Diversidade genética do novo coronavírus SARS-CoV-2 (COVID-19) em Portugal

Mais informações em **https://insaflu.insa.pt/covid19/**

Relatório de situação — 21 de Agosto de 2020

O **Instituto Nacional de Saúde Doutor Ricardo Jorge, I.P. (INSA)** analisou até à data **1785 sequências do genoma do novo coronavírus SARS-CoV-2**, obtidas de amostras colhidas em 63 laboratórios/hospitais/instituições representando 159 concelhos (**Figura 1**).

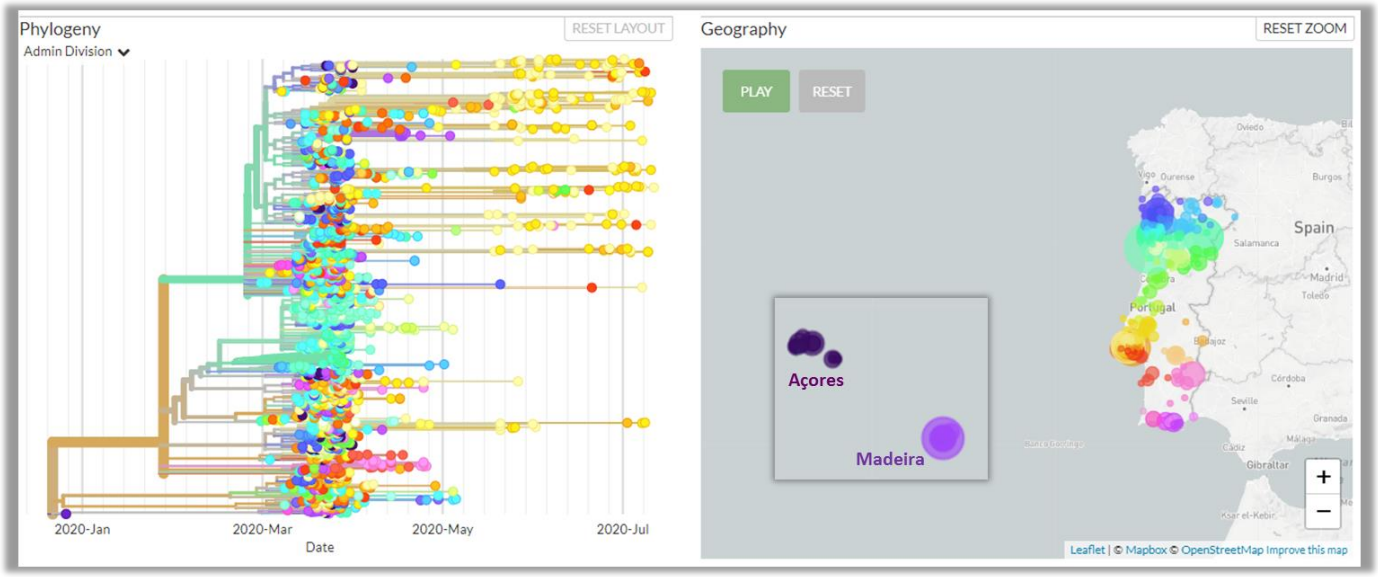

Figura 1. Visão global da diversidade genética e dispersão geotemporal do vírus SARS-CoV-2 em Portugal. Os diferentes genomas (representados por círculos no painel à esquerda) estão coloridos de acordo com o local de residência, com a mesma tonalidade no mapa – o tamanho dos círculos no mapa é proporcional ao número de genomas sequenciados por localidade (consultar o site https://insaflu.insa.pt/covid19/ para mais detalhes).

Desde o último relatório (24 de Junho), o site foi sendo actualizado, contemplando agora mais 809 genomas virais, perfazendo um total de 1785 sequências. Destas, 397 sequências foram geradas no Instituto de Gulbenkian de Ciência (IGC), instituição colaboradora do INSA neste estudo.

Distribuição por "Clade"

- **A distribuição por clade** (**Tabela 1**) **é, em termos gerais, semelhante àquela que é observada a nível Europeu** (https://nextstrain.org/ncov/europe)**. A maioria dos vírus** (91,2%) **integra o braço filogenético contendo os clades 20A, 20B e 20C, os quais apresentam, entre outros marcadores genéticos, a mutação "aa D614G"** ("nt A23403G") **na proteína "Spike (S)".** Esta proteína é responsável pela entrada do vírus SARS-CoV-2 nas células humanas, sendo também o principal antigénio deste vírus pandémico.

Tabela 1. Distribuição dos genomas por "Clade"

Clade	n.º de genomas	%
19A	114	6,4%
19B	43	2,4%
20A	654	36,6%
20B	928	52%
20C	46	2,6%

- A **distribuição dos diferentes clades por Distrito e por Concelho** pode ser explorada na plataforma Microreact (Nota: estes links dão também acesso a um filme que mostra a distribuição filogeográfica ao longo do tempo).

Núcleo de Bioinformática | Departamento de Doenças Infecciosas
INSTITUTO NACIONAL DE SAÚDE DOUTOR RICARDO JORGE
Avenida Padre Cruz, 1649-016 Lisboa, PORTUGAL

Mutações – balanço geral

O **número médio de mutações (SNPs) por genoma** (comparando com o primeiro genoma sequenciado na China; MN908947.3) **é de 8 (variando entre 1 e 20) mutações**, o que se enquadra dentro da taxa de mutação prevista para este vírus (i.e., cerca de 2 mutações por genoma por mês). No conjunto dos 1785 genomas analisados até à data, observam-se SNPs em 1274 dos 29903 nucleótidos que constituem o genoma do novo coronavírus SARS-CoV-2, sendo que **103 são SNPs não-sinónimas que alteram a proteína "Spike (S)".**

PRINCIPAL DESTAQUE

O principal destaque desta actualização vem no âmbito da divulgação (artigo submetido para publicação) dos **primeiros grandes resultados deste estudo de colaboração** entre os mais de 60 laboratórios/hospitais espalhados pelo país. De forma simplificada, deixamos aqui os resultados mais relevantes deste primeiro estudo:

- O ínicio da pandemia em Portugal, com foco no Norte e Centro do país, foi caracterizado pelo **espalhamento massivo de um variante do SARS-CoV-2 com uma mutação (D839Y) na proteína Spike** (Figura 2);

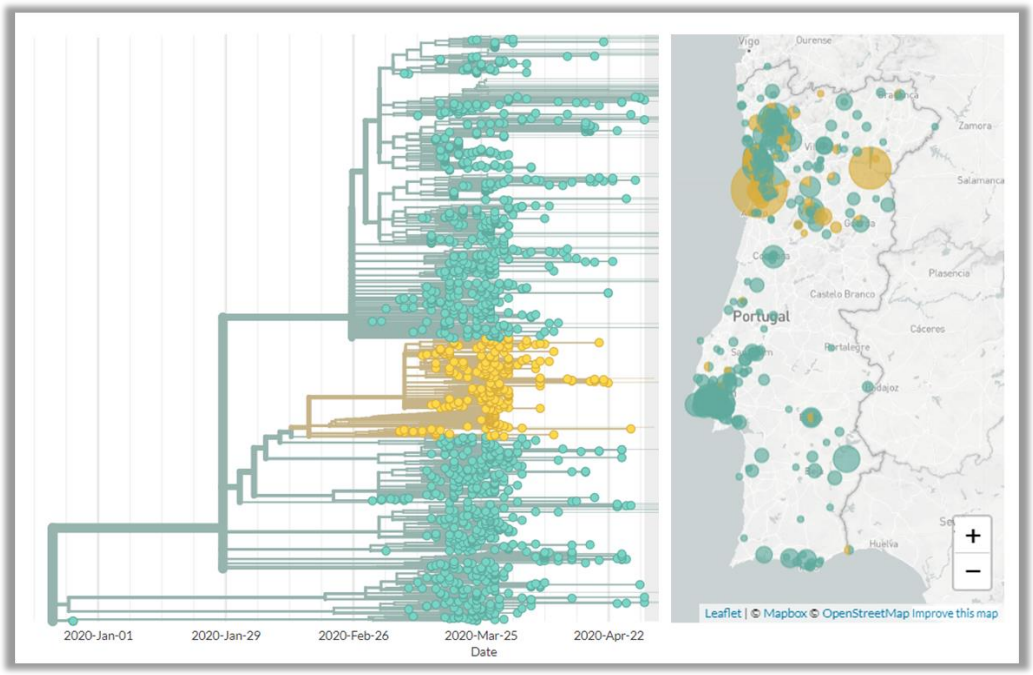

Figura 2. Enquadramento e dispersão geográfica do sub-clade com a mutação D839Y na proteína Spike no contexto do global dos genomas analisados em Portugal até ao final de Abril. Os genomas com e sem a mutação D839Y estão coloridos a amarelo e verde, respetivamente, sendo o tamanho dos círculos no mapa proporcional ao número de genomas sequenciados por localidade.

- Esta mutação tinha sido já sinalizada internacionalmente como de particular interesse dado estar **localizada precisamente num domínio (*fusion peptide*) da Spike crítico para a entrada do vírus nas células humanas**. Esta mutação foi já detetada em 13 países de quatro continentes;

- Este variante do **SARS-CoV-2 terá sido importado de Itália, região da Lombardia, pouco depois do meio de Fevereiro**, tendo circulado em Portugal pelo menos uma semana antes de terem sido reportados os primeiros casos de COVID-19 (2 de Março) no nosso país, originando potencialmente cadeias de transmissão não detectadas;

PRINCIPAL DESTAQUE (continuação)

- Durante a fase exponencial da epidemia em Portugal, observou-se uma incidência inicial em distritos do litoral das regiões Norte e Centro, tendo esta variante disseminado para distritos no interior das mesmas regiões (**Figura 3 e Figura 4**). Esta variante foi responsável pelo **grande surto em Ovar**, tendo levado à cerca sanitária entre 17 de Março e 17 de Abril;

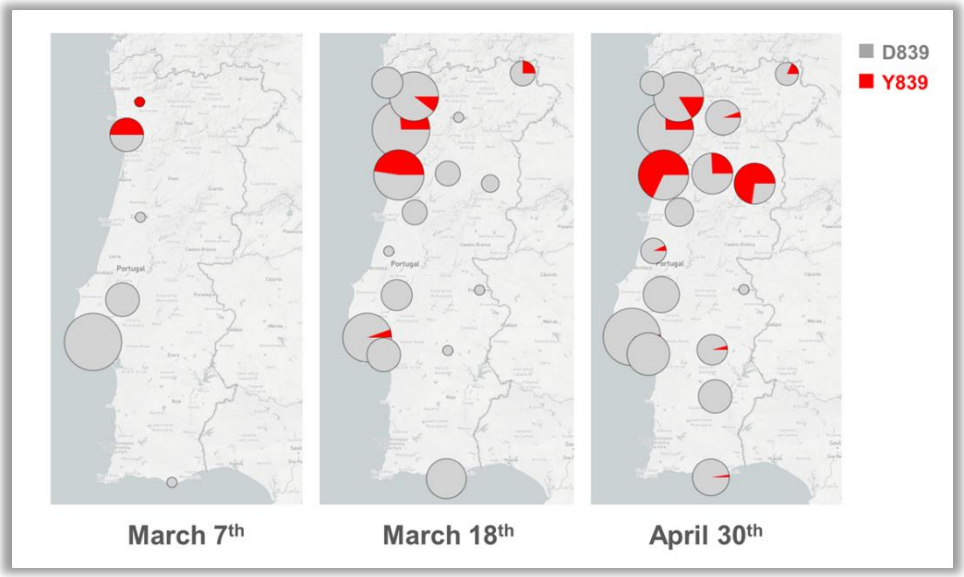

Figura 3. Dispersão geográfica, por distrito, das frequências relativas da mutação D839Y (a vermelho) na proteína Spike, em três pontos temporais: 7 de Março (deteção dos primeiros genomas com a mutação D839Y), **18 de Março** (quando foi declarado o estado de emergência) e **30 de Abril de 2020.** As Regiões Autónomas dos Açores e da Madeira não são exibidas pois o variante D839Y não foi detectado nas mesmas.

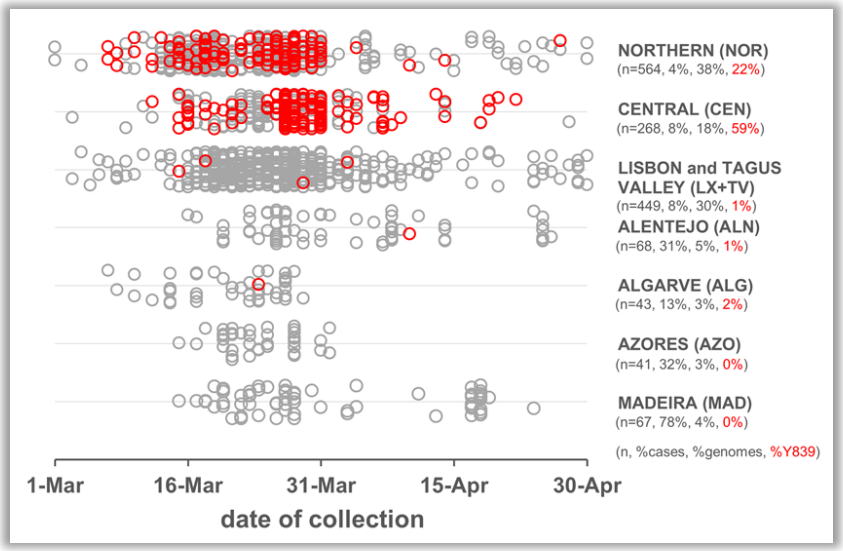

Figura 4. Dispersão temporal, por região, do SARS-CoV-2 com a mutação D839Y na proteína Spike. Distribuição das sequências dos genoma analisados (n=1500, até 30 de Abril) por data de colheita e Região, destacando os casos de COVID-19 causados pelo variante Spike D839Y (a vermelho). No eixo dos y, é indicado, para cada Região, o número de sequências analisadas (n), a percentagem de casos confirmados com dados de genoma SARS-CoV-2 (*% cases*), a percentagem de sequências de cada Região no conjunto total de genomas sequenciados (*% genomes*) e a percentagem de sequências do variante D839Y (% Y839, a vermelho), a 30 de Abril de 2020.

- A **frequência relativa** deste variante aumentou a uma **taxa estimada de 12,1%** (6.1%-18.2%, CI 95%) **a cada três dias, entre 14 de Março e 9 de Abril**. Estima-se que, durante esse período, este variante tenha sido responsável por cerca de 3800 casos de COVID-19 em Portugal, correspondendo a cerca de um em cada quatro casos de COVID-19 no país;

Perante estes resultados, levantamos **duas hipóteses** que poderão, separada ou concomitantemente, ter contribuído para a elevada frequência do variante D839Y durante a fase exponencial da epidemia em Portugal:

- Por um lado, **este variante poderá ter tido maior oportunidade de propagação (*founder effect*)** dado ter sido introduzido no país mais de uma semana antes dos primeiros casos detetados de COVID-19;
- Por outro lado, o aumento significativo da sua frequência relativa poderá dever-se a um **maior *fitness*, em particular em termos da sua capacidade de transmissão.**

Salientamos que será necessária a realização de estudos funcionais que testem o impacto da mutação D839Y na infectividade e propagação do SARS-CoV-2.

Nota: O artigo submetido pode ser consultado em: doi: https://doi.org/10.1101/2020.08.10.20171884

FUTURO

Encontramo-nos em fase de análise de resultados com vista a uma completa caracterização do **início da pandemia em Portugal**, nomeadamente na **identificação geotemporal das múltiplas introduções do SARS-CoV-2** no nosso país.

NOTAS ADICIONAIS (também disponível em https://insaflu.insa.pt/covid19/)

Objectivos gerais do estudo:

- Determinação dos **padrões de disseminação do vírus nas diferentes regiões de Portugal** e em diferentes grupos populacionais.
- Determinação dos perfis mutacionais do SARS-CoV-2 para **identificação e monitorização de cadeias de transmissão**, bem como **identificação de novas introduções do vírus** em Portugal.
- Prever o início da transmissão na comunidade e aferir o impacto das medidas de contenção, avaliando a **contribuição da transmissão local *versus* importações do vírus**.
- Determinação do **grau de variabilidade genética de antigénios ou alvos de fármacos antivirais** com possível impacto no desenvolvimento / eficiência de medidas profiláticas (vacinas) e terapêuticas, bom como **monitorização das mutações em alvos genéticos de testes de diagnóstico**.
- Determinação de possíveis **associações entre perfis genéticos (mutacionais)** do SARS-CoV-2 e **determinadas manifestações clínicas** (ex. diferentes graus severidade da COVID-19) ou diferente **capacidade de transmissão** do vírus.
- Estudar os **mecanismos evolutivos do vírus** e a sua relação com os perfis de disseminação em diferentes regiões de Portugal e em diferentes grupos populacionais.
- **Contribuir para a avaliação da relevância funcional e fenotípica** de mutações particulares.

Métodos

- **Procedimento Pre-NGS:** adaptado da Artic Network (https://artic.network/ncov-2019, https://www.protocols.io/view/ncov-2019-sequencing-protocol-bbmuik6w)
- **Procedimento NGS:** Nextera XT e MiSeq (Illumina)
- **Dos "reads" às sequências do genoma:** INSaFLU
- **Das sequências do genoma à "filogeografia*":** Nextstrain (mais detalhes sobre o método em em https://nextstrain.org/ncov e https://github.com/nextstrain/ncov) e Microreact.

** O posicionamento geográfico reflecte o local (Concelho – "Admin Division" ou Freguesia – "Location") de residência ou, caso não exista informação, local de ocorrência ou da entidade que enviou a amostra. Apenas são indicadas as freguesias com população residente superior a 5000 pessoas (Fonte: CENSOS 2011 - Instituto Nacional de Estatística). Para as restantes freguesias, por motivos de confidencialidade, é apenas indicado o Concelho.*

Agradecimentos

- A todos os laboratórios nacionais que enviam amostras clínicas (suspeitas ou positivas para SARS-CoV-2) para o Laboratório Nacional de Referência da Gripe e outros vírus respiratórios do INSA.
- À Sara Hill e Nuno Faria (Universidade de Oxford) e Joshua Quick e Nick Loman (Universidade de Birmingham) por nos terem gentilmente cedido os primers usados no ínicio deste estudo.
- Às equipas dos projectos Nextstrain e Microreact pela libertação de algumas ferramentas de bioinformática usadas neste estudo.
- Ao Miguel Pinheiro (iBiMED / Universidade de Aveiro) pelo seu trabalho na atualização da plataforma INSaFLU para o novo coronavírus SARS-CoV-2.
- À Infraestrutura Nacional de Computação Distribuída (INCD), por ter fornecido recursos computacionais para testar a plataforma INSaFLU. O INCD foi financiado pela FCT e FEDER sob o projeto 22153-01 / SAICT / 2016.

Este estudo é co-financiado pela Fundação para a Ciência e Tecnologia (234_596874175) no âmbito da "call" Research 4 COVID-19.

Núcleo de Bioinformática | Departamento de Doenças Infecciosas
INSTITUTO NACIONAL DE SAÚDE DOUTOR RICARDO JORGE
Avenida Padre Cruz, 1649-016 Lisboa, PORTUGAL