UNIVERSIDADE FEDERAL DE MINAS GERAIS
PROGRAMA INTERUNIDADES DE PÓS-GRADUAÇÃO EM BIOINFORMÁTICA

Fábio Rodrigues Martins

DESENVOLVIMENTO DE UMA METODOLOGIA COMPUTACIONAL PARA ANÁLISE DE VARIANTES GÊNICAS NO LOCUS DE CADEIA PESADA DE IMUNOGLOBULINAS HUMANAS

Belo Horizonte
2022

Fábio Rodrigues Martins

DESENVOLVIMENTO DE UMA METODOLOGIA COMPUTACIONAL PARA ANÁLISE DE VARIANTES GÊNICAS NO LÓCUS DE CADEIA PESADA DE IMUNOGLOBULINAS HUMANAS

Tese apresentada ao Programa Interunidades de Pós-Graduação em Bioinformática da Universidade Federal de Minas Gerais, como requisito parcial à obtenção do título de Doutor em Bioinformática.

Orientadora: Liza Figueiredo Felicori Vilela
Coorientador: Tiago Antônio de Oliveira Mendes

Belo Horizonte
2022

043 Martins, Fábio Rodrigues.
Desenvolvimento de uma metodologia computacional para análise de variantes gênicas no lócus de cadeia pesada de imunoglobulinas humanas [manuscrito] / Fábio Rodrigues Martins. – 2022.
193 f. : il. ; 29,5 cm.

Orientadora: Liza Figueiredo Felicori Vilela. Coorientador: Tiago Antônio de Oliveira Mendes.
Tese (doutorado) – Universidade Federal de Minas Gerais, Instituto de Ciências Biológicas. Programa Interunidades de Pós-Graduação em Bioinformática.

1. Bioinformática. 2. Alótipos de Imunoglobulina. 3. Bases de Dados Genéticas. I. Vilela, Liza Figueiredo Felicori. II. Mendes, Tiago Antônio de Oliveira. III. Universidade Federal de Minas Gerais. Instituto de Ciências Biológicas. IV. 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 Interunidades de Pós-Graduação em Bioinformática da UFMG

FOLHA DE APROVAÇÃO

FÁBIO RODRIGUES MARTINS

"Desenvolvimento de Uma Metodologia Computacional Para Análise de Variantes Gênicas No Lócus de Cadeia Pesada de Imunoglobulinas Humanas"

Tese aprovada pela banca examinadora constituída pelos Professores:

Profa. Liza Figueiredo Felicori Vilela - Orientadora
UFMG

Prof. Tiago Antonio de Oliveira Mendes - Coorientador
UFV

Prof. José Miguel Ortega
UFMG

Prof. Gabriel da Rocha Fernandes
FIOCRUZ

Prof. Marcelo de Macedo Brigido
UNB

Prof. Raony Guimarães Corrêa Do Carmo Lisboa Cardenas
BIO BUREAU

Belo Horizonte, 20 de julho de 2022.

Documento assinado eletronicamente por **Tiago Antônio de Oliveira Mendes**, **Usuário Externo**, em 20/07/2022, às 19:03, 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 **Gabriel da Rocha Fernandes**, **Usuário Externo**, em 20/07/2022, às 19:04, 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 **Marcelo de Macedo Brígido**, **Usuário Externo**, em 20/07/2022, às 19:04, 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 **Raony Guimaraes Correa do Carmo Lisboa Cardenas**, **Usuário Externo**, em 20/07/2022, às 19:06, 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 **Liza Figueiredo Felicori Vilela**, **Professora do Magistério Superior**, em 22/07/2022, às 11:13, 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 **Jose Miguel Ortega**, **Servidor(a)**, em 22/07/2022, às 12:27, 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 **1587316** e o código CRC **B47856C5**.

Referência: Processo nº 23072.240034/2022-19 SEI nº 1587316

AGRADECIMENTOS

Primeiramente, agradeço a Deus pela saúde, pela vida, por todas as bênçãos e por me dar forças para superar os dias difíceis.

Agradeço a minha mãe Lindaura e ao meu pai Francisco, por todo apoio.

A minha esposa Ezilane e minha filha Joana, pelo amor e pelo incentivo.

Aos meus orientadores, Liza e Tiago Mendes, por acreditar na minha capacidade de desenvolver este trabalho, muito obrigado.

Aos colegas do Synbiom, Bruno, Carlena, Christina, Iago, Igor, João, Laís, Luana, Lucas Minto, Lucas Pontes, Luciana, Manuela, Marcella, Maura, Milene, Raniele, Regina, Vivi e Yala, muito obrigado pela convivência nos corredores e papos no café, essa troca de ideia contribui para a resolução dos problemas encontrados durante a pesquisa.

Agradeço em especial ao Lucas Pontes, Luciana e a Marcele que colaborou com as revisões da tese.

A todo o corpo docente do programa de Pós-graduação em Bioinformática, aprendi baste durante estes 4 anos que estive no departamento.

Agradeço a todos os funcionário do ICB, ao pessoal da limpeza, da manutenção dos equipamentos de internet e etc. Em especial, ao pessoal da secretaria do programa de Bioinformática, Sheila e o Tiago, que auxilia os alunos com profissionalismo e dedicação, parabéns pelo atendimento.

Aos colegas do departamento de informática do IFMG compus São João Evangelista, obrigado pelo incentivo.

À Universidade Federal de Minas Gerais.

Enfim, agradeço a todos que direta ou indiretamente, contribuíram para esta tese. Do fundo do meu coração obrigado.

RESUMO

As células B da resposta imune adaptativa de humanos e outros vertebrados possuem uma diversidade enorme de receptores de antígenos. A diversidade desses receptores é gerada por alguns processos como: variabilidade do lócus de imunoglobulinas, processo de recombinação e de hipermutação somática. Estes processos influenciam na identificação de novos alelos de imunoglobulina humana, pois dada uma variante não é uma tarefa simples dizer se a mesma veio realmente da linhagem germinativa do indivíduo ou se foi gerada pelos processos citados acima. Porém, identificar novos alelos de imunoglobulina humana é importante para estudos de maturação e afinidade destas moléculas. A identificação correta dos alelos pode auxiliar nos estudos de diversas doenças humanas associadas ao repertório de anticorpos e no desenvolvimento de novos terapêuticos por meio de técnicas de engenharia de anticorpos. Para contribuir com a descoberta de novos alelos, o advento de *Next Generation Sequencing* (NGS) possibilitou que genomas humanos fossem sequenciados com boa cobertura. Com isto, várias ferramentas, bancos de dados e abordagens que estudam repertório de anticorpos humanos têm sido desenvolvidos para o mapeamento e identificação de novos alelos de imunoglobulina. A metodologia desenvolvida neste trabalho identificou 10.550 alelos de genes IGHV, sendo que 10.156 são novos alelos putativos. Também foram identificados 524 variantes de genes IGHD e 670 de gene IGHJ. Um banco de dados integrado a uma plataforma *web* foi criado (YVr-DB) para armazenar e tornar acessível às prováveis novas variantes encontradas.

PALAVRAS-CHAVE: Variantes de Imunoglobulina, Genes de Anticorpos, gnomAD, GWAS, Banco de Dados

ABSTRACT

The B and T cells of the adaptive immune response of humans and other vertebrates have an enormous diversity of receptors. The diversity of these receptors is generated by some processes such as: variability of the immunoglobulin locus, recombination process and somatic hypermutation. These processes influence the identification of new human immunoglobulin alleles, because given a variant it is not a simple task to say if it really came from the individual's germ line or if it was generated by the processes mentioned above. However, identifying new alleles of human immunoglobulin is important for studies of maturation and affinity of these molecules. The correct identification of alleles can help in the studies of several human diseases associated with the antibody repertoire and in the development of new therapeutics through antibody engineering techniques. To contribute to the process of discovering new alleles, the advent of Next Generation Sequencing (NGS) made it possible for human genomes to be sequenced with good coverage. With this, several tools, databases and approaches that study the human antibody repertoire have been developed for the mapping and identification of new immunoglobulin alleles. The methodology developed in this work identified 10,550 alleles of IGHV genes, of which 10,156 are new putative alleles. We also identified 524 alleles of IGHD genes and 670 alleles of IGHJ gene. A database integrated to a web platform was created (YVr-DB) to store and make accessible the likely new variants found.

KEYWORDS: Immunoglobulin Variants, Antibody Genes, gnomAD, GWAS, Database

LISTA DE FIGURAS

Figura 1: Esquema da estrutura de um anticorpo humano. 16
Figura 2: Representação do lócus de IGH de humano na banda 14q32.33 18
Figura 3: Etapas da recombinação VDJ de cadeia pesada de anticorpos 19
Figura 4: Identificação de novos VH alelos no IgDiscover. 34
Figura 5: Nomenclatura do IMGT com as regiões dos genes V 48
Figura 6: Alinhamento múltiplo dos segmentos gênicos IGHV1-18. 54
Figura 7: Tela inicial da Plataforma *web* YVr-DB..... 65
Figura 8: Tela que apresenta o resultado de uma busca submetida à plataforma YVr-DB. 66
Figura 9: Tela de detalhes da variante no YVr-DB. 66
Figura 10: Gráfico da Plataforma *web* YVr-DB, guia *Summary*..... 67
Figura 11: Diagrama de tabelas relacionas (DTR). 69
Figura 12: Variantes IGHV descritas neste trabalho e também encontradas em outras GLDB. 71
Figura 13: Número de variantes por posição. 76
Figura 14: Número de variantes de substituição não-sinônimas (*missense*) e sinônimas presentes nos diferentes genes IGHV 78
Figura 15: Número de deleções (..... 80
Figura 16: Número de variantes população-específicas ou compartilhadas por segmento gênico IGHV. 83
Figura 17: Métrica VQSLOD versus métrica QD dos grupos G1, G2, G3 e G4. 88
Figura 18: Métrica VQSLOD versus métrica QD do grupo com as variantes reportadas em outros bancos e o grupo das variantes novas. 90
Figura 19: Teste *post-hoc* de *Dunn* da métrica QD entre os grupos G1, G2, G3 e G4. 92
Figura 20: Teste *post-hoc* de *Dunn* da métrica VQSLOD entre os grupos G1, G2, G3 e G4. 94
Figura 21: Quantidade de variantes por gene IGHD. 96
Figura 22: Quantidade de variantes por gene IGHJ. 98

LISTA DE TABELAS

Tabela 1: Trabalhos que identificam alelos a partir de dados de Rep-seq..... 31
Tabela 2: Trabalhos que identificam alelos de Imunoglobulina a partir de dados de genoma 36
Tabela 3: Comparação das posições dos genes IGHV utilizando anotações do NCBI e do GENCODE. 62
Tabela 4: Número de vezes que cada variante putativa foi sequenciada no exoma (contagem de alelos)..... 73
Tabela 5: IGHV variantes por subgrupo (família de genes)..... 74
Tabela 6: As 10 variantes de IGHV mais prevalentes e não presentes no IMGT e IgPDB 81
Tabela 7: Número de variantes únicas por população e a variante mais frequente da população..... 83
Tabela 8: Variantes presentes no Catálogo GWAS..... 86
Tabela 9: Número de vezes que cada variante putativa do gene D foi identificada em genoma ou exoma (contagem de alelos)..... 94
Tabela 10: Número de vezes que cada variante putativa do gene IGHJ foi identificada em genoma ou exoma (contagem de alelos)..... 97

LISTA DE ABREVIATURAS E SIGLAS

AD	*Allele Depth*
AID	*Activation-induced Cytidine Deaminase*
AIRR	*Adaptive Immune Receptor Repertoire*
AIRR-seq	*Adaptative Immune Receptor Repertoire Sequencing*
BCR	*B Cell Receptors*
BQSR	*Recalibração Base Quality Score*
CCS	*Circular Consensus Sequencing*
CDRs	*Complementarity-Determining Regions*
CNV	*copy number variations*
CSR	*Class-switch Recombination*
CSS	*Cascading Style Sheets*
CSV	*Comma-separated values*
dbSNP	*Database for Short Genetic Variations*
DDL	Linguagem de Definição de Dados
DML	Linguagem de Manipulação de Dados
DP	*Depth*
DTR	Diagrama de Tabelas Relacionais
EMBL	*European Molecular Biology Laboratory*
Fab	*Fragment antigen-binding*
FWR	*Frameworks Regions*
G1K	Projeto 1000 genomas
GATK	*Genome Analysis Toolkit*
GFF	*General Feature Format*
GLDB	*Germline Database*
gnomAD	*Genome Aggregation Database*
GQ	*Genotype Quality*
HiFi	*high-fidelity*
HTML	*HyperText Markup Language*
HGNC	*Nomenclature Committee*
IDE	*Integrated Development Environment* - Ambiente de Desenvolvimento Integrado

Ig	Imunoglobulina
IGH	*Immunoglobulin heavy locus*
IGHV	*Immunoglobulin heavy variable gene*
IGHD	*Immunoglobulin heavy diversit gene*
IGHJ	*Immunoglobulin heavy joining gene*
IGHC	*Immunoglobulin heavy constant gene*
IMGT	*International ImMunoGeneTics information system*
JSON	*JavaScript Object Notation*
NCBI	*National Center for Biotechnology Information*
NGS	*Next Generation Sequencing*
NGS	Sequenciamento de Nova Geração
OGRDB	*Open Germline Receptor Database*
ORF	*Open Reading Frame*
PacBio	*Pacific Biosciences, Menlo Park, CA, United States*
PCR	*Polymerase chain reaction*
PHP	*Hypertext Preprocessor*
QD	*Quality by Depth*
RAG	Enzima de recombinação (*Recombination-Activating Gene*)
Rep-seq	R*epertoire Sequencing*
RF	*Random Forest*
scFv	*Single-chain variable Fragment*
SGBD	Sistema de Gerenciamento de Banco de Dados
SHM	Hipermutação somática
SMRT	*Single-molecule Real-time*
SNP	*Single Nucleotide Polymorphisms*
SQL	*Structured Query Language* – Linguagem Estruturada de Consulta
SV	*Structural Variants*
TCGA	*The Cancer Genome Atlas*
TIgGER	*Tool for Immunoglobulin Genotype Elucidation* via Rep-Seq
UCSC	*University of California, Santa Cruz*
URL	*Uniform Resource Locator*
VEP	*Variant Effect Predictor*
VQSLOD	*Variant quality Score Log-odds ration*

VQSR	*Variant Quality Score Recalibration*
WEB	*World Wide Web*
WES	Sequenciamento do Exoma Completo
WGS	Sequenciamento do Genoma Completo
YVr-DB	*Antibody Variants Database*

Sumário

1 **INTRODUÇÃO** 15

1.1 Resposta imune adaptativa 15

1.2 Estrutura do anticorpo 15

1.3 Lócus da cadeia pesada de imunoglobulina 17

1.4 Sequenciamento de Nova Geração (NGS) aplicado à imunologia 20

1.4.1 Chamada de variantes em dados de NGS 22

1.4.2 Descoberta de características de repertório de imunoglobulina usando Rep-seq 25

1.4.3 Limitações do uso de *short reads* para imunoglobulina 26

1.5 Banco de dados de imunoglobulinas e alelos 28

1.5.1 IMGT/GENE-DB banco de dados de imunoglobulinas e receptores de célula T 28

1.5.2 Banco de dados de alelos de imunoglobulinas 29

1.6 Estratégias para descoberta de novos alelos de genes de imunoglobulina 30

1.6.1 Descoberta de alelos de imunoglobulina em dados de Rep-seq *short reads* 30

1.6.2 Descoberta de alelos de imunoglobulina em dados de Genoma *short reads* 36

1.6.3 Descoberta de alelos de imunoglobulina em dados de Genoma *long reads* 39

1.6.4 Banco de dados Genome Aggregation Database (gnomAD) 40

1.7 Justificativa 42

2 **OBJETIVOS** 44

2.1 Objetivo Geral 44

2.2 Objetivos específicos 44

3 **MATERIAL E MÉTODOS** 45

3.1 Obtenção de segmentos gênicos IGHV 45

3.2 Comparação das posições das variantes nos genes IGHV no GENCODE *versus* NCBI 45

3.3 Extração de variantes de genes IGHV do GNOMAD 48

3.4 Selecionar somente as variantes presentes na V-REGION dos genes IGH IMGT 49

3.5 Posicionar as variantes do gnomAD nas sequências V-REGION 50
3.6 Verificação de alteração do gene referência a partir da análise de SNP ou indel 50
3.7 Critérios de veto de variantes 50
3.8 Pesquisa de variantes associadas a doenças 51
3.9 Verificar a presença das variantes identificadas em outros bancos de dados 52
3.10 Método utilizado para criar o banco de dados YVr-DB 54
3.11 Método para criar a plataforma web YVr-DB versão 1.0 55
3.12 Consulta e formatação dos dados para gerar gráficos ou tabelas 56
3.13 Método estatístico 57
3.14 Desenvolvimento de uma nova metodologia para filtrar genes VDJ 58
4 **RESULTADOS** 61
4.1 Resultado da comparação das posições dos genes IGHV do NCBI *versus* GENCODE 61
4.2 Plataforma web YVr-DB versão 1.0 64
4.2.1 Funcionalidades disponíveis na plataforma web YVr-DB versão 1.0 64
4.2.2 Construção da base de dados da plataforma YVr-DB versão 1.0 68
4.3 Mineração dos genes IGHV do exoma humano em larga escala revelou 10.550 variantes putativas 69
4.4 A maioria das variantes do IGHV estão no *Framework* 3 75
4.5 A maioria das variantes IGHV são *missense* 77
4.6 A maioria das variantes são de população específicas 81
4.7 Variantes IGHV identificadas no Catálogo GWAS 84
4.8 Avaliação da distribuição das variantes de acordo com as métricas VQSLOD e QD 86
4.9 Resultado da busca de variantes nos segmentos gênicos IGHD 94
4.10 Resultado da busca de variantes nos segmentos gênicos IGHJ 97
5 **DISCUSSÃO** 99
6 **CONCLUSÕES** 102

7 **PERSPECTIVAS** 104

REFERÊNCIAS 105

ANEXO I 112

ANEXO II 148

1 INTRODUÇÃO

1.1 Resposta imune adaptativa

A resposta imune adaptativa de humanos e outros vertebrados é formada principalmente pelas células B e células T que podem expressar uma grande diversidade de receptores proteicos (GIUDICELLI et al., 2017; MARKS; DEANE, 2020). Neste trabalho vamos focar no estudo dos receptores de células B.

Grande parte da diversidade destes receptores é formada pelo processo de recombinação de segmentos gênicos V(D)J. O repertório de receptores de antígenos de um indivíduo é estimado em cerca de 2 x 10^{12} diferentes sequências distintas de BCR (*B Cell Receptors*) (LEFRANC; LEFRANC, 2020).

O entendimento da diversidade do repertório de imunoglobulinas é importante na elucidação de informações relativas às características do sistema imune em diversas condições fisiológicas, como no envelhecimento, e também em infecções virais, doenças crônicas e autoimunes, e vacinação (GADALA-MARIA et al., 2015; MARKS; DEANE, 2020; YU; CEREDIG; SEOIGHE, 2017).

1.2 Estrutura do anticorpo

A forma secretada dos BCRs é denominada “anticorpo” ou “imunoglobulina”, cujo papel é essencial na resposta imune adaptativa. Os mamíferos são capazes de gerar um grande número de anticorpos com diferentes especificidades, que reconhecem (ligam) diferentes antígenos com alta afinidade e assim induzem a neutralização do patógeno (MARKS; DEANE, 2020).

O anticorpo humano é formado por quatro cadeias, sendo duas pesadas (*heavy*, H) cujo lócus está presente no cromossomo humano 14, e duas cadeias leves (*light*, L), cujos loci estão presentes nos cromossomos 2 (lócus IGK *kappa*) e 22 (lócus IGL *lambda*). As cadeias possuem uma região constante (C) e uma região variável (V). A região constante é responsável pelas funções efetoras, ou seja, envolvida no mecanismo de eliminação dos patógenos. Já a região (V) é responsável pelo reconhecimento específico do antígeno e recebe este nome devido a sua grande variabilidade entre as moléculas de anticorpo (LEFRANC; LEFRANC, 2020; MURPHY, 2014). A Figura 1 apresenta a estrutura de um anticorpo.

As regiões variáveis dos anticorpos são compostas por três *loops* hipervariáveis, também chamado de *Complementarity-Determining Regions* (regiões de determinação de complementariedade) CDR1, CDR2 e CDR3. Essas regiões são intercaladas por quatro *frameworks regions* (FWR1, FWR2, FWR3, FWR4) (LEES; SHEPHERD, 2017).

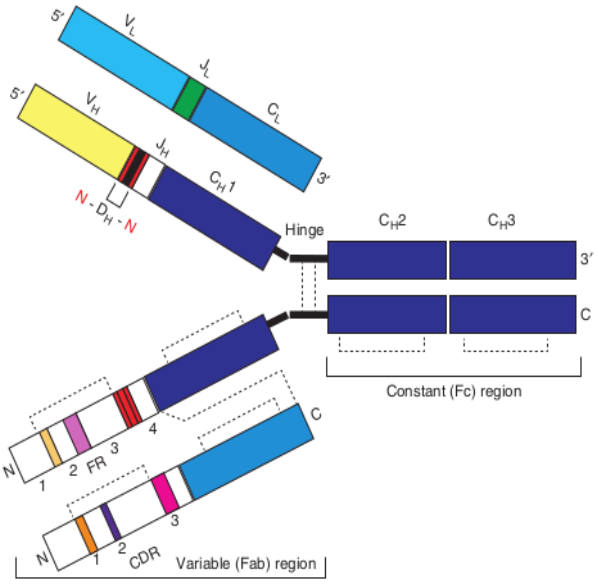


Figura 1: Esquema da estrutura de um anticorpo humano. A figura apresenta as cadeias leves (L) e pesadas (H), os segmentos gênicos V, (D), J, que fazem parte da região variável, juntamente com a região constante C. As regiões de adição de N-nucleotídeos (nucleotídeos não codificados) são indicadas em vermelho. Na porção superior da imagem, estão representados os segmentos gênicos que constituem as cadeias do anticorpo, e na porção inferior da imagem, os *motifs* estruturais que resultam destes segmentos (FWR e CDRs). As linhas tracejadas indicam as pontes dissulfeto, e tanto a cadeia pesada como a leve apresentam as regiões de FWRs e CDRs (GEORGIOU et al., 2014).

A região variável da cadeia pesada de imunoglobulina é composta por três segmentos gênicos (também chamados "genes"): variável (V), diversidade (D) e de junção (J). Os *frameworks* FWR1, FWR2, FWR3, juntamente com as regiões determinantes de complementariedade CDR1, CDR2 e uma parte do CDR3 são originados do segmento gênico V. As partes remanescentes do CDR3 advêm do segmento gênico D e de uma parte do J. Por fim, o *framework* 4 (FWR4) é formado pela parte restante do segmento gênico J (GUPTA; VISWANATHA; PATEL, 2020; LEES; SHEPHERD, 2017).

A região variável da cadeia pesada é formada pelos genes V, D e J. Já essas regiões nas cadeias leves de imunoglobulina *kappa* ou *lambda* são produzidas com apenas os segmentos V e J. Essas duas cadeias (leve e pesada) formam um receptor completo de célula B (GUPTA; VISWANATHA; PATEL, 2020; LEES; SHEPHERD, 2017).

1.3 Lócus da cadeia pesada de imunoglobulina

Classicamente, são unidos um segmento gênico V, um D, um J e um C para a montagem da cadeia pesada do BCR. De acordo com o sistema de informação *Internacional ImMunoGeneTics information system* (IMGT), o lócus da cadeia pesada da imunoglobulina (*Immunoglobulin heavy locus*) IGH de humanos se encontra na banda 14q32.33 dentro da região telomérica do cromossomo 14, e na sua forma germinativa (isto é, na sua forma não recombinada), compreende principalmente 38-46 genes IGHV (*Immunoglobulin heavy variable gene*) funcionais, 23 IGHD (*Immunoglobulin heavy diversit gene*), 6 IGHJ (*Immunoglobulin heavy joining gene*) e 9 genes IGHC (*Immunoglobulin heavy constant gene*) (LEFRANC, 2001), como apresentado na Figura 2. Deste modo, o fato de selecionar ao acaso um segmento gênico de cada tipo já contribui para a grande diversidade na região variável do anticorpo (MURPHY, 2014).

No lócus, também se encontram diversos segmentos gênicos que não são funcionais (pseudogenes). Estes segmentos apresentam principalmente um *codon* de parada (*stop codon*) e/ou mutação *frameshift* em sua sequência o que inviabiliza a produção de um receptor funcional (ver Figura 2).

O lócus IGH de humanos possui grupos espacialmente separados de segmentos gênicos VH, DH, JH e CH, dispostos em *tandem*. Os segmentos VH de humanos são agrupados em 7 famílias/subgrupos, cujos membros compartilham ao menos 80% de identidade na sequência de DNA. Notavelmente, o lócus de imunoglobulina apresenta uma grande densidade de sequências repetitivas, incluindo duplicações de muitos segmentos gênicos (MURPHY, 2014).

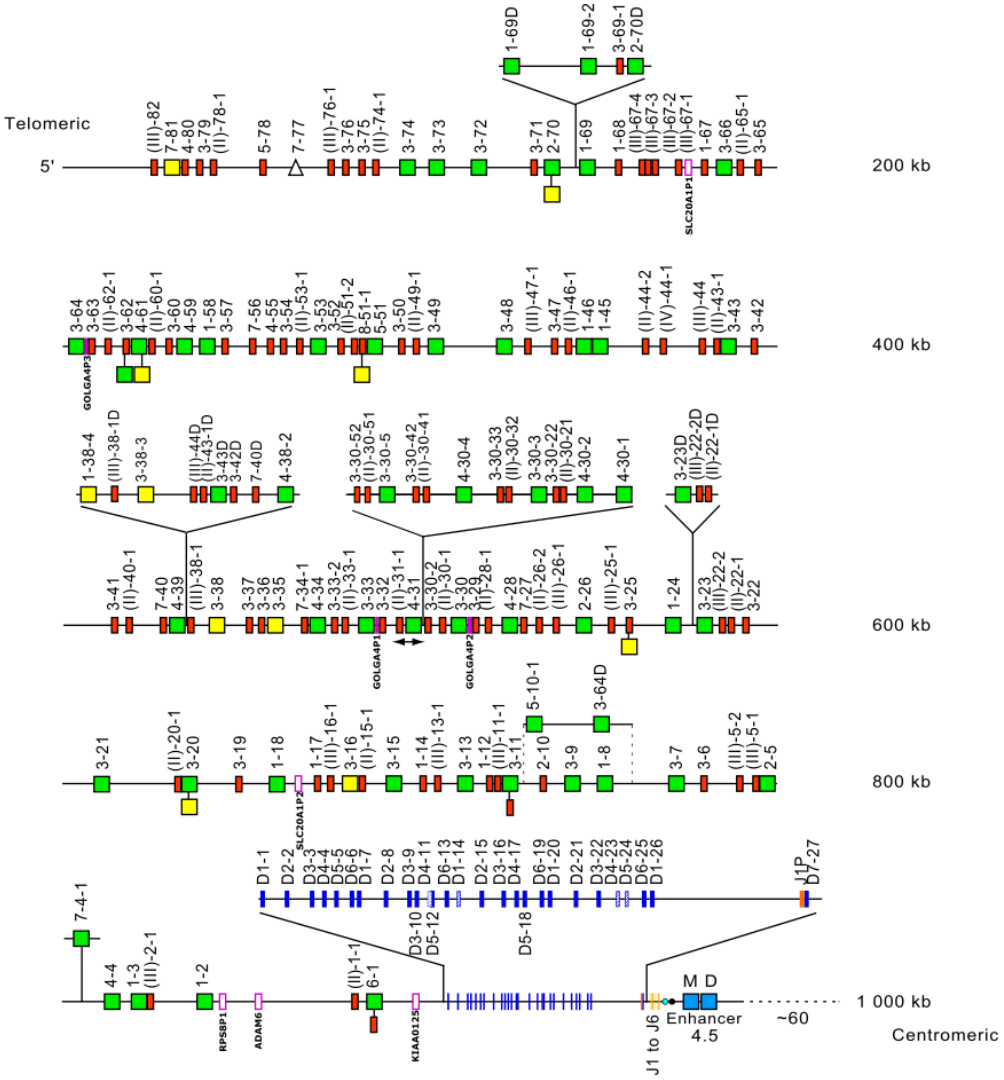


Figura 2: Representação do lócus de IGH de humano na banda 14q32.33. As caixas representam os genes. Para os genes VH, caixas verdes representam genes funcionais, as amarelas representam *Open Reading Frame* (ORF) e as vermelhas representam pseudogene. Segmentos gênicos que não foram associados a uma família específica são identificados por um algarismo romano entre parênteses. As caixas azuis se referem aos genes D funcionais, e as caixas azul tracejado às ORFs. Os genes J estão em amarelo, e localizados após os genes D (J1 até J6); Caixas com contorno de rosa são pseudogenes não relacionados ao lócus (LEFRANC, 2001).

Existem duas formas principais de geração de diversidade do repertório de anticorpos. A diversidade primária, que ocorre na medula óssea antes da exposição do anticorpo ao antígeno, ocorre através de mecanismos de recombinação (isto é, rearranjo e junção) dos segmentos gênicos V(D)J, como apresentado anteriormente. Já a diversidade secundária acontece pela hipermutação somática (SHM) e recombinação de troca de isotipo (GEORGIOU et al., 2014).

Durante o processo de formação de diversidade primária, ocorre a recombinação V(D)J, que é mediada por vários eventos de reorganização e quebra de DNA (Figura 3). As enzimas RAG1 e RAG2 (*Recombination-Activating Genes* 1 e 2), reconhecem as regiões RSS (*Recombination Signal Sequences*) nas extremidades dos genes V, D e J, se ligam aos RSS e realizam a clivagem do DNA. Posteriormente a enzima artemis juntamente com um complexo de enzimas que incluem a TdT fazem o reparo do DNA (junção), realizando a adição de nucleotídeos. (SCHATZ; SWANSON, 2011).

Os segmentos VH, DH e JH presentes na sequência recombinada da cadeia pesada são rearranjados durante o desenvolvimento inicial e sofrem deleção e inserção de nucleotídeos nas regiões de junção, ou seja, na junção D-J e V-D. São adicionados vários nucleotídeos (N) *non-templated* e nucleotídeos palíndromos (P), o que aumenta a diversidade. Dessa forma, são geradas as sequências específicas da região variável da cadeia pesada para cada linfócito B. Caso ocorra a geração de um rearranjo improdutivo de IGH, a célula B sofrerá apoptose (GUPTA; VISWANATHA; PATEL, 2020; SCHEIJEN et al., 2019) (Figura 3).

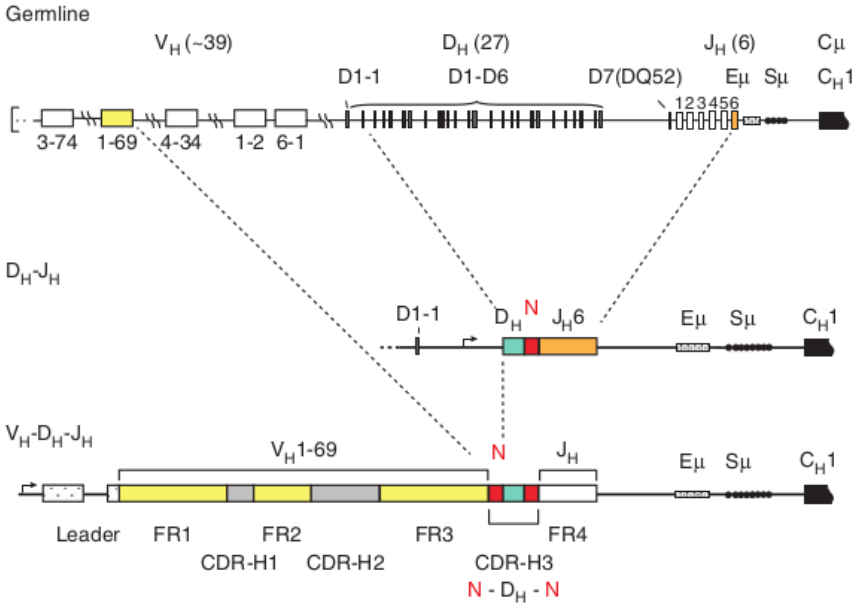


Figura 3: Etapas da recombinação VDJ de cadeia pesada de anticorpos. O repertório primário de anticorpo é criado pelo rearranjo dos segmentos, e pela junção dos segmentos através da adição de N-nucleotídeos (nucleotídeos não codificados), o que gera variabilidade na região variável. A região de reconhecimento dos antígenos (cadeia pesada) é formada pelas regiões determinantes de complementariedade (CDRs) (GEORGIOU et al., 2014).

No processo de formação da diversidade secundária, após a ativação da célula B pelo antígeno nos órgãos linfóides secundários como baço e linfonodo, ocorre a SHM (isto é, a introdução de mutações ao acaso na região variável das cadeias), seguida da proliferação (expansão clonal) das células cuja afinidade foi aumentada. A SHM é catalisada por enzimas como a *Activation-induced Cytidine Deaminase* (AID). Assim, a afinidade das células B por seus antígenos é aumentada na formação da resposta imune adaptativa (maturação de afinidade) (GEORGIOU et al., 2014; GUPTA; VISWANATHA; PATEL, 2020; RECALDIN; FEAR, 2016).

O evento de *Class-switch Recombination* (CSR) ou recombinação por troca de classe altera o tipo de região constante da cadeia pesada do anticorpo. A alteração é feita quando o gene CH que compõe a região constante (C) da cadeia pesada da imunoglobulina é substituído por outro gene CH. Desse modo, a célula B muda o subtipo de imunoglobulina expresso (MATTHEWS AJ, ZHENG S, DIMENNA LJ, 2014; RECALDIN; FEAR, 2016).

A compreensão dos processos de formação de diversidade de anticorpos pode auxiliar na identificação ou no prognóstico de diversas doenças autoimunes. Um dos exemplos é o caso da leucemia linfocítica crônica, um tipo muito comum de leucemia de células B maduras. Foi observado que uma menor taxa de SHM na população clonal dominante de células B está associada à agressividade da doença (GUPTA; VISWANATHA; PATEL, 2020).

1.4 Sequenciamento de Nova Geração (NGS) aplicado à imunologia

Para estudos de diversidade de alelos de segmentos gênicos de imunoglobulina e também dos repertórios de anticorpos, faz-se necessário o uso de técnicas que permitam a obtenção dessas sequências com alta confiabilidade e cujo rendimento ou profundidade seja proporcional à diversidade esperada, principalmente para identificação de sequências raras. Um estudo que compara o sequenciamento Sanger e NGS, destacou algumas vantagens do NGS como: processamento em paralelo (maior eficiência), baixo custo comparado ao Sanger, determinação direta de clones, e possibilidade de identificar na maioria das vezes, mais de um rearranjo clonal dominante de IGH (GUPTA; VISWANATHA; PATEL, 2020).

O termo NGS designa plataformas de sequenciamento com alto rendimento (isto é, capazes de sequenciar uma grande quantidade de moléculas diferentes de DNA), alta cobertura e alta profundidade. Estas plataformas apresentam diferenças quanto a quantidade

de bases lidas por sequência: um primeiro grupo, capaz de ler sequências curtas (chamadas de *short reads*), inclui as plataformas Illumina *MiSeq* (300 pb *paired-end*), *HiSeq* (250 pb *paired-end*) e *Roche 454 GS FLX* (até 1kb); o segundo grupo, capaz de ler sequências maiores (chamadas de *long reads*), contém as plataformas *PacBio* (250 bp – 40 kb) e *MinIOn* (*Oxford Nanopore*). Cada uma dessas plataformas apresenta distintas taxas de cobertura de *reads* (ROUET et al., 2018).

As técnicas de NGS simplificaram e reduziram drasticamente os custos de sequenciamento do DNA. O exemplo mais notável é o sequenciamento do genoma humano: o primeiro genoma humano levou 12 anos para ser montado (com o uso da técnica de sequenciamento Sanger) e custou aproximadamente 3 bilhões de dólares; cerca de 15 anos depois, com o uso de técnicas de NGS, este mesmo sequenciamento pode ser feito de um a três dias, e custa em torno de 1.000 dólares (KUMAR; COWLEY; DAVIS, 2019; ROUET et al., 2018).

Além disso, as técnicas de NGS permitiram a introdução de novas abordagens, como o sequenciamento de *amplicons*, cujo objetivo é obter as sequências de regiões específicas do genoma, de um único gene ou de múltiplos genes. Para esta tarefa, desenha-se iniciadores (*primers*) específicos que amplifiquem apenas a região de interesse (KUMAR; COWLEY; DAVIS, 2019).

Além disso, existem painéis clínicos (painéis de iniciadores) para estudo de variações genéticas associadas à fenótipos específicos. Um exemplo de uso destes painéis é o tipo de Sequenciamento do Exoma Completo (WES). Esta abordagem, com base nos iniciadores do painel, acessa de forma direcionada às regiões dos genes que codificam proteínas (isto é, os *éxons*), que correspondem a 2% do genoma. A importância desta técnica deriva do fato de que grande parte das variantes causadoras de doenças é encontrada nos *éxons* (KUMAR; COWLEY; DAVIS, 2019).

Outro tipo de sequenciamento do NGS é o Sequenciamento do Genoma Completo (WGS), onde tanto as regiões do genoma que são codificantes de proteínas (*éxon*) quanto as que não codificam proteínas (*íntron*) são sequenciadas. Por meio desta abordagem é possível investigar as regiões não codificantes e como é realizado o sequenciamento geral (*éxon/íntron*), também é possível analisar os *éxons* (WES) por meio de ferramentas de bioinformática (*in silico*) (KUMAR; COWLEY; DAVIS, 2019).

É importante destacar que o aumento de sequenciamentos utilizando estas tecnologias impulsionou o desenvolvimento de ferramentas mais robustas de bioinformática para analisar

grandes quantidades de dados de sequenciamento, fato que influenciou diretamente na acurácia dos bancos de dados de referência (SCHEIJEN et al., 2019; YU; CEREDIG; SEOIGHE, 2017), uma vez que, diversas ferramentas específicas para identificação de linhagem germinativa V(D)J e detecção de rearranjos gênicos se baseiam na qualidade e quantidade de *reads* geradas no sequenciamento (SCHEIJEN et al., 2019).

Porém, o tamanho dos fragmentos dos anticorpos é um desafio para as plataformas baseadas em *short reads* e que dependem de processo de montagem de dados, uma vez que o domínio variável tem de 300 a 400 pb. Já o scFv (*Single-chain variable Fragment*) composto pelo conjunto VH-linker-VL e o Fab (*Fragment antigen-binding*) composto pelo VH-VL e CH1-CL que são geralmente sequenciados, os tamanho dos fragmentos variam de 700 a 800 pb, podendo chegar a 1.500 pb. Nota-se que, mesmo fazendo *paired-end* com *short reads*, não é possível produzir uma sequência que cubra o tamanho total destes segmentos. Para este caso, ou quando se queira cobrir toda a extensão do lócus de imunoglobulina é recomendado o sequenciamento com *reads* longas, que são produzidas pela plataforma MinIOn ou *Pacific Biosciences, Menlo Park, CA, United States* (PacBio) (ROUET et al., 2018).

É importante observar que, no início, a tecnologia PacBio gerava *long reads*, sem necessidade de fragmentação de alvo e montagem de sequências, com uma taxa elevada de erros (FORD et al., 2020). Posteriormente, com a otimização da técnica de sequenciamento de consenso circular (*Circular Consensus Sequencing* - CCS), a precisão do *Single-molecule Real-time* (SMRT), foi elevada para 99,8%, ou seja, *high-fidelity* (HiFi). Portanto, são geradas *reads* com alta fidelidade e com média de tamanho de 13,5 kb (WENGER et al., 2019).

No entanto, o custo do sequenciamento por base ainda é elevado e a quantidade de dados sequenciados especificamente do lócus codificadores de imunoglobulinas com as plataformas de *long reads* ainda é restrito e limitado (RODRIGUEZ et al., 2020).

1.4.1 Chamada de variantes em dados de NGS

Apesar do avanço na plataforma de sequenciamento de *long reads* é importante destacar que os sequenciamentos de *short reads* como WES e WGS deram e ainda dão uma grande contribuição para identificação de variantes associadas a doenças e estudos

populacionais, como no caso do Projeto 1000 genomas - (G1K) (1000 GENOMES PROJECT et al., 2015).

Identificar variantes em dados de WES e WGS, em teoria, seria realizado apenas alinhando as sequências na referência e verificando os *mismatches*. Contudo, na prática diversos problemas podem surgir como: viés de amplificação, erro de sequenciamento, error de montagem de *contigs* baseados em *reads*, erro no *software* e mapeamento de artefatos principalmente associado ao reduzido tamanho dos lócus e elevado padrão de polimorfismo (VAN DER AUWERA et al., 2014).

Para distinguir variantes reais de artefatos e erros, foram criados protocolos que disponibilizam diversas métricas. Como exemplo de um protocolo básico, seria a execução dos cinco passos a seguir:

1) mapear as *reads* na referência (BWA);

2) Formatar os dados para entrar no programa *Genome Analysis Toolkit* (GATK) (MCKENNA et al., 2010) (Picard);

3) realinhar as inserções e deleções (indel);

4) Fazer a *Recalibração Base Quality Score* (BQSR) no GATK (MCKENNA et al., 2010);

5) Compactar os dados GATK (MCKENNA et al., 2010).

Diversas alternativas podem ser consideradas e o protocolo para chamada de variantes e avaliação da qualidade das mesmas pode ser alterado (VAN DER AUWERA et al., 2014). Foram selecionadas duas métricas importantes para o contexto desta tese: *Variant quality Score Log-odds ration* (VQSLOD) e *Quality by Depth* (QD).

Os comandos do GATK (MCKENNA et al., 2010), *HaplotypeCaller* e *UnifiedGenotyper* são utilizados para realizar a chamada de variante e possuem a característica de atingir um alto grau de sensibilidade. Por um lado, é bom recuperar uma grande quantidade de variantes reais, todavia é necessário dispor de uma estratégia para restringir variantes falso positivas (VAN DER AUWERA et al., 2014).

Uma boa estratégia é o uso da métrica *Variant Quality Score Recalibration* (VQSR), que utiliza um algoritmo de aprendizagem de máquina aplicado em duas etapas. A primeira usa todo o conjunto de dados para atribuir uma probabilidade a cada variante. Já a segunda etapa, aplica um filtro no conjunto para gerar um subconjunto com a qualidade desejada (VAN DER AUWERA et al., 2014).

É importante mencionar que, para classificar as variantes como sendo verdadeiras, é utilizado dados de variantes de alta qualidade obtidos de vários bancos de dados, como omni, 1000 Genomas e *Hapmap*, *Database for Short Genetic Variations* (dbSNP), entre outros. O algoritmo analisa as variantes verdadeiras e aprende como caracterizá-las, sendo este procedimento também realizado para variantes consideradas falsas (VAN DER AUWERA et al., 2014).

O aprendizado de variantes verdadeiras e falsas é o conjunto de treinamento utilizado na etapa citada anteriormente. Como temos a probabilidade da variante ser verdadeira ao invés de um artefato de sequenciamento ou erro, é possível atribuir, para cada variante, a métrica de qualidade VQSLOD, ou seja, a chance de ser verdadeira dividido pela chance de ser falsa (VAN DER AUWERA et al., 2014).

Já a métrica QD é obtida pelo processo de divisão da métrica QUAL que diz sobre a confiança da variante ser verdadeira, pela profundidade da variante *Allele Depth* (AD), ou seja, por quantas *reads* a variante é suportada. Ao realizar esta normalização é possível ter mais clareza da real qualidade da variante. Além disso, para os arquivos VCFs com dados de muitas amostras, o cálculo do QD utiliza os dados QUAL/QD de alelos não homozigotos com a referência (0/0: homozigoto referência; 0/1: heterozigoto; 1/1: homozigoto alelo alternativo).

Ainda sobre a qualidade na chamada de variantes, o trabalho apresentado por Wenger e colaboradores (WENGER et al., 2019) utilizou um conjunto de dados de variantes de alta qualidade do *Genome in a Bottle* (GIAB) *benchmark* (ZOOK et al., 2019), para verificar a acurácia na chamada de variantes em dados de *long reads* CCS, comparado com dados gerados pelo sequenciamento na plataforma Ilumina (NovaSeq) *short reads*. Foram utilizados SNVs e pequenos indels (<50 pb) na comparação.

As variantes do cromossomo 20 foram analisadas, e utilizando-se as ferramentas *DeepVariant* (POPLIN et al., 2018) e GATK (MCKENNA et al., 2010), foram obtidos os seguintes resultados de acurácia:

a) PacBio (CCS) - *DeepVariant* (CCS *model*) 99.914%, ao recalibrar 99.959%;

b) Illumina (NovaSeq) - *DeepVariant* (Illumina *model*) 99.925%, ao recalibrar 99.940%;

c) PacBio (CCS) - GATK *HaplotypeCaller* (*hard filter*) 99.408%, ao recalibrar 99.531%;

d) Illumina (NovaSeq) - GATK *HaplotypeCaller* (*no filter*) 99.824%, ao recalibrar 99.920%.

Os dados apresentados acima apontam que a chamada de variantes provenientes de Illumina (NovaSeq) apresenta uma acurácia maior em relação ao CCS de PacBio. Porém, como são utilizados modelos de treinamento para direcionar o programa, esses modelos podem introduzir algum viés para *short reads* (WENGER et al., 2019).

Outro trabalho realizou a comparação da abordagem *long reads* com o *framework* IGenotyper contra o G1K, para recuperar SNVs. O uso do IGenotyper melhorou a acurácia da chamada de variantes em mais de 35% em SNPs do G1K verdadeiros positivos e em torno de 97% para os dados de falsos positivos (RODRIGUEZ et al., 2020).

1.4.2 Descoberta de características de repertório de imunoglobulina usando Rep-seq

A tecnologia NGS *short reads* foi inicialmente concebida para ser utilizada em genômica, com o sequenciamento WGS. Porém, devido à capacidade de sequenciar grandes quantidades de dados com boa resolução (profundidade de *reads*) foi bastante utilizada em imunologia, no sequenciamento em larga escala do repertório imune R*epertoire Sequencing* (Rep-seq) (ROUET et al., 2018). Diversos trabalhos adotaram esta técnica para investigar diversos aspectos dos repertórios em diversas condições, como será descrito abaixo.

Um ponto importante a ser considerado no estudo de repertório é que a diversidade teórica V(D)J é em torno de 10^{13} e que ocorre após os processos moleculares de recombinação (CALIS; ROSENBERG, 2014); todavia, apenas 2% dos linfócitos são acessíveis na circulação em um determinado momento (TREPEL, 1974; WESTERMANN; PABST, 1992). A alta diversidade e o acesso a uma quantidade restrita de dados limita a análise de repertórios. No entanto, com as plataformas de sequenciamento de *high-throughput*, fornecidas pelo NGS e com as abordagens estatísticas para análise, por exemplo, do CDR3, possibilitaram estudar mais detalhadamente o repertório de anticorpos (Rep-seq) produzido por um indivíduo, em um determinado momento (CHAUDHARY; WESEMANN, 2018).

Outro estudo de análise de Rep-seq investigou o comportamento de um desafio com antígeno em relação à restrição ao sono, evidenciando que a resposta ao antígeno pela célula B, ou seja, a produção de anticorpos é prejudicada (atrasada). Esta observação não é tão relevante na situação de infecção por patógenos, já que a apresentação do antígeno é contínua

e as células B e T que forem ativadas podem reverter a situação inicial. No entanto, o impacto deste atraso na resposta inicial pode agravar os sintomas e prolongar a doença (TUNE et al., 2021).

Os resultados apresentados também indicaram que a eficiência da vacinação é reduzida quando existe a privação de sono na noite anterior, em comparação a noite posterior. Logo, uma investigação mais detalhada em humanos e modelos animais com o intuito de desvendar a relevância clínica e como a restrição do sono diminui a resposta de célula B dependente de célula T seria necessária (TUNE et al., 2021).

A análise de Rep-seq foi também realizada em portadores de leucemia linfocítica crônica (CLL). O estudo, com duração de oito anos, contou com a inclusão de 138 pacientes, 81 homens e 57 mulheres, com média de idade de 63 anos (intervalo de 33 a 84). Os pacientes estavam em diferentes estágios da doença, dos quais 65 (47%) evoluíram para terapia e 24 (17%) vieram a óbito.

As amostras desses pacientes foram sequenciadas por meio de *amplicons* provenientes da região V(D)J. Os dados foram alinhados no banco de dados IMGT (GIUDICELLI; CHAUME; LEFRANC, 2005) e os pacientes foram separados em dois grupos: os que apresentavam mutações na região V(D)J, classificados com M-CLL (n=78) e os que não continham mutações U-CLL (n=60). Posteriormente, foi verificado as propriedades físico-químicas de cada aminoácido de cada HCDR3. O acompanhamento das alterações nestas regiões apontou que os U-CLL tiveram sobrevida reduzida e a situação clínica mais agressiva comparada a M-CLL. Portanto, a avaliação destas mutações é fundamental para prognósticos em leucemia linfocítica crônica. podendo ser utilizada para diagnóstico de várias outras doenças (RODRÍGUEZ-CABALLERO et al., 2021).

Diversas técnicas estão sendo utilizadas para determinar clones específicos em dados de Rep-seq por meio de diversos algoritmos de agrupamento de sequências. Isso pode ser empregado para identificar grandes quantidades de anticorpos monoclonais (PAROLA; NEUMEIER; REDDY, 2018).

1.4.3 Limitações do uso de *short reads* para imunoglobulina

Com o surgimento da tecnologia NGS, especificamente o sequenciamento de *amplicons* de regiões V(D)J utilizando *short reads*, o estudo de Rep-seq tem sido cada vez mais utilizado para diferentes propósitos. No entanto, para realizar as análises e estimar a

diversidade do repertório de anitcorpos é necessário alinhar as sequências V(D)J obtidas a um banco de dados de referência que contenha as sequencias da linhagem germinativa (*Germline Database* -GLDB) dos genes V(D)J como o IMGT (GIUDICELLI; CHAUME; LEFRANC, 2005) e determinar o CDR3, que por sua vez, viabiliza estimar a diversidade do repertório (YU; CEREDIG; SEOIGHE, 2017).

Entretanto, existem diversas limitações quanto ao uso de *short reads* para dados de Ig/TCR, visto que, a natureza repetitiva do *loci*, aliada ao reduzido tamanho e elevado polimorfismo, pode levar a um mapeamento incorreto das *reads* e associação de variantes para genes incorretos, bem como dificultar a detecção destes erros (WATSON et al., 2017).

O *loci* de IG possuem mais de 40 genes V, D (IGH, TRB) e J, com a fase de leitura aberta e que possuem uma alta similaridade de sequências, podendo chegar a 100% (gene duplicado) (GIUDICELLI; CHAUME; LEFRANC, 2005).

Como os bancos de dados contendo alelos da linhagem germinativa são incompletos, o entendimento se o alelo é oriundo de IG não foi totalmente elucidado. Isto é um problema principalmente para o lócus IGK, já que os genes V estão em dois blocos subsequentes (em *tandem*), nos quais já foram relatados eventos de conversão direta de genes (WATSON et al., 2015) . Conversão gênica é quando uma parte da sequência de um gene é substituída por outra parte de outro gene – sem recombinação.

Além das duplicações, alguns genes não estão presentes nas referências GRCh37 e GRCh38, o que impossibilita anotar alelos para eles (YU; CEREDIG; SEOIGHE, 2017). Existem 16 genes V classificados como funcionais/ORF e 220 kbp de sequências presentes em haplótipos humanos que não estão na referência GRCh37 (WATSON et al., 2013).

Devido à anotação incompleta de genes V na referência GRCh37, o mapeamento é comprometido, as *reads* são mapeadas para os genes que estão presentes, introduzindo novos CNVs e alelos que poderiam ser dos genes não presentes. Exemplo para SNP heterozigoto, mas o gene foi deletado, isso mascara o homozigoto e pode produzir heterozigoto artificial. Em algumas populações, o haplótipo alternativo (não presente na referência) aparece como principal e ao fazer o alinhamento com a referência não é encontrado, pois na referência está o principal (WATSON et al., 2013).

Além disso, um ponto relevante é a completude e acurácia do banco de dados referência IMGT (GIUDICELLI; CHAUME; LEFRANC, 2005) , uma vez que problemas nas sequências depositadas no banco podem impactar em todas as análises realizadas nos dados de Rep-seq.

As limitações apresentadas acima, não se aplicam apenas aos genes da linha germinativa de humanos, elas podem ser estendidas para outras espécies como modelos murinos e primatas não humanos (COLLINS et al., 2015; CORCORAN et al., 2016).

1.5 Banco de dados de imunoglobulinas e alelos

Com a geração de grandes quantidades de dados de sequenciamento das regiões de imunoglobulina por NGS, a acurácia e completude dos bancos de dados de referência, como o IMGT pode ser verificada (SCHEIJEN et al., 2019; YU; CEREDIG; SEOIGHE, 2017).

1.5.1 IMGT/GENE-DB banco de dados de imunoglobulinas e receptores de célula T

No banco de dados de referência de imunoglobulina (Ig) e receptores de célula T (TCR) IMGT/GENE-DB (GIUDICELLI; CHAUME; LEFRANC, 2005) estão armazenadas as informações de genes de imunoglobulinas de humanos, ratos e outros vertebrados. As nomenclaturas dos genes de imunoglobulinas utilizadas por este banco de dados são definidas em colaboração com *HUGO Nomenclature Committee* (HGNC) e podem ser acessadas pelo site https://www.genenames.org/.

O IMGT/GENE-DB, disponibiliza informações como polimorfismo alélicos, números de alelos, sequências de referência entre outros sobre o lócus de imunoglobulina de diversos organismos (GIUDICELLI; CHAUME; LEFRANC, 2005). Para realizar as anotações dos segmentos gênicos V, D e J, inúmeros métodos alinham as sequências (nucleotídeos e aminoácidos) das amostras dos usuários contra um banco de dados de referência, sendo o IMGT/GENE-DB o mais utilizado. Ainda, o *National Center for Biotechnology Information* (NCBI) utiliza o IMGT/GENE-DB como referência para anotação de imunoglobulinas Ig e TR (GADALA-MARIA et al., 2015).

Em julho de 2020, o IMGT/GENE-DB (http://www.imgt.org/) tinha catalogado os genes V de humanos (56 funcionais e 306 alelos); D (23 funcionais e 30 alelos) e J (6 funcionais e 13 alelos).

A completude e acurácia do banco de dados IMGT (GIUDICELLI; CHAUME; LEFRANC, 2005) impacta diretamente no contexto de Ig e TCR na definição da expressão de repertório, uma vez que as sequências para identificar os clones do repertório são alinhadas no GLDB. Portanto, manter a qualidade deste banco contribui diretamente para o

entendimento do repertório e consequentemente em pesquisa de saúde humana (XOCHELLI et al., 2015).

1.5.2 Banco de dados de alelos de imunoglobulinas

Diversos estudos têm sido realizados para identificar alelos de Ig e TCR que não atendem os critérios exigidos para serem depositados no IMGT. Estes alelos são catalogados, baseado no alinhamento de suas sequências contra o IMGT e bancos de dados são criados para armazenar estes alelos que alguns autores chamam de alelos putativos. Seguem alguns dos principais bancos de dados que armazenam alelos putativos.

O banco de dados *Open Germline Receptor Database* (OGRDB), disponível no site https://ogrdb.airr-community.org, foi criado para armazenar sequências de alelos da linhagem germinativa, juntamente com suas evidências. O objetivo é garantir a qualidade dos dados, juntamente com o acompanhamento do progresso de análise e desta forma deixar acessíveis os resultados para a comunidade científica. Em uma pesquisa no banco de dados em julho/2020, foi possível verificar a existência de 12 alelos de genes IGHV humanos (LEES et al., 2020).

O banco de dados IgPdb, disponível no endereço https://cgi.cse.unsw.edu.au/~ihmmune/IgPdb/information.php é um repertório de variantes alélicas putativas de genes de imunoglobulina humana. O banco de dados ficou inativo nos últimos anos, mas é uma boa fonte para acessar informações passadas. Em um acesso no site do IgPdb em julho/2020, foram identificados 228 putativos alelos do gene IGHV de humano (LEES et al., 2020). É importante observar que o IgPDB não contém a sequência completa (LPART1-LPART2-V-D-J-C) dos genes de Ig (KHATRI et al., 2021).

O banco de dados VBASE2 possui sequências germinativas de genes V de humanos e ratos que são relacionadas com as sequências de genes V no *Ensembl Genome Browser*, acessível no site https://www.ensembl.org/index.html. Para cada sequência do gene IGHV, todas as variações são obtidas do banco de dados de sequência e nucleotídeos do *European Molecular Biology Laboratory* (EMBL), até mesmo rearranjos V(D)J. O VBASE2 classifica alelos em três diferentes níveis de confiança, como exemplo os alelos IGHV de humanos presentes no banco de dados: Classe 1 com 59 alelos: evidencia genômica e rearranjo evidente; Classe2 com 204 alelos: evidência genômica apenas e Classe3 com 3 alelos: apenas evidência de rearranjo. Sendo no total 266 IGHV genes humanos (RETTER et al., 2005).

O banco de dados Lym1K foi criado a partir de dados obtidos do projeto 1000 genomas fase 3 (1000 GENOMES PROJECT et al., 2015), que sequenciou amostra de 2.504 humanos de 26 populações em quatro continentes: África (AFR), Leste da Ásia (EAS), Europa (EUR), Sul da Ásia (SAS) e nas Américas (AMR). Os alelos depositados no Lym1K foram montados com as sequências do banco G1K. Foram identificados 3.609 novos alelos putativos. A estratégia do Lym1K seria montar um banco germline, e foram identificados 3.609 novos alelos putativos cuja base de dados pode ser acessada em http://maths.nuigalway.ie/biocluster/database/ (YU; CEREDIG; SEOIGHE, 2017). Porém, em uma tentativa de acessar em 16 de abril de 2022, não foi possível, pois o sistema estava fora do ar.

O banco de dados pmIG, também foi criado a partir das sequências obtidas do projeto 1000 genoma fase 3 (1000 GENOMES PROJECT et al., 2015). Foram identificados 409 novos alelos putativos em IGHV, 179 em IGKV e 199 em IGLV. Os alelos são suportados por no mínimo 4 indivíduos, o banco de dados está disponível no GitHub pode ser acessado pelo *link* https://github.com/InduKhatri/pmIG (KHATRI et al., 2021).

1.6 Estratégias para descoberta de novos alelos de genes de imunoglobulina

Diversas abordagens têm sido utilizadas para descoberta de novos alelos de genes de imunoglobulina. A forma com que os dados de sequenciamento foram gerados contribui para esta atividade. Atualmente encontramos abordagens que buscam novos alelos em dados de WGS ou WES, ou em dados de Rep-seq, isto é quando são desenhados *primers* específicos para amplificar somente a região de interesse (*amplicons*), que podem ser por exemplo regiões codificadoras das imunoglobulinas.

1.6.1 Descoberta de alelos de imunoglobulina em dados de Rep-seq *short reads*

O grande desafio dos trabalhos que buscam identificar novos alelos em dados de Rep-seq é diferenciar mutações presentes na linhagem germinativa do indivíduo das geradas pelo processo de SHM. Diversas ferramentas computacionais foram desenvolvidas nos últimos anos na tentativa de identificar novos alelos em dados de Rep-Seq como mostra a Tabela 1.

Tabela 1: Trabalhos que identificam alelos a partir de dados de Rep-seq.

Referência	Anotação	Novos alelos V	Pipeline	Ferramenta	Estratégia Identificar Variante
GADALA-MARIA et al., 2015	IMGT/HighV-QUEST	11	R	TigGER	Freq. mutação Regressão Linear
CORCORAN et al., 2016	IgBLAST	2	Python	IgDiscover	Freq. mutação Distribuição Binominal Cluster UPGMA
WENDEL et al., 2017	IgBLAST	17	SeqPrep TigGER	-	TigGER
MIKOCZIOVA et al., 2020	IgBLAST	25	TigGER, IgDiscover Bayesiana	-	TigGER, IgDiscover Abordagem Bayesiana

O TIgGER (*Tool for Immunoglobulin Genotype Elucidation)* via Rep-seq baseia-se na análise de padrões de mutações em dados de Rep-seq para identificação de novos alelos dos segmentos V, determinação de genótipo de Ig específico do indivíduo e revisão de atribuição de alelos do *germline* V(D)J (GADALA-MARIA et al., 2015).

O princípio do TIgGER baseia-se na informação de que alelos que apresentam polimorfismo em uma posição, geram repertórios que possuem uma alta frequência de sequências que apresentam esse polimorfismo contado como uma mutação, independente da quantidade total de mutações que cada sequência apresente. Já mutações oriundas de SHM em uma determinada posição, aparecem em poucas sequências, e sua frequência depende da quantidade total de mutações que a sequência apresenta (GADALA-MARIA et al., 2015).

O fluxo de trabalho do TIgGER segue os seguintes passos: (1) anotação das sequências com a ferramenta IMGT/HighV-QUEST; (2) descartar as sequências onde o

número de mutações é maior que 10, separar grupos com o mesmo número total de mutações e calcular a porcentagem de mutação em na posição "i", submeter os dados para a regressão linear para separar mutações advindas de um alelo novo e de SHM; (3) alinhar novamente IMGT/HighV-QUEST, analisar o score e adicionar à lista (TIgGER) de possíveis alelos putativos do indivíduo; (4) montar o conjunto de alelos específicos do indivíduo; (5) alinhar o repertório com os alelos específicos do indivíduo. No trabalho apresentado pelo autor o TIgGER identificou 11 novos putativos alelos.

Outra estratégia para identificação de novos alelos é o IgDiscover que trabalha com a distribuição de Poisson e para alguns casos utiliza o algoritmo de agrupamento hierárquico (cluster) de sequências UPGMA como descrito abaixo.

Os dados são anotados com o IgBLAST, um script externo (expressão regular) detecta o CDR3 das sequências no nível de aminoácido. A qualidade das sequências é verificada com os seguintes critérios: não possuir *stop-codon*; *E-value* no máximo 10^{-3}; as regiões comparadas com a referência devem cobrir 90% do segmento gênico VH e 60% do JH para reduzir o número de acertos espúrios. Assim, o banco de dados inicial é gerado (CORCORAN et al., 2016).

O IgDiscover faz interações no banco de dados para descoberta de novos alelos (no trabalho foi utilizado n=3). No início de cada iteração o IgBLAST é utilizado para identificar e classificar os genes V e J. Uma observação fundamental é que novos alelos da linhagem germinativa são atribuídos para sequências similares do banco de dados (CORCORAN et al., 2016).

A identificação de novos VH alelos no IgDiscover é feita pela análise do padrão da distribuição de Poisson que consideram no eixo X percentual de diferença das sequências pela frequência em que as sequências aparecem, Figura 4A. Quando o alelo da referência é idêntico ao alelo expresso no indivíduo as diferenças percentuais na distribuição de Poisson se encontram à esquerda do gráfico da função.

Os novos alelos são identificados de três formas: (1) se a sequência mais próxima presente no banco de dados não foi expressa no indivíduo, a distribuição Poisson dos percentuais de diferença é deslocada para direita, Figura 4A; (2) se o alelo do banco de dados e o candidato a novo alelo é atribuído ao mesmo gene ocorre a distribuição binomial combinada (dois grupos na binomial). O grupo da esquerda corresponde ao alelo existente no *germline*, o grupo da direita ao novo alelo candidato, Figura 4B; (3) quando vários alelos candidatos são associados a vários alelos do *germline* ocorre a sobreposição de picos no

histograma não sendo possível resolver como nas etapas anteriores. Então se utiliza um algoritmo de agrupamento hierárquico (*cluster*) de sequências UPGMA. Os subgrupos de sequências similares gerados pelo algoritmo são destacados ao longo da diagonal de uma matriz (quadrado claro). Como apresentado na Figura 4C.

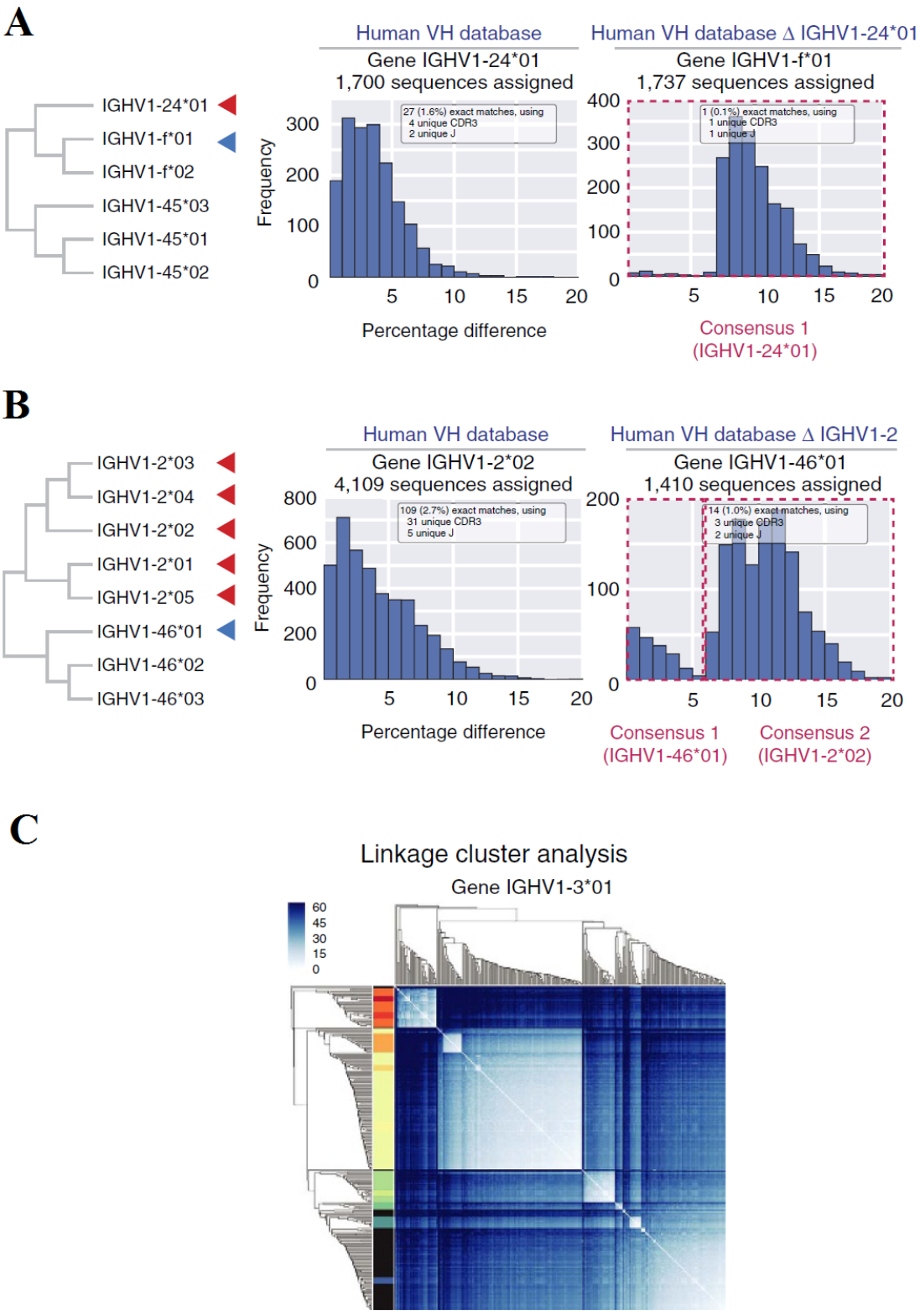


Figura 4: Identificação de novos VH alelos no IgDiscover. A) Representação da distribuição de Poisson, quando alelo igual ao do IMGT e quando é diferente, no eixo X o percentual de diferença, no eixo Y a frequência

de alelos. B) Representação de uma distribuição binomial, o gráfico apresenta duas distribuições, na figura estão separada por uma linha tracejada. C) Diagrama de matriz mostra a análise de agrupamento de ligação de sequências atribuídas a IGHV1-3*01. Os agrupamentos estão identificados com cores mais claras. Fonte: Adaptado de (CORCORAN et al., 2016).

Após a identificação dos alelos candidatos é realizado um filtro (*pre-germline*) para eliminar falsos positivos, mutações *hostspot* ou mutações oriundas de PCR (*Polymerase chain reaction*). Juntamente com os alelos candidatos é gerado um banco de dados que será utilizado na etapa de identificação de novos alelos.

Ao término das iterações de descoberta de novos alelos do gene IGHV, um filtro (*germline filter*) é aplicado para verificar a qualidade das sequências. Um banco de dados individualizado é criado com as sequências que atenderem os critérios do filtro (CORCORAN et al., 2016). No trabalho do IgDiscovery, foram encontrados dois novos alelos do IGHV-3-21*1, sendo validados por sequenciamento (VH3-21*01_S4693 e VH3-21-DEL).

A abordagem apresentada por Wendel e colaboradores (2017), foi criada para trabalhar com pouca amostra de sangue (interessante para avaliar o estado do sistema imune em crianças). Segundo o autor, apesar de existirem diversas ferramentas para identificação de novos alelos, existe a carência de um método simples na situação em que a quantidade de material biológico é limitante (WENDEL et al., 2017).

Os passos para a predição dos novos alelos neste trabalho consistem em:

(1) Após o sequenciamento (Illumina Miseq 2 x 250 PE) obtém-se apenas sequências únicas, sendo o conjunto de dados inicial, para minimizar os efeitos da SHM e expansão clonal. Estas sequências são mais prováveis de pertencerem a uma célula B *naive*, pois tem menos SHM que outros isotipos.

(2) Alinhar as sequências a um banco de dados de alelos de referência como IMGT (IgBlast), sendo atribuído a cada sequência o alelo mais próximo.

(3) Determinar grupos de sequências com 1, 2, 3 e 4 *mismatches*, considerados como possíveis *single nucleotide polymorphisms* (SNP). Sequências que não apresentam mutações ou que não possuem uma cobertura de mutações na posição são desconsideradas para análise de novo alelo.

(4) Sequências com 2 e 1 mutações foram inspecionadas para verificar padrões como por exemplo se as mutações idênticas aparecem em pelo menos 20% das sequências únicas. Como exemplo, considerando todas as sequências do gene IGHV1-8, o percentual da mutação G-T na posição x, deve ser menor que 20%. Além disso, é verificado se pelo menos

2% das sequências têm mutações diferentes na posição. Atendendo a estes critérios, o alelo é considerado um SNP e é indicado com possível novo alelo.

O método utilizado neste trabalho encontrou oito alelos que já foram reportados pelo TIgGER. De uma forma geral, este trabalho identificou 17 novos alelos que foram preditos a partir de seis genes em oito indivíduos.

No trabalho apresentado por Mikocziova e colaboradores (2020), utilizou-se TigGER v0.3.1, IgDiscover v0.11 e uma abordagem Bayesiana. Foram relatados 25 novos alelos de IGHV da linhagem germinativa de imunoglobulinas de cadeia pesada. Foram analisadas 98 amostras de dados do AIRR-seq (*Adaptative Immune Receptor Repertoire Sequencing*) de célula B *naive* sequenciadas na plataforma Illumina MiSeq (2 × 300 pb). Treze alelos foram submetidos à validação. Porém, dez foram confirmados por amplificação e sequenciamento Sanger. Foi detectado muita variabilidade *upstream* da região V-REGION, mais especificamente na 5' UTR L-PART1 e L-PART2.

Dos 25 alelos encontrados, 22 foram identificados pelas duas ferramentas TigGER e IgDiscover. Porém, dois foram identificados exclusivamente pelo IgDiscover e um foi identificado exclusivamente pelo TigGER.

Segundo o autor, TigGER e IgDiscover são mais adequados e geram dados de saída compatíveis com análise no R Studio 3.6.0 (MIKOCZIOVA et al., 2020).

1.6.2 Descoberta de alelos de imunoglobulina em dados de Genoma *short reads*

Os trabalhos que reportaram novos alelos putativos de imunoglobulinas em dados de genomas utilizaram o projeto 1000 genomas (1000 GENOMES PROJECT et al., 2015) (Tabela 2), que descreve variantes presentes no genoma humano de indivíduos e fornece informações da diversidade genética com o intuito de entender possíveis doenças.

Na fase 3 do G1K, foi realizado o sequenciamento de 2.504 indivíduos em 26 populações. Nesta fase, foram reportadas 88 milhões de variantes sendo: 84,7 milhões de SNPs; 3,6 milhões de pequenas inserções e deleções (*indel*); e 60.000 variantes estruturais. A plataforma Illumina foi a metodologia utilizada para o sequenciamento (1000 GENOMES PROJECT et al., 2015).

Tabela 2: Trabalhos que identificam alelos de Imunoglobulina a partir de dados de genoma

Referência	Sequenciamento	Anotação	Novos alelos V	Ferramenta
YU et. al., 2017	Illumina	G1K fase 3	3.609	Pipeline (AlleleMine)
KHATRI et al., 2020	-	G1K	409	Python, R (pmIG)

O trabalho descrito por Yu e colaboradores (2017), apresentou um *pipeline* de bioinformática desenvolvido em Java chamado AlleleMiner, que identificou possíveis novos alelos de Ig e TCR. Foram utilizadas sequências extraídas dos arquivos VCFs do projeto 1000 genomas (fase 3) para gerar os haplótipos depositados no Lym1K.

O AlleleMiner funciona da seguinte forma: primeiramente são obtidos os arquivos .VCF do G1K versão GRCh37. Além disso, realiza-se uma busca das posições dos genes TCR e Ig no BioMart (GRCh38.p2) e assim faz-se a correlação das posições para a versão 38 do genoma nos arquivos .VCF (YU; CEREDIG; SEOIGHE, 2017).

Em seguida, para cada gene Ig e TCR, o AlleleMiner recupera a sequência no genoma de referência da *University of California Santa Cruz* (UCSC) e obtém o SNP do arquivo VCF. Desta forma, é obtido todos os alelos de cada gene para o haplótipo (montagem do haplótipo). Na ocorrência de haplótipos idênticos, apenas um é considerado. Durante este processo, o *software* faz a contagem da quantidade de vezes que o mesmo haplótipo aparece; com essa contagem o usuário pode indicar um ponto de corte (*threshold*). Isto para eliminar haplótipos com ocorrência única, pois possuem alta probabilidade de serem erros de sequenciamento. No entanto, para recuperar toda a diversidade do banco de dados G1K, foi executado o AlleleMiner com threshold igual a 1 (YU; CEREDIG; SEOIGHE, 2017).

Adicionalmente, os haplótipos são submetidos a uma etapa de filtro. Esta etapa consiste em primeiro alinhar com o banco de dados IMGT, onde um alinhamento com 100% de identidade permite a adição no Lym1K. Para aqueles que não atenderem ao primeiro quesito, é verificado se o banco de dados dbSNP (SHERRY et al., 2001), dá suporte aos SNPs do haplótipo; caso positivo, este é adicionado no Lym1K com *status* de novo alelo putativo (YU; CEREDIG; SEOIGHE, 2017). Ao final, realiza-se um alinhamento com o

banco de dados VBASE2 (RETTER et al., 2005), com o intuito de verificar o nível de confiança dos alelos putativos.

O trabalho desenvolvido por KHATRI e colaboradores (2021), utilizou os dados dos arquivos VCFs do G1K e criou o banco de dados, o pmIG. Para a construção deste banco de dados, foi utilizada a versão GRCh37 do genoma de referência e foram analisadas as informações das 26 populações do G1K (KHATRI et al., 2021).

Para esta metodologia, os arquivos VCFs foram filtrados para serem considerados apenas: SNP, inserção e deleção (*indel*), excluindo-se variantes do tipo *copy number variations* (CNV). Para esta tarefa, foram desenvolvidos dois scripts, um em Python e outro em R, que ao serem executados recuperou-se 5.008 haplótipos de genes do lócus de imunoglobulina independentes de 2.504 amostras (KHATRI et al., 2021). Os alelos foram ordenados de forma decrescente. Posteriormente, com a ferramenta *Muscle* (EDGAR, 2004), foi realizado um alinhamento múltiplo em três bancos de dados diferentes: IMGT (GIUDICELLI; CHAUME; LEFRANC, 2005), IgPdb (https://cgi.cse.unsw.edu.au/~ihmmune/IgPdb/information.php) e VBASE2 (RETTER et al., 2005). Realizou-se então uma verificação manual dos dados, para identificar a acurácia do alinhamento (KHATRI et al., 2021).

É importante observar que a comparação dos genes V foi realizada com o tamanho total do gene (LPART1-LPART2-V-éxon), onde mutações na região líder são consideradas como um alelo diferente. As sequências líder foram incluídas pois indicam o uso do éxon V e, deste modo, influencia diretamente na composição do repertório (PRAMANIK et al., 2011).

Com este alinhamento é possível verificar o nível de confiança dos alelos que foram classificado como (KHATRI et al., 2021):

a) AS1 - alelos com suporte a quatro haplótipos identificados no IMGT, IgPDB e ou VBASE2 (classe 1 - alto nível de confiança);
b) AS2 - frequência de novos alelos - alelos com o mínimo de 19 haplótipos (mínimo de 10 indivíduos);
c) AS3 - novos alelos raros - que tem suporte entre 7 e 18 haplótipos (mínimo de 4 indivíduos).

Os alelos da categoria AS3, são considerados de baixa confiança com relação a haplótipos e são chamados de alelos raros. Contudo, apesar da baixa confiança, acredita-se que são alelos verdadeiros, uma vez que a ocorrência de sete haplótipos idênticos de 5,008

haplótipos independentes exclui a possibilidade de erro no sequenciamento (KHATRI et al., 2021).

Como os alelos do AS2 e AS3 podem ser duplicados ou divergentes, foram criadas mais três subcategorias;

a) Alelos por grupo de genes (GA): Genes marcados como duplicados no lócus IG: IGHV1-69, IGHV1-69D, IGHV3-43, IGHV3-43D, IGHV3-23, IGHV3-23D, IGHV3-64, IGHV3-64D, IGHV2-70 e IGHV2-70D.
b) Alelos por operações indistinguíveis (IO): Como existem genes V parálogos, o mapeamento de *short reads* pode ser errôneo. Esses genes podem ser organizados com base na sua similaridade. Foi gerada uma árvore *neighbord-joining* (NJ) para todos os genes V das IGH, IGK, IGL separadamente. Os genes que compartilham um clado com um pequeno tamanho de 0.02 são chamados de IO genes.
c) (SE) são alelos que não são anotados como GA ou IO: Alelos da categoria AS1, conhecidos ou de GA ou alelos OI também podem conter falsos positivos.

Por fim, foi identificada uma alta diversidade de genes oriundos da população Africana. O trabalho identificou 409 IGHV, 179 IGKV e 199 IGLV.

1.6.3 Descoberta de alelos de imunoglobulina em dados de Genoma *long reads*

No trabalho apresentado por (RODRIGUEZ et al., 2020), foi utilizada uma abordagem de sequenciamento de *long reads,* com a plataforma *Pacific Biosciences, SMRTbell* (*PacBio*, *Menlo Park, CA, United States*) e uma ferramenta de bioinformática chamada IGenotyper, para fazer uma caracterização do lócus de Ig. A ferramenta foi aplicada em dados de sequenciamento de trios (pai mãe e filho). Como resultado, gerou uma referência de alta qualidade, sendo mais de 98% completa e mais de 99% acurada.

Adicionalmente, foram identificadas duas novas variantes estruturais (*indel* > 50pb) no lócus de Ig. Quinze novos alelos de IGH foram identificados em apenas dois indivíduos, fato que confere uma grande contribuição para a completude do banco de dados IMGT (RODRIGUEZ et al., 2020).

Outro trabalho que investiga a diversidade alélica do lócus de imunoglobulinas, desenvolvido por Ford e colaboradores (FORD et al., 2020) apresentou o ImmunoTyper que tem a finalidade fazer a genotipagem do IGHV e identificar CNVs e variantes estruturais

(*Structural Variants*- SVs) em dados de *long reads* com a plataforma PacBio. Esta é a primeira ferramenta que trabalha com sequências completas de WGS de *long reads* aplicada ao lócus de imunoglobulina.

Com o uso de *long reads*, foi possível obter informações detalhadas das regiões codificantes e não codificantes do gene IGHV. É importante destacar que, apesar do ImmunoTyper trabalhar com os alelos conhecidos do IGH, em versão futura será disponibilizada uma função para a identificação de novos alelos (FORD et al., 2020). O ImmunoTyper foi eficiente em reportar sequências não codificantes, sequências que flanqueiam os genes IGHV e alelos juntamente com CNVs para maioria dos pseudogenes IGHV. Uma exceção é o gene 1-69, onde os alelos *01 e *06 se diferenciam em apenas uma base, fato que dificulta a identificação. Uma observação é que o gene 4-31 foi identificado, apesar de apresentar duas cópias (FORD et al., 2020).

Outro trabalho não específico de sequenciamento de imunoglobulina é apresentado por (EBERT et al., 2021), que teve como objetivo investigar as SV, ou seja, variantes maiores que 50pb (inserção, deleção ou duplicação) em dados de sequenciamento de *short reads* e *long reads*. Para *short reads* foram utilizadas as amostras do projeto G1K (plataforma Ilumina) que reportou 69.000 SVs em 2.504 amostras. Ao analisar 32 amostras, incluindo as do G1K sequenciadas com *long reads*, o trabalho reportou 107.136 SVs, superior ao reportado pelo G1K.

O trabalho apresentado por (NURK et al., 2022) utilizou principalmente PacBio (Hi-Fi) e Oxford Nanopore, para gerar uma nova referência do genoma humano T2T-CHM3, que fechou gaps em todos os cromossomos (com exceção do cromossomo Y) e adicionou na referência 200 milhões de pares de bases e corrigiu erros das referência anteriores. Esta referência já está disponível na *University of California, Santa Cruz* (UCSC).

1.6.4 Banco de dados Genome Aggregation Database (gnomAD)

Atualmente existem poucos estudos dedicados a investigar a diversidade das imunoglobulinas Ig/TCR, que utilizam o sequenciamento de *long reads* (EBERT et al., 2021; RODRIGUEZ et al., 2020). Além disso, outro ponto fundamental a ser considerado é que o custo do sequenciamento *long read* ainda é elevado (RODRIGUEZ et al., 2020).

Portanto, para investigar a existência de novo alelos putativos de imunoglobulinas, foi utilizado neste trabalho o catálogo de variantes *Genome Aggregation Database*

(gnomAD) versão v2.1.1, que sequenciou inicialmente amostras de 199.588 WES e 20.314 WGS de diversas populações. O conjunto de dados é formado de estudos de controle de doenças comuns como: doenças cardiovasculares, diabetes tipo 2 e transtornos psiquiátricos. Os dados brutos sequenciados são em torno de 1,3 e 1,6 *petabytes* (KARCZEWSKI et al., 2020).

Este grande volume de dados faz do gnomAD o maior banco de genomas e exomas da atualidade. Os dados foram processados com um pipeline BWA-Picard-GATK, as *reads* foram mapeadas na versão GRCh37 e GRCh38 do genoma de referência. As amostras com baixa qualidade foram desconsideradas, ou seja, baixa qualidade de sequenciamento, amostra com segundo grau de parentesco ou mais próximo, amostras que tenham doenças severas e outras (KARCZEWSKI et al., 2020).

Também foram removidos indivíduos com tipos específicos de doenças, como câncer, presente no *The Cancer Genome Atlas* (TCGA) e com distúrbios neurológicos, reportados pelo *Bravo TOPMed* (https://bravo.sph.umich.edu). Após o filtro, foram eleitos 125.748 WES e 15.708 WGS, com alta qualidade de sequenciamento de indivíduos únicos não relacionados (KARCZEWSKI et al., 2020).

A qualidade das variantes presentes no gnomAD foi verificada em exomas e genomas separadamente usando o mesmo pipeline. No entanto, a eliminação das variantes sem qualidade foi realizada em duas etapas. Primeiro um filtro "*hard*" de qualidade e posteriormente um modelo de *Random Forest* (RF) (KARCZEWSKI et al., 2020).

O primeiro filtro é a verificação das informações de qualidade de alinhamento das variantes. Foram excluídas variantes com i) excesso de heterozigotos "diferente da referência", definidos pelo *inbreeding coefficient* < -0.3; ii) variantes que não possuem alta cobertura "*depth*" DP>=10; iii) *genotype quality* GQ>= 20; e iv) *minor allele fraction* >=0.2, para todos os alelos não referência de heterozigotos (KARCZEWSKI et al., 2020).

As variantes que obtiveram sucesso no primeiro filtro de qualidade, são submetidas a um modelo de RF, com o objetivo de separar variantes verdadeiras de artefatos. O modelo considera SVNs e *indels* juntos e verifica cada variante separada. Este modelo superou o índice de qualidade de variantes do *GATK* (KARCZEWSKI et al., 2020).

No site https://gnomad.broadinstitute.org/help/variant-qc são apresentados detalhes da configuração da RF, já que este método utiliza algumas métricas geradas pelo GATK que são utilizadas como *features* para modelagem da RF. No entanto, é necessário um conjunto

de treinamento e neste sentido foram utilizados os dados dos bancos de dados (omni, 1000 *Genomes high-quality* site e outros).

Além disso, foram utilizados dados de sequenciamento das amostras do NA12878 e CHM1-CHM13 que é uma mistura de DNA (50,7% / 49,3%) de dois haplóides da linhagem celular CHM, sequenciadas na plataforma PacBio, para aumentar a qualidade e determinar o *cutoff* da RF. É importante observar que a RF teve um desempenho superior ao GATK; variantes que não passaram no teste foram identificadas como RF. Neste procedimento foram removidas 12,2% de SNVs (RF *probability* >= 0,1) e 24,7% de *indels* (RF *probability* >= 0,2) em dados de exoma. Em genomas foram removidos 10,7% de SNVs e 22,3% de *indels*, e ambos com RF *probability* >= 0,4. É importante observar que a *feature* mais importante é o QD.

As ferramenta *Variant Effect Predictor* (VEP) versão 85, foi utilizada para anotar as variantes contra o Gencode v19 (KARCZEWSKI et al., 2020). O site https://gnomad.broadinstitute.org/faq#should-i-switch-to-the-latest-version-of-gnomad, apresenta a informação de que, na criação do gnomAD foram incluídas amostras de exomas do projeto G1K. Porém, não foram adicionados dados de WGS, uma vez que, os dados tem uma baixa cobertura.

1.7 Justificativa

A identificação de novos alelos codificadores de imunoglobulina contribui diretamente para o conhecimento da variabilidade dos segmentos gênicos do lócus, essencial para desvendarmos a geração de diversidade dos repertórios de anticorpos de humanos (WANG et al., 2011).

Identificar novos alelos é uma tarefa desafiadora, uma vez que o lócus de imunoglobulina apresenta diversas limitações técnicas em sua elucidação, como segmentos gênicos muito similares, duplicações, inserções e deleções (WANG et al., 2011).

O crescente número de estudos tem relatado que os dados do IMGT ainda não representam a diversidade de alelos presente no lócus de humanos. Consequentemente, a identificação e análise da diversidade de Ig e TCR e sua correlação com doenças estão inexploradas (GADALA-MARIA et al., 2015; WANG et al., 2011).

Adicionalmente, existem alelos no IMGT que apresentam problemas. Em uma análise de 226 genes/alelos IGHV, foram identificados 104 alelos como falso positivo, fato que

revela a existência de erros inerentes ao sequenciamento (WANG et al., 2008). Diversos trabalhos identificam alelos putativos de gene IGHV (KHATRI et al., 2020; WANG et al., 2011; WATSON; BREDEN, 2012; YU; CEREDIG; SEOIGHE, 2017), corroborando para a ideia de que o banco de dados IMGT está incompleto e com inconsistências.

No contexto de incompletude e acurácia do banco de dados IMGT, este trabalho propõe identificar e caracterizar variantes de imunoglobulinas nos dados disponibilizados de WES e WGS pelo gnomAD (KARCZEWSKI et al., 2020). O gnomAD, além de ser o banco mais completo de WGS e WES, utiliza dois critérios de qualidade altamente rigorosos (GATK e *Random Forest*), fatos que aumentam a probabilidade de obter variantes verdadeiras. Desta forma, hipotetizamos neste trabalho, que será possível encontrar um grande quantidade de variantes de genes V(D) e J de imunoglobulina utilizando o banco de dados gnomAD.

2 OBJETIVOS

2.1 Objetivo Geral

Identificar novas variantes gênicas nos segmentos VDJ do lócus codificador da cadeia pesada de imunoglobulinas utilizando o gnomAD.

2.2 Objetivos específicos

Os objetivos específicos são:

1. Desenvolver uma metodologia para busca de novas variantes do gene IGHV de imunoglobulinas;
2. Verificar a presença das prováveis novas variantes encontradas em bancos de dados existentes;
3. Verificar o número de prováveis novas variantes por gene IGHV;
4. Verificar se existe preferência das variantes encontradas por posição;
5. Verificar os tipos de mutações mais frequentes;
6. Verificar as frequências e exclusividades das prováveis novas variantes em diferentes populações;
7. Desenvolver um banco de dados integrado a uma plataforma web para armazenar e tornar acessível às prováveis novas variantes encontradas;
8. Desenvolver uma metodologia para busca de novas variantes dos genes IGHD e IGHJ de imunoglobulinas;
9. Criar critérios para separar variantes mais confiáveis e menos confiáveis;
10. Automatizar a metodologia de busca de novos alelos;
11. Correlacionar novas variantes com doenças.

3 MATERIAL E MÉTODOS

3.1 Obtenção de segmentos gênicos IGHV

As sequências dos segmentos gênicos V humanos foram obtidas do IMGT/GENE-DB (GIUDICELLI; CHAUME; LEFRANC, 2005) que é um banco de dados de imunoglobulinas. Ao acessar o site https://www.imgt.org/genedb/, foram realizados os filtros: *Specie*, *Gene type*, *Functionality* e *Molecular component*. Foram selecionadas respectivamente as seguintes opções: *Homo sapiens*, *variable*, *functional* e *IG*. Ao aplicar este filtro foram retornadas as sequências de nucleotídeos de 55 genes V funcionais e 288 alelos destes genes.

Ainda no site, foi selecionada a opção de obter estes dados em formato de texto da sequência completa do gene IGHV, genes e alelos. Posteriormente a mesma consulta foi realizada, porém para obter apenas as sequências V-REGION dos genes e alelos. Após isso, um programa escrito em Java versão 1.8 foi criado para organizar os dados em um arquivo FASTA, já que a sequência obtida do site vem em múltiplas linhas, este programa também gera um arquivo no formato texto com os nomes dos genes. É importante observar que esta consulta foi realizada na versão 3.1.22 do banco de dados IMGT, em 11 de abril de 2019.

3.2 Comparação das posições das variantes nos genes IGHV no GENCODE *versus* NCBI

O banco de dados gnomAD em sua versão v2.1.1, utilizou a anotação do GENCODE v19 (FRANKISH et al., 2019) da versão GRCh37.p13 do genoma de referência para a chamada de variantes. Este é um consórcio para anotação de genomas humanos e de camundongos (FRANKISH et al., 2019). Porém como será utilizado o IMGT/GENE-DB (GIUDICELLI; CHAUME; LEFRANC, 2005) para verificar a presença de novos alelos putativos, foi necessário verificar as posições dos segmentos gênicos no lócus de imunoglobulina da anotação do GENCODE v19, já que o IMGT utiliza a anotação do *National Center for Biotechnology Information* (NCBI), *RefSeq Reference Genome Annotation from build* GRCh37 (CHURCH et al., 2011).

Inicialmente foram obtidas as posições dos genes no arquivo de anotação tabular *General Feature Format* (*GFF)* do GENCODE, disponível para *download* no *link*: https://www.gencodegenes.org/human/release_19.html.

O mesmo procedimento de filtragem foi realizado para o NCBI GRCh37, arquivo GFF disponível no *link*: https://www.ncbi.nlm.nih.gov/projects/genome/guide/human/index.shtml, dado que o NCBI é a anotação de referência utilizada pelo IMGT. É importante observar que apenas os alelos *01 estão na referência. Nota-se que, a comparação é realizada com o tamanho total do gene. Logo, foram observadas inconsistências em alguns genes como: tamanho, posição inicial e final, conforme apresentado na TABELA 3 no capítulo 4. Afim de recuperar as posições dos genes IGHV anotadas pelo NCBI e GENCODE, foi utilizado o comando (*java -jar bin/filtrarGenesVGenomaGff.jar data/referencia/imgt/genesV.txt data/referencia/gencodev19/gencode.v19.annotation.txt > data/referencia/gencodev19/posicao-geneV-GenCode-GRCh37.txt*) para executar o programa Java que filtra as posições dos genes. O programa *filtrarGenesVGenomaGff.jar* recebe como parâmetro de entrada um arquivo com o nome dos genes e o arquivo GFF da referência, realiza o filtro e gera o arquivo com as posições na referência indicada.

Para averiguar a diferença entre as anotações, foram recuperadas as sequências de nucleotídeos do arquivo formato FASTA que são disponibilizados com a extensão FNA ou FA, do GENCODE e do NCBI. Foi utilizada a ferramenta *Samtools* (LI et al., 2009) na versão 1.6 *download* no site (http://www.htslib.org/) que disponibiliza a função para indexar o arquivo FASTA com o comando *samtools faidx <filereference.fa>*. Com o arquivo indexado é possível fazer buscas por regiões específicas, ao informar o intervalo e recuperar as sequências de nucleotídeos, via linha de comando.

Foi criado um programa Java versão 1.8 (*java -jar bin/gerarSciptSamtoolsGeneCode.jar data/referencia/ncbi-imgt/posicao-geneV-IMGT-NCBI-GRCh37.txt > data/referencia/ncbi-imgt/scriptSamtools.sh*), que recebe como parâmetro de entrada o arquivo com as posições dos genes e gera um outro arquivo *shell* Linux, que internamente possui os comandos *Samtools* que recupera as sequências de nucleotídeos. Este é um exemplo de um comando para recuperar os nucleotídeos (*samtools faidx GRCh37-chromossomo14.fna NC_000014.8:106494134-106494577*). No entanto, todas as sequências são recuperadas ao executar o arquivo *shell Linux* que contém todos os comandos. Exemplo do comando utilizado para recuperar as sequências de nucleotídeos dos genes IGHV do NCBI (*./scriptSamtools.sh > sequenciaGenesV-GenomaReferencia-NCBI-IMGT.txt*).

Primeiramente as sequências foram comparadas manualmente, obtendo a sequência de um arquivo e contrapondo com a sequência do mesmo gene do outro arquivo (GENCODE x NCBI). Ficou evidente a diferença, que também foi confirmada ao submeter estas sequências para o IMGT/V-Quest (GIUDICELLI; CHAUME; LEFRANC, 2004).

Portanto, devido às inconsistências, as análises realizadas neste trabalho utilizaram apenas os segmentos V-REGION dos genes V funcionais (Figura 5). Como validação fizemos um alinhamento global das sequências V-REGION contra as sequências do GENCODE.

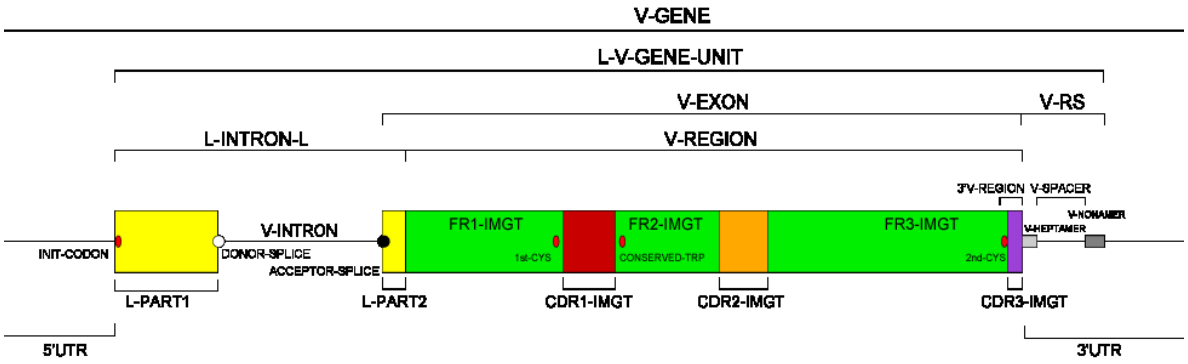


Figura 5: Nomenclatura do IMGT com as regiões dos genes V. L-V-GENE-UNIT compreende a região total de gDNA do gene, na configuração da linha germinativa e que é composta por L-PART1, V-INTRON, V-EXON e V-RS. As abreviações L são as regiões *leaders*, já a abreviação RS (*recombination signal*) é a região de recombinação de sinal. Fonte: (LEFRANC; LEFRANC, 2020).

Para esta tarefa foi utilizado o algoritmo "*Needleman-Wunsch Global Align Nucleotide Sequences*", do pacote Biopython versão 1.74 (COCK et al., 2009). Considerou-se apenas os alinhamentos com 100% de identidade. Foi utilizada a função *pairwise2.align.globalxx()* e para algumas situações foram penalizadas a abertura e extensão de *gaps* com o uso dos parâmetros *pairwise2.align.globalms(x, y, 2,-1,-0.5,-.1)*.

Este alinhamento foi feito com as sequências V-REGIONs obtidas do IMGT. Como as sequências V-REGIONs estão na orientação complementar reversa, foi necessário fazer esta conversão nas sequências do GENCODE. Para isso, foi desenvolvido um programa Java 1.8, o programa *gerarComplementarEReverso.jar*. Por fim, foi utilizado o s*cript* Python versão 2.7.17 e o BioPython para o alinhamento, com este comando (*python bin/AlinhamentoGlobalGenes.py data/referencia/gencodev19/sequenciaGenesV-GR-AAL-final-Fmt-CR.txt data/referencia/imgt/V-REGION-Fmt.fasta > data/referencia/gencodev19/sequenciaGenesV-GR-AAL-final-Fmt-CR-ALVREGION.txt*).

Após o alinhamento foi possível identificar a posição inicial e final da sequência V-REGION do IMGT, na referência GENCODE. Esta informação foi adicionada no cabeçalho de cada sequência dentro do arquivo FASTA. Logo, é possível identificar mais facilmente posições entre as referências.

3.3 Extração de variantes de genes IGHV do GNOMAD

O procedimento de extração das variantes do gnomAD (KARCZEWSKI et al., 2020), ocorreu de duas forma: Primeiramente foi realizado um filtro no site

(https://gnomad.broadinstitute.org/) com o nome de cada gene obtido do IMGT. Após o filtro do site apresentar o resultado, então, foi realizado o *download* de um arquivo no formato *Comma-separated values* (CSV), contendo as informações: nome do gene, cromossomo, posição na referência e o nucleotídeo alterado. Exemplo de um SNP: gene IGHV1-18, cromossomo 14, posição na referência GENCODE 106641585, substituição de uma guanina por uma adenina (G-A).

Posteriormente, foram utilizadas as informações do arquivo CSV para criar a *Uniform Resource Locator* (URL) de cada variante do gnomAD. Este arquivo serve como parâmetro de entrada para um *web scraping* desenvolvido em Java versão 1.8, utilizando o pacote (*lerPaginaAjax.jar*) e a classe (*lerpaginaajax. LerPaginaAjax*) e a biblioteca *Selenium*, que pode ser obtida no site https://www.selenium.dev.

Essa técnica é utilizada para coletar dados da *World Wide Web* (WEB) de maneira automatizada. Desta forma, para cada variante, foi realizado o *download* do arquivo HTML (*HyperText Markup Language* - Linguagem de Marcação de Hipertexto), contendo informações mais detalhadas de cada variante, informações como: quantidade de genoma e exomas que a variante está presente, informações populacionais e outras.

Os arquivos HTML foram processados com um programa desenvolvido em Java versão 1.8 pacote (*lerPaginaAjax.jar*), classe (*lerpaginaajax.TratarArquivosGerados*), sendo gerado como resultado deste processamento um arquivo final contendo as informações detalhadas de todas as variantes dos 40 genes. O comando utilizado para processar o arquivos HTML foi: (*java -cp bin/lerPaginaAjax.jar lerpaginaajax.TratarArquivosGerados data/dadosGnomAD/paginasGnomAD/ > data/dadosGnomAD/resultQtdeGenomaExoma.txt*).

3.4 Selecionar somente as variantes presentes na V-REGION dos genes IGH IMGT

Como foi delimitado o estudo para investigar somente as variantes presentes na V-REGION foi necessário filtrar do arquivo CSV apenas as variantes presentes nesta região. No processo de alinhamento das V-REGIONs com a referência, foi adicionado na identificação das sequências qual a posição que alinhou na referência (informação no arquivo FASTA). Então, foi criado um programa para acessar os dados das variantes no arquivo CSV, e para cada gene foi verificado no arquivo FASTA se a variante está no intervalo indicado na identificação da sequência. Este processo foi realizado para todas as variantes.

3.5 Posicionar as variantes do gnomAD nas sequências V-REGION

Como no arquivo FASTA das sequências V-REGION já existem as posições iniciais e finais com relação à referência GENCODE, a substituição das variantes obtidas no gnomAD a cada sequência V-REGION ficou mais simples, já que, a relação de substituição é direta (cada variante vai gerar uma sequência V-REGION, com o nucleotídeo substituído).

Dito isto, foi desenvolvido um programa em Java versão 1.8 (*posicionarVariantesNovaRefGeneCode.jar*), que, para cada variante recuperada do banco de dados gnomAD, os nucleotídeos foram substituídos em uma sequência V-REGION do gene em questão. Como exemplo, para o gene IGHV1-18 106641585-G-A, o programa obtém a V-REGION do gene 1-18 (e como já tem a informação do início na referência GENCODE) faz a substituição diretamente, gerando uma sequência já contendo a mutação. Isto é realizado para todas as variantes.

É importante observar que, após a adição/remoção/substituição de nucleotídeos e geração das sequências resultantes, essas já estão na orientação complementar reversa.

3.6 Verificação de alteração do gene referência a partir da análise de SNP ou indel

Para realizar a substituição dos SNPs e *indels* foram utilizadas as sequências *01 de cada gene do IMGT. Porém, essa alteração na sequência poderia por exemplo resultar em uma troca de família de um gene para outro. Para investigar esta situação, todas as sequências foram submetidas ao IgBlast (YE et al., 2013), e não foram identificadas situações como esta.

Antes de submeter as sequências para o IgBlast, foi necessário atualizar o banco de dados da aplicação, por meio do *makeblastdb* disponível pelo próprio programa, disponível dentro da pasta *bin*. Além disso, é importante evidenciar que foi utilizado o parâmetro (-*outfmt* 19) no comando para anotar as sequências. Este parâmetro faz com que os dados sejam gerados no formato de tabela, de acordo com os padrões especificados pelo *Adaptive Immune Receptor Repertoire* (AIRR) *Community* (https://docs.airr-community.org/en/latest/datarep/rearrangements.html).

3.7 Critérios de veto de variantes

Para filtrar variantes com alta qualidade, foi desenvolvido um *script* R (BUNN; KORPELA, 2013) na (*Integrated Development Environment* - Ambiente de Desenvolvimento Integrado) IDE RStudio (RSTUDIO TEAM, 2020), que remove as variantes do conjunto com base em três critérios:

a) variantes oriundas do WGS, dado à menor cobertura do sequenciamento para dados de genoma em relação aos de exoma;

b) variantes que não passaram no teste de qualidade gnomAD *Random Forest.* Foram selecionadas apenas variantes com a opção *PASS*, já que esse filtro do gnomAD remove do conjunto de dados 12,2% de SNV e 24,7% de variantes *indels,* em dados de exoma e finalmente;

c) variantes com inserção ou deleção maior que 50pb - variantes estruturais (EBERT et al., 2021).

Após o veto, o conjunto de dados ficou com um total de 10.550 variantes, apenas de exomas. Essas variantes foram divididas em quatro grupos, com base no número de variantes compartilhadas entre os exomas: grupo 1 contém variantes encontradas em um exoma, grupo 2, variantes presentes em 2 a 6 alelos de exomas, grupo 3, presentes em 7 a 18 alelos dos exomas e grupo 4, variantes presentes em mais que 18 alelos de exomas sequenciados.

3.8 Pesquisa de variantes associadas a doenças

Com o intuito de identificar a associação das variantes aqui encontradas com doenças, foi realizada uma busca nas bases de dados ClinVar, disponibilizada em formato VCF v4.1 de 01 de maio de 2021 (LANDRUM et al., 2018) (32) e GWAS Catalog (BUNIELLO et al., 2019), acessado em 26 de maio de 2021.

As variantes do banco de dados ClinVar foram baixadas no formato VCF via FTP no *link* https://ftp.ncbi.nlm.nih.gov/pub/clinvar/, versão GRCh37 da referência NCBI (CHURCH et al., 2011). Posteriormente, foi criado um programa Java 1.8, que filtrou, no arquivo VCF, as variantes do lócus de imunoglobulina, posição [106052774-107288051] de acordo com o NCBI.

Em uma busca no site do GWAS por estudos relacionados aos genes de imunoglobulinas, foi usado o intervalo 14:106052774-107288051 para filtrar as variantes. Para baixar os dados utilizamos a API RESTful (www.ebi.ac.uk/gwas/rest/api) disponibilizada pela plataforma.

Com o comando CURL, programa disponível no Linux, foi passado como parâmetro para API GWAS a região de interesse 14:106052774-107288051. Assim foi retornado um arquivo no formato JSON (*JavaScript Object Notation*), com todos os SNPs juntamente com os *rsIds* do banco de dados dbSNP (SHERRY et al., 2001) associados.

Posteriormente, foi desenvolvido um programa em Java versão 1.8, para filtrar os *rsIds* de cada SNP e adicionar em um arquivo *shell script* Linux, que ao ser executado faz o *download* de cada estudo associado ao SNP e salva em um arquivo JSON. Por fim, foi verificado se o *rsId* provenientes do GWAS e do ClinVar estavam presentes nas variantes aqui identificadas.

3.9 Verificar a presença das variantes identificadas em outros bancos de dados

Para verificar se as variantes filtradas nas etapas anteriores estavam presentes em outros bancos de dados como IMGT (GIUDICELLI; CHAUME; LEFRANC, 2005) e IgPDB (https://cgi.cse.unsw.edu.au/~ihmmune/IgPdb/index.php), realizou-se uma averiguação por posição. Primeiramente foram obtidos os dados dos respectivos bancos, como descrito nas seguintes etapas:

a) foi baixado novamente as sequências do banco de dados IMGT, agora na versão 3.1.34 de 31 de maio de 2021. Nesta versão a consulta por genes IGHV funcionais apresentou 57 genes e 311 alelos. Três desses alelos não possuíam sequências, só a numeração;

b) os dados do IgPDB estão armazenados no site https://cgi.cse.unsw.edu.au/~ihmmune/IgPdb/download.php. O próprio site disponibiliza a opção de fazer *download* de todo os dados do banco em formato FASTA. Essas sequencias foram obtidas e ao aplicar um filtro no arquivo foram identificados 194 alelos de IGHV, reportados por este banco de dados.

Em seguida, foi criado um *script* em Python 3.0 para executar, após obtenção dos alelos dos dois banco de dados, a separação em um arquivo FASTA de cada gene IGHV e seus respectivos alelos (*GroupAlleleGenes.py*). Este *script* foi executado nos dados de cada banco de dados separadamente. Em seguida, foi utilizado outro *script* (*MergeFiles.py*) para juntar os arquivos FASTA de cada gene/alelo dos dois bancos de dados. Então, cada arquivo de cada gene IGHV com seus alelos, contém todas as sequências dos dois bancos de dados.

Ainda, na identificação de cada sequência, dentro do arquivo FASTA, foi adicionada a informação da origem da sequência sendo IMGT ou IgPDB. Como indicado pelo asterisco em vermelho na Figura 6.

Após a organização dos arquivos, foi utilizado o MAFFT versão 7.310 (KATOH et al., 2002) para fazer o alinhamento múltiplo de cada arquivo (Figura 6). Ao término, uma curadoria manual foi realizada em cada arquivo para averiguar se o alelo *01 (alelo do IMGT) de cada gene tinha sofrido adição de *gaps*. Isso foi observado no arquivo do IGHV3-11. Nesta situação foi removido do arquivo FASTA o alelo 3-11*02 do IMGT e o *p07 do IgPDB, e realizado o alinhamento novamente, sem *gaps* desta vez.

Por fim, após o alinhamento, para cada variante identificada neste trabalho, foi observado na posição/coluna, se outros banco de dados (IMGT, IgPDB ou ambos) já haviam reportado o nucleotídeo (SNP). É importante observar que, como se trata de mais de uma sequência alinhada, a verificação é realizada em todos os nucleotídeos da coluna. Como exemplo de um SNP, observe na Figura 6 a coluna 6. Nas sequências do IMGT existem alelos dando suporte a timina T, já no IgPDB existem alelos dando suporte a *gap*, T, C e G. Portanto, esta posição só permite como novo SNP uma adenina (A).

Além disso, quando a variante tiver mais de um nucleotídeo, exemplo A-CTG, remoção de uma adenina e inclusão de uma citosina, timina e uma guanina, a verificação é realizada para os três nucleotídeos, cada um em sua posição/coluna. A variante será considerada reportada em outro banco de dados se ocorre *match* dos três nucleotídeos. O procedimento de verificar as posições é realizado pelo *script VerifyPositionVarAANC.py*.

Portanto, desta forma é verificado se a variante já foi reportada em outros bancos de dados. Esta informação é incluída no arquivo CSV das variantes. Este procedimento de validação, foi desenvolvido com base no trabalho apresentado por (MIKOCZIOVA et al., 2020), que realizou parte deste procedimento manualmente.

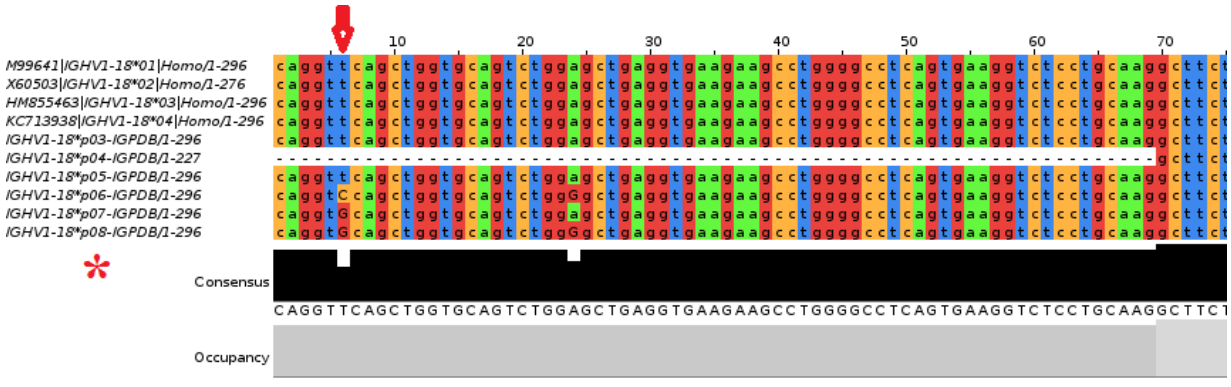


Figura 6: Alinhamento múltiplo dos segmentos gênicos IGHV1-18. Foram alinhados todos os alelo do IMGT e os segmentos deste gene do banco de dados IgPDB. A visualização do alinhamento foi gerada no Jalview 2.11.1.3, onde as cores para representar os nucleotídeos são: adenina verde, tirosina azul, citosina laranja e guanina vermelha. A figura apresenta uma seta indicando uma coluna, onde todos os nucleotídeos da coluna são considerados para comparação. Já o asterisco (*) em vermelho indica a informação da sequência que é utilizada para obter no nome do banco de dados que reportou a sequência.

3.10 Método utilizado para criar o banco de dados YVr-DB

Nas etapas anteriores deste trabalho produziu-se um arquivo no formato FASTA, que contém as sequências V-REGION, já com a substituição dos SNP, *indel* e também um arquivo CSV com as informações detalhadas de cada variante. Com a combinação destes 2 arquivos temos diversas informações como: nome do gene, posição no genoma de referência, de qual população a variante é derivada, presença em outros bancos de dados e as sequências de nucleotídeos da V-REGION com as substituições.

Entretanto, para facilitar a recuperação das informações, não é viável utilizar arquivos FASTA, já que, relacionar e aplicar filtros nos dados pode ser uma tarefa árdua, além de problemas de segurança da informação, já que com uma grande manipulação dos arquivos eles podem corromper e por último, a disponibilização dos dados para acesso de multiusuários.

Portanto, nesta situação utiliza-se a SQL (*Structured Query Language* – Linguagem Estruturada de Consulta), para relacionar e recuperar as informações de uma forma mais simples.

Então, foi criado um banco de dados no Sistema de Gerenciamento de Banco de Dados (SGBD) MySQL versão 8.0.20 (https://www.mysql.com/). O SGBD disponibiliza diversas funções como a Linguagem de Definição de Dados (DDL), que é utilizada para criar o esquema conceitual do banco de dados, ou seja, a definição das tabelas do banco de dados, juntamente com os atributos e os relacionamentos. Esta função engloba os comandos *create table, alter table, primary key* e comandos relacionados com a estrutura do banco de dados (ELMASRI; NAVATHE, 2011).

No entanto, para facilitar a modelagem das tabelas, utilizou-se a ferramenta MySQL Workbench 8.0.20, disponível no site https://dev.mysql.com/downloads/workbench/. Esta ferramenta é integrada ao MySQL e auxilia na criação da estrutura do banco de dados, fazendo a geração das tabelas fisicamente de forma automática. Inclusive, pequenas

alterações no modelo, a ferramenta atualiza sem a necessidade do usuário interagir com os comandos de criação e atualização de tabelas.

Outra função importante do SGBD é a Linguagem de Manipulação de Dados (DML), que disponibiliza diversos comandos como: *insert*, *select*, *update* e *delete*, que são utilizados para inserir dados, selecionar dados, atualizar dados e remover dados, respectivamente (ELMASRI; NAVATHE, 2011). Desta forma é possível cruzar as informações do banco e responder as perguntas com maior agilidade.

Desta maneira, foi então criado o banco de dados *online* de variantes de IGHV de imunoglobulinas denominado YVr-DB.

3.11 Método para criar a plataforma web YVr-DB versão 1.0

Após a implementação física do banco de dados, foi desenvolvida uma plataforma web. No processo de desenvolvimento desta plataforma, utilizou-se diversas tecnologias, tanto na criação da camada de apresentação, também chamada de *front-end*, que é a parte visual onde o usuário acessa os componentes do site, ou funções disponibilizadas pelo site, como também na camada de acesso aos dados, que é chamada de *back-end,* que é executada no servidor.

No *front-end*, utilizou-se a linguagem de marcação HTML, que tem como característica o uso de *tags*, que por sua vez, determina como o conteúdo é exibido no navegador *web*. Também foi utilizado o *Cascading Style Sheets* (CSS), que tem a função de aplicar estilos no HTML.

No entanto, com o uso do Bootstrap (https://getbootstrap.com/), foi utilizado pouco CSS, já que o *framework* disponibiliza diversos componentes e *layouts* prontos, fato que agiliza o desenvolvimento de um site. Nas situações em que as funcionalidades do site demandaram um comportamento dinâmico, utilizou-se o JavaScript (https://www.javascript.com), que é uma linguagem de programação criada para imputar este comportamento a um site web.

Para o *back-end*, foi utilizada a linguagem de pré-processamento de texto *Hypertext Preprocessor* (PHP) versão 5.5.9, disponível no site (https://www.php.net/). Esta linguagem, executada no lado do servidor, executa uma consulta SQL no banco de dados e retorna os dados para serem apresentados no *front-end*. A aplicação está hospedada em um servidor Apache versão 2.4.7 (https://www.apache.org/).

3.12 Consulta e formatação dos dados para gerar gráficos ou tabelas

Após os dados serem armazenados no banco de dados da plataforma YVr-DB, os procedimentos para recuperação das informações e realização das análises, foram realizados. Primeiro é criada uma consulta SQL, que seleciona os atributos de análise com os relacionamentos e os critérios de filtro desejados. Ao executar este comando na ferramenta Workbench 8.0.20, que disponibiliza uma função para consulta na base de dados é gerado um arquivo CSV da consulta. É importante observar que grande parte do trabalho já pode ser realizado no próprio comando SQL, como exemplo: agrupamento de dados com a função (*group by*), soma (*sum*), média (*avg*) dentre outras. Uma relação completa das funções disponíveis no SGBD pode ser acessada no *link*: https://dev.mysql.com/doc/.

Com o arquivo CSV já organizado, o próximo passo é importar os dados no R (BUNN; KORPELA, 2013), o que pode ser feito com o comando *read.csv("variants-imgt-new.csv", sep=";", dec=",", header = TRUE).* Porém, para facilitar o desenvolvimento dos gráficos, neste trabalho foi utilizada a IDE RStudio (RSTUDIO TEAM, 2020) e para plotar efetivamente os gráficos foi utilizada a biblioteca ggplot2. Esta sequência de procedimentos é realizada para grande maioria dos gráficos apresentados neste trabalho.

Para gerar alguns gráficos foi necessário realizar algumas transformações nos dados, com o intuito de melhorar a legibilidade e o entendimento dos mesmos. Este é o caso dos gráficos que apresentam no eixo X as coordenadas dos *frameworks* e CDRs, como as Figuras 13 e 15.

Dito isto, primeiramente foi necessário realizar a normalização das posições das variantes dentro da V-REGION, já que na figura a quantidade de variantes é apresentada em um escala de 1 a 311 e com identificação dos *frameworks* e CDRs.

A normalização das posições foi obtida da seguinte forma: como a V-REGION já foi alinhada ao genoma de referência anteriormente, para cada gene, é verificado no arquivo onde iniciou o alinhamento da V-REGION e obtém-se a posição inicial da V-REGION na referência. Este valor é considerado 1, ou seja, primeiro nucleotídeo da V-REGION. Com este valor, a posição de qualquer variante dentro do segmento é determinada pela subtração do valor final pelo inicial.

Como exemplo, considere a V-REGION do gene IGHV3-74 com início na posição 107218678, ou seja 107218678=1 na V-REGION. Suponha que, dado uma variante na

posição 107218684 (107218684−107218678=6), passou a ser a posição 7 dentro da V-REGION (6+1). Todas as posições das variantes foram normalizadas com base neste critério.

Porém, para visualizar todas as variantes foi necessário fazer um alinhamento múltiplo, com o software MUSCLE (EDGAR, 2004) de todas as sequencias V-REGION. Além disso, foi necessário ajustar a numeração obtida no passo anterior, já que foram inseridos *gaps* pelo alinhamento múltiplo, utilizando um programa em Java 1.8 (*posicionarVariantesAMGraficoR.jar*). Desta forma, como existem posições conservadas na V-REGION, posições que delimitam as regiões de CDRs e *frameworks*, foi possível identificar em qual segmento a variante se encontrava.

Então, foi executada uma consulta SQL no YVr-DB e recuperação de todas as variantes, sendo gerado um arquivo CSV com as posições das variantes na referência e a sequência completa. Posteriormente, foi executado o procedimento acima, onde o resultado seria um arquivo CSV já com as posições delimitadas de 1 a 311. Com este arquivo já estruturado foi feita a importação destes dados na IDE RStudio e foram plotados os gráficos que necessitam dessa informações no eixo X.

3.13 Método estatístico

Após a separação dos dados em quatro grupos, sendo o G1 aquelas variantes que foram sequenciadas em apenas 1 alelo, G2 para variantes que foram sequenciadas em 2 a 6 alelos, G3 variantes que foram sequenciadas em 7 a 18 alelos e G4 aquelas que foram sequenciadas em mais de 18 alelos, foram utilizados alguns métodos estatísticos com o objetivo de verificar se as variantes de grupo G1 eram aleatórias, considerando as métricas QD e VQSLOD do GATK, todos os testes foram realizados no R na IDE RStudio. Primeiramente, utilizou-se o teste de *Shapiro-Wilk* para verificar se os dados atendem uma distribuição normal ou não. Para isso, a hipótese H0 (hipótese nula) é que os dados seguem uma distribuição normal (*p-value* > 0,05). Já o H1 é quando os dados não seguem uma distribuição normal (*p-value* < 0,05). Portanto, ao executar este teste em cada grupo individualmente das métricas QD e VQSLOD, o resultado foi que o valor de $p < 0,05$, indicando que a hipótese nula foi rejeitada, ou seja, não segue uma distribuição normal.

Dado este resultado, foi utilizado o teste de *Kruskal-Wallis* nas métricas QD e VQSLOD, que é um teste não paramétrico para comparação de três ou mais grupos. Neste caso, o teste verifica se existe diferença na distribuição dos dados. Então a hipótese H0 é que

a mediana do G1 é igual a mediana do G2 que é igual a mediana do G3 que é igual a mediana do G4 (medianas G1=G2=G3= G4), *p-value* > 0,05. A hipótese H1, hipótese alternativa é que a mediana dos grupos são diferentes (medianas G1# G2#G3#G4), *p-value* <= 0,05.

Como o teste de *Kruskal-Wallis* apresentou diferença na distribuição de QD e VQSLOD entre os grupos, é importante verificar se a diferença é significativa. Para isto, foi utilizado o teste de *post-hoc* de *Dunn* com correção de *Bonferroni*. As hipótese são: H0 indica que os grupos comparados não tem diferença significativa nas medianas, *p.adj* > 0,05, hipótese nula. Já o H1, indica que houve diferença significativa nas medianas, *p.adj* <= 0,05, hipótese alternativa.

Todos os programas criados para busca e análise dos segmentos gênicos V estão disponíveis no *github* e podem ser acessados pelo *link:* https://github.com/frmmartins/yvrv. É importante observar que este método de busca de variantes diretamente nas páginas *web* (*web scraping*) do gnomAD versão 2.1.1, foi criado em 03 de setembro de 2019. Porém, nesta data não estavam disponíveis no site do gnomAD os arquivos VCFs.

3.14 Desenvolvimento de uma nova metodologia para filtrar genes VDJ

Como já tinha sido desenvolvido o *pipeline* para filtrar as variantes presentes nos segmentos gênicos V, foi utilizada esta implementação como base para o desenvolvimento do procedimento que filtra os genes D e J. Porém, a grande mudança é que ao acessar o site do gnomAD em 06 de dezembro de 2020, o grupo de pesquisa já havia disponibilizado o arquivo VCF com as informações de todas as variantes da versão 2.1.1.

Então, foi criada esta nova metodologia que filtra as variantes dos genes V, D e J. O *pipeline* já contempla a busca dos três segmentos gênicos, mas como os dados do V já foram processados, foi executada a busca apenas para os segmentos *REGIONS* D e J.

Além de alterar a forma de busca das variantes, já que foi disponibilizado o arquivo VCF, o *pipeline* foi dividido em duas partes. A primeira parte é a comparação das referências GENCODE x NCBI. Todos os *scripts* utilizados nesta etapa que tem a finalidade de obter a relação das posições das D-REGIONS e V-REGIONs estão organizados por ordem de execução em um *script shell Linux* (*runMainVerifyReference.sh*). Um ponto importante é que foi utilizado uma estrutura de pastas para facilitar o uso do *pipeline* e deste modo para executar é necessário baixar os *scripts* com estrutura de pasta completa. Os arquivos

referentes às *REGIONS*, já podem ser baixados juntamente e estão disponíveis na pasta (*data/reference/imgt/*).

As sequências dos genes D e J que estão na distribuição no *github* foram baixadas do site do IMGT em 06 de dezembro de 2020, versão 3.1.30. Foi realizado o filtro como indicado no método anterior, onde foram obtidas 23 sequências do gene D e 6 sequências do gene J, ambos para segmentos gênicos funcionais. Para atualizar os arquivos, basta fazer o filtro no IMGT, baixar as sequências novamente e substituir o arquivo na pasta *data/reference/imgt*.

Os arquivos GFF (arquivos de anotação) e FNA (arquivo com a sequência de nucleotídeos), do GENCODE e NCBI deverão ser adicionados na pasta (*data/reference/gencodev19*) e (*data/reference/ncbi-imgt*), respectivamente. Estes arquivos não foram incluídos na distribuição (*github*) devido ao tamanho.

Ainda, neste procedimento de verificar a posição nas duas referências, foi implementada uma funcionalidade que observa se o alinhamento das sequências de nucleotídeos GENCODE x NCBI estão com alinhamento de 100% de identidade. Caso não esteja, como o objetivo é verificar a posição do segmento NCBI no GENCODE, o programa verifica os nucleotídeos que faltaram no início ou no fim da sequência GENCODE e complementa a sequência GENCODE obtendo estes nucleotídeos do arquivo FNA e compara novamente com a sequência do NCBI.

Este procedimento é executado três vezes e caso não tenha 100% de identidade a sequência é gravada em um arquivo texto para que o usuário verifique manualmente dentro da pasta *data/reference/error-alignment/*. Para os genes V, D e J o programa fez o deslocamento da sequência NCBI na sequência GENCODE e recuperou todas com 100% de identidade. O arquivo *shell Linux runSubVerifyReference.sh* contém o conjunto de *scripts* que faz esta verificação e é utilizado internamente pelo *script* principal de verificação da referência (*runMainVerifyReference.sh*).

A segunda parte do pipeline, tem o objetivo de realizar a busca das variantes com base nas posições identificadas na etapa anterior. Como esta versão possui uma estrutura de pasta, então para realizar a busca, basta adicionar na pasta *vcfs/genome* ou *vcfs/exome* para que a busca seja realizada. Estas separação permite aos *scripts* do *pipeline* (*runMainMiningVar.sh*) adicionar a origem da variante (genoma ou exoma) no arquivo de resultado.

Além disso, é possível adicionar vários arquivos VCFs dentro das pastas. O *script* (*runMainMiningVar*) identifica esta situação e faz a busca em múltiplos arquivos. Para efetivamente aplicar o filtro no arquivo VCF foi utilizado o programa TABIX (LI, 2011), que busca variantes baseado em um intervalo em arquivo VCF (*tabix gnomad.exomes.r2.1.1.sites.14.vcf.bgz 14:106331761-106331771*). No entanto, é importante observar que o arquivo VCF deve estar indexado, já que a maioria das plataformas disponibilizam os arquivos VCFs compactados no formato gz e o index no formato TBI ou CSI. Porém, caso não tenha o arquivo indexado, basta executar o comando (*tabix -p vcf myfile.vcf.gz*).

Portanto, ao executar o script para filtrar as variantes (*./runMainMiningVar.sh data/reference/gencodev19/position-geneD-GenCode-FF.txt GENCODEDREGION data/reference/gencodev19/seqGeneD-Rev-final-fmt.txt*), com os parâmetros: arquivo que contém, para cada gene a posição que se deseja filtrar, nome do arquivo a ser gerado no *output*, arquivo com as sequências de nucleotídeos de cada gene revisado, arquivo gerado no processo de verificação da referência (*runMainVerifyReference.sh*).

O arquivo VCF é processado e produz como resultado um arquivo CSV com todas as variantes que atendem os critérios do filtro. Este arquivo é gerado dentro da pasta *vcfs/resultgroupvcfs/*. Também é gerado dentro desta pasta um arquivo com as variantes encontradas agrupadas por gene.

Além disso, para cada variante do arquivo CSV de resultado gerado no procedimento anterior, o *script* recupera a sequência do gene e faz a substituição da variante na sequência. E por fim, gera a sequência complementar reversa como é o padrão dos alelos depositados no banco de dados IMGT.

Com relação aos vetos, já está implementado no *script* para filtrar apenas as variantes que passaram em todos os testes, ou seja, que apresenta na coluna INFO a informação *PASS*. É importante observar que o resultado só contempla este filtro. Estão sendo considerados dados de genoma e exoma e não foram filtradas as variantes SV.

Todos os *scripts* criados nesta nova metodologia que filtra variantes presentes nos segmentos gênicos V, D e J estão disponíveis para *download* no *link*: https://github.com/frmmartins/yvrvdj. Para rodar a script é necessário ter o Java na versão 1.8 ou superior instalado na máquina e também é requisito ter o Python 3.

4 RESULTADOS

4.1 Resultado da comparação das posições dos genes IGHV do NCBI *versus* GENCODE

Afim de comparar os genes IGHV do genoma de referência GRCh37 provenientes de duas anotações distintas, NCBI e GENCODE v.19, foi feita uma análise das posições iniciais e finais de cada um dos genes IGHV (Tabela 3). Além das posições iniciais e finais é possível identificar na tabela o nome do gene na coluna "Gene" e o tamanho total do gene na coluna "T".

De um total de 55 genes IGHV funcionais presentes no IMGT foram recuperados apenas 40 genes no GENCODE. Portanto, não encontramos 15 genes descritos no IMGT na anotação do GENCODE.

Inicialmente ao considerar o tamanho total do gene, foram recuperadas 16.884 variantes do gnomAD. Ao filtrar para serem consideradas apenas as variantes que estão posicionadas dentro do segmento V-REGION, permaneceram 10.909 variantes.

Tabela 3: Comparação das posições dos genes IGHV utilizando anotações do NCBI e do GENCODE.

NCBI - GRCh37.p13				GENCODE v.19		
Gene	P. início	P. fim	T	P. início	P. fim	T
IGHV6-1	106405609	106406056	447	106405611	106406108	497
IGHV1-2	106452669	106453106	437	106452671	106453170	499
IGHV1-3	106471244	106471681	437	106471246	106471723	477
IGHV4-4	106478108	106478539	431	106478110	106478603	493
IGHV2-5	106494134	106494577	443	106494135	106494597	462
IGHV3-7	106518398	106518853	455	106518400	106518932	532
IGHV1-8	0	0	0	106539079	106539577	498
IGHV3-9	0	0	0	106552285	106552809	524
IGHV3-11	106573231	106573680	449	106573233	106573800	567
IGHV3-13	106586135	106586587	452	106586137	106586667	530
IGHV3-15	106610311	106610772	461	106610313	106610852	539
IGHV1-18	106641561	106641997	436	106641563	106642056	493
IGHV3-20	106667579	106668033	454	106667581	106668095	514
IGHV3-21	106691671	106692124	453	106691673	106692203	530
IGHV3-23	106725199	106725654	455	106725201	106725733	532
IGHV1-24	106733142	106733579	437	106733144	106733639	495
IGHV2-26	106757649	106758092	443	106757650	106758116	466
IGHV4-28	106780511	106780945	434	106780513	106781017	504
IGHV3-30	106791003	106791456	453	106791005	106791536	531
IGHV4-31	0	0	0	106805209	106805716	507
IGHV3-33	106815720	106816173	453	106815722	106816253	531
IGHV4-34	106829592	106830024	432	106829594	106830076	482
IGHV4-39	106877617	106878055	438	106877619	106878126	507
IGHV3-43	106926187	106926644	457	106926188	106926724	536
IGHV1-45	106962929	106963366	437	106962931	106963424	493
IGHV1-46	106967047	106967484	437	106967049	106967788	739
IGHV3-48	106993812	106994267	455	106993814	106994346	532
IGHV3-49	107012936	107013397	461	107012938	107013477	539
IGHV5-51	107034727	107035162	435	107034729	107035221	492
IGHV3-53	107048670	107049120	450	107048672	107049341	669
IGHV1-58	107078371	107078808	437	107078373	107078869	496
IGHV4-59	107083254	107083685	431	107081806	107083830	2.024
IGHV4-61	107095124	107095561	437	107095126	107095662	536
IGHV3-64	107113739	107114194	455	107113741	107114274	533
IGHV3-66	107131031	107131481	450	107131033	107131560	527
IGHV1-69	107169929	107170367	438	107169931	107170428	497
IGHV2-70	107178819	107179262	443	107178820	107179338	518
IGHV3-72	107198930	107199391	461	107198932	107199471	539
IGHV3-73	107210930	107211391	461	107210932	107211471	539
IGHV3-74	107218674	107219129	455	107218676	107219365	689
IGHV1-69-2	0	0	0	0	0	0
IGHV1-69D	0	0	0	0	0	0
IGHV2-70D	0	0	0	0	0	0
IGHV3-23D	0	0	0	0	0	0
IGHV3-30-3	0	0	0	0	0	0
IGHV3-30-5	0	0	0	0	0	0
IGHV3-43D	0	0	0	0	0	0
IGHV3-64D	0	0	0	0	0	0
IGHV3-NL1	0	0	0	0	0	0
IGHV4-30-1	0	0	0	0	0	0
IGHV4-30-2	106805207	106805644	437	0	0	0
IGHV4-30-4	0	0	0	0	0	0
IGHV4-38-2	0	0	0	0	0	0
IGHV5-10-1	0	0	0	0	0	0
IGHV7-4-1	0	0	0	0	0	0

4.2 Plataforma web YVr-DB versão 1.0

Foi desenvolvido neste trabalho uma plataforma web denominada *Antibody Variants Database* (YVr-DB), que pode ser acessada pelo seguinte endereço web (http://bioinfo.icb.ufmg.br/yvr/). No banco de dados desta plataforma estão armazenadas todas as variantes obtidas neste trabalho, juntamente com as informações relacionadas. A plataforma também disponibiliza diversos filtros e opções de *download* das informações do banco de dados.

4.2.1 Funcionalidades disponíveis na plataforma web YVr-DB versão 1.0

A plataforma desenvolvida apresenta em seu menu diversas funções como *Home*, *Summary*, *Help* e *About*. Neste tópico serão apresentadas apenas as funções disponibilizadas nos menus *Home* e *Summary*, já que o *Help* apresenta informações de como utilizar a plataforma e o *About* apresenta apenas as informações de contato com os responsáveis pela manutenção da plataforma.

Na tela *Home*, ou tela inicial, são apresentados os filtros que a plataforma disponibiliza para recuperar os dados, onde é possível informar o nome do gene no campo *Gene* (na Tabela 3 consta o nome dos genes a ser utilizados para busca), qual a população de interesse *Population* (*African, Ashkenazi Jewish, East Asian, European (no-Finnish), European (Finnish), Latino, South Asian, Other*), o tipo de mutação *Type of Mutation* (*frameshift, inframe_deletion, inframe_insertion, missense, protein_altering, stop_gained, synonymous*). Também é possível informar se a variante está presente em outros bancos de dados *Presence in DB*, ou seja, se a variante que se deseja filtrar é uma variante nova *New*, ou se já foi reportada por outros bancos de dados Others. Para esta versão os outros bancos de dados seriam o IMGT ou o IgPDB.

Ainda nesta tela, é possível informar em qual quantidade de alelos essa variante foi sequenciada. Este valor é informado no campo *Allele Count*, que contém uma indicação se o valor a ser pesquisado é maior ou menor, ou seja, ao informar 15, pode ser indicado mais de 15 ou menos que 15 alelos. Todas estas opções estão disponíveis como apresentado na Figura 7.

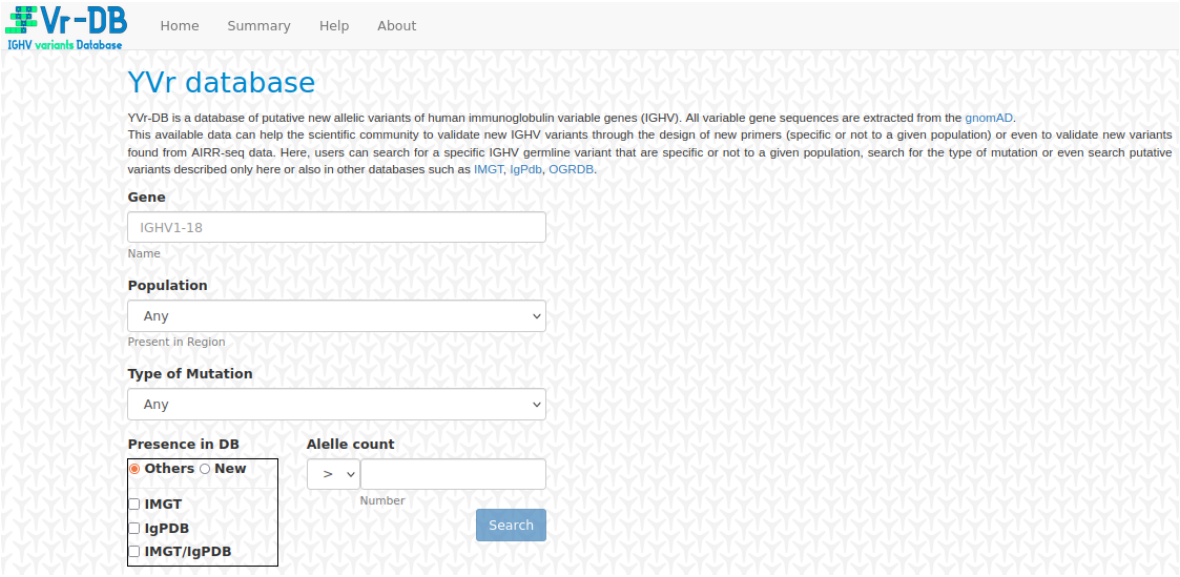


Figura 7: Tela inicial da Plataforma *web* YVr-DB. A figura apresenta todas as opções de filtro para variantes IGHV, que a plataforma disponibiliza, na versão 1.0.

Ao aplicar o filtro (pressionar o botão *Search*) o usuário é direcionado para a página de resultados, onde são apresentadas todas as variantes presentes no banco de dados que satisfaçam os critérios informados no filtro, feitos na *Home*.

Nesta tela, de resultados, além de apresentar os dados das variantes de interesse, também é apresentado um botão que possibilita ao usuário fazer o *download* das informações. É possível selecionar no botão a opção de *Nucleotides* ou *Aminoacid.* Após a seleção da opção, um arquivo no formato FASTA, com todas as informações das variantes é baixado automaticamente para o computador do usuário. A tela de resultados, é apresentada na Figura 8.

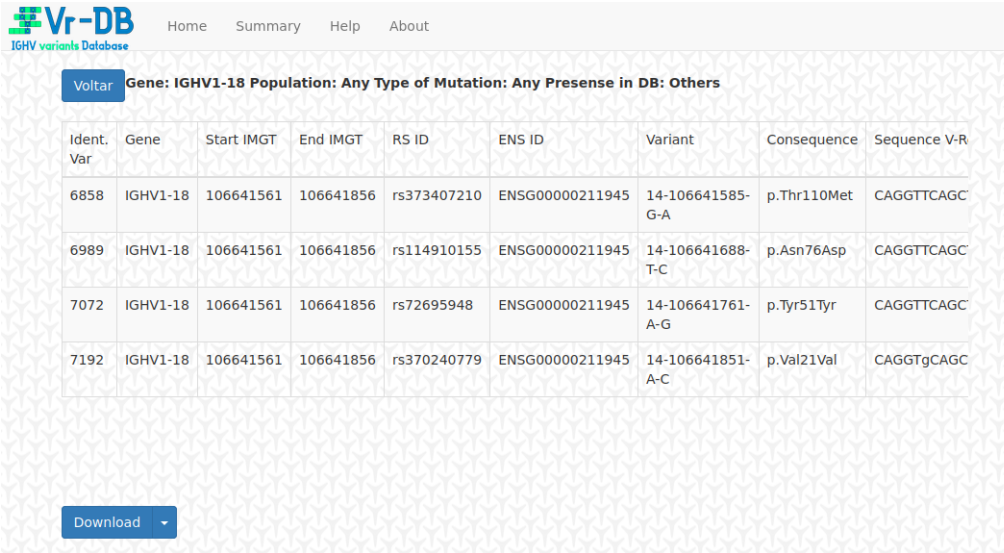


Figura 8: Tela que apresenta o resultado de uma busca submetida à plataforma YVr-DB. No cabeçalho da tela é possível verificar quais as informações de filtro.

Ainda na tela de resultado da busca submetida à plataforma YVr-DB é possível visualizar detalhes das variantes, clicando na linha onde se encontra a variante de interesse. Após isso, uma nova tela com detalhes da variante será apresentada. Nesta tela é possível verificar a posição em que a variante se encontra dentro do segmento gênico, mostrada em letre minúscula, onde é possível visualizar se a mesma está em uma região de CDR ou de *framework*. É importante observar que o segmento apresentado está na orientação complementar reversa, T-C, então a variante na complementar é um G, como apresentado na Figura 9, no CDR2.

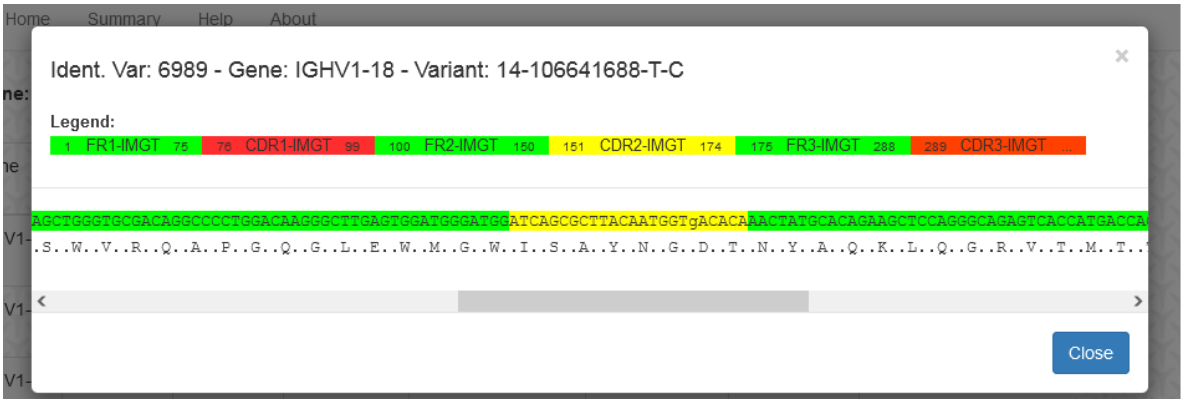


Figura 9: Tela de detalhes da variante no YVr-DB. Esta Figura apresenta os detalhes de uma variante selecionada pelo usuário. Mostra o identificador da variante no YVr-DB, o gene ao qual a variante pertence, juntamente com a identificação do cromossomo e o nucleotídeo alterado. Na legenda é apresentada a informação dos *frameworks* e CDRs. Cada segmento com uma cor específica. Os *frameworks* na cor verde e os

CDRs na cor vermelha, amarela e laranja para CDR1, CDR2 e início do CDR3 respectivamente. Também é apresentada a informação de início e fim de cada região. Esta informação, juntamente com a tradução da sequência de nucleotídeos para aminoácido, foi obtida ao submeter a sequência completa para a IgBLAST (YE et al., 2013). O nucleotídeo alterado pode ser observado em caixa baixa (letra minúscula) no CDR2.

Outra funcionalidade disponibilizada pela plataforma é o sumário *Summary*, que apresenta em forma de gráfico a quantidade de variantes que o banco de dados contém. A quantidade de variantes é categorizada por gene e contém a informação da quantidade de variantes novas (*New*) ou já reportadas em outros bancos de dados (*Other*), como apresentado na Figura 10.

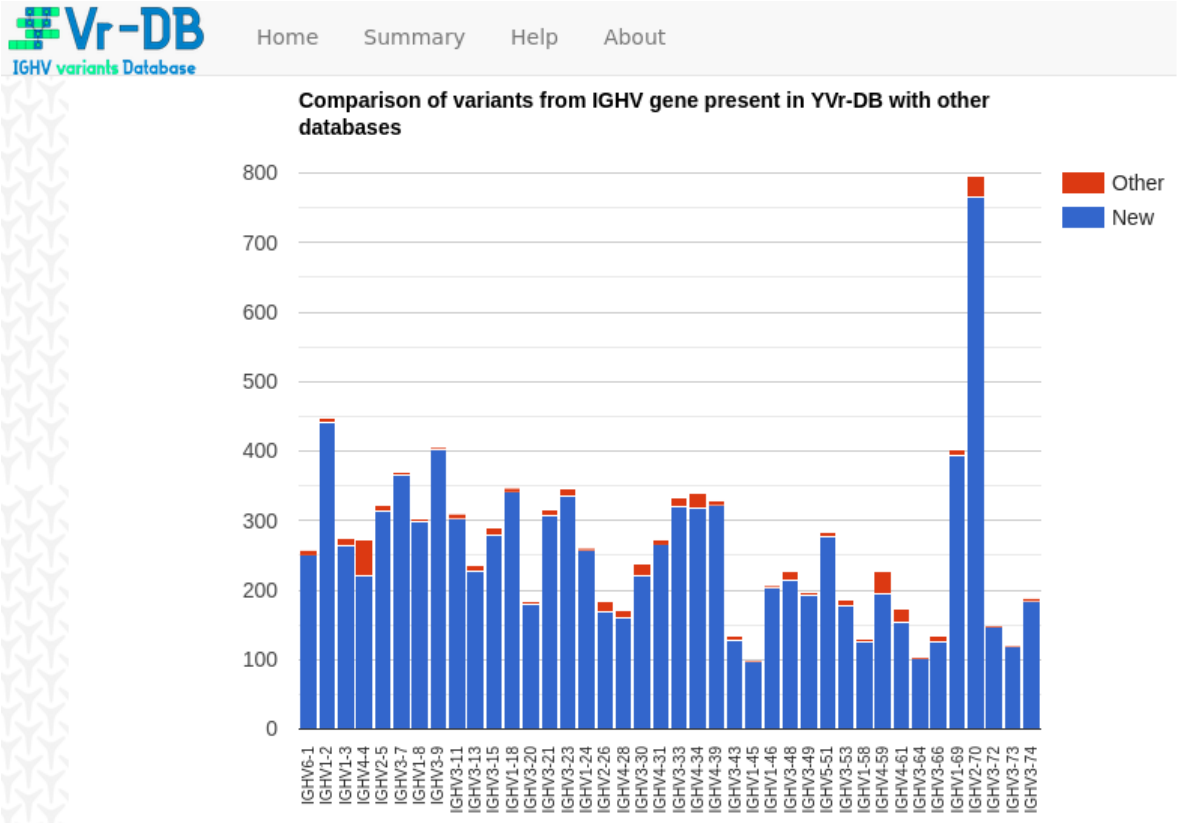


Figura 10: Gráfico da Plataforma *web* YVr-DB, guia *Summary*. No eixo X do gráfico são mostrados todos os genes já no eixo Y a quantidade de variantes presente no banco de dados, cada barra do gráfico possui duas informações uma indicando variantes novas e outra indicando variantes reportadas em outros bancos de dados. Os dados do eixo X estão ordenados no sentido 5'- 3', complementar reverso, sendo que o segmento gênico IGHV3-74 se encontra próximo à região telomérica.

Ainda na guia *Summary*, também está disponível uma tabela com a quantidade de variantes classificadas pelas famílias dos genes, e desta forma é apresentado a quantidade de

variantes encontradas para cada gene dentro da família e no rodapé da tabela é apresentado o valor total de variantes encontradas na família.

4.2.2 Construção da base de dados da plataforma YVr-DB versão 1.0

A Figura 11, apresenta o Diagrama de Tabelas Relacionais (DTR), modelo criado na ferramenta Workbench 8.0.20, que é incorporada ao MySQL. Esta ferramenta cria fisicamente as tabelas, os atributos e os relacionamentos do banco de dados da plataforma YVr-DB.

As tabelas são as representações das entidades que armazenam os dados. Já os atributos são as propriedades específicas de cada tabela. E por fim, os relacionamentos entre cada tabela, que é utilizado para garantir a navegabilidade, ou seja a ligação entre as tabelas. Para essa navegabilidade é importante ressaltar o uso das chaves primárias (*primary key*) de cada tabela, que além de ter o papel de identificar unicamente cada registro é utilizada para fazer a junção ou união entre as tabelas.

Cada tabela do YVr-DB, armazena informações específicas e assim evita que existam dados redundantes no banco de dados. As tabelas criadas nesta versão da plataforma são: *TBvariant* armazena as informações principais dos alelo, *TBgene* contém as informações dos genes relacionados a cada variante, juntamente com sua posição inicial e final na referência. *TBregionvar* armazena as informações referente as regiões geográficas, ou seja, o nome das populações, pois a *TBvariant_TBregionvar* armazena os dados que relacionam cada variante a uma população ou populações, sendo esta uma tabela de ligação. Já a tabela *TBtype_annotation* é uma tabela que armazena os tipos de anotações. A *TBdbexternal* contém a identificação dos bancos que a variante já foi reportada, IMGT e/ou IgPDB. E por fim, a *TBtranslator* armazena todas as informações obtidas do resultado do comando IgBlast (YE et al., 2013), já que o comando foi utilizado com o parâmetro (-*outfmt* 19), que apresenta como resultado uma tabela formatada no padrão AIRR, contendo a anotação dos segmentos gênicos V.

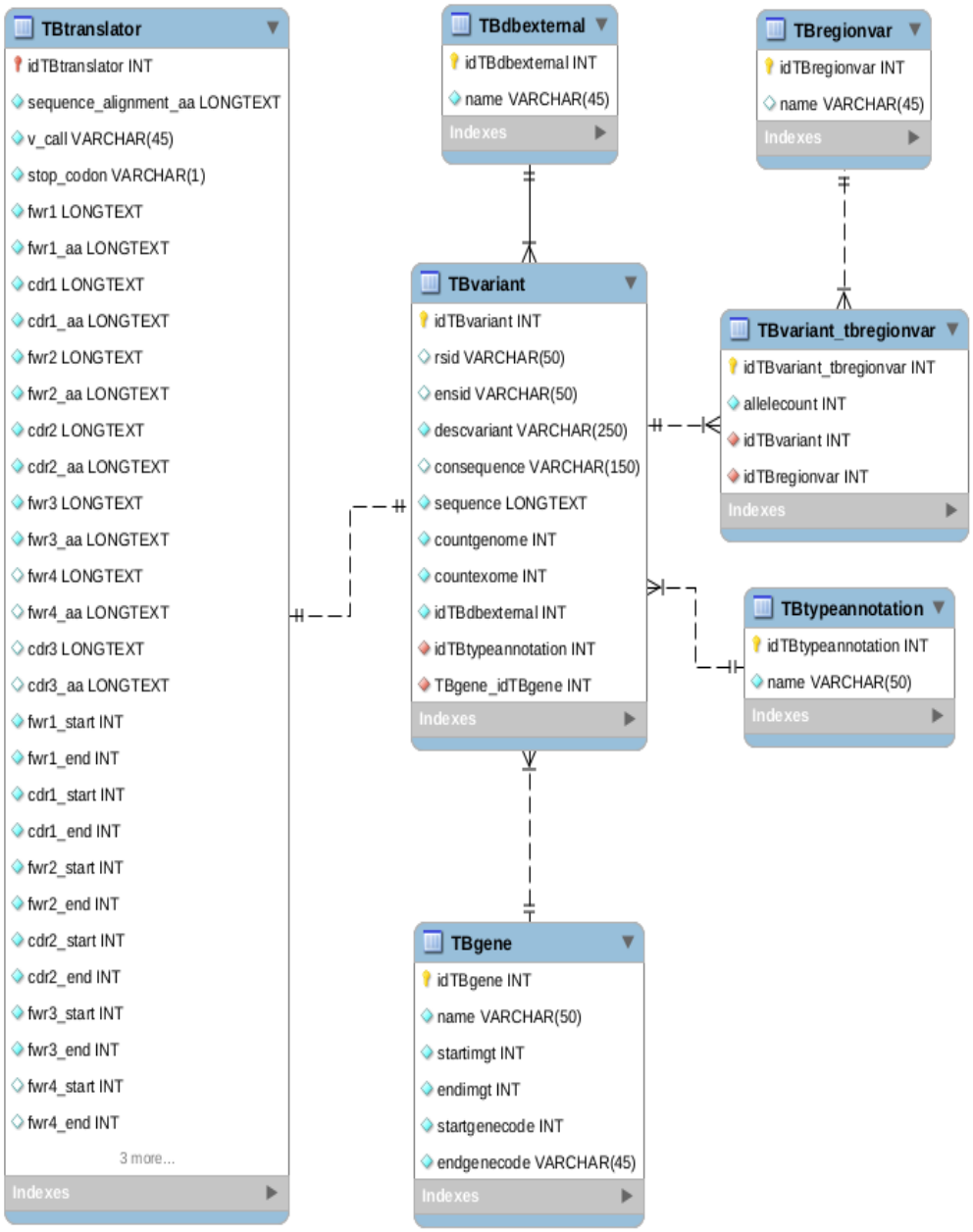


Figura 11: Diagrama de tabelas relacionas (DTR). Neste diagrama são apresentadas as tabelas que armazenam as informações das variantes da plataforma YVr-DB.

4.3 Mineração dos genes IGHV do exoma humano em larga escala revelou 10.550 variantes putativas

Foram encontrados 40 (quarenta) genes IGHV funcionais do gnomAD na anotação GENCODE (FRANKISH et al., 2019). Após filtrar as variantes exclusivas da V-REGION, foram encontradas 10.909 variantes. Em seguida, foram filtradas as variantes presentes no genoma, no caso 325, as que não preenchiam os critérios de qualidade da *Random Forest*, 28

variantes e 6 variantes estruturais, todas elas excluídas da análise, permanecendo na mesma apenas variantes de alta qualidade e variantes exclusivas de exoma.

Após os filtros, ficaram um total de 10.550 variantes de IGHV presentes nos exomas gnomAD conforme apresentado na Tabela 4. Algumas variantes descritas também foram encontradas em outros bancos de dados de linhagem germinativa de imunoglobulinas, como 278 em IMGT/GENE-DB (GIUDICELLI; CHAUME; LEFRANC, 2005), 75 em IgPdb (https://cgi.cse.unsw.edu.au/~ihmmune/IgPdb/information.php) e 41 em ambos os bancos de dados, resultando em apenas 394 variantes já descritas.

A maioria das novas variantes putativas descritas aparecem apenas uma vez nos exomas presentes no gnomAD (4.949 das 10.505 variantes), sendo 81 delas também detectadas no IMGT e/ou IgPDB, conforme apresentado na Tabela 4.

Notavelmente, 3.991 novas variantes putativas foram sequenciadas em 2 a 6 alelos diferentes, 813 aparecem em 7 a 18 e 797 foram sequenciadas em mais de 18 alelos (Tabela 4). Das 797 variantes mais frequentes, 575 não estão descritas nas outras bases de dados de genes V de linhagem germinativas (GLDB). Pelo menos uma variante de cada gene IGHV foi descrita em outros GLDB, como IMGT e/ou IgPDB (Figura 12).

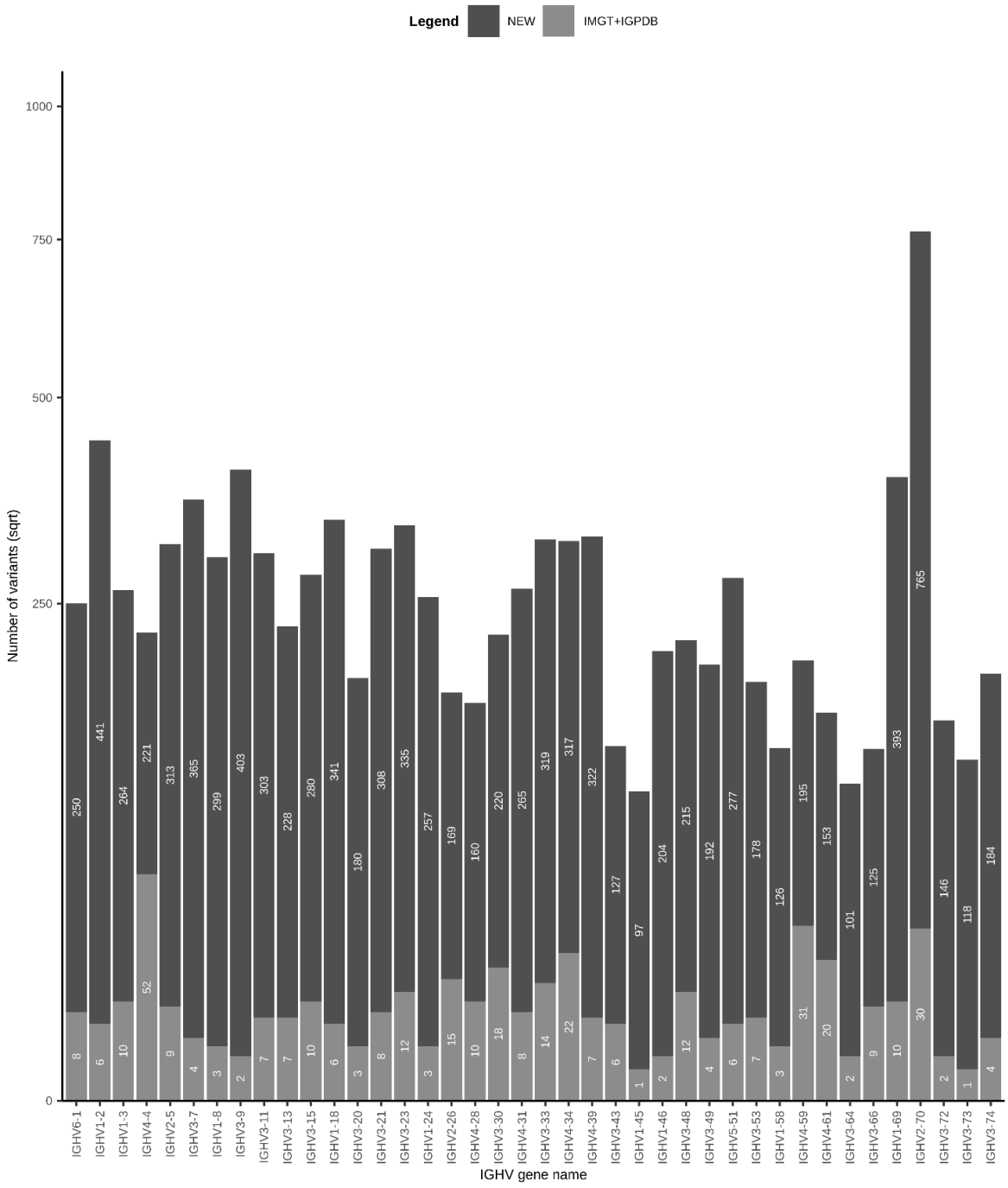


Figura 12: Variantes IGHV descritas neste trabalho e também encontradas em outras GLDB. As barras cinza claro representam variantes que já foram reportadas no IMGT ou IgPDB. No eixo X são apresentados os genes IGHV e no Y a quantidade de variantes em cada grupo. Foi utilizada a escala *sqrt* (raiz quadrada) para representar o tamanho das barras, a indicação da cor cinza claro são as variantes reportadas no IMGT e IGPDB, já cinza escuro são as variantes novas. Os dados do eixo X estão ordenados no sentido 5'- 3', complementar reverso, sendo que o segmento gênico IGHV3-74 se encontra próximo à região telomérica.

As variantes foram encontradas distribuídas ao longo de todos os segmentos gênicos pesquisados do lócus IGHV. No entanto, observamos uma variabilidade considerável no número de variantes de diferentes genes de IGHV, variando de 97 novas variantes para IGHV1-45 a 765 para IGHV2-70 (Figura 12). Além disso, 42,26% das variantes (4.459), foram encontradas no maior subgrupo IGHV, o IGHV3, como apresentado na Tabela 5. É importante observar que na busca inicial identificamos no IMGT 55 genes IGHV, porém 15 não foram encontrados no arquivo de anotação do GENCODE. Portanto, constam na figura apenas 40 genes.

Tabela 4: Número de vezes que cada variante putativa foi sequenciada no exoma (contagem de alelos)

		Presence of variants in other GLDB			Type of Mutation						
Allele Count	**Number of Variants**	**IMGT**	**IgPDB**	**IMGT/IgPDB**	**Missense**	**Synonymous**	**Stop_gained**	**Frameshift**	**Deletion**	**Insertion**	**Protein_altering***
1	**4,949**	**67**	**13**	**1**	**3,344**	**1,198**	**162**	**185**	**32**	**23**	**5**
2-6	**3,991**	**59**	**8**	**2**	**2,728**	**1,105**	**95**	**49**	**7**	**7**	**0**
7-18	**813**	**16**	**4**	**2**	**533**	**250**	**15**	**13**	**0**	**2**	**0**
>18	**797**	**136**	**50**	**36**	**476**	**278**	**18**	**17**	**4**	**4**	**0**
Total	**10,550**	**278**	**75**	**41**	**7,081**	**2,831**	**290**	**264**	**43**	**36**	**5**

**Protein Altering* refere-se a variantes em *frame* ou *frameshift* (inserções/deleções). Aqui, todas as 5 contêm inserção do quadro classificados pelo gnomAD como uma variante de alteração de proteína.

Tabela 5: IGHV variantes por subgrupo (família de genes)

IGHV subgroups (gene name, number of functional IGHV genes, and number of variants found in this study)											
IGHV1		IGHV2		IGHV3		IGHV4		IGHV5		IGHV6	
V1-18 (4F)	347	V2-26 (4F)	184	V3-11 (5F, 1P)	310	V4-28 (7F)	170	V5-51 (7F)	283	V6-1 (2F, 1P)	258
V1-2 (7F)	447	V2-5 (2F)	322	V3-13 (5F)	235	V4-31 (11F)	273				
V1-24 (1F)	260	V2-70 (17F, 1ORF)	795	V3-15 (8F)	290	V4-34 (13F)	339				
V1-3 (5F)	274			V3-20 (2F, 2ORF)	183	V4-39 (7F)	329				
V1-45 (3F)	98			V3-21 (6F)	316	V4-4 (9F)	273				
V1-46 (4F)	206			V3-23 (5F)	347	V4-59 (13F)	226				
V1-58 (3F)	129			V3-30 (19F)	238	V4-61 (9F, 1ORF)	173				
V1-69 (19F)	403			V3-33 (7F)	333						
V1-8 (3F)	302			V3-43 (2F)	133						
				V3-48 (4F)	227						
				V3-49 (5F)	196						
				V3-53 (5F)	185						
				V3-64 (6F)	103						
				V3-66 (4F)	134						

		V3-7 (5F) 369 V3-72 (2F) 148 V3-73 (2F) 119 V3-74 (3F) 188 V3-9 (3F) 405			
2,466	1,301	4,459	1,783	283	258

4.4 A maioria das variantes do IGHV estão no *Framework* 3

É importante observar que, a maioria das variantes (84,3%) estão presentes nas regiões de *framewok*: 20,9% (2.208) no FWR1, 22% (2.313) no FWR2 e a grande maioria no FWR3 (4.374 variantes ou 41,4%). Também foram encontradas variantes nas regiões CDRs como CDR1 (719 variantes, 6,8%), CDR2 (422 variantes, 4%) ou no início do CDR3 (514 variantes, 4,8%).

Não foram observadas tendências por uma posição específica na V-REGION: a frequência de 1 variante por posição foi de 35,5%, a de 2 variantes foi de 38,37% e a de 3 foi de 26%. No entanto, poucas variantes, como as derivadas do gene IGHV4-4, apresentam cinco variantes diferentes na mesma posição no FWR3. Conforme ilustra a Figura 13.

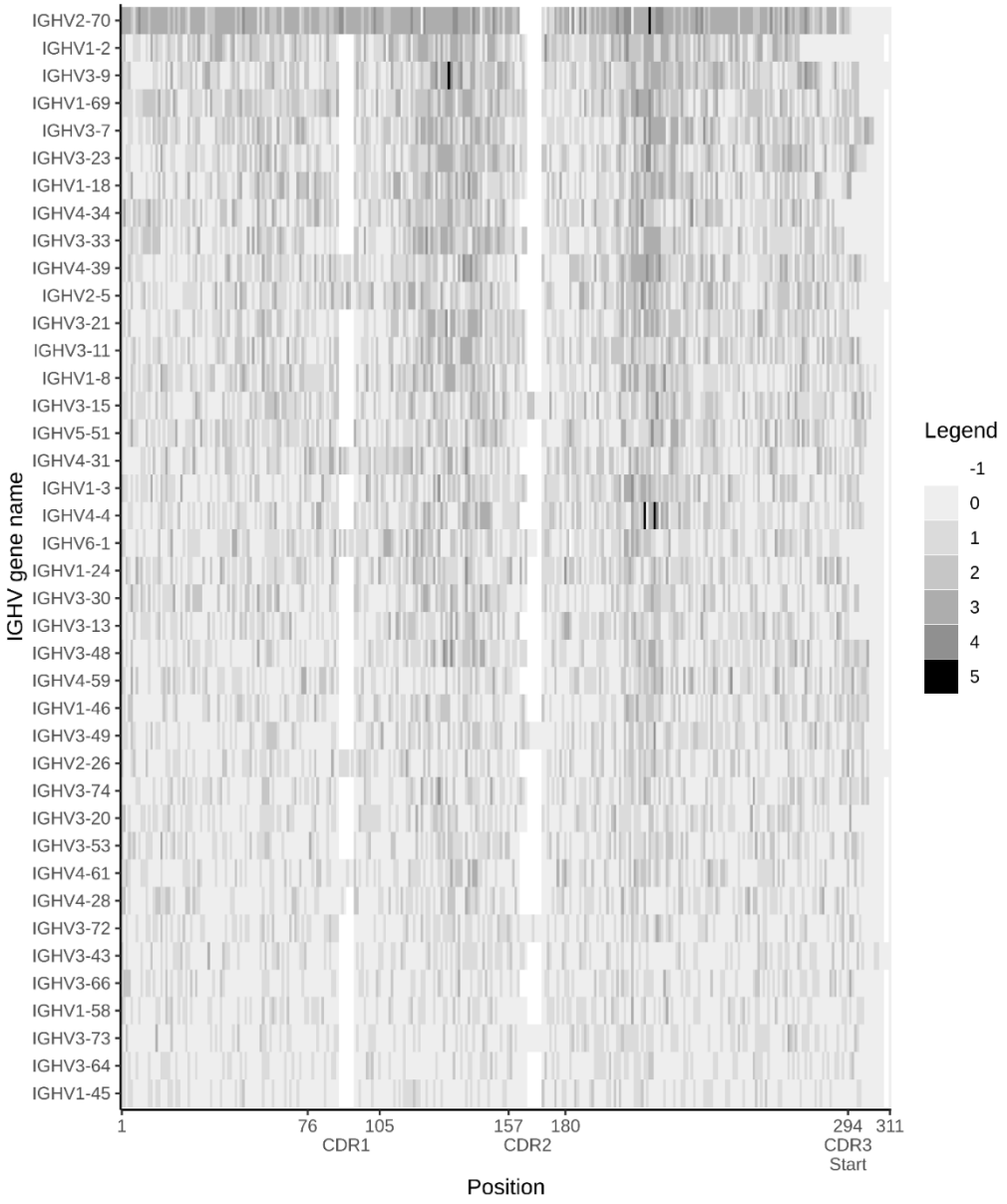


Figura 13: Número de variantes por posição. O *heatmap* apresenta a quantidade de variantes existentes em cada posição dos segmentos gênicos VH. A legenda indica a quantidade de variantes na posição. Sendo que -1 indica gap inserido pelo alinhamento múltiplo dos segmentos gênicos que vieram dos 40 genes do IMGT, de 0 a 5 a quantidade de variantes na posição. Estes valores vieram da quantificação das variantes do gnomAD em cada posição. O *heatmap* foi organizado em ordem crescente da quantidade de variantes que o segmento gênico possui. Sendo o IGHV2-70 o segmento gênico com maior quantidade de variantes. Também foi indicado no eixo X as posições de referência para o CDR1, CDR2 e o início do CDR3. Com estas marcações é possível ter o entendimento do FWR1-CDR1-FWR2-CDR2 e o final do FWR3.

4.5 A maioria das variantes IGHV são *missense*

Das 10.550 variantes descritas neste trabalho, a maioria são variantes *missense* (7.081 variantes: 67,1%) e sinônimas (2.831: 26,8%) e de nucleotídeo único (SNP), como apresentado na Tabela 4. Os genes IGHV2-70, IGHV1-2 e IGHV1-69, possuem o maior número de variações *missense* e sinônimas (Figura 14). No total, 290 variantes putativas encontradas neste trabalho resultam em códon de parada (*stop_gained*), incluindo 18 presentes no grupo de mais de 18 alelos. Além disso, mais de 17 *frameshifted*, *in-frame deletion* ou *insertion* aparecem no grupo de mais de 18 alelos (Tabela 4).

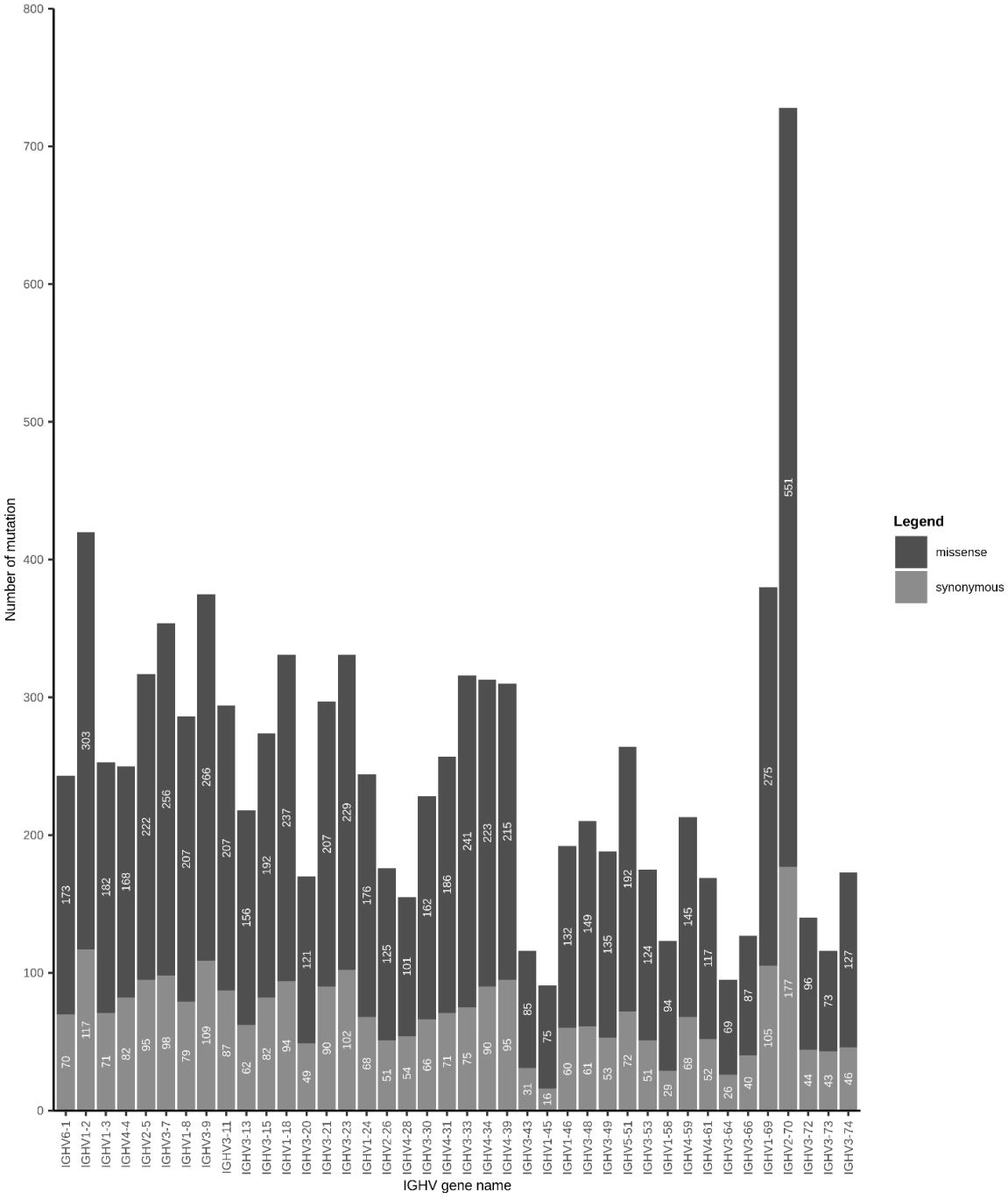


Figura 14: Número de variantes de substituição não-sinônimas (*missense*) e sinônimas presentes nos diferentes genes IGHV. As barras cinza-claro representam variantes sinônimas e as barras cinza-escuras representam variantes *missense*. No eixo X são apresentados os nomes dos genes, no Y o número de mutações

em cada gene, a indicação da cor cinza claro são as variantes reportadas no IMGT e IGPDB, já cinza escuro são as variantes novas.

As maiores deleções foram encontradas no segmento gênico IGHV2-70 (36 nucleotídeos), localizadas no FWR3, IGHV3-21 (35 nucleotídeos) e IGHV3-13 (33 nucleotídeos), localizadas no FWR2. Por outro lado, grandes inserções foram observadas nas variantes derivadas de IGHV1-18 (com uma inserção de 42 nucleotídeos) e IGHV1-2 (36 nucleotídeos), como apresentado na Figura 15.

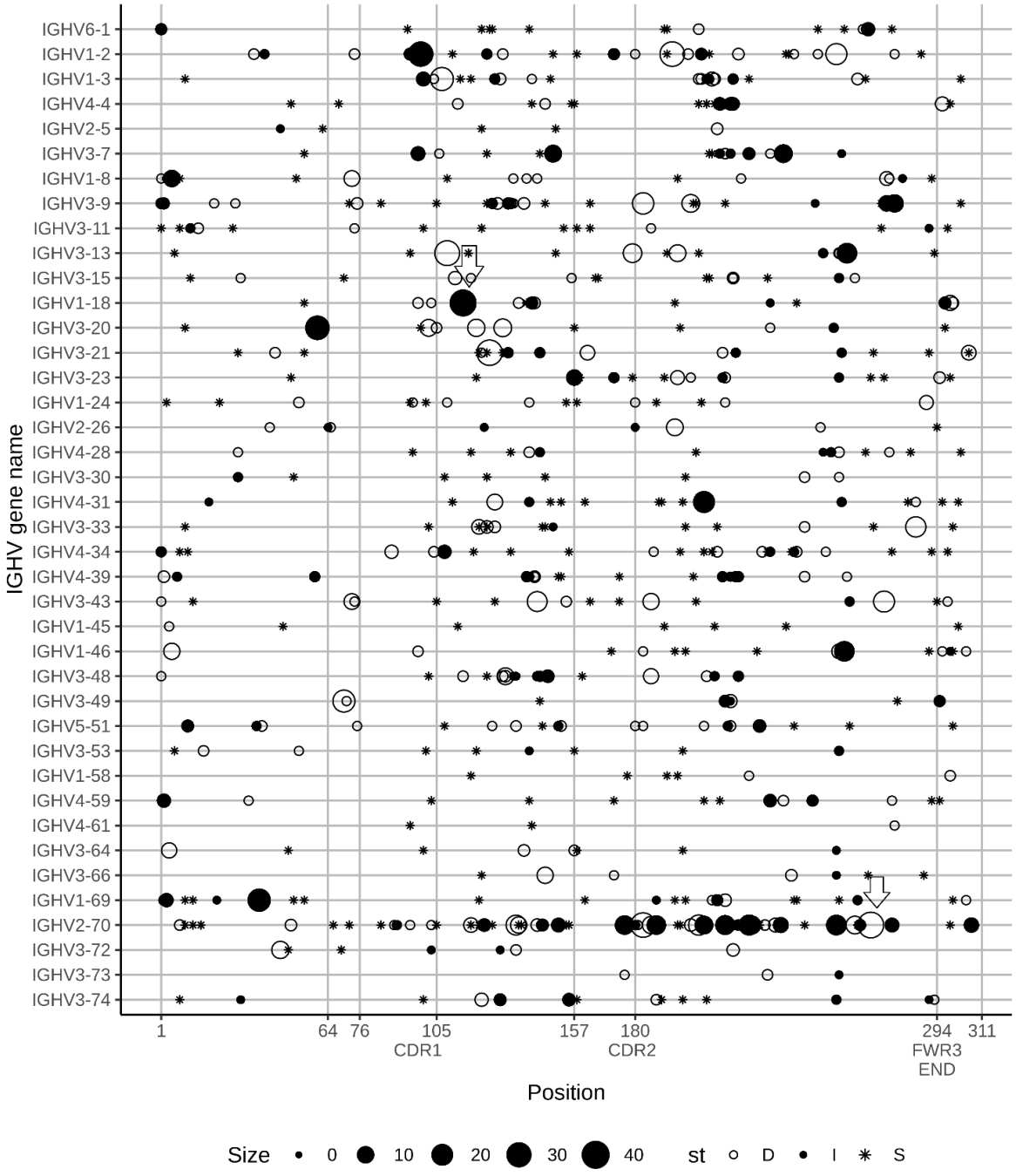


Figura 15: Número de deleções (o), inserções (•), ou mutações que levam ao aparecimento de *stop_codon* (*) por posição no segmento gênico IGHV. FWR1 (1 a 75), CDR1 (76 a 105), FWR2 (106-156), CDR2 (157 a 180), FWR3 (181-293) e início do CDR3 (294-311). A primeira cisteína conservada (64) também é destacada na Figura. Diferentes tamanhos de círculo representam o número de nucleotídeos inseridos ou

deletados. Quanto maior o círculo maior a inserção ou deleção. As duas setas indicam a maior deleção e maior inserção.

4.6 A maioria das variantes são de população específicas

Apenas 262 variantes aqui descritas foram compartilhadas por todas as populações representadas no gnomAD (Africano, Latino, Leste Asiático, Sul Asiático, Europeu (*finnish* e *no-finnish*) e *Ashkenazi Jewish*). A maioria dessas variantes são altamente prevalentes, conforme mostrado na Tabela 6. As variantes predominantes não incluídas em outros bancos de dados são provenientes dos subgrupos IGHV3 e IGHV4. Variantes nos genes IGHV4-31, IGHV4-4, IGHV4-39, IGHV3-33 são as mais frequentes.

Tabela 6: As 10 variantes de IGHV mais prevalentes e não presentes no IMGT e IgPDB

IGHV gene name	Variant Description	Allele Number	Allele frequency
IGHV4-31	14-106805381-T-G	82,924	0.3793
IGHV4-4	14-106478194-C-T	78,412	0.4481
IGHV4-31	14-106805395-G-A	63,483	0.2869
IGHV4-39	14-106877753-A-G	61,869	0.2703
IGHV3-53	14-107048767-G-T	37,893	0.185
IGHV4-39	14-106877796-G-C	37,715	0.1938
IGHV3-43	14-106926311-T-G	37,463	0.1543
IGHV3-33	14-106815862-A-G	26,121	0.1088
IGHV3-33	14-106815868-C-A	26,057	0.1086
IGHV3-33	14-106815970-T-C	22,555	0.09422

Foi detectado um aumento do número de variantes presente exclusivamente em algumas populações, embora presente em baixas frequências como pode ser observado na Figura 16 e na Tabela 7.

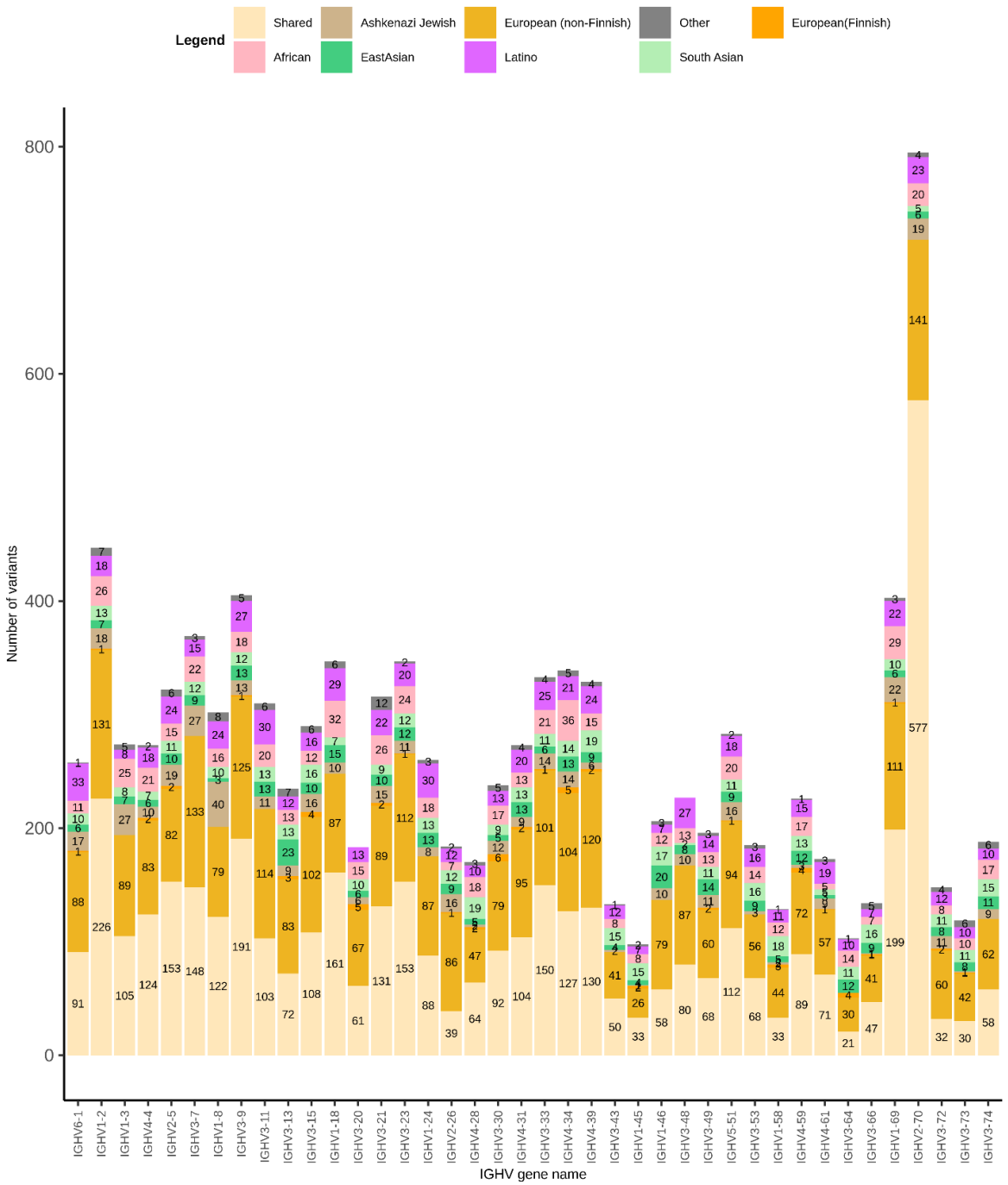

Legend
Shared
Ashkenazi Jewish
European (non-Finnish)
Other
European(Finnish)
African
EastAsian
Latino
South Asian
Number of variants
0
200
400
600
800
IGHV6-1
IGHV1-2
IGHV1-3
IGHV4-4
IGHV2-5
IGHV3-7
IGHV1-8
IGHV3-9
IGHV3-11
IGHV3-13
IGHV3-15
IGHV1-18
IGHV3-20
IGHV3-21
IGHV3-23
IGHV1-24
IGHV2-26
IGHV4-28
IGHV3-30
IGHV4-31
IGHV3-33
IGHV4-34
IGHV4-39
IGHV3-43
IGHV1-45
IGHV1-46
IGHV3-48
IGHV3-49
IGHV5-51
IGHV3-53
IGHV1-58
IGHV4-59
IGHV4-61
IGHV3-64
IGHV3-66
IGHV1-69
IGHV2-70
IGHV3-72
IGHV3-73
IGHV3-74
IGHV gene name

Figura 16: Número de variantes população-específicas ou compartilhadas por segmento gênico IGHV. O número de variantes por segmento gênico do IGHV pode ser exclusivo para as diferentes populações analisadas, como europeia (*finnish* e *no-finnish*), leste e sul da Ásia, latina, africana, outras populações não definidas, ou compartilhadas entre pelo menos duas populações diferentes.

Tabela 7: Número de variantes únicas por população e a variante mais frequente da população

Population	Number of unique variants	Most frequent IGHV	Variant Description	Allele Count	Allele Frequency
African	668	IGHV1-18	14-106641779-A-T	22	0.00008928
Latino	704	IGHV1-2	14-106452903-G-C	61	0.0002478
EastAsian	371	IGHV3-9	14-106552504-T-G	60	0.0002946
South Asian	475	IGHV3-53	14-107048714-G-A	39	0.0001582
European (non-Finnish)	3,286	IGHV3-21	14-106691736-G-A	24	0.00009735
Ashkenazi Jewish	460	IGHV1-2	14-106452694-T-C	15	0.00006103
European (Finnish)	63	IGHV3-13	14-106586381-C-A	10	0.00004069
Other	154	IGHV1-8	14-106539202-G-T	3	0.00001479
Total	6,181				

Foram observadas 668 variantes únicas de IGHV em africanos, 704 em americanos mistos (latinos), 846 em asiáticos (475 sul da Ásia + 381 no leste da Ásia), 3.809 em europeus (3.286 *non-Finnish*, 63 *Finnish* e 460 judeus *Ashkenazi*) e 154 em uma população não definida (Outros). Esses resultados indicam que em torno de 58,5% das variantes de IGHV encontradas neste trabalho (6.181 de 10.550) são específicas da população.

A maioria das variantes únicas (3.286) foram encontradas em europeus (*non-finnish*), a população com mais indivíduos sequenciados. A variante única mais prevalente

nesta população é derivada do gene IGHV3-21, com frequência de 0,00009735, como ilustrado na Tabela 7. A variante única mais frequente foi encontrada no gene IGHV1-2 da população latina (61 alelos sequenciados, 0,0002478) e IGHV3-9 da população do leste asiático (60 alelos, 0,0002946). Além da baixa frequência de variantes populacionais específicas, a maioria das variantes encontradas estão presentes em todos os segmentos gênicos do IGHV analisados Figura 16.

4.7 Variantes IGHV identificadas no Catálogo GWAS

Para identificar variantes genéticas de IGHV associadas a doenças foi realizada uma busca por variantes encontradas nos repositórios NHGRI-EBI GWAS (BUNIELLO et al., 2019) (33,34) e ClinVar (LANDRUM et al., 2018). Embora a pesquisa ClinVar não tenha recuperado nenhum resultado, 14 variantes de IGHV foram associadas a doenças/características no Catálogo GWAS. Essas variantes foram encontradas em sete posições diferentes (rs2073668, rs201076896, rs201691548, rs200931578, rs202166511, rs202117805, rs11845244) de três segmentos de genes IGHV (IGHV3-73, IGHV3-61, IGHV1-69).

A mutação sinônima rs2073668 (G>A; p.Ser26Ser) no IGHV3-73 tem uma frequência comum (0,4518) em todo o mundo, com a maior frequência observada em sul-asiáticos (0,6810). O polimorfismo (rs2073668) está associado (*in trans*) com os níveis de proteína da subunidade beta da ATPase transportadora de potássio em europeus (SUN et al., 2019). Esta variante também é descrita no IMGT (GIUDICELLI; CHAUME; LEFRANC, 2005).

Por outro lado, o alelo T da variante *missense* rs11845244 (C>T; p.Gly69Arg) no IGHV1-69 está associado, *in trans*, com o nível de Beta-defensina 119 (36). A variante mais comum nesta posição é descrita no IMGT (GIUDICELLI; CHAUME; LEFRANC, 2005) e IgPDB (https://cgi.cse.unsw.edu.au/~ihmmune/IgPdb/information.php), mas aqui descrevemos duas outras variantes de baixa frequência na mesma posição.

Ainda, dez (10) das 14 variantes de IGHV encontradas no catálogo GWAS foram associadas a um risco aumentado de doença cardíaca reumática (*Rheumatic heart disease*). Essas variantes estão presentes em 4 posições diferentes que estão em proximidade no segmento gênico IGHV4-61. As variantes mais frequentes foram descritas no IMGT, com indicado com asterisco na tabela. Destas 14 variantes, sete delas não estão presentes no IMGT (Tabela 8).

Tabela 8: Variantes presentes no Catálogo GWAS

rsid	Allele Count	IGHV gene	Trait
rs2073668*	112340	IGHV3-73	*Potassium-transporting ATPase subunit beta levels*
rs201076896*	2189	IGHV4-61	*Rheumatic heart disease*
rs201691548	1	IGHV4-61	*Rheumatic heart disease*
rs201691548	7	IGHV4-61	*Rheumatic heart disease*
rs201691548*	1978	IGHV4-61	*Rheumatic heart disease*
rs200931578*	1375	IGHV4-61	*Rheumatic heart disease*
rs202166511	3	IGHV4-61	*Rheumatic heart disease*
rs202166511*	1675	IGHV4-61	*Rheumatic heart disease*
rs202166511	2	IGHV4-61	*Rheumatic heart disease*
rs202117805	4	IGHV4-61	*Rheumatic heart disease*
rs202117805*	28144	IGHV4-61	*Rheumatic heart disease*
rs11845244	1	IGHV1-69	*Beta defensin 119 level*
rs11845244	1	IGHV1-69	*Beta defensin 119 level*
rs11845244*	97894	IGHV1-69	*Beta defensin 119 level*

* Variantes presentes no IMGT

4.8 Avaliação da distribuição das variantes de acordo com as métricas VQSLOD e QD

Com o objetivo de avaliar se as variantes mapeadas neste trabalho apresentam *scores* muito diferentes de qualidade, avaliamos aqui a distribuição dos *scores* de VQSLOD e QD das variantes deste trabalho, para os grupos G1, G2, G3 e G4 (Tabela 4). Como podemos observar na Figura 17 onde é apresentada a dispersão das *scores* das variantes com relação às métricas VQSLOD e QD, grande parte das variantes estão concentrada em valores mais positivos de VQSLOD 97,33% (10,268 > -20) e valores de QD abaixo de 20, independente do grupo avaliado.

É importante observar na figura que foi delimitado o valor inferior, ou seja, os valores de VQSLOD menores que -100 foram omitidos, para facilitar a visualização. Porém, é importante observar que podemos encontrar variantes no intervalo de 636,20 negativo a 6,79 positivo. Em uma análise no intervalo 636,20 negativo a -20 negativo, onde apresenta uma maior dispersão foram identificadas 282 variantes com os seguintes percentuais normalizado pela quantidade total de variantes em cada grupo: G1 0,029% (142 de 4.949), G2 0,018% (74 de 3.991), G3 0,0172% (14 de 813) e G4 0,0652% (52 de 797).

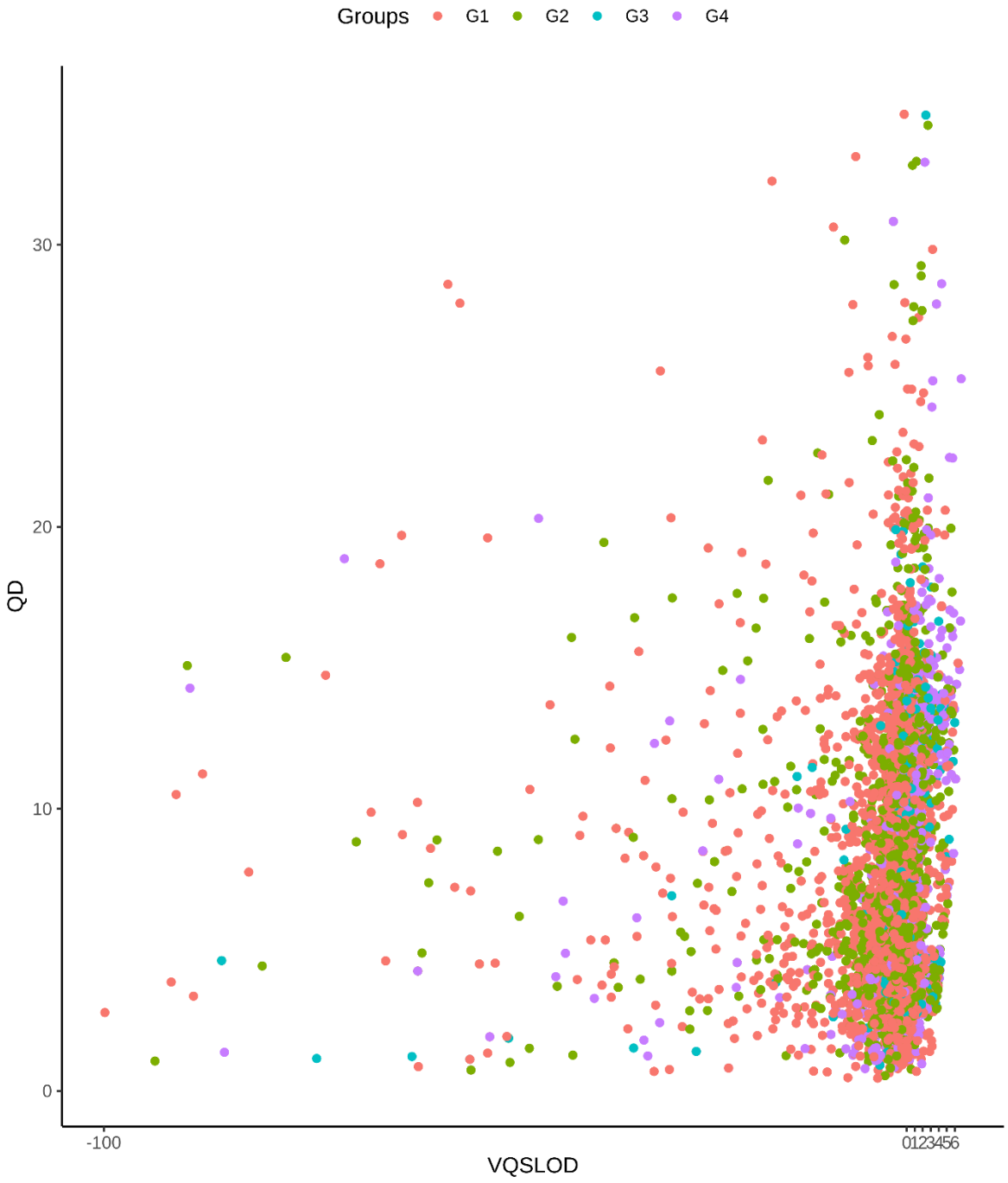


Figura 17: Métrica VQSLOD versus métrica QD dos grupos G1, G2, G3 e G4. Cada ponto na figura é a representação da combinação das métricas, os pontos possuem as cores vermelho, verde, azul e roxo, que representam os grupos G1, G2, G3 e G4, respectivamente.

A mesma comparação das métricas foi realizada, porém, no conjunto de dados onde foram criados dois grupos, o primeiro são as variantes que foram reportadas em outros bancos de dados como IMGT, IgPDB ou em ambos e o outro grupo são as variantes classificadas como novas. É importante observar na Figura 18 que a distribuição dos dados segue o mesmo padrão do gráfico anterior, não sendo possível definir alguma correlação entre os valores de VQSLOD versus QD. Em uma análise das 282 variantes no intervalo de -636,20 a -20, valores mais baixos e dispersos de VQSLOD, percebemos que 255 (0,025% : 255 de 10.156) destas eram novas variantes (não descritas no IMGT e IgPDB) e 27 (0,0685% : 27 de 394) delas já descritas pelo IMGT-IGPDB.

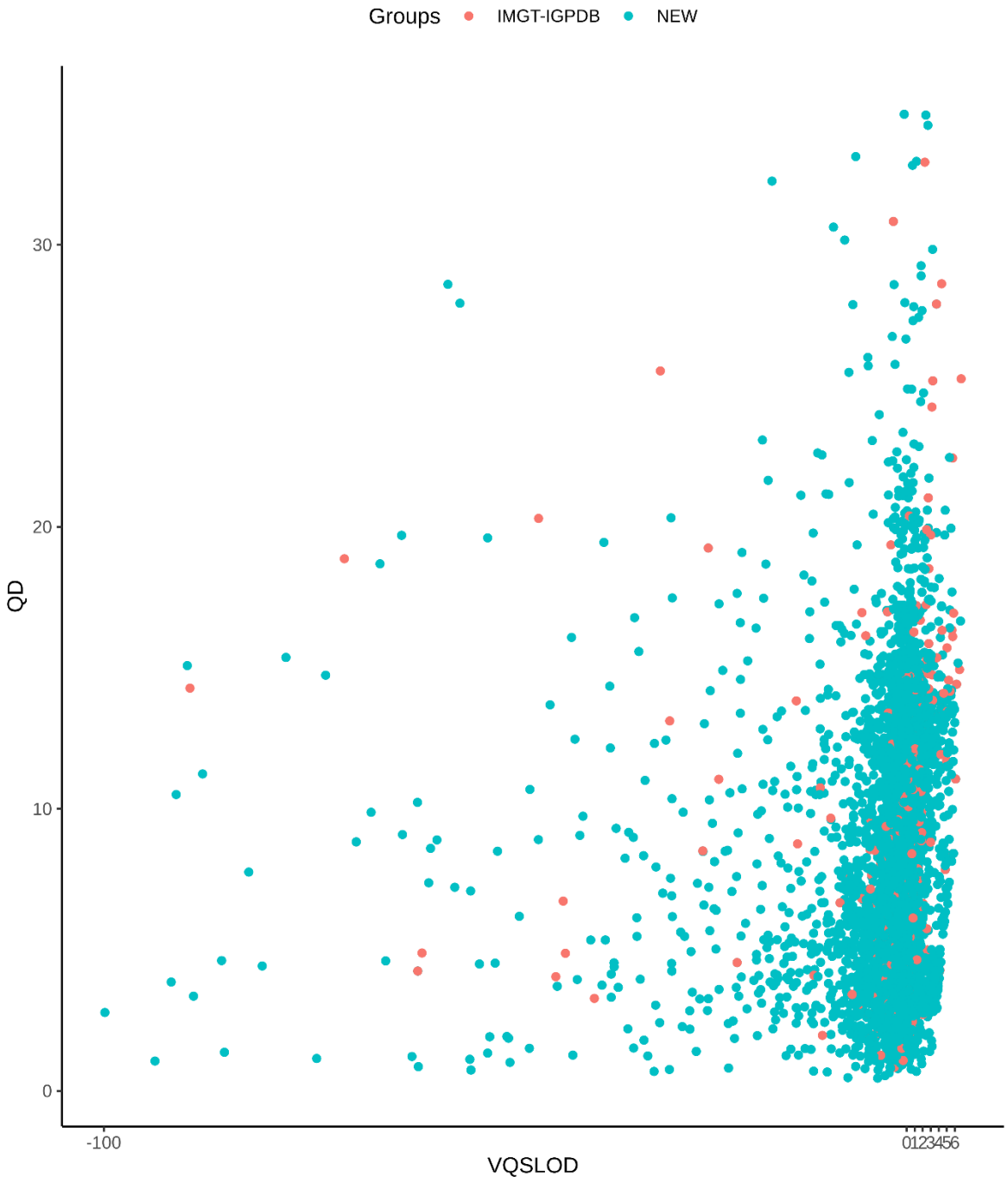


Figura 18: Métrica VQSLOD versus métrica QD do grupo com as variantes reportadas em outros bancos e o grupo das variantes novas. Cada ponto na figura é a representação da combinação das métricas, os pontos possuem as cores azul e vermelho, que representam os grupos IMGT-IgPDB e *NEW*, respectivamente.

Após realizar a separação das variantes em 4 grupos como apresentado na seção de metodologia, foi realizada uma análise estatística para verificar se existe diferença estatística na mediana das métricas destes grupos. Na Figura 19 é apresentado o *boxplot* com a comparação dos quatro grupos. Foram comparados todos contra todos e podemos observar que G1, G2 e G3, foram classificados com a letra (a), ou seja, estes três grupos não apresentam diferença estatística entre si com relação à mediana da métrica QD. Já o grupo G4, foi identificado com a letra (b), portanto, com relação a mediana este grupo se difere dos demais, apresentando uma maior mediana.

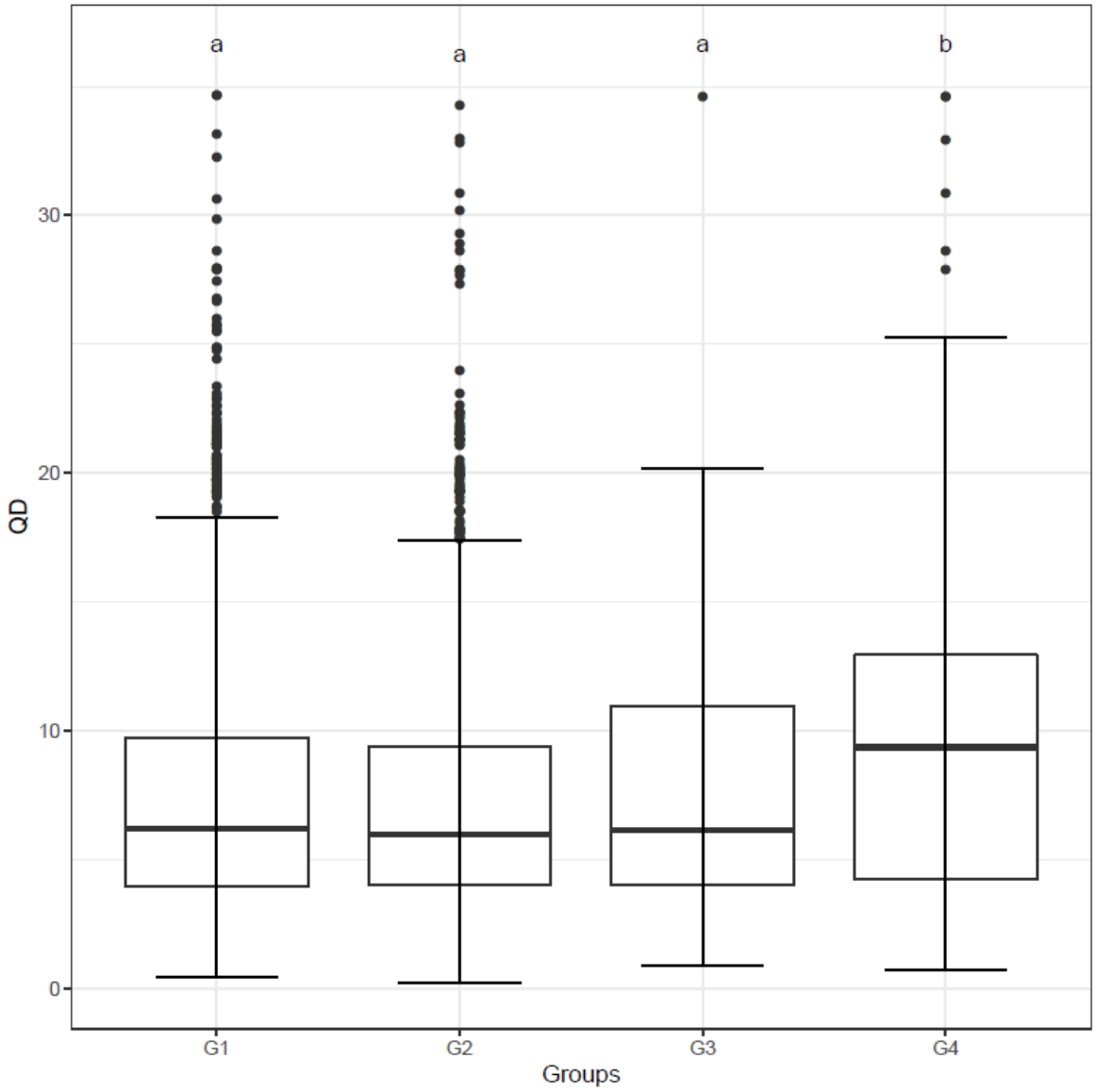


Figura 19: Teste *post-hoc* de *Dunn* da métrica QD entre os grupos G1, G2, G3 e G4. Cada *boxplot* na figura representa um grupo que está indicado no eixo X. O eixo Y são os valores da métrica QD, as letras no topo do *boxplot* indicam se ocorreu diferença significativa, sendo que mesma letra indica não ter diferença entre os grupos.

Este teste estatístico para verificar diferenças significativas na mediana, também foi utilizado para comparar os mesmos grupos, porém desta vez, para a métrica VQSLOD. Então, o objetivo é verificar se existe diferença entre as medianas dos grupos G1, G2, G3 e G4, ao comparar os valores das medianas dos grupos com os dados da métrica VQSLOD.

O resultado dessa comparação está ilustrado na Figura 20. Foram comparados todos grupos contra todos e a indicação de diferença significativa pode ser vista no cabeçalho do *boxplot*. Nota-se que o grupo G1 foi indicado com a letra (a), o G2 com a letra (b), já o G3 e G4 com a letra (c). Então, o G1 tem diferença significativa na mediana comparado com os outros grupos e o G2 também possui diferença significativa comparado com os outros grupos. Já o G3 e G4 não possuem diferença significativa na mediana comparados entre si. Porém são diferentes comparados ao G1 e G2.

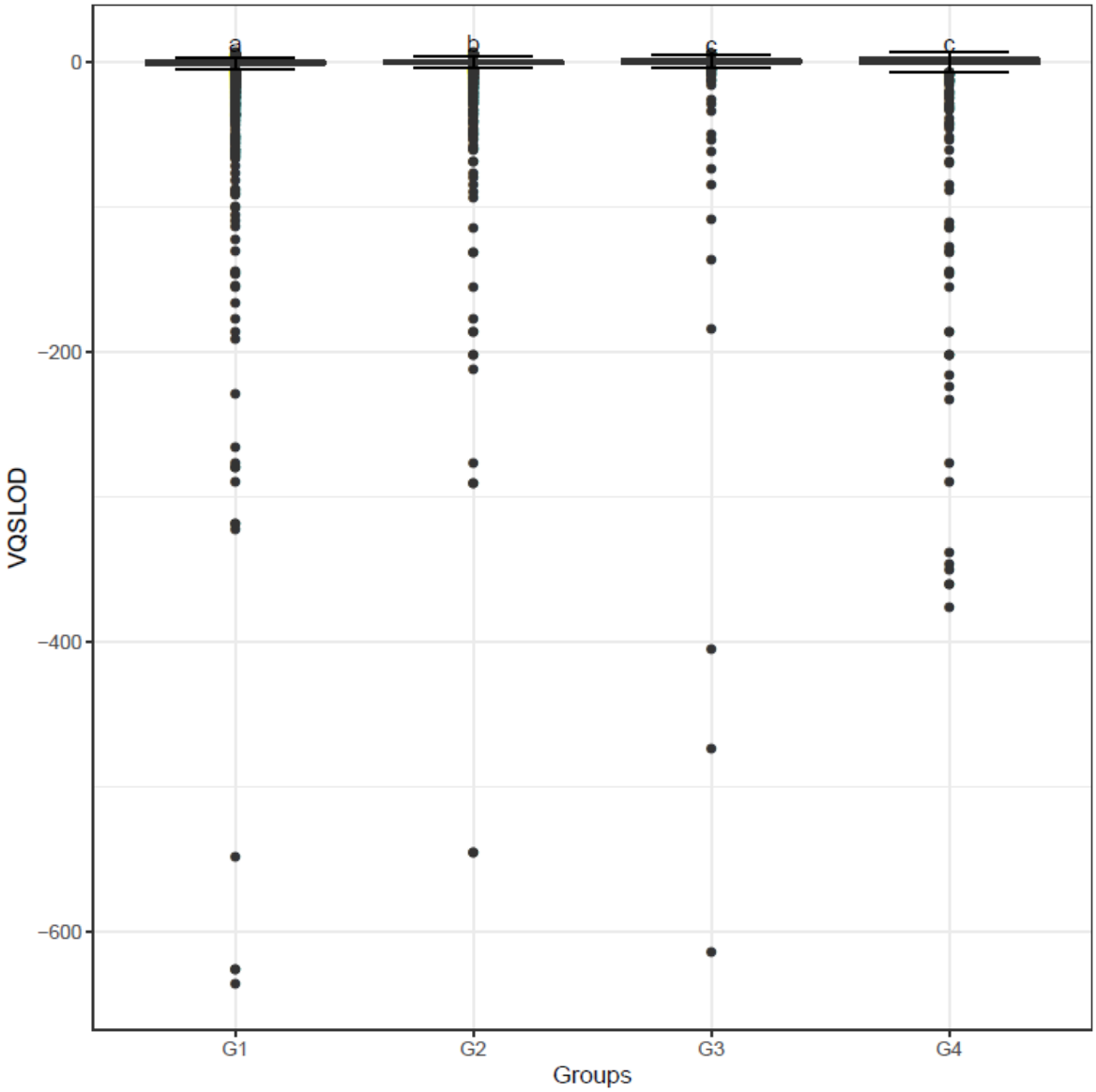


Figura 20: Teste *post-hoc* de *Dunn* da métrica VQSLOD entre os grupos G1, G2, G3 e G4. Cada *boxplot* na figura representa um grupo que está indicado no eixo X. O eixo Y são os valores da métrica VQSLOD, as letras no topo do *boxplot* indicam se ocorreu diferença significativa, mesma letra não tem diferença entre os grupos.

4.9 Resultado da busca de variantes nos segmentos gênicos IGHD

Após rodar o programa para filtrar as variantes do banco de dados gnomAD, com as configurações para o gene D, foram recuperadas 524 variantes de IGHD que estão presentes nos genomas e exomas do banco de dados gnomAD, conforme apresentado na Tabela 9. Deste total de variantes, 46% (241) foram encontradas em apenas um genoma ou exoma, 38% (199) das variantes putativas estão presente em 2 a 6 alelos diferentes, 7% (37) aparecem em 7 a 18 e 9% (47) são encontradas em mais de 18 alelos diferentes.

Em uma consulta ao banco de dados IMGT(GIUDICELLI; CHAUME; LEFRANC, 2005) versão 3.1.36 em 08 de junho de 2022, por genes D funcionais, foram retornados pelo banco de dados 23 genes D funcionais e 30 alelos.

Tabela 9: Número de vezes que cada variante putativa do gene D foi identificada em genoma ou exoma (contagem de alelos)

Groups Allele Count gene D	
Allele Count	**Number of Variants**
1	**241**
2-6	**199**
7-18	**37**
>18	**47**
Total	**524**

Um outro resultado que o *pipeline* disponibiliza é o agrupamento das variantes em cada segmento gênico. Esta informação pode ser observada na Figura 21, onde é apresentado um gráfico com a quantidade de variantes separadas por segmento gênico

IGHD. É importante observar que, dos 23 segmentos gênicos presentes no banco de dados IMGT, foram identificadas variantes para 22 dos segmentos, o único gene que não foi encontrada nenhuma variante foi o IGHD7-27.

A maioria das variantes estão presentes no gene IGHD2-2 com 8,3% (55), já o segmento com a menor quantidade de variantes é o IGHD1-26 com 0,8% (5). Os genes estão ordenados na ordem crescente da quantidade de variantes.

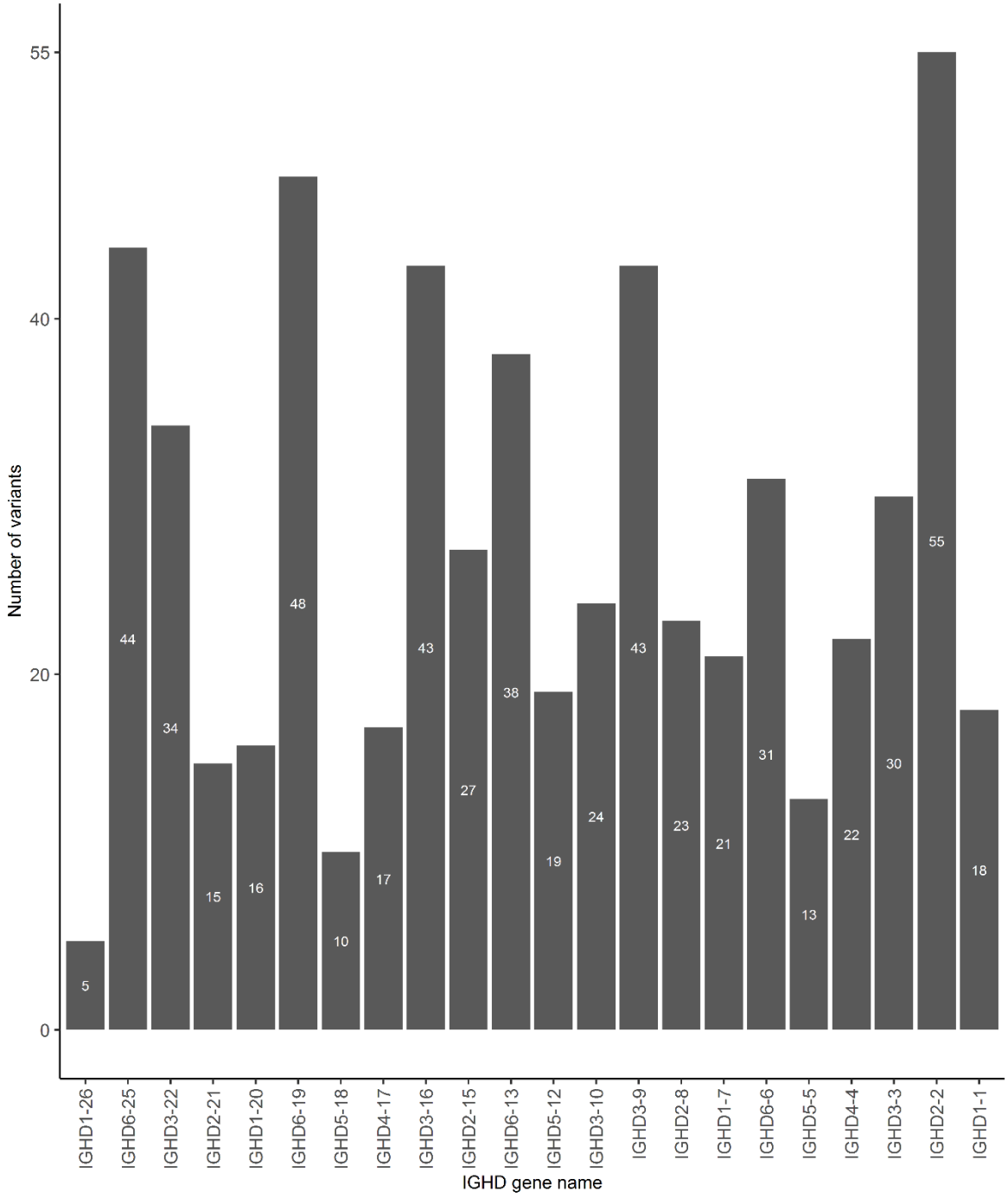


Figura 21: Quantidade de variantes por gene IGHD. A figura apresenta a quantidade de variantes para cada segmento gênico IGHD. No eixo X estão representados os nomes dos genes e no eixo Y a quantidade de variantes de cada segmento gênico. Os dados do eixo X estão ordenados no sentido 5'- 3', complementar reverso, sendo que o segmento gênico IGHD1-1 se encontra próximo à região telomérica.

4.10 Resultado da busca de variantes nos segmentos gênicos IGHJ

Ao executar o programa com as configurações para o gene J, foram recuperadas 670 variantes de IGHJ que estão presentes nos genomas e exomas do banco de dados gnomAD. A quantidade de variantes do segmento gênico J, separadas por grupo pode ser observada na Tabela 10. Destas variantes, 28% (187) foram encontradas em apenas um genoma ou exoma, 46,6% (312) das variantes putativas estão presente em 2 a 6 alelos diferentes, 18,5% (123) aparece em 7 a 18 e 7,2% (48) são encontradas em mais de 18 alelos diferentes.

Para o gene J, também foi verificada a quantidade de genes e alelos presentes no banco de dados IMGT(GIUDICELLI; CHAUME; LEFRANC, 2005) versão 3.1.36. A consulta a este banco foi realizada em 08 de junho de 2022 e foram retornados 06 genes J funcionais e 13 alelos.

Tabela 10: Número de vezes que cada variante putativa do gene IGHJ foi identificada em genoma ou exoma (contagem de alelos)

Groups Allele Count gene J	
Allele Count	**Number of Variants**
1	**187**
2-6	**312**
7-18	**123**
>18	**48**
Total	**670**

Foram identificadas variantes em todos os 6 genes IGHJ presentes no IMGT (Figura 22). Nota-se que a maioria das variantes estão presentes no gene IGHJ5 28,5% (190), já o segmento com a menor quantidade de variantes é o IGHJ1 8,7% (58).

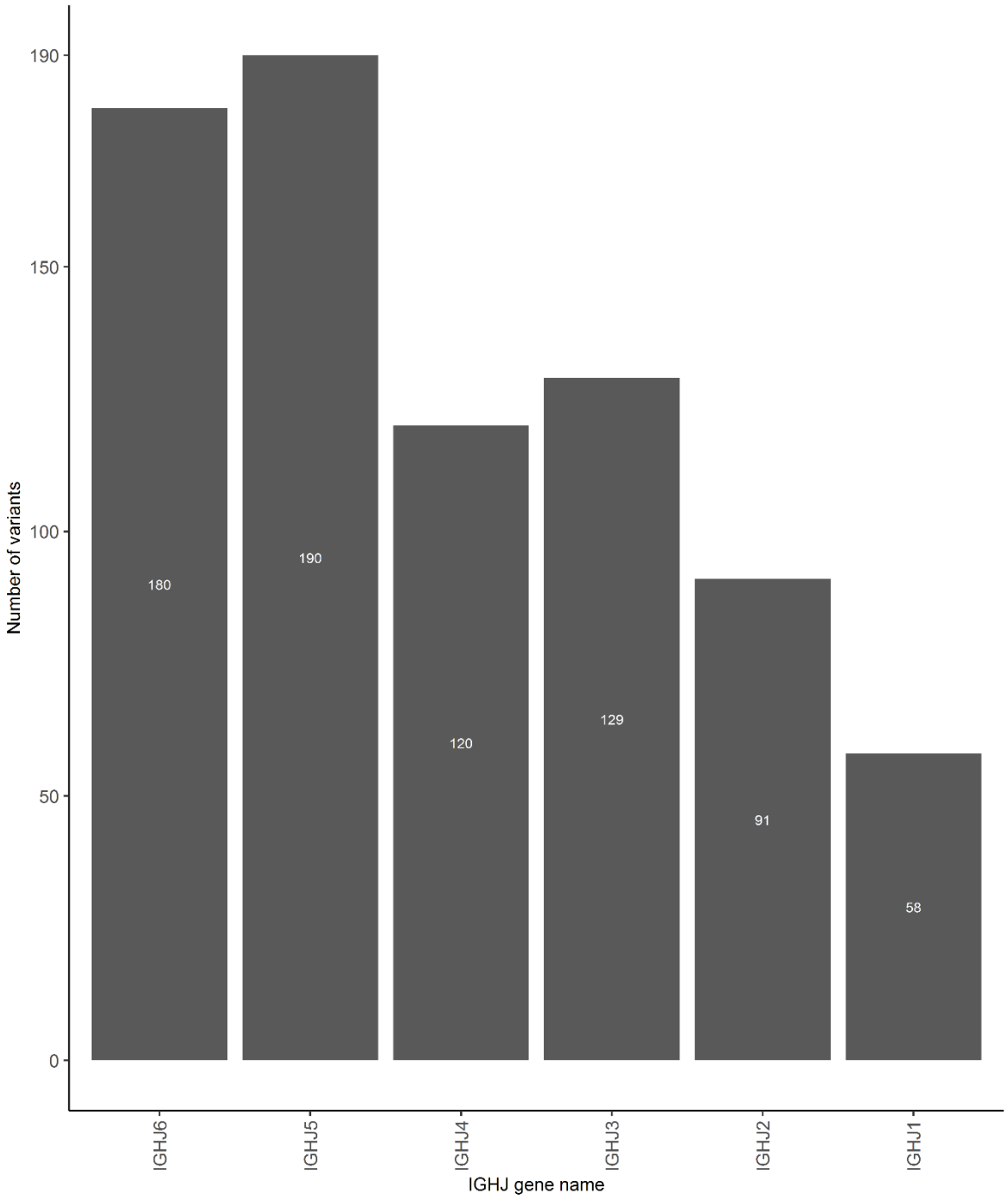


Figura 22: Quantidade de variantes por gene IGHJ. A figura apresenta a quantidade de variantes identificada para cada gene J. No eixo X estão representados os nomes dos genes. No eixo Y é a quantidade de variantes de cada segmento gênico. Os genes estão ordenados na ordem crescente da quantidade de variantes. Os dados

do eixo X estão ordenados no sentido 5'- 3', complementar reverso, sendo que o segmento gênico IGHJ1-1 se encontra próximo à região telomérica.

5 DISCUSSÃO

Dados de sequenciamento genômico do loci de imunoglobulinas sugerem que IGH é uma das regiões mais polimórficas no genoma humano (WATSON et al., 2013). Nosso conhecimento sobre essa diversidade é limitado, conforme indicado pelos 319 alelos IGHV descritos no IMGT/GENE-DB (GIUDICELLI; CHAUME; LEFRANC, 2005), versão 3.1.35.

Mais evidências indicam que essa diversidade é subestimada, incompleta e inconsistente (GADALA-MARIA et al., 2015; WANG et al., 2011). Um estudo em 28 indígenas sul-africanos identificou 122 alelos não reportados pelo IMGT em segmentos gênicos de IGHV (SCHEEPERS et al., 2015). A descoberta de novos alelos a partir de dados de repertório de 7 indivíduos identificou 11 novos alelos de IGHV (GADALA-MARIA et al., 2015), e esse número aumentou para 28 novos alelos de IGHV quando 26 indivíduos foram analisados (GADALA-MARIA et al., 2019). Usando alelos germinativos de 2.504 indivíduos, (KHATRI et al., 2021; YU; CEREDIG; SEOIGHE, 2017) descreveram 3.609 de IGH e IGL e 410 novos putativos alelos de IGHV, respectivamente.

Embora os desafios técnicos para detectar variantes em um lócus que possui variantes estruturais complexas, como o IGH (WATSON et al., 2017) usando *short-reads* sejam inegáveis, a evolução do sequenciamento, montagem e análise de genomas e exomas permitiu o desenvolvimento de bancos de dados de alta qualidade, como o gnomAD (KARCZEWSKI et al., 2020). E ainda, a busca de novas variantes em dados de WGS, embora limitada, pode ser validada posteriormente em laboratório utilizando PCR *target* específica e sequências de Sanger direcionado que permite uma melhor cobertura da região e evita erros de montagem e artefatos de sequenciamento.

Neste trabalho, usando informações do gnomAD de 125.748 exomas, identificamos 10.550 variantes putativas de IGHV, o número mais significativo de variantes de IGHV

descritas atualmente. A maioria dessas variantes estão presentes em apenas um exoma sequenciado (4.949) e pode representar, para alguns achados, um resultado de artefatos de sequenciamento.

No entanto, o gnomAD usa filtros muito rigorosos, para selecionar *reads* com parâmetros de qualidade *Phred* acima de 20 (uso de *reads* com apenas 1% de chance de erro) e as variantes cobertas por pelo menos 10 *reads*, isso minimiza o número de variantes falso-positivas no conjunto de dados (GUDMUNDSSON et al., 2021). Além disso, todas as variantes selecionadas apresentam um *score* de confiança de alta qualidade, VSQLOD (uma métrica que é a probabilidade logarítmica da razão de chances da variante ser positiva ou negativa - *log odds ratio*) e passaram por uma *Random Forest* para distinguir variantes genéticas verdadeiras de artefatos.

Além disso, 81 dessas variantes encontradas em apenas um exoma foram reportadas na mesma posição das variantes encontradas em outros bancos de dados (67 no IMGT, 13 no IgPDB e 1 em ambos). Nota-se que, 5.601 das variantes descritas foram encontradas em 2 ou mais alelos disponíveis nos dados do gnomAD, e 797 delas foram encontradas em mais de 18 alelos, nos quais 575 deles não foram descritos anteriormente nas bases de dados analisadas. Como essas variantes foram encontradas em muitos exomas, a probabilidade dessa variante ser um falso positivo é mínima.

Interessantemente, foi observado que a maioria das variantes são população-específicas e apareceram em baixas frequências, indicando uma possível diversidade individual ou poucos IGHVs únicos de compartilhamento de indivíduos. Por exemplo, se analisarmos as 3.349 variantes específicas da população europeia de IGHV, a mais frequente está presente em apenas 24 alelos dos dados do gnomAD v2.1. Essa pode ser uma das razões para as respostas de diferentes indivíduos a doenças infecciosas, como as causadas pelo HIV, influenza e SARS-COV2.

No entanto, o conhecimento sobre variantes de segmentos gênicos de imunoglobulinas é muito escasso, e ainda mais rara é sua associação com suscetibilidade a doenças. Foram encontradas apenas 14 variantes em segmentos gênicos funcionais de

IGHV relacionados a doenças/características no catálogo GWAS (BUNIELLO et al., 2019).

Algumas variantes do gene IGHV4-61 estão associadas a um risco aumentado de doença cardíaca reumática (MUHAMED; PARKS; SLIWA, 2020). Outras variantes encontradas têm efeitos inesperados, *in trans*, sobre os níveis de proteínas do sangue, como o nível da subunidade beta da ATPase transportadora de potássio e o nível de beta-defensina 119 (SUN et al., 2019). Essa associação entre variantes de IGHV e níveis de proteínas envolvidas na neurotransmissão e na resposta imune inata é fascinante e abre novas perspectivas para o impacto e importância das variantes de IGHV.

Além disso, cada vez mais evidências associam as variantes do IGHV à suscetibilidade a doenças, como esclerose múltipla (IGHV2), artrite reumatoide (IGHV3-30, IGHV4-31, IGHV1-69), lúpus eritematoso sistêmico (IGHV3-30, IGHV4- 31), diabetes tipo 1 (IGHV2, IGHV4, IGHV5) e doença de *Kawasaki* (IGHV1-69, IGHV2-70), conforme revisado por (MARIE J. KIDD ET AL., 2012). Mas também a resposta a doenças infecciosas como gripe (AVNIR et al., 2016) e AIDS (YACOOB et al., 2016).

Curiosamente, os genes IGHV observados com o maior número de variantes (por exemplo, IGHV2-70, IGHV1-69) são encontrados com 1-4 cópias por indivíduo, o que pode impactar esse maior número de variantes (WATSON et al., 2013).

A maioria das variantes foram descritas nas regiões de FWR, especialmente na FWR3. Mutações nessas regiões podem afetar as propriedades biofísicas do anticorpo, como reconhecimento de antígeno, rendimento de expressão do anticorpo, pH e estabilidade térmica (CNUDDE et al., 2020). Enquanto os FWRs ajudam a estabilizar o sítio de ligação ao antígeno e definem as conformações dos loops CDRs, os papéis das mutações/polimorfismos nos FWRs não são bem compreendidos. Além dessas propriedades biofísicas, as variantes genéticas do IGHV também podem estar associadas à variação no uso de genes observada entre os indivíduos (KENTER; WATSON; SPILLE, 2021).

O escopo do conjunto de dados de variantes fornecido por este trabalho tem algumas restrições: primeiro, o genoma de referência GENCODE GRCh37 inclui apenas

uma parte dos genes IGHV conhecidos (40 de 55) e famílias (6 de 7 - ver Tabela 5); segundo, a informação genômica em nível individual que constrói o banco de dados gnomAD não é acessível, o que prejudica o genótipo individual; terceiro, a análise foi restrita às sequências V-REGION, estreitando a análise da funcionalidade do alelo para verificar a presença de *stop codons* antes do último códon; quarto, os tipos de células utilizados para todo o genoma e sequenciamento de exoma para cada indivíduo no conjunto de dados são desconhecidos.

Apenas dois estudos de associação genômica ampla (GWAS) identificaram uma associação com genes funcionais de IGHV, sugerindo uma desconexão potencial entre a diversidade de haplótipos de IGHV conhecidos e as ferramentas atuais de genotipagem de alto rendimento. Portanto, é fornecido o acesso para todas as variantes encontradas neste trabalho (correspondentes a 1.262 rsIDs) no banco de dados YVr-DB.

É importante observar que, a expansão do banco de dados de variantes germinativas produzida por este trabalho pode: 1) beneficiar os softwares de análise de genes IGHV, superando problemas relacionados à incompletude do banco de dados de referência ou ajudar a validar a descoberta de novos alelos; 2) cooperar na compreensão das variantes individuais ou únicas da população (incluindo africanos, latinos e asiáticos sub-representados) e suas relações com suscetibilidades a doenças; 3) permitir a identificação de sequências de consenso completas entre diferentes alelos para auxiliar no desenho de *primers* para estudos de repertório facilitando a validação destas variantes; 4) habilitar o *link* SNP IGHV com GWAS.

É importante destacar que, como apresentou o trabalho de (WENGER et al., 2019), onde a chamada de variante do GATK em dados de *short reads* alcançou uma acurácia superior aos dados de *long reads*, o uso de dados de *short reads* para identificar variantes genéticas de imunoglobulina apresenta grande potencial.

6 CONCLUSÕES

Em conclusão, este trabalho descreve uma coleção de 10.505 variantes de genes IGHV, sendo 46,9% (4.949) presente em apenas um exoma. Do total de variantes, 10.156 são novas, ou seja, não identificadas em outros bancos de dados. Este é o conjunto mais abrangente de variantes putativas de IGHV disponível atualmente.

Dessa forma, mesmo considerando aquelas variantes identificadas em mais de 18 alelos sequenciados, fica evidente a alta variabilidade da região IGHV e até mesmo a possível diversidade no nível individual. Além disso, também é apresentado neste trabalho a identificação de 524 variantes putativas do gene IGHD e 670 variantes do gene IGHJ.

Essa diversidade pode afetar a suscetibilidade a doenças, alterar o nível de proteínas não relacionadas ao IGHV no plasma sanguíneo, provavelmente afetar o uso do gene IGHV e a produção de anticorpos secretados.

Todas as variantes IGHV identificas neste trabalho para o segmento gênico IGHV podem ser acessadas pela plataforma *web* YVr-DB, que disponibiliza diversos filtros para auxiliar na investigação do pesquisador.

7 PERSPECTIVAS

- Analisar com mais detalhes, as variantes dos segmentos gênicos D e J, obtidos através de uma nova metodologia;
- Ajustar os *scripts* e recuperar as variantes da versão do genoma de referência CRCh38 e T2T-CHM3;
- Atualizar a Plataforma *web*, para possibilitar filtrar as variantes dos genes IGHD e IGHJ;
- Validar experimentalmente, por sequenciamento Sanger a variante 14-107083516-G-A do gene IGHV4-59. Esta variante não foi reportada pelos bancos de dados IMGT e IgPDB, porém foi identificada no sequenciamento WES de dois pacientes com hemofilia.

REFERÊNCIAS

1000 GENOMES PROJECT, C. et al. A global reference for human genetic variation. **Nature**, p. 68–74, 2015.

AVNIR, Y. et al. IGHV1-69 polymorphism modulates anti-influenza antibody repertoires, correlates with IGHV utilization shifts and varies by ethnicity. **Scientific Reports**, v. 6, n. November 2015, 2016.

BUNIELLO, A. et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. **Nucleic Acids Research**, v. 47, n. D1, p. D1005–D1012, 2019.

BUNN, A.; KORPELA, M. A language and environment for statistical computing. **Foundation for Statistical Computing**, v. 2, p. 1–12, 2013.

CALIS, J. J. A.; ROSENBERG, B. R. Characterizing immune repertoires by high throughput sequencing: strategies and applications. **Trends Immunol**, p. 581–590, 2014.

CHAUDHARY, N.; WESEMANN, D. R. Analyzing immunoglobulin repertoires. **Frontiers in Immunology**, v. 9, n. MAR, p. 1–18, 2018.

CHURCH, D. M. et al. Modernizing reference genome assemblies. **PLoS Biology**, v. 9, n. 7, p. 1–5, 2011.

CNUDDE, T. et al. Exploration and Modulation of Antibody Fragment Biophysical Properties by Replacing the Framework Region Sequences. **Antibodies**, v. 9, n. 2, p. 9, 2020.

COCK, P. J. A. et al. Biopython: Freely available Python tools for computational molecular biology and bioinformatics. **Bioinformatics**, v. 25, n. 11, p. 1422–1423, 2009.

COLLINS, A. M. et al. The mouse antibody heavy chain repertoire is germline-focused and highly variable between inbred strains. **Philosophical Transactions of the Royal Society B: Biological Sciences**, v. 370, n. 1676, 2015.

CORCORAN, M. M. et al. Production of individualized v gene databases reveals high levels of immunoglobulin genetic diversity. **Nature Communications**, v. 7, 2016.

EBERT, P. et al. Haplotype-resolved diverse human genomes and integrated analysis of structural variation. **Science**, v. 372, 2021.

EDGAR, R. C. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. **Nucleic Acids Research**, v. 32, n. 5, p. 1792–1797, 2004.

ELMASRI, R.; NAVATHE, S. B. **Sistemas de banco de dados**. 6. ed. [s.l: s.n.].

FORD, M. et al. Genotyping and Copy Number Analysis of Immunoglobin Heavy Chain Variable Genes Using Long Reads. **iScience**, v. 23, n. 3, 2020.

FRANKISH, A. et al. GENCODE reference annotation for the human and mouse genomes. **Nucleic Acids Research**, v. 47, n. D1, p. D766–D773, 2019.

GADALA-MARIA, D. et al. Automated analysis of high-throughput B-cell sequencing data reveals a high frequency of novel immunoglobulin V gene segment alleles. **Proceedings of the National Academy of Sciences of the United States of America**, v. 112, n. 8, p. E862–E870, 2015.

GADALA-MARIA, D. et al. Identification of subject-specific immunoglobulin alleles from expressed repertoire sequencing data. **Frontiers in Immunology**, v. 10, n. FEB, p. 1–12, 2019.

GEORGIOU, G. et al. The promise and challenge of high-throughput sequencing of the antibody repertoire. **Nature Biotechnology**, v. 32, n. 2, p. 158–168, 2014.

GIUDICELLI, V. et al. IG and TR single chain fragment variable (scFv) sequence analysis : a new advanced functionality of IMGT / V-QUEST and IMGT /. p. 1–13, 2017.

GIUDICELLI, V.; CHAUME, D.; LEFRANC, M. P. IMGT/V-QUEST, an integrated software program for immunoglobulin and T cell receptor V-J and V-D-J rearrangement analysis. **Nucleic Acids Research**, v. 32, n. WEB SERVER ISS., p. 435–440, 2004.

GIUDICELLI, V.; CHAUME, D.; LEFRANC, M. P. IMGT/GENE-DB: A comprehensive database for human and mouse immunoglobulin and T cell receptor genes. **Nucleic Acids Research**, v. 33, n. DATABASE ISS., p. 256–261, 2005.

GUDMUNDSSON, S. et al. Variant interpretation using population databases: Lessons from gnomAD. **Human Mutation**, n. November 2021, 2021.

GUPTA, S. K.; VISWANATHA, D. S.; PATEL, K. P. Evaluation of Somatic Hypermutation Status in Chronic Lymphocytic Leukemia (CLL) in the Era of Next Generation Sequencing. **Frontiers in Cell and Developmental Biology**, v. 8, n. May, p. 1–12, 2020.

KARCZEWSKI, K. J. et al. The mutational constraint spectrum quantified from variation in 141,456 humans. **Nature**, v. 581, n. 7809, p. 434–443, 2020.

KATOH, K. et al. MAFFT: A novel method for rapid multiple sequence alignment based

on fast Fourier transform. **Nucleic Acids Research**, v. 30, n. 14, p. 3059–3066, 2002.

KENTER, A. L.; WATSON, C. T.; SPILLE, J.-H. Igh Locus Polymorphism May Dictate Topological Chromatin Conformation and V Gene Usage in the Ig Repertoire. **Frontiers in Immunology**, v. 12, n. May, p. 1–7, 18 maio 2021.

KHATRI, I. et al. Population matched (PM) germline allelic variants of immunoglobulin (IG) loci: New pmIG database to better understand IG repertoire and selection processes in disease and vaccination. **bioRxiv**, 2020.

KHATRI, I. et al. Population matched (pm) germline allelic variants of immunoglobulin (IG) loci: Relevance in infectious diseases and vaccination studies in human populations. **Genes and Immunity**, v. 22, n. 3, p. 172–186, 2021.

KUMAR, K. R.; COWLEY, M. J.; DAVIS, R. L. Next-Generation Sequencing and Emerging Technologies. **Seminars in Thrombosis and Hemostasis**, v. 45, n. 7, p. 661–673, 2019.

LANDRUM, M. J. et al. ClinVar: Improving access to variant interpretations and supporting evidence. **Nucleic Acids Research**, v. 46, n. D1, p. D1062–D1067, 2018.

LEES, W. et al. OGRDB: A reference database of inferred immune receptor genes. **Nucleic Acids Research**, v. 48, n. D1, p. D964–D970, 2020.

LEES, W. D.; SHEPHERD, A. J. Studying Antibody Repertoires with Next-Generation Sequencing. **Bioinformatics. Methods in Molecular Biology**, v. 1526, p. 257–270, 2017.

LEFRANC, M. IMGT Locus in Focus Nomenclature of the Human Immunoglobulin Heavy (IGH) Genes. **Exp Clin Immunogenet**, v. 18, p. 100–116, 2001.

LEFRANC, M. P.; LEFRANC, G. Immunoglobulins or antibodies: IMGT® bridging genes, structures and functions. **Biomedicines**, v. 8, n. 9, 2020.

LI, H. et al. The Sequence Alignment/Map format and SAMtools. **Bioinformatics**, v. 25, n. 16, p. 2078–2079, 2009.

LI, H. Tabix: Fast retrieval of sequence features from generic TAB-delimited files. **Bioinformatics**, v. 27, n. 5, p. 718–719, 2011.

MARIE J. KIDD ET AL. Chain V Region Gene Loci by Analysis of VDJ Gene. **J Immunol**, v. 188, n. 3, p. 1333–1340, 2012.

MARKS, C.; DEANE, C. M. How repertoire data are changing antibody science.

Journal of Biological Chemistry, v. 295, n. 29, p. 9823–9837, 2020.

MATTHEWS AJ, ZHENG S, DIMENNA LJ, C. J. Regulation of Immunoglobulin Class-Switch Recombination: Choreography of Noncoding Transcription, Targeted DNA Deamination, and Long-Range DNA Repair. **Adv Immunol**, p. 1–57, 2014.

MCKENNA, A. et al. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. **Genome Res**, p. 1297–303, 2010.

MIKOCZIOVA, I. et al. Polymorphisms in human immunoglobulin heavy chain variable genes and their upstream regions. **Nucleic Acids Research**, v. 48, n. 10, p. 5499–5510, 2020.

MUHAMED, B.; PARKS, T.; SLIWA, K. Genetics of rheumatic fever and rheumatic heart disease. **Nature Reviews Cardiology**, v. 17, n. 3, p. 145–154, 2020.

MURPHY, K. **Imunobiologia de Janeway**. ArtMed ed. Porto Alegre: [s.n.].

NURK, S. et al. The complete sequence of a human genome. **Science**, 2022.

PAROLA, C.; NEUMEIER, D.; REDDY, S. T. Integrating high-throughput screening and sequencing for monoclonal antibody discovery and engineering. **Immunology**, v. 153, n. 1, p. 31–41, 2018.

POPLIN, R. et al. A universal snp and small-indel variant caller using deep neural networks. **Nature Biotechnology**, v. 36, n. 10, p. 983, 2018.

PRAMANIK, S. et al. Segmental duplication as one of the driving forces underlying the diversity of the human immunoglobulin heavy chain variable gene region. **BMC Genomics**, v. 12, n. 1, p. 78, 2011.

RECALDIN, T.; FEAR, D. J. Transcription factors regulating B cell fate in the germinal centre. **Clinical and Experimental Immunology**, v. 183, n. 1, p. 65–75, 2016.

RETTER, I. et al. VBASE2, an integrative V gene database. **Nucleic Acids Research**, v. 33, p. 671–674, 2005.

RODRÍGUEZ-CABALLERO, A. et al. The Hydropathy Index of the HCDR3 Region of the B-Cell Receptor Identifies Two Subgroups of IGHV-Mutated Chronic Lymphocytic Leukemia Patients With Distinct Outcome. **Frontiers in Oncology**, v. 11, n. October, p. 1–12, 2021.

RODRIGUEZ, O. L. et al. A Novel Framework for Characterizing Genomic Haplotype Diversity in the Human Immunoglobulin Heavy Chain Locus. **Frontiers in**

Immunology, v. 11, n. September, p. 1–16, 2020.

ROUET, R. et al. Next-generation sequencing of antibody display repertoires. **Frontiers in Immunology**, v. 9, n. FEB, p. 1–5, 2018.

RSTUDIO TEAM. **RStudio: Integrated Development Environment for R**. Disponível em: <http://www.rstudio.com/>.

SCHATZ, D. G.; SWANSON, P. C. V(D)J recombination: Mechanisms of initiation. **Annual Review of Genetics**, v. 45, n. D, p. 167–202, 2011.

SCHEEPERS, C. et al. Ability To Develop Broadly Neutralizing HIV-1 Antibodies Is Not Restricted by the Germline Ig Gene Repertoire. **The Journal of Immunology**, v. 194, n. 9, p. 4371–4378, 2015.

SCHEIJEN, B. et al. Next-generation sequencing of immunoglobulin gene rearrangements for clonality assessment: a technical feasibility study by EuroClonality-

NGS. **Leukemia**, v. 33, n. 9, p. 2227–2240, 2019.

SHERRY, S. T. et al. DbSNP: The NCBI database of genetic variation. **Nucleic Acids Research**, v. 29, n. 1, p. 308–311, 2001.

SUN, B. B. et al. Europe PMC Funders Group Europe PMC Funders Author Manuscripts Europe PMC Funders Author Manuscripts Genomic atlas of the human plasma proteome Europe PMC Funders Author Manuscripts Europe PMC Funders Author Manuscripts. v. 558, n. 7708, p. 73–79, 2019.

TREPEL, F. Number and distribution of lymphocytes in man. A critical analysis. **Klinische Wochenschrift**, v. 52, n. 11, p. 511–515, 1974.

TUNE, C. et al. Sleep restriction prior to antigen exposure does not alter the T cell receptor repertoire but impairs germinal center formation during a T cell-dependent B cell response in murine spleen. **Brain, Behavior, & Immunity - Health**, v. 16, n. July, p. 100312, 2021.

VAN DER AUWERA, G. et al. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. **Curr Protoc Bioinformatics**, 2014.

WANG, Y. et al. Many human immunoglobulin heavy-chain IGHV gene polymorphisms have been reported in error. **Immunology and Cell Biology**, v. 86, n. 2, p. 111–115, 2008.

WANG, Y. et al. Genomic screening by 454 pyrosequencing identifies a new human

IGHV gene and sixteen other new IGHV allelic variants. **Immunogenetics**, v. 63, n. 5, p. 259–265, 2011.

WATSON, C. T. et al. Complete haplotype sequence of the human immunoglobulin heavy-chain variable, diversity, and joining genes and characterization of allelic and copy-number variation. **American Journal of Human Genetics**, v. 92, n. 4, p. 530–546, 2013.

WATSON, C. T. et al. Sequencing of the human IG light chain loci from a hydatidiform mole BAC library reveals locus-specific signatures of genetic diversity. **Genes and Immunity**, v. 16, n. 1, p. 24–34, 2015.

WATSON, C. T. et al. Comment on "A Database of Human Immune Receptor Alleles Recovered from Population Sequencing Data". **The Journal of Immunology**, v. 198, n. 9, p. 3371–3373, 2017.

WATSON, C. T.; BREDEN, F. The immunoglobulin heavy chain locus: Genetic variation, missing data, and implications for human disease. **Genes and Immunity**, v. 13, n. 5, p. 363–373, 2012.

WENDEL, B. S. et al. A streamlined approach to antibody novel germline allele prediction and validation. **Frontiers in Immunology**, v. 8, n. SEP, 2017.

WENGER, A. M. et al. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. **Nature Biotechnology**, v. 37, n. 10, p. 1155–1162, 2019.

WESTERMANN, J.; PABST, R. Distribution of lymphocyte subsets and natural killer cells in the human body. **The Clinical Investigator**, v. 70, n. 7, p. 539–544, 1992.

XOCHELLI, A. et al. Immunoglobulin heavy variable (IGHV) genes and alleles: new entities, new names and implications for research and prognostication in chronic lymphocytic leukaemia. **Immunogenetics**, v. 67, n. 1, p. 61–66, 2015.

YACOOB, C. et al. Differences in Allelic Frequency and CDRH3 Region Limit the Engagement of HIV Env Immunogens by Putative VRC01 Neutralizing Antibody Precursors. **Cell Reports**, v. 17, n. 6, p. 1560–1570, 2016.

YE, J. et al. IgBLAST: an immunoglobulin variable domain sequence analysis tool. **Nucleic acids research**, v. 41, n. Web Server issue, p. 34–40, 2013.

YU, Y.; CEREDIG, R.; SEOIGHE, C. A Database of Human Immune Receptor Alleles Recovered from Population Sequencing Data. **The Journal of Immunology**, v. 198, n. 5, p. 2202–2210, 2017.

ZOOK, J. M. et al. An open resource for accurately benchmarking small variant and reference calls. **Nature Biotechnology**, v. 37, n. 5, p. 561–566, 2019.

ANEXO I

A seguir o artigo científico elaborado durante o período do doutorado, este artigo está depositado no bioRxiv doi: https://doi.org/10.1101/2021.01.15.426262

Description of 10,909 putative human immunoglobulin heavy chain IGHV variants

Fabio R Martins[1,2], Lucas Alves de Melo Pontes[1], Tiago Antônio de Oliveira Mendes[3], Liza F. Felicori[1*]

[1]Laboratory of Synthetic Biology and Biomimetics, Departamento de Bioquímica e Imunologia, Universidade Federal de Minas Gerais, Belo Horizonte, MG, Brazil

[2]Departamento de Informática, Instituto Federal de Minas Gerais (IFMG) - Campus São João Evangelista, São João Evangelista, MG, Brazil

[3]Departamento de Bioquímica e Biologia molecular, Universidade Federal de Viçosa, MG, Brazil

Correspondence:

*Liza F. Felicori

Email: liza@icb.ufmg.br

Keywords: immunoglobulin variants, gnomAD, antibody genes, GWAS

Abstract

The identification of immunoglobulin locus diversity can assist the understanding of the immune system, human diseases, and the development of new therapies. The advent of massively parallel sequencing enabled the development of tools to map and identify new immunoglobulin variants, mainly using the 1,000 Genomes (G1K) data. However, G1K data present caveats, such as low coverage, inaccurate variant call, and a limited sample size of 2,504 individuals. To overcome these limitations, we performed a computational analysis of IGHV variants in gnomAD, the most comprehensive high-quality catalog of variation from human exomes and genomes.

From 10,909, a total of 10,550 putative high-quality IGHV variants were identified, evidencing the impressive polymorphic characteristics of this region. Interestingly, 1,366 new variants were found in more than seven alleles, and 575 of them in more than 18 alleles sequenced available on gnomAD. The IGHV2-70 was the gene segment with the highest number of variants described. Most of the variants were i) missense mutations, ii) located in framework 3, and ii) population or even individual-specific. Some of the variants were associated with autoimmune and infectious diseases and correlated, in trans, with the potassium-transporting ATPase subunit beta and beta-defensin 119 proteins in plasma. These variants can link the adaptive and innate immune systems and the neuroimmune systems.

In order to provide storage and accessibility to those new variants, a database named YVr-DB was developed. This available data can shed light on the role of IGHV variants in disease susceptibility and the generation of antibody diversity.

Introduction

According to the international ImMunoGeneTics information system (IMGT), the human immunoglobulin heavy chain (IGH) locus at 14q32.33 comprises mostly 38-46 functional IGHV genes belonging to 7 subgroups, 23 IGHD, 6 IGHJ, and 9 IGHC genes (1).

Variations of these individual immunoglobulin germline genes can influence the B cell receptor repertoire, giving rise to differences in immune response, which may impact disease susceptibility or development. For example, previous studies have shown the association of specific germline genotypes with susceptibility to autoimmune diseases such as rheumatoid arthritis (RA), systemic lupus erythematosus (SLE), and multiple sclerosis (MS) (2–5).

Our understanding of sequence variation for IGHV gene alleles is limited and inaccurate (6). IGHV variant analysis is hampered due to the repetitive nature of locus, the similarity between the genes, the difficulty distinguishing polymorphisms from somatic hypermutations, high fragmentation of genome data available, and finally, IGHV sequence diversity have been based on a restricted ethnic heterogeneity (7,8).

Some methods have been developed over the years to discover new IGHV variants, either utilizing Adaptive Immune Receptor Repertoire sequencing (AIRR-seq) or genomic data (9–14). One of these methods significant limitations is the differentiation between somatic hypermutation and new alleles, which requires further validation. In addition, immunoglobulin new variant discovery from genome studies use 1,000 Genomes (G1K) data, which analyzed 2,504 genomes from 26 populations (15). However, genomes sequenced in the G1K project presented low coverage and variant calls with higher failure rates using alternative variant discovery/genotyping methods (16).

This work used Genome Aggregation Database (gnomAD) data, the most extensive high-quality catalog of variations performed variant call using a standardized BWA-Picard-GATK pipeline. This database initially aggregated whole-exome sequencing data from 199,558 individuals and whole-genome sequencing data from 20,314 individuals. However, after stringent sample quality control, the gnomAD v2.1.1 contains data from

125,748 exomes and 15,708 human genomes, superior to G1K which contained only genome data from 2,504 individuals (17).

Using this high-quality dataset, we described 10,909 putative IGHV variants, most of them from exome data, whose data can be accessed in our database YVr-DB. This database is the most comprehensive collection of putative IGHV variants currently available. This available data can help the scientific community to validate new IGHV variants, including the design of new primers to amplify specific or families of variants, to understand individual differences in response to infectious diseases, to understand better the link between the IGHV polymorphisms and the innate and the nervous systems among several other diseases and systems linking.

Methods

Development of a pipeline for IGHV variants discovery from exome

We aimed to identify possible variable (V) gene variants from the human heavy chain immunoglobulin (IGHV) germline gene using the gnomAD (17). Since gnomAD adopts the GENCODE release 19 for gene annotation and the standard Immunoglobulin annotation uses the IMGT/GENE-DB (18) as a database for germline sequences, a comparison was made to extract the correct corresponding position between both datasets. Briefly, a search in the ".gff" file for IGHV positions was done for both NCBI (RefSeq Reference Genome Annotation from build GRCh37 – ref. (19) that uses IMGT standards) and GENCODE (Comprehensive gene annotation from build GRCh37.p13 – ref. (20)). Since differences were observed between the positions of each V gene in both databases, all IGHV nucleotide sequences from both datasets were recovered from the related ".fa" files using Samtools v1.6 (http://www.htslib.org/). The sequences for all IGHV from both datasets were aligned using the Needleman-Wunsch algorithm through Biopython v1.74 (21). Only the V-REGION (encoded by a V gene: without leader and RSS) was filtered and analyzed. Variant data (sequence and position) were obtained from gnomAD using Selenium library, a web scraping application developed in Java 1.8. To check whether the YVr database variants were present in IMGT (18), IgPDB (22), or both, the position of the suspected polymorphism was inspected in these databases. For that, the nucleotide variant recovered from gnomAD v2.1.1 was replaced in the corresponding V-REGION using a Java 1.8 program. The sequences were complementary reversed after nucleotide deletion, addition, or substitution in V-REGION, as standardized by IMGT/V-Quest (23). This methodology is summarized in Figure 1A.

Sequences from the IMGT/GENE-DB database version 3.1.34 of (31 May 2021) were downloaded, where a query for functional V genes resulted in 57 genes and 311 alleles (of which 3 had no sequences). IGHV sequences from the IgPDB database were also downloaded. Only IGHV alleles "heavy chain V genes" were filtered out, with 194 alleles reported in this database.

A program was created to merge the alleles of the two databases into a family, as an example: all the alleles of the IGHV1-69 gene would be in just one FASTA file. Then MAFFT v7.310 (24) was used to make a multiple alignments of each file. It was manually checked if the *01 alleles of each gene had suffered the addition of a gap. This was the case of the IGHV3-11 gene, 3-11*02 allele of the IMGT, and the *p07 of the IgPDB. These alleles were not taken into account in our analysis.

With the data merged and aligned, for each variant reported by YVr, it was verified in a specific position or in a group of sequential nucleotides if the nucleotide was reported by IMGT/IgPDB as described by (25).

Since individual-level data were not available, it was impossible to infer haplotypes, underestimating our comparison.

Variant vetting criteria

To filter for variants with high quality, an R script was developed in the Rstudio software, which removes the variants from the set based on three criteria: a) variants from whole-genome sequencing due to lower genome coverage compared to exome sequences; b) variants that did not pass the gnomAD Random Forest quality test (we selected only the PASS option already filtered by gnomAD after removing from their dataset 12.2% of SNV and 24.7% of indels variants in exome data) (17) and finally c) variants with insertion or deletion bigger than 50 bp (structural variants). After the veto, the dataset had a total of 10,550 exome-only variants. This dataset was then divided into four groups, based on the number of shared variants between the exomes: group 1 contain variants found in one exome, group 2, variants present in 2 to 6 alleles from exomes, group 3, the ones present

in 7 to 18 alleles from the sequenced exomes, and group 4, variants present in more than 18 alleles sequenced.

Database Generation

A database named YVr-DB was created. The YVr-DB contain IGHV gene identification, chromosome position (based on GRCh37 – ref. (19)), and variant information including nucleotide sequence, amino acid sequence (obtained through IgBLAST (26)), FWR and CDR position, type of mutation classified according to Ensembl (missense, synonymous, frameshift, in-frame deletion or insertion, protein_altering and stop_gained), number of times the variants appear on genome and exomes sequenced in gnomAD, dbSNP and Ensemble IDs (27,28), the presence of the variant in other databases and the occurrence in different populations (Figure 1B e 1C).

The database YVr-DB was created using MySQL version 8.0.20 (https://www.mysql.com/) and a Java 1.8 program in a Linux system (x86_64). For table modeling, MySQL Workbench 8.0.20 was used. YVr-DB is available on: http://bioinfo.icb.ufmg.br/yvr/

Unique and shared population variants

An SQL query was performed directly on the YVr-DB database, where information regarding the occurrence of variations in one population (unique variants) or shared among populations, along with the gene information and the population description, were obtained and stored in two .csv files.

The above files were imported into the R (29) software, Rstudio (30), where unique and shared variant information was merged. The data were grouped by genes/population and plotted using the R package ggplo2 (31).

Disease-associated variants search

A search was carried out in the ClinVar (32) and GWAS Catalog (33) databases to detect the association of the variants found here with diseases. Variants from the ClinVar database were downloaded in VCF format via FTP. The GRCh37 – ref. (19) reference genome was used (https://ftp.ncbi.nlm.nih.gov/pub/clinvar/), and an in-house Java 1.8 program filtered the variants of the immunoglobulin locus, using locus location 106052774-107288051 in the VCF according to NCBI.

In a search on the GWAS website for studies related to immunoglobulin genes, the range 14:106052774-107288051 was used to filter out the variants. The RESTful API (www.ebi.ac.uk/gwas/rest/api) was used to download the data and obtain SNP's in JSON format with the associated rsId dbSNP (27) through Linux CURL command. An in-house Java 1.8 program was used to filter the rsId of each SNP and create a Linux shell script, which downloads each associated study and saves it in a JSON file. Finally, we checked if the rsId from GWAS and ClinVar were present in the YVr-DB database.

Results

1. Large-scale human exome IGHV-gene mining revealed 10,550 putative IGHV variants

Forty functional IGHV genes were found in gnomAD GENCODE annotation (20). After filtering for variants unique to V-REGION, 10,909 variants were found. After that, variants present in genome 325, the ones that did not fill the criteria of GATK quality 28, and structural variants 6 were excluded from our analysis since we considered here only high-quality variants and exome-exclusive variants. After these filters, we describe here 10,550 IGHV variants present in gnomAD exomes (Table 1).

Some of the variants described here were also found in other immunoglobulin germline databases, such as 278 in IMGT/GENE-DB (18), 75 in IgPdb (22), and 41 in both databases, resulting in only 394 already described variants. Most of the new putative

variants described here appear only once in exomes present in gnomAD 4,949, being 81 of them also detected on IMGT and/or IgPDB (Table 1).

Notably, 3,991 new putative variants appear in 2 to 6 different alleles, 813 appeared in 7 to 18, and 797 appear in more than 18 different alleles (Table 1). From the 797 most frequent variants, 575 are not described in the other germline databases (GLDB). At least one variant from each IGHV gene was described in other germline databases such as IMGT and IgPDB (Supplementary Figure 1). Variants were distributed along IGHV locus; however, considerable variability in the number of variants from different IGHV genes is observed, ranging from 97 variants for IGHV1-45 to 765 for IGHV2-70 (Supplementary Figure 1). Besides that, 42.26**%** of the variants (4,459), were found in the largest IGHV subgroup, the IGHV3 (Table 2).

2. The majority of the IGHV variants are in the Framework 3 region

Interestingly, the majority (84,3%) of variants was present on framework regions: 20.9% (2,208) in FWR1, 22% (2,313) in the FWR2, and the large majority in FWR3 (4,374 variants or 41.4%). Some of the variants were also found in the regions corresponding to CDR1 (719 variants, 6.8%), CDR2 (422 variants, 4%), or the beginning of CDR3 (514 variants, 4.8%). No trend was observed for variants in a specific position in V-REGION (Figure 2). The frequency of 1 variant (35,5%), 2 variants (38,37%), or even 3 (26%) variants at the same position is elevated. However, few variants, like those derived from the IGHV4-4 gene, present five different variants at the same position in the FWR3 (Figure 2).

3. Most of the IGHV variants are missense

From 10,550 variants described, the majority are missense (7,081 variants: 67.1%) and synonymous (2,831: 26.8%) single nucleotide variants (SNVs) (Table 1). IGHV2-70, IGHV1-2, and IGHV1-69 are the genes with the highest number of missense and synonymous variations (Supp Figure 1B). In total, 290 putative variants found in this work result in a stop codon (stop_gained), including 18 present in more than 18 alleles. In

addition, more than 17 frameshifted and in-frame deletion or insertion appeared in more than 18 alleles (Table 1). The largest deletions were on IGHV2-70 (36 nucleotides), localized at FWR3, IGHV3-21 (35 nucleotides), and IGHV3-13 (33 nucleotides), localized at FWR2. Conversely, large insertions were observed in the variants derived from IGHV1-18 (with an insertion of 42 nucleotides) and IGHV1-2 (36 nucleotides) (Supp Figure 2).

4. Most of the variants are population-specific

Only 262 variants were shared by all the populations represented in gnomAD (African, Latino, East Asian, South Asian, European (Finish and non-Finish), and Ashkenazi Jewish). Most of these variants are highly prevalent, as shown in Table 3. The predominant variants not included in other databases come from IGHV3 and IGHV4 subgroups. Variants in the genes IGHV4-31, IGHV4-39, IGHV3-33 are more frequent. We detected an increased number of variants occurring exclusively to some populations, although present at low frequencies (Figure 3, Table 4). We observed 668 unique IGHV variants in Africans, 704 in admixed Americans (Latinos), 846 in Asians (475 South Asia + 381 East Asia), 3,809 in Europeans (3,286 non-Finnish, 63 Finnish, and 460 Ashkenazi Jewish), and 154 in a non-defined population (Other). These results indicate that almost 58,5% of the IGHV variants found in this work (6,181 out of 10,550) are population-specific.

Most of the unique variants (3,286) were found in Europeans (non-Finish), the population with more sequenced individuals. The unique variant higher prevalent in this population is derived from the IGHV3-21 gene, with a frequency of 0.00009735% (Table 4). The most frequent unique variant was found in the IGHV1-2 gene from Latino (61 alleles, 0.0002478%) and IGHV3-9 from East Asian population (60 alleles, 0.0002946%. Interestingly, besides the low frequency of population-specific variants, most of the variants found are present in all of the IGHV gene segments analyzed (Figure 3).

5. IGHV variants identified in GWAS Catalog

To identify IGHV genetic variants associated with common diseases or traits, a search for variants found in this work in NHGRI-EBI GWAS Catalog (33,34) and ClinVar (35) was

performed. Although ClinVar search retrieved no results, 14 IGHV variants were associated with diseases/traits in GWAS Catalog. These variants were found in seven different positions (rs2073668, rs201076896, rs201691548, rs200931578, rs202166511, rs202117805, rs11845244) from three IGHV gene segments (IGHV3-73, IGHV3-61, IGHV1-69).

The rs2073668 (G>A; p.Ser26Ser) synonymous mutation at the IGHV3-73 has a common frequency (0.4518) worldwide, with the highest frequency observed in South Asians (0.6810). The polymorphism (rs2073668) is associated (in trans) with protein levels of Potassium-transporting ATPase subunit beta in Europeans (36). This variant is also described in IMGT (18).

Conversely, the T allele of the missense variant rs11845244 (C>T; p.Gly69Arg) at the IGHV1-69 is associated, in trans, with the level of Beta-defensin 119 (36). The most common variant in this position is described in IMGT (18) and IgPDB (22), but here we described two other low-frequency variants in the same position.

Ten out of 14 IGHV variants found in the GWAS catalog were associated with an increased risk of Rheumatic heart disease. These variants are present in 4 different positions that are in proximity in the IGHV4-61 gene segment. The most frequent variants were described in IMGT, but 4 of them, with lower frequency, were not previously described.

Of these 14 variants, six of them are not present in IMGT (Table 5).

Discussion

Immunoglobulin loci genomic sequencing data suggests that IGH is one of the most polymorphic regions in the human genome (37). Our knowledge regarding this diversity is unexplored, as indicated by the 319 IGHV alleles described in the IMGT/GENE-DB, version 3.1.35 (18). More evidence indicates that this diversity is underestimated, incomplete, and inconsistent (7,9). A study on 28 indigenous South Africans identified 122 non-IMGT IGHV alleles (38). The discovery of new alleles from repertoire data from 7 subjects identified 11 novel IGHV alleles (9), and this number increased to 28 novels IGHV alleles when 26 individuals were analyzed (39). Using germline alleles from 2,504

individuals, Yu et al and Kahtri et al described 3,609 and 410 new IGHV alleles, respectively. Although the technical challenges for detecting variants in complex structural variant locus such as the IGH (40) using short-read data is undeniable, the evolution of the short-reads genome sequencing, assembly, and analysis enabled the development of high-quality databases such as gnomAD (17).

In this work, using gnomAD information from 125,748 exomes, we identified 10,550 putative IGHV variants, the more significant number of IGHV variants described nowadays. The majority of these variants are present in only one exome sequenced (4,949) and might represent, for some findings, a result of sequencing artifacts. However, gnomAD uses very stringent filters, selecting reads with a Phred quality parameters above 20 (use of reads with only 1% chance of error) and the variants covered by at least 10 reads, minimizing the number of false-positive variants found (41). In addition, all the variants present a high-quality confidence score, VSQLOD (a metric that is the log odds ratio probability of the variant being positive or negative), and passed the random forest approach to distinguish true genetic variants from artifacts.

In addition, 81 of these variants found in only one exome were found in the same position as the variants found in other databases (67 on IMGT, 13 on IgPDB, and 1 in both). Interestingly, 5,601 of the variants described here were found in 2 or more alleles available on gnomAD data, and 797 of them were found in more than 18 alleles, in which 575 of them were not previously described in the databases analyzed. Since these variants were found in many exomes, the likelihood of this variant being a false positive is minimal.

Interestingly, we observed that most of the variants are population-specific and appeared at low frequencies, indicating a possible individual-level or few individual-sharing unique IGHVs. For instance, if we analyze the 3,349 IGHV European population-specific variants, the most frequent one is present only in 24 alleles from gnomAD v2.1 data. This can be one reason for different individuals' responses to infectious diseases such as the ones caused by HIV, influenza, and SARS-COV2.

However, knowledge about immunoglobulin gene segment variants is very scarce, and even more rare is their association with diseases susceptibility. We found only 14 variants

in functional IGHV gene segments related to diseases/traits in GWAS Catalog (33). Some of these variants in the IGHV4-61 gene are associated with an increased risk of rheumatic heart disease (42). Other variants found have unexpected effects, in trans, on blood protein levels, such as the level of potassium-transporting ATPase subunit beta and the level of Beta-defensin 119 (36). This association between IGHV variants and protein levels involved in neurotransmission and in the innate immune response is fascinating and opens new perspectives for the impact and importance of IGHV variants.

In addition, more and more evidence associate IGHV variants with the susceptibility to diseases, such as multiple sclerosis (IGHV2), rheumatoid arthritis (IGHV3-30, IGHV4-31, IGHV1-69), systemic lupus erythematosus (IGHV3-30, IGHV4-31), type 1 diabetes (IGHV2, IGHV4, IGHV5), and Kawasaki disease (IGHV1-69, IGHV2-70) (as reviewed by (43)) and also the response to infectious diseases as flu (44) and AIDS (45).

Interestingly, the IGHV genes observed with the higher number of variants (e.g. IGHV2-70, IGHV1-69) are found with 1-4 copies per individual, impacting these higher number of variants (37).

Most of the variants were described in FWR regions, especially in the FWR3. Mutations in these regions can impact antibody's biophysical properties such as antigen recognition, antibody expression yield, pH, and thermal stability (46). While the FWRs help stabilize the antigen-binding site and defines the conformations of the CDR loops, the roles of mutations/polymorphisms in the FWR are not well-understood. Besides these biophysical properties, genetic IGHV variants can also be associated with the variation in gene usage observed between individuals (47).

The scope of the variant dataset provided by this work has some restrictions: first, the GENCODE GRCh37 reference genome includes only a part of the known IGHV genes (40 of 55) and families (6 of 7 - see Table 2); second, the individual-level genomic information that builds the gnomAD database is not accessible, which impairs the individual genotype; third, the analysis was restricted to the V-REGION sequences, narrowing the allele's functionality analysis to check for the presence of stop codons before the last codon; fourth,

the cell types utilized for the whole genome and exome sequencing for each individual in the dataset are unknown.

Only two genome-wide association studies (GWAS) have identified an association with IGHV functional genes, suggesting a potential disconnection between known IGHV haplotype diversity and current high-throughput genotyping tools. Therefore, we provided accession for all variants found in this work (corresponding to 1,262 rsIDs) on the YVr-DB database. We believe that the germline variant database expansion produced by this work can: 1) benefit IGHV gene analysis software, overcoming problems related to the incompleteness of the reference database or helping to validate novel allele discovery; 2) cooperate in understanding the individual or unique population variants (including underrepresented Africans, Latinos, and Asians) and their relationships with diseases susceptibilities; 3) allow the identification of complete consensus sequences between different alleles to help primers design for repertoire studies; 4) enable IGHV SNP link with GWAS.

In conclusion, this work describes the most comprehensive collection of putative IGHV variants currently available, indicating this region's high variability and uniqueness at the individual level. This diversity can impact disease susceptibility, alter the level of non IGHV related proteins in blood plasma, probably impact IGHV gene usage and the yield of secreted antibodies.

Author Contributions

F.R.M. performed research, generated all the data and figures, created the database. L.F.F and T.A.O.M: designed research; L.F.F and L.A.M.P. discussed and analyzed the results, wrote the paper

Conflict of Interest

The authors declare that the research was conducted without any commercial or financial relationships that could be construed as a potential conflict of interest.

Fundings

Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brazil (CAPES) [grant numbers 88887.506611/2020-00, 88887.504420/2020-00 and 935/19 (COFECUB)]; Fundação de Amparo a Pesquisa de Minas Gerais (FAPEMIG) [grant numbers PPM-00615-18, Rede Mineira de Imunobiologicos grant #REDE-00140-16]; Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) [Pq to LFF]; National Institutes of Health (NIH) [grant number 1R01AI143552-02]; Pro-Reitoria de Pesquisa da Universidade Federal de Minas Gerais

Acknowledgments

Lucas Silva for Figure 1 and YVr-DB design. Dr. Luciana Zuccherato for paper review and SynBiom group for fruitful discussions.

References

1. Lefranc Chairperson M-P, Lefranc M-P. IMGT Locus in Focus Nomenclature of the Human Immunoglobulin Heavy (IGH) Genes. *Exp Clin Immunogenet* (2001) **18**:100–116. doi:10.1159/000049189

2. Watson CT, Breden F. The immunoglobulin heavy chain locus: Genetic variation, missing data, and implications for human disease. *Genes Immun* (2012) **13**:363–373. doi:10.1038/gene.2012.12

3. Olee T, Yang PM, Siminovitch KA, Olsen NJ, Hillson J, Wu J, Kozin F, Carson DA, Chen PP. Molecular basis of an autoantibody-associated restriction fragment length polymorphism that confers susceptibility to autoimmune diseases. *J Clin Invest* (1991) **88**:193–203. doi:10.1172/JCI115277

4. Cho M La, Chen PP, Seo Y Il, Hwang SY, Kim WU, Min DJ, Park SH, Cho CS. Association of homozygous deletion of the Humhv3005 and the VH3-30.3 genes with renal involvement in systemic lupus erythematosus. *Lupus* (2003) **12**:400–405. doi:10.1191/0961203303lu385oa

5. Walter MA, Gibson WT, Ebers GC, Cox DW. Susceptibility to multiple sclerosis is associated with the proximal immunoglobulin heavy chain variable region. *J Clin Invest* (1991) **87**:1266–1273. doi:10.1172/JCI115128

6. Wang Y, Jackson KJL, Sewell WA, Collins AM. Many human immunoglobulin heavy-chain IGHV gene polymorphisms have been reported in error. *Immunol Cell Biol* (2008) **86**:111–115. doi:10.1038/sj.icb.7100144

7. Wang Y, Jackson KJ, Gäeta B, Pomat W, Siba P, Sewell WA, Collins AM. Genomic screening by 454 pyrosequencing identifies a new human IGHV gene and sixteen other new IGHV allelic variants. *Immunogenetics* (2011) **63**:259–265. doi:10.1007/s00251-010-0510-8

8. Romo-González T, Morales-Montor J, Rodríguez-Dorantes M, Vargas-Madrazo E. Novel substitution polymorphisms of human immunoglobulin VH genes in Mexicans. *Hum Immunol* (2005) **66**:731–739. doi:10.1016/j.humimm.2005.03.002

9. Gadala-Maria D, Yaari G, Uduman M, Kleinstein SH. Automated analysis of high-throughput B-cell sequencing data reveals a high frequency of novel immunoglobulin V gene segment alleles. *Proc Natl Acad Sci U S A* (2015) **112**:E862–E870. doi:10.1073/pnas.1417683112

10. Corcoran MM, Phad GE, Bernat NV, Stahl-Hennig C, Sumida N, Persson MAA, Martin M, Hedestam GBK. Production of individualized v gene databases reveals high levels of immunoglobulin genetic diversity. *Nat Commun* (2016) **7**: doi:10.1038/ncomms13642

11. Zhang W, Wang IM, Wang C, Lin L, Chai X, Wu J, Bett AJ, Dhanasekaran G, Casimiro DR, Liu X. IMPre: An accurate and efficient software for prediction of T- and B-cell receptor germline genes and alleles from rearranged repertoire data. *Front Immunol* (2016) **7**:1–14. doi:10.3389/fimmu.2016.00457

12. Wendel BS, He C, Crompton PD, Pierce SK, Jiang N. A streamlined approach to antibody novel germline allele prediction and validation. *Front Immunol* (2017) **8**: doi:10.3389/fimmu.2017.01072

13. Yu Y, Ceredig R, Seoighe C. A Database of Human Immune Receptor Alleles Recovered from Population Sequencing Data. *J Immunol* (2017) **198**:2202–2210. doi:10.4049/jimmunol.1601710

14. Khatri I, Berkowska MA, van den Akker EB, Teodosio C, Reinders MJT, van Dongen JJM. Population matched (PM) germline allelic variants of immunoglobulin (IG) loci: New pmIG database to better understand IG repertoire and selection processes in disease and vaccination. *bioRxiv* (2020) doi:10.1101/2020.04.09.033530

15. Birney E, Soranzo N. Human genomics: The end of the start for population sequencing. *Nature* (2015) **526**:52–53. doi:10.1038/526052a

16. 1000 Genomes Project Consortium, Auton, A., L. D. Brooks, R. M. Durbin, E. P. Garrison, H. M. Kang, J. O. Korbel, J. L. Marchini, S. McCarthy, G. A. McVean and GRA. A global reference for human genetic variation. *Nature* (2015)68–74. doi:10.1038/nature15393

17. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, Collins RL, Laricchia KM, Ganna A, Birnbaum DP, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* (2020) **581**:434–443. doi:10.1038/s41586-020-2308-7

18. Giudicelli V, Chaume D, Lefranc MP. IMGT/GENE-DB: A comprehensive database for human and mouse immunoglobulin and T cell receptor genes. *Nucleic Acids Res* (2005) **33**:256–261. doi:10.1093/nar/gki010

19. Church DM, Schneider VA, Graves T, Auger K, Cunningham F, Bouk N, Chen HC, Agarwala R, McLaren WM, Ritchie GRS, et al. Modernizing reference genome assemblies. *PLoS Biol* (2011) **9**:1–5. doi:10.1371/journal.pbio.1001091

20. Frankish A, Diekhans M, Ferreira AM, Johnson R, Jungreis I, Loveland J, Mudge JM, Sisu C, Wright J, Armstrong J, et al. GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res* (2019) **47**:D766–D773. doi:10.1093/nar/gky955

21. Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck T, Kauff F, Wilczynski B, et al. Biopython: Freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* (2009) **25**:1422–1423. doi:10.1093/bioinformatics/btp163

22. The Immunoglobulin Polymorphism Database.https://cgi.cse.unsw.edu.au/~ihmmune/IgPdb/informa. Available at: https://cgi.cse.unsw.edu.au/~ihmmune/IgPdb/informa

23. Giudicelli V, Chaume D, Lefranc MP. IMGT/V-QUEST, an integrated software program for immunoglobulin and T cell receptor V-J and V-D-J rearrangement analysis. *Nucleic Acids Res* (2004) **32**:435–440. doi:10.1093/nar/gkh412

24. Katoh K, Misawa K, Kuma KI, Miyata T. MAFFT: A novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res* (2002) **30**:3059–3066. doi:10.1093/nar/gkf436

25. Mikocziova I, Gidoni M, Lindeman I, Peres A, Snir O, Yaari G, Sollid LM. Polymorphisms in human immunoglobulin heavy chain variable genes and their upstream regions. *Nucleic Acids Res* (2021) **48**:5499–5510. doi:10.1093/NAR/GKAA310

26. Ye J, Ma N, Madden TL, Ostell JM. IgBLAST: an immunoglobulin variable domain sequence analysis tool. *Nucleic Acids Res* (2013) **41**:34–40. doi:10.1093/nar/gkt382

27. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K. DbSNP: The NCBI database of genetic variation. *Nucleic Acids Res* (2001) **29**:308–311. doi:10.1093/nar/29.1.308

28. Zerbino DR, Achuthan P, Akanni W, Amode MR, Barrell D, Bhai J, Billis K, Cummins C, Gall A, Girón CG, et al. Ensembl 2018. *Nucleic Acids Res* (2018) **46**:D754–D761. doi:10.1093/nar/gkx1098

29. Bunn A, Korpela M. A language and environment for statistical computing. *Found Stat Comput* (2013) **2**:1–12.

30. RStudio Team. RStudio: Integrated Development Environment for R. (2020) Available at: http://www.rstudio.com/

31. Villanueva RAM, Chen ZJ. ggplot2: Elegant Graphics for Data Analysis (2nd ed.). *Meas Interdiscip Res Perspect* (2019) **17**:160–167. doi:10.1080/15366367.2019.1565254

32. Landrum MJ, Lee JM, Benson M, Brown GR, Chao C, Chitipiralla S, Gu B, Hart J, Hoffman D, Jang W, et al. ClinVar: Improving access to variant interpretations and supporting evidence. *Nucleic Acids Res* (2018) **46**:D1062–D1067. doi:10.1093/nar/gkx1153

33. Buniello A, Macarthur JAL, Cerezo M, Harris LW, Hayhurst J, Malangone C, McMahon A, Morales J, Mountjoy E, Sollis E, et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res* (2019) **47**:D1005–D1012. doi:10.1093/nar/gky1120

34. MacArthur J, Bowler E, Cerezo M, Gil L, Hall P, Hastings E, Junkins H, McMahon A, Milano A, Morales J, et al. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res* (2017) **45**:D896–D901. doi:10.1093/nar/gkw1133

35. Landrum MJ, Lee JM, Benson M, Brown G, Chao C, Chitipiralla S, Gu B, Hart J, Hoffman D, Hoover J, et al. ClinVar: Public archive of interpretations of clinically relevant variants. *Nucleic Acids Res* (2016) **44**:D862–D868. doi:10.1093/nar/gkv1222

36. Sun BB, Maranville JC, Peters JE, Stacey D, James R, Blackshaw J, Burgess S, Jiang T, Paige E, Oliver-williams C, et al. Europe PMC Funders Group Europe PMC Funders Author Manuscripts Europe PMC Funders Author Manuscripts Genomic atlas of the human plasma proteome Europe PMC Funders Author Manuscripts Europe PMC Funders Author Manuscripts. (2019) **558**:73–79. doi:10.1038/s41586-018-0175-2.Genomic

37. Watson CT, Steinberg KM, Huddleston J, Warren RL, Malig M, Schein J, Willsey AJ, Joy JB, Scott JK, Graves TA, et al. Complete haplotype sequence of the human immunoglobulin heavy-chain variable, diversity, and joining genes and characterization of allelic and copy-number variation. *Am J Hum Genet* (2013) **92**:530–546. doi:10.1016/j.ajhg.2013.03.004

38. Scheepers C, Shrestha RK, Lambson BE, Jackson KJL, Wright IA, Naicker D, Goosen M, Berrie L, Ismail A, Garrett N, et al. Ability To Develop Broadly Neutralizing HIV-1

Antibodies Is Not Restricted by the Germline Ig Gene Repertoire. *J Immunol* (2015) **194**:4371–4378. doi:10.4049/jimmunol.1500118

39. Gadala-Maria D, Gidoni M, Marquez S, Vander Heiden JA, Kos JT, Watson CT, O'Connor KC, Yaari G, Kleinstein SH. Identification of subject-specific immunoglobulin alleles from expressed repertoire sequencing data. *Front Immunol* (2019) **10**:1–12. doi:10.3389/fimmu.2019.00129

40. Watson CT, Matsen FA, Jackson KJL, Bashir A, Smith ML, Glanville J, Breden F, Kleinstein SH, Collins AM, Busse CE. Comment on "A Database of Human Immune Receptor Alleles Recovered from Population Sequencing Data." *J Immunol* (2017) **198**:3371–3373. doi:10.4049/jimmunol.1700306

41. Gudmundsson S, Singer-Berk M, Watts NA, Phu W, Goodrich JK, Solomonson M, Rehm HL, MacArthur DG, O'Donnell-Luria A. Variant interpretation using population databases: Lessons from gnomAD. *Hum Mutat* (2021) doi:10.1002/humu.24309

42. Muhamed B, Parks T, Sliwa K. Genetics of rheumatic fever and rheumatic heart disease. *Nat Rev Cardiol* (2020) **17**:145–154. doi:10.1038/s41569-019-0258-2

43. Marie J. Kidd et al. Chain V Region Gene Loci by Analysis of VDJ Gene. *J Immunol* (2012) **188**:1333–1340. doi:10.4049/jimmunol.1102097

44. Avnir Y, Watson CT, Glanville J, Peterson EC, Tallarico AS, Bennett AS, Qin K, Fu Y, Huang CY, Beigel JH, et al. IGHV1-69 polymorphism modulates anti-influenza antibody repertoires, correlates with IGHV utilization shifts and varies by ethnicity. *Sci Rep* (2016) **6**: doi:10.1038/srep20842

45. Yacoob C, Pancera M, Vigdorovich V, Oliver BG, Glenn JA, Feng J, Sather DN, McGuire AT, Stamatatos L. Differences in Allelic Frequency and CDRH3 Region Limit the Engagement of HIV Env Immunogens by Putative VRC01 Neutralizing Antibody Precursors. *Cell Rep* (2016) **17**:1560–1570. doi:10.1016/j.celrep.2016.10.017

46. Cnudde T, Lakhrif Z, Bourgoin J, Boursin F, Horiot C, Henriquet C, di Tommaso A, Juste MO, Jiacomini IG, Dimier-Poisson I, et al. Exploration and Modulation of Antibody Fragment Biophysical Properties by Replacing the Framework Region Sequences. *Antibodies* (2020) **9**:9. doi:10.3390/antib9020009

47. Kenter AL, Watson CT, Spille J-H. Igh Locus Polymorphism May Dictate Topological Chromatin Conformation and V Gene Usage in the Ig Repertoire. *Front Immunol* (2021) **12**:1–7. doi:10.3389/fimmu.2021.682589

Tables

Table 1: Number of times each putative variants were sequenced in exome (allele count)

		Presence of variants in other GLDB					Type of Mutation				
Allele Count	**Number of Variants**	**IMGT**	**IgPDB**	**IMGT/IgPDB**	**Missense**	**Synonymous**	**Stop_gained**	**Frameshift**	**Deletion**	**Insertion**	**Protein_altering***
1	**4,949**	**67**	**13**	**1**	**3,344**	**1,198**	**162**	**185**	**32**	**23**	**5**
2-6	**3,991**	**59**	**8**	**2**	**2,728**	**1,105**	**95**	**49**	**7**	**7**	**0**
7-18	**813**	**16**	**4**	**2**	**533**	**250**	**15**	**13**	**0**	**2**	**0**
>18	**797**	**136**	**50**	**36**	**476**	**278**	**18**	**17**	**4**	**4**	**0**
Total	**10,550**	**278**	**75**	**41**	**7,081**	**2,831**	**290**	**264**	**43**	**36**	**5**

***Protein Altering** refers to in frame or frameshifted (insertions/deletions) variants. Here, all 5 are in frame insertion classified by gnomAD as a protein-altering variant.

Table 2: IGHV variants by subgroup

IGHV subgroups (gene name, number of functional IGHV genes, and number of variants found in this study)											
IGHV1		IGHV2		IGHV3		IGHV4		IGHV5		IGHV6	
V1-18 (4F)	347	V2-26 (4F)	184	V3-11 (5F, 1P)	310	V4-28 (7F)	170	V5-51 (7F)	283	V6-1 (2F, 1P)	258
V1-2 (7F)	447	V2-5 (2F)	322	V3-13 (5F)	235	V4-31 (11F)	273				
V1-24 (1F)	260	V2-70 (17F, 1ORF)	795	V3-15 (8F)	290	V4-34 (13F)	339				
V1-3 (5F)	274			V3-20 (2F, 2ORF)	183	V4-39 (7F)	329				
V1-45 (3F)	98			V3-21 (6F)	316	V4-4 (9F)	273				
V1-46 (4F)	206			V3-23 (5F)	347	V4-59 (13F)	226				
V1-58 (3F)	129			V3-30 (19F)	238	V4-61 (9F, 1ORF)	173				
V1-69 (19F)	403			V3-33 (7F)	333						
V1-8 (3F)	302			V3-43 (2F)	133						
				V3-48 (4F)	227						
				V3-49 (5F)	196						

		V3-53 (5F) 185 V3-64 (6F) 103 V3-66 (4F) 134 V3-7 (5F) 369 V3-72 (2F) 148 V3-73 (2F) 119 V3-74 (3F) 188 V3-9 (3F) 405			
2,466	1,301	4,459	1,783	283	258

Table 3: Top 10 most prevalent IGHV variants not present on IMGT and IgPDB

IGHV gene name	Variant Description	Allele Number	Allele frequency
IGHV4-31	14-106805381-T-G	82,924	0.3793
IGHV4-4	14-106478194-C-T	78,412	0.4481
IGHV4-31	14-106805395-G-A	63,483	0.2869
IGHV4-39	14-106877753-A-G	61,869	0.2703
IGHV3-53	14-107048767-G-T	37,893	0.185
IGHV4-39	14-106877796-G-C	37,715	0.1938
IGHV3-43	14-106926311-T-G	37,463	0.1543
IGHV3-33	14-106815862-A-G	26,121	0.1088
IGHV3-33	14-106815868-C-A	26,057	0.1086
IGHV3-33	14-106815970-T-C	22,555	0.09422

Table 4: Number of unique variants per population and the most frequent variant of the population

Population	Number of unique variants	Most frequent IGHV	Variant Description	Allele Count	Allele Frequency
African	668	IGHV1-18	14-106641779-A-T	22	0.00008928
Latino	704	IGHV1-2	14-106452903-G-C	61	0.0002478
EastAsian	371	IGHV3-9	14-106552504-T-G	60	0.0002946
South Asian	475	IGHV3-53	14-107048714-G-A	39	0.0001582
European (non-Finnish)	3,286	IGHV3-21	14-106691736-G-A	24	0.00009735
Ashkenazi Jewish	460	IGHV1-2	14-106452694-T-C	15	0.00006103
European (Finnish)	63	IGHV3-13	14-106586381-C-A	10	0.00004069
Other	154	IGHV1-8	14-106539202-G-T	3	0.00001479
Total	6,181				

Table 5: Variants present in GWAS Catalog

rsid	Allele Count	IGHV gene	Trait
rs2073668*	112340	IGHV3-73	Potassium-transporting ATPase subunit beta levels
rs201076896	2189	IGHV4-61	Rheumatic heart disease
rs201691548	1	IGHV4-61	Rheumatic heart disease
rs201691548	7	IGHV4-61	Rheumatic heart disease
rs201691548	1978	IGHV4-61	Rheumatic heart disease
rs200931578	1375	IGHV4-61	Rheumatic heart disease
rs202166511	3	IGHV4-61	Rheumatic heart disease
rs202166511	1675	IGHV4-61	Rheumatic heart disease
rs202166511	2	IGHV4-61	Rheumatic heart disease
rs202117805	4	IGHV4-61	Rheumatic heart disease
rs202117805	28144	IGHV4-61	Rheumatic heart disease
rs11845244	1	IGHV1-69	Beta defensin 119 level
rs11845244	1	IGHV1-69	Beta defensin 119 level
rs11845244	97894	IGHV1-69	Beta defensin 119 level

* Only this mutation is synonymous, the others are missense

Figure Legends:

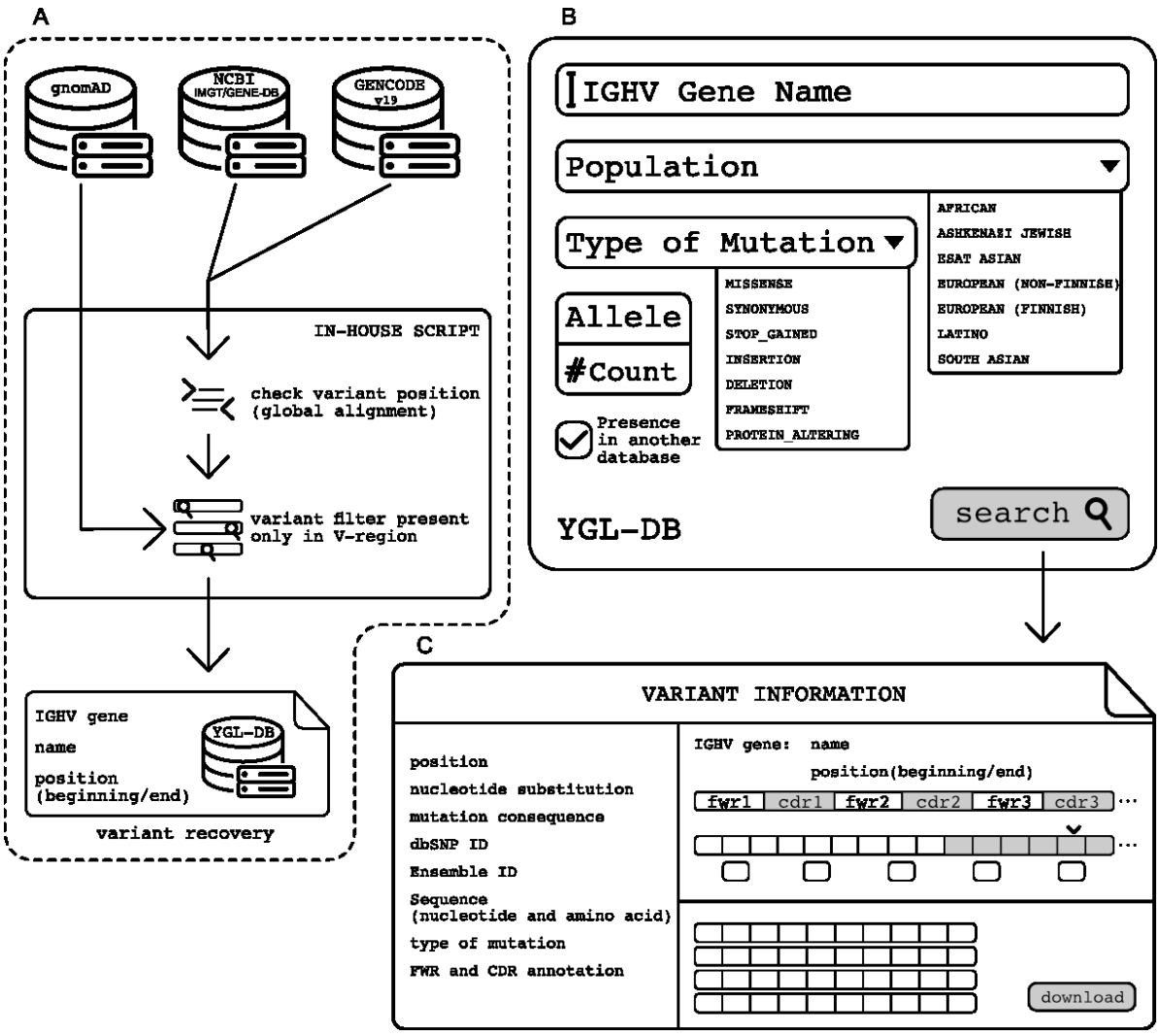


Figure 1: IGHV germline variants discovery strategy and YVr-DB. A) IGHV genes names, sequences, and positions were extracted from NCBI (that uses IMGT-Gene-DB) and from GENCODE (used by gnomAD). The sequences obtained were compared using global alignment. After manual checking, all the 40 IGHV on gnomAD, matched with IMGT sequence, were filtered only by its V-REGION, and the variants were recovered by replacing the corresponding mutant nucleotide in V-REGION. B) YVr-DB input where the user can search the database by IGHV gene name, population, type of mutation, the number of times the variants were sequenced, or if the

variants are present in other databases. C) YVr-DB output with the corresponding IGHV gene name and position and several variant information: position, nucleotide substitution and consequence in the protein, type of mutation, variant sequence, dbSNP and Ensembl identifier, FWR and CDR annotation in the variants, and allele count (number of times this variant were sequenced).

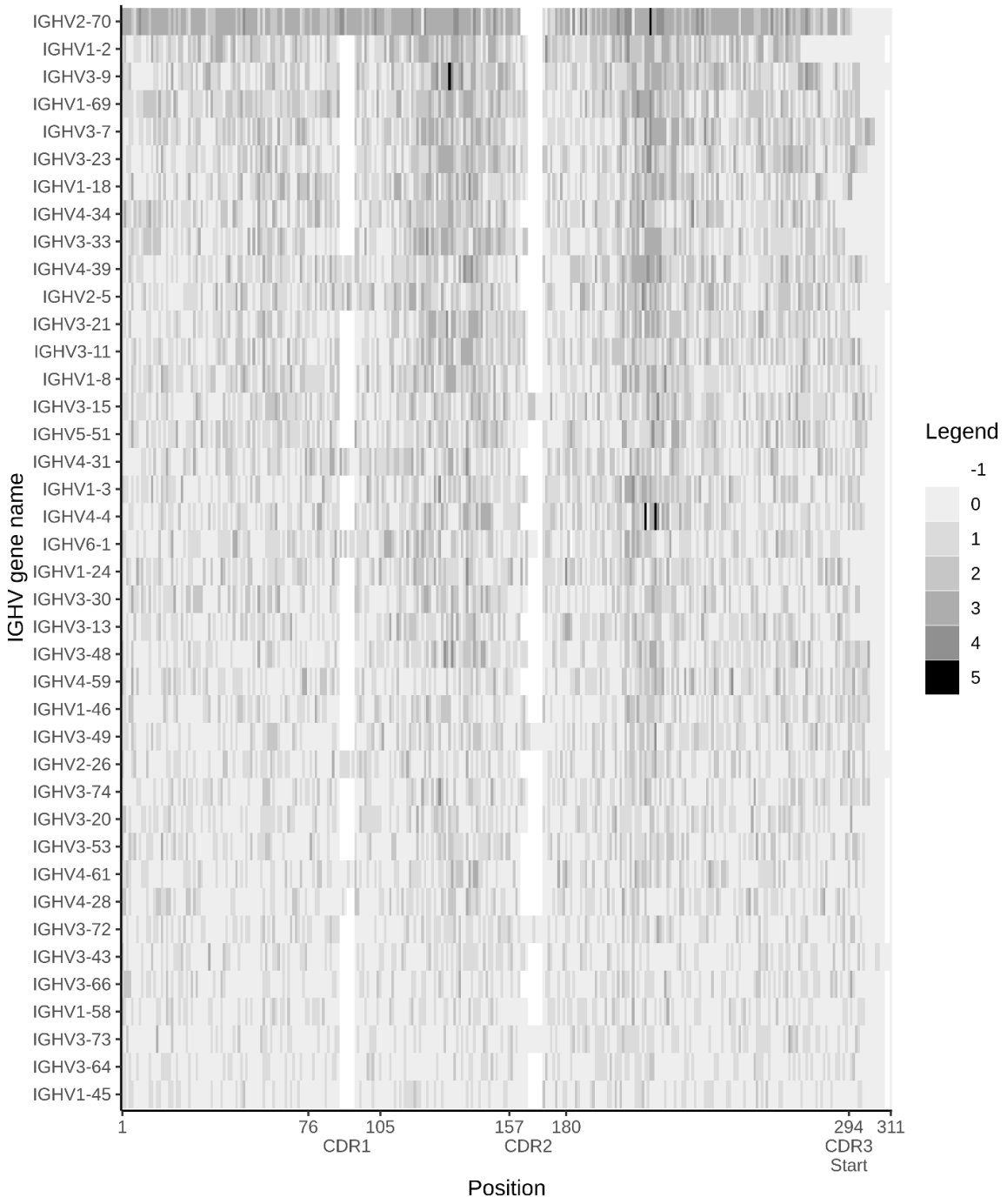


Figure 2: Number of variants in each of the IGHV genes per position. Heatmap presenting the number of each variant per IGHV position. Zero (0) means no variant in a given position, 1 one variant, 2 for two variants, 3 for three variants, 4 for four variants and 5 for five variants in a given position. In this heatmap, genes presenting more variants are at the top of the list, and those with fewer variants are at the bottom.

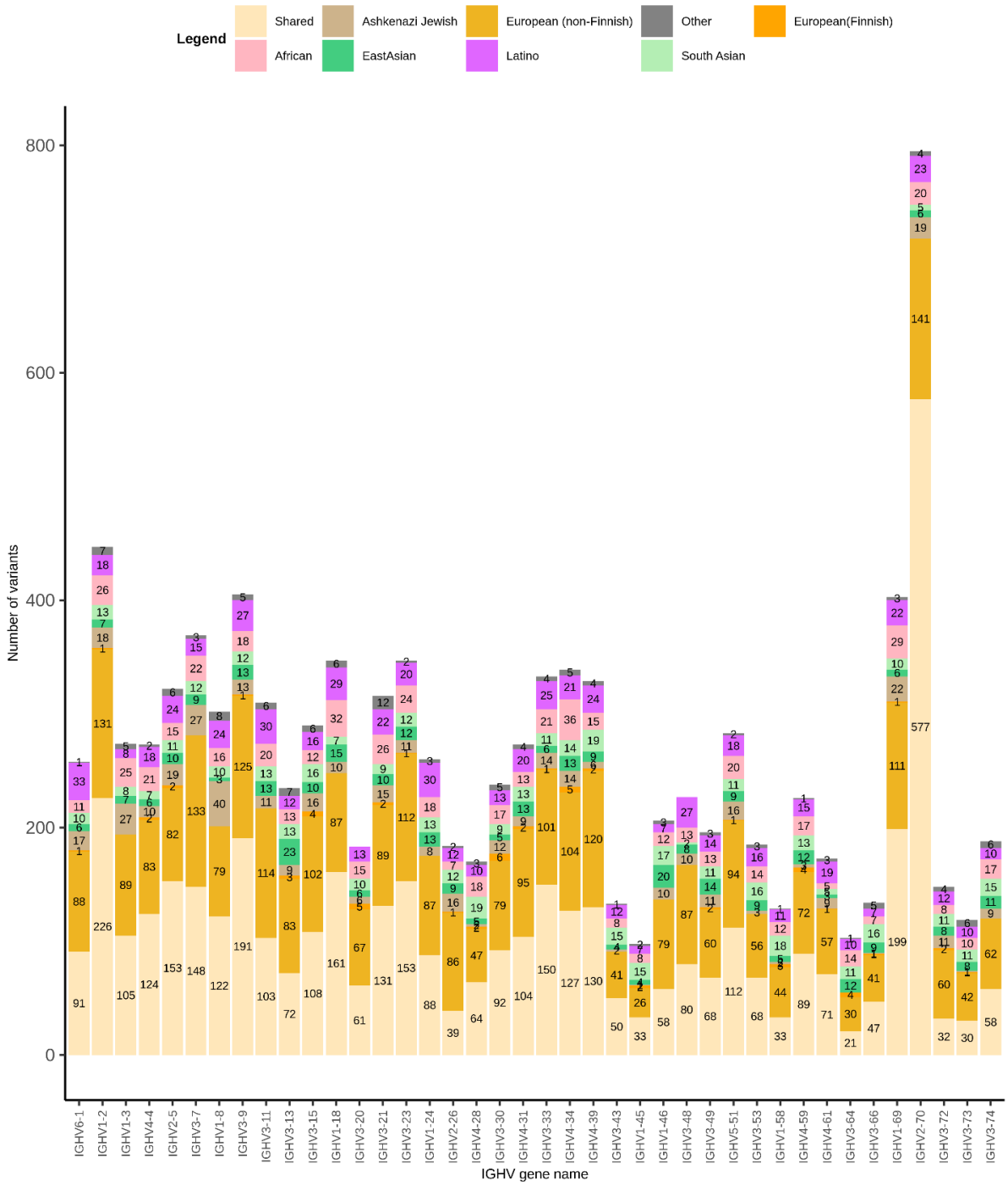
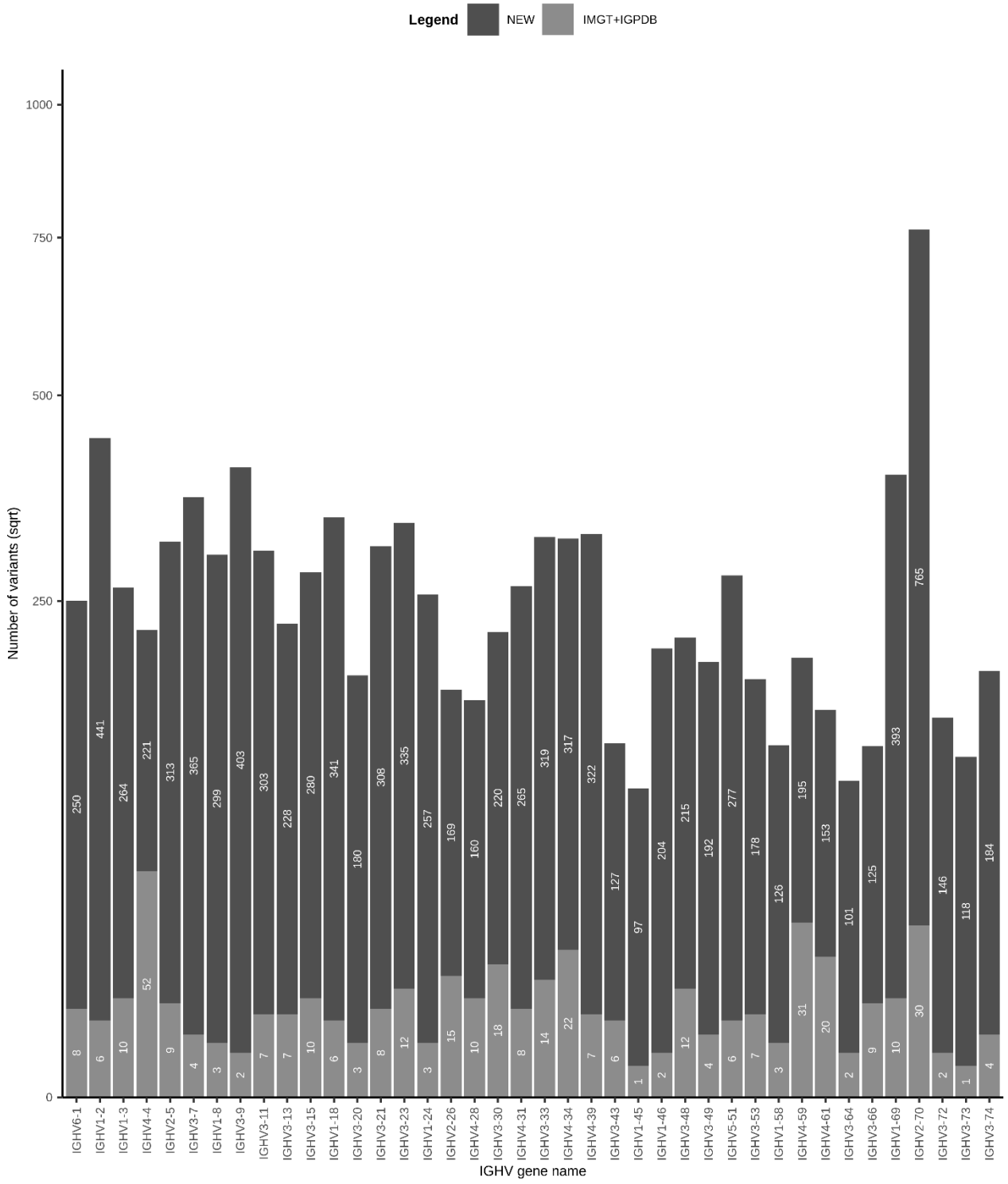


Figure 3: Number of shared or unique population-specific variants per IGHV gene segment. The number of variants per IGHV gene segment found in this work unique to the different populations analyzed such as European (non-Finnish and Finnish), East and South Asian, Latino,

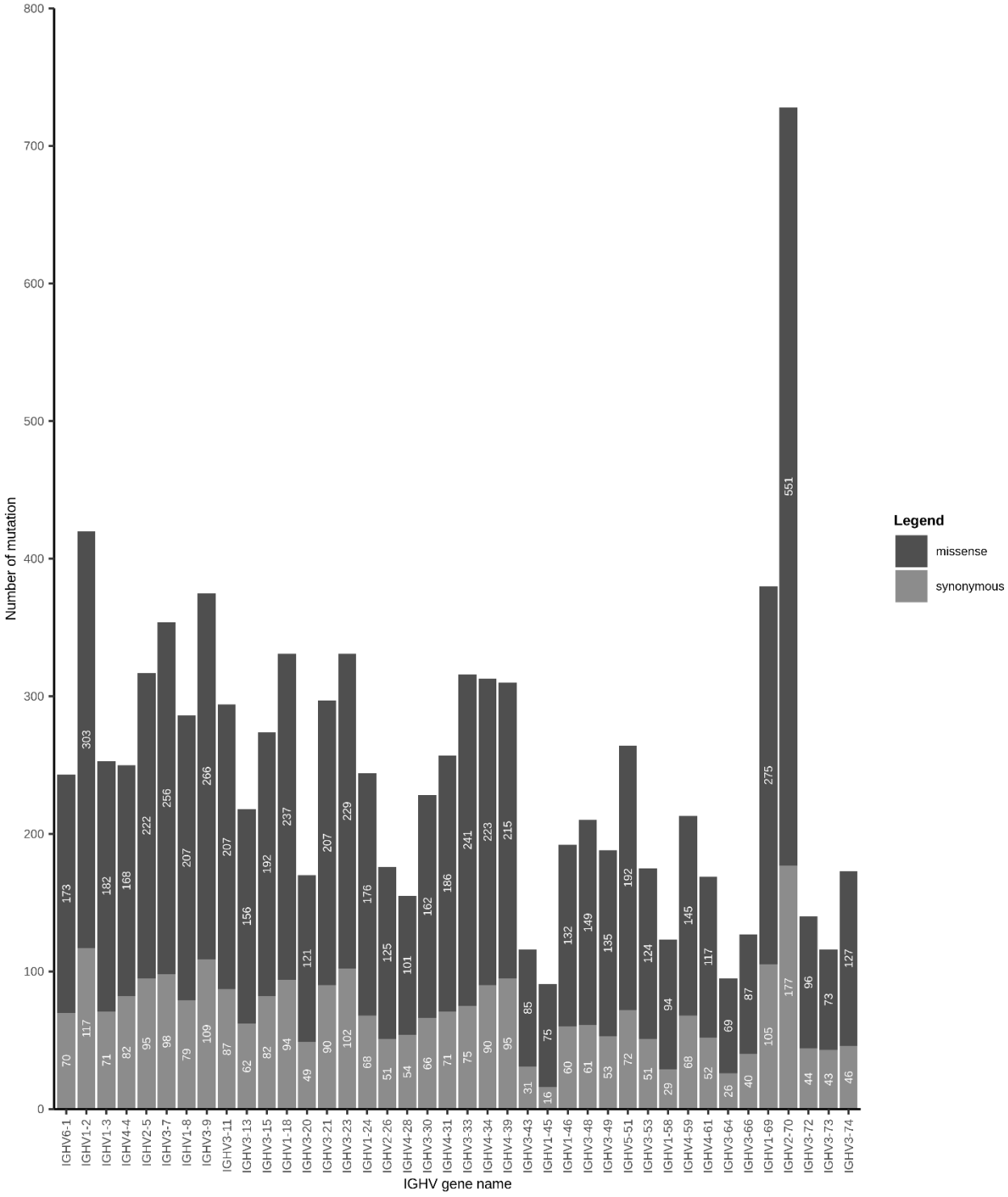
African, Other non-defined populations or shared between at least two different populations are represented in this figure.

Supplementary material

A)



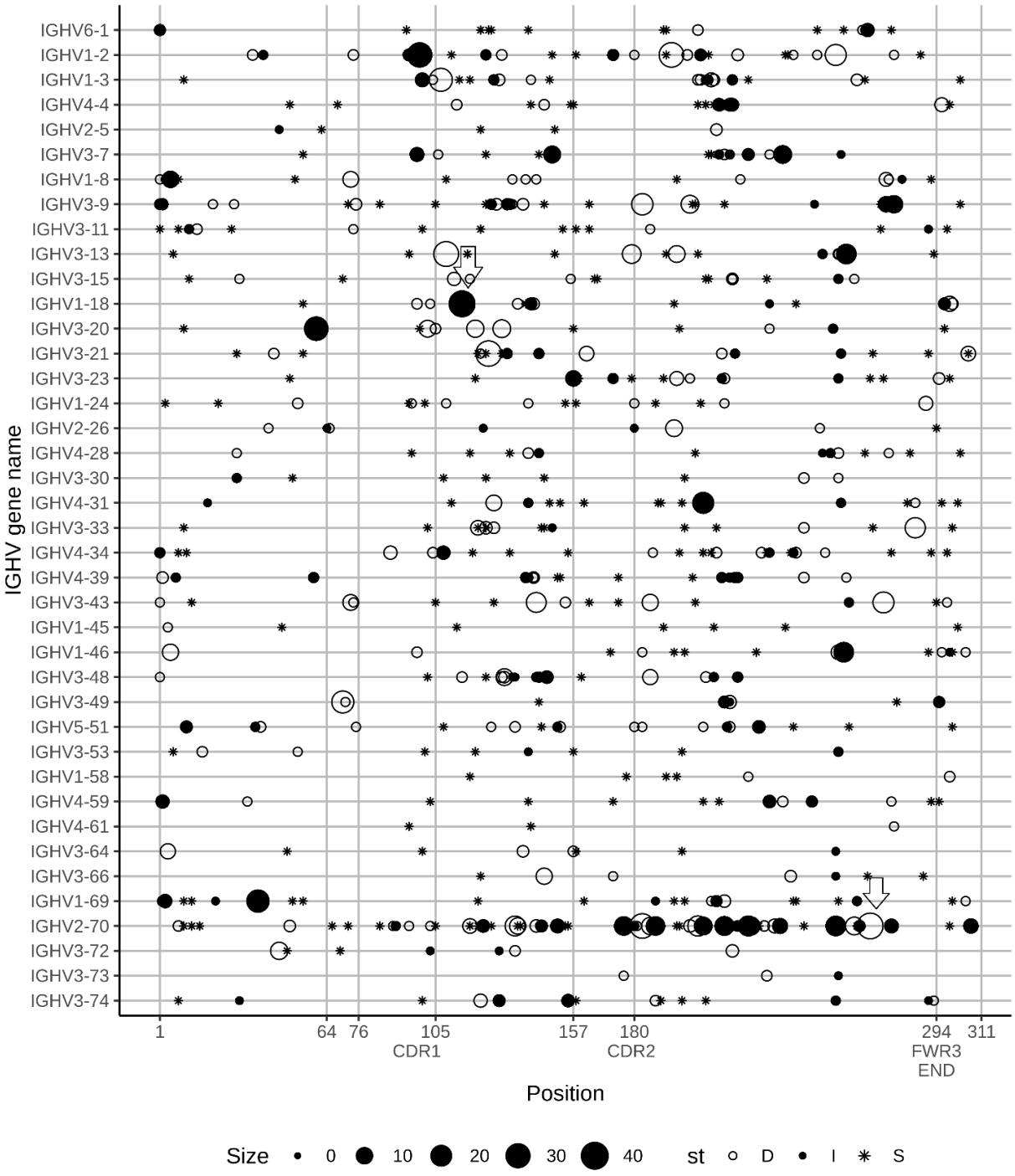
B)


Number of mutation
IGHV gene name
Legend
missense
synonymous
IGHV6-1
IGHV1-2
IGHV1-3
IGHV4-4
IGHV2-5
IGHV3-7
IGHV1-8
IGHV3-9
IGHV3-11
IGHV3-13
IGHV3-15
IGHV1-18
IGHV3-20
IGHV3-21
IGHV3-23
IGHV1-24
IGHV2-26
IGHV4-28
IGHV3-30
IGHV4-31
IGHV3-33
IGHV4-34
IGHV4-39
IGHV3-43
IGHV1-45
IGHV1-46
IGHV3-48
IGHV3-49
IGHV5-51
IGHV3-53
IGHV1-58
IGHV4-59
IGHV4-61
IGHV3-64
IGHV3-66
IGHV1-69
IGHV2-70
IGHV3-72
IGHV3-73
IGHV3-74

Supplementary Figure 1: Variants found in each IGHV gene segment from exome data. The 40 IGHV gene segment are presented in this figure from centromeric to telomeric (3' to 5').

A) Variants from IGHV gene described in this work and the ones also find in other GLDB databases. Light gray bars represent variants that were already found in IMGT and IgPDB. Sqrt (square root).

B) Missense and synonymous variant number present in the different IGHV genes. Light gray bars represent synonymous variants and dark gray bars represents missense variants.



Supplementary Figure 2: Number of deletions (□), insertions (□), or stop_codon (□) mutations by IGHV gene position. FWR1 (1 to 75), CDR1 (76 to 105), FWR2 (106-156),

CDR 2 (157 to 180), FWR3 (181-293) and beginning of the CDR3 (294-311). The first conserved cysteine position (64) is also highlighted in the Figure. Different circle sizes represent the number of inserted or deleted nucleotides, the bigger the circle larger the insertion or deletion. The two arrows indicate the larger deletion and insertion.

ANEXO II

A seguir o artigo que participei durante o período de doutorado. Colaborei desenvolvendo alguns algoritmos complementares ao *pipeline* e também na elaboração de algumas figuras.

Este artigo está depositado no bioRxiv doi: https://biorxiv.org/cgi/content/short/2022.06.20.496904v1

Molecular Immunology

THE MAJOR ROLE OF JUNCTIONAL DIVERSITY IN THE HORSE ANTIBODY REPERTOIRE

--Manuscript Draft--

Manuscript Number:	
Article Type:	Full Length Article
Keywords:	horse; antibody repertoire; BCR-seq; junctional diversity
Corresponding Author:	Liza Felicori Universidade Federal de Minas Gerais Belo Horizonte, BRAZIL
First Author:	Carlena Navas
Order of Authors:	Carlena Navas
	Taciana Manso
	Fabio Martins
	Lucas Minto
	Rennan Moreira
	João Minozzo
	Bruno Antunes
	André Vale
	Jonathan R. McDaniel
	Gregory Ippolito
	Liza Felicori
Abstract:	The sequencing of the antibody repertoire (Rep-seq) revolutionized the diversity of antigen B cell receptor studies, allowing deep and quantitative analysis to decipher the role of adaptive immunity in health and disease. Particularly, horse (Equus caballus) polyclonal antibodies have been produced and used since the century XIX to treat and prophylaxis of diphtheria, tuberculosis, tetanus, pneumonia, and, more recently, COVID-19. However, our knowledge about the horse B cell receptors repertories is minimal. We present a deep horse antibody heavy chain repertoire (IGH) characterization of non-immunized horses using HTS technology. In this study, we obtained a mean of 248,169 unique IgM clones and 66,141 unique IgG clones from four domestic adult horses. Rarefaction analysis showed sequence coverage was between 52 and 82% in IgM and IgG isotypes. We observed that besides horses antibody can use all of the functional IGHV genes, around 80% of their antibodies use only three IGHV gene segments, and around 55% use only one IGHJ gene segment. This limited VJ diversity seems to be compensated by the junctional diversity of these antibodies. We observed that the junctional diversity in horses antibodies is highly frequent, present in more than 90% of horse antibodies. Besides this, the length of this region seems to be higher in horse antibodies than in other species. N1 and N2 nucleotides addition range from 0 to 111 nucleotides. In addition, around 45% of the antibody clones have more than ten nucleotides in both N1 and N2 junction regions. This diversity mechanism may be one of the most important in providing variability to the equine antibody repertoire. This study provides new insights regarding horse antibody composition, diversity generation, and particularities compared to other species, such as the frequency and length of N nucleotide addition. This study also points out the urgent need to better characterize TdT in horses and in other species to better understand antibody repertoire characteristics.
Suggested Reviewers:	Rebecca L. Tallmadge rlt8@cornell.edu Expert in horse antibodies. Published previous works characterizing horse antibodies repertoire using Sanger sequencing

	Antti Iivanainen antti.iivanainen@helsinki.fi Published a work characterizing cattle antibodies repertoire (PMID: 24926997).
	Harry W Schroeder Jr hwsj@uab.edu Expert in IGHD gene characterization

Cover Letter

Universidade Federal de Minas Gerais
Curso de Pós-Graduação em Bioinformática ICB/UFMG
Av. Antônio Carlos, 6627 – Pampulha
31270-901 - Belo Horizonte – MG
e-mail: bicinfo@icb.ufmg.br (31)3409-2615

Dear Editor,

We are pleased to send to you a manuscript entitled **"THE MAJOR ROLE OF JUNCTIONAL DIVERSITY IN THE HORSE ANTIBODY REPERTOIRE"** which we would like to have published in Molecular Immunology. This article reports high-throughput sequencing of the antibody repertoire for the IgM and IgG isotypes of PBMC from non-immunized horses. A previously study of our group published in Molecular Immunology characterized horse antibodies, in a high-throuput way, for the first time (PMID: 30562645). Now, with a depper analysis of 4 horses instead of 2 previsously analysed, we observed that horses antibodies has a restrict VJ usage in their antibodies. We observed that around 80% of IgM and IgG repertoires are composed by only 3 different IGHV4 gene segments, and 60% of them use only the IGHJ6 gene segment. This restrict gene usage in horses antibodies seems to be compensated by a high frequent and in many times, long junctional diversity of N1 and N2 regions, that can have from 0 to 111 nucleotides. In this way, this study provides new insights regarding horse antibody composition, diversity generation, and particularities compared to other species, such as the frequency and length of N nucleotide addition.

For that, we believe that our results are very relevant and of general interest. In addition, studies of the antibody repertoire of dogs (PMID: 24509215), salmon (PMID: 30593934), pigs (https://doi.org/10.1016/j.molimm.2005.10.017) among other species have also been published in Molecular Immunology. Given the interest of the Molecular Inmunology community in understand the characteristics of antibody repertoires from different species, we believe this work present a depper and significant contribution for a better understanding of horse antibody characteristics.

Looking forward to hearing from you,

Yours truly,

Felicori

Liza Felicori
Associate Professor |
Principal investigator
Synbiom Lab
Coordinator of the IdeaReal
lab: www.ideareal.org

SYNBIOM
Laboratório de Biologia
Sintética e Biomiméticos
Departamento de
Bioquímica e Imunologia |
ICB
Universidade Federal de
Minas Gerais

+55 (31) 99330-4420

liza@icb.ufmg.br

Av. Antonio Carlos, 6627 –
Bloco N4 – sala 202 –
Pampulha
Belo Horizonte – MG –
Brazil
CEP 31270-901

Highlights

- Horse antibodies have a restrict VJ gene usage
- Only 3 different IGHV4 and the IGHJ6 genes are used for most of the horses antibodies
- The junctional diversity in horse antibodies looks bigger compared to other species
- More than 80% of horse antibodies use IGHD reading frame 1

Abstract

The sequencing of the antibody repertoire (Rep-seq) revolutionized the diversity of antigen B cell receptor studies, allowing deep and quantitative analysis to decipher the role of adaptive immunity in health and disease. Particularly, horse (*Equus caballus*) polyclonal antibodies have been produced and used since the century XIX to treat and prophylaxis of diphtheria, tuberculosis, tetanus, pneumonia, and, more recently, COVID-19. However, our knowledge about the horse B cell receptors repertories is minimal. We present a deep horse antibody heavy chain repertoire (IGH) characterization of non-immunized horses using HTS technology. In this study, we obtained a mean of 248,169 unique IgM clones and 66,141 unique IgG clones from four domestic adult horses. Rarefaction analysis showed sequence coverage was between 52 and 82% in IgM and IgG isotypes. We observed that besides horses antibody can use all of the functional IGHV genes, around 80% of their antibodies use only three IGHV gene segments, and around 55% use only one IGHJ gene segment. This limited VJ diversity seems to be compensated by the junctional diversity of these antibodies. We observed that the junctional diversity in horses antibodies is highly frequent, present in more than 90% of horse antibodies. Besides this, the length of this region seems to be higher in horse antibodies than in other species. N1 and N2 nucleotides addition range from 0 to 111 nucleotides. In addition, around 45% of the antibody clones have more than ten nucleotides in both N1 and N2 junction regions. This diversity mechanism may be one of the most important in providing variability to the equine antibody repertoire. This study provides new insights regarding horse antibody composition, diversity generation, and particularities compared to other species, such as the frequency and length of N nucleotide addition. This study also points out the urgent need to better characterize TdT in horses and in other species to better understand antibody repertoire characteristics.

THE MAJOR ROLE OF JUNCTIONAL DIVERSITY IN THE HORSE ANTIBODY REPERTOIRE

Carlena Navas[a, b], Taciana Manso[a,c], Fabio Martins[a], Lucas Minto[a], Rennan Moreira[d], João Minozzo[e], Bruno Antunes[e], André Vale[f], Jonathan R. McDaniel[g], Gregory C. Ippolito[g], Liza F. Felicori[a*]

[a] Laboratory of Synthetic Biology and Biomimetics, Departamento de Bioquímica e Imunologia, Instituto de Ciências Biológicas - ICB, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil.

[b] University of Carabobo, Faculty of Health Sciences, School of Biomedical and Technological Sciences Department of Morphological and Forensic Sciences, Valencia Venezuela.

[c] The International Immunogenetics Information System / IMGT Institut de Génétique Humaine / IGH – CNRS Montpellier / France.

[d] Multi-users Laboratories Center, Instituto de Ciências Biológicas - ICB, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil

[e]Production and Research Centre of Immunobiological Products, Department of Health of the State of Paraná, Piraquara 83302-200, Brazil.

[f] Program in Immunobiology, Carlos Chagas Filho Institute of Biophysics, Federal University of Rio de Janeiro, Rio de Janeiro, Brazil

[g] Department of Molecular Biosciences, The University of Texas at Austin, 100 E. 24th Street, Stop A5000, Austin, TX, 78712, USA

Correspondence:

*Liza F. Felicori

Email: liza@icb.ufmg.br

Keywords: horse, antibody repertoire, BCR-seq, junctional diversity

Abstract

The sequencing of the antibody repertoire (Rep-seq) revolutionized the diversity of antigen B cell receptor studies, allowing deep and quantitative analysis to decipher the role of adaptive immunity in health and disease. Particularly, horse (*Equus caballus*) polyclonal antibodies have been produced and used since the century XIX to treat and prophylaxis of diphtheria, tuberculosis, tetanus, pneumonia, and, more recently, COVID-19. IIowever, our knowledge about the horse B cell receptors repertories is minimal. We present a deep horse antibody heavy chain repertoire (IGH) characterization of non-immunized horses using HTS technology. In this study, we obtained a mean of 248,169 unique IgM clones and 66,141 unique IgG clones from four domestic adult horses. Rarefaction analysis showed sequence coverage was between 52 and 82% in IgM and IgG isotypes. We observed that besides horses antibody can use all of the functional IGHV genes, around 80% of their antibodies use only three IGHV gene segments, and around 55% use only one IGHJ gene segment. This limited VJ diversity seems to be compensated by the junctional diversity of these antibodies. We observed that the junctional diversity in horses antibodies is highly frequent, present in more than 90% of horse antibodies. Besides this, the length of this region seems to be higher in horse antibodies than in other species. N1 and N2 nucleotides addition range from 0 to 111 nucleotides. In addition, around 45% of the antibody clones have more than ten nucleotides in both N1 and N2 junction regions. This diversity mechanism may be one of the most important in providing variability to the equine antibody repertoire. This study provides new insights regarding horse antibody composition, diversity generation, and particularities compared to other species, such as the frequency and length of N nucleotide addition. This study also points out the urgent need to better characterize TdT in horses and in other species to better understand antibody repertoire characteristics.

Introduction

The effective humoral immune response depends partly on having a variety of B cells with different B cell receptors (BCRs) capable of recognizing and binding to many different antigens. The entire set of B cells with different BCRs is called the antibody repertoire (Glanville et al., 2009) . In humans, it has the theoretical potential to reach a size of up to 10^{16}-10^{18} unique antibody sequences (Briney et al., 2019).

The sequencing of the repertoire (Rep-seq) revolutionized the antigen B-cell receptors studies, allowing deep and quantitative analysis to decipher the role of adaptive immunity in health and disease (Georgiou et al., 2014). However, besides being less common, antibody repertoire analysis in species such as chicken, sheep, pig, cattle, and horses revealed new insights into the many different mechanisms that can create antibodies diversity in vertebrates (Butler et al., 2009; Liljavirta et al., 2014; Reynaud et al., 1989; Sun et al., 2012).

Particularly, horses (*Equus caballus*) polyclonal antibodies have been produced and used since the century XIX for the treatment and prophylaxis of diseases such as diphtheria, tuberculosis, tetanus, and pneumonia (ANDERSON, 1955; Cole & Moore, 1917; Glatman-Freedman & Casadevall, 1998; Gonçalves et al., 2007; Lang et al., 2000) to the present day. It is even being used in the current COVID-19 pandemic as a treatment in some countries (Cunha et al., 2020; Zylberman et al., 2020).

Similar to other vertebrates, horses have three types of immunoglobulin chains: light lambda (IGL), light kappa (IGK), and heavy (IGH). The horse antibody V(D)J gene segments were annotated by Sun et al. (2010) and reviewed by Walter et al. (2015), using an EquCab 2.0 genome composed of several scaffolds. After that, the EquCab3.0 genome was published, and the international ImMunoGeneTics information system® (IMGT®) annotated the IG locus. In this annotation, the horse IGH locus present on chromosome 24 has 104 IGHV (21 functional, 74 pseudogenes, and nine ORFs), 44 IGHD (16 functional, twenty-eight ORFs), and nine IGHJ (six functional, and three ORFs).

So far, analyzes of the repertoire of equine antibodies have been carried out by different methodologies, most of them by Sanger sequencing, with low deepness (Almagro et al., 2006; Tallmadge et al., 2013, 2014) . Only in 2019, our group carried out a deeper horse antibody

repertoire analysis using the new generation technology (NGS) was carried out, showing some characteristics of 45,000 IGH clones and 30,000 IGL clones as new gene transcripts (IGHV6S1 and IGLV4S2) and the amino acids composition and features of CDR-H3 (Manso et al., 2019). However, some essential horse antibody repertoire characteristics are still unclear, such as somatic hypermutation frequency and the characteristics of the junction, among others. Furthermore, the fraction of the potential repertoire expressed in an individual is unknown, and how similar repertoires are between individuals who have lived in similar environments.

We present a deeper horse antibody heavy chain repertoire (IGH) characterization of non-immunized horses using HTS technology, where we obtained a mean of 248,169 unique IgM clones and 66,141 unique IgG clones from four domestic adult horses. Sequence coverage was between 52 and 82% in IgM and IgG isotypes. We observed that the IGHV4 subgroup is expressed in around 80% of horse's antibodies, and between 50% and 56% use IGHJ6 indicating limited use of combinations of gene segments. However, most horse antibody IgM and IgG clones (~91%) present N-nucleotide addition, reaching 78 nucleotides in N1 and 62 in N2 regions for IgM and 111 nucleotides in N1 and 104 in N2 for IgG. These results suggest a major role of junctional diversity in generating equine antibody repertoire variability.

MATERIALS AND METHODS

Horse blood samples

The peripheral blood samples from four healthy, mixed male breed adult horses, aged 5 to 9 years old, were obtained in partnership with the Immunobiological Research and Production Center (CPPI) of the State of Paraná.

About 35 ml of peripheral blood was obtained from each animal using Vacutainer tubes with EDTA anticoagulant. The PBMC were isolated by Ficoll-PaqueTM gradient centrifugation. The cells ($1x10^7$ cells) were cryopreserved in FBS 90%/ DMSO 10% at -80 °C until use.

The Ethics Committee approved the experimental design on the Use of Animals of the Federal University of Minas Gerais (CEUA - UFMG) under protocol number 190/2018.

Amplification of the horse antibody BCR repertoire

Mononuclear cells (PBMCs) were isolated for RNA extraction and subsequent cDNA synthesis. Total RNA extraction was performed by the TRIzol method (Rio et al., 2010), and the RNA

concentrations were verified by the Qubit RNA BR Assay kit (Thermo Fisher Scientific). According to the manufacturer's instructions, approximately 500 ng of RNA was used for cDNA synthesis using the SuperScript IV enzyme (Thermo Fisher Scientific). The IGH amplification of the gene segments V and the constant region was carried out by multiplex PCR. A set of forward specific primers (F) for the heavy chain variable region (Manso et al., 2019)was used with new reverse specific primers (R) for the heavy chain constant region designed in this study: IgM isotype 5' ATGACGTTGGGTAGGAAGTCCCG 3' and IgG isotype 5' CCACCGTGGMGTCAGAYGTG 3'. All primers have incorporated the Illumina overhang adaptors sequence to prepare the Illumina library.

Multiplex PCR reactions were conducted to obtain IGH amplicons from each of the four horses. All reactions were prepared to contain 10X High Fidelity buffer, 50 mM MgSO4, 10 mM dNTPs, 0.5 µM of each F primer, 0.5 µM of each R primer, and 0.5 U Taq DNA polymerase Platinum High Fidelity (Thermo Fisher Scientific). The cycling parameters were 94 °C for 2 min; 4 cycles of 94 °C for 1 min, 50 °C for 1 min and 72 °C for 1 min; 4 cycles of 94 °C for 1 min, 55 °C for 1 min and 72 °C for 1 min; 26 cycles of 94 °C for 1 min, 63 °C for 1 min and 72 °C for 1 min, and 72 °C for 7 min. The amplifications were analyzed on 1% agarose gels and stained with Sybr Safe (Invitrogen). The bands were excised, and purified with PCR clean-up Gel extraction (NucleoSpin).

Library preparation and sequencing

The purified cDNAs were quantified by Qubit DNA High Sensitivity kit (Thermo Fisher Scientific). Then, it was used for sequencing libraries prepared by the Nextera XT DNA Library Prep kit (Illumina) according to the manufacturer's instructions. The P5 and P7 indexes and adapters were incorporated into the 500 bp amplicons by the overhang adapters added to the primers. The library concentration was verified using Qubit DNA High Sensitivity kit (Thermo Fisher Scientific), and the size and quality of amplicons were confirmed with the Bioanalyzer High Sensitivity DNA Analysis (Agilent).

The IGH samples (18 pM) from the four equines were sequenced using Illumina MiSeq platform 2 × 300 bp read length. The sequences have been deposited at the NIH SRA (Sequence Read Archive) under accession number PRJNA851406.

Bioinformatic analysis of the horse immunoglobulin heavy chain (IgH) variable-region repertoire

The reads were preprocessed by the pRESTO pipeline (vander Heiden et al., 2014), and the IG genes were annotated using IMGT/HighV-QUEST (Alamyar et al., 2012) . The unproductive V(D)J rearrangements were eliminated from the dataset, as well as the productive sequences containing insertions, deletions (indels), or stop codons in V- and J-gene segments. The sequences with the same VJ segment and identical CDR H3 size were grouped using the IMGT/StatClonotype (Aouinti et al., 2016) for clonotype analyses.

After processing the sequences, analyses of the antibodies diversity were conducted, evaluating the frequencies of gene segment usage, gene subgroups, and the combination of V(D)J genes in each animal using IMGT/StatClonotype. The size, composition, and amino acids groups (defined by Crooks et al., 2004) of CDR-H3 amino acid sequences (numbered according to IMGT) (Lefranc et al., 2003) were analyzed using R studio. R studio was also used to get public repertoire antibodies defined as different horses containing antibodies with the same V and J and CDR3. To determine the reading frame (RF) of IGHD genes, we first determined the hydrophobicity index, according to Kyte-Doolittle scale, of each frame using R studio peptide package. The most hydrophilic reading frame was definied as RF1, the most hydrophobic as RF2, and the one hydrophobic with stop codons was defined as RF3 (Ivanov et al., 2002)..

Other parameters such as somatic hypermutation and junction were analyzed from data extracted from IMGT/HighV-QUEST.

Rarefaction analysis and constructing species-richness curves clonotypes.

We used the program iNEXT (Hsieh et al., 2016) to subsample populations of clonotypes from immunoglobulin heavy chains that belonged to four horses based on the frequency of their occurrence in productive reads. iNEXT was also used to extrapolate beyond the number of experimentally observed productive reads that we might expect with additional sequencing.

Recon (Kaplinsky & Arnaout, 2016) was used to estimate the number of missing clonotypes in the immunoglobulin heavy-chain datasets

Statistical analysis

To compare antibody isotypes (IgG and IgM) differences, we used the Shapiro-Wilk test followed by the Mann-Whitney test.

For all analyses, the media of each horse-specific parameter followed by the average media of all four horses, and also the standard deviation of the average media of the 4 horses were used.

RESULTS

Restricted VJ gene usage in horse antibodies

We analyzed IgG and IgM variable heavy chains from four individual horses. Overall, the mean of raw reads per IgM samples was 1,082,148 (354,598- 1,558,035) and 347,302 (210,964-481,150) for IgG samples. We obtained 31-64% of productive reads and between 40,018 to 328,300 horse antibody clones (Table 1).

The EquCab3.0 horse's genome includes 21 IGHV, 16 IGHD, and 6 IGHJ functional gene segments, leading to 2,016 possible germline coding antibodies. However, in our study, we observed a strong preference for IGHV4 subgroup gene segments, where the IGHV4-21, IGHV4-29, and IGHV4-22 are used by 80% of the horse antibodies in both IgM and IgG isotypes (Figure 1A). In addition, only 13 (of which three are present in less than 0.1% of the antibodies) from 21 IGHV seem to be used in horse antibodies.

Similarly, IGHJ4 and IGHJ6 are the preferred J gene used by both IgM and IgG isotypes (Figure 1B). Interestingly, IGHJ6 is present in almost 60% of all horse antibody clones, showing a restricted use of IGHV and IGHJ gene segments. All the 16 functional IGHD genes seem to be used by horses' antibodies (Supplementary Figure 1).

The most frequent VDJ combinations used by horse antibodies were IGHV4-21, IGHD2-26, and IGHJ6-1, found in 2.8% (±0.9) of the IgM and 2.6% (±0.6) IgG isotypes (Figure 1C).

Our analysis also showed that more rare clones were observed in IgM samples than in IgG since the rarefaction curves do not begin to plateau, indicating that we were unlikely to capture this population's full diversity (Figures 2A and 2B). However, we were able to capture between 52 to 66% and between 62 to 82% of all IgM and IgG, respectively (Figure 2C), with no statistical difference between IgM and IgG horse antibodies diversity compared to Shannon's (Figure 2D) and Simpsons test (Figure 2E).

Public horse antibody repertoire is enriched in shorter CDR-H3

In this study, we observed that the four horses shared (public repertoire) shared only 0.05% (44 clones) of their IgM repertoire and 0.0099% of the IgG repertoire (4 clones) (Figures 3A and 3B). For the IgM public repertoire, most of the clones present the IGHV4-21 gene (77%) of the IGHV genes in the IgM public repertoire, while in the total repertoire, it represents approximately 32% (Figure 1A). In the case of IGHJ genes, we observed an increased gene usage of IGHJ4 (from

4% to 9%) and IGHJ5 (from 29 to 32%) in the public IgM repertoire (Figure 1B).

Interesting to note that more than 90% of the CDR-H3 found in the IgM public antibody repertoire presented only five amino acids length (Figure 3C). In general, the CDR-H3 size distribution of horses follows a bi-modal pattern, with sizes ranging from 4 to 51 amino acids residues with a median length of 16 residues for both IgM and IgG isotypes (Figure 3C).

Interestingly, polar amino acids such as glycine (G) and tyrosine (Y) are increased in IgM public repertoire, differently from the acidic aspartic (D) and glutamic acids (E) and the hydrophobic phenylalanine (F), tryptophan (W) and alanine (A) that decreases (Figure 3D).

Our results suggest that the horse IGHV repertoire appears to be derived from limited germline gene families.

Characterization of somatic hypermutation (SHM) frequency and pattern found in horse antibodies

Based on our previous results, we hypothesized that the biggest horse immunoglobulin diversity comes from somatic hypermutation and junctional diversity.

Here, we observed that the frequency of mutations in IgG isotype (media: 7.22%) was similar to IgM (media: 6.46%) compared to their germline mapped on EquCab3.0 genome (Figure 4A). The majority of mutations were found in CDR regions, especially at positions 32 (CDR1), 50, 52 and 58, from CDR2 and 88 (FR3) for both IgM and IgG isotypes (Figure 4B). We also observed that an average of 16.79% of nucleotides are mutated in CDRs of region IGHV, from which the majority of them (45 to 51%) are present in AID motifs (RGYW and complementary WRCY nucleotide motifs) (Supplementary Table 1).

Characterization of Horse Antibodies Junctional Diversity

An essential source of antibody diversification is the addition and deletion of nucleotides between VDJ junctions. Therefore, we analyzed the occurrence of the P/N nucleotide addition and exonuclease trimming for both IgM and IgG horse antibodies. We observed very similar characteristics in all the junctional regions of IgM and IgG horse antibodies (Figures 5A and 5B). N1 and N2 nucleotides media vary from 8.6 to 9.2 present in around 92 % of the IgM and IgG antibody clone (n = 998,756 clones for IgM, n = 264,566 clones for IgG (Figura 5A and 5B, Table 2). Interestingly, 43-44% of the antibodies have N1 (ranging from 10 to 111 nt) and N2

junctions (ranging from 10 to 104 nt) with 10 or more nucleotides, and 6.9 to 9.8% of these regions with 22 or more nucleotides (Table 2). It was also possible to observe that half of the 10 biggest CDR3 present cysteines (Supplementary table 4). We also noticed that N1 region is highly enriched in G (35.59%), and the N2 region is enriched in G and T (30%) for both isotypes (Supplementary Table 2).

Similarly, exonuclease trimming was observed in around 70% to 97% of the Ig clones, with the biggest number of nucleotides trimmed in the IGHJ gene ends (mean of 10 for IgM and 11 for IgG) (Figure 5A and 5B). When analyzing the components that contribute to the length of the CDR-H3, we observed that an average of 15 nucleotides from IGHD gene segment contribute to the length of the IgM and IgG CDR3s with contain an average of 54 nucleotides (around 26-27% of IGHD contribution to the CDR3). Surprisingly, when the 5 biggest CDR-H3 of the IgG clones where anlysed we observed a contribution of 12 to 22 nucleotides of IGHD genes which represents only 9% to 12% of the CDR3. The biggest contribution in these cases came from N1 addition that can contribute with 111 nucleotides of a CDR3 with 150 nucleotides (74%) or the N2 addition that can contribute with 91 nucleotides of a CDR3 containing 129 nucleotides (70%) (Figure 5C).

We also observed a preference for the use of RF1 (more of 80%) in the horse's antibodies (Supplementary Figure 2), enriched in polar amino acids such as tyrosine and glycine (Table 2). Of the 96 possible sets of IGHD amino acid sequences (16 x 6 = 96), 36 (37.5%) include one or more tyrosine, while only 13 (13.5%) have one or two arginines (Supplementary Table 3).

Discussion

In this work, we investigate the antibody-heavy chain repertoire of four different horses, presenting the largest collection of adaptive immune receptor sequences described to date for horses. We analyzed 40,018 to 328,300 horse IgG or IgM clones, with good coverage, from 52 to 82% of the repertoire. Similar to this work, it was observed a difference in depths for IgM (36%) when compared to IgG (64%) for human antibodies (Galson et al., 2015) . Although not much difference was observed between IgM and IgG isotypes, this is the first high-throughput sequencing study that characterizes both isotype's horse repertoires.

Interestingly, we found that approximately 80% of the IgM and IgG antibodies present the IGHV4 group as a gene segment used in their antibodies, corroborating works from, Similar results were also observed by Tallmadge and collaborators (2013) using 5' RACE, in which they found a strong preference (80%) for the IGHV4-29 (previously called IGHV2S3) and IGHJ6-1 (55%) (previously called IGHJ1S5) usage in adult horses' antibodies (Chaudhary & Wesemann, 2018; Manso et al., 2019; Sun et al., 2010; Tallmadge et al., 2013). Similar, humans antibodies also have a preference for the IGHV4 family(Arnaout et al., 2011), differing from other organisms like cattle, dogs, and mice that present a predominance for IGHV1 genes in their antibodies and cats with the presence of IGHV3 genes (Pasman et al., 2017; Rettig et al., 2018; Steiniger et al., 2014, 2017). We also observed a predominance of IGHJ6 in horse antibodies, while dogs and cats antibodies se mostly IGHJ4, and mouse, the IGHJ1 group (Arnaout et al., 2011; Steiniger et al., 2014, 2017). It is interesting to note that horse antibody repertoire is highly dominated by only a few IGHV and IGHJ genes, even if they can use all of their theoretical germline combinations. We observed a high frequency of antibodies (2.8 ±0.9 %) containing IGHV4-21, IGHD2-26, and IGHJ6-1 gene segments combination, also identified in previous work on non-immunized horses (Manso et al., 2019; Tallmadge et al., 2013). This result

is not different from human antibody repertoires (Arnaout et al., 2011), where 0.1% to 2.7% of sequences have the same V(D)J combinations.

In addition, even for a few clones, we observed the presence of a public horse antibody repertoire in the absence of any specific immune stimulation. We found more clones in the public IgM (0.05%) repertoire than in the public IgG repertoire (0.009%). The small number of shared public antibodies clones can be due to the high diversity of horses antibodies but can also be due to an artifact of the clonotyping method used in this work that considers the same clone only antibodies with identical CDH3 region (IMGT/HighVQuest). The observation of more public IgM (1.4%) than IgG (0.3%) and IgA (0.5%) was also observed in the human antibody (Galson et al., 2015. Interestingly, even with a smaller number of public antibodies, we found differences between the CDR-H3 length of the public and the entire repertoire, observing a higher percentage of short CDR-H3 in the public repertoire, also observed in human (Briney et al., 2019; Galson et al., 2014; Soto et al., 2019). It is supposed that B cells expressing receptors with short CDR-H3 are selected because they increase their affinity for the antigen, make clonal expansion, and differentiate in plasma cells or memory B cells (Rosner et al., 2001).

In this work, we also evaluated the SHMs in the IGHV gene segment of the animals (FR1, CDR1, FR2, CDR2, and FR3 regions). The IgG sequences showed a similar mutation frequency than the IgM sequences, probably due to limited pathogen stimulation since they are not immunized. To better understand how mutations are distributed along the IGHV gene segment, we evaluated the number of mutations present in each position for both studied Ig isotypes. We observed a similar mutation profile between the IgG and IgM, with more mutations found in the CDR regions, even in these non-immunized animals. This corroborates previous studies in adult horses (Tallmadge et al., 2013) and healthy and HIV patients (Bowers et al., 2014).

The mechanism of variability that produces SHM is carried out by the enzyme cytidine deaminase (AID) induced by activation, by deamination of the cytosine base, creating a U:G mismatch. The AID targets SHM mutations on "hotspots" (complementary RGYW and WRCY nucleotide motifs) (Spencer & Dunn-Walters, 2005). This work observed that around 19% of IGHV sequences present AID motifs, similar to human IGHV, presenting 17.8% of these motifs (Bowers et al., 2014). In addition, as in the human repertoires, we found a higher percentage of

motifs in the CDR (average of 13.5%) than in the FR (average of 5.2%), as well as a higher number of mutated nucleotides in this region (Bowers et al., 2014).

It is important to highlight that, besides a very similar VJ gene usage in IgM and IgG antibodies, and also a very similar profile of SHM, the IgM and IgG clones looks very dissimilar according to Bray-Curtis dissimilarity index (data not show).

Very little has been described the characteristics of horse junctional diversity in horse antibodies. Here, we observed N1 and N2 nucleotide additions in most IgM and IgG clones, and also observed exceptionally long N nucleotide additions for both N1 and N2 IgG (10–111 bp) in around 44% of the antibodies. Non-template additions to IGH genes have been reported in humans and mice (Shi et al., 2014), pigs (Šinkora et al., 2003), and cattle (Liljavirta et al., 2014). The mean number of nucleotide addition in humans is 6.6 ± 4.3 in N1 and 6.4 ± 4.6 in N2, while in mice, it is 2.4 ± 2.2 and 2.1 ± 1.8 in N1 and N2, respectively (Shi et al., 2014). The average number of nucleotide additions in cattle that present ultralong CDR3s is not particularly high (2.5 in N1 and 2.6 in N2), reflecting the high frequency (35%) of unions with zero additions (Liljavirta et al., 2014). Such frequency and length of N additions have not been reported in other species, suggesting that this diversity mechanism is essential to generating variability in equine immunoglobulins.

Extensive trimming of IGHD genes in horse antibodies (mean value of 6 and 7.9 nucleotides for the 3D and 5D junction, respectively) observed is not very different from cattle antibodies (trimming of 5 to 6 nucleotides) (Liljavirta et al., 2014). It is interesting to note that the larger trimming followed in horse antibodies was observed in the IGHJ gene, ranging between 10 and 11 nucleotides, while in other species, such as cattle, humans, and mice, have respectively 2, 6, and 4 nucleotides trimmed in this region (Liljavirta et al., 2014, Shi et al., 2014). In our data, between 70% to 97% of antibody clones had nucleotides deleted anywhere in the junction, similar to other species (Liljavirta et al., 2014; Shi et al., 2014).

This impressive N nucleotide addition frequency and length can be due to differences in Terminal deoxynucleotidyl transferase (TdT) enzyme activity in horses compared to other species.

These enzymes are composed of mainly two regions, a catalytic core composed of finger, palm, and thumb domains at the C-terminus and a BRCA1 (breast cancer susceptibility protein) C-terminal (BRCT) domain at the N-terminus. When comparing the horse sequence of TdT

isorform 1 and 2 with human, mouse, pig, and cattle TdT we observed the conservation of the catalytic aspartic acids and the substrate-specific loop 1 (Figure 6). Interestingly, the palm domain region, between the first aspartic acids and the loop1, is one of the most different regions between the TdT from other species.

In addition to this region, we can also observe a very dissimilar region between the BRCT domain and TdT catalytic core. For the best of all knowledge, it is unclear how this non-enzymatic domain contributes to the unique biological function of TdT. Interesting to note that this interdomain region is enriched in Proline amino acids. It looks like for DNA polymerase lambda, which presents a bigger Proline-rich domain in this region compared to TdT, this region can impact DNA polymerase fidelity and with BRCT domain can act cooperatively to promote primer/template realignment between DNA strands of limited sequence homology (Fiala et al., 2006; Taggart et al., 2014). Since the TdT template and untemplated activities (Loc'h et al., 2016) are proposed to be essential for antibodies diversity, future studies need to investigate the role of the interdomain region in these activities, as well as the role of the region in palm domain in between aspartic acids and loop1.

This study also observed that more of 80% of the horse antibodies use reading frame 1 (RF1) for both the IgM and IgG isotypes. Several species, such as humans, mice, and sharks, produce antibodies using IGHD reading frame 1 (RF1) (Schroeder et al., 2010). However, similarly to other species RF1 used for horses antibodies is strongly enriched in tyrosine, representing 39.55% of the IGHD amino acids. We know that tyrosine is the amino acid that typically makes the most significant contribution to binding affinity at protein ligand-receptor interfaces (Bogan & Thorn, 1998). This suggests that natural selection was operating on immunoglobulin diversity gene segments to restrict and control their evolution in such a way as to influence the composition and range of diversity of immunoglobulin antigen-binding sites (Burnet, 1976).

Conclusions

This is the first high-throughput sequencing study that characterizes IgM and IgG isotype horse repertoire to the best of our knowledge. We showed a highly restrict use of IGHV and IGHJ genes in horse antibody repertoire in which around 80% of the antibodies are composed by only 3 IGHV gene segments and almost 60% of them with the same IGHJ gene segment. We

observed a complex and diverse repertoire for IGH, given mainly by the junctional diversity, much bigger and frequent than the one present in other species. Our study on the equine antibody repertoire contributes to understanding the generation of their diversity and open up new questions about horse TdT particularities to generate such diversity.

Author Contributions

CN: designed the study, did all the experiments, analyzed the data and wrote the paper; TM: designed and validated primers, designed the study and discussed the results; FM: help in data analysis; LM: helped in the initial analysis of the data; RM: helped Illumina library preparison and sequenced the samples; JM, BA: collected horse samples; AV: discussed the results; JMD, GI: help primer design, study design and wrote the paper; LFF: designed the study, discussed and analyzed the data and wrote the paper

Conflict of Interest

The authors declare that the research was conducted without any commercial or financial relationships that could be construed as a potential conflict of interest.

Fundings

Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brazil (CAPES) [grant numbers 88887.506611/2020-00, 88887.504420/2020-00 and 935/19 (COFECUB)]; Fundação de Amparo a Pesquisa de Minas Gerais (FAPEMIG) [grant numbers PPM-00615-18, Rede Mineira de Imunobiologicos grant #REDE-00140-16]; Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) [Pq to LFF]; National Institutes of Health (NIH) [grant number 1R01AI143552-02]; Pro-Reitoria de Pesquisa da Universidade Federal de Minas Gerais.

Acknowledgments

SynBiom group for fruitful discussions, specially Dr. Marcella Nunes de Mello-Braga and Dra. Marcele Rocha Neves Rocha. A special acknowledge to Regina Maria Fernandes for project management.

References

Alamyar, E., Giudicelli, V., Li, S., Duroux, P., & Lefranc, M. P. (2012). IMGT/Highv-quest: The IMGT® web portal for immunoglobulin (IG) or antibody and T cell receptor (TR) analysis from NGS high throughput and deep sequencing. Immunome Research, 8(1), 1–15. https://doi.org/10.4172/1745-7580.1000056

Almagro, J. C., Martinez, L., Smith, S. L., Alagon, A., Estevez, J., & Paniagua, J. (2006). Analysis of the horse VH repertoire and comparison with the human IGHV germline genes, and sheep, cattle and pig VH sequences. Molecular Immunology, 43(11), 1836–1845. https://doi.org/10.1016/j.molimm.2005.10.017

ANDERSON, C. G. (1955). The distribution of diphtheria antitoxin in pepsin-digested horse antiserum. The Biochemical Journal, 59(1), 47–52. https://doi.org/10.1042/bj0590047

Aouinti, S., Giudicelli, V., Duroux, P., Malouche, D., Kossida, S., & Lefranc, M. P. (2016). IMGT/statclonotype for pairwise evaluation and visualization of NGS IG and TR IMGT clonotype (AA) diversity or expression from IMGT/HighV-QUEST. Frontiers in Immunology, 7(SEP), 1–14. https://doi.org/10.3389/fimmu.2016.00339

Arnaout, R., Lee, W., Cahill, P., Honan, T., Sparrow, T., Weiand, M., Nusbaum, C., Rajewsky, K., & Koralov, S. B. (2011). High-resolution description of antibody heavy-chain repertoires in humans. PLoS ONE, 6(8). https://doi.org/10.1371/journal.pone.0022365

Bogan, A. A., & Thorn, K. S. (1998). Anatomy of hot spots in protein interfaces. Journal of Molecular Biology, 280(1), 1–9. https://doi.org/10.1006/jmbi.1998.1843

Bowers, E., Scamurra, R. W., Asrani, A., Beniguel, L., MaWhinney, S., Keays, K. M., Thurn, J. R., & Janoff, E. N. (2014). Decreased mutation frequencies among immunoglobulin G variable region genes during viremic HIV-1 infection. PLoS ONE, 9(1), 1–13. https://doi.org/10.1371/journal.pone.0081913

Briney, B., Inderbitzin, A., Joyce, C., & Burton, D. R. (2019). Commonality despite exceptional diversity in the baseline human antibody repertoire. Nature, 566(7744), 393–397. https://doi.org/10.1038/s41586-019-0879-y

Burnet, F. M. (1976). A Modification of Jerne's Theory of Antibody Production using the Concept of Clonal Selection. CA: A Cancer Journal for Clinicians, 26(2), 119–121. https://doi.org/10.3322/canjclin.26.2.119

Butler, J. E., Wertz, N., Deschacht, N., & Kacskovics, I. (2009). Porcine IgG: Structure, genetics, and evolution. Immunogenetics, 61(3), 209–230. https://doi.org/10.1007/s00251-008-0336-9

Chaudhary, N., & Wesemann, D. R. (2018). Analyzing immunoglobulin repertoires. Frontiers in Immunology, 9(MAR), 1–18. https://doi.org/10.3389/fimmu.2018.00462

Cole, R., & Moore, H. F. (1917). The production of antipneumococcic serum. Journal of Experimental Medicine, 26(4), 537–561. https://doi.org/10.1084/jem.26.4.537

Crooks, G., Hon, G., Chandonia, J., & Brenner, S. (2004). NCBI GenBank FTP Site\nWebLogo: a sequence logo generator. Genome Res, 14, 1188–1190. https://doi.org/10.1101/gr.849004.1

Cunha, L. E. R., Stolet, A. A., Strauch, M. A., Pereira, V. A. R., Dumard, C. H., Souza, P. N. C., Fonseca, J. G., Pontes, F. E., Meirelles, L. G. R., Albuquerque, J. W. M., Sacramento, C. Q., Fintelman-Rodrigues, N., Lima, T. M., Alvim, R. G. F., Zingali, R. B., Oliveira, G. A. P., Souza, T. M. L., Tanuri, A., Gomes, A. M. O., … Silva, J. L. (2020). Equine hyperimmune globulin raised against the SARS-CoV-2 spike glycoprotein has extremely high neutralizing titers. BioRxiv. https://doi.org/10.1101/2020.08.17.254375

Delarue, M., Boule J.B, Lescar J, Expert-Bezancacon N, Jourdan N, Sukumar N, Rougeon F, & Papanicolaou C. (2002). Crystal structures of a template-independentDNA polymerase: murine terminaldeoxynucleotidyltransferase. The EMBO Jourrnal, 21, 427–439. https://doi.org/0.1093/emboj/21.3.427

Fiala, K. A., Duym, W. W., Zhang, J., & Suo, Z. (2006). Up-regulation of the fidelity of human DNA polymerase λ by its non-enzymatic proline-rich domain. Journal of Biological Chemistry, 281(28), 19038–19044. https://doi.org/10.1074/jbc.M601178200

Galson, J. D., Pollard, A. J., Trück, J., & Kelly, D. F. (2014). Studying the antibody repertoire after vaccination: Practical applications. Trends in Immunology, 35(7), 319–331. https://doi.org/10.1016/j.it.2014.04.005

Galson, J. D., Trück, J., Fowler, A., Münz, M., Cerundolo, V., Pollard, A. J., Lunter, G., & Kelly, D. F. (2015). In-depth assessment of within-individual and inter-individual variation in the B cell receptor repertoire. Frontiers in Immunology, 6(OCT), 1–13. https://doi.org/10.3389/fimmu.2015.00531

Georgiou, G., Ippolito, G. C., Beausang, J., Busse, C. E., Wardemann, H., & Quake, S. R. (2014). The promise and challenge of high-throughput sequencing of the antibody repertoire. Nature Biotechnology, 32(2), 158–168. https://doi.org/10.1038/nbt.2782

Glanville, J., Zhai, W., Berka, J., Telman, D., Huerta, G., Mehta, G. R., Ni, I., Mei, L., Sundar, P. D., Day, G. M. R., Cox, D., Rajpal, A., & Pons, J. (2009). Precise determination of the diversity of a combinatorial antibody library gives insight into the human immunoglobulin repertoire. Proceedings of the National Academy of Sciences of the United States of America, 106(48), 20216–20221. https://doi.org/10.1073/pnas.0909775106

Glatman-Freedman, A., & Casadevall, A. (1998). Serum therapy for tuberculosis revisited: Reappraisal of the role of antibody-mediated immunity against Mycobacterium tuberculosis. Clinical Microbiology Reviews, 11(3), 514–532. https://doi.org/10.1128/cmr.11.3.514

Gonçalves, E. S., Salomão, M. G., & Almeida-santos, S. M. de. (2007). O uso do monitoramento espaço-temporal da expansão urbana no diagnóstico de áreas passíveis de risco epidemiológico peçonhento em Guarulhos-Estado de São Paulo, Brasil. Anais Do XIII Simpósio Braileiro de Sensoriamento Remoto, 3171–3178.

Hsieh, T. C., Ma, K. H., & Chao, A. (2016). iNEXT: an R package for rarefaction and extrapolation of species diversity (Hill numbers). Methods in Ecology and Evolution, 7(12), 1451–1456. https://doi.org/10.1111/2041-210X.12613

Ivanov I, Link J., Ippolito G.C., Schoroeder, H. W. J. (2002). Constraints on the Hydropathicity of HCDR3 are conserved across evolution. Chapter 3, 43-67. The Antibodies (Zaneti Maurizio &

Capra Donald, Eds.; Vol 7).

Kaplinsky, J., & Arnaout, R. (2016). Robust estimates of overall immune-repertoire diversity from high-throughput measurements on samples. Nature Communications, 7(May). https://doi.org/10.1038/ncomms11881

Lang, J., Kamga-Fotso, L., Peyrieux, J. C., Blondeau, C., Lutsch, C., & Forrat, R. (2000). Safety and immunogenicity of a new equine tetanus immunoglobulin associated with tetanus-diphtheria vaccine. American Journal of Tropical Medicine and Hygiene, 63(5–6), 298–305. https://doi.org/10.4269/ajtmh.2000.63.298

Lefranc, M. P., Pommié, C., Ruiz, M., Giudicelli, V., Foulquier, E., Truong, L., Thouvenin-Contet, V., & Lefranc, G. (2003). IMGT unique numbering for immunoglobulin and T cell receptor variable domains and Ig superfamily V-like domains. Developmental and Comparative Immunology, 27(1), 55–77. https://doi.org/10.1016/S0145-305X(02)00039-3

Liljavirta, J., Niku, M., Pessa-Morikawa, T., Ekman, A., & Iivanainen, A. (2014). Expansion of the preimmune antibody repertoire by junctional diversity in Bos taurus. PLoS ONE, 9(6). https://doi.org/10.1371/journal.pone.0099808

Loc'h, J., Rosario, S., & Delarue, M. (2016). Structural Basis for a New Templated Activity by Terminal Deoxynucleotidyl Transferase: Implications for V(D)J Recombination. Structure, 24(9), 1452–1463. https://doi.org/10.1016/j.str.2016.06.014

Lu, S., Wang, J., Chitsaz, F., Derbyshire, M. K., Geer, R. C., Gonzales, N. R., Gwadz, M., Hurwitz, D. I., Marchler, G. H., Song, J. S., Thanki, N., Yamashita, R. A., Yang, M., Zhang, D., Zheng, C., Lanczycki, C. J., & Marchler-Bauer, A. (2020). CDD/SPARCLE: The conserved domain database in 2020. Nucleic Acids Research, 48(D1), D265–D268. https://doi.org/10.1093/nar/gkz991

Manso, T. C., Groenner-Penna, M., Minozzo, J. C., Antunes, B. C., Ippolito, G. C., Molina, F., & Felicori, L. F. (2019). Next-generation sequencing reveals new insights about gene usage and CDR-H3 composition in the horse antibody repertoire. Molecular Immunology, 105(August 2018), 251–259. https://doi.org/10.1016/j.molimm.2018.11.017

Pasman, Y., Merico, D., & Kaushik, A. K. (2017). Preferential expression of IGHV and IGHD encoding antibodies with exceptionally long CDR3H and a rapid global shift in transcriptome characterizes development of bovine neonatal immunity. Developmental and Comparative Immunology, 67, 495–507. https://doi.org/10.1016/j.dci.2016.08.020

Rettig, T. A., Ward, C., Bye, B. A., Pecaut, M. J., & Chapes, S. K. (2018). Characterization of the naive murine antibody repertoire using unamplified high-throughput sequencing. PLoS ONE, 13(1), 1–20. https://doi.org/10.1371/journal.pone.0190982

Reynaud, C. A., Dahan, A., Anquez, V., & Weill, J. C. (1989). Somatic hyperconversion diversifies the single VH gene of the chicken with a high incidence in the D region. Cell, 59(1), 171–183. https://doi.org/10.1016/0092-8674(89)90879-9

Rio, D. C., Ares, M., Hannon, G. J., & Nilsen, T. W. (2010). Purification of RNA using TRIzol (TRI Reagent). Cold Spring Harbor Protocols, 5(6), 1–4. https://doi.org/10.1101/pdb.prot5439

Rosner, K., Winter, D. B., Tarone, R. E., & Skovgaard, G. L. (2001). Third complementarity-determining region of mutated V H immunoglobulin genes contains shorter V , D , J , P , and N components than non-mutated genes.

Schroeder, H. W., Zemlin, M., Khass, M., Nguyen, H. H., & Schelonka, R. L. (2010). Genetic control of DH reading frame and its effect on B-cell development and antigen-specifc antibody production. Critical Reviews in Immunology, 30(4), 327–344. https://doi.org/10.1615/critrevimmunol.v30.i4.20

Shi, B., Ma, L., He, X., Wang, X., Wang, P., Zhou, L., & Yao, X. (2014). Comparative analysis of human and mouse immunoglobulin variable heavy regions from IMGT/LIGM-DB with IMGT/HighV-QUEST. Theoretical Biology and Medical Modelling, 11(1), 1–11. https://doi.org/10.1186/1742-4682-11-30

Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Söding, J., Thompson, J. D., & Higgins, D. G. (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Molecular Systems Biology, 7. https://doi.org/10.1038/msb.2011.75

Šinkora, M., Sun, J., Šinkorová, J., Christenson, R. K., Ford, S. P., & Butler, J. E. (2003). Antibody Repertoire Development in Fetal and Neonatal Piglets. VI. B Cell Lymphogenesis Occurs at Multiple Sites with Differences in the Frequency of In-frame Rearrangements. The Journal of Immunology, 170(4), 1781–1788. https://doi.org/10.4049/jimmunol.170.4.1781

Soto, C., Bombardi, R. G., Branchizio, A., Kose, N., Matta, P., Sevy, A. M., Sinkovits, R. S., Gilchuk, P., Finn, J. A., & Crowe, J. E. (2019). High frequency of shared clonotypes in human B cell receptor repertoires. Nature, 566(7744), 398–402. https://doi.org/10.1038/s41586-019-0934-8

Spencer, J., & Dunn-Walters, D. K. (2005). Hypermutation at A-T Base Pairs: The A Nucleotide Replacement Spectrum Is Affected by Adjacent Nucleotides and There Is No Reverse Complementarity of Sequences Flanking Mutated A and T Nucleotides. The Journal of Immunology, 175(8), 5170–5177. https://doi.org/10.4049/jimmunol.175.8.5170

Steiniger, S. C. J., Dunkle, W. E., Bammert, G. F., Wilson, T. L., Krishnan, A., Dunham, S. A., Ippolito, G. C., & Bainbridge, G. (2014). Fundamental characteristics of the expressed immunoglobulin VH and VL repertoire in different canine breeds in comparison with those of humans and mice. Molecular Immunology, 59(1), 71–78. https://doi.org/10.1016/j.molimm.2014.01.010

Steiniger, S. C. J., Glanville, J., Harris, D. W., Wilson, T. L., Ippolito, G. C., & Dunham, S. A. (2017). Comparative analysis of the feline immunoglobulin repertoire. Biologicals, 46, 81–87. https://doi.org/10.1016/j.biologicals.2017.01.004

Sun, Y., Liu, Z., Ren, L., Wei, Z., Wang, P., Li, N., & Zhao, Y. (2012). Immunoglobulin genes and diversity: What we have learned from domestic animals. Journal of Animal Science and Biotechnology, 3(1), 1–5. https://doi.org/10.1186/2049-1891-3-18

Sun, Y., Wang, C., Wang, Y., Zhang, T., Ren, L., Hu, X., Zhang, R., Meng, Q., Guo, Y., Fei, J., Li, N., & Zhao, Y. (2010). A comprehensive analysis of germline and expressed immunoglobulin repertoire in the horse. Developmental and Comparative Immunology, 34(9), 1009–1020. https://doi.org/10.1016/j.dci.2010.05.003

Taggart, D. J., Dayeh, D. M., Fredrickson, S. W., & Suo, Z. (2014). N-terminal domains of human DNA polymerase lambda promote primer realignment during translesion DNA synthesis. DNA Repair, 22, 41–52. https://doi.org/10.1016/j.dnarep.2014.07.008

Tallmadge, R. L., Tseng, C. T., & Felippe, M. J. B. (2014). Diversity of immunoglobulin lambda light chain gene usage over developmental stages in the horse. Developmental and Comparative Immunology, 46(2), 171–179. https://doi.org/10.1016/j.dci.2014.04.001

Tallmadge, R. L., Tseng, C. T., King, R. A., & Felippe, M. J. B. (2013). Developmental progression of equine immunoglobulin heavy chain variable region diversity. Developmental and Comparative Immunology, 41(1), 33–43. https://doi.org/10.1016/j.dci.2013.03.020

vander Heiden, J. A., Yaari, G., Uduman, M., Stern, J. N. H., O'connor, K. C., Hafler, D. A., Vigneault, F., & Kleinstein, S. H. (2014). PRESTO: A toolkit for processing high-throughput sequencing raw reads of lymphocyte receptor repertoires. Bioinformatics, 30(13), 1930–1932. https://doi.org/10.1093/bioinformatics/btu138

Waterhouse, A. M., Procter, J. B., Martin, D. M. A., Clamp, M., & Barton, G. J. (2009). Jalview Version 2-A multiple sequence alignment editor and analysis workbench. Bioinformatics, 25(9), 1189–1191. https://doi.org/10.1093/bioinformatics/btp033

Zylberman, V., Sanguineti, S., Pontoriero, A. v., Higa, S. v., Cerutti, M. L., Seijo, S. M. M., Pardo, R., Muñoz, L., Intrieri, M. E. A., Alzogaray, V. A., Avaro, M. M., Benedetti, E., Berguer, P. M., Bocanera, L., Bukata, L., Bustelo, M. S., Campos, A. M., Colonna, M., Correa, E., … Goldbaum, F. A. (2020). Development of a hyperimmune equine serum therapy for covid-19 in Argentina. Medicina, 80, 1–6.

Tables

Table 1: Overview of the IgM and IgG heavy chain variable sequencing results from four non-immunized horses

Horse Sample	Ig Isotype	Raw Reads	Preprocessed Reads	Annotated Reads	Productive Reads	Clones
1	IgM	1,558,035	779,236	526,221	484,236	294,600
2	IgM	1,499,997	900,306	891,250	790,308	328,300
3	IgM	915,963	690,915	529,584	485,888	283,600
4	IgM	354,598	268,020	203,623	193,644	86,177
1	IgG	210,964	187,229	182,918	175,686	43,861
2	IgG	288,611	233,147	219,402	209,248	40,018
3	IgG	408,483	306,844	304,394	271,204	91,136
4	IgG	481,141	390,341	382,438	361,320	89,551

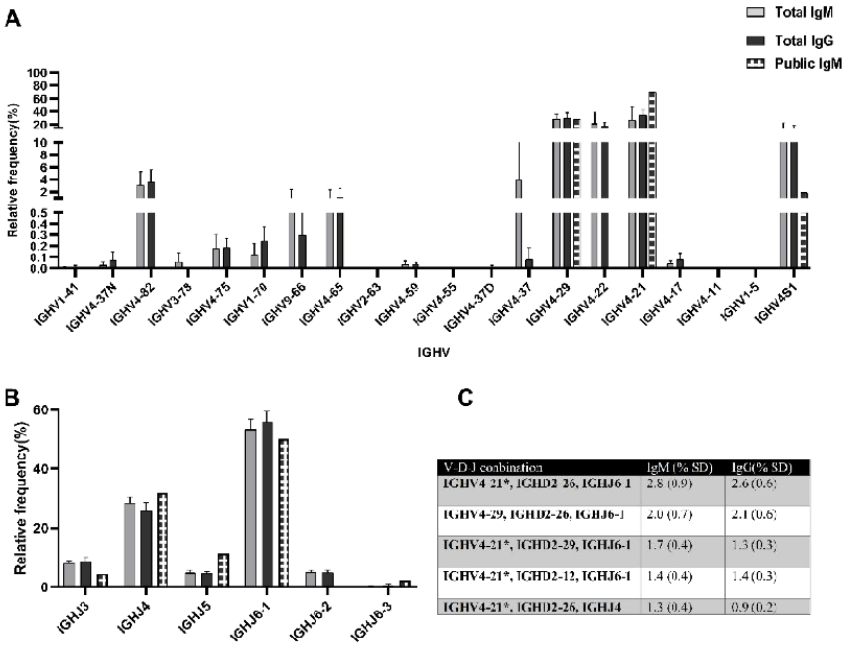
Table 2: Analysis of nucleotide additions in Horse Antibodies

Sample	Number of clones	Mean number of N1 nucleotides (range)	Mean number of N2 nucleotides (range)	Clones with N1 additions longer than 10 nucleotides/ and longer than 22 nt (%)	Clones with N2 additions longer than 10 nucleotides/and longer than 22 nt (%)
IgM	998,756	9.54 (0-78)	9.71 (0-62)	43.95/7.4	43.74/8.6
IgG	264,566	8.72(0-111)	9.24 (0-104)	43.67/7.6	45.17/10.0

Table 3: Amino acid percentage composition per reading frame of horse D gene

	RF1	RF2	RF3	iRF1	iRF2	iRF3	
D	**5.22**	0.75	0	0	0	0	Acid
E	0.74	0	2.90	0	0	3.00	
R	0.74	1.50	4.34	5.97	0.74	0	Basic
K	0.74	0	0	0.74	0	0	
H	0	0	0.72	21.64	1.49	0	
F	0	0	0.72	0	0	0.75	Hydrophobic
P	0	2.25	0	2.24	4.48	17.29	
V	0	27.82	0.72	0.74	29.10	3.00	
L	0	5.26	42.75	1.49	1.49	18.04	
I	2.23	6.77	1.44	0.74	18.66	3.00	
A	5.97	3.00	1.45	3.54	5.26	2.93	
W	3.73	0	13.76	0	0	0	
M	0	18.04	0	0	0	0.75	
N	5.97	6.66	0.72	15.67	2.99	0	Neutral
Q	0	0	2.90	1.49	0	4.51	
C	0.74	0	5.80	5.97	0	2.94	Polar
S	**14.92**	0	0	23.88	4.48	6.77	
G	**18.65**	4.5	1.45	0	5.97	1.50	
Y	**39.55**	0	2.90	11.94	0.74	2.25	
T	0.74	29.32	0	0.74	26.12	1.50	
*	0	0	**17.39**	4.47	0.74	**32.33**	
AA Media by RF	7.88	7.82	8.11	7.88	7.88	7.82	

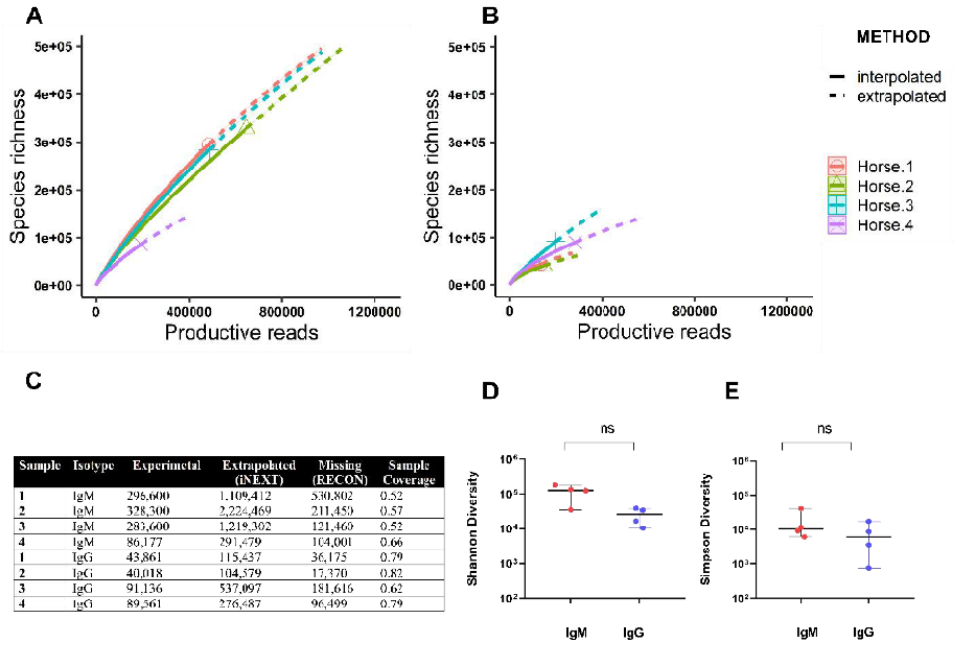
Figures



V-D-J combination	IgM (% SD)	IgG(% SD)
IGHV4-21*, IGHD2-26, IGHJ6-1	2.8 (0.9)	2.6 (0.6)
IGHV4-29, IGHD2-26, IGHJ6-1	2.0 (0.7)	2.1 (0.6)
IGHV4-21*, IGHD2-29, IGHJ6-1	1.7 (0.4)	1.3 (0.3)
IGHV4-21*, IGHD2-12, IGHJ6-1	1.4 (0.4)	1.4 (0.3)
IGHV4-21*, IGHD2-26, IGHJ4	1.3 (0.4)	0.9 (0.2)

Figure 1: IGHV and IGHJ gene segments frequency present in public IgM and total IgM and IgG horses' antibodies.

Median of relative frequency (%) of IGHV (A), IGHJ (B) and V(D)J more frequent combination (C) in IgM and IgG isotype from four horses. The genes are organized in the order that it appears in the EquCab 3.0 genome (5'- 3').



Sample	Isotype	Experimetal	Extrapolated (iNEXT)	Missing (RECON)	Sample Coverage
1	IgM	296,600	1,109,412	530,802	0.52
2	IgM	328,300	2,224,469	211,459	0.57
3	IgM	283,600	1,219,302	121,469	0.52
4	IgM	86,177	291,479	104,001	0.66
1	IgG	43,861	115,437	36,175	0.79
2	IgG	40,018	104,579	17,370	0.82
3	IgG	91,136	537,097	181,616	0.62
4	IgG	89,561	276,487	96,499	0.79

Figure 2: Antibody heavy chain repertoire richness and diversity estimates for IgM and IgG in the four non-immunized horses.

Interpolation and extrapolation of species richness were obtained using iNEXT for IgM (A) and IgG (B). Solid lines correspond to the interpolation (based on experimental data), and the dashed lines belong to the extrapolated data. Summary of estimates for repertoire size, including missing clones (C). The comparison of Shannon's (D) and Simpsons (E) diversity between the IgM and IgG isotypes ($p < 0.05$).

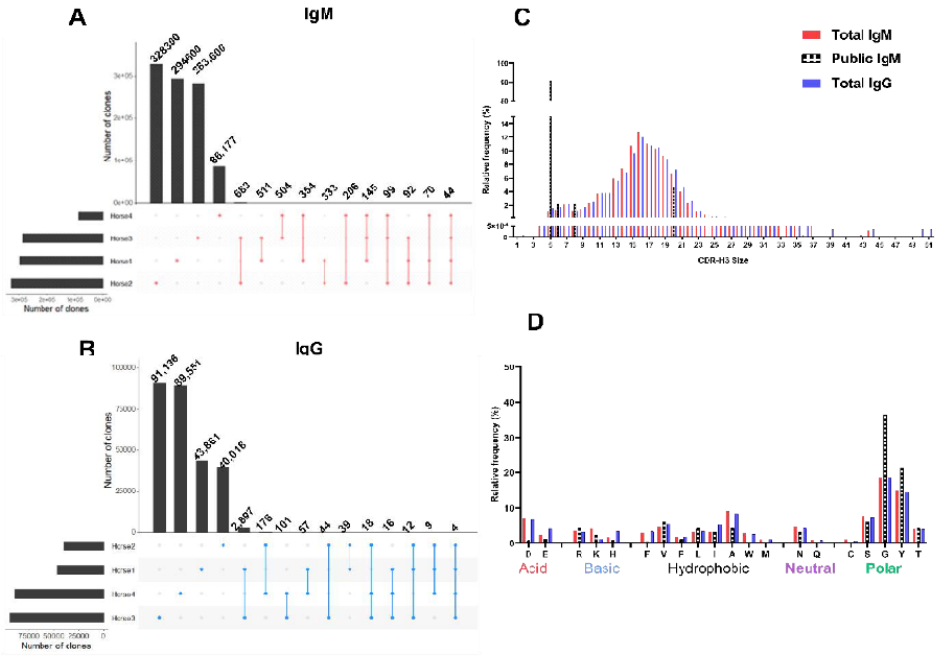


Figure 3: Horse public and private heavy chain variable region repertoire.

The number of antibody clones presented by the different horses and the shared number of clones between the 2, 3, or 4 horses in this study for the IgM (A) and IgG (B). Comparison of the CDR-H3 length (C) and amino acid composition (D) between the Total IgM, Total IgG, and the public IgM repertoire.

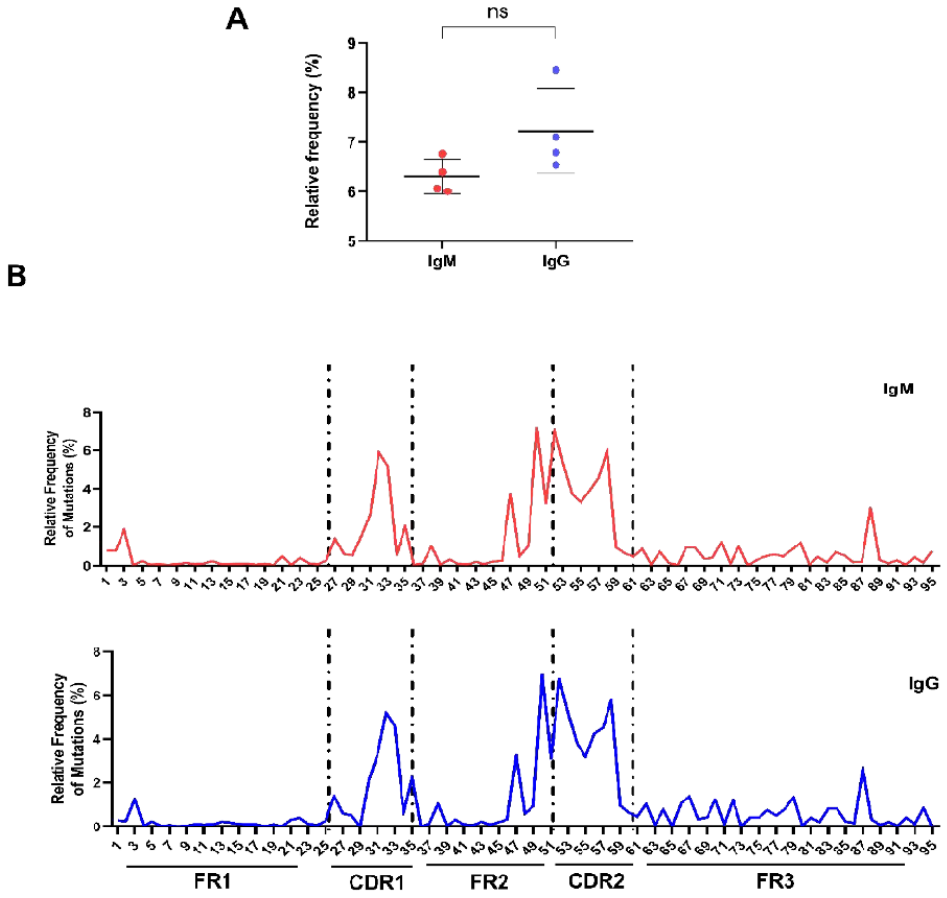


Figure 4: Somatic Hypermutation (SHM) characteristics of Horse IgG and IgM heavy chain variable region.

(A) Media of SHM frequency (%) at the IGHV gene segment from IgM and IgG isotypes ($p < 0.05$). (B) The number of mutations by amino acid position in the IGHV gene segment of the horses' heavy chain (According to the IMGT numbering, without gaps). The dotted lines delimit the FR and CDR regions . IgM is shown in red and IgG in blue.

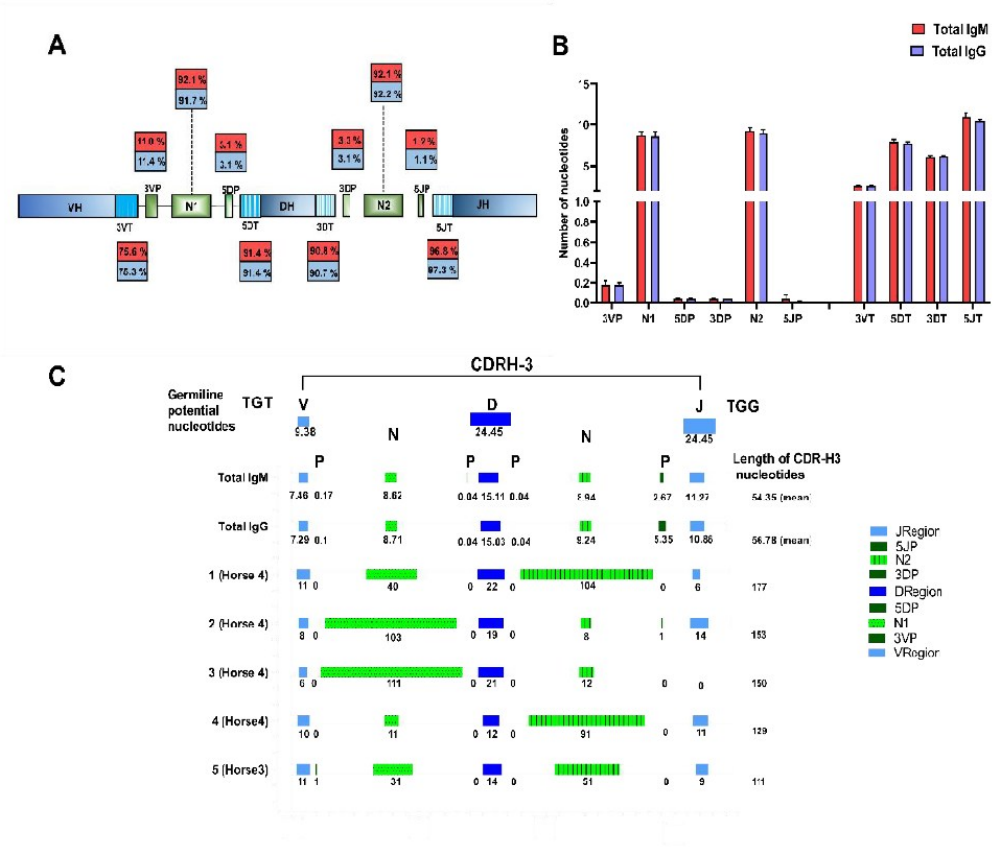


Figure 5: VDJ junction analysis.

(A) Junctional modifications schema during VDJ rearrangement, showing the locations and the occurrence of different types of junctional modifications. Into the box are the mutation frequency (%) of each junctional segment in all the antibodies clones analyzed: 3VP and 3VT for 3′V region, 5DP and 5DT for 5′D genes, 3DP and 3DT for 3′D region, and 5JP and 5JT for 5′J genes, where P means palindromic nucleotides additions, and T means exonuclease trimmings; N1 the non-template randomized nucleotides additions at the 3′V and the 5′D genes; N2 for N additions at the 3′D and the 5′J genes. (B) The median number of nucleotides per junction region added or trimmed.(C) Deconstruction of the components that contribute to the length of the CDR-H3 in

the media of total of the IgM and IgG clones, as well as the 5 biggest CDR-H3 of the IgG clones. The mean of nucleotides of the germline sequence of the VH gene segment, P and N junctions, the DH gene segment, and the JH gene segment to the CDR-H3 length is illustrated.

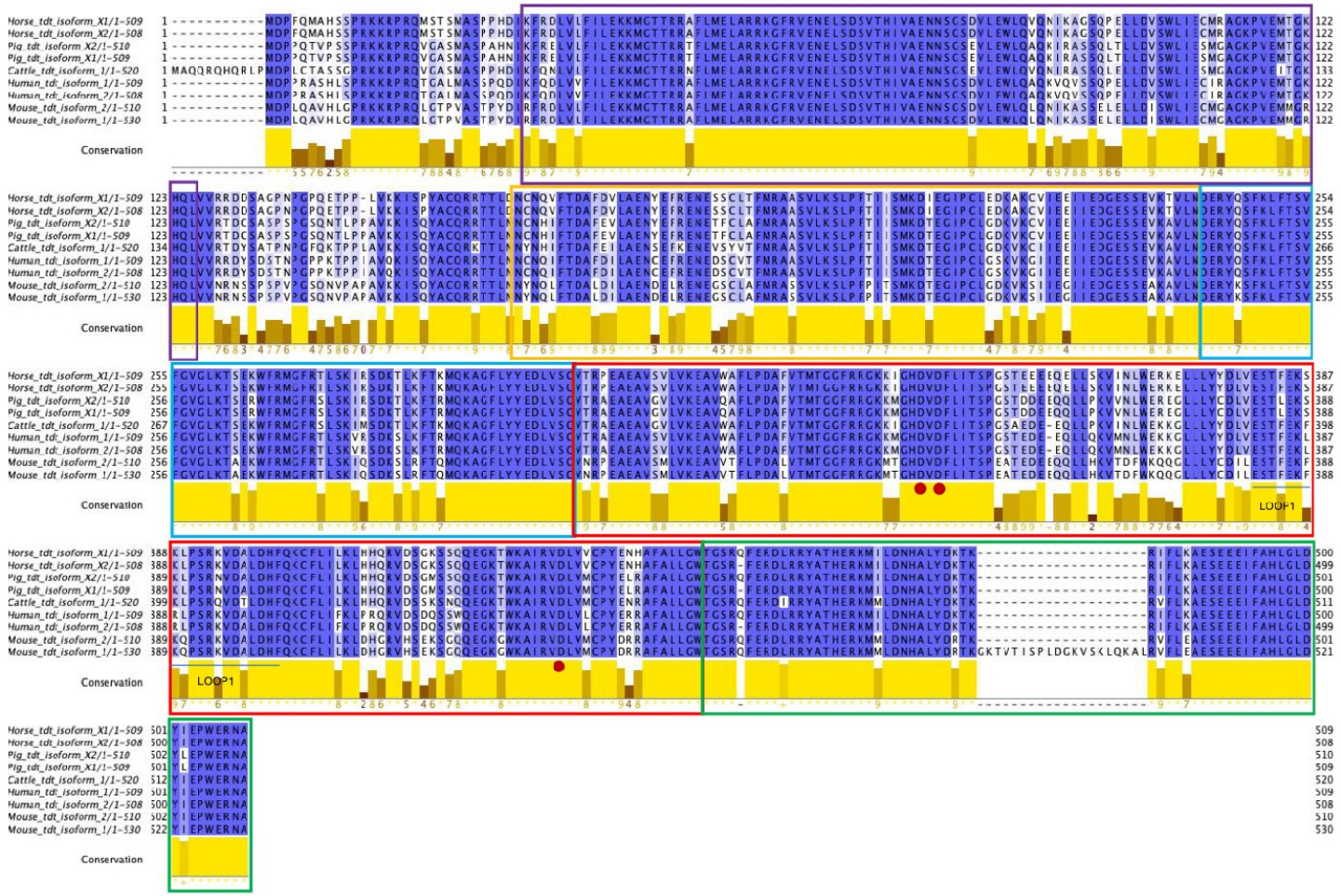


Figure 6: Multiple Sequence alignment of TdT from different species. An amino acid sequence alignment of the two equine TdT isoforms with other vertebrate TdT was done using the Clustal Omega program (Sievers et al., 2011) present in Jalview package (Waterhouse et al., 2009)that also provided Conservation analysis. The different box colors represent TdT domains, BRCT domain annotated by CDD (Lu et al., 2020) and TdT

catalytic core as described(Delarue et al., 2002) Clique ou toque aqui para inserir o texto. : Purple: BRCA1 C Terminus (BRCT) domain (CL0459), Yellow: Helix-hairpin-helix domain (HHH_8), Blue: Fingers domain of DNA polymerase lambda, Red: DNA polymerase beta palm, highlighting the 3 catalytic Aspartic Acids (red circle) and loop1, Green: DNA polymerase beta thumb.

NCBI Reference Sequences: Horse_tdt_isoform_X1 (XP_005602408.1); Horse_tdt_isoform_X2 (XP_001501812.3); Pig_tdt_isoform_X2 (XP_003133204.1); Pig_tdt_isoform_X1 (XP_005671421.1); Cattle_tdt_isoform_1 (NP_803461.1); Human_tdt_isoform_1 (NP_004079.3); Human_tdt_isoform_2 (NP_001017520.1); Mouse_tdt_isoform_2 (NP_001036693.1); Mouse_tdt_isoform_1 (NP_033371.2)

Supplementary Table 1: AID (RGYW/WRCY) motifs and targeted mutation frequencies in CDR and FR regions

	IgM	IgG
Number of RGYW/WRCY motifs per IGHV segment	19 (1.18)	20 (1.17)
% of CDR nucleotides mutated	16.79	18.81
% of CDR mutations present in RGYW/WRCY motifs	51.61	45.85
% of FR nucleotides mutated	4.24	4.71
% of FR mutations present in RGYW/WRCY motifs	16.92	18.76
% of all nucleotides mutated	6.37	7.12
% of all mutations present in RGYW/WRCY motifs	24.20	25.59

Supplementary Table 2: Percentage of nucleotides present at the N1 and N2 junctions of horse IgM and IgG antibodies

	%A	%T	%G	%C	
IgM	24.82	20.21	**35.59**	19.37	**N1.REGION**
IgG	24.28	21.47	**34.90**	19.35	
IgM	21.29	**28.20**	**30.32**	20.19	**N2.REGION**
IgG	21.57	**29.60**	**28.87**	19.96	

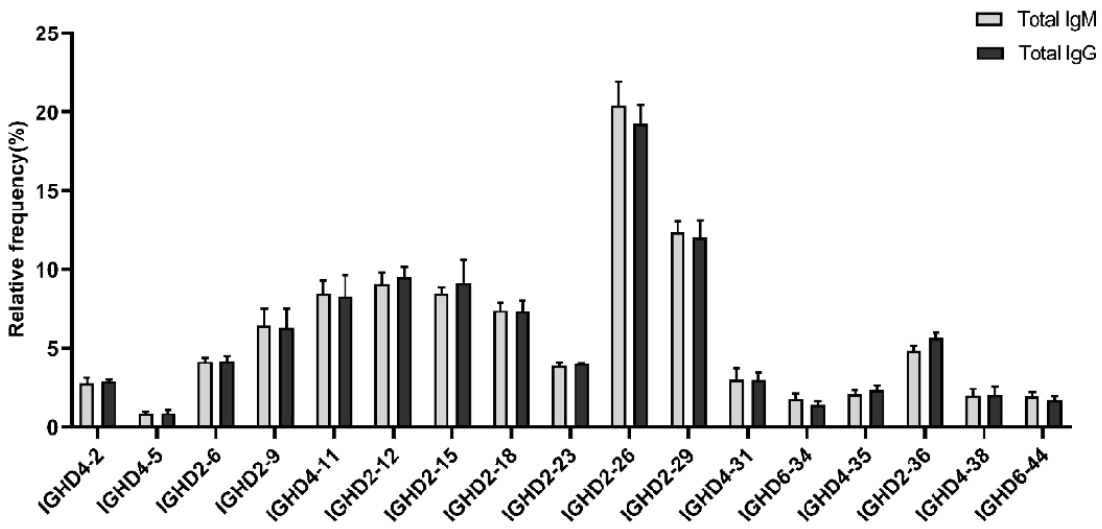
Supplementary Table 3: Amino Acid Composition per reading frame of IGHD functional gene segments in horse antibodies.

IMGT Group	IMGT gene	RF1	RF2	RF3	iRF1	iRF2	iRF3
IGHD2	IGHD2-6	GYYSRSCYT	MVTIVGVAI	WLL**ELLY	YSNSYYSNH	GIATPTIVT	V*QLLL**P
	IGHD2-9	GYYASGYDY	MVTMLVVMTI	WLLC*WL*L	CSHNH*HSNH	V VITISIVI	*S*PLA**P
	IGHD2-12	YSYGSYYA	TIVMVVTM	L*LW*LLC	HSNYHNYS	GIVTTITI	A*LP*L*
	IGHD2-15	CYYGSYYSSYYA	MVTMVVTTVVTM	WLLW*LLQ*LLC	HSNYCSNYHSNH	GIVTTVVTTIVT	A*LL**LP**P
	IGHD2-18	GYYAGSYYA	MVTMLVVIM	WLLCW*LLC	HSNYQHSNH	GIVTTSIVT	A*LPA**P
	IGHD2-23	DYYGISDSY	MITMVLVTPI	*LLWY**LL	CRSH*YHSNH	VGVTNTIVI	*ESLIP**S
	IGHD2-26	YGYGGAYYY	TMVMVVLITI	LWLWWCLLL	CSSKHHHNHS	VVVSTIIII	***APP*P*
	IGHD2-29	SYYGGSSWYS	TVTMVVVPGI	QLLWW*FLVL	STRNYHHSNC	GVPGTITIVI	EYQELPP**L
	IGHD2-36	DYYGAIDYI	MIIMVLLIT*	*LLWCY*LHN	LCSQ*HHNNH	YVVNSTIII	VM*SIAP**S
IGHD4	IGHD4-2	YYGWGN	TMAGV	LLWLG*	YPSHS	VTPAIV	LPQP**
	IGHD4-5	YYNYNY	TTTIT	LLQL*L	SYSCS	VVIVVV	*L*L*
	IGHD4-11	NYGYGYA	TTVMVML	*LRLWLCY	VA*P*P*L	*ENENRS	SITITVV
	IGHD4-31	YDDGYYN	TMTDTT	LR*RILQ	CSIRHR	VVVSVIV	L*YPSS*
	IGHD4-35	NYGSYNYY	TMAPIIT	*LWLL*LL	SNYRSHS	VVIIGAIV	**L*EP*L
	IGHD4-38	EKSWSN	RRVGV	GEELE*	YSNSS	VTPTLL	LLQLFS
IGHD6	IGHD6-34	YGSGW	TVAVG	LR*RLA	ANRYR	PTATV	GQPLP*
	IGHD6-44	YGSGW	TVVVG	IR*WLA	ANHYR	PTTTV	GQPLPY

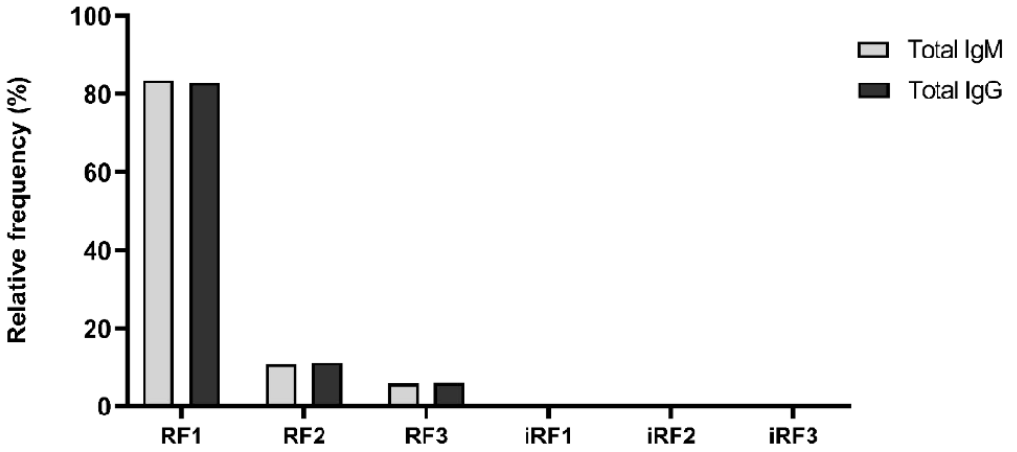
* means Stop codon

Supplementary Table 4: Top 10 bigger CDRH3 found in this study

Horse	ID	IGHV gene and allele	IGHJ gene and allele	CDR3 amino acids	CDR3 length (aa)	N1 number (nt)	N2 nt number
4	1	IGHV4-21*01 ORF	IGHJ4*01 F	TGGKRGGSFQLKEGGGGDGGSYSSGTGSPQRDRYFGYWGQDTPV KAVAQSVETEYYYTY	59	40	104
4	2	IGHV4-21*01 ORF	IGHJ4*01 F	AGGEVCEEGCEEKCYDSGYSITVQEEKCVRRSVNDSEYYSRSCCC RYYFAY	51	103	8
4	3	IGHV4S1*01	IGHJ6-1*01 F	AGADYGGTMHGIKFWGQGILVTVSSGESHSPSLYCCTGADYGGTY HGIKF	50	111	12
4	4	IGHV4-21*01 ORF	IGHJ4*01 F	AGVWGDWKGLVYAIDEWGPGILSTVSSGESHDDRGGLLYSIDY	43	11	91
1	5	IGHV4-21*01 ORF	IGHJ2*01 ORF	AGGNMVGYCAMMRCGIEYCVQGILGTVSSWESRSTEN	37	31	51
2	6	IGHV4-29*01	IGHJ4*01 F	GASLTVVGELPPGPLLETGVADDYDDTFAFTESEVY	36	62	12
2	7	IGHV4-21*01 ORF	IGHJ6-1*01 F	SGGEGRVKDSTIYADEAIMEGRVKDSTVSVDEAILY	36	16	66
4	8	IGHV4-37D*01	IGHJ6-1*01 F	ATALAQVVLPDWPRWYCLKNVLLGYKLLVYWGINS	35	13	62
4	9	IGHV4-37D*01	IGHJ6-1*01 F	ATALAQVVLPDWPRWYCLKNVLLGYKFLVYWGINS	35	13	62
4	10	IGHV4S1*01	IGHJ6-1*01 F	KGLVARDAGGSESLRRRRELSLRIMPSVYYSVNY	34	12	56



Supplementary Figure 1: Median of Relative frequency of functional IGHD gene segments in horses in IgM and IgG repertoires.



Supplementary Figure 2- Reading Frame (RF) preference for IGHD gene segment in non-immunized horses. Relative frequency of the six possible open reading frames for the IGHD gene segment. RF1, RF2, and RF3 are generated by deletion, while iRF1, iRF2, and iRF3 are generated by inversion.