa, Á. N. O'Toole, V. Hill, S. Rooke, S. L. Hong, S. Dellicour, O. Majyambere, S. Bontems, B. Boujemla, J. Quick, P. C. Resende, N. Loman, E. Umumararungu, A. Kabanda, M. M. Murindahabi, P. Tuyisenge, M. Gashegu, N. Rujeni, Genomic sequencing of SARS-CoV-2 in Rwanda: Evolution and regional dynamics. medRxiv 2021.04.02.21254839 [Preprint] (2021); https://doi.org/10.1101/2021.04.02.21254839.
23. C. N Agoti, G. Githinji, K. S Mohammed, A. W Lambisia, Z. R de Laurent, M. W Mburu, E. M Ong'era, J. M Morobe, E. Otieno, H. Abdou Azali, K. Said Abdallah, A. Diarra, A. Ahmed Yahaya, P. Borus, N. Gumede Moeletsi, D. Fred Athanasius, B. Tsofa, P. Bejon, D. James Nokes, L. Isabella Ochola-Oyier, Detection of SARS-CoV-2 variant 501Y.V2 in Comoros Islands in January 2021. *Wellcome Open Res.* **6**, 192 (2021). doi:10.12688/wellcomeopenres.16889.1 Medline
24. J. M. Morobe, B. Pool, L. Marie, D. Didon, A. W. Lambisia, T. Makori, K. S. Mohammed, Z. R. de Laurent, L. Ndwiga, M. W. Mburu, E. Moraa, N. Murunga, J. Musyoki, J. Mwacharo, L. Nyamako, D. Riako, P. Ephnatus, F. Gambo, J. Naimani, J. Namulondo, S. Z. Tembo, E. Ogendi, T. Balde, F. A. Dratibi, A. A. Yahaya, N. Gumede, R. A. Achilla, P. K. Borus, D. W. Wanjohi, S. K. Tessema, J. Mwangangi, P. Bejon, D. J. Nokes, L. I. Ochola-Oyier, G. Githinji, L. Biscornet, C. N. Agoti, Genomic Epidemiology of SARS-CoV-2 in Seychelles, 2020-2021. *Viruses* **14**, 1318 (2022). doi:10.3390/v14061318 Medline
25. C. M. Morang'a, J. M. Ngoi, J. Gyamfi, D. S. Y. Amuzu, B. D. Nuertey, P. M. Soglo, V. Appiah, I. A. Asante, P. Owusu-Oduro, S. Armoo, D. Adu-Gyasi, N. Amoako, J. Oliver-Commey, M. Owusu, A. Sylverken, E. D. Fenteng, V. V. M'cormack, F. Tei-Maya, E. B. Quansah, R. Ayivor-Djanie, E. K. Amoako, I. T. Ogbe, B. K. Yemi, I. Osei-Wusu, D. N. A. Mettle, S. Saiid, K. Tapela, F. Dzabeng, V. Magnussen, J. Quaye, P. C. Opurum, R. A. Carr, P. T. Ababio, A. K. Abass, S. K. Akoriyea, E. Amoako, F. Kumi-Ansah, O. D. Boakye, D. K. Mibut, T. Odoom, L. Ofori-Boadu, E. Allegye-Cudjoe, S. Dassah, V. Asoala, K. P. Asante, R. O. Phillips, M. Y. Osei-Atweneboana, J. O. Gyapong, P. Kuma-Aboagye, W. K. Ampofo, K. O. Duedu, N. T. Ndam, Y. Bediako, P. K. Quashie, L. N. Amenga-Etego, G. A. Awandare, Genetic diversity of SARS-CoV-2 infections in Ghana from 2020-2021. *Nat. Commun.* **13**, 2494 (2022). doi:10.1038/s41467-022-30219-5 Medline
26. C. N. Agoti, L. I. Ochola-Oyier, S. Dellicour, K. S. Mohammed, A. W. Lambisia, Z. R. de Laurent, J. M. Morobe, M. W. Mburu, D. O. Omuoyo, E. M. Ongera, L. Ndwiga, E. Maitha, B. Kitole, T. Suleiman, M. Mwakinangu, J. K. Nyambu, J. Otieno, B. Salim, J. Musyoki, N. Murunga, E. Otieno, J. N. Kiiru, K. Kasera, P. Amoth, M. Mwangangi, R. Aman, S. Kinyanjui, G. Warimwe, M. Phan, A. Agweyu, M. Cotten, E. Barasa, B. Tsofa, D. J. Nokes, P. Bejon, G. Githinji, Transmission networks of SARS-CoV-2 in Coastal Kenya during the first two waves: A retrospective genomic study. *eLife* **11**, e71703 (2022). doi:10.7554/eLife.71703 Medline
27. S. P. C. Brand, J. Ojal, R. Aziza, V. Were, E. A. Okiro, I. K. Kombe, C. Mburu, M. Ogero, A. Agweyu, G. M. Warimwe, J. Nyagwange, H. Karanja, J. N. Gitonga, D. Mugo, S. Uyoga, I. M. O. Adetifa, J. A. G. Scott, E. Otieno, N. Murunga, M. Otiende, L. I. Ochola-Oyier, C. N. Agoti, G. Githinji, K. Kasera, P. Amoth, M. Mwangangi, R. Aman, W. Ng'ang'a, B. Tsofa, P. Bejon, M. J. Keeling, D. J. Nokes, E. Barasa, COVID-19 transmission dynamics underlying epidemic waves in Kenya. *Science* **374**, 989–994 (2021). doi:10.1126/science.abk0414 Medline
28. G. Githinji, Z. R. de Laurent, K. S. Mohammed, D. O. Omuoyo, P. M. Macharia, J. M. Morobe, E. Otieno, S. M. Kinyanjui, A. Agweyu, E. Maitha, B. Kitole, T. Suleiman, M. Mwakinangu, J. Nyambu, J. Otieno, B. Salim, K. Kasera, J. Kiiru, R. Aman, E. Barasa, G. Warimwe, P. Bejon, B. Tsofa, L. I. Ochola-Oyier, D. J. Nokes, C. N. Agoti, Tracking the introduction and spread of SARS-CoV-2 in coastal Kenya. *Nat. Commun.* **12**, 4809 (2021). doi:10.1038/s41467-021-25137-x Medline
29. H. Tegally, E. Wilkinson, R. J. Lessells, J. Giandhari, S. Pillay, N. Msomi, K. Mlisana, J. N. Bhiman, A. von Gottberg, S. Walaza, V. Fonseca, M. Allam, A. Ismail, A. J. Glass, S. Engelbrecht, G. Van Zyl, W. Preiser, C. Williamson, F. Petruccione, A. Sigal, I. Gazy, D. Hardie, N. Y. Hsiao, D. Martin, D. York, D. Goedhals, E. J. San, M. Giovanetti, J. Lourenço, L. C. J. Alcantara, T. de Oliveira, Sixteen novel lineages of SARS-CoV-2 in South Africa. *Nat. Med.* **27**, 440–446 (2021). doi:10.1038/s41591-021-01255-3 Medline
30. W. H. Roshdy, M. K. Khalifa, J. E. San, H. Tegally, E. Wilkinson, S. Showky, D. P. Martin, M. Moir, A. Naguib, N. Elguindy, M. R. Gomaa, M. Fahim, H. A. Elsood, A. A. Mohsen, R. Galal, M. Hassany, R. J. Lessells, A. A. Al Karmalawy, R. E. L. Shesheny, A. M. Kandeil, T. de Oliveira, SARS-CoV-2 Genetic diversity and lineage dynamics of in Egypt. medRxiv 2022.01.05.22268646 [Preprint] (2022); https://doi.org/10.1101/2022.01.05.22268646.
31. C. Scheepers, J. Everatt, D. G. Amoako, H. Tegally, C. K. Wibmer, A. Mnguni, A. Ismail, B. Mahlangu, B. E. Lambson, D. P. Martin, E. Wilkinson, J. E. San, J. Giandhari, N. Manamela, N. Ntuli, P. Kgagudi, S. Cele, S. I. Richardson, S. Pillay, T. Mohale, U. Ramphal, Y. Naidoo, Z. T. Khumalo, G. Kwatra, G. Gray, L.-G. Bekker, S. A. Madhi, V. Baillie, W. C. Van Voorhis, F. K. Treurnicht, M. Venter, K. Mlisana, N. Wolter, A. Sigal, C. Williamson, N. Y. Hsiao, N. Msomi, T. Maponga, W. Preiser, Z. Makatini, R. Lessells, P. L. Moore, T. de Oliveira, A. von Gottberg, J. N. Bhiman,

Emergence and phenotypic characterization of the global SARS-CoV-2 C.1.2 lineage. *Nat. Commun.* **13**, 1976 (2022). doi:10.1038/s41467-022-29579-9 Medline

32. G. Dudas, S. L. Hong, B. I. Potter, S. Calvignac-Spencer, F. S. Niatou-Singa, T. B. Tombolomako, T. Fuh-Neba, U. Vickos, M. Ulrich, F. H. Leendertz, K. Khan, C. Huber, A. Watts, I. Olendraitė, J. Snijder, K. N. Wijnant, A. M. J. J. Bonvin, P. Martres, S. Behillil, A. Ayouba, M. F. Maidadi, D. M. Djomsi, C. Godwe, C. Butel, A. Šimaitis, M. Gabrielaitė, M. Katėnaitė, R. Norvilas, L. Raugaitė, G. W. Koyaweda, J. K. Kandou, R. Jonikas, I. Nasvytienė, Ž. Žemeckienė, D. Gečys, K. Tamušauskaitė, M. Norkienė, E. Vasiliūnaitė, D. Žiogienė, A. Timinskas, M. Šukys, M. Šarauskas, G. Alzbutas, A. A. Aziza, E. K. Lusamaki, J. M. Cigolo, F. M. Mawete, E. L. Lofiko, P. M. Kingebeni, J. M. Tamfum, M. R. D. Belizaire, R. G. Essomba, M. C. O. Assoumou, A. B. Mboringong, A. B. Dieng, D. Juozapaitė, S. Hosch, J. Obama, M. O. Ayekaba, D. Naumovas, A. Pautienius, C. D. Rafaï, A. Vitkauskienė, R. Ugenskienė, A. Gedvilaitė, D. Čereškevičius, V. Lesauskaitė, L. Žemaitis, L. Griškevičius, G. Baele, Emergence and spread of SARS-CoV-2 lineage B.1.620 with variant of concern-like mutations and deletions. *Nat. Commun.* **12**, 5769 (2021). doi:10.1038/s41467-021-26055-8 Medline
33. D. L. Bugembe, M. V. T. Phan, I. Ssewanyana, P. Semanda, H. Nansumba, B. Dhaala, S. Nabadda, Á. N. O'Toole, A. Rambaut, P. Kaleebu, M. Cotten, Emergence and spread of a SARS-CoV-2 lineage A variant (A.23.1) with altered spike protein in Uganda. *Nat. Microbiol.* **6**, 1094–1101 (2021). doi:10.1038/s41564-021-00933-9 Medline
34. H. Tegally, M. Moir, J. Everatt, M. Giovanetti, C. Scheepers, E. Wilkinson, K. Subramoney, Z. Makatini, S. Moyo, D. G. Amoako, C. Baxter, C. L. Althaus, U. J. Anyaneji, D. Kekana, R. Viana, J. Giandhari, R. J. Lessells, T. Maponga, D. Maruapula, W. Choga, M. Matshaba, M. B. Mbulawa, N. Msomi, Y. Naidoo, S. Pillay, T. J. Sanko, J. E. San, L. Scott, L. Singh, N. A. Magini, P. Smith-Lawrence, W. Stevens, G. Dor, D. Tshiabuila, N. Wolter, W. Preiser, F. K. Treurnicht, M. Venter, G. Chiloane, C. McIntyre, A. O'Toole, C. Ruis, T. P. Peacock, C. Roemer, S. L. Kosakovsky Pond, C. Williamson, O. G. Pybus, J. N. Bhiman, A. Glass, D. P. Martin, B. Jackson, A. Rambaut, O. Laguda-Akingba, S. Gaseitsiwe, A. von Gottberg, T. de Oliveira, NGS-SA Consortium, Emergence of SARS-CoV-2 Omicron lineages BA.4 and BA.5 in South Africa. *Nat. Med.* (2022). doi:10.1038/s41591-022-01911-2 Medline
35. S. A. Madhi, G. Kwatra, J. E. Myers, W. Jassat, N. Dhar, C. K. Mukendi, A. J. Nana, L. Blumberg, R. Welch, N. Ngorima-Mabhena, P. C. Mutevedzi, Population immunity and Covid-19 severity with omicron variant in South Africa. *N. Engl. J. Med.* **386**, 1314–1326 (2022). doi:10.1056/NEJMoa2119658 Medline
36. N. Wolter, W. Jassat, S. Walaza, R. Welch, H. Moultrie, M. Groome, D. G. Amoako, J. Everatt, J. N. Bhiman, C. Scheepers, N. Tebeila, N. Chiwandire, M. du Plessis, N. Govender, A. Ismail, A. Glass, K. Mlisana, W. Stevens, F. K. Treurnicht, Z. Makatini, N. Y. Hsiao, R. Parboosing, J. Wadula, H. Hussey, M. A. Davies, A. Boulle, A. von Gottberg, C. Cohen, Early assessment of the clinical severity of the SARS-CoV-2 omicron variant in South Africa: A data linkage study. *Lancet* **399**, 437–446 (2022). doi:10.1016/S0140-6736(22)00017-4 Medline
37. H. C. Lewis, H. Ware, M. Whelan, L. Subissi, Z. Li, X. Ma, A. Nardone, M. Valenciano, B. Cheng, K. Noel, C. Cao, M. Yanes-Lane, B. L. Herring, A. Talisuna, N. Ngoy, T. Balde, D. Clifton, M. D. Van Kerkhove, D. Buckeridge, N. Bobrovitz, J. Okeibunor, R. K. Arora, I. Bergeri, UNITY Studies Collaborator Group, SARS-CoV-2 infection in Africa: A systematic review and meta-analysis of standardised seroprevalence studies, from January 2020 to December 2021. *BMJ Glob. Health* **7**, e008793 (2022). doi:10.1136/bmjgh-2022-008793 Medline
38. R. M. Barber, R. J. D. Sorensen, D. M. Pigott, C. Bisignano, A. Carter, J. O. Amlag, J. K. Collins, C. Abbafati, C. Adolph, A. Allorant, A. Y. Aravkin, B. L. Bang-Jensen, E. Castro, S. Chakrabarti, R. M. Cogen, E. Combs, H. Comfort, K. Cooperrider, X. Dai, F. Daoud, A. Deen, L. Earl, M. Erickson, S. B. Ewald, A. J. Ferrari, A. D. Flaxman, J. J. Frostad, N. Fullman, J. R. Giles, G. Guo, J. He, M. Helak, E. N. Hulland, B. M. Huntley, A. Lazzar-Atwood, K. E. LeGrand, S. S. Lim, A. Lindstrom, E. Linebarger, R. Lozano, B. Magistro, D. C. Malta, J. Månsson, A. M. Mantilla Herrera, A. H. Mokdad, L. Monasta, M. Naghavi, S. Nomura, C. M. Odell, L. T. Olana, S. M. Ostroff, M. Pasovic, S. A. Pease, R. C. Reiner Jr., G. Reinke, A. L. P. Ribeiro, D. F. Santomauro, A. Sholokhov, E. E. Spurlock, R. Syailendrawati, R. Topor-Madry, A. T. Vo, T. Vos, R. Walcott, A. Walker, K. E. Wiens, C. S. Wiysonge, N. A. Worku, P. Zheng, S. I. Hay, E. Gakidou, C. J. L. Murray, COVID-19 Cumulative Infection Collaborators, Estimating global, regional, and national daily and cumulative infections with SARS-CoV-2 through Nov 14, 2021: A statistical analysis. *Lancet* **399**, 2351–2380 (2022). doi:10.1016/S0140-6736(22)00484-6 Medline
39. J. Quick, nCoV-2019 sequencing protocol v3 (LoCost) (2020).
40. J. R. Tyson, P. James, D. Stoddart, N. Sparks, A. Wickenhagen, G. Hall, J. H. Choi, H. Lapointe, K. Kamelian, A. D. Smith, N. Prystajecky, I. Goodfellow, S. J. Wilson, R. Harrigan, T. P. Snutch, N. J. Loman, J. Quick, Improvements to the ARTIC multiplex PCR method for SARS-CoV-2 genome sequencing using nanopore. bioRxiv 2020.09.04.283077 [Preprint] (2020); https://doi.org/10.1101/2020.09.04.283077.
41. M. Cotten, D. Lule Bugembe, P. Kaleebu, M. V T Phan, Alternate primers for whole-genome SARS-CoV-2 sequencing. *Virus Evol.* **7**, veab006 (2021). doi:10.1093/ve/veab006 Medline
42. H. Tegally, M. Ramuth, D. Amoaka, C. Scheepers, E. Wilkinson, M. Giovanetti, R. J. Lessells, J. Giandhari, A. Ismail, D. Martin, E. J. San, M. Crawford, R. S. Daniels, R. Harvey, S. Bahadoor, J. Sonoo, M. Timol, L. Veerapa-Mangroo, A. von Gottberg, J. Bhiman, T. de Oliveira, S. Manraj, A novel and expanding SARS-CoV-2 variant, B.1.1.318, dominates infections in Mauritius. medRxiv 2021.06.16.21259017 [Preprint] (2021); https://doi.org/10.1101/2021.06.16.21259017.
43. A. N. Zekri, A. A. Bahnasy, M. M. Hafez, Z. K. Hassan, O. S. Ahmed, H. K. Soliman, E. R. El-Sisi, M. H. S. E. Dine, M. S. Solimane, L. S. A. Latife, M. G. Seadawy, A. S. Elsafty, M. Abouelhoda, Characterization of the SARS-CoV-2 genomes in Egypt in first and second waves of infection. *Sci. Rep.* **11**, 21632 (2021). doi:10.1038/s41598-021-99014-4 Medline
44. C. Nasimiyu, D. Matoke-Muhia, G. K. Rono, E. Osoro, D. O. Obado, J. M. Mwangi, N. Mwikwabe, K. Thiongâ O, J. Dawa, I. Ngere, J. Gachohi, S. Kariuki, E. Amukoye, M. Mureithi, P. Ngere, P. Amoth, I. Were, L. Makayotto, V. Nene, E. O. Abworo, M. K. Njenga, S. N. Seifert, S. O. Oyola, Imported SARS-COV-2 variants of concern drove spread of infections across Kenya during the second year of the pandemic. *COVID* **2**, 586–598 (2022). doi:10.3390/covid2050044
45. M. U. G. Kraemer, V. Hill, C. Ruis, S. Dellicour, S. Bajaj, J. T. McCrone, G. Baele, K. V. Parag, A. L. Battle, B. Gutierrez, B. Jackson, R. Colquhoun, Á. O'Toole, B. Klein, A. Vespignani, E. Volz, N. R. Faria, D. M. Aanensen, N. J. Loman, L. du Plessis, S. Cauchemez, A. Rambaut, S. V. Scarpino, O. G. Pybus, COVID-19 Genomics UK (COG-UK) Consortium, Spatiotemporal invasion dynamics of SARS-CoV-2 lineage B.1.1.7 emergence. *Science* **373**, 889–895 (2021). doi:10.1126/science.abj0113 Medline
46. S. J. Lycett, J. Hughes, M. P. McHugh, A. da Silva Filipe, R. Dewar, L. Lu, T. Doherty, A. Shepherd, R. Inward, G. Rossi, D. Balaz, R. R. Kao, S. Rooke, S. Cotton, M. D. Gallagher, C. B. Lopez, Á. O'Toole, E. Scher, V. Hill, J. T. McCrone, R. M. Colquhoun, B. Jackson, T. C. Williams, K. A. Williamson, N. Johnson, K. Smollett, D. Mair, S. Carmichael, L. Tong, J. Nichols, K. Brunker, J. G. Shepherd, K. Li, E. Aranday-Cortes, Y. A. Parr, A. Broos, K. Nomikou, S. E. McDonald, M. Niebel, P. Asamaphan, I. Starinskij, N. Jesudason, R. Shah, V. B. Sreenu, T. Stanton, S. Shaaban, A. MacLean, M. Woolhouse, R. Gunson, K. Templeton, E. C. Thomson, A. Rambaut, M. T. G. Holden, D. L. Robertson, COVID-19 Genomics UK (COG-UK) Consortium, Epidemic waves of COVID-19 in Scotland: a genomic perspective on the impact of the introduction and relaxation of lockdown on SARS-CoV-2. medRxiv 2021.01.08.20248677 [Preprint] (2021); https://doi.org/10.1101/2021.01.08.20248677.
47. P. W. G. Mallon, F. Crispie, G. Gonzalez, W. Tinago, A. A. Garcia Leon, M. McCabe, E. de Barra, O. Yousif, J. S. Lambert, C. J. Walsh, J. G. Kenny, E. Feeney, M. Carr, P. Doran, P. D. Cotter, Whole-genome sequencing of SARS-CoV-2 in the Republic of Ireland during waves 1 and 2 of the pandemic. medRxiv 2021.02.09.21251402 [Preprint] (2021); https://doi.org/10.1101/2021.02.09.21251402.
48. H. Tegally, E. Wilkinson, C. L. Althaus, M. Giovanetti, J. E. San, J. Giandhari, S. Pillay, Y. Naidoo, U. Ramphal, N. Msomi, K. Mlisana, D. G. Amoako, J. Everatt, T. Mohale, A. Nguni, B. Mahlangu, N. Ntuli, Z. T. Khumalo, Z. Makatini, N. Wolter, C. Scheepers, A. Ismail, D. Doolabh, R. Joseph, A. Strydom, A. Mendes, M. Davis, S. H. Mayaphi, Y. Ramphal, A. Maharaj, W. A. Karim, D. Tshiabuila, U. J. Anyaneji, L. Singh, S. Engelbrecht, V. Fonseca, K. Marais, S. Korsman, D. Hardie, N. Hsiao, T. Maponga, G. van Zyl, G. Marais, A. Iranzadeh, D. Martin, L. C. J. Alcantara, P. A. Bester, M. M. Nyaga, K. Subramoney, F. K. Treurnicht, M. Venter, D. Goedhals, W. Preiser, J. N. Bhiman, A. Gottberg, C. Williamson, R. J. Lessells, T. de Oliveira, Rapid replacement of the Beta variant by the Delta variant in South Africa.

medRxiv 2021.09.23.21264018 [Preprint] (2021); https://doi.org/10.1101/2021.09.23.21264018.
49. S. Chang, E. Pierson, P. W. Koh, J. Gerardin, B. Redbird, D. Grusky, J. Leskovec, Mobility network models of COVID-19 explain inequities and inform reopening. *Nature* **589**, 82–87 (2021). doi:10.1038/s41586-020-2923-3 Medline
50. M. Chinazzi, J. T. Davis, M. Ajelli, C. Gioannini, M. Litvinova, S. Merler, A. Pastore Y Piontti, K. Mu, L. Rossi, K. Sun, C. Viboud, X. Xiong, H. Yu, M. E. Halloran, I. M. Longini Jr., A. Vespignani, The effect of travel restrictions on the spread of the 2019 novel coronavirus (COVID-19) outbreak. *Science* **368**, 395–400 (2020). doi:10.1126/science.aba9757 Medline
51. M. U. G. Kraemer, C.-H. Yang, B. Gutierrez, C.-H. Wu, B. Klein, D. M. Pigott, L. du Plessis, N. R. Faria, R. Li, W. P. Hanage, J. S. Brownstein, M. Layan, A. Vespignani, H. Tian, C. Dye, O. G. Pybus, S. V. Scarpino, Open COVID-19 Data Working Group, The effect of human mobility and control measures on the COVID-19 epidemic in China. *Science* **368**, 493–497 (2020). doi:10.1126/science.abb4218 Medline
52. P. Nouvellet, S. Bhatia, A. Cori, K. E. C. Ainslie, M. Baguelin, S. Bhatt, A. Boonyasiri, N. F. Brazeau, L. Cattarino, L. V. Cooper, H. Coupland, Z. M. Cucunuba, G. Cuomo-Dannenburg, A. Dighe, B. A. Djaafara, I. Dorigatti, O. D. Eales, S. L. van Elsland, F. F. Nascimento, R. G. FitzJohn, K. A. M. Gaythorpe, L. Geidelberg, W. D. Green, A. Hamlet, K. Hauck, W. Hinsley, N. Imai, B. Jeffrey, E. Knock, D. J. Laydon, J. A. Lees, T. Mangal, T. A. Mellan, G. Nedjati-Gilani, K. V. Parag, M. Pons-Salort, M. Ragonnet-Cronin, S. Riley, H. J. T. Unwin, R. Verity, M. A. C. Vollmer, E. Volz, P. G. T. Walker, C. E. Walters, H. Wang, O. J. Watson, C. Whittaker, L. K. Whittles, X. Xi, N. M. Ferguson, C. A. Donnelly, Reduction in mobility and COVID-19 transmission. *Nat. Commun.* **12**, 1090 (2021). doi:10.1038/s41467-021-21358-2 Medline
53. C. Xiong, S. Hu, M. Yang, W. Luo, L. Zhang, Mobile device data reveal the dynamics in a positive relationship between human mobility and COVID-19 infections. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 27087–27089 (2020). doi:10.1073/pnas.2010836117 Medline
54. S. Pillay, J. Giandhari, H. Tegally, E. Wilkinson, B. Chimukangara, R. Lessells, Y. Moosa, S. Mattison, I. Gazy, M. Fish, L. Singh, K. S. Khanyile, J. E. San, V. Fonseca, M. Giovanetti, L. C. Alcantara Jr., T. de Oliveira, Whole genome sequencing of SARS-CoV-2: Adapting Illumina protocols for quick and accurate outbreak investigation during a pandemic. *Genes* **11**, 949 (2020). doi:10.3390/genes11080949 Medline
55. L. Singh, J. E. San, H. Tegally, P. M. Brzoska, U. J. Anyaneji, E. Wilkinson, L. Clark, J. Giandhari, S. Pillay, R. J. Lessells, D. P. Martin, M. Furtado, A. M. Kiran, T. de Oliveira, Targeted Sanger sequencing to recover key mutations in SARS-CoV-2 variant genome assemblies produced by next-generation sequencing. *Microb. Genom.* **8**, (2022). doi:10.1099/mgen.0.000774 Medline
56. A. J. Page, A. E. Mather, T. Le-Viet, E. J. Meader, N.-F. Alikhan, G. L. Kay, L. de Oliveira Martins, A. Aydin, D. J. Baker, A. J. Trotter, S. Rudder, A. P. Tedim, A. Kolyva, R. Stanley, M. Yasir, M. Diaz, W. Potter, C. Stuart, L. Meadows, A. Bell, A. V. Gutierrez, N. M. Thomson, E. M. Adriaenssens, T. Swingler, R. A. J. Gilroy, L. Griffith, D. K. Sethi, D. Aggarwal, C. S. Brown, R. K. Davidson, R. A. Kingsley, L. Bedford, L. J. Coupland, I. G. Charles, N. Elumogo, J. Wain, R. Prakash, M. A. Webber, S. J. L. Smith, M. Chand, S. Dervisevic, J. O'Grady, The Covid-Genomics Uk Cog-Uk Consortium, Large-scale sequencing of SARS-CoV-2 genomes from one region allows detailed epidemiology and enables local outbreak management. *Microb. Genom.* **7**, (2021). doi:10.1099/mgen.0.000589 Medline
57. N. De Maio, C. Walker, R. Borges, L. Weilguny, G. Slodkowicz, N. Goldman, Issues with SARS-CoV-2 sequencing data, Virological.org (2020); https://virological.org/t/issues-with-sars-cov-2-sequencing-data/473.
58. N. E. Freed, M. Vlková, M. B. Faisal, O. K. Silander, Rapid and inexpensive whole-genome sequencing of SARS-CoV-2 using 1200 bp tiled amplicons and Oxford Nanopore Rapid Barcoding. *Biol. Methods Protoc.* **5**, bpaa014 (2020). doi:10.1093/biomethods/bpaa014 Medline
59. J.-S. Eden, R. Rockett, I. Carter, H. Rahman, J. de Ligt, J. Hadfield, M. Storey, X. Ren, R. Tulloch, K. Basile, J. Wells, R. Byun, N. Gilroy, M. V. O'Sullivan, V. Sintchenko, S. C. Chen, S. Maddocks, T. C. Sorrell, E. C. Holmes, D. E. Dwyer, J. Kok, 2019-nCoV Study Group, An emergent clade of SARS-CoV-2 linked to returned travellers from Iran. *Virus Evol.* **6**, veaa027 (2020). doi:10.1093/ve/veaa027 Medline
60. A. S. Gonzalez-Reiche, M. M. Hernandez, M. J. Sullivan, B. Ciferri, H. Alshammary, A. Obla, S. Fabre, G. Kleiner, J. Polanco, Z. Khan, B. Alburquerque, A. van de Guchte, J. Dutta, N. Francoeur, B. S. Melo, I. Oussenko, G. Deikus, J. Soto, S. H. Sridhar, Y.-C. Wang, K. Twyman, A. Kasarskis, D. R. Altman, M. Smith, R. Sebra, J. Aberg, F. Krammer, A. García-Sastre, M. Luksza, G. Patel, A. Paniz-Mondolfi, M. Gitman, E. M. Sordillo, V. Simon, H. van Bakel, Introductions and early spread of SARS-CoV-2 in the New York City area. *Science* **369**, 297–301 (2020). doi:10.1126/science.abc1917 Medline
61. K. Itokawa, T. Sekizuka, M. Hashino, R. Tanaka, M. Kuroda, Disentangling primer interactions improves SARS-CoV-2 genome sequencing by multiplex tiling PCR. *PLOS ONE* **15**, e0239403 (2020). doi:10.1371/journal.pone.0239403 Medline
62. A. X. Han, A. Toporowski, J. A. Sacks, M. Perkins, S. Briand, M. van Kerkhove, E. Hannay, S. Carmona, B. Rodriguez, E. Parker, B. E. Nichols, C. A. Russell, Low testing rates limit the ability of genomic surveillance programs to monitor SARS-CoV-2 variants: a mathematical modelling study. medRxiv 2022.05.20.22275319 [Preprint] (2022); https://doi.org/10.1101/2022.05.20.22275319.
63. H. Wang, K. R. Paulson, S. A. Pease, S. Watson, H. Comfort, P. Zheng, A. Y. Aravkin, C. Bisignano, R. M. Barber, T. Alam, J. E. Fuller, E. A. May, D. P. Jones, M. E. Frisch, C. Abbafati, C. Adolph, A. Allorant, J. O. Amlag, B. Bang-Jensen, G. J. Bertolacci, S. S. Bloom, A. Carter, E. Castro, S. Chakrabarti, J. Chattopadhyay, R. M. Cogen, J. K. Collins, K. Cooperrider, X. Dai, W. J. Dangel, F. Daoud, C. Dapper, A. Deen, B. B. Duncan, M. Erickson, S. B. Ewald, T. Fedosseeva, A. J. Ferrari, J. J. Frostad, N. Fullman, J. Gallagher, A. Gamkrelidze, G. Guo, J. He, M. Helak, N. J. Henry, E. N. Hulland, B. M. Huntley, M. Kereselidze, A. Lazzar-Atwood, K. E. LeGrand, A. Lindstrom, E. Linebarger, P. A. Lotufo, R. Lozano, B. Magistro, D. C. Malta, J. Månsson, A. M. Mantilla Herrera, F. Marinho, A. H. Mirkuzie, A. T. Misganaw, L. Monasta, P. Naik, S. Nomura, E. G. O'Brien, J. K. O'Halloran, L. T. Olana, S. M. Ostroff, L. Penberthy, R. C. Reiner Jr., G. Reinke, A. L. P. Ribeiro, D. F. Santomauro, M. I. Schmidt, D. H. Shaw, B. S. Sheena, A. Sholokhov, N. Skhvitaridze, R. J. D. Sorensen, E. E. Spurlock, R. Syailendrawati, R. Topor-Madry, C. E. Troeger, R. Walcott, A. Walker, C. S. Wiysonge, N. A. Worku, B. Zigler, D. M. Pigott, M. Naghavi, A. H. Mokdad, S. S. Lim, S. I. Hay, E. Gakidou, C. J. L. Murray, COVID-19 Excess Mortality Collaborators, Estimating excess mortality due to the COVID-19 pandemic: A systematic analysis of COVID-19-related mortality, 2020-21. *Lancet* **399**, 1513–1536 (2022). doi:10.1016/S0140-6736(21)02796-3 Medline
64. F. Wu, S. Zhao, B. Yu, Y.-M. Chen, W. Wang, Z.-G. Song, Y. Hu, Z.-W. Tao, J.-H. Tian, Y.-Y. Pei, M.-L. Yuan, Y.-L. Zhang, F.-H. Dai, Y. Liu, Q.-M. Wang, J.-J. Zheng, L. Xu, E. C. Holmes, Y.-Z. Zhang, A new coronavirus associated with human respiratory disease in China. *Nature* **579**, 265–269 (2020). doi:10.1038/s41586-020-2008-3 Medline
65. I. Aksamentov, C. Roemer, E. Hodcroft, R. Neher, Nextclade: Clade assignment, mutation calling and quality control for viral genomes. *J. Open Source Softw.* **6**, 3773 (2021). doi:10.21105/joss.03773
66. D. P. Martin, A. Varsani, P. Roumagnac, G. Botha, S. Maslamoney, T. Schwab, Z. Kelz, V. Kumar, B. Murrell, RDP5: A computer program for analyzing recombination in, and removing signals of recombination from, nucleotide sequence datasets. *Virus Evol.* **7**, veaa087 (2020). doi:10.1093/ve/veaa087 Medline
67. M. N. Price, P. S. Dehal, A. P. Arkin, FastTree 2—Approximately maximum-likelihood trees for large alignments. *PLOS ONE* **5**, e9490 (2010). doi:10.1371/journal.pone.0009490 Medline
68. J. Felsenstein, Confidence limits on phylogenies: An approach using the bootstrap. *Evolution* **39**, 783–791 (1985). doi:10.1111/j.1558-5646.1985.tb00420.x Medline
69. A. Rambaut, T. T. Lam, L. Max Carvalho, O. G. Pybus, Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). *Virus Evol.* **2**, vew007 (2016). doi:10.1093/ve/vew007 Medline
70. A.-A. Popescu, K. T. Huber, E. Paradis, ape 3.0: New tools for distance-based phylogenetics and evolutionary analysis in R. *Bioinformatics* **28**, 1536–1537 (2012). doi:10.1093/bioinformatics/bts184 Medline
71. P. Sagulenko, V. Puller, R. A. Neher, TreeTime: Maximum-likelihood phylodynamic analysis. *Virus Evol.* **4**, vex042 (2018). doi:10.1093/ve/vex042 Medline
72. S. Wang, X. Xu, C. Wei, S. Li, J. Zhao, Y. Zheng, X. Liu, X. Zeng, W. Yuan, S. Peng, Molecular evolutionary characteristics of SARS-CoV-2 emerging in the United States. *J. Med. Virol.* **94**, 310–317 (2022). doi:10.1002/jmv.27331 Medline
73. G. Yu, Using ggtree to visualize data on tree-like structures. *Curr. Protoc. Bioinformatics* **69**, e96 (2020). doi:10.1002/cpbi.96 Medline
74. H. Wickham, ggplot2. *Wiley Interdiscip. Rev. Comput. Stat.* **3**, 180–185 (2011).

doi:10.1002/wics.147
75. J. Hadfield, C. Megill, S. M. Bell, J. Huddleston, B. Potter, C. Callender, P. Sagulenko, T. Bedford, R. A. Neher, Nextstrain: Real-time tracking of pathogen evolution. *Bioinformatics* **34**, 4121–4123 (2018). doi:10.1093/bioinformatics/bty407 Medline
76. S. E. James, CERI-KRISP/SARS-CoV-2-epidemic-in-Africa: Expanding Africa SARS-CoV-2 sequencing capacity in a fast evolving pandemic analysis. Zenodo (2022); https://doi.org/10.5281/zenodo.7006806.
77. T. Ward, A. Johnsen, Understanding an evolving pandemic: An analysis of the clinical time delay distributions of COVID-19 in the United Kingdom. *PLOS ONE* **16**, e0257978 (2021). doi:10.1371/journal.pone.0257978 Medline

ACKNOWLEDGMENTS

First and foremost, we acknowledge authors in institutions in Africa and beyond who have made invaluable contributions toward specimen collection and sequencing to produce and share, via GISAID, SARS-CoV-2 genomic data. We also acknowledge the authors from the originating and submitting laboratories worldwide, who generated and shared SARS-CoV-2 sequence data, via GISAID, from other regions in the world, which was used to contextualize the African genomic data. A full list of GISAID sequence IDs used in the current study are available in table S4. **Funding:** Sequencing efforts in the African Union Member States were supported by the Africa Center for Disease Control (Africa CDC) - Africa Pathogen Genomics Initiative (Africa PGI), and the World Health Organization Regional Office for Africa (WHO AFRO) through the transfer of laboratory infrastructure, the provision of reagents and training. The Africa PGI is supported by the African Union, Centers for Disease Control and Prevention (CDC), Bill and Melinda Gates Foundation (BMGF), Illumina Inc, Oxford Nanopore Technologies (ONT) and other partners. In addition, all Institut Pasteur organizations and CERMES in Niger are part of the PEPAIR COVID-19-Africa project which is funded by the French Ministry for European and Foreign Affairs. KRISP and CERI is supported in part by grants from WHO, the Abbott Pandemic Defense Coalition (APDC), the National Institute of Health USA (U01 AI151698) for the United World Antivirus Research Network (UWARN) and the INFORM Africa project through IHVN (U54 TW012041), H3BioNet Africa (Grant # 2020 HTH 062), the South African Department of Science and Innovation (SA DSI) and the South African Medical Research Council (SAMRC) under the BRICS JAF #2020/049. ILRI is also supported by the Ministry for Economic Cooperation and Federal Development of Germany (BMZ). Work conducted at ACEGID is made possible by support provided to ACEGID by a cohort of generous donors through TED's Audacious Project, including the ELMA Foundation, MacKenzie Scott, the Skoll Foundation, and Open Philanthropy. Work at ACEGID was also partly supported by grants from the National Institute of Allergy and Infectious Diseases (https://www.niaid.nih.gov), NIH-H3Africa (https://h3africa.org) (U01HG007480 and U54HG007480), the World Bank (projects ACE-019 and ACE-IMPACT), the Rockefeller Foundation (Grant #2021 HTH), the Africa CDC through the African Society of Laboratory Medicine (ASLM; Grant #INV018978), the Wellcome Trust (Project 216619/Z/19/Z) and the Science for Africa Foundation. Sequencing efforts at the National Institute for Communicable Diseases (NICD) was also supported by a conditional grant from the South African National Department of Health as part of the emergency COVID-19 response; a cooperative agreement between the NICD of the National Health Laboratory Service (NHLS) and the United States Centers for Disease Control and Prevention (FAIN# U01IP001048; NU51IP000930); the South African Medical Research Council (SAMRC, project number 96838); the ASLM and the Bill and Melinda Gates Foundation grant number INV-018978; the UK Foreign, Commonwealth and Development Office and Wellcome (Grant no 221003/Z/20/Z); and the UK Department of Health and Social Care and managed by the Fleming Fund and performed under the auspices of the SEQAFRICA project. Funding for sequencing efforts in Angola was supported through Projecto Bongola (N.º 11/MESCTI/PDCT/2020) and OGE INIS (2020/2021). Botswana's sequencing efforts led by the Botswana Harvard AIDS Institute Partnership was supported by: Foundation for Innovative New Diagnostics(FINDdx); BMGF, H3ABioNet [U41HG006941], Sub-Saharan African Network for TB/HIV Research Excellence (SANTHE) and Fogarty International Center (Grant # 5D43TW009610) . H3ABioNet is an initiative of the Human Health and Heredity in Africa Consortium (H3Africa) program of the African Academy of Science (AAS). HHS/NIH/National Institute of Allergy and Infectious Diseases (NIAID) (5K24AI131928-04; 5K24AI131924-04); SANTHE is a DELTAS Africa Initiative [grant # DEL-15-006]. The DELTAS Africa Initiative is an independent funding scheme of the African Academy of Sciences (AAS)'s Alliance for Accelerating Excellence in Science in Africa (AESA) and supported by the New Partnership for Africa's Development Planning and Coordinating Agency (NEPADAgency) with funding from the Wellcome Trust [grant #107752/Z/15/Z] and the UK government. From Brazil, Joicymara Santos Xavier was funded by Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brazil (CAPES) - Finance Code 001. Sequencing efforts from Côte d'Ivoire were funded by the Robert Koch Institute and the German Federal Ministry of Education and Research (BMBF). Sequencing efforts in the Democratic Republic of the Congo were funded by the Bill & Melinda Gates Foundation under grant INV-018030 awarded to CBP and further supported by funding from the Africa CDC through ASLM for Accelerating SARS-CoV-2 Genomic Surveillance in Africa, the US Centre for Disease Control and Prevention (US CDC), USAMRIID, IRD/Montepellier, UCLA and SACIDS FIND. Efforts from Egypt were funded by the Egyptian Ministry of Health, the Egyptian Academy for Scientific Research and Technology (ASRT) JESOR project #3046 (Center for Genome and Microbiome Research), the Cairo University anti COVID-19 fund and the Science and Technology Development Fund (STDF), Project ID: 41907. The sequencing effort in Equatorial Guinea was supported by a public–private partnership, the Bioko Island Malaria Elimination Project, composed of the government of Equatorial Guinea Ministries of Mines and Hydrocarbons, and Health and Social Welfare, Marathon EG Production Limited, Noble Energy, Atlantic Methanol Production Company, and EG LNG. Analysis for the Gabon strains was supported by the Science and Technology Research Partnership for Sustainable Development (SATREPS), Japan International Cooperation Agency (JICA), and Japan Agency for Medical Research and Development (AMED) (grant number JP21jm0110013) and a grant from AMED (grant number JP21wm0225003). CIRMF (Gabon) is funded by the Gabonese Government and TOTAL Energy inc. CIRMF is a member of CANTAM supported by EDCTP. The work at WACCBIP (Ghana) was funded by a grant from the Rockefeller Foundation (2021 HTH 006), an Institut de Recherche pour le Développement (IRD) grant (ARIACOV), African Research Universities Alliance (ARUA) Vaccine Development Hubs grant with funds from Open Society Foundation, National Institute of Health Research (NIHR) (17.63.91) grants using UK aid from the UK Government for a global health research group for Genomic surveillance of malaria in West Africa (Wellcome Sanger Institute, UK) and the World Bank African Centers of Excellence Impact grant (WACCBIP-NCDs: Awandare). In addition to the funding sources from ILRI, KEMRI (Kenyan) contributions to sequencing efforts was supported in part by the National Institute for Health Research (NIHR) (project references 17/63/82 and 16/136/33) using UK aid from the UK Government to support global health research, and The UK Foreign, Commonwealth and Development Office (FCDO) and Wellcome (grant# 220985/Z/20/Z) and the Kenya Medical Research Institute Grant # KEMRI/COV/SPE/012. Contributions from Lesotho were supported by the Africa CDC, ALSM and SA NICD. Liberian efforts was funded by the Africa CDC through a subaward from the Bill and Melinda Gates Foundations, while efforts from Madagascar were funded by the French Ministry for Europe and Foreign Affairs through the REPAIR COVID-19-Africa project coordinated by the Pasteur International Network association. Sequencing from Malawi was supported by Wellcome Trust. Contributions from Mali was supported by Fogarty International Center and National Institute of Allergy and Infectious Diseases sections of the National Institutes of Health under Leidos-15X051, award numbers U2RTW010673 for the West African Center of Excellence for Global Health Bioinformatics Research Training and U19AI089696 and U19AI129387 for the West Africa International Center of Excellence for Malaria Research. Funding for surveillance, sampling and testing in Madagascar: World Health Organization (WHO), the US Centers for Disease Control and Prevention (US CDC: Grant#U5/IP000812-05), the United States Agency for International Development (USAID: Cooperation Agreement 72068719CA00001), the Office of the Assistant Secretary for Preparedness and Response in the U.S. Department of Health and Human Services (DHHS: grant number IDSEP190051-

01-0200). Funding for sequencing: Bill & Melinda Gates Foundation (GCE/ID OPP1211841), Chan Zuckerberg Biohub, and the Innovative Genomics Institute at UC Berkeley. Mozambique acknowledges support from the Mozambican Ministry of Health and the President's Emergency Plan for AIDS Relief (PEPFAR) through the U.S. Centers for Disease Control and Prevention (CDC) under the terms of [grant number GH002021, GH001944], and the Bill & Melinda Gates Foundation, #OPP1214435. Namibian efforts was supported by Africa CDC through a subaward from the Bill and Melinda Gates Foundations. Efforts from the country Niger were supported by the French Ministry for Europe and Foreign Affairs through the REPAIR COVID-19-Africa project coordinated by the Pasteur International Network association. In addition to the funding support for ACEGID already listed, Nigeria's contributions were made possible by support from Flu Lab and a cohort of donors through the Audacious Project, a collaborative funding initiative housed at TED, including the ELMA Foundation, MacKenzie Scott, the Skoll Foundation, and Open Philanthropy. Efforts from the Republic of the Congo was supported by the European and Developing Countries Clinical Trials Partnership (EDCTP) IDs: PANDORA, CANTAM and German Academic Exchange Service (DAAD) IDs: PACE-UP; DAAD Project ID: 5759234. Rwanda's contributions were made possible by funding from the African Network for improved Diagnostics, Epidemiology and Management of common Infectious Agents (ANDEMIA) was granted by the German Federal Ministry of Education and Research (BMBF grant 01KA1606, 01KA2021 and 01KA2110B) and the National Institute of Health Research (NIHR) Global Health Research program (16/136/33) using UK aid from the UK Government. In addition to the South African institutions listed above, the University of Cape Town's work was supported by the Wellcome Trust [Grant # 203135/Z/16/Z], EDCTP RADIATES (RIA2020EF-3030), the South African Department of Science and Innovation (SA DSI) and the South African Medical Research Council (SAMRC), Stellenbosch University's contributions by the South African Medical Research Council (SA-MRC), and the University of Pretoria's contributions funded by the G7 Global Health Fund and a BMBF ANDEMIA grant. Funding from the Fleming Fund supported sequencing in Sudan. The Ministry of Higher Education and Scientific Research of Tunisia provided funding for sequencing from Tunisia. UVRI (Uganda) acknowledge support from the Wellcome Trust and FCDO - Wellcome Epidemic Preparedness – Coronavirus (AFRICO19, grant agreement number 220977/Z/20/Z), from the MRC (MC_UU_1201412) and from the UK Medical Research Council (MRC/UKRI) and FCDO (DIASEQCO, grant agreement number NC_PC_19060). Research at the FredHutch institute which supported bioinformatics analyses of sequences in the present study was supported by the Bill and Melinda Gates foundation (#INV-018979). Research support from Broad Institute colleagues was made possible by support from Flu lab and a cohort of generous donors through TED's Audacious Project, including the ELMA Foundation, MacKenzie Scott, the Skoll Foundation, Open Philanthropy, the Howard Hughes Medical Institute and NIH (U01AI151812 and U54HG007480) (P.C.S.). Work from Quadram Institute Bioscience was funded by The Biotechnology and Biological Sciences Research Council Institute Strategic Programme Microbes in the Food Chain BB/R012504/1 and its constituent projects BBS/E/F/000PR10348, BBS/E/F/000PR10349, BBS/E/F/000PR10351, and BBS/E/F/000PR10352 and by the Quadram Institute Bioscience BBSRC funded Core Capability Grant (project number BB/CCG1860/1). Sequences generated in Zambia through PATH were funded by the BMGF and Africa CDC. The content and findings reported herein are the sole deduction, view and responsibility of the researcher/s and do not reflect the official position and sentiments of the funding agencies. **Author contributions:** Conceptualization: HT, CB, SKT, TdO, RL, EW; Methodology: HT, JES, MC, BT, GM, DPM, AWL, DAR, LMK, GG, TdO, RL, EW; Genomic Data Generation: HT, JES, MC, MM, BT, GM, DPM, AWL, AD, DGA, MMD, AS, ANZ, ASG, AKS, AO, AS, AOM, AKS, AGA, AL, AK, AEA, AAJ, AF, AOO, AAA, AJ, AK, AM, AR, AS, AK, AB, AC, AJT, AC, AKK, AK, AB, AS, AA, AN, AVG, AN, AJP, AY, AV, ANH, AC, AI, AM, ALB, AI, AAS, AG, AF, AES, BM, BLS, BSO, BB, BD, BLH, BT, BL, BM, BN, BTM, BAK, BK, BA, BP, BM, CB, CW, CN, CA, CBP, CS, CGA, CNA, CMM, CL, CKO, CI, CNM, CP, CG, CEO, CDR, CMM, CE, DBL, DJB, DM, DP, DB, DJN, DS, DT, DSA, DG, DSG, DOO, DM, DWW, EF, EKL, ES, EMO, ENN, EOA, EO, ES, EB, EBA, EAA, EL, EM, EP, EB, ES, EAA, FL, FMT, FW, FA, FTT, FD, FVA, FT, FO, FN, FMM, FER, FAD, FI, GKM, GT, GLK, GOA, GUvZ, GAA, GS, GPM, HCR, HEO, HO, HA, HK, HN, HT, HAAK, HE, HG, HM, HK, IS, IBO, IMA, IO, IBB, IAM, IS, IW, ISK, JWAH, JA, JS, JCM, JMT, JH, JGS, JG, JM, JN, JNU, JNB, JY, JM, JK, JDS, JH, JKO, JMM, JOG, JTK, JCO, JSX, JG, JFW, JHB, JN, JE, JN, JMN, JN, JUO, JCA, JJL, JJHM, JO, KJS, KV, KTA, KAT, KSC, KSM, KD, KGM, KOD, LF, LS, LMK, LB, LdOM, LC, LO, LDO, LLD, LIO, LT, MM, MR, MM, ME, MM, MIM, MK, MD, MM, MdLLM, MV, MFP, MF, MMN, MM, MD, MWM, MGM, MO, MRW, MYT, MOA, MA, MAB, MGS, MKK, MMM, MK, MS, MBM, MM, MA, MVP, NA, NR, NA, NI, NE, NMT, ND, NM, NH, NBS, NMF, NS, NB, NM, NG, NW, NS, NN, NAA, NT, NM, NHR, NI, NM, OCK, OS, OF, OMA, OT, OAO, OF, OEO, O-EO, OF, PS, PO, PC, PN, PS, PEO, PA, PKQ, POO, PB, PD, PAB, PKM, PK, PA, RE, RJ, RKA, RGE, RA, RN, ROP, RG, RAK, RMND, RAA, RAC, SG, SM, SB, SS, SIM, SF, SM, SH, SKK, SM, ST, SHA, SWM, SD, SM, SA, SSA, SMA, SE, SM, SL, SG, SJ, SFA, SO, SG, SL, SP, SO, SvW, SFS, SK, SA, SR, SP, SN, SB, SLB, SvdW, TM, TM, TL, TPV, TS, TGM, TB, UJA, UC, UR, UEG, VE, VN, VG, WHR, WAK, WKA, WP, WTC, YAA, YR, YB, YN, YB, ZRdL, AEO, AvG, GG, MM, OT, PCS, AAS, SOO, YKT, SKT, TdO, CH, RL, JN, EW; Data Analysis: HT, JES, MC, MM, BT, GM, DPM, AWL, AlE, DAR, EM, GSK, SvW, GG, TdO, RL, EW; Funding acquisition: AEO, AvG, GG, MM, OT, AAS, SOO, YKT, SKT, TdO, CH; Project administration: GM, AD, DGA, MMD, AC, DWW, HO, SWM, AEO, AvG, GG, MM, OT, PCS, AAS, SOO, YKT, SKT, TdO, CH, RL, JN, EW; Supervision: AEO, AvG, GG, MM, OT, PCS, AAS, SOO, YKT, SKT, TdO, CH, RL, JN, EW; Writing – original draft: HT, JES, MC, GM, DPM, CB, SKT, TdO, RL, EW; Writing – review and editing: HT, JES, MC, MM, BT, GM, DPM, CB, AWL, AD, DGA, MMD, AS, ANZ, ASG, AKS, AO, AS, AOM, AKS, AIE, AL, AK, AEA, AAJ, AF, AOO, AAA, AJ, AK, AM, AR, AS, AK, AB, AC, AJT, AC, AKK, AK, AB, AS, AA, AVG, AJP, AY, AV, ANH, AC, AI, AM, ALB, AI, AAS, AG, AF, AES, BM, BLS, BSO, BB, BD, BLH, BT, BL, BM, BN, BTM, BAK, BK, BA, BP, BM, CB, CW, CA, CBP, CS, CGA, CNA, CMM, CL, CKO, CI, CNM, CP, CEO, CDR, CMM, CE, DBL, DJB, DM, DP, DB, DJN, DS, DT, DSA, DG, DSG, DOO, DM, DWW, EF, EKL, ES, EMO, ENN, EOA, EO, ES, EB, EBA, EL, EM, EP, EB, ES, EAA, EM, FL, FMT, FW, FA, FTT, FD, FVA, FT, FO, FN, FMM, FER, FAD, FI, GKM, GT, GLK, GOA, GUvZ, GAA, GSK, GS, GPM, HCR, HEO, HO, HA, HK, HN, HT, HAAK, HE, HG, HM, HK, IS, IBO, IMA, IO, IBB, IS, IW, ISK, JWAH, JA, JS, JCM, JMT, JH, JGS, JG, JM, JNU, JNB, JY, JM, JK, JDS, JH, JKO, JMM, JOG, JTK, JCO, JSX, JG, JHB, JN, JE, JN, JMN, JN, JUO, JCA, JJL, JO, KJS, KV, KTA, KAT, KSC, KSM, KD, KGM, KOD, LF, LS, LB, LdOM, LC, LO, LLD, LIO, MM, MR, MM, ME, MM, MIM, MK, MD, MM, MdLLM, MV, MFP, MF, MMN, MM, MD, MWM, MGM, MO, MRW, MYT, MOA, MA, MAB, MGS, MKK, MMM, MK, MS, MBM, MM, MVP, NA, NR, NI, NMT, ND, NM, NH, NBS, NMF, NS, NB, NM, NG, NW, NS, NN, NAA, NT, NM, NHR, NI, NM, OCK, OS, OF, OMA, OT, OAO, OF, OEO, OF, PS, PO, PC, PN, PS, PEO, PA, PKQ, POO, PB, PD, PAB, PKM, PK, PA, RE, RJ, RKA, RGE, RA, RN, ROP, RG, RAK, RAA, RAC, SG, SM, SS, SIM, SF, SM, SH, SKK, SM, ST, SHA, SWM, SD, SM, SA, SSA, SMA, SE, SM, SL, SG, SJ, SFA, SO, SG, SL, SP, SO, SvW, SFS, SK, SA, SR, SP, SN, SB, SLB, SvdW, TM, TM, TL, TPV, TS, TGM, TB, UJA, UC, UR, UEG, VE, VN, VG, WHR, WAK, WKA, WP, WTC, YAA, YR, YB, YN, YB, ZRdL, AEO, AvG, GG, MM, OT, PCS, AAS, SOO, YKT, SKT, TdO, CH, RL, JN, EW. **Competing interests:** With the exception of Pardis Sabeti who is a co-founder of and consultant to Sherlock Biosciences and a Board Member of Danaher Corporation and who holds equity in the companies, we the authors have no conflicts of interest to declare. **Data and materials availability:** All of the SARS-CoV-2 whole genome sequences that were analyzed in the present study are all publicly available on the GISAID sequence database. A full list of the African sequences as well as global references are presented and acknowledged in table S4 and in our github repository (https://github.com/CERI-KRISP/SARS-CoV-2-epidemic-in-Africa) (*76*). The repositories also contain all of the metadata, raw and time scaled ML tree topologies, annotated tree topologies as well as data analysis and visualization scripts used here which will allow for the independent reproduction of results. Furthermore, the repositories also contain all Institutional Review Board (IRB) and Material Transfer Agreements (MTA). Please refer to the Ethics Statement in the Methods section for more details.

Full list of authors and affiliations

Houriiyah Tegally[1,2]†, James E. San[1,2], Matthew Cotten[3,4], Monika Moir[1], Bryan Tegomoh[5,6], Gerald Mboowa[7], Darren P. Martin[8,9], Cheryl Baxter[1,10], Arnold W. Lambisia[11], Amadou Diallo[12], Daniel G. Amoako[13,14], Moussa M. Diagne[12], Abay Sisay[15,16], Abdel-Rahman N. Zekri[17], Abdou Salam Gueye[18], Abdoul K. Sangare[19], Abdoul-Salam Ouedraogo[20], Abdourahmane Sow[21], Abdualmoniem O. Musa[22,23,24], Abdul K. Sesay[25], Abe G. Abias[26], Adam I. Elzagheid[27], Adamou Lagare[28], Adedotun-Sulaiman Kemi[29], Aden Elmi Abar[30], Adeniji A. Johnson[31,32], Adeola Fowotade[33,34], Adeyemi O. Oluwapelumi[35,36], Adrienne A. Amuri[37,38], Agnes Juru[39], Ahmed Kandeil[40], Ahmed Mostafa[40], Ahmed Rebai[41], Ahmed Sayed[42], Akano Kazeem[43,44], Aladje Balde[45,46], Alan Christoffels[7,47], Alexander J. Trotter[48], Allan Campbell[49], Alpha K. Keita[50,51], Amadou Kone[52], Amal Bouzid[41,53], Amal Souissi[41], Ambrose Agweyu[11], Amel Naguib[54], Ana V. Gutierrez[48], Anatole Nkeshimana[55], Andrew J. Page[48], Anges Yadouleton[56], Anika Vinze[57], Anise N. Happi[43], Anissa Chouikha[58,59], Arash Iranzadeh[8,9], Arisha Maharaj[1], Armel L. Batchi-Bouyou[60,61], Arshad Ismail[13], Augustina A. Sylverken[62,63], Augustine Goba[64,65], Ayoade Femi[43,44], Ayotunde E. Sijuwola[43], Baba Marycelin[66,67], Babatunde L. Salako[29,32], Bamidele S. Oderinde[66], Bankole Bolajoko[43], Bassirou Diarra[52], Belinda L. Herring[18], Benjamin Tsofa[11], Bernard Lekana-Douki[68,69], Bernard Mvula[70], Berthe-Marie Njanpop-Lafourcade[18], Blessing T. Marondera[71], Bouh Abdi Khaireh[72,73], Bourema Kouriba[19], Bright Adu[74], Brigitte Pool[75], Bronwyn McInnis[17], Cara Brook[76,77], Carolyn Williamson[9,10,78], Cassien Nduwimana[55], Catherine Anscombe[79,80], Catherine B. Pratt[81], Cathrine Scheepers[13,82], Chantal G. Akoua-Koffi[83,84], Charles N. Agoti[11,85], Chastel M. Mapanguy[60,86], Cheikh Loucoubar[12], Chika K. Onwuamah[87], Chikwe Ihekweazu[88], Christian N. Malaka[89], Christophe Peyrefitte[12], Chukwa Grace[43,44], Chukwuma E. Omoruyi[33,34], Clotaire D. Rafai[90], Collins M. Morang'a[91], Cyril Erameh[92], Daniel B. Lule[3], Daniel J. Bridges[93], Daniel Mukadi-Bamuleka[37], Danny Park[57], David A. Rasmussen[94,95], David Baker[48], David J. Nokes[11,96], Deogratius Ssemwanga[3,97], Derek Tshiabuila[2], Dominic S. Y. Amuzu[91], Dominique Goedhals[98], Donald S. Grant[64,65,99], Donwilliams O. Omuoyo[11], Dorcas Maruapula[100], Dorcas W. Wanjohi[7], Ebenezer Foster-Nyarko[48], Eddy K. Lusamaki[37,38,51], Edgar Simulundu[101], Edidah M. Ong'era[11], Edith N. Ngabana[37,38], Edward O. Abworo[102], Edward Otieno[11], Edwin Shumba[71], Edwine Barasa[11], El Bara Ahmed[103,104], Elhadi A. Ahmed[23], Emmanuel Lokilo[37], Enatha Mukantwari[105], Eromon Philomena[43], Essia Belarbi[106], Etienne Simon-Loriere[107], Etilé A. Anoh[83], Eusebio Manuel[108], Fabian Leendertz[106], Fahn M. Taweh[109], Fares Wasfi[58], Fatma Abdelmoula[41,110], Faustinos T. Takawira[39], Fawzi Derrar[111], Fehintola V. Ajogbasile[43], Florette Treurnicht[112,113], Folarin Onikepe[43,44], Francine Ntoumi[60,114], Francisca M. Muyembe[37,38], Frank E. Z. Ragomzingba[115], Fred A. Dratibi[116,117], Fred-Akintunwa Iyanu[43], Gabriel K. Mbunsu[38], Gaetan Thilliez[48], Gemma L. Kay[48], George O. Akpede[92], Gert U. van Zyl[118,119], Gordon A. Awandare[91], Grace S. Kpeli[120,121], Grit Schubert[106], Gugu P. Maphalala[122], Hafaliana C. Ranaivoson[77], Hannah E. Omunakwe[123], Harris Onywera[7], Haruka Abe[124], Hela Karray[125], Hellen Nansumba[126], Henda Triki[58], Herve Albéric Adje Kadjo[127], Hesham Elgahzaly[128], Hlanai Gumbo[39], Hota Mathieu[129], Hugo Kavunga-Membo[37], Ibtihel Smeti[41], Idowu B. Olawoye[43], Ifedayo M. O. Adetifa[88,130], Ikponmwosa Odia[92], Ilhem Boutiba-Ben Boubaker[131,132], Iluoreh Ahmed Mohammad[43], Isaac Ssewanyana[126], Isatta Wurie[133], Iyaloo S. Konstantinus[134], Jacqueline Wemboo Afiwa Halatoko[135], James Ayei[26], Janaki Sonoo[136], Jean-Claude C. Makangara[37,38], Jean-Jacques M. Tamfum[37,38], Jean-Michel Heraud[12,77], Jeffrey G. Shaffer[137], Jennifer Giandhari[2], Jennifer Musyoki[11], Jerome Nkurunziza[138], Jessica N. Uwanibe[43], Jinal N. Bhiman[13,113], Jiro Yasuda[124], Joana Morais[139,140], Jocelyn Kiconco[97], John D. Sandi[64,65], John Huddleston[141], John K. Odoom [74], John M. Morobe[11], John O. Gyapong[120], John T. Kayiwa[3], Johnson C. Okolie[43], Joicymara S. Xavier[1,142,143], Jones Gyamfi[120], Joseph F. Wamala[144], Joseph H. K. Bonney[74], Joseph Nyandwi[55,145], Josie Everatt[13], Joweria Nakaseegu[97], Joyce M. Ngoi[91], Joyce Namulondo[97], Judith U. Oguzie[43,44], Julia C. Andeko[68], Julius J. Lutwama[3], Juma J. H. Mogga[144], Justin O'Grady[48], Katherine J. Siddle[57], Kathleen Victoir[146], Kayode T. Adeyemi[43,44], Kefentse A. Tumedi[147], Kevin S. Carvalho[148], Khadija Said Mohammed[11], Koussay Dellagi[146], Kunda G. Musonda[149], Kwabena O. Duedu[120,121], Lamia Fki-Berrajah[125], Lavanya Singh[2], Lenora M. Kepler[94,95], Leon Biscornet[75], Leonardo de Oliveira Martins[48], Lucious Chabuka[150], Luicer Olubayo[8], Lul Deng Ojok[26], Lul Lojok Deng[26], Lynette I. Ochola-Oyier[11], Lynn Tyers[9], Madisa Mine[151], Magalutcheemee Ramuth[136], Maha Mastouri[152,153], Mahmoud ElHefnawi[154], Maimouna Mbanne[12], Maitshwarelo I. Matsheka[147], Malebogo Kebabonye[155], Mamadou Diop[12], Mambu Momoh[64,65,156], Maria da Luz Lima Mendonça[148], Marietjie Venter[157], Marietou F. Paye[57], Martin Faye[12], Martin M. Nyaga[158], Mathabo Mareka[159], Matoke-Muhia Damaris[160], Maureen W. Mburu[11], Maximillian G. Mpina[161,162,163], Michael Owusu[164], Michael R. Wiley[81,165], Mirabeau Y. Tatfeng[166], Mitoha Ondo'o Ayekaba[162], Mohamed Abouelhoda[167,168], Mohamed Amine Beloufa[111], Mohamed G. Seadawy[169,170], Mohamed K. Khalifa[171], Mooko Marethabile Matobo[159], Mouhamed Kane[12], Mounerou Salou[172], Mphaphi B. Mbulawa[155], Mulenga Mwenda[93], Mushal Allam[173], My V. T. Phan[3], Nabil Abid[152,174], Nadine Rujeni[175,176], Nadir Abuzaid[177], Nalia Ismael[178], Nancy Elguindy[54], Ndeye Marieme Top[12], Ndongo Dia[12], Nédio Mabunda[178], Nei-yuan Hsiao[9,78], Nelson Boricó Silochi[162], Ngiambudulu M. Francisco[139], Ngonda Saasa[179], Nicholas Bbosa[3], Nickson Murunga[11], Nicksy Gumede[18], Nicole Wolter[13,113], Nikita Sitharam[1], Nnaemeka Ndodo[88], Nnennaya A. Ajayi[180], Noël Tordo[181], Nokuzola Mbhele[9], Norosoa H. Razanajatovo[77], Nosamiefan Iguosadolo[43], Nwando Mba[88], Ojide C. Kingsley[182], Okogbenin Sylvanus[92], Oladiji Femi[183], Olubusuyi M. Adewumi[31,32], Olumade Testimony[43,44], Olusola A. Ogunsanya[43], Oluwatosin Fakayode[184], Onwe E. Ogah[185], Ope-Ewe Oludayo[43], Ousmane Faye[12], Pamela Smith-Lawrence[155], Pascale Ondoa[71], Patrice Combe[186], Patricia Nabisubi[187,188], Patrick Semanda[126], Paul E. Oluniyi[43], Paulo Arnaldo[178], Peter Kojo Quashie[91], Peter O. Okokhere[92,189], Philip Bejon[11], Philippe Dussart[77], Phillip A. Bester[190], Placide K. Mbala[37,38], Pontiano Kaleebu[3,97], Priscilla Abechi[43,44], Rabeh El-Shesheny[40,191], Rageema Joseph[9], Ramy Karam Aziz[192,193], René G. Essomba[194,195], Reuben Ayivor-Djanie[91,120,121], Richard Njouom[196], Richard O. Phillips[63], Richmond Gorman[63], Robert A. Kingsley[48], Rosa Maria D. E. S. A. Neto Rodrigues[197,198], Rosemary A. Audu[29], Rosina A. A. Carr[120,121], Saba Gargouri[125], Saber Masmoudi[41], Sacha Bootsma[144], Safietou Sankhe[12], Sahra Isse Mohamed[199], Saibu Femi[43], Salma Mhalla[132,200], Salome Hosch[161,201], Samar Kamal Kassim[128], Samar Metha[57], Sameh Trabelsi[202], Sara Hassan Agwa[128], Sarah Wambui Mwangi[7], Seydou Doumbia[52], Sheila Makiala-Mandanda[37,38], Sherihane Aryeetey[63], Shymaa S. Ahmed[54], Side Mohamed Ahmed[103], Siham Elhamoumi[57], Sikhulile Moyo[100,203], Silvia Lutucuta[139], Simani Gaseitsiwe[100,203], Simbirie Jalloh[64,65], Soa Fy Andriamandimby[77], Sobajo Oguntope[43], Solène Grayo[181], Sonia Lekana-Douki[68], Sophie Prosolek[48], Soumeya Ouangraoua[204,205], Stephanie van Wyk[1], Stephen F. Schaffner[57], Stephen Kanyerezi[187,188], Steve Ahuka-Mundeke[37,38], Steven Rudder[48], Sureshnee Pillay[2], Susan Nabadda[126], Sylvie Behillil[206], Sylvie L. Budiaki[159], Sylvie van der Werf[206], Tapfumanei Mashe[39,207], Thabo Mohale[13], Thanh Le-Viet[48], Thirumalaisamy P. Velavan[114,208], Tobias Schindler[161,162,201], Tongai G. Maponga[118], Trevor Bedford[141,209], Ugochukwu J. Anyaneji[2], Ugwu Chinedu[43,44], Upasana Ramphal[2,10,210], Uwem E. George[43], Vincent Enouf[206], Vishvanath Nene[102], Vivianne Gorova[211,212], Wael H. Roshdy[54], Wasim Abdul Karim[1], William K. Ampofo[213], Wolfgang Preiser[118,119], Wonderful T. Choga[100,214], Yahaya Ali Ahmed[18], Yajna Ramphal[1], Yaw Bediako[91,215], Yeshnee Naidoo[2], Yvan Butera[175,216,217], Zaydah R. de Laurent[11], Africa Pathogen Genomics Initiative (Africa PGI), Ahmed E. O. Ouma[7], Anne von Gottberg[13,113], George Githinji[11,218], Matshidiso Moeti[18], Oyewale Tomori[43], Pardis C. Sabeti[57], Amadou A. Sall[12], Samuel O. Oyola[102], Yenew K. Tebeje[7], Sofonias K. Tessema[7], Tulio de Oliveira[1,2,10,219]*, Christian Happi[43,44], Richard Lessells[2], John Nkengasong[7], Eduan Wilkinson[1,2]†*

[1]Centre for Epidemic Response and Innovation (CERI), School of Data Science and Computational Thinking, Stellenbosch University, Stellenbosch, South Africa. [2]KwaZulu-Natal Research Innovation and Sequencing Platform (KRISP), Nelson R Mandela School of Medicine, University of KwaZulu-Natal, Durban, South Africa. [3]MRC/UVRI and LSHTM Uganda Research Unit, Entebbe, Uganda. [4]MRC-University of Glasgow Centre for Virus Research, Glasgow, UK. [5]The Biotechnology Centre of the University of Yaoundé I, Yaoundé, Cameroon. [6]CDC Foundation, Atlanta, Georgia, Nebraska Department of Health and Human Services, Lincoln, NE, USA. [7]Institute of Pathogen Genomics, Africa Centres for Disease Control and Prevention (Africa CDC), Addis Ababa, Ethiopia. [8]Institute of Infectious Diseases and Molecular Medicine, Department of Integrative Biomedical Sciences, Computational Biology Division, University of Cape Town, Cape Town, South Africa. [9]Division of Medical Virology, Wellcome Centre for Infectious Diseases in Africa, Institute of Infectious Disease and Molecular Medicine, University of Cape Town, Cape Town, South Africa. [10]Centre for the AIDS Programme of Research in South Africa (CAPRISA), Durban, South Africa. [11]KEMRI-Wellcome Trust Research Programme, Kilifi, Kenya. [12]Virology Department, Institut Pasteur de Dakar, Dakar, Senegal. [13]National Institute for Communicable Diseases (NICD) of the National Health Laboratory Service (NHLS), Johannesburg, South Africa. [14]School of Health Sciences, College of Health Sciences, University of KwaZulu-Natal, Durban, KwaZulu-Natal, South Africa. [15]Department of Medical Laboratory Sciences, College of Health Sciences, Addis Ababa University, Addis Ababa, Ethiopia. [16]Department of Microbial, Cellular and Molecular Biology, College of Natural and Computational Sciences, Addis Ababa University, Addis Ababa, Ethiopia. [17]Cancer Biology Department, Virology and Immunology Unit, National Cancer Institute, Cairo University, Cairo,

Egypt. [18]World Health Organization, Africa Region, Brazzaville, Republic of the Congo. [19]Centre d'Infectiologie Charles Mérieux-Mali (CICM-Mali), Bamako, Mali. [20]Bacteriology and Virology Department Souro Sanou University Hospital, Bobo-Dioulasso, Burkina Faso. [21]West African Health Organisation, Bobo-Dioulasso, Burkina Faso. [22]Faculty of Medicine and Health Sciences, Kassala University, Kassala City, Sudan. [23]Department of Microbiology, Faculty of Medical Laboratory Sciences, University of Gezira, Gezira, Sudan. [24]General Administration of Laboratories and Blood Banks, Ministry of Health, Kassala State, Sudan. [25]MRC Unit The Gambia at LSHTM, Fajara, Gambia. [26]National Public Health Laboratory, Ministry of Health, Juba, Republic of South Sudan. [27]Libyan Biotechnology Research Center, Tripoli, Libya. [28]Center for Medical and Sanitary Research (CERMES), Niamey, Niger. [29]The Nigerian Institute of Medical Research, Yaba, Lagos, Nigeria. [30]Laboratoire de la Caisse Nationale de Sécurité Sociale, Djibouti, Republic of Djibouti. [31]Department of Virology, College of Medicine, University of Ibadan, Ibadan, Nigeria. [32]Infectious Disease Institute, College of Medicine, University of Ibadan, Ibadan, Nigeria. [33]Medical Microbiology and Parasitology Department, College of Medicine, University of Ibadan, Ibadan, Nigeria. [34]Biorepository Clinical Virology Laboratory, College of Medicine, University of Ibadan, Ibadan, Nigeria. [35]Department of Medical Microbiology and Parasitology, Faculty of Basic Clinical Sciences, College of Health Sciences, University of Ilorin, Ilorin, Kwara State, Nigeria. [36]The Pirbright Institute, Woking, UK. [37]Pathogen Sequencing Lab, Institut National de Recherche Biomédicale (INRB), Kinshasa, the Democratic Republic of the Congo. [38]Université de Kinshasa (UNIKIN), Kinshasa, the Democratic Republic of the Congo. [39]National Microbiology Reference Laboratory, Harare, Zimbabwe. [40]Center of Scientific Excellence for Influenza Viruses, National Research Centre (NRC), Cairo, Egypt. [41]Laboratory of Molecular and Cellular Screening Processes, Centre of Biotechnology of Sfax, University of Sfax, Sfax, Tunisia. [42]Genomics and Epigenomics Program, Research Department CCHE57357, Cairo, Egypt. [43]African Centre of Excellence for Genomics of Infectious Diseases (ACEGID), Redeemer's University, Ede, Osun State, Nigeria. [44]Department of Biological Sciences, Faculty of Natural Sciences, Redeemer's University, Ede, Osun State, Nigeria. [45]Laboratório de Biologia Molecular Jean Piaget, Bissau, Guinea-Bissau. [46]University Jean Piaget in Guinea-Bissau, Bissau, Guinea-Bissau. [47]SAMRC Bioinformatics Unit, SA Bioinformatics Institute, University of the Western Cape, Cape Town, South Africa. [48]Quadram Institute Bioscience, Norwich, UK. [49]Central Public Health Reference Laboratories, Freetown, Sierra Leone. [50]Centre de Recherche et de Formation en Infectiologie de Guinée (CERFIG), Université de Conakry, Conakry, Guinea. [51]TransVIHMI, Institut de Recherche pour le Développement, Institut National de la Santé et de la Recherche Médicale (INSERM), Montpellier University, 34090, Montpellier, France. [52]University Clinical Research Center (UCRC), University of Sciences, Techniques and Technology of Bamako, Bamako, Mali. [53]Sharjah Institute for Medical Research, College of Medicine, University of Sharjah, Sharjah, United Arab Emirates. [54]Central Public Health Laboratories (CPHL), Cairo, Egypt. [55]National Institute of Public Health, Bujumbura, Burundi. [56]Laboratoire des Fièvres Hémorragiques Virales du Benin, Cotonou, Benin. [57]Infectious Disease and Microbiome Program, Broad Institute of Harvard and MIT, Cambridge, MA, USA. [58]Laboratory of Clinical Virology, WHO Reference Laboratory for Poliomyelitis and Measles in the Eastern Mediterranean Region, Pasteur Institute of Tunis, University Tunis El Manar (UTM), Tunis 1002, Tunisia. [59]Research Laboratory "Virus, Vectors and Hosts: One Health Apporach and Technological Innovation for a Better Health", LR20IPT02, Pasteur Institute, Tunis 1002, Tunisia. [60]Fondation Congolaise pour la Recherche Médicale, Brazzaville, Republic of the Congo. [61]Marien Ngouabi, Brazzaville, Republic of the Congo. [62]Kwame Nkrumah University of Science and Technology, Department of Theoretical and Applied Biology, Kumasi, Ghana. [63]Kumasi Centre for Collaborative Research in Tropical Medicine, Kwame Nkrumah University of Science and Technology, Kumasi, Ghana. [64]Viral Haemorrhagic Fever Laboratory, Kenema Government Hospital, Kenema, Sierra Leone. [65]Ministry of Health and Sanitation, Freetown, Sierra Leone. [66]Department of Immunology, University of Maiduguri Teaching Hospital, P.M.B. 1414, Maiduguri, Nigeria. [67]Department of Medical Laboratory Science, College of Medical Sciences, University of Maiduguri, P.M.B. 1069, Maiduguri, Borno State, Nigeria. [68]Centre Interdisciplinaires de Recherches Medicales de Franceville (CIRMF), Franceville, Gabon. [69]Département de Parasitologie-Mycologie Université des Sciences de la Santé (USS), Libreville, Gabon. [70]National HIV Reference Laboratory, Community Health Sciences Unit, Ministry of Health, Lilongwe, Malawi. [71]African Society for Laboratory Medicine, Addis Ababa, Ethiopia. [72]National Medical and Molecular Biology Laboratory, Ministry of Health, Djibouti, Republic of Djibouti. [73]Africa CDC, Rapid Responder, Team Djibouti, Djibouti, Djibouti. [74]Noguchi Memorial Institute for Medical Research, University of Ghana, Legon, Ghana. [75]Seychelles Public Health Laboratory, Public Health Authority, Ministry of Health Seychelles, Victoria, Seychelles. [76]Department of Ecology and Evolution, University of Chicago, Chicago, IL, USA. [77]Virology Unit, Institut Pasteur de Madagascar, Antananarivo, Madagascar. [78]National Health Laboratory Service (NHLS), Cape Town, South Africa. [79]Malawi-Liverpool-Wellcome Trust Clinical Research Programme, Blantyre, Malawi. [80]Liverpool School of Tropical Medicine, Liverpool, UK. [81]University of Nebraska Medical Center (UNMC), Omaha, NE, USA. [82]SAMRC Antibody Immunity Research Unit, School of Pathology, University of the Witwatersrand, Johannesburg, South Africa. [83]CHU de Bouaké, Laboratoire/Unité de Diagnostic des Virus des Fièvres Hémorragiques et Virus Émergents, Bouaké, Côte d'Ivoire. [84]UFR Sciences Médicales, Universite Alassane Ouattara, Bouaké, Côte d'Ivoire. [85]School of Public Health, Pwani University, Kilifi, Kenya. [86]Faculty of Science and Techniques, University Marien Ngouabi, Brazzaville, Republic of the Congo. [87]Centre for Human Virology and Genomics, Nigerian Institute of Medical Research, Yaba, Lagos, Nigeria. [88]Nigeria Centre for Disease Control and Prevention, Abuja, Nigeria. [89]Laboratoire des Arbovirus, Fièvres Hémorragiques virales, Virus Emergents et Zoonoses, Institut Pasteur de Bangui, Bangui, Central African Republic. [90]Le Laboratoire National de Biologie Clinique et de Santé Publique (LNBCSP), Bangui, Central African Republic. [91]West African Centre for Cell Biology of Infectious Pathogens (WACCBIP), College of Basic and Applied Sciences, University of Ghana, Accra, Ghana. [92]Institute of Lassa Fever Research and Control, Irrua Specialist Teaching Hospital, Irrua, Nigeria. [93]PATH, Lusaka, Zambia. [94]Department of Entomology and Plant Pathology, North Carolina State University, Raleigh, NC, USA. [95]Bioinformatics Research Center, North Carolina State University, Raleigh, NC, USA. [96]School of Life Sciences and Zeeman Institute for Systems Biology and Infectious Disease Epidemiology Research (SBIDER), University of Warwick, Coventry, UK. [97]Uganda Virus Research Institute, Entebbe, Uganda. [98]PathCare Vermaak, Pretoria, South Africa and Division of Virology, University of the Free State, Bloemfontein, South Africa. [99]College of Medicine and Allied Health Sciences, University of Sierra Leone, Freetown, Sierra Leone. [100]Botswana Harvard AIDS Institute Partnership and Botswana Harvard HIV Reference Laboratory, Gaborone, Botswana. [101]Macha Research Trust, Choma, Zambia. [102]International Livestock Research Institute (ILRI), Nairobi, Kenya. [103]INRSP, Nouakchott, Mauritania. [104]Faculté de Médecine de Nouakchott, Nouakchott, Mauritani. [105]Rwanda National Reference Laboratory, Kigali, Rwanda. [106]Robert Koch-Institute, Berlin, Germany. [107]G5 Evolutionary Genomics of RNA Viruses, Institut Pasteur, Paris, France. [108]Direcção Nacional da Saúde Pública, Ministério da Saúde, Luanda, Angola. [109]National Public Health Reference Laboratory–National Public Health Institute of Liberia, Monrovia, Liberia. [110]Faculty of Pharmacy of Monastir, Monastir, Tunisia. [111]National Influenza Centre, Institut Pasteur d'Algérie, Algiers, Algeria. [112]Department of Virology, National Health Laboratory Service (NHLS), Charlotte Maxeke Johannesburg Academic Hospital, Johannesburg, South Africa. [113]School of Pathology, Faculty of Health Science, University of the Witwatersrand, Johannesburg, South Africa. [114]Institute of Tropical Medicine, Universitätsklinikum Tübingen, Tübingen, Germany. [115]Ministère de Santé Publique et de la Solidarité Nationale, Ndjamena, Chad. [116]WHO Int Comoros, Moroni, Union of Comoros. [117]World Health Organization, Africa Region, Brazzaville, Republic of the Congo. [118]Division of Medical Virology, Faculty of Medicine and Health Sciences, Stellenbosch University, Tygerberg, Cape Town, South Africa. [119]National Health Laboratory Service (NHLS), Tygerberg, Cape Town, South Africa. [120]UHAS COVID-19 Testing and Research Centre, University of Health and Allied Sciences, Ho, Ghana. [121]Department of Biomedical Sciences, University of Health and Allied Sciences, PMB 31, Ho, Ghana. [122]Ministry of Health, COVID-19 Testing Laboratory, Mbabane, Kingdom of Eswatini. [123]Satellite Molecular Laboratory, Rivers State University Teaching Hospital, Port Harcourt, Nigeria. [124]Department of Emerging Infectious Diseases, Institute of Tropical Medicine, Nagasaki University, Nagasaki, Japan. [125]CHU Habib Bourguiba, Laboratory of Microbiology, Faculty of Medicine of Sfax, University of Sfax, Sfax, Tunisia. [126]Central Public Health Laboratories (CPHL), Kampala, Uganda. [127]Institut Pasteur de Côte d'Ivoire, Departement des Virus Epidemiques, Abidjan, Côte d'Ivoire. [128]Faculty of Medicine Ain Shams Research Institute (MASRI), Ain Shams University, Cairo, Egypt. [129]Doctoral School of Technical and Environmental Sciences, Department of Biology and Human Health, N'Djamena, Chad. [130]Department of Infectious Diseases Epidemiology, London School of Hygiene and Tropical Medicine, London, UK. [131]Charles Nicolle Hospital, Laboratory of Microbiology, National Influenza Center, Tunis, Tunisia. [132]University of Tunis El Manar, Faculty of Medicine of Tunis, Research Laboratory LR99ES09, Tunis, Tunisia. [133]College of Medicine and Allied Health Science, University of Sierra Leone,

Freetown, Sierra Leone. [134]Namibia Institute of Pathology, Windhoek, Namibia. [135]National Institute of Hygiene, Lomé, Togo. [136]Virology/Molecular Biology Department, Central Health Laboratory, Victoria Hospital, Ministry of Health and Wellness, Port Louis, Mauritius. [137]Department of Biostatistics and Data Science, School of Public Health and Tropical Medicine, Tulane University, New Orleans, LA, USA. [138]WHO Burundi, Gitega, Burundi. [139]Grupo de Investigação Microbiana e Imunológica, Instituto Nacional de Investigação em Saúde (National Institute for Health Research), Luanda, Angola. [140]Departamento de Bioquímica, Faculdade de Medicina, Universidade Agostinho Neto, Luanda, Angola. [141]Vaccine and Infectious Disease Division, Fred Hutchinson Cancer Center, Seattle, WA, USA. [142]Universidade Federal de Minas Gerais, Belo Horizonte, Brazil. [143]Institute of Agricultural Sciences, Universidade Federal dos Vales do Jequitinhonha e Mucuri, Unaí, Brazil. [144]WHO South Sudan, Juba, South Sudan. [145]Faculty of Medicine, University of Burundi, Bujumbura, Burundi. [146]Pasteur Network, Institut Pasteur, Paris, France. [147]Botswana Institute for Technology Research and Innovation, Gaborone, Botswana. [148]Instituto Nacional de Saúde Pública, Praia, Cape Verde. [149]Zambia National Public Health Institute, Lusaka, Zambia. [150]Public Health Institute of Malawi, Lilongwe, Malawi. [151]National Health Laboratory, Gaborone, Botswana. [152]Laboratory of Transmissible Diseases and Biologically Active Substances (LR99ES27), Faculty of Pharmacy, University of Monastir, Monastir, Tunisia. [153]Laboratory of Microbiology, University Hospital of Monastir, Monastir, Tunisia. [154]Biomedical Informatics and Chemoinformatics Group, Informatics and Systems Department, National Research Centre, Cairo, Egypt. [155]Ministry of Health and Wellness, Gaborone, Botswana. [156]Eastern Technical University of Sierra Leone, Kenema, Sierra Leone. [157]Zoonotic Arbo and Respiratory Virus Program, Centre for Viral Zoonoses, Department of Medical Virology, University of Pretoria, Pretoria, South Africa. [158]Next Generation Sequencing Unit and Division of Virology, Faculty of Health Sciences, University of the Free State, Bloemfontein, South Africa. [159]National Reference Laboratory Lesotho, Maseru, Lesotho. [160]Centre for Biotechnology Research and Development, Kenya Medical Research Institute, Nairobi, Kenya. [161]Swiss Tropical and Public Health Institute, Basel, Switzerland. [162]Laboratorio de Investigaciones de Baney, Baney, Equatorial Guinea. [163]Ifakara Health Insitute, Ifakara, Tanzania. [164]Department of Medical Diagnostics, Kumasi Centre for Collaborative Research in Tropical Medicine, Kwame Nkrumah University of Science and Technology, Kumasi, Ghana. [165]PraesensBio, Lincoln, NE, USA. [166]Department of Medical Laboratory Science, Niger Delta University, Bayelsa State, Nigeria. [167]Systems and Biomedical Engineering Department, Faculty of Engineering, Cairo University, Cairo, Egypt. [168]King Faisal Specialist Hospital and Research Center, Riyadh, Kingdom of Saudi Arabia. [169]Biological Prevention Department, Ministry of Defence, Cairo, Egypt. [170]Faculty of Science, Fayoum University, Fayoum, Egypt. [171]Molecular Pathology Lab, Children's Cancer Hospital, Cairo, Egypt. [172]Laboratoire Biolim FSS/Université de Lomé, Lomé, Togo. [173]Department of Genetics and Genomics, College of Medicine and Health Sciences, United Arab Emirates University, Abu Dhabi, United Arab Emirates. [174]High Institute of Biotechnology of Monastir, University of Monastir, Rue Taher Haddad 5000, Monastir, Tunisia. [175]Rwanda National Joint Task Force COVID-19, Rwanda Biomedical Centre, Ministry of Health, Kigali, Rwanda. [176]School of Health Sciences, College of Medicine and Health Sciences, University of Rwanda, Kigali, Rwanda. [177]Department of Microbiology, Faculty of Medical Laboratory Sciences, Omdurman Islamic University, Sudan. [178]Instituto Nacional de Saúde (INS), Marracuene, Mozambique. [179]Department of Disease Control, School of Veterinary Medicine, University of Zambia, Lusaka, Zambia. [180]Internal Medicine Department, Alex Ekwueme Federal University Teaching Hospital, Abakaliki, Nigeria. [181]Institut Pasteur de Guinée, Conarky, Guinea. [182]Virology Laboratory, Alex Ekwueme Federal University Teaching Hospital, Abakaliki, Nigeria. [183]Department of Epidemiology and Community Health, Faculty of Clinical Sciences. College of Health Sciences. University of Ilorin, Ilorin, Kwara State, Nigeria. [184]Department of Public Health, Ministry of Health, Ilorin, Kwara State, Nigeria. [185]Alex Ekwueme Federal University Teaching Hospital, Abakaliki, Nigeria. [186]Mayotte Hospital Center, Mayotte, France. [187]The African Center of Excellence in Bioinformatics and Data-Intensive Sciences, The Infectious Diseases Institute, Kampala, Uganda. [188]Immunology and Molecular Biology, Makerere University, Kampala, Uganda. [189]Department of Medicine, Faculty of Clinical Sciences, College of Medicine, Ambrose Alli University, Ekpoma, Edo State, Nigeria. [190]Division of Virology, National Health Laboratory Service and University of the Free State, Bloemfontein, South Africa. [191]Infectious Hazards Preparedness, World Health Organization, Eastern Mediterranean Regional Office, Cairo, Egypt. [192]Department of Microbiology and Immunology, Faculty of Pharmacy, Cairo University, Cairo, Egypt. [193]Microbiology and Immunology Research Program, Children's Cancer Hospital Egypt, Cairo, Egypt. [194]National Public Health Laboratory, Ministry of Public Health of Cameroon, Yaoundé, Cameroon. [195]Faculty of Medicine and Biomedical Sciences, University of Yaoundé, Yaoundé, Cameroon. [196]Virology Service, Centre Pasteur of Cameroun, Yaounde, Cameroon. [197]Coordenadora da rede do Diagnóstico Tuberculose/HIV/COVID-19 na Instituição - Laboratório Nacional de Referência da Tuberculose em São Tomé e Príncipe, São Tomé, São Tomé and Principe. [198]Ponto focal para Melhoria da qualidade dos Laboratórios (SLIPTA) ao nível de São Tomé e Príncipe, São Tomé, São Tomé and Principe. [199]National Public Health Reference Laboratory (NPHRL), Mogadishu, Somalia. [200]Faculty of Medicine of Monastir, University of Monastir, Monastir, Tunisia. [201]University of Basel, Basel, Switzerland. [202]Clinical and Experimental Pharmacology Lab, LR16SP02, National Center of Pharmacovigilance, University of Tunis El Manar, Tunis, Tunisia. [203]Harvard T.H. Chan School of Public Health, Boston, MA, USA. [204]Centre MURAZ, Ouagadougou, Burkina Faso. [205]National Institute of Public Health of Burkina Faso (INSP/BF), Ouagadougou, Burkina Faso. [206]National Reference Center for Respiratory Viruses, Molecular Genetics of RNA Viruses, UMR 3569 CNRS, Université Paris Cité, Institut Pasteur, Paris, France. [207]World Health Organization, Harare, Zimbabwe. [208]Vietnamese-German Center for Medical Research, Hanoi, Vietnam. [209]Howard Hughes Medical Institute, Fred Hutchinson Cancer Center, Seattle, WA, USA. [210]Sub-Saharan African Network For TB/HIV Research Excellence (SANTHE), Durban, South Africa. [211]World Health Organization, WHO Lesotho, Maseru, Lesotho. [212]Med24 Medical Centre, Ruwa, Zimbabwe. [213]Department of Virology, Noguchi Memorial Institute for Medical Research, University of Ghana, Legon, Ghana. [214]Division of Human Genetics, Department of Pathology, University of Cape Town, Cape Town, South Africa. [215]Yemaachi Biotech, Accra, Ghana. [216]Center for Human Genetics, College of Medicine and Health Sciences, University of Rwanda, Kigali, Rwanda. [217]Laboratory of Human Genetics, GIGA Research Institute, Liège, Belgium. [218]Department of Biochemistry and Biotechnology, Pwani University, Kilifi, Kenya. [219]Department of Global Health, University of Washington, Seattle, WA, USA.

†These authors contributed equally to this work.
*Corresponding author. Email: tulio@sun.ac.za (T.d.O.); ewilkinson@sun.ac.za (E.W.)

SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.abq5358
Figs. S1 to S16
Tables S1 to S4
Reference (*77*)
MDAR Reproducibility Checklist

Submitted 14 April 2022; accepted 12 September 2022
Published online 15 September 2022
10.1126/science.abq5358

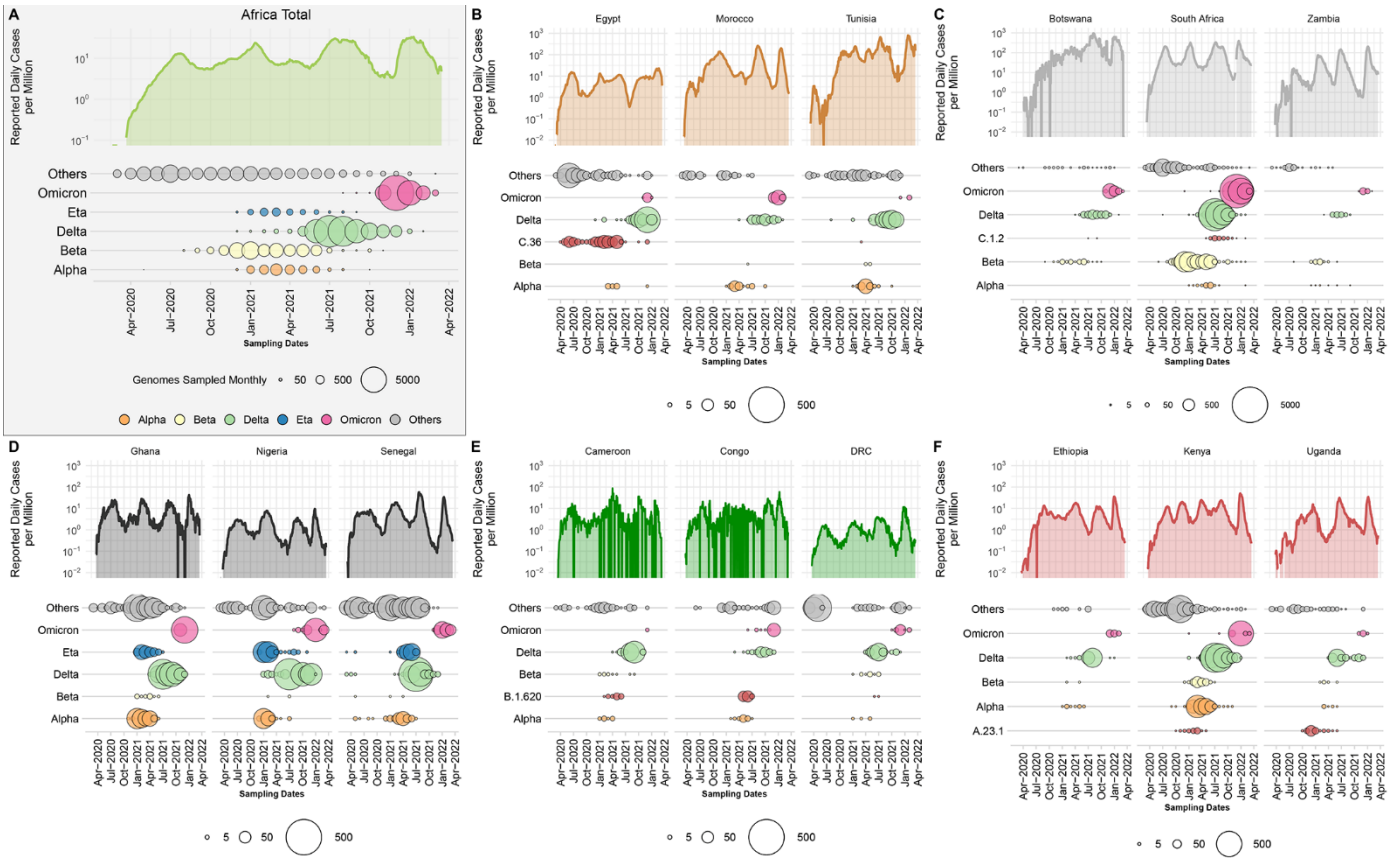


Fig. 1. Epidemiological progression of the COVID-19 pandemic on the African continent. (**A**) Total reported new case counts per million inhabitants in Africa (Data Source: Our World in Data; OWID; log-transformed) along with the distribution of VOCs, the Eta VOI and other lineages through time (size of circles proportional to the number of genomes sampled per month for each category). (**B** to **F**) Breakdown of reported new cases per million (Data Source: Our World in Data; OWID; log-transformed) and monthly sampling of VOCs, regional variant or lineage of interest and other lineages for three selected countries for North, Southern, West, Central and East Africa respectively. For each region, a different variant or lineage of interest is shown, relevant to that region (C.36, C.1.2, Eta, B.1.620 and A.23.1, respectively).

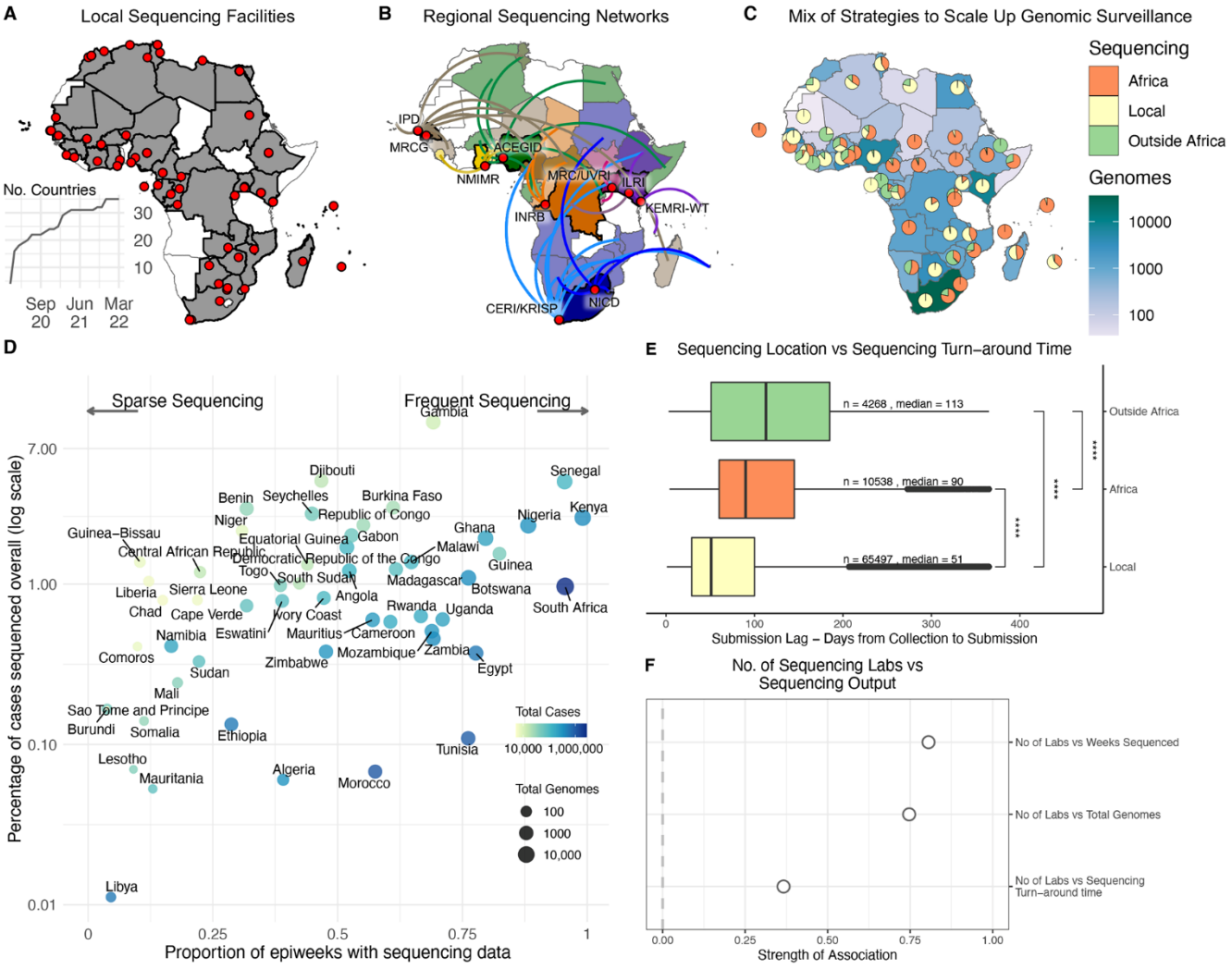


Fig. 2. Sequencing strategies and outputs in Africa. (**A**) Geographical representation of all countries (shaded in gray) and institutions (red dots) in Africa with their own on-site sequencing facilities. The inset graph shows the number of countries in Africa able to carry out sequencing locally over time. (**B**) Key regional sequencing hubs and networks in Africa showing countries (shaded in bright colors) and institutions (red dots) that have sequenced for other countries (shaded in corresponding light colors and linking curves) on the continent. CERI: Centre for Epidemic Response and Innovation; KRISP: KwaZulu-Natal Research Innovation and Sequencing Platform; NICD: National Institute for Communicable Diseases; KEMRI-WT: Kenya Medical Research Institute - Wellcome Trust; ILRI: International Livestock Research Institute; MRC/UVRI: Medical Research Council/Uganda Virus Research Institute; INRB: Institut National de Recherche Biomédicale; ACEGID: African Centre of Excellence for Genomics of Infectious Diseases; NMIMR: Noguchi Memorial Institute for Medical Research; MRCG: Medical Research Council Unit - The Gambia; IPD: Institut Pasteur de Dakar. (**C**) Geographical representation of the total number of SARS-CoV-2 whole genomes produced over the course of the pandemic in each country, as well as the proportion of those sequences that were produced locally, regionally or abroad. (**D**) Correlation of the proportion of COVID-19 positive cases that have been sequenced and the corresponding number of epidemiological weeks since the start of the pandemic that are represented with genomes for each African country. The color of each circle represents the number of cases and its size the number of genomes. (**E**) Comparison of sequencing turn-around times (lag times from sample collection to sequence submission) for the three strategies of sequencing in Africa, showing a significant difference in the means (p-value<0.0001). The box and whisker plot denote the lower quartile, median and upper quartile (box), the minimum and maximum values (whisker), and outliers (black dots). (**F**) Pearson correlations of the total number of sequencing laboratories per country against key sequencing outputs.

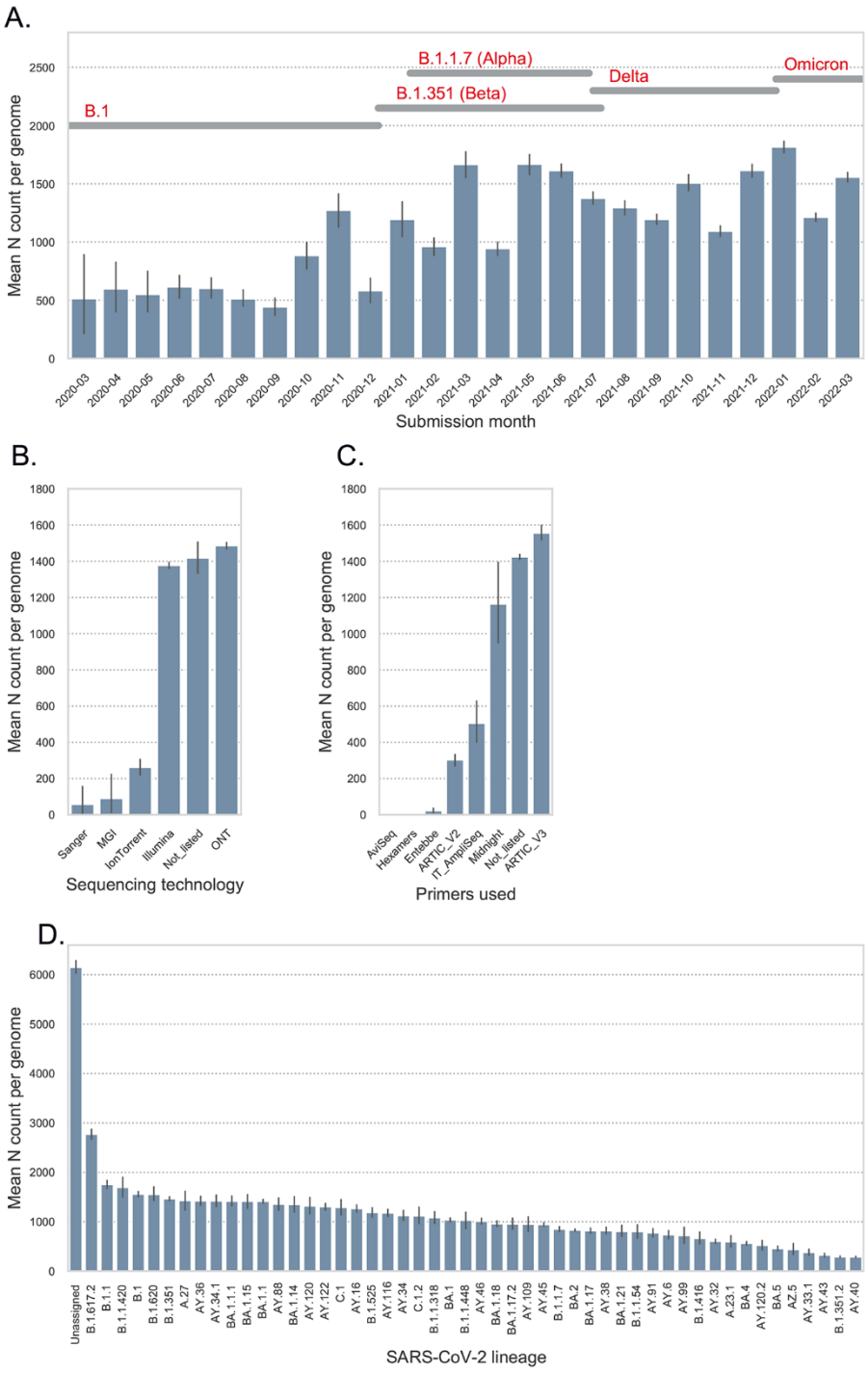


Fig. 3. Genome gap analysis. (**A**) Shows the mean N count per genome by month of submission to GISAID. The dates for the detection of important SARS-CoV-2 lineages are indicated at the top of the figure. (**B**) Illustrates the mean N count per genome stratified by sequencing technology. (**C**) Shows the mean N count per genome stratified by the sequencing primers sets used. For panels A to C, error bars indicate 95% confidence intervals. (**D**) Mean N count per genome by lineage. The mean N data were stratified by SARS-CoV-2 lineages to investigate lineage-specific frequency of genome gaps, an indirect measure of primer mismatch. All lineages present at least 100 times in the genome data were presented.

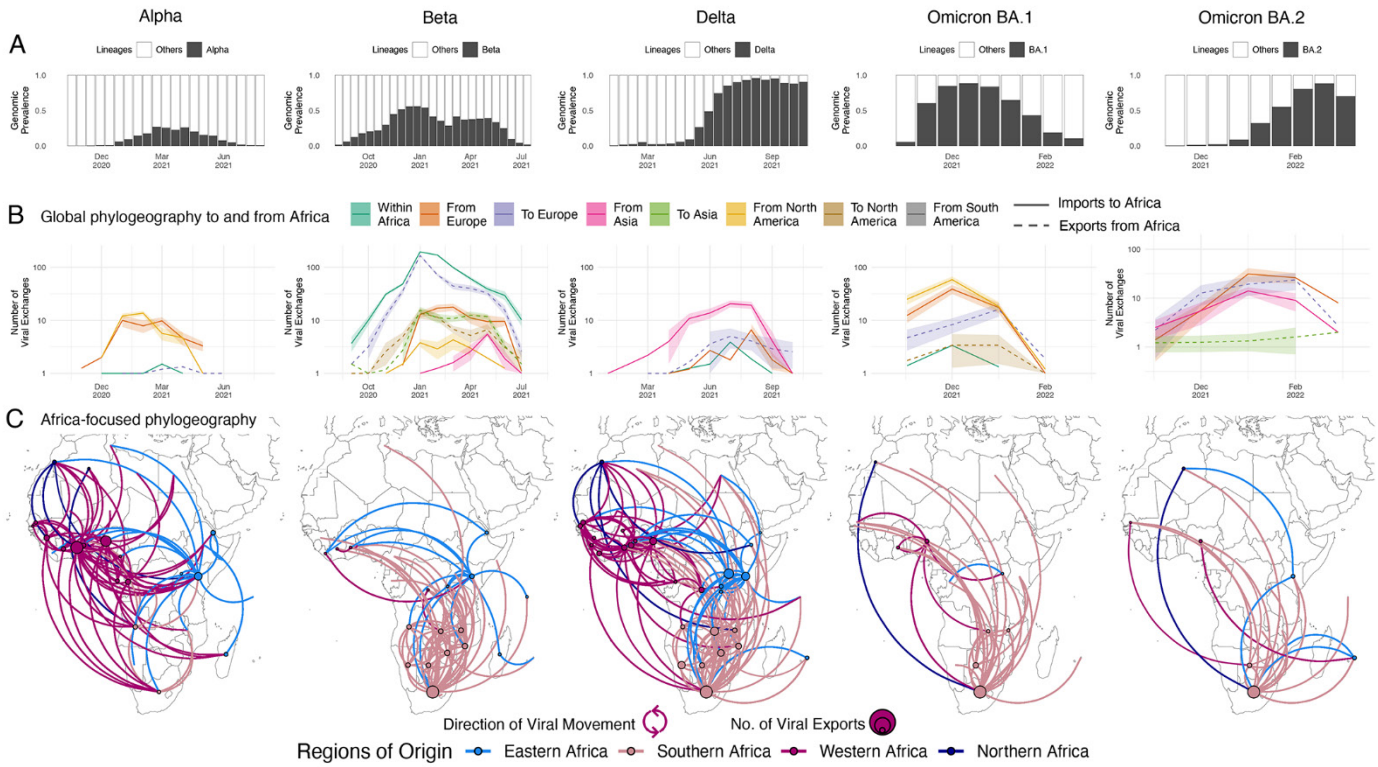


Fig. 4. Inferred viral dissemination patterns of VOCs within Africa. (**A**) Genomic prevalence of VOCs Alpha, Beta, Delta and Omicron in Africa over time. (**B**) Inferred viral exchange patterns to, from and within the Africa continent for the four VOCs (Omicron as BA.1 and BA.2) based on case-sensitive phylogeographic inference. Introductions and viral transitions within Africa are shown in solid lines and exports from Africa are shown in dotted lines and these are colored by continent. The shaded areas around the lines represent uncertainty of this analysis from ten replicates (+/− s.d.). (**C**) Dissemination patterns of the VOCs within Africa, from inferred ancestral state reconstructions performed on Africa enriched datasets, annotated and colored by region in Africa. The countries of origin of viral exchange routes are also shown with dots and the curves go from country of origin to destination country in an anti-clockwise direction.

Science

The evolving SARS-CoV-2 epidemic in Africa: Insights from rapidly expanding genomic surveillance

Houriiyah TegallyJames E. SanMatthew CottenMonika MoirBryan TegomohGerald MboowaDarren P. MartinCheryl BaxterArnold W. LambisiaAmadou DialloDaniel G. AmoakoMoussa M. DiagneAbay SisayAbdel-Rahman N. ZekriAbdou Salam GueyeAbdoul K. SangareAbdoul-Salam OuedraogoAbdourahmane SowAbdualmoniem O. MusaAbdul K. SesayAbe G. AbiasAdam I. ElzagheidAdamou LagareAdedotun-Sulaiman KemiAden Elmi AbarAdeniji A. JohnsonAdeola FowotadeAdeyemi O. OluwapelumiAdrienne A. AmuriAgnes JuruAhmed KandeilAhmed MostafaAhmed RebaiAhmed SayedAkano KazeemAladje BaldeAlan ChristoffelsAlexander J. TrotterAllan CampbellAlpha K. KeitaAmadou KoneAmal BouzidAmal SouissiAmbrose AgweyuAmel NaguibAna V. GutierrezAnatole NkeshimanaAndrew J. PageAnges YadouletonAnika VinzeAnise N. HappiAnissa ChouikhaArash IranzadehArisha MaharajArmel L. Batchi-BouyouArshad IsmailAugustina A. SylverkenAugustine GobaAyoade FemiAyotunde E. SijuwolaBaba MarycelinBabatunde L. SalakoBamidele S. OderindeBankole BolajokoBassirou DiarraBelinda L. HerringBenjamin TsofaBernard Lekana-DoukiBernard MvulaBerthe-Marie Njanpop-LafourcadeBlessing T. MaronderaBouh Abdi KhairehBourema KouribaBright AduBrigitte PoolBronwyn McInnisCara BrookCarolyn WilliamsonCassien NduwimanaCatherine AnscombeCatherine B. PrattCathrine ScheepersChantal G. Akoua-KoffiCharles N. AgotiChastel M. MapanguyCheikh LoucoubarChika K. OnwuamahChikwe IhekweazuChristian N. MalakaChristophe PeyrefitteChukwa GraceChukwuma E. OmoruyiClotaire D. RafaïCollins M. Morang'aCyril EramehDaniel B. LuleDaniel J. BridgesDaniel Mukadi-BamulekaDanny ParkDavid A. RasmussenDavid BakerDavid J. NokesDeogratius SsemwangaDerek TshiabuilaDominic S. Y. AmuzuDominique GoedhalsDonald S. GrantDonwilliams O. OmuoyoDorcas MaruapulaDorcas W. WanjohiEbenezer Foster-NyarkoEddy K. LusamakiEdgar SimulunduEdidah M. Ong'eraEdith N. NgabanaEdward O. AbworoEdward OtienoEdwin ShumbaEdwine BarasaEl Bara AhmedElhadi A. AhmedEmmanuel LokiloEnatha MukantwariEromon PhilomenaEssia BelarbiEtienne Simon-LoriereEtilé A. AnohEusebio ManuelFabian LeendertzFahn M. TawehFares WasfiFatma AbdelmoulaFaustinos T. TakawiraFawzi DerrarFehintola V. AjogbasileFlorette TreurnichtFolarin OnikepeFrancine NtoumiFrancisca M. MuyembeFrank E. Z. RagomzingbaFred A. DratibiFred-Akintunwa IyanuGabriel K. MbunsuGaetan ThilliezGemma L. KayGeorge O. AkpedeGert U. van ZylGordon A. AwandareGrace S. KpeliGrit SchubertGugu P. MaphalalaHafaliana C. RanaivosonHannah E. OmunakweHarris OnyweraHaruka AbeHela KarrayHellen NansumbaHenda TrikiHerve Albéric Adje KadjoHesham ElgahzalyHlanai GumboHota MathieuHugo Kavunga-MemboIbtihel SmetiIdowu B. OlawoyeIfedayo M. O. AdetifaIkponmwosa OdiaIlhem Boutiba Ben BoubakerIluoreh Ahmed MohammadIsaac SsewanyanaIsatta WurieIyaloo S. KonstantinusJacqueline Wemboo Afiwa HalatokoJames AyeiJanaki SonooJean-Claude C. MakangaraJean-Jacques M. TamfumJean-Michel HeraudJeffrey G. ShafferJennifer GiandhariJennifer MusyokiJerome NkurunzizaJessica N. UwanibeJinal N. BhimanJiro YasudaJoana MoraisJocelyn KiconcoJohn D. SandiJohn HuddlestonJohn K. OdoomJohn M. MorobeJohn O. GyapongJohn T. KayiwaJohnson C. OkolieJoicymara S. XavierJones GyamfiJoseph F. WamalaJoseph H. K. BonneyJoseph NyandwiJosie EverattJoweria NakaseeguJoyce M. NgoiJoyce NamulondoJudith U. OguzieJulia C. AndekoJulius J. LutwamaJuma J. H. MoggaJustin O'GradyKatherine J. SiddleKathleen VictoirKayode T. AdeyemiKefentse A. TumediKevin S. CarvalhoKhadija Said MohammedKoussay DellagiKunda G. MusondaKwabena O. DueduLamia Fki-BerrajahLavanya SinghLenora M. KeplerLeon BiscornetLeonardo de Oliveira MartinsLucious ChabukaLuicer OlubayoLul Deng OjokLul Lojok DengLynette I. Ochola-OyierLynn TyersMadisa MineMagalutcheemee RamuthMaha MastouriMahmoud ElHefnawiMaimouna MbanneMaitshwarelo I. MatshekaMalebogo KebabonyeMamadou DiopMambu MomohMaria da Luz Lima MendonçaMarietjie VenterMarietou F. PayeMartin FayeMartin M. NyagaMathabo MarekaMatoke-Muhia DamarisMaureen W. MburuMaximillian G. MpinaMichael OwusuMichael R. WileyMirabeau Y. TatfengMitoha Ondo'o AyekabaMohamed AbouelhodaMohamed Amine BeloufaMohamed G. SeadawyMohamed K. KhalifaMooko Marethabile MatoboMouhamed KaneMounerou SalouMphaphi B. MbulawaMulenga MwendaMushal AllamMy V. T. PhanNabil AbidNadine RujeniNadir AbuzaidNalia IsmaelNancy ElguindyNdeye Marieme TopNdongo DiaNédio MabundaNei-yuan HsiaoNelson Boricó SilochiNgiambudulu M. FranciscoNgonda SaasaNicholas BbosaNickson MurungaNicksy GumedeNicole WolterNikita SitharamNnaemeka NdodoNnennaya A. AjayiNoël TordoNokuzola MbheleNorosoa H. RazanajatovoNosamiefan IguosadoloNwando MbaOjide C. KingsleyOkogbenin SylvanusOladiji FemiOlubusuyi M. AdewumiOlumade TestimonyOlusola A. OgunsanyaOluwatosin FakayodeOnwe E. OgahOpe-Ewe OludayoOusmane FayePamela Smith-LawrencePascale OndoaPatrice CombePatricia NabisubiPatrick SemandaPaul E. OluniyiPaulo ArnaldoPeter Kojo QuashiePeter O. OkokherePhilip BejonPhilippe DussartPhillip A. BesterPlacide K. MbalaPontiano KaleebuPriscilla AbechiRabeh El-SheshenyRageema JosephRamy Karam AzizRené G. EssombaReuben Ayivor-DjanieRichard NjouomRichard O. PhillipsRichmond GormanRobert A. KingsleyRosa Maria D. E. S. A. Neto RodriguesRosemary A. AuduRosina A. A. CarrSaba GargouriSaber MasmoudiSacha BootsmaSafietou SankheSahra Isse MohamedSaibu FemiSalma MhallaSalome HoschSamar Kamal KassimSamar MethaSameh TrabelsiSara Hassan AgwaSarah Wambui MwangiSeydou DoumbiaSheila Makiala-MandandaSherihane AryeeteyShymaa S. AhmedSide Mohamed AhmedSiham ElhamoumiSikhulile MoyoSilvia LutucutaSimani GaseitsiweSimbirie JallohSoa Fy AndriamandimbySobajo OguntopeSolène GrayoSonia Lekana-DoukiSophie ProsolekSoumeya OuangraouaStephanie van WykStephen F. SchaffnerStephen KanyereziSteve Ahuka-MundekeSteven RudderSureshnee PillaySusan NabaddaSylvie BehillilSylvie L. BudiakiSylvie van der WerfTapfumanei MasheThabo MohaleThanh Le-VietThirumalaisamy P. VelavanTobias SchindlerTongai G. MapongaTrevor BedfordUgochukwu J. AnyanejiUgwu ChineduUpasana RamphalUwem E. GeorgeVincent EnoufVishvanath

Science (ISSN) is published by the American Association for the Advancement of Science. 1200 New York Avenue NW, Washington, DC 20005. The title *Science* is a registered trademark of AAAS.

Science, **Ahead of Print** • DOI: 10.1126/science.abq5358

B INFORMAÇÕES ADICIONAIS DESENVOLVIMENTO THEMOMUTDB

Neste anexo, são apresentadas informações que, devido a limitação do número de páginas, não foram incluídas no artigo, sendo elas, a modelagem do banco de dados e o projeto de interface da aplicação. O modelo do banco de dados foi desenvolvido utilizando o software MySQL Workbench[3] e pode ser visualizado na Figura B1. O modelo compreende tanto os dados extraídos da literatura, quanto os dados que foram derivados de outras bases de dados.

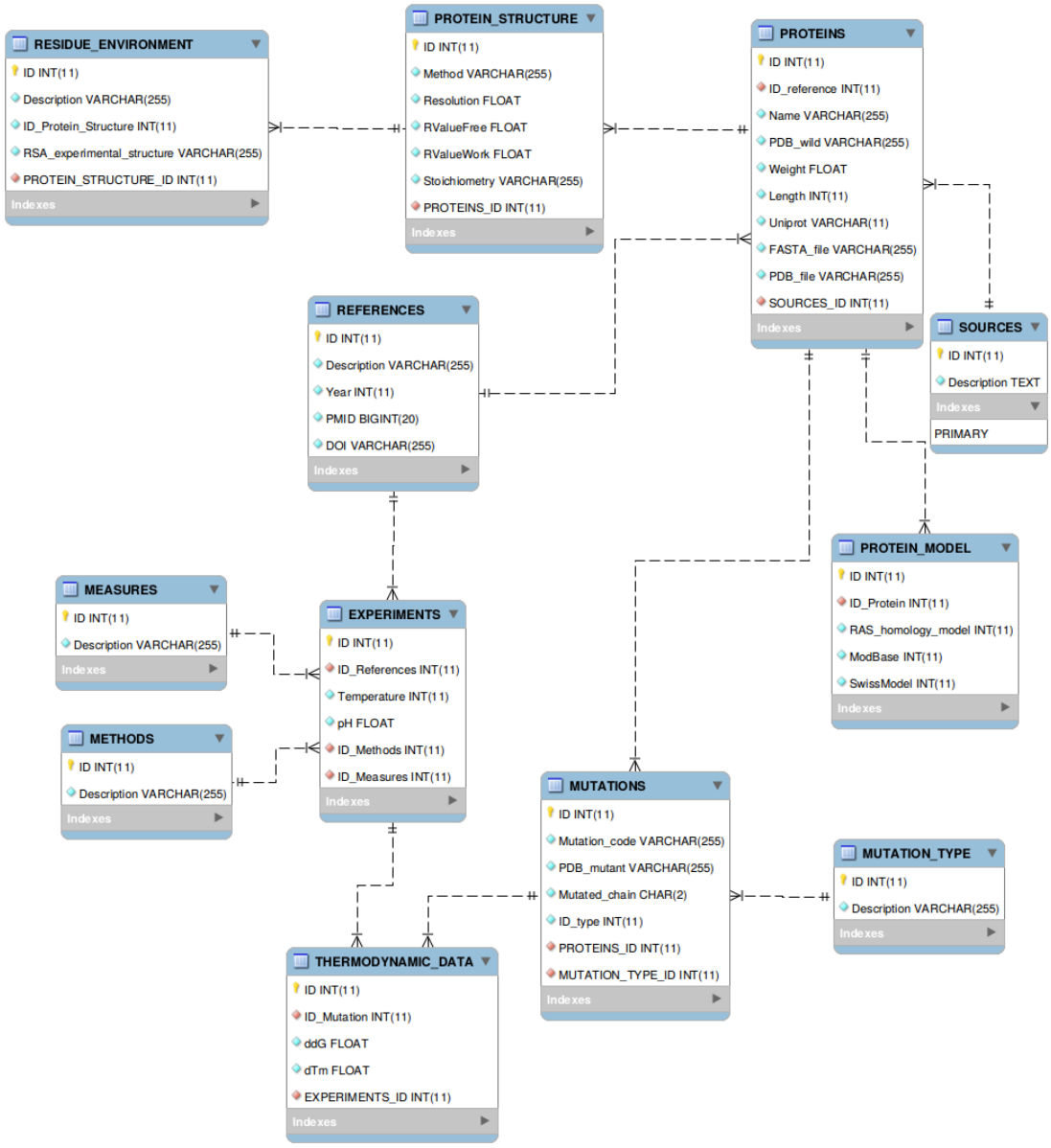


Figura B1 : Modelo do Banco de Dados do ThermoMutDB

[3] https://www.mysql.com/products/workbench/

A identidade visual foi pensada para uma paleta de cores (Figura B2) e formatos que refletissem o movimento que a termodinâmica representa, o que originou a logo e a interface gráfica apresentada no artigo.

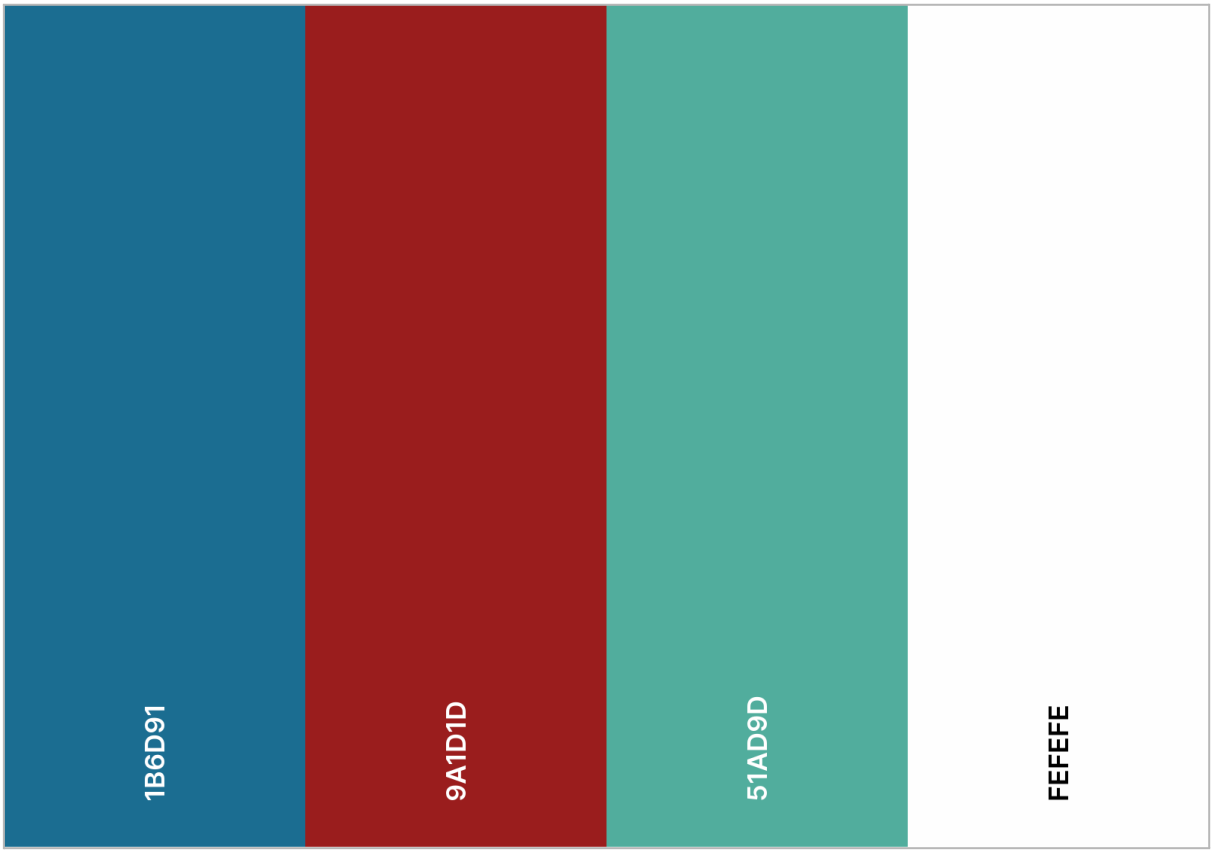


Figura B2 : Paleta de cores utilizada na interface do ThermoMutDB

ANEXO A - RELATÓRIOS DE VISITAS THERMOMUTDB

Neste anexo são apresentados os relatórios de utilização do ThermoMutDB exportados do Google Analytics (https://analytics.google.com/) de 1 de Outubro de 2020 (poucos dias antes da publicação do artigo) até 9 de Outubro de 2022 (data da escrita). No primeiro relatório é possível observar o crescimento do número de usuários através do tempo, total de usuários no período, porcentagem de usuários por país, entre outras informações. No segundo relatório, é mostrado visualmente o alcance do ThermoMutDB no mundo e as taxas de aquisição de usuários, comportamento e conversão em detalhes.

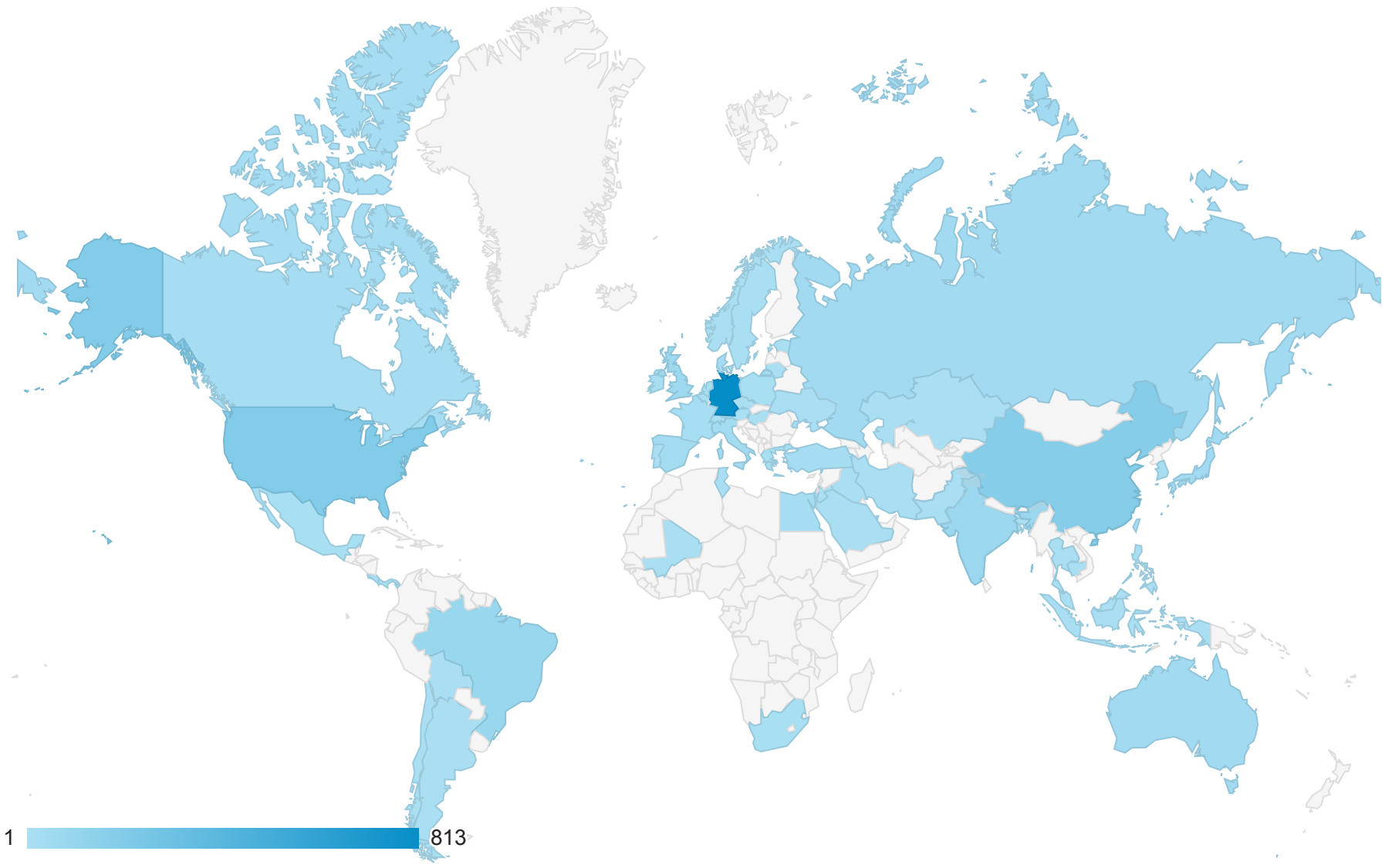
Localização

Todos os usuários
100,00% Usuários

1 de out. de 2020 - 9 de out. de 2022

Cobertura regional

Dados resumidos



País	Aquisição			Comportamento			Conversões		
	Usuários ↓	Novos usuários	Sessões	Taxa de rejeição	Páginas / sessão	Duração média da sessão	Taxa de conversão de meta	Conclusões de meta	Valor da meta
	1.864 Porcentagem do total: 100,00% (1.864)	**1.869** Porcentagem do total: 100,27% (1.864)	**3.091** Porcentagem do total: 100,00% (3.091)	**59,95%** Média de visualizações: 59,95% (0,00%)	**2,49** Média de visualizações: 2,49 (0,00%)	**00:03:20** Média de visualizações: 00:03:20 (0,00%)	**0,00%** Média de visualizações: 0,00% (0,00%)	**0** Porcentagem do total: 0,00% (0)	**US$ 0,00** Porcentagem do total: 0,00% (US$ 0,00)
1. Germany	**813** (43,13%)	810 (43,34%)	817 (26,43%)	97,43%	1,08	00:00:07	0,00%	0 (0,00%)	US$ 0,00 (0,00%)
2. United States	**193** (10,24%)	191 (10,22%)	366 (11,84%)	51,64%	2,34	00:02:41	0,00%	0 (0,00%)	US$ 0,00 (0,00%)
3. China	**178** (9,44%)	172 (9,20%)	313 (10,13%)	54,63%	2,46	00:03:00	0,00%	0 (0,00%)	US$ 0,00 (0,00%)
4. Brazil	**84** (4,46%)	81 (4,33%)	370 (11,97%)	38,11%	5,19	00:11:03	0,00%	0 (0,00%)	US$ 0,00 (0,00%)
5. India	**75** (3,98%)	77 (4,12%)	106 (3,43%)	46,23%	2,58	00:04:38	0,00%	0 (0,00%)	US$ 0,00 (0,00%)
6. United Kingdom	**60** (3,18%)	60 (3,21%)	127 (4,11%)	52,76%	2,39	00:02:23	0,00%	0 (0,00%)	US$ 0,00 (0,00%)
7. Japan	**46** (2,44%)	45 (2,41%)	88 (2,85%)	42,05%	2,82	00:03:10	0,00%	0 (0,00%)	US$ 0,00 (0,00%)
8. Australia	**41** (2,18%)	37 (1,98%)	100 (3,24%)	47,00%	2,67	00:02:33	0,00%	0 (0,00%)	US$ 0,00 (0,00%)
9. Russia	**40** (2,12%)	41 (2,19%)	141 (4,56%)	37,59%	3,03	00:04:06	0,00%	0 (0,00%)	US$ 0,00 (0,00%)
10. Italy	**37** (1,96%)	37 (1,98%)	94 (3,04%)	44,68%	3,33	00:05:16	0,00%	0 (0,00%)	US$ 0,00 (0,00%)

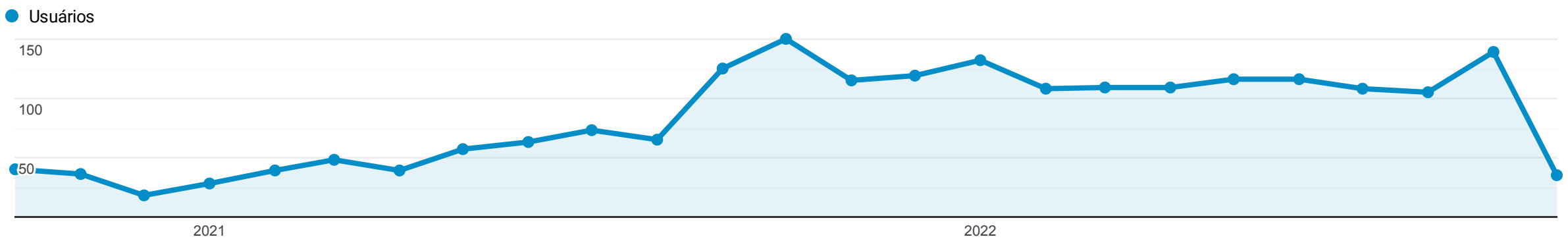
Linhas 1 - 10 de 61

Visão geral do público-alvo

Todos os usuários
100,00% Usuários

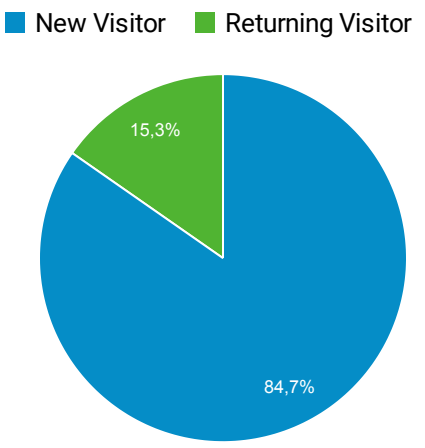
1 de out. de 2020 - 9 de out. de 2022

Visão geral



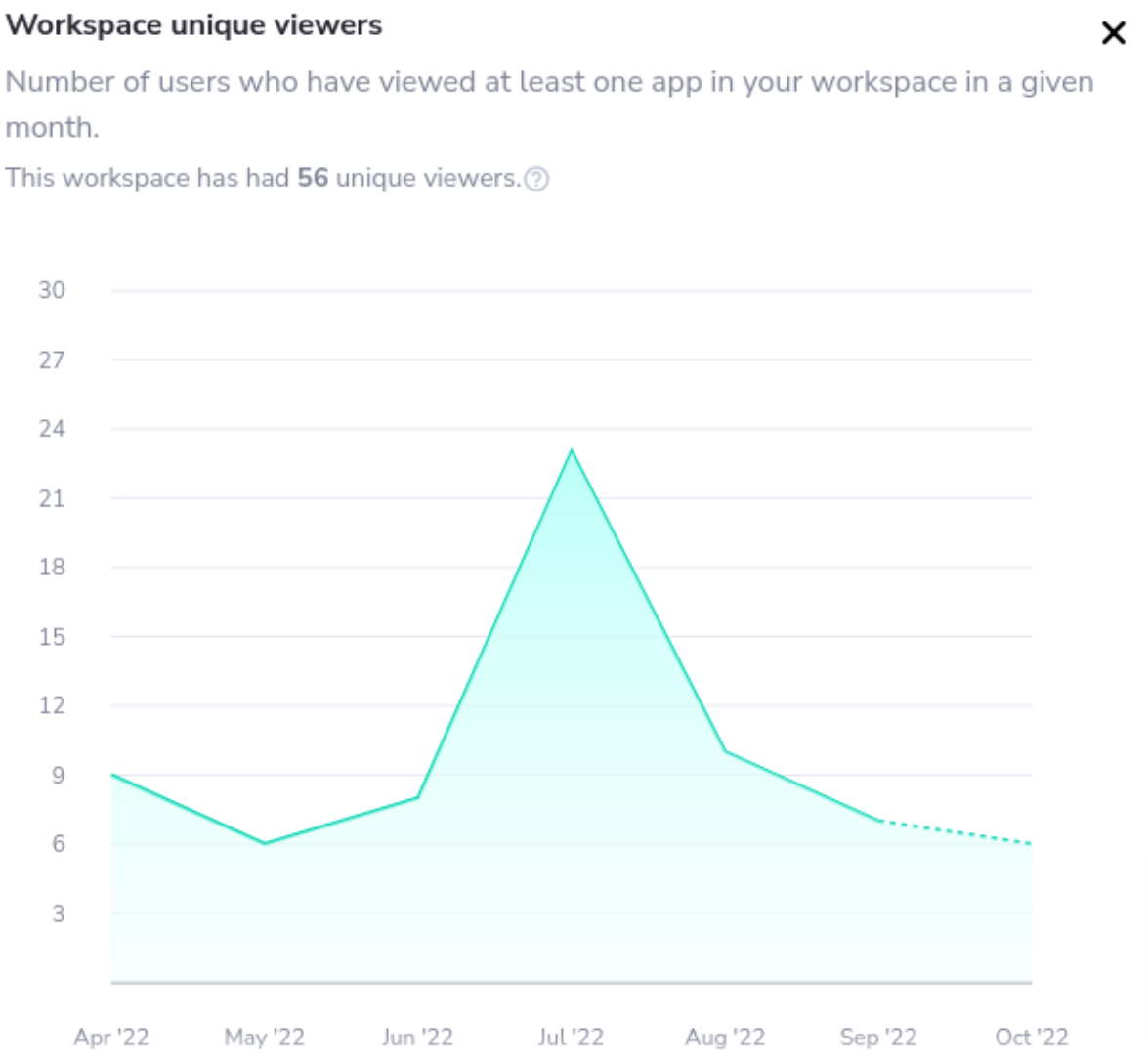
Usuários	Novos usuários	Sessões	Número de sessões por usuário
1.864	1.864	3.091	1,66

Visualizações de página	Páginas / sessão	Duração média da sessão	Taxa de rejeição
7.690	2,49	00:03:20	59,95%



	País	Usuários	Porcentagem do Usuários
1.	Germany	813	43,13%
2.	United States	193	10,24%
3.	China	178	9,44%
4.	Brazil	84	4,46%
5.	India	75	3,98%
6.	United Kingdom	60	3,18%
7.	Japan	46	2,44%
8.	Australia	41	2,18%
9.	Russia	40	2,12%
10.	Italy	37	1,96%

ANEXO B - RELATÓRIOS DE VISITAS SARS-CoV-2 AFRICA DASHBOARD



ANEXO C - DIVULGAÇÃO DA APRESENTAÇÃO DOS VENCEDORES PRÊMIO TRIMESTRAL BAKER DE PUBLICAÇÃO DE EXCELÊNCIA

Quarterly Research Prizes for Publication Excellence

Arpeeta Sharma: Diabetes & Metabolism

Specific NLRP3 Inhibition Protects Against Diabetes-Associated Atherosclerosis.

Yow Keat Tham: Cardiac Hypertrophy

Novel Lipid Species for Detecting and Predicting Atrial Fibrillation in Patients with Type 2 Diabetes

Simon Bond: Molecular Metabolism & Ageing

Deletion of Trim28 in committed adipocytes promotes obesity but preserves glucose tolerance

Joicymara Xavier: Computational Biology and Clinical Informatics

ThermoMutDB: a thermodynamic database for missense mutations

Chair: Prof Murray Esler

Students and ECS are reminded that their attendance is required.

Please note: Baker staff and students only — this is an internal seminar and may contain confidential data

ANEXO D - CERTIFICADO DE SEGUNDO MELHOR POSTER MÓDULO NGS NO VEME WORKSHOP

26th Bioinformatic Workshop on
Virus Evolution and Molecular Epidemiology

CERTIFICATE

The second best poster of the NGS module is entitled **"An interactive tool for visualizing COVID-19 genomics data"**, presented by **Joicymara Santos Xavier** from **Brazil**.

Panama, August 26th, 2022

Luiz Carlos Junior Alcantara
Oswaldo Cruz Foundation

Alexander A. Martínez C. MT.
Instituto Conmemorativo Gorgas de Estudios en Salud.

Tulio de Oliveira
Stellenbosch University, South Africa

PAHO | IGORGAS | CERI | Stellenbosch | krisp

ANEXO E - PRODUÇÃO ACADÊMICA

Nesta seção são apresentados a produção acadêmica durante o período do doutorado da autora (2018-2022) assim como os prêmios obtidos.

Artigos publicados como primeira autora:

1. Xavier, J. S., Rezende, P. M., Velloso, J. P. L., Nguyen, T.-B., Karmarkar, M., Portelli, S., Ascher, D. B., et al. (2021). ThermoMutDB: a thermodynamic database for missense mutations. *Nucleic Acids Research*, *49*(D1), D475–D479.

2. Xavier, J. S., Moir, M., Tegally, H., Sitharam, N., Karim, A., San, J., et al. SARS-CoV-2 Africa Dashboard for real-time COVID-19 information. *Nature Microbiology*. 2022. *(Aceito para publicação)*

Demais artigos:

1. Pires, D. E. V., Veloso, W. N. P., Myung, Y., Rodrigues, C. H. M., Silk, M., Rezende, P. M., Silva, F., Xavier, J.S., et al. (2020). EasyVS: a user friendly web based tool for molecule library selection and structure-based virtual screening. *Bioinformatics*.

2. da Silva, A. L.*, Abreu, A. P. de, Mariano, D.*, Santos, F. B*., Lage, F. S. D.*, Quintanilha-Peixoto, G.*, Hilario, H.O.*, Xavier, J.S.*, et al. (2021). From In-Person to the Online World: Insights Into Organizing Events in Bioinformatics. *Frontiers in Bioinformatics*, *1*.
   ** Esses autores contribuíram igualmente e compartilham a primeira autoria.*

3. Rezende, P. M., Xavier, J. S., Ascher, D. B., Fernandes, G. R., & Pires, D. E. V. (2022). Evaluating hierarchical machine learning approaches to classify biological databases. *Briefings in Bioinformatics*, 23(4).

4. Tegally, H., San, J. E., Cotten, M., Moir, M., Tegomoh, B., Mboowa, G., Martin, D. P., et al. (2022). The evolving SARS-CoV-2 epidemic in Africa: Insights from rapidly expanding genomic surveillance. *Science*, eabq5358.

5. Poongavanan, J., Xavier, J., Dunaiski, M., Tegally, H., Oladejo, S., Ayorinde, O., Wilkinson, E., et al. (2022). Managing and assembling population-scale data streams,

tools and workflows to plan for future pandemics within the INFORM Africa Consortium Authors. *Unpublished*.

Livros/capítulos de livros:

1. Pires, D. E. V., Rodrigues, C. H. M., Albanaz, A. T. S., Karmakar, M., Myung, Y., Xavier, J., Michanetzi, E.-M., et al. (2019). Exploring protein supersecondary structure through changes in protein folding, stability, and flexibility. *Methods in Molecular Biology*, *1958*, 173–185.

2. Pires, D. E. V., Portelli, S., Rezende, P. M., Veloso, W. N. P., Xavier, J. S., Karmakar, M., Myung, Y., et al. (2020). A Comprehensive Computational Platform to Guide Drug Development Using Graph-Based Signature Methods. *Methods in Molecular Biology*, *2112*, 91–106.

3. Airey, E., Portelli, S., Xavier, J. S., Myung, Y. C., Silk, M., Karmakar, M., Velloso, J. P. L., et al. (2021). Identifying Genotype-Phenotype Correlations via Integrative Mutation Analysis. *Methods in Molecular Biology*, *2190*, 1–32.

4. Mariano, D., Dezordi, F. Z., Martins, P., Xavier, J., Sousa, T. de J., Lima, L., & Santos, L. H. (Eds.). (2021). *BIOINFO - Revista Brasileira de Bioinformática e Biologia Computacional*. Alfahelix.

Artigos apresentados em eventos:

1. Marinho Rezende, P., & Santos Xavier, J. (2019). Uso de aprendizado de máquina para criação de um Banco de Dados com informações taxonômicas de gene de rRNA 16s. *Anais do 14º Simpósio Brasileiro de Automação Inteligente*. Presented at the ANAIS DO 14º SIMPóSIO BRASILEIRO DE AUTOMAÇÃO INTELIGENTE, Galoa.

2. Figueiredo, L. A., Dias, A. B. A., Fagundes, L. A. G., Silva, V. D. P., Almeida, S. G. M., Magalhães, M. L. C., Rezende, P. M., … & Xavier, J.S. (2020). Código X em Casa: um relato de experiência sobre o ensino remoto de computação desplugada para meninas em situação de vulnerabilidade socioeconômica, em tempos de distanciamento social. *Anais do XXVI Workshop de Informática na Escola (WIE*

2020) (pp. 279–288). Presented at the Workshop de Informática na Escola, Sociedade Brasileira de Computação - SBC.

Prêmios obtidos:

1. **Quartely Research Prizes for Publication Excellence**. Baker Heart and Diabetes Institute. Artigo: *ThermoMutDB: a thermodynamic database for missense mutations* na categoria Computational Biology and Clinical Informatics. (Anexo C)
2. **2nd Best Poster**. 26th Bioinformatic Workshop on Virus Evolution and Molecular Epidemiology. Poster: An Interactive tool for visualizing COVID-19 genomics data. Categoria: NGS module. (Anexo D)