# Análise da Prevalência de Alelos HLA em Pacientes com COVID-19

# Analysis of the Prevalence of HLA Alleles in Patients with COVID-19

# Análisis de la Prevalencia de Alelos HLA en Pacientes con COVID-19

Gabriel P. Mendes[1], Luís C. M. S. Pôrto[2], Cristiano Lima[3], Helton Santiago[4], Stephanie Almeida[4], Alexandre C. Sena[1]

1 Departamento de Informática e Ciência da Computação - Universidade do Estado do Rio de Janeiro (UERJ)
2 Laboratório de Histocompatibilidade e Criopreservação (UERJ)
3 Departamento de Cirurgia - Universidade Federal de Minas Gerais
4 Departamento de Bioquímica e Imunologia - Universidade Federal de Minas Gerais

Autor correspondente: Alexandre da Costa Sena
E-mail: asena@ime.uerj.br

**Resumo**

O objetivo do trabalho foi analisar a prevalência de alelos HLA em pacientes com COVID-19 através de um Estudo de Caso-Controle (ECC). A metodologia adotada foi utilizar os sistemas de informação SIVEP-Gripe, que registra casos de Síndrome Respiratória Aguda Grave (SRAG), e e-SUS, que registra casos suspeitos ou confirmados de COVID-19, para compor a base de dados para o ECC. Em seguida, foi realizada uma consulta na base de dados de doadores de órgãos (REDOME) para se obter os alelos HLA. Para gerar uma base de controle homogênea para o ECC a partir das características dos registros da base de casos, foi implementado um algoritmo de pareamento. Por fim, foi realizada à análise da prevalência de alelos HLA nos pacientes com COVID-19. Os resultados mostram a escolha balanceada do algoritmo proposto e a análise dos alelos mostrou diferenças entre a distribuição dos grupos alélicos em função da etnia/raça.

**Descritores:** Estudo Caso-Controle; alelos HLA; Algoritmo de Pareamento

**Abstract**

The aim of this work was to analyze the prevalence of HLA alleles in patients with COVID-19 through a Case-Control Study (CCS). The methodology used the information systems SIVEP-Gripe, which records cases of Severe Acute Respiratory Syndrome (SARS), and e-SUS, which records suspected or confirmed cases of COVID-19, to compose the database for the CCS. Then, a query was performed on the organ donor database (REDOME) to obtain the HLA alleles. To obtain a homogeneous control base for the CCS from the characteristics of the case base records, a matching algorithm was implemented. Finally, the analysis of the HLA alleles in patients with COVID-19 was performed. The results show the balanced choice of the algorithm and the allele analysis showed differences between the distribution of allelic groups as a function of ethnicity/race.

**Keywords:** Case-Control Study; HLA alleles; Matching Algorithm

**Resumen**

El objetivo del trabajo fue analizar la prevalencia de alelos HLA en pacientes con COVID-19 através de un Estudio de Casos y Controles (ECC). La metodología usó los sistemas de información SIVEP-Gripe, que registra casos de Síndrome Respiratorio Agudo Severo (SARS), y e-SUS, que registra casos sospechosos o confirmados de COVID-19, para componer la base de datos del ECC. Luego, se realizó una consulta a la base de datos de donantes de órganos (REDOME) para obtener los alelos HLA. Para generar una base de control homogénea para la ECC a partir de las características de los registros de la base de casos, se implementó un algoritmo de emparejamiento. Finalmente, se realizó el análisis de prevalencia de alelos HLA en pacientes con COVID-19. Los resultados muestran la elección equilibrada del algoritmo propuesto y el análisis de alelos mostró diferencias entre la distribución de los grupos alélicos en función de la etnia/raza.

**Descriptores:** Estudio de casos y controles; alelos HLA; Algoritmo de coincidencia

## Introdução

O controle epidemiológico da COVID-19 foi um dos problemas sanitários no qual se evidenciou a importância de se integrar eficientemente os sistemas de informação

usados pelo SUS. Durante o combate a pandemia no Brasil, dois destes sistemas de informação se destacaram para a gestão de saúde pública: SIVEP-Gripe[1] e e-SUS[2]. Ambos são usados para registrar notificações e apoiar no processo de monitoramento de casos de COVID-19. O SIVEP-Gripe é um sistema de vigilância que já era usado para registrar casos de Síndrome Respiratória Aguda Grave (SRAG) que passou também a incluir casos graves de infecções causados especificamente pelo Sars-CoV-2. Por sua vez, o sistema e-SUS foi criado durante a pandemia para registrar casos suspeitos ou confirmados de COVID-19.

Por outro lado, o Ministério da Saúde do Brasil também conta com uma base de dados com as características genéticas que permitem a busca de possíveis doadores de células tronco hematopoéticas (Registro Nacional de Doadores Voluntários de Medula Óssea-REDOME) para pacientes que necessitam receber essas células como tratamento (Registro Nacional de Receptores de Medula Óssea - REREME). As compatibilidades desses genes além de serem definidoras do processo de rejeição estão implicadas nas respostas imunológicas para patógenos (e.g. vírus e bactérias) e dependendo da combinação de um ou mais alelos desses genes estão associados com a susceptibilidade de determinadas doenças e uma resposta imunológica efetiva após a vacinação.

Assim, o objetivo central deste trabalho é propor e implementar uma abordagem para criação de uma base de dados confiável para Estudo Caso-Controle (ECC) da prevalência de alelos HLA em pacientes com COVID-19, a partir dos dados das bases SIVEP-Gripe e e-SUS, combinadas com os alelos HLA da base de dados do REDOME. Ainda, a partir da base de dados gerada pelo algoritmo de pareamento proposto, analisar a prevalência de alelos HLA em pacientes com COVID-19.

Para atingir esse objetivo, as seguintes etapas foram necessárias: tratamento, uniformização e integração das bases de dados; busca das informações sobre os alelos na base de dados do REDOME; criação de uma base Estudo Caso-Controle (ECC)[4] confiável; avaliação da prevalência de alelos. Todas essas etapas são descritas neste trabalho, em especial, o algoritmo de pareamento para criação de uma base ECC. O algoritmo de pareamento tem o objetivo de criar um arquivo de controle a partir das características da base de casos. O balanceamento entre a base de controle e a base de casos é muito importante, especialmente quando se considera a etnia dos pacientes que

tem uma influência direta no alelo que pode ser encontrado, ainda mais quando se considera populações miscigenadas, que traçam sua ascendência através de diferentes regiões geográficas. Por exemplo, o trabalho apresentado em [5] mostrou que pacientes com ascendência africana enfrentam maior dificuldade em localizar potenciais doadores em função da maior variabilidade HLA das populações Africanas e da maior presença de doadores brancos no banco de dados de doadores de órgãos avaliado no estudo. Por outro lado, o ECC para a prevalência de alelos pode ajudar a identificar alelos que possam ajudar a fornecer uma proteção ou um risco maior a COVID-19. Por exemplo, o trabalho em [3] identificou alelos relacionados a disseminação da COVID-19 na Itália.

O algoritmo proposto tem potencial para ajudar bastante a comunidade científica, não só permitindo a criação de uma base de controle balanceada em relação aos casos, mas também por evitar a necessidade do pesquisador ter que escolher os casos da base de controle, evitando assim um possível viés de seleção que é um dos problemas da técnica de ECC [4]. A avaliação do algoritmo de pareamento proposto mostrou que ele foi capaz de produzir um arquivo de Controle homogêneo, especialmente para os campos de maior prioridade. Por sua vez, o ECC realizado apontou diferenças na distribuição dos grupos alélicos em função da autodeclaração raça/etnia. Ainda, a análise identificou que o alelo B*51 oferece uma maior proteção, enquanto que o alelo A*36 aumenta o risco.

## Métodos

O estudo foi aprovado pela Comissão Nacional de Ética em Pesquisa - Conep (CAAE - 40921320.1.0000.5259). Os pesquisadores envolvidos assinaram termo de confidencialidade sobre o uso dos dados pessoais utilizados para cruzamento das bases de dados e a base de estudo gerada foi oferecida para uso já com uma codificação anonimizada dos registros, em consonância também com a Lei Geral de Proteção de Dados Pessoais (LGPD).

A metodologia adotada para criação de uma base de Estudo Caso-Controle (ECC) para a análise da prevalência de alelos HLA em pacientes com COVID-19 pode ser dividida em três etapas. A seguir são descritas cada uma dessas etapas.

**Etapa 1 - Aquisição e Tratamento dos Dados**

A etapa 1 define as bases de dados de onde serão extraídos os pacientes que tiveram casos confirmados ou suspeitos da COVID-19. Para isso, dois Sistemas de Informação em Saúde (SIS) estão sendo de suma importância durante a pandemia de COVID-19 para ajudar a monitorar os casos da doença. O SIVEP-Gripe[1], que é utilizado para armazenar todas as notificações referentes a Síndromes Respiratórias Agudas Graves (SRAG), encerradas ou não. O e-SUS[2] que foi criado durante a pandemia para registrar casos suspeitos ou confirmados especificamente de COVID-19. Esta primeira etapa se caracteriza, inicialmente, pela exportação dos dados armazenados nas bases SIVEP e e-SUS para arquivos no formato CSV.

Em seguida. foi feita uma análise nas duas bases de dados para investigar possíveis inconsistências nos dados, duplicação de registros, erros de digitação, entre outros problemas. Foram encontrados os seguintes problemas: campos com valores em branco; campos com valores com formato divergentes (cpf e datas); campos com valores divergentes para um mesmo paciente como, por exemplo, duas notificações referentes ao mesmo paciente apresentarem endereços distintos; registros duplicados. Por fim, esses arquivos foram tratados e armazenados de forma adequada para serem utilizados nas etapas seguintes.

**Etapa 2 - Aquisição dos Alelos HLA**

As bases SIVEP-Gripe e e-SUS, geradas na etapa 1, contêm apenas informações sobre a situação dos pacientes com sintomas ou que tiveram a COVID-19, como por exemplo, o resultado do teste, tipo do teste, data de notificação, entre outras. A etapa 2 apresenta a estratégia adotada para se conseguir os alelos HLA desses pacientes. Uma possibilidade seria realizar o exame de tipificação HLA de todos (ou um grupo) os pacientes das bases SIVEP e e-SUS. Essa abordagem é impraticável por dois motivos principais. Primeiro, o alto custo envolvido na realização de exame de tipificação HLA (i.e. exame que identifica os alelos de um paciente) para cada um dos pacientes que foram utilizados na pesquisa. Em segundo lugar, muitos pacientes presentes nessas bases

faleceram de COVID-19. Assim, a abordagem proposta para se conseguir os alelos HLA para os pacientes das bases SIVEP e e-SUS foi realizar um cruzamento de dados com o REDOME, uma vez que a compatibilidade HLA entre paciente-doador é a principal informação armazenada nesta base de dados. Além dos alelos HLA, foi obtido também o campo raça/etnia do REDOME, uma vez que esses dados estavam mais consistentes que os dados raça/etnia das bases SIVEP e e-SUS.

**Etapa 3 – Criação da Base de Controle**

A última etapa, consistiu em gerar a base de controle em função das características dos pacientes da base de casos. A abordagem proposta é a criação de um algoritmo de pareamento que escolha os registros para a base de controle de acordo com as características dos registros da base de casos. A utilização de um algoritmo para escolha dos registros tem dois objetivos. Em primeiro lugar, criar uma base de Controle com características similares a base de Casos, especialmente com relação ao campo raça/etnia que tem influência no alelo que pode ser encontrado. Em segundo lugar, evitar que o pesquisador tenha que participar da escolha dos registros de forma não randômica no pareamento entre casos e controles, evitando um possível viés de seleção.

O arquivo SIVEP final gerado, composto de pacientes de SRAG que testaram positivos para COVID-19, será a base de Casos da análise. Dentro dessa base, os registros são categorizados em 4 grupos de acordo com a evolução do estado do paciente: Recuperado, Internado, UTI e Óbito. Os pacientes com evolução do caso Recuperado não precisaram ser internados e conseguiram se recuperar da doença. Por sua vez, os pacientes com evolução do caso Internado tiveram que ser internados, mas não precisaram de Unidade de Terapia Intensiva (UTI) e conseguiram se recuperar da doença. Já, os pacientes com evolução do caso UTI tiveram que ser internados em UTIs, mas conseguiram se recuperar da doença. Por fim, os pacientes com evolução do caso Óbito morreram de COVID-19.

Por sua vez, a base de controle será composta por dados provenientes da base e-SUS, que contêm casos suspeitos ou confirmados de COVID-19. A maior dificuldade em um Estudo Caso-Controle é a seleção do grupo de controle. Para um estudo confiável

da prevalência de alelos uma base homogênea é fundamental, uma vez que, principalmente, a raça/etnia e a região podem ter uma influência significativa no alelo que pode ser encontrado. Para resolver este problema este trabalho propõe um algoritmo para criar uma base de Controle, a partir das características dos registros da base de Casos. Assim, este trabalho implementa um algoritmo para realizar o pareamento entre uma base de Casos e uma base de Controle, que consiste em encontrar para cada registro do SIVEP um ou mais registros (Positivo e Negativo) do e-SUS que apresentem semelhanças considerando características como: etnia, sexo, região, entre outras.

O objetivo do algoritmo é parear cada registro do SIVEP com uma quantidade 2 x N de registros do e-SUS, sendo N registros com resultado do teste de COVID-19 Positivo e N Negativo. Mais importante, é necessário que o algoritmo escolha registros do e-SUS com características similares as do registro do SIVEP para que a base de controle seja homogênea. Como a base de controle a ser gerada será utilizada para um estudo sobre a prevalência de alelos HLA para pacientes com COVID-19, os campos escolhidos para serem pareados foram: etnia; data de notificação; município; região; sexo; tipo do teste.

Outra característica importante é a idade do paciente. Porém, encontrar pacientes na base e-SUS com exatamente a mesma idade não seria viável. Logo, os registros da base de casos SIVEP foram divididos em 3 faixas etárias distintas através do cálculo do percentil 33 e 67 para as idades da base de Casos (SIVEP). Ou seja, a primeira faixa etária é composta do valor da menor idade da base até o valor do percentil 33% (arredondado para baixo). A segunda faixa é composta do valor do percentil 33% (arredondado para cima) até a o valor do percentil 67% (arredondado para baixo). Por fim, a terceira faixa é composta do valor do percentil 67% (arredondado para cima) até o valor da maior idade da base. No momento do pareamento apenas registros com as idades dentro da faixa são selecionados.

O código implementado pode ser visto no Algoritmo 1, que recebe como entrada os arquivos de caso (arqSivep) e controle (arqSus), além da quantidade N que deseja parear. Por exemplo, se N=3 o algoritmo irá parear 3 registros Positivos e 3 registros Negativos da base de controle para cada registro do SIVEP. O valor de N é uma informação importante que o pesquisador deve fornecer de acordo com o tamanho da

amostra e o tamanho da base de dados disponível. A saída do algoritmo é o arquivo final que contém a base de controle (listaPar).

Algoritmo 1 – Pareamento Caso-Controle

```
Input: arqSivep, arqSus, N
1: for all reg ∈ arqSivep do
2:     filtro ← 1; semanas ← 1;
3:     listaPos ← ∅; listaNeg ← ∅; listaPar ← ∅;
4:     while (listaPos.tam < N ∨ listaNeg.tam < N) ∧ filtro < 10 do
5:         listaSus ← arqSus - listaPar
6:         if filtro < 9 then
7:             listaSus ← FILTRAR (listaSus, reg.etnia)
8:         if filtro < 8 then
9:             listaSus ← FILTRAR (listaSus, reg.dataNotificacao, semanas)
10:        if filtro < 4 then
11:            listaSus ← FILTRAR (listaSus, reg.municipio)
12:        if filtro == 4 then
13:            listaSus ← FILTRAR (listaSus, reg.regiao)
14:        if filtro < 3 then
15:            listaSus ← FILTRAR (listaSus, reg.sexo)
16:        if filtro < 2 then
17:            listaSus ← FILTRAR (listaSus, reg.tipoTeste)
18:        if listaPos.tam < N then
19:            listaPos ← listaPos ∪ RECEBEPOSITIVOS (listaSus, N)
20:        if listaNeg.tam < N then
21:            listaNeg ← listaNeg ∪ RECEBENEGATIVOS (listaSus, N)
22:        if listaPos.tam == N ∧ listaNeg.tam == N then
23:            filtro ← 10;
24:        if filtro > 4 ∧ filtro < 8 then
25:            semanas ← semanas + 1
26:        filtro ← filtro + 1
27:    end while
28:    listaPar ← listaPar ∪ listaPos
29:    listaPar ← listaPar ∪ listaNeg
30: end for
```

O pareamento é realizado das linhas 1 a 30 do algoritmo, que executa o pareamento para cada registro, reg, da base SIVEP. Na linha 2, as variáveis filtro e semanas são inicializadas. As listas que guardam os registros positivos encontrados (listaPos), os registros negativos encontrados (listaNeg) e lista final contendo todos os

registros da base de controle a ser gerada (listaPar) são inicializados com valor nulo (linha 3). O processo de pareamento de um registro acontece da linha 4 a 27 através de um laço do tipo **while**. Esta ação só termina quando forem encontrados N registros Positivos e N registros Negativos ou se não existirem N registros mesmo após todos os filtros serem liberados. Para cada registro a ser pareado a lista de registros do e-SUS que podem ser selecionados (listaSus) é inicializada com todos os registros da base e-SUS (arqSus) menos os registros que já foram selecionados pelo algoritmo de pareamento que ficam armazenados na lista listaPar (linha 5).

A próxima etapa do algoritmo é aplicar os filtros que definem as características a serem pareadas. A aplicação ou não do filtro é definida pela variável filtro, inicializada com valor 1. Ou seja, inicialmente todos os filtros serão aplicados. Assim, a variável que contém os registros do e-SUS que podem ser selecionados (listaSus) primeiro é filtrada pelo campo etnia (linhas 6 e 7) presente no registro SIVEP (e.g. apenas registros com etnia PARDA serão selecionados caso a etnia do registro SIVEP seja PARDA). Em seguida, é aplicado o filtro pela data de notificação (linha 8 e 9). Ou seja, inicialmente, apenas os registros que tiverem data de notificação uma semana para frente ou para trás da data de notificação do registro do SIVEP serão selecionados. Repare que esse filtro é aplicado em cima do filtro da Etnia (e.g. apenas registros da etnia PARDA no período de uma semana para frente ou para trás da data de notificação serão selecionados). O próximo filtro a ser aplicado é o do município que seleciona apenas registros do mesmo município que o registro do SIVEP (linhas 10 e 11). Como selecionar pacientes de um mesmo município pode ser inviável, o algoritmo está preparado para, caso não encontre os N registros Positivos e Negativos de um mesmo município, filtrar por região (linhas 12 e 13). O processo de filtragem continua aplicando o filtro sobre o sexo (linhas 14 e 15) e, depois, sobre o tipo do teste (linhas 16 e 17). Em seguida, o algoritmo armazena na lista de positivos (listaPos) os N primeiros registros Positivos após toda a filtragem (linhas 18 e 19) e na listaNeg os primeiros registros Negativos após toda a filtragem. Caso tenha encontrado registros suficientes, o pareamento do registro atual termina atribuindo o valor 10 para variável filtro (linhas 22 e 23). Repare que neste caso o pareamento realizado foi extremamente adequado, pois todos os filtros desejados foram aplicados.

Caso contrário, o pareamento do registro atual irá continuar, porém, com um filtro a menos pois a variável filtro é incrementada de 1 (linha 26). Nesse caso, o processo irá continuar até conseguir a quantidade de registros Positivos e Negativos desejada, sendo que a cada nova iteração um novo filtro é retirado.

## Resultados e Discussão

Esta seção descreve os resultados obtidos. Inicialmente, são apresentados os resultados da avaliação do algoritmo de pareamento proposto. Em seguida, é apresentada a análise da prevalência de alelos HLA para pacientes com COVID-19 das bases SIVEP e e-SUS. Para avaliar os resultados foi utilizado o *software* estatístico Epi Info projetado pelo CDC (*Centers for Disease Control and Prevention*) para a comunidade global de médicos e pesquisadores da saúde pública.

### Análise do Algoritmo de Pareamento

Para comparar os resultados produzidos pelo algoritmo proposto foram criados outros dois algoritmos. O algoritmo chamado de *Sem Filtro* é basicamente o mesmo algoritmo apresentado, porém sem aplicar nenhum filtro. Ou seja, ele escolhe os N primeiros registros Positivos e Negativos disponíveis para cada registro do SIVEP. Por sua vez, o algoritmo chamado de *Aleatório* seleciona para cada registro do SIVEP N registros Positivos e Negativos aleatoriamente. Ainda, para verificar o desempenho do algoritmo de acordo com o tamanho de N, cada um dos 3 algoritmos foi executado com os valores 2, 3 e 4.

Inicialmente, foi avaliada a distribuição do campo idade. Conforme descrito na seção anterior, o arquivo SIVEP foi dividido em 3 faixas etárias, de maneira que de acordo com a idade do registro do SIVEP apenas os registros do e-SUS com a idade dentro da faixa prevista ficam disponíveis. A menor idade, maior idade e idade média para cada um dos arquivos de controle gerado pelos algoritmos podem ser visto na Tabela 1.

*Tabela 1 - Valores máximo, médio e mínimo para o campo idade*

| Idade | SIVEP | e-SUS | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Com Filtro | | | Sem Filtro | | | Aleatório | | |
| | | 2x2 | 3x3 | 4x4 | 2x2 | 3x3 | 4x4 | 2x2 | 3x3 | 4x4 |
| Mínima | 22 | 23 | 22 | 22 | 19 | 19 | 19 | 20 | 19 | 19 |
| Média | 49 | 48 | 47 | 48 | 40 | 40 | 41 | 42 | 41 | 41 |
| Máxima | 75 | 73 | 73 | 73 | 72 | 72 | 72 | 69 | 74 | 74 |

Quando se compara os valores gerados por cada um dos 3 algoritmos com os valores do SIVEP, fica claro o melhor desempenho do algoritmo proposto (*Com Filtro*). Não só as idades mínima e máxima ficaram muito próximas, mas, principalmente a média aritmética das idades. Com relação a variação do valor de N, não houve mudança significativa. O campo mais importante do algoritmo de pareamento proposto é o referente a etnia do paciente, uma vez que a essa característica tem uma influência significativa no alelo que pode ser encontrado. Em função disso, o primeiro filtro a ser aplicado é o de etnia, conforme descrito na Seção Métodos. Ou seja, ele é o último filtro a ser ignorado no algoritmo de pareamento. Os resultados da frequência (em porcentagem) para cada uma das etnias presentes no arquivo de controle gerado pelos algoritmos, assim como o valor do chi-quadrado, podem ser visualizados na Tabela 2.

*Tabela 2 - Valores em percentual (%) para frequência dos valores para o campo Etnia*

| Etnia | SIVEP | e-SUS | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Com Filtro | | | Sem Filtro | | | Aleatório | | |
| | | 2x2 | 3x3 | 4x4 | 2x2 | 3x3 | 4x4 | 2x2 | 3x3 | 4x4 |
| Amarela | 1,39 | 1,43 | 1,42 | 1,43 | 1,65 | 1,39 | 1,34 | 1,30 | 1,30 | 1,47 |
| Branca | 59,62 | 59,88 | 60,51 | 60,66 | 64,25 | 63,78 | 63,58 | 62,61 | 62,71 | 61,63 |
| Ignorado | 0,0 | 0,0 | 0,0 | 0,11 | 1,47 | 1,42 | 1,67 | 1,56 | 1,73 | 1,71 |
| Indigena | 1,04 | 1,08 | 0,92 | 0,76 | 0,35 | 0,35 | 0,41 | 0,43 | 0,23 | 0,37 |
| Não Inf. | 2,95 | 2,56 | 1,99 | 1,58 | 0,17 | 0,35 | 0,45 | 0,26 | 0,35 | 0,19 |
| Parda | 18,89 | 18,93 | 19,04 | 19,24 | 18,80 | 19,61 | 19,95 | 21,36 | 20,85 | 20,93 |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Preta | 16,12 | 16,12 | 16,12 | 16,23 | 13,30 | 13,11 | 12,59 | 12,48 | 12,82 | 13,69 |
| chi-2 | | 0,28 | 2,25 | 6,78 | 66,23 | 64,83 | 63,83 | 60,51 | 71,63 | 96,19 |

Ao se comparar os valores de frequência para cada uma das etnias dos arquivos de controle gerados pelos algoritmos de pareamento com os valores da base de casos (SIVEP), pode-se perceber a melhor distribuição produzida pelo algoritmo *Com Filtro*. Por exemplo, ao se considerar a etnia Ignorado, a distribuição para N=2 e N=3 foi exatamente igual à do SIVEP (0.0), enquanto que para N=4 foi muito similar (0.11). Por outro lado, quando analisamos os valores produzidos pelos outros dois algoritmos se percebe um percentual muito maior. Outra observação importante, é que a distribuição produzida pelo algoritmo *Com Filtro* piora a medida que o valor de N aumenta. Esse comportamento é esperado uma vez que, ao se aumentar valor de N, a cada passo do algoritmo existirá menos opções de registros para serem usadas no próximo pareamento, o que pode fazer com que o algoritmo não consiga aplicar um ou mais filtros. Por exemplo, com N=4 após parear o primeiro registro do SIVEP, 8 registros a menos estarão disponíveis (4 Positivos e 4 Negativos), enquanto que com N=2 apenas 4 registros não estarão disponíveis. Esse mesmo comportamento não acontece nos outros dois algoritmos, pois a escolha dos registros não utiliza nenhum tipo de prioridade.

Além da frequência dos valores para cada etnia, como esse campo é muito relevante para a análise da prevalência de alelos HLA, foi realizado o teste chi-quadrado para comparar duas variáveis categóricas e verificar se são homogêneas entre si. Neste caso, é comparado a variável etnia do SIVEP com a variável etnia do arquivo de controle gerado pelo algoritmo. Os valores para chi-2 na tabela mostram claramente que o algoritmo *Com Filtro* produziu os resultados mais homogêneos, uma vez que quanto menor o valor do chi-2 melhor. Mais importante, a observação que a probabilidade do chi-2 obtido foi 99,79% para N=2 e 81,29% para N=3 confirma a homogeneidade dos valores obtidos. Por sua vez, para N=4, a probabilidade encontrada foi de 34%, mostrando que para esse valor de N, a distribuição não está tão homogênea. Por outro lado, quando analisamos o valor do chi-2 para os outros dois algoritmos, é possível observar que os resultados produzidos estão completamente desbalanceados.

Todos os outros campos foram analisados e o algoritmo de pareamento proposto sempre produziu um balanceamento muito superior aos outros dois algoritmos. Entretanto, como a cada novo filtro a prioridade é menor, o balanceamento tem a tendência de piorar de acordo com a diminuição da prioridade do filtro.

**Análise da prevalência de Alelos HLA em Pacientes com COVID-19**

A análise da prevalência de alelos HLA foi realizada utilizando os dados dos arquivos SIVEP e e-SUS de pacientes residentes no estado de Minas Gerais. Após todas as etapas descritas na seção Métodos, a base de casos ficou composta de 1858 registros e a base e-SUS para geração da base de controle com 31.625 registros.

Foi utilizado o arquivo de controle com N=2 gerado pelo algoritmo de pareamento proposto. O arquivo gerado é balanceado, especialmente com relação a etnia, e o pareamento ficou o mais próximo possível dentro das opções dos registros disponíveis. Foi realizada a análise que comprovou diferenças entre a distribuição dos grupos alélicos em função da autodeclaração raça/etnia. Mais especificamente, não foram encontradas diferenças entre a distribuição dos grupos alélicos nos locos A e DRB1 em pacientes com raça/etnia Branca, nos locos A, B e DRB1 em pacientes com raça/etnia Parda e nos locos B e DRB1 com raça/etnia Preta. Por sua vez, quando comparados os 4 grupos de acordo com a evolução do estado do paciente (Recuperado, Internado, UTI e Óbito) existem diferenças na frequência dos grupos alélicos do loco B em pacientes com raça/etnia Branca e do loco A em pacientes com raça/etnia Preta.

A comparação entre a frequência alélica do Alelo A*36 foi maior nos pacientes autodeclarados com raça/etnia Preta internados com COVID-19 em UTI que no grupo Recuperado (razão de chance 9,6 IC: 2,6-34,8; p=0,001) ou no Grupo Internado (razão de chance 7,8 IC: 1,9-32,5; p=0,005), indicando que este alelo oferece um risco maior para pacientes com COVID-19. Por outro lado, a comparação entre a frequência alélica do alelo B*51 foi significativamente maior nos pacientes autodeclarados com raça/etnia Branca sem internação com COVID-19 do que nos Internados (razão de chance 0,7 IC: 0,5-0,9; p=0,021) ou do Grupo Óbito (razão de chance 0,5 IC: 0,2-0,9; p=0,014), indicando que este alelo oferece uma proteção maior a COVID-19.

## Conclusão

Este trabalho apresentou uma abordagem para criação de uma base de Estudo Caso-Controle para avaliação da prevalência de alelos HLA em pacientes com COVID-19. Em especial, foi proposto e implementado um algoritmo de pareamento capaz de criar uma base de controle homogênea, a partir das características dos registros da base de casos. Em seguida, foi realizada análise da prevalência de alelos HLA em pacientes com COVID-19 residentes no estado de Minas Gerais.

Os resultados apresentados mostraram a escolha balanceada do algoritmo proposto, em especial, para os campos com maior prioridade. A comparação dos resultados com outros dois algoritmos que não utilizam nenhum tipo de prioridade, mostra claramente os benefícios da utilização do algoritmo proposto. Por fim, a análise da prevalência de alelos HLA mostrou que existe diferenças na distribuição dos grupos alélicos em função do campo raça/etnia, onde o alelo B*51 tem uma chance maior de oferecer proteção, enquanto que o alelo A*36 aumenta o risco para a COVID-19.

## Agradecimentos

Os autores agradecem o apoio da CAPES e FAPERJ através do edital APQ1 26/2021.

## Referências

1. Brasil, Ministério da Sáude. e-SUS Notifica [Internet]. Available from: https://www.gov.br/saude/pt-br/composicao/svs/sistemas-de-informacao/e-sus-notifica
2. Brasil, Ministério da Sáude. SIVEP-Gripe [Internet]. Available from: https://sivepgripe.saude.gov.br/sivepgripe/
3. Correale P, Mutti L, Pentimalli F, et al. HLA-B*44 and C*01 Prevalence Correlates with Covid19 Spreading across Italy. Int J Mol Sci. 2020; 21(15).
4. Tenny S, Kerndt CC, Hoffman MR Case Control Studies. StatPearls [Internet]. 2022.
5. Nunes K, Aguiar V, Silva M, et al. How Ancestry Influences the Chances of Finding Unrelated Donors: An Investigation in Admixed Brazilians Front. Immunol. 11:584950. doi: 10.3389/fimmu.2020.584950