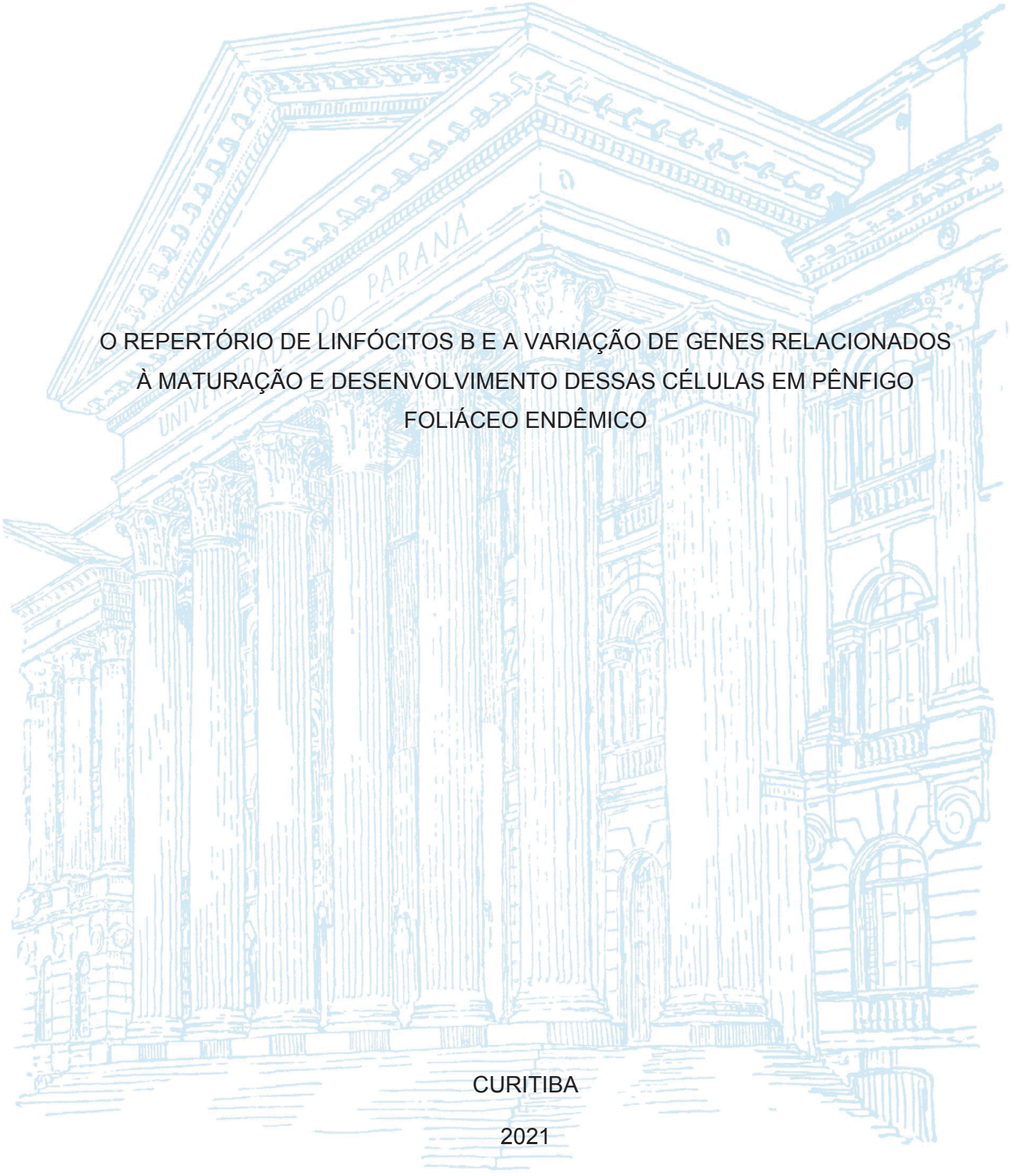
UNIVERSIDADE FEDERAL DO PARANÁ

VERÓNICA CALONGA SOLÍS



O REPERTÓRIO DE LINFÓCITOS B E A VARIAÇÃO DE GENES RELACIONADOS À MATURAÇÃO E DESENVOLVIMENTO DESSAS CÉLULAS EM PÊNFIGO FOLIÁCEO ENDÊMICO

CURITIBA

2021

VERÓNICA CALONGA SOLÍS

O REPERTÓRIO DE LINFÓCITOS B E A VARIAÇÃO DE GENES RELACIONADOS À MATURAÇÃO E DESENVOLVIMENTO DESSAS CÉLULAS EM PÊNFIGO FOLIÁCEO ENDÊMICO

Tese de doutorado apresentada ao Programa de Pós-Graduação em Genética, Setor de Ciências Biológicas, Universidade Federal do Paraná, como requisito parcial para obtenção do título de doutora em genética.

Orientador: Dr. Danillo Gardenal Augusto
Coorientadora: Profª Dra. Danielle Malheiros Ferreira.

CURITIBA

2021

Universidade Federal do Paraná. Sistema de Bibliotecas.
Biblioteca de Ciências Biológicas.
(Rosilei Vilas Boas – CRB/9-939).

Solís, Verónica Calonga.
O repertório de linfócitos B e a variação de genes relacionados à maturação e desenvolvimento dessas células em pênfigo foliáceo endêmico. / Verónica Calonga Solís. – Curitiba, 2021.
177 f. : il.

Orientador: Danillo Gardenal Augusto.
Coorientadora: Danielle Malheiros Ferreira.

Tese (Doutorado) – Universidade Federal do Paraná, Setor de Ciências Biológicas. Programa de Pós-Graduação em Genética.

1. Pênfigo. 2. Genética. 3. Imunoglobulinas. 4. Doenças. 5. Autoimunidade. I. Título. II. Ferreira, Danielle Malheiros. III. Augusto, Danillo Gardenal. IV. Universidade Federal do Paraná. Setor de Ciências Biológicas. Programa de Pós-Graduação em Genética.

CDD (20. ed.) 616.5

MINISTÉRIO DA EDUCAÇÃO
SETOR DE CIÊNCIAS BIOLÓGICAS
UNIVERSIDADE FEDERAL DO PARANÁ
PRÓ-REITORIA DE PESQUISA E PÓS-GRADUAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO GENÉTICA - 40001016006P1

TERMO DE APROVAÇÃO

Os membros da Banca Examinadora designada pelo Colegiado do Programa de Pós-Graduação em GENÉTICA da Universidade Federal do Paraná foram convocados para realizar a arguição da tese de Doutorado de **VERÓNICA CALONGA SOLÍS** intitulada: **O REPERTÓRIO DE LINFÓCITOS B E A VARIAÇÃO DE GENES RELACIONADOS À MATURAÇÃO E DESENVOLVIMENTO DESSAS CÉLULAS EM PÊNFIGO FOLIÁCEO ENDÊMICO**, sob orientação do Prof. Dr. DANILLO GARDENAL AUGUSTO, que após terem inquirido a aluna e realizada a avaliação do trabalho, são de parecer pela sua APROVAÇÃO no rito de defesa.

A outorga do título de doutor está sujeita à homologação pelo colegiado, ao atendimento de todas as indicações e correções solicitadas pela banca e ao pleno atendimento das demandas regimentais do Programa de Pós-Graduação.

CURITIBA, 25 de Junho de 2021.

Assinatura Eletrônica
05/07/2021 14:39:59.0
DANILLO GARDENAL AUGUSTO
Presidente da Banca Examinadora

Assinatura Eletrônica
28/06/2021 16:23:23.0
LIANA ALVES DE OLIVEIRA
Avaliador Externo (FACULDADES INTEGRADAS DO BRASIL)

Assinatura Eletrônica
05/08/2021 13:29:48.0
ANDRE MACEDO VALE
Avaliador Externo (UNIVERSIDADE FEDERAL DO RIO DE JANEIRO)

Assinatura Eletrônica
28/06/2021 21:58:28.0
KARIN BRAUN PRADO
Avaliador Interno (UNIVERSIDADE FEDERAL DO PARANÁ)

Assinatura Eletrônica
28/06/2021 14:16:47.0
EDUARDO LANI VOLPE DA SILVEIRA
Avaliador Externo (UNIVERSIDADE DE SÃO PAULO)

Aos meus pais,
María Magdalena e Ignacio.

AGRADECIMENTOS

Agradeço ao meu orientador Danillo Augusto, que ao longo destes seis anos de trabalho juntos me guiou, me apoiou e me incentivou a procurar sempre a excelência. Obrigada pela grande paciência. A minha coorientadora Danielle Malheiros, por estar sempre que precisei, pelo carinho e pela amizade. A professora Maria Luiza Petzl-Erler, que me recebeu no laboratório LGMH quando eu era ainda uma estrangeira desconhecida que ligava com insistência para saber se teriam vaga no laboratório. Tenho muita honra, e muita sorte, de ter sido orientada por vocês três.

Aos meus supervisores de intercâmbio na Alemanha, Hauke Busch e Anke Fähnrich, obrigada por me acolherem profissional e pessoalmente, e por me apoiarem em todo este período.

Aos meus pais, por todo o amor e carinho, e por sempre apoiarem todos os meus projetos. As minhas irmãs Valeria e Victoria, que junto com meus pais, são meu grande suporte na vida, obrigada por todos estarem presentes em cada detalhe da minha vida. A minha sobrinha Agustina, que todos os dias nos enche de amor e alegria. Aos demais membros da minha família, obrigada por sempre torcerem por mim, e por compreender que nem sempre posso estar perto fisicamente, mas que sempre estou acompanhando tudo com o coração.

Aos meus amigos do laboratório. Obrigada pelas inúmeras ajudas, não só no laboratório, mas sempre que foi preciso. Obrigada pelos *happy hours*, pelas viagens de congressos, pelos almoços no RU, pelos passeios nos botânicos. Graças a vocês até os momentos mais complicados foram mais fáceis de superar.

Aos meus amigos da Argentina, que mesmo na distância estão sempre presentes e me fazem levar a vida com mais leveza. Aos meus amigos do Paraguai, que comemoram cada conquista e amortecem os momentos mais difíceis.

Aos colegas do laboratório na Alemanha, que me receberam com amizade e carinho, e com quem o tempo de pandemia foi um pouco mais leve.

À Universidade Federal do Paraná e a Universidade de Lübeck (Alemanha). Aos institutos de fomento CNPq, CAPES e Fundação Araucária pelo financiamento dos projetos pesquisas. A CAPES pela bolsa de doutorado, e ao DAAD e a CAPES por co-financiarem o meu período de intercâmbio na Alemanha.

“A educação é a arma mais poderosa que
você pode usar para mudar o mundo”
Nelson Mandela

RESUMO

O pênfigo foliáceo (PF) é uma doença autoimune mediada por células B e autoanticorpos contra a desmogleina 1, uma molécula de adesão celular dos queratinócitos. Apesar de rara no mundo, apresenta alta incidência e prevalência em algumas regiões do Brasil, nas quais essa doença é endêmica. Por ser uma doença de etiologia complexa, diversos fatores genéticos e ambientais estão envolvidos na patogênese de PF. O primeiro objetivo desse estudo foi avaliar o impacto da variação de genes que codificam moléculas envolvidas no desenvolvimento das células B e na produção de anticorpos na suscetibilidade ao PF. Analisamos 3.336 polimorfismos de nucleotídeos únicos (SNPs) em 167 genes candidatos em um estudo do tipo caso-controle em uma amostra de 227 pacientes de PF e 193 controles. Encontramos 10 variantes intrônicas ou intergênicas associadas à uma maior suscetibilidade ao PF (OR > 1,56; p < 0,005): *rs6657275*G* (*TGFB2*); *rs1818545*A* (*RAG1/RAG2/IFTAP*); *rs10781530*A* (*PAXX*), *rs10870140*G* e *rs10781522*A* (*TRAF2*); *rs535068*A* (*TNFRSF1B*); *rs324011*A* (*STAT6*); *rs6432018*C* (*YWHAQ*); *rs17149161*C* (*YWHAG*); e *rs2070729*C* (*IRF1*). A partir de análises *in silico* constatamos que estes SNPs foram anteriormente associados à expressão diferencial, indicando que podem ter um impacto no funcionamento normal das vias moleculares nas quais essas moléculas participam. O segundo objetivo deste trabalho foi caracterizar o repertório de imunoglobulinas em pacientes de PF e controles. Estudamos três subgrupos de pacientes (n = 16), controles da área endêmica (n = 6) e controles de uma área não endêmica (n = 4). Após o isolamento de RNA total a partir de células mononucleadas do sangue periférico, realizamos o sequenciamento da região variável (exon VDJ) da cadeia pesada de imunoglobulinas IgG e IgM. Encontramos que tanto pacientes quanto indivíduos saudáveis que vivem em áreas endêmicas apresentam uma diversidade reduzida do repertório de células B em comparação com indivíduos saudáveis de uma região não endêmica, sugerindo uma possível alteração do repertório por características ambientais específica da região endêmica. Identificamos que os segmentos gênicos IGHV3-23 e IGHV3-30 têm expressão reduzida e aumentada, respectivamente, em pacientes com doença ativa comparados à indivíduos sem a doença (p < 0,04). Observamos que os pacientes apresentaram sequências da região CDR3 que desviaram da distribuição normal, com maior frequência de sequências mais longas. Ainda, realizamos uma cuidadosa análise de similaridades de sequências CDR3 e identificamos dois agrupamentos específicos de PF, o que pode ser informativo quanto aos anticorpos relevantes para a doença e podem ser possíveis alvos terapêuticos no futuro. Pela primeira vez, mostramos que os polimorfismos nos genes envolvidos no desenvolvimento de células B e na produção de anticorpos conferem suscetibilidade diferencial ao PF endêmico. Ainda, fomos os primeiros a demonstrar diferenças no repertório de células B em pacientes e controles e também diferenças potencialmente causadas por fatores ambientais da área endêmica de PF. Por ser uma doença negligenciada e pouco compreendia, a caracterização do repertório de células B em PF pode contribuir na elucidação da patogênese dessa doença, assim como também ajudar na identificação dos fatores ambientais que desencadeiam a doença em indivíduos geneticamente susceptíveis. Nossos resultados representam um grande avanço e podem servir de base para estudos futuros que venham a focar no desenvolvimento de terapias mais eficazes e com menos efeitos adversos do que as convencionais corticoterapias.

Palavras-chaves: Pênfigo foliáceo endêmico. Associação genética. Repertório de imunoglobulinas.

ABSTRACT

Pemphigus foliaceus (PF) is an autoimmune disease mediated by B cells and autoantibodies against desmoglein 1, a cell adhesion molecule of keratinocytes. Although rare in the world, PF exhibits high incidence and prevalence in some endemic regions of Brazil. Because it is a complex disease, several genetic and environmental factors are involved in the pathogenesis of PF. The first aim of this study was to evaluate the impact of variation in genes encoding molecules involved in B-cell development and antibody production on susceptibility to PF. We analyzed 3,336 single nucleotide polymorphisms (SNPs) in 167 candidate genes in a case-control study in a sample of 227 FP patients and 193 controls. We found 10 intronic or intergenic variants associated with increased susceptibility to PF (OR > 1.56; $p < 0.005$): *rs6657275*G* (*TGFB2*); *rs1818545*A* (*RAG1/RAG2/IFTAP*); *rs10781530*A* (*PAXX*), *rs10870140*G* and *rs10781522*A* (*TRAF2*); *rs535068*A* (*TNFRSF1B*); *rs324011*A* (*STAT6*); *rs6432018*C* (*YWHAQ*); *rs17149161*C* (*YWHAG*); and *rs2070729*C* (*IRF1*). According to our *in-silico* analyses, these SNPs were previously associated with differential expression of the genes, indicating that they may have an impact on the functioning of their molecular pathways. The second aim of this work was to characterize the immunoglobulin repertoire in PF patients and controls. We studied three subgroups of patients (n = 16), controls from the endemic area (n = 6) and controls from a non-endemic area (n = 4). After isolation of total RNA from peripheral blood mononucleated cells, we performed sequencing of the variable region (exon VDJ) of the IgG and IgM immunoglobulin heavy chain. We found that both patients and healthy individuals living in endemic areas show reduced B-cell repertoire diversity compared with healthy individuals from the non-endemic region, suggesting a possible alteration of the repertoire caused by environmental characteristics of the endemic region. We identified that the IGHV3-23 and IGHV3-30 gene segments have reduced and increased expression, respectively, in patients with active disease compared to individuals without the disease ($p < 0.04$). We observed that patients the CDR3 region sequence distribution deviated from the normal distribution, with higher frequencies of longer sequences. In addition, we performed a careful similarity analysis of CDR3 sequences and identified two specific clusters in PF patients, which may be informative about antibodies relevant to the disease and may be possible therapeutic targets in the future. For the first time, we showed that polymorphisms in genes involved in B-cell development and antibody production confer differential susceptibility to endemic PF. Furthermore, we were the first to demonstrate differences in the B-cell repertoire in patients and controls and also differences potentially caused by environmental factors in the PF endemic area. As a neglected and poorly understood disease, the characterization of the B-cell repertoire in FP may contribute to elucidate the pathogenesis of this disease, as well as help in the identification of environmental factors that trigger the disease in genetically susceptible individuals. Our results represent an advance in the study of this diease and may serve as a basis for future studies focusing on the development of more effective therapies with fewer adverse effects than the conventional corticotherapies.

Keywords: Endemic pemphigus foliaceus. Genetic association. Immunoglobulin repertoire.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SIGLAS E ABREVIATURAS

AID	Citidina desaminase induzida por ativação
APC	Células apresentadoras de antígenos
BCR	Receptores de células T
BER	Reparo do DNA por excisão de base
BMP	Proteína morfogênica óssea
CAD	Doença por aglutininas a frio
cDNA	DNA complementario
CDR	Regiões determinantes da complementariedade
CMV	Citomegalovirus
CSR	Mudança de classe por recombinação
DSG	Desmogleina
EGF	Fator de crecimento epidérmico
Gm	Variações alotípicas encontradas na cadeia constante de IgG
HLA	Genes dos Antígenos Leucocitários Humanos
Ig	Molécula de imunoglobulina
IG	Gene de imunoglobulina
IGH	Gene da cadeia pesada das imunoglobulinas
IGHV	Segmento gênico variável da cadeia pesada das imunoglobulinas
IGHD	Segmento gênico de diversidade da cadeia pesada das imunoglobulinas
IGHJ	Segmento gênico de junção da cadeia pesada das imunoglobulinas
IGHC	Segmento gênico constante da cadeia pesada das imunoglobulinas
IGK	Gene da cadeia leve kappa das imunoglobulinas
IGL	Gene da cadeia leve lambda das imunoglobulinas
IL	Interleucina
Km	Variações alotípicas encontradas na cadeia constante de IgK
MAC	Complexo de ataque à membrana
MMR	Reparo de mal pareamento do DNA
mRNA	RNA mensageiros
MZ	Zona marginal
NHEJ	Reparação de ADN não-homólogo
nt	Nucleotídeos

pb	Pares de bases
PBMC	Células mononucleadas do sangue periférico
PF	Pênfigo foliáceo
PV	Pênfigo vulgar
RAG	Genes ativadores de recombinação
RSS	Sequencias sinais de recombinação
SHM	Hipermutação somática
SLE	Lúpus eritematoso sistêmico
TCR	Receptores de células B
Treg	Linfócitos T reguladores
VHS	Vírus da herpes simples
VVZ	Vírus da varicela-zoster

SUMÁRIO

1 INTRODUÇÃO

As imunoglobulinas são moléculas produzidas pelos linfócitos B, que quando ancoradas na superfície destas células, servem como receptores de células B (BCR, do inglês *B Cell Receptor*) que medeiam o desenvolvimento das respostas humorais através do reconhecimento de antígenos. Após o reconhecimento de antígenos, os linfócitos B sofrem expansão clonal e diferenciação para plasmócitos, que secretarão imunoglobulinas contra antígenos específicos. A essas imunoglobulinas secretadas pelos plasmócitos dá-se o nome de anticorpos.

Os genes codificadores de imunoglobulinas possuem uma organização diferente dos demais genes do genoma, necessitando de rearranjos somáticos em cada linfócito B para a produção de moléculas funcionais. Estes processos incluem a recombinação V(D)J, a mudança de classe por recombinação e a hipermutação somática. Em cada etapa, um conjunto de moléculas e vias de sinalização perfeitamente orquestradas geram o imenso repertório de imunoglobulinas formado por cada indivíduo. A enorme diversidade tem o potencial de reconhecer os antígenos de todos os possíveis patógenos. No entanto, uma parcela desses anticorpos pode eventualmente reconhecer antígenos próprios e levar ao desenvolvimento de doenças autoimunes.

Apesar da grande implicação em doenças, pouco se conhece sobre os mecanismos que levam à quebra da autotolerância. Graças ao advento do sequenciamento de nova geração, hoje é possível a análise do repertório de imunoglobulinas no contexto de doenças, uma área promissora dentro da imunogenética.

Dentro desse contexto, o pênfigo foliáceo (PF) representa uma das doenças autoimunes mediadas por células B, na qual a produção de autoanticorpos contra proteínas de adesão celular resulta no aparecimento de bolhas na pele. O PF ocorre de forma esporádica e com baixa incidência no mundo. Porém, apresenta incidência e prevalência altas em algumas regiões do Brasil, que são consideradas regiões endêmicas da doença. A existência de regiões endêmicas sugere que fatores ambientais podem atuar como desencadeadores da doença; ao mesmo tempo que os estudos de associação genética apontam que fatores genéticos aumentam o risco de desenvolver a doença.

Na primeira etapa deste trabalho, realizamos um estudo caso-controle em indivíduos da área endêmica de PF no Brasil, e encontramos que polimorfismos de nucleotídeos únicos dos genes que codificam as moléculas envolvidas nas diferentes fases de desenvolvimento das células B e produção de anticorpos influenciam o risco de desenvolver a doença.

Em uma segunda fase desse trabalho, utilizamos sequenciamento de nova geração para fazer a primeira caracterização do repertório de imunoglobulinas em pacientes com PF endêmico e em controles de regiões endêmica e não endêmica, contribuindo com novos resultados relevantes para a compreensão da patogênese dessa doença.

2 REVISÃO BIBLIOGRÁFICA

2.1 SISTEMA IMUNE ADAPTATIVO, O PAPEL DOS LINFÓCITOS B E IMUNOGLOBULINAS

O sistema imune dos vertebrados é constituído pelo sistema imune inato e o adaptativo. O sistema imune inato é constituído majoritariamente por células e moléculas que reconhecem padrões moleculares nos patógenos e iniciam a primeira resposta do organismo. Já o sistema imune adaptativo é constituído por células e moléculas capazes de formar respostas eficazes de maneira específica para cada um desses patógenos e, ainda, consegue desenvolver memória celular capaz de gerar uma resposta imune mais rápida caso haja uma exposição posterior ao mesmo patógeno. Os linfócitos B e T, junto com seus respectivos receptores (BCR – receptores de células B; TCR – receptores de células T), estão envolvidos na resposta imune adaptativa, e são capazes de reconhecer e reagir de forma específica aos diferentes antígenos (MURPHY; WEAVER, 2016).

Os BCR são moléculas ligadas à membrana plasmática das células B, constituídas por uma molécula de imunoglobulina (Ig) e pelas moléculas acessórias Igα/Igβ (CD79a/CD79b) (CHU; ARBER, 2001). Quando os linfócitos B reconhecem seus antígenos alvo eles são ativados e secretam aos espaços extracelulares imunoglobulinas que reconhecem os mesmos antígenos, que passam a ser chamadas de anticorpos, e produzem um tipo de resposta denominada resposta imune humoral (MURPHY; WEAVER, 2016). A princípio, o repertório de imunoglobulinas de cada indivíduo é capaz de reconhecer todas as moléculas existentes (NOSSAL, 2003). Os genes de imunoglobulinas serão denominados IG e as moléculas codificada por eles de Ig.

2.1.1 Imunoglobulinas: estrutura gênica e proteica

As Ig são compostas de duas cadeias pesadas e duas cadeias leves, ligadas de forma covalente entre si. A molécula pode ser dividida em dois domínios funcionais: o domínio variável (V) que reconhece e se liga aos antígenos, e o domínio constante

(C) que exerce as funções efetoras (MURPHY; WEAVER, 2016; SCHROEDER; CAVACINI, 2010).

Em humanos, os genes de IG encontram-se nos cromossomos 14 (cadeia pesada – *IGH*, do inglês *immunoglobulin heavy locus*), 2 (cadeia leve kappa – *IGK*, *immunoglobulin kappa locus*) e 22 (cadeia leve lambda – *IGL*, *immunoglobulin lambda locus*) (CROCE et al., 1979; MCBRIDE et al., 1982). Esses genes possuem uma configuração particular, sendo constituídos por repetições em tandem de segmentos gênicos (TABELA 1 e FIGURA 1). O gene que codifica a cadeia pesada possui os segmentos gênicos variável (V_H, *variable*), diversidade (D_H, *diversity*), junção (J_H, *joining*) e constante (C_H, *constant*); e os da cadeia leve possuem somente segmentos gênicos variável (V_L), junção (J_L) e constante (C_L). No domínio variável de ambas cadeias (pesadas e leves) encontram-se três regiões determinantes da complementariedade (CDR, do inglês *complementary determining region*), que reconhecem e se ligam aos antígenos de forma altamente específica (FIGURA 1). As primeiras duas regiões (CDR1 e CDR2) estão codificadas na linhagem germinativa dos segmentos *IGHV*, e a região CDR3 é gerada em cada rearranjo, na junção dos segmentos gênicos V(D)J (MURPHY; WEAVER, 2016; SCHROEDER; CAVACINI, 2010).

TABELA 1: NÚMERO DE SEGMENTOS GÊNICOS NOS GENES DE IMUNOGLOBULINAS

Segmento gênico	**IGH**		**IGK**		**IGL**	
	Total	Funcional	Total	Funcional	Total	Funcional
V	123-129	38-46	76	34-38	73-74	29-33
D	27	23	0	0	0	0
J	9	6	5	5	7-11	4-5
C	11	9	1	1	7-11	4-5

Na contagem total de genes, além dos segmentos gênicos funcionais estão incluídos os pseudo-segmentos gênicos e as sequências de fase de leitura aberta (ORF, do inglês *open reading frame*). Dados atualizados por última vez no banco de dados IMGT no dia 30/01/2020 (LEFRANC et al., 2015; LEFRANC; LEFRANC, 2001).

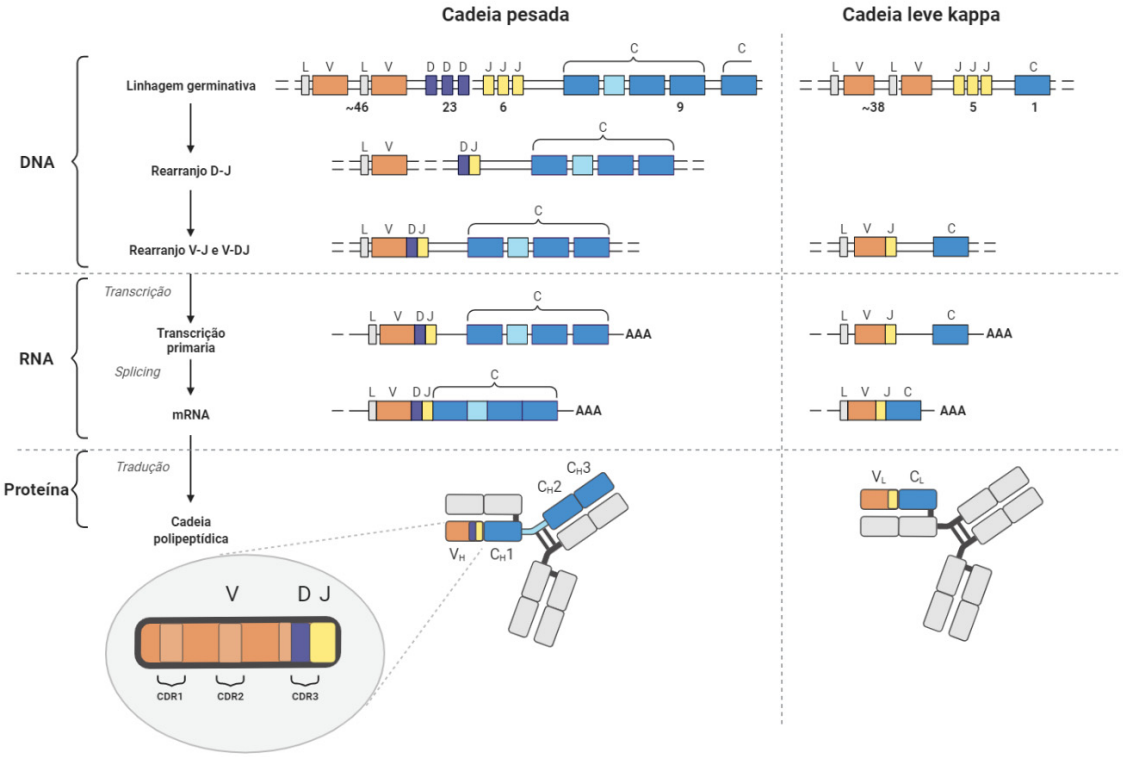


FIGURA 1: ESQUEMATIZAÇÃO DA ESTRUTURA DOS GENES DE *IGH* E *IGK* E DO REARRANJO V(D)J. A linhagem germinativa dos genes está constituída por segmentos gênicos repetidos em tandem, e suas correspondências nos domínios variável ($V_{H\ e\ L}$) e constante ($C_{H\ e\ L}$) na molécula de imunoglobulinas. Os números embaixo dos segmentos gênicos representam a quantidade de segmentos gênicos funcionais de acordo com o banco de dados IMGT (LEFRANC et al., 2015; LEFRANC; LEFRANC, 2001). No rearranjo V(D)J acontece uma seleção de um segmento gênico V, um D e um J, que constituirão o exon V(DJ). Este rearranjo acontece primeiro na cadeia pesada (primeiro D-J e logo V-DJ) e posteriormente na cadeia leve. Em destaque, a localização das regiões determinantes de complementariedade da cadeia pesada: CDR1 e CDR2 estão codificadas no segmento gênico V e CDR3 é formado pela junção de V(D)J. As cadeias pesadas estão unidas entre si e com as cadeias leves por pontes dissulfeto. Imagem modificada de MURPHY; WEAVER (2016) e realizada com BioRender (www.biorender.com).

Os segmentos gênicos constantes da cadeia pesada determinam as distintas classes de Ig (IgM, IgD, IgG, IgA e IgE), e compreendem nove segmentos gênicos funcionais: *IGHM*, *IGHD*, *IGHG1*, *IGHG2*, *IGHG3*, *IGHG4*, *IGHA1*, *IGHA2* e *IGHGE*, e dois pseudo segmentos-gênicos: *IGHGP* e *IGHEP1* (LEFRANC; LEFRANC, 2001)

2.1.2 Rearranjos somáticos

Na linhagem germinativa, os genes de imunoglobulinas possuem uma estrutura complexa, com múltiplas sequências de segmentos gênicos semelhantes.

Essa estrutura gênica é encontrada em todas as células do organismo. Entretanto, para que se tornem funcionais, estes genes devem antes passar por um complexo processo de rearranjos somáticos, para que então possam ser transcritos para RNA mensageiros funcionais (mRNA) e traduzidos para proteínas (Revisado por JUNG et al., 2006). Esse processo é chamado de rearranjo V(D)J e acontece durante o desenvolvimento dos linfócitos B na medula óssea. Posteriormente, os genes sofrem outras modificações que alteraram a afinidade e a função dos anticorpos através dos processos de hipermutação somática (SHM) e de mudança de classe por recombinação (CSR), respectivamente (Revisado por HWANG; ALT; YEAP, 2015).

2.1.2.1 Recombinação V(D)J

No processo de recombinação V(D)J são selecionados apenas um dos segmentos gênicos V, um dos segmentos D, e um J em cada linfócito, formando o éxon V(D)J, que codificará a região variável da molécula (FIGURA 1). Desta forma, um gene funcional de imunoglobulina existe apenas nos linfócitos B, após o processo de recombinação somática, e cada célula possui um gene funcional de sequência diferente (Revisado por JUNG et al., 2006).

A regulação da recombinação V(D)J requer uma sincronização de modificações na estrutura da cromatina, de fatores de transcrição, e de mecanismos de quebra e reparo do DNA(Revisado por SCHATZ; JI, 2011). A seguir será descrito este processo para a cadeia pesada de imunoglobulina (IGH), o qual acontece antes do rearranjo da cadeia leve, e se diferencia dele por incluir a seleção do segmento gênico IGHD.

Em linfócitos em desenvolvimento, os genes *RAG1* e *RAG2* (genes ativadores de recombinação 1 e 2) são expressos, e seus produtos iniciam o processo de reconhecimento e clivagem do DNA. Esta clivagem acontece em sequências específicas do gene de imunoglobulinas chamadas de sequências sinais de recombinação (RSSs, *recombination signal sequences*) que se encontram adjacentes aos segmentos gênicos *IGHV*, *IGHD* e *IGHJ*, e estão constituídas por sequências altamente conservadas de sete nucleotídeos (heptâmero) e de nove nucleotídeos (nonâmero), separadas por uma sequência pouco conservada de 12 ou 23 nucleotídeos (nt) (FIGURA 2). O complexo RAG1/2 somente reconhece e cliva o DNA em RSS com um espaçador de 12 nt e o outro com um espaçador de 23 nt, processo

chamado de regra 12/23 que garante que o processo aconteça de forma precisa, unindo *IGHD* a *IGHJ*, e logo *IGHV* a *IGHD*/*IGHJ* (EASTMAN; LEU; SCHATZ, 1996).

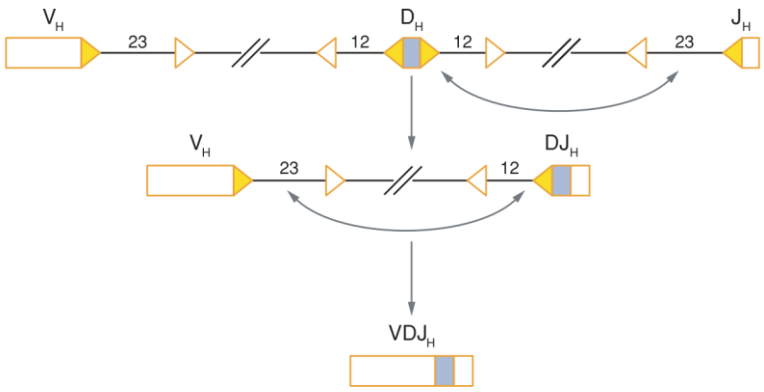


FIGURA 2: REGRA 12/23 NO REARRANJO V(D)J. O complexo RAG1/2 deve reconhecer a sequências sinais de recombinação (RSS) com um espaçador de 23 nt e outro de 12 nt. No gene *IGH*, na primeira etapa é produzida a quebra e união na fita dupla entre *IGHD* e *IGHJ*, e posteriormente uma quebra e união entre *IGHV* e *IGHD*/*IDHJ*. Imagem modificado de JUNG et al. (2006).

A ordem desses eventos é regulada e direcionada por sequências promotoras de transcrição a montante dos segmentos gênicos IGHV e IGHD, às quais unem-se a fatores de transcrição, que promovem uma transcrição “estéril” do segmento gênico, isto é, que não resulta em um mRNA funcional e não será traduzido a proteína, mas permite a abertura da cromatina, o acesso do complexo RAG1/2 e a clivagem do DNA nas sequências RSSs (SCHATZ; JI, 2011).

O complexo RAG1/2 cliva a fita dupla de DNA e reúne cada fita de forma covalente em uma alça em forma de grampo, que é novamente clivada pelas enzimas Artemis e DNA-PKcs (do inglês, *DNA-dependent protein kinase, catalytic subunit*) em uma posição de quatro a cinco nucleotídeos de distância do vértice da alça, tornando uma fita de DNA mais comprida do que a outra. A fita mais curta será preenchida por DNA polimerases, gerando assim pequenos trechos de sequências palindrômicas chamados de nucleotídeos P e, adicionalmente, a enzima TdT (do inglês, *terminal deoxynucleotidyl transferase*) adiciona nucleotídeos não-moldes, chamados de nucleotídeos N. Finalmente, estas pontas são unidas através das vias de reparo de DNA não-homólogo (NHEJ, do inglês *non-homologous end joining*), envolvendo outras proteínas como Ku70, Ku80, XRCC4, DNA Ligase IV e XLF. Desta forma, a junção dos segmentos V, D e J resulta na região CDR3, a de maior diversidade nas

imunoglobulinas, e que se une com maior especificidade ao antígeno (FIGURA 3) (Revisado em MALU et al., 2012).

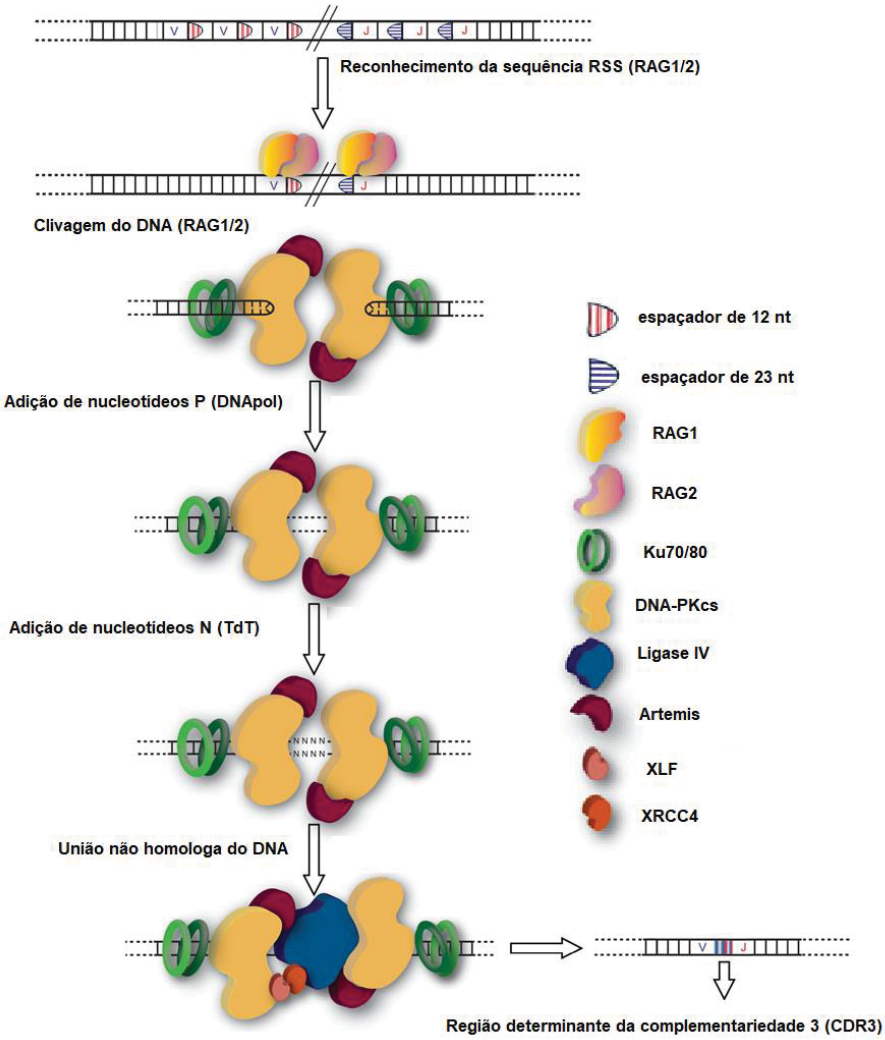


FIGURA 3: PASSOS E MOLÉCULAS ENVOLVIDAS NO REARRANJO V(D)J. Primeiramente o complexo RAG1/RAG2 reconhece e cliva as sequências sinais de recombinação (RSS). Posteriormente a DNA polimerase adiciona sequências palindrômicas (nucleotídeos P) nas junções entre os segmentos gênicos, e a enzima TdT adiciona nucleotídeos aleatórios (nucleotídeos N). Finalmente, o DNA é unido através das vias de reparo do DNA não-homólogo (NHEJ) pelas enzimas Ku70, Ku80, XRCC4, DNA Ligase IV e XLF. Esta junção resulta na região CDR3, de maior diversidade nas imunoglobulinas. Imagem adaptada de MALU et al. (2012).

Este rearranjo posiciona o novo éxon VDJ adjacente ao segmento constante IGHM e, assim, o promotor encontrado na sequência germinativa de *IGHV* é posicionado muito próximo de um forte elemento intensificador da transcrição

(chamado iEμ), que se encontra no íntron entre o *IGHJ* e *IGHM*. Desta forma, inicia-se a transcrição do gene funcional, começando no promotor do segmento VH selecionado no rearranjo e terminando nos segmentos constantes *IGHM* e *IGHD*, resultando em um mRNA com a região variável rearranjada e as regiões constantes IGHM e IGHD (YANCOPOULOS; ALT, 1985). Posteriormente, o mRNA passa por splicing alternativo que permite a tradução de IgM ou IgD (GEISBERGER; LAMERS; ACHATZ, 2006).

Se um rearranjo é bem sucedido, a recombinação no gene de imunoglobulina no cromossomo homólogo é bloqueada por um processo chamado de exclusão alélica. Desta forma, cada clone de célula B possui um único gene rearranjado de imunoglobulina (OUTTERS et al., 2015).

Quando os linfócitos B terminam o processo de rearranjo V(D)J, eles saem da medula óssea através da corrente sanguínea e migram para a baço. Neste momento são chamados de linfócitos B transicionais tipo 1 (T1), e em seguida de linfócitos B T2. Os linfócitos B T2 que migram para a zona marginal do baço (MZ, do inglês *marginal zone*) e constituirão as células B MZ, e aquelas que recirculam para outros órgãos linfoides secundários são chamados de linfócitos B maduros foliculares.

2.1.2.2 Hipermutação somática

Os linfócitos B podem encontrar seus antígenos específicos nos linfonodos secundários, onde podem ser ativados e coestimulados pelos linfócitos T, pelas células dendríticas foliculares e pelas citocinas liberadas por estas células. Este processo resulta na formação dos centros germinativos nos linfonodos, na qual os linfócitos B iniciam uma expansão celular clonal (MAK; SAUNDERS; JETT, 2014). Em cada divisão celular e replicação do DNA, os éxons V(D)J previamente rearranjados sofrem substituições nucleotídicas pontuais, em um processo chamado de hipermutação somática, que acontece numa taxa de uma em mil pares de bases (pb) por divisão celular, frequência expressivamente maior do que a taxa de mutação normal (Revisado por HWANG; ALT; YEAP, 2015). Este processo aumenta a diversidade de imunoglobulinas e permite uma seleção positiva de células B com maior afinidade de ligação ao antígeno (RAJEWSKY; FÖRSTER; CUMANO, 1987).

Este processo é conduzido pela enzima citidina desaminase induzida por ativação (AID, do inglês *activation-induced cytidine deaminase*), uma molécula expressa unicamente em linfócitos B (MURAMATSU et al., 2000), que promove a desaminação dos resíduos de citosina (C) convertendo-os em uracila (U), a qual posteriormente será substituída por outro nucleotídeo (Revisado por HWANG; ALT; YEAP, 2015). A desaminação acontece em DNA de fita simples, com preferência em citosinas dentro do motivo DGYW (D = A/G/T, Y = C/T, W = A/T), ou WRCH na fita complementar (W = A/T, R = A/G, H = T/C/A). Estas sequências se encontram enriquecidas no éxon V(D)J das Ig, e são consideradas "pontos quentes" de SHM (ROGOZIN; DIAZ, 2004).

A ação de AID precisa da abertura de dupla fita de DNA, que acontece por meio de mecanismos baseados na transcrição gênica, e que permitem que a AID tenha acesso às sequências alvo no DNA na sua conformação de fita simples (Revisado por HWANG; ALT; YEAP, 2015).

As mutações acontecem por falhas no reparo da desaminação, o qual pode acontecer por três mecanismos (FIGURA 4). No primeiro, uma DNA polimerase pode simplesmente parear U com uma adenina (A) (transição da mutação: C:G > U:G > A:G > A:T; FIGURA 4a); ou pode envolver mecanismos normalmente usados para reparo de danos no DNA, porém, com a participação de enzimas de baixa fidelidade, deixando a correção mais propensa à erros e modificando a sequência de DNA. Um desses mecanismos é o reparo por excisão de base (BER, do inglês *basic excision repair*), no qual a uracila pode ser removida pela enzima uracila-DNA-glicosilase (UNG, do inglês *Uracil-DNA-Glycosylase*) deixando a posição sem base, e uma polimerase de baixa fidelidade insere nucleotídeos de forma aleatória (transição da mutação: C:G > U:G > 0:G > N:N; sendo 0 um sitio abásico e N qualquer nucleotídeo; FIGURA 4b). O outro mecanismo é o reparo de mal pareamento do DNA (MMR, do inglês *mismatch repair*), no qual o par de base U:G é reconhecido pelo heterodímero MSH2-MSH6, que recruta o complexo MutLα e a exoclease-1 (Exo1) que em conjunto clivam e removem a U e os nucleotídeos adjacentes da fita de DNA. Neste momento, polimerases de baixa fidelidade, como a Pol-η, adicionam nucleotídeos de forma aleatória nas posições abásicas resultantes (transição da mutação: C:G/T:A > U:G/T:A > 0:G/0:A > N:N/N:N; sendo 0 um sitio abásico e N qualquer nucleotídeo; FIGURA 4c) (Revisado por HWANG; ALT; YEAP, 2015; MURPHY; WEAVER, 2016).

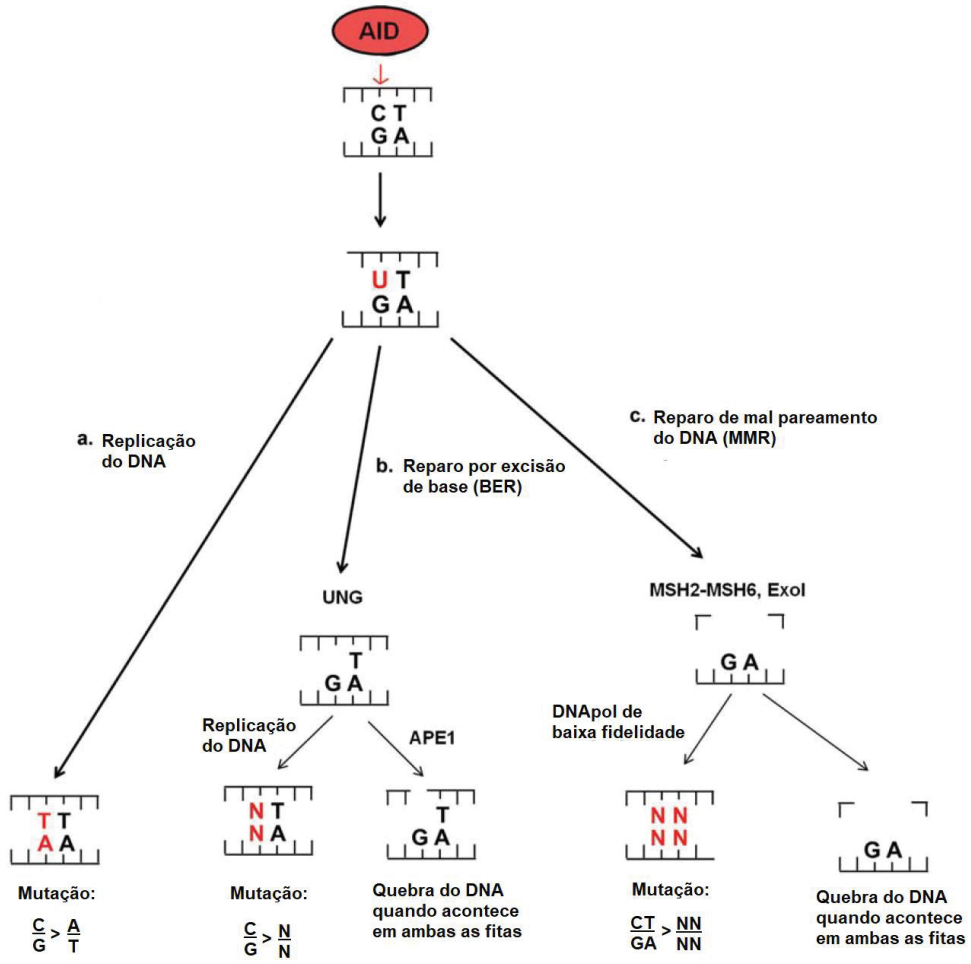


FIGURA 4: MECANISMOS DE REPARO E CLIVAGEM DO DNA NOS PROCESSOS DE HIPERMUTAÇÃO SOMÁTICA (SHM) E MUDANÇA DE CLASSE POR RECOMBINAÇÃO (CSR). Após a desaminação pela enzima AID dos resíduos de citosina (C) para uracila (U) na região variável durante a hipermutação somática (SHM) e na região constante dos genes *IGH* durante a mudança de classe por recombinação (CSR), respectivamente. (a) Replicação do DNA. Na SHM a DNA polimerase pareia a uracila (U) do sitio desaminado com uma adenina (A). (b) Reparo por excisão de base (BER): Na SHM a enzima UNG remove a U deixando o sítio abásico, e na replicação do DNA um nucleotídeo aleatório (N) é adicionado. Na CSR enzimas UNGs deixam o sitio abásico em ambas fitas do DNA e acontece a quebra da dupla fita pela endonuclease APE1. (c) Reparo de mal pareamento do DNA (MMR): Na SHM o heterodímero MSH2-MSH6 e a Exo1 removem a U e o sítio adjacente, deixando ambas as posições abásicas, à qual são adicionados nucleotídeos aleatórios (N) por uma polimerase de baixa fidelidade. Na CSR a remoção de nucleotídeos acontece em ambas as fitas do DNA, levando à quebra da dupla fita. Imagem modificada de HWANG; ALT; YEAP (2015).

Este processo leva a uma maior diversificação do exon V(D)J, principalmente na região de ligação aos antígenos, CDR1, CDR2 e CDR3, e permite uma seleção positiva dos linfócitos com maior afinidade pelos seus antígenos, levando assim ao desenvolvimento de anticorpos com maior afinidade (RAJEWSKY; FÖRSTER; CUMANO, 1987). Além da seleção positiva dos linfócitos com poucas mudanças nas regiões de moldura das imunoglobulinas (que não incluem as CDR), existem outros

fatores ainda não conhecidos que concentram as mutações nas regiões CDR (Revisado por HWANG; ALT; YEAP, 2015).

2.1.2.3 Mudança de classe por recombinação (CSR)

Uma vez rearranjado o éxon V(D)J, ele se encontra adjacente aos segmentos gênicos *IGHM* e *IGHD*, e assim, os linfócitos B expressam IgM e/ou IgD, por meio do processo de splicing alternativo do RNA mensageiro (FIGURA 5). Imunoglobulinas de classe IgM são importantes na primeira fase da resposta imune, porém, para atingir respostas mais especializadas é necessário que os linfócitos B expressem outras classes de Ig com funções diferentes. Para isto, ainda nos centros germinativos, o gene da cadeia pesada das imunoglobulinas sofrem um processo chamado de mudança de classe por recombinação (CSR, do inglês *class switch recombination*), no qual o isotipo expresso muda para IgA, IgG ou IgE (DUDLEY et al., 2005; STAVNEZER; GUIKEMA; SCHRADER, 2008).

Assim como na SHM, na CSR acontece uma transcrição estéril dos segmentos gênicos constantes, que muda a conformação da cromatina e permite o acesso ao DNA das demais moléculas necessárias para a CSR (Revisado por STAVNEZER; GUIKEMA; SCHRADER, 2008). A jusante dos promotores e a montante de cada segmento gênico constante, existem sequências altamente repetitivas chamadas de regiões S (switch) (FIGURA 5) também constituídas pelo motivo DGYW/WRCH, em especial pela sequência palindrômica AGCT. As regiões S são reconhecidas por moléculas adaptadoras chamadas de 14-3-3, as quais recrutam e estabilizam a AID e outras moléculas, como a proteína quinase PKA-Cα, a UNG e a endonuclease APE1 (HWANG; ALT; YEAP, 2015; XU et al., 2012).

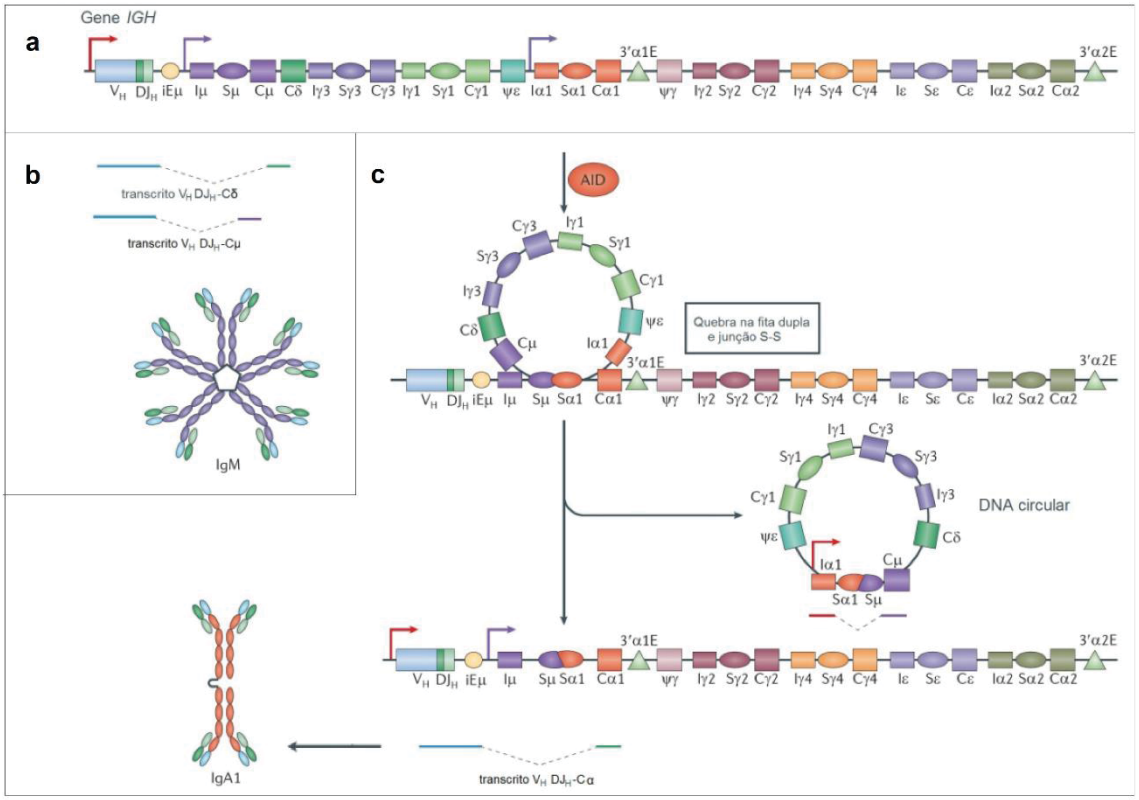


FIGURA 5: TROCA DO ISOTIPO IGM PARA IGA NA MUDANÇA DE CLASSE POR RECOMBINAÇÃO. (a) Disposição do gene *IGH* após a recombinação VDJ. O éxon VDJ rearranjado se encontra adjacente aos segmentos gênicos da região constante *IGHM* (Cμ) e *IGHD* (Cδ), seguida dos demais segmentos gênicos *IGHC* (Cα,γ,ε). (b) Por meio de splicing alternativo, o segmento gênico *IGHM* é mantido no mRNA e IgM é expresso pelas células B. (c) A enzima AID promove a quebra da dupla fita do DNA nas sequências Sμ e Sα a montante dos segmentos gênicos *IGHM* e *IGHA*, respectivamente. Posteriormente o DNA é unido nas regiões Sμ-Sα, formando um DNA circular com a porção clivada do gene, e aproximando o exon VDJ ao segmento gênico *IGHA* que será transcrito pelas células B, que passarão a expressar IgA. Imagem modificada de XU et al. (2012).

Na CSR, a AID também produz desaminação de citosinas para uracilas, contudo, os mecanismos de reparo BER e MMR finalizaram na quebra da dupla fita, e sua posterior junção através da união de extremidade não-homóloga (NHEJ, do inglês *non-homologous end joining*). Este processo aproxima o éxon VDJ rearranjado ao segmento constante selecionado. A CSR resulta na expressão e secreção de anticorpos de isotipos diferentes e com diferentes funções efetoras. O segmento gênico utilizado dependerá das citocinas envolvidas na ativação dos linfócitos, por exemplo, interleucina-4 (IL-4) irá ativar o promotor de IGHG1 e IGHE, e TGF-β o promotor de IGHG2 e IGHA (Revisado em HWANG; ALT; YEAP, 2015).

2.1.3 Níveis de geração de diversidade no repertório de imunoglobulinas

A estrutura dos genes de imunoglobulina, assim como os rearranjos somáticos pelos quais eles passam, são importantes para a geração da grande diversidade que o repertório de anticorpos precisa para reconhecer virtualmente qualquer antígeno.

O primeiro nível dessa diversidade se encontra no grande conjunto de segmentos gênicos que os genes possuem (TABELA 1), e nos próprios polimorfismos desses segmentos. O segundo nível é a diversidade combinatória fornecido pela junção aleatória dos segmentos V, (D), J, seguida da diversidade juncional (terceiro nível) pela adição dos nucleotídeos não moldes N e P. A seguir, a combinação das duas cadeias rearranjadas (pesadas e leves), formando o domínio variável da proteína leva ao quarto nível de geração de diversidade das Ig, já que a recombinação que ocorre na cadeia pesada é independente da que ocorre na cadeia leve. Dessa forma, cada indivíduo consegue gerar uma diversidade enorme de moléculas de imunoglobulinas, que teoricamente pode alcançar valores de 10^{16} (MURPHY; WEAVER, 2016; SCHROEDER; CAVACINI, 2010).

O processo de hipermutação somática, que introduz mutações pontuais na região variável, constitui o quinto nível de aumento da variabilidade das imunoglobulinas e eleva ainda mais a diversidade das imunoglobulinas (MURPHY; WEAVER, 2016; SCHROEDER; CAVACINI, 2010), que é capaz de reconhecer um universo infinito de antígenos (NOSSAL, 2003). O conjunto de imunoglobulinas rearranjadas e expressas de um indivíduo, capazes de reconhecer potencialmente qualquer antígeno, é chamado de repertório de Ig.

2.1.4 Desenvolvimento dos linfócitos B - etapa independente do antígeno

Os linfócitos B são inicialmente produzidos no saco vitelino, posteriormente, durante a vida fetal, no fígado e finalmente na medula óssea (FIGURA 6). Nesses órgãos os linfócitos B originam-se, amadurecem e adquirem sua especificidade antigênica em um processo ainda independente do encontro com os antígenos (MURPHY; WEAVER, 2016).

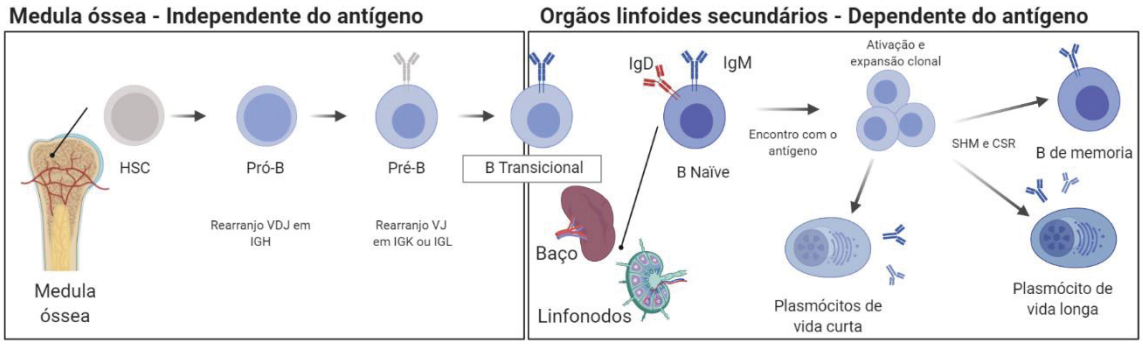


FIGURA 6: REPRESENTAÇÃO SIMPLIFICADA DO DESENVOLVIMENTO DOS LINFÓCITOS B. Em adultos, o desenvolvimento inicia na medula óssea, a partir das células-tronco hematopoiéticas (HSC). Em uma etapa independente do antígeno, nos linfócitos pró-B, o éxon VDJ do gene IGH é rearranjado. Na seguinte etapa, os linfócitos pré-B expressam o receptor de pré-B, constituído pela cadeia pesada rearranjada e uma cadeia leve substituta. Nesta etapa um gene da cadeia leve é rearranjado. Posteriormente, os linfócitos B migram para os órgãos linfoides periféricos no estágio de células B transicional, onde podem ser ativados pelos seus antígenos alvo, e coestimulados por linfócitos T auxiliares. Neste momento as células podem se diferenciar imediatamente em plasmócitos de vida curta, que promovem uma resposta rápida, ou podem migrar aos centros germinativos, onde seus genes de IG podem sofrer mudança de classe (CSR) e hipermutação somática (SHM), e os linfócitos B podem diferenciar-se em plasmócitos de vida longa e em células B de memória. Imagem criada com BioRender (www.biorender.com).

Os linfócitos B originam-se a partir das células-tronco hematopoiéticas (HSC, do inglês *hematopoietic stem cells*) que se diferenciam para dar origem aos progenitores linfoides comuns quando estimuladas pela sinalização da proteína morfogênica óssea (BMP), e pela supressão da estimulação do fator de crescimento transformador beta (TGF-β), que pelo contrário, estimularia o desenvolvimento da linhagem mieloide (NAKA; HIRAO, 2017). As primeiras etapas do desenvolvimento são promovidas pelas células estromais da medula óssea que produzem moléculas sinalizadoras como o ligante de FLT3, a interleucina-7 (IL-7), o fator de células-tronco (SCF) e a quimiocina CXCL12. A diferenciação até o estágio pró-B é regulada pela expressão de diferentes fatores de transcrição, como PU.1, E2B, EBF, IRF8 e Ikaros, que levam a produção de proteínas importantes para o comprometimento com a linhagem de células B, como RAG1 e RAG2, essenciais para o rearranjo V(D)J das imunoglobulinas. Além destes, passa a expressar o fator de transcrição Pax5, responsável pela repressão de um grande conjunto de genes que impedem o comprometimento para outras linhagem celulares, e pela expressão de genes específicos de linfócitos B, como Igα (CD79a), CD19 e CD21 (BONILLA; OETTGEN, 2010; FUXA; SKOK, 2007; LEBIEN; TEDDER, 2008).

No estágio pró-B acontece o rearranjo V-D-J (primeiro D-J, e logo V-DJ) na cadeia pesada das imunoglobulinas. Uma vez finalizado o rearranjo, as células entram

no estágio pré-B e passam a expressar o receptor de células pré-B, constituído pela cadeia pesada rearranjada e uma cadeia leve substituta, em conjunto com o heterodímero Igα/Igβ necessário para a transdução de sinais intracelulares. Se o rearranjo V-D-J não for satisfatório, acontece o rearranjo a partir do outro alelo de IGH. Se o rearranjo em ambos cromossomos falharem, processo que ocorre em cerca da metade das células pré-B, a célula sofre morte por apoptose (MURPHY; WEAVER, 2016). A expressão do receptor de pré-B induz o rearranjo V-J na cadeia leve de imunoglobulinas. Neste momento, as células passam a expressar IgM na sua superfície e são chamadas de células B imaturas. A expressão do BCR (IgM junto com Igα/Igβ), na superfície da célula B envia um sinal intracelular para a finalização dos rearranjos nos genes das imunoglobulinas (GRAWUNDER et al., 1995).

Os linfócitos B migram para os órgãos linfoides periféricos, momento em que passam ao estágio de células B transicional tipo 1 (T1), e ao chegar no baço passam ao estágio T2 após receberem os sinais gerados pelo fator de ativação das células B (BAFF, do inglês *B lymphocyte activating factor belonging to the TNF family*). As células B T2 podem recircular pelo sistema linfático iniciando o estágio de células B maduras foliculares, ou podem migrar para a zona marginal do baço, e se transformar em células B MZ (do inglês, *marginal zone*). Em ambos os casos, os linfócitos passam a expressar IgD, além de IgM, através do splicing alternativo do mRNA da cadeia pesada das imunoglobulinas (XU et al., 2012).

Assim, as células B maduras, também chamadas de células B naïve, circulam pelos sistemas sanguíneo e linfático ou são mantidas nos órgãos linfoides periféricos, onde suas respostas imunes são iniciadas após encontrarem seus antígenos alvo, momento em que começa a fase de diferenciação. Todos os dias, a medula óssea humana adulta gera cerca de 10^9 células B maduras, entretanto, cada linfócito tem uma chance em 10^5 de ser estimulado por seu antígeno específico (MAK; SAUNDERS; JETT, 2014).

2.1.5 Ativação dos linfócitos B - etapa dependente do antígeno

Após o encontro com o antígeno, a ativação das células B pode ser independente ou dependente de linfócitos T, dependendo do tipo de antígeno reconhecido. Posteriormente, as células podem se diferenciar em plasmócitos produtores de anticorpos ou em células B de memória (MURPHY; WEAVER, 2016).

2.1.5.1 Resposta independente de células T

A ativação de linfócitos B independente da estimulação das células T auxiliares (Th, do inglês *T helper*) confere uma rápida resposta contra os patógenos. Neste tipo de resposta estão envolvidos dois tipos de antígenos, os Ti-1 e os Ti-2. Geralmente os linfócitos B MZ estão envolvidos nesta resposta (MAK; SAUNDERS; JETT, 2014; MURPHY; WEAVER, 2016).

Os antígenos Ti-1 atuam como mitógenos (estimuladores da mitose), ligando-se de forma não específica aos receptores presentes nas membranas dos linfócitos B (não somente imunoglobulinas), e enviando fortes sinais intracelulares de proliferação celular. Como a união aos receptores não depende da especificidade ao sítio de união ao antígeno, muitos clones de linfócitos B podem ser ativados, ou seja, há uma ativação policlonal (MAK; SAUNDERS; JETT, 2014; MURPHY; WEAVER, 2016).

Por outro lado, os Ti-2 são grandes antígenos que possuem elementos estruturais repetitivos, na maioria das vezes são proteínas e polissacarídeos, mas também podem ser lipídeos e ácidos nucleicos. Estes antígenos são geralmente provenientes de bactérias ou vírus e, pelo seu grande tamanho, podem ligar-se a diversos BCRs na superfície celular, enviando um forte sinal intracelular de ativação, proliferação e diferenciação. Esta união ao antígeno é chamada de união interligada (ou *cross-linked*) porque as moléculas de Ig se encontram ligadas umas com as outras por meio das suas uniões aos antígenos (VOS et al., 2000).

Como estas respostas não dependem da interação com linfócitos T, os linfócitos B não recebem sinais coestimulatórios que induzem à mudança de classe e a hipermutação somática, e consequentemente, não atingem um nível alto de afinidade ao antígeno e geralmente não produzem células de memória. Entretanto, os linfócitos ativados por estes antígenos secretam anticorpos IgM, que adotarão uma estrutura pentamérica e se ligarão de forma multivalente aos antígenos, compensando a falta de maturação da afinidade dos anticorpos. Além disso, confere uma rápida resposta contra os patógenos que confere uma janela de tempo para recrutar células T para uma resposta imune posterior mais efetiva (VOS et al., 2000).

2.1.5.2 Resposta dependente de células T

A ativação dos linfócitos B dependente da interação com linfócitos T requer que linfócitos Th CD4+ reconheçam o mesmo antígeno, mas não necessariamente o mesmo epítopo desse antígeno. Estes antígenos são chamados de Td (dependentes

de células T). Primeiramente, os linfócitos B reconhecem um antígeno proteico, por exemplo, de vírus ou bactérias, por meio dos seus BCRs. Os BCRs enviam um sinal intracelular que ativa fatores de transcrição necessários para a expressão de moléculas HLA de classe II e de receptores de citocinas necessários para receber a ajuda dos linfócitos Th, e para a posterior proliferação celular (MAK; SAUNDERS; JETT, 2014; PARKER, 1993).

Os antígenos são interiorizados por meio de endocitose, e posteriormente fragmentados em cadeias polipeptídicas menores que serão apresentadas nas fendas das moléculas de HLA de classe II aos linfócitos Th CD4+, cujos TCR reconhecem o mesmo antígeno (processo chamado de reconhecimento ligado). O correceptor CD4 das células T se liga a uma região distal das moléculas HLA de classe II, permitindo uma união mais estreita entre os linfócitos T e B. Os linfócitos T são ativados quando reconhecem seus peptídeos cognatos nas moléculas de HLA de classe II em células apresentadoras de antígenos (APC, do inglês, *antigen presenting cells*), o que aumenta a expressão da molécula CD40L na superfície celular. Já os linfócitos B são ativados após a união dos receptores CD40 em sua superfície aos ligantes CD40L dos linfócitos T. A partir desde momento, outros receptores de citocinas, como por exemplo, receptores de IL-1, IL-4, IL-5, IL-10, são expressos na superfície dos linfócitos B. Estas citocinas são produzidas e secretadas pelas células Th ativadas (MAK; SAUNDERS; JETT, 2014; PARKER, 1993).

2.1.5.3 Diferenciação das células B

Imediatamente após o encontro com o antígeno, tanto Ti como Td, uma porcentagem de linfócitos B ativados não sofrerão hipermutação somática nem mudança de classe. Desta forma, estas células se diferenciam imediatamente em plasmócitos de vida curta, que promoverão uma rápida resposta contra os patógenos através de anticorpos IgM de baixa afinidade (LEBIEN; TEDDER, 2008).

Quando os linfócitos B são coestimulados pelas células Th nos centros germinativos dos órgãos linfoides secundários, eles passam por uma expansão clonal, e seus genes de imunoglobulina sofrem a mudança de classe por recombinação que gera anticorpos com uma função diferente, e hipermutação somática, que permite a seleção positiva de células com especificidade aumentada contra esses antígenos (HOLLING; SCHOOTEN; VAN DEN ELSEN, 2004).

Posteriormente, os linfócitos B se diferenciam em plasmócitos de vida longa ou em células B de memória. Os plasmablastos de vida curta podem migrar dos centros germinativos para a medula óssea ou também para os linfonodos, onde passam a ser plasmablastos de vida longa e a produzir altas quantidades de anticorpos de alta afinidade (até 40% das suas proteínas podem ser imunoglobulinas), que são secretadas à corrente sanguínea ou outros tecidos. Os linfócitos B de memória são células pequenas, residentes nos órgãos linfoides onde foram originalmente ativadas, e onde provavelmente serão necessitadas novamente, ou também podem circular pelos gânglios linfáticos, mantendo uma vigilância periférica contra o antígeno (MAK; SAUNDERS; JETT, 2014).

2.1.6 Checagem da autorreatividade

O sistema imune é adaptado para reconhecer e reagir contra patógenos e fatores externos, sem causar danos aos próprios tecidos. Durante o desenvolvimento das células B existem diversos mecanismos para prevenir que receptores autorreativos proliferem e se mantenham no sistema. Esta ausência ou baixa resposta ao próprio é chamada de autotolerância. A checagem da tolerância acontece em duas instâncias, a tolerância central na medula óssea nos primeiros estágios de desenvolvimento das células B, e a tolerância periférica nos órgãos linfáticos secundários que impedem a ativação dos linfócitos autorreativos (MURPHY; WEAVER, 2016).

2.1.6.1 Tolerância central

Na medula óssea, as células estromais apresentam autoantígenos às células B imaturas. Se os BCR dessas células reconhecerem com alta afinidade os autoantígenos, as células B autorreativas sofrem uma edição do receptor, na qual os linfócitos podem ser recuperados através de um rearranjo secundário da região variável dos genes das imunoglobulinas, por meio da reativação da expressão do complexo RAG. Nas cadeias leves acontecerá um novo rearranjo enquanto existam segmentos V_L e J_L a serem utilizados, entretanto, nas cadeias pesadas, como após o primeiro rearranjo já não existem outros segmentos gênicos D_H disponíveis, o rearranjo acontece somente com a mudança do segmento gênico V_H (HALVERSON; TORRES; PELANDA, 2004). Se este mecanismo falhar, e os linfócitos B continuarem

autorreativos, eles poderão ser selecionados negativamente e eliminados por apoptose, em um processo chamado de deleção clonal (PELANDA; TORRES, 2012).

Por outro lado, quando as células B reconhecem fracamente um autoantígeno entram em um estado permanente de irresponsividade, chamado de anergia, e que pode eventualmente resultar em morte celular. Entretanto, elas ainda podem migrar para os órgãos linfoides secundários, mas não poderão ser estimuladas ao encontrarem seus antígenos alvo (NEMAZEE, 2017). Finalmente, as células B com potencial para reconhecerem autoantígenos, mas que não o encontrarem na medula óssea, ou o encontrarem com uma concentração tão baixa que não permite sua ativação, permanecem em estado de ignorância e, portanto, escapam aos pontos de checagem centrais e passam aos órgãos linfáticos secundários (APLIN et al., 2003).

2.1.6.2 Tolerância periférica

Assim como na medula óssea, os linfócitos B podem se tornar anergizados se reconhecerem fortemente autoantígenos nos órgãos linfáticos secundários sem receberem sinais coestimulatórios de células Th, o que posteriormente pode resultar em deleção clonal por apoptose. É possível que linfócitos B autorreativos encontrem linfócitos T que reconhecem o mesmo autoantígeno, mas que foram previamente anergizados e, nesse caso, as células B também se tornam irresponsivas devido à falta de coestimulação, resultando posteriormente em morte celular. Contudo, algumas células B autorreativas podem escapar ao controle dos mecanismos de seleção negativa, e formar parte do repertório de linfócitos maduros. De fato, um nível baixo destas células é necessário para o funcionamento normal do sistema imune, já que de outra forma o repertório poderia ser muito limitado (MAK; SAUNDERS; JETT, 2014).

A tolerância periférica dos linfócitos B também pode ser mantida pela interação com os linfócitos T. Foi observado que pacientes com atividade defeituosa de linfócitos T reguladores (Treg) também apresentam um acumulo de linfócitos B autorreativos (KINNUNEN et al., 2013) e que pacientes com expressão defeituosa de CD40L e moléculas de HLA de classe II também desenvolveram células B auto reativas (HERVÉ et al., 2007).

2.2 DIFICULDADES NO ESTUDO DA DIVERSIDADE DOS GENES DE IMUNOGLOBULINAS

Por muito tempo e até recentemente, os estudos da diversidade populacional de imunoglobulinas eram limitados aos métodos sorológicos que identificavam alótipos de Ig. Os alótipos de Ig são variações nas cadeias polipeptídicas da região constante dos anticorpos, localizadas tanto nas cadeias pesadas dos anticorpos de isotipo IgG e IgA como na cadeia leve kappa. Estas variações são reflexos da diversidade alélica contida nos segmentos gênicos da cadeia pesada *IGHG1*, *IGHG2*, *IGHG3*, *IGHA2* (chamados de alótipos G1m, G2m, G3m e A2m, respectivamente) e da cadeia leve *IGKC* (alótipos Km) (JEFFERIS; LEFRANC, 2009; LEFRANC; LEFRANC, 2012; PANDEY; LI, 2013).

A limitação da análise desses segmentos gênicos acontece devido à grande complexidade metodológica imposta pela estrutura do gene que está constituído por repetições de segmentos gênicos com alta similaridade (TABELA 1), e pela presença de duplicações e deleções de regiões dentro do gene. Isto dificulta o alinhamento de sequencias curtas (de entre 50 e 300 pb) obtidas através das técnicas de sequenciamento de segunda geração (WATSON; BREDEN, 2012), a mais utilizadas na atualidade. O sequenciamento da totalidade dos genes tem sido realizada por técnicas de clonagem por meio BAC (cromossomo artificial bacteriano), YAC (cromossomo artificial de levedura), entre outros (MATSUDA et al., 1998; WATSON et al., 2013), porém o alto custo e as dificuldades metodológicas dessa abordagem inviabilizam sua aplicação em escala populacional.

Outra dificuldade no estudo da diversidade destes genes deve-se ao fato de que muitas das amostras de DNA utilizadas pelos principais bancos de dados genômicos foram construídos a partir de linhagens de células B. Essas células sofrem recombinação somática nos genes de Ig e, portanto, não possuem o gene completo, o que impede a genotipagem de todos os segmentos gênicos da linhagem germinativa nos indivíduos estudados causando uma subestimativa de frequências alélicas de diversas variantes (WATSON et al., 2013).

Apesar destas dificuldades, diversos pesquisadores fizeram esforços para investigar os efeitos da diversidade desses genes na susceptibilidade de desenvolver doenças infecciosas, autoimunes, câncer, ou na concentração de anticorpos nos tecidos biológicos, etc. Estes trabalhos foram feitos por meio das técnicas sorológicas

que identificam os alótipos de Ig e, mais recentemente por técnicas de biologia molecular que identificam os SNPs que servem de marcadores para esses alótipos (TABELA 2). Entretanto, estas associações não podem ser corroboradas ou refutadas em estudos de associação de genoma completo já que variantes desses genes se encontram pouco representadas nos microarranjos de genotipagem (PANDEY; LI, 2013), devido a limitações do desenho de sondas dos microarranjos de genotipagem para essas regiões que apresentam alta identidade de sequência e variação estrutural. Um exemplo direto dessa limitação será demonstrado nos resultados da primeira parte dessa tese, no qual encontramos uma quantidade muito baixa de variantes em genes *IGH*, *IGK* e *IGL* utilizando um microarranjo de mais de 500 mil SNPs genômicos.

2.3 ESTUDO DAS CARACTERÍSTICAS DO REPERTÓRIO DE IMUNOGLOBULINAS

Até o momento focamos na configuração dos genes de imunoglobulina na sua linhagem germinativa, e os rearranjos que esses genes sofrem nos linfócitos B. Cada sequência de mRNA transcrito a partir dos genes rearranjados em cada clone de célula B é chamada de clonotipo e possui uma combinação única de segmentos gênicos IGHV, IGHD, IGHJ e IGHC associados à uma sequência específica de CDR3. Cada célula B produz apenas um clonotipo (HERSHBERG; LUNING PRAK, 2015). Ao conjunto de clonotipos produzidos pela totalidade de células B, damos o nome de repertório de imunoglobulinas ou repertório de células B. Esse repertório é responsável por codificar o conjunto de anticorpos expressos em cada indivíduo.

Apesar das dificuldades técnicas em estudar os genes de imunoglobulinas na linhagem germinativa, é possível analisar o repertório de imunoglobulinas a partir dos seus transcritos (mRNA) através de técnicas de sequenciamento de nova geração. A análise do repertório permite compreender a dinâmica do desenvolvimento de clones de células B, as diferenças entre indivíduos saudáveis e pacientes, e permite compreender o mecanismo do desenvolvimento de imunidade após imunizações e da patogênese de doenças, alergias, entre outros (BASHFORD-ROGERS; SMITH; THOMAS, 2018a).

TABELA 2: ASSOCIAÇÕES DA VARIABILIDADE GENÉTICA DAS IMUNOGLOBULINAS COM DIVERSAS CARACTERÍSTICAS

Doença	**Alótipo associado**	**Método de detecção**	**Associação**	**Referências**
Esclerose múltipla	Gm1,17;21	Sorologia	RR = 3,6; *p* = 0,02	(PANDEY et al., 1981)
Hepatite crônica autoimune	Gm1,2 / HLA-B8	Sorologia	RR > 39	(WHITTINGHAM et al., 1981)
Imunidade após vacinação Meningococo C	Km1	Sorologia	*P* = 0,005	(PANDEY et al., 1979).
Imunidade após vacinação contra *Haemophilus influenzae* tipo b	Km1-/1-	Sorologia	*P* = 0,008	(PANDEY et al., 1979).
Alta concentração de IgG no líquido cefalorraquidiano em pacientes com esclerose múltipla	Gm21;1,2,17;(..)	Sorologia	$p < 1x10^{-5}$	(BUCK et al., 2013).
Maior concentração média de IgG4	Gm23	Sorologia	*p* < 0,02	(STEINBERG; MORELL, 1973)
Ligação aumentada a proteínas do citomegalovírus humano (HCMV):	Gm3	Sorologia	p = 0.0005	(NAMBOODIRI; PANDEY, 2011)
Meningite por *Haemophilus* e epiglotite	Gm1,3,17;5,13,21;23 sem *HLA-B*51* ou *HLA-B*52*	Sorologia	OR < 0,1; p < 0,004	(GRANOFF et al., 1984)
Meningite por *Haemophilus* e epiglotite	Gm1,3,17;5,13,21 com HLA-DR3	Sorologia	OR = 11; p = 0,02	(GRANOFF et al., 1984)
Esquizofrenia	Gm3;(..)	Molecular	OR = 3,4, p = 0,0002	(PANDEY; NAMBOODIRI; ELSTON, 2016)
Câncer de mama	Gm3	Molecular	OR = 2.07; p = 0.0147	(PANDEY et al., 2012)

Gm: Variações alotípicas encontradas na cadeia constante de IgG.
Km: Variações alotípicas encontradas na cadeia constante de IgK.
HLA: Genes dos Antígenos Leucocitários Humanos.

Para analisar o repertório de imunoglobulinas, realiza-se o preparo de bibliotecas de sequenciamento, que pode ser feito a partir de mRNA total da amostra. O mRNA pode ser analisado em populações de células (em inglês, *bulk sequencing*) e mais recentemente através do sequenciamento de células individuais (em inglês, *single-cell sequencing*). Ainda, é possível analisar células a partir de diferentes populações celulares, como a a população total de células mononucleadas do sangue periférico (PBMC, do inglês *peripheral blood mononuclear cell*) ou diretamente de subpopulações de linfócitos B (WARDEMANN; BUSSE, 2017).

Na abordagem mais frequentemente utilizada, o cDNA (DNA complementar) é transcrito a partir do mRNA total e posteriormente os genes de imunoglobulinas são enriquecidos a partir de amplificação utilizando oligonucleotideos iniciadores (*primers*) específicos que reconhecem a extremidade 5' do éxon V(D)J rearranjado (TURCHANINOVA et al., 2016; VÁZQUEZ BERNAT et al., 2019). Essa abordagem, portanto, foca no sequenciamento unicamente da sequência codificadora, após remoção dos íntrons.

As bibliotecas mais abrangentes permitem a identificação dos segmentos gênicos V, D, J e C utilizados no rearranjo de imunoglobulinas (cadeias pesadas e leves), entretanto muitos estudos focam unicamente no sequenciamento das regiões CDR3, por se tratar da região de maior importância no reconhecimento do antígeno (GLANVILLE et al., 2015). Em nosso trabalho, fizemos a análise do repertório de maneira abrangente, sequenciando as cadeias pesadas completas no contexto da doença autoimune pênfigo foliáceo.

2.3.1.1 Uso diferencial dos segmentos gênicos

A utilização dos segmentos gênicos V, D e J no rearranjo dos genes de IG não acontece de forma aleatória. Isto foi evidenciado por experimentos mostrando que variações nas sequências sinais de recombinação (RSSs) e nas suas regiões espaçadoras afetam tanto a ligação e clivagem pelo complexo RAG1/2, e também a posterior união e reparo do DNA (MONTALBANO et al., 2003). Além disso, certos segmentos gênicos são mais usados do que outros no processo normal de rearranjo (KITAURA et al., 2017), assim como em algumas em doenças (BASHFORD-ROGERS et al., 2019b; BASHFORD-ROGERS; SMITH; THOMAS, 2018b).

Os segmentos IGHV codificam a maior parte da região variável das Ig e determinam os motivos estruturais das regiões CDR1 e CDR2 (FIGURA 1), que se ligam aos antígenos (SCHROEDER; CAVACINI, 2010); portanto, é de se esperar que o segmento gênico utilizado no rearranjo tenha um impacto funcional nos anticorpos. Por exemplo, na doença por aglutininas a frio (CAD, do inglês *cold agglutinin disease*) foi observado que os segmentos gênicos *IGHV4-34* e *IGKV3-20* codificam motivos estruturais que reconhecem e se ligam ao antígeno I dos glóbulos vermelhos, e poderiam determinar o fenótipo clínico da doença (MAŁECKA et al., 2016).

2.3.1.2 Características da região CDR3

As regiões determinantes de complementariedade 3 (CDR3) dos anticorpos é a região mais variável nas moléculas de Ig, e a mais importante para o reconhecimento e ligação aos antígenos. O comprimento de CDR3 apresenta uma distribuição Gaussiana no repertório de indivíduos saudáveis em condições fisiológicas normais, isto é, que não se encontram diante de uma resposta imunológica intensa como as causadas por vacinação, infecção, entre outros. Por isso, um desvio da distribuição de comprimento de regiões CDR3 foi sugerida como relacionada a alterações imunológicas causadas por patologias ou por respostas direcionadas do sistema imunitário (MIQUEU et al., 2007).

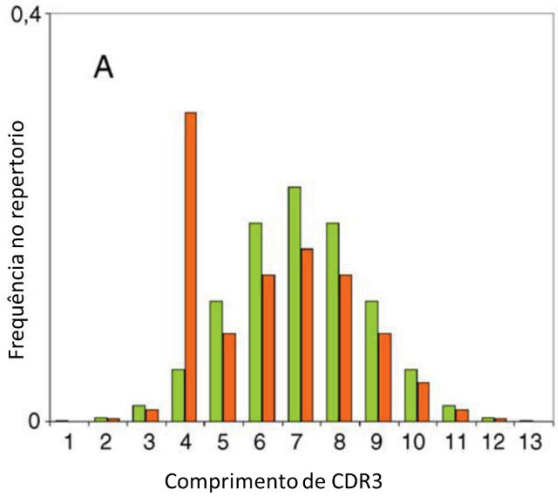


FIGURA 7: DISTRIBUIÇÃO DO COMPRIMENTO DE SEQUÊNCIAS CDR3. No eixo X são representados os diferentes comprimentos da região CDR3 em aminoácidos e no eixo Y a frequência de clonotipos contendo esse tamanho de CDR3. O modelo teórico Gaussiano (em verde) é contrastado com um repertório alterado (em laranja) que apresenta expansão de clonotipos com comprimento de CDR3 de 4 aminoácidos. Imagem adaptada de MIQUEU et al., 2007.

Um dos motivos que causam esse desvio é a hiperexpansão de determinados linfócitos após uma resposta imune, aumentando a frequência de clonotipos especificamente relacionados a essa resposta (MIQUEU et al., 2007). Por outro lado, genes de imunoglobulina rearranjados propensos à causar autorreatividade sofrem um rearranjo secundário que substitui o segmento IGHV e deixa um vestígio de aminoácidos na região CDR3, resultando em sequências CDR3 mais longas (LANGE et al., 2014). Foi identificado que as frequências de substituição de IGHV são significativamente elevadas nos clonotipos encontrados em diferentes doenças autoimunes, incluindo artrite reumatoide e lúpus eritematoso sistémico, e rinite alérgica, e nos clonotipos que codificam anticorpos antivirais (LANGE et al., 2014).

2.3.1.3 Avaliação da expansão clonal no repertório de imunoglobulinas

Outra característica de interesse é a frequência relativa dos clonotipos no repertório (i.e., a quantidade de cópias de cada clonotipo em relação ao repertório total). Esta análise permite avaliar a existência de expansão clonal de células B específicas. Por exemplo, o repertório de células B naïve de sangue periférico em indivíduos saudáveis tende a ter uma distribuição uniforme do número de cópias dos clonotipos. Já em neoplasias de linfócitos B, um clonotipo pode ser muitas vezes mais frequente que os demais clonotipos do repertório (Revisado em HERSHBERG; LUNING PRAK, 2015). Em doenças autoimunes, o repertório pode ser composto por uma quantidade menor de clonotipos, mas que são encontrados em alta frequência. Ou seja, clonotipos hiperexpandidos são comuns em doenças autoimunes, como relatado na síndrome de Sjogren (HERSHBERG et al., 2014) ou em lúpus eritematoso sistêmico (SFIKAKIS et al., 2009).

2.4 PÊNFIGO FOLIÁCEO

O pênfigo consiste em um conjunto de doenças autoimunes da epiderme caracterizado clinicamente pelo aparecimento de bolhas na pele e mucosas. Além disso, é caracterizado imunologicamente pela presença de autoanticorpos contra proteínas de adesão celular dos queratinócitos, principalmente desmogleina (DSG), e consequente pela perda de adesão entre células da epiderme (acantólise) que leva ao desenvolvimento de bolhas e erosões na pele (CULTON *et al*., 2008).

Existem duas formas mais comuns de pênfigo: pênfigo foliáceo (PF), que acomete as camadas mais superficiais da pele, e pênfigo vulgar (PV), que acomete as camadas mais profundas da epiderme, afetando mucosa e pele. Ambas as formas de pênfigo ocorrem de forma esporádica no mundo; o PV é a forma com maior incidência, de 1,7, 10, 16,1 e 94,8 casos por milhão por ano segundo reportes em populações da França, Irã, Israel (Judeus) e Alemanha, respectivamente (BASTUJI-GARIN et al., 1995; CHAMS-DAVATCHI et al., 2005; HÜBNER et al., 2016; PISANTI et al., 1974), e é também a forma mais estudada. O PF é uma doença considerada rara no panorama mundial, com incidência média de 1,1 casos por milhão de habitantes por ano na França, Suíça, Estados Unidos e Índia (BASTUJI-GARIN et al., 1995; HANS-FILHO et al., 1996; JOLY; LITROWSKI, 2011; KUMAR, 2008; MARAZZA et al., 2009; SCHMIDT; KASPERKIEWICZ; JOLY, 2019b), mas apresenta alta incidência e prevalência em algumas regiões geográficas que são consideradas áreas endêmicas da doença. Foram descritos focos de pênfigo foliáceo endêmico no Peru, Colômbia, Equador, Paraguai, Venezuela, Tunísia e Brasil (Revisado em AOKI et al., 2015; JAIRO; OCAMPO; LOPERA, 2011).

Existem duas formas clínicas principais de PF: a forma localizada e a forma generalizada (FIGURA 8). Com a forma localizada, aparecem pequenas bolhas nas áreas seborreicas da pele: na face, cabeça, pescoço e parte superior do tronco, e erosões e crostas ocorrem como resultado do rompimento das bolhas que aparecem nessas áreas. Na forma generalizada, as lesões se encontram por todo o tronco, abdômen e membros superiores e inferiores, além da face e couro cabeludo. A expressão máxima da forma generalizada do PF ocorre na fase eritrodérmica, na qual aparecem eritema e lesões esfoliativas em toda a pele dos pacitentes. A fase mais aguda é a fase de invasão bolhosa, que geralmente é acompanhada por febre,

artralgias e sensação de queimação e ardor cutâneo (CAMPBELL et al., 2001; AOKI et al., 2011).

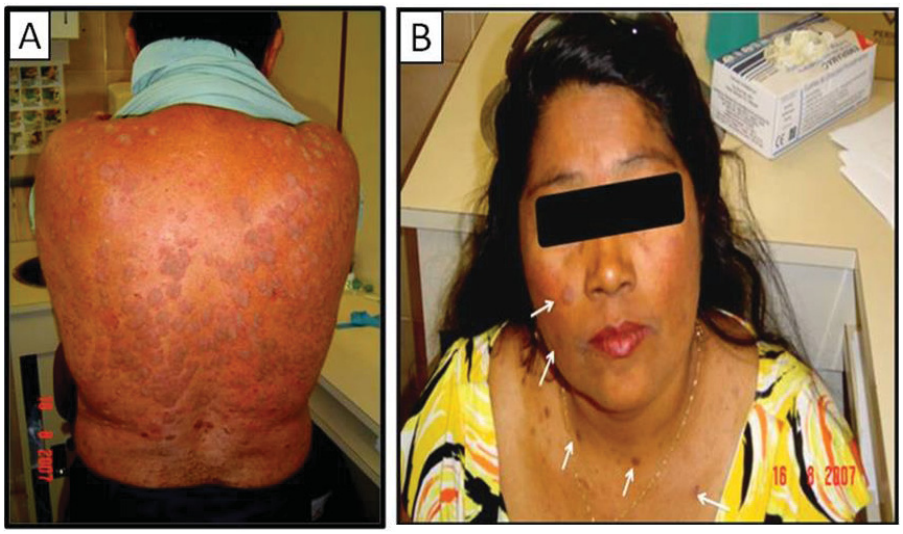


FIGURA 8: FORMAS DE MANIFESTAÇÃO DE LESÕES CARACTERÍSTICAS DO PÊNFIGO FOLIÁCEO (PF). a) forma generalizada com bolhas em todo o corpo, as costas em destaque; b) forma localizada, apresentando eritema na face e pescoço. Fonte: Acervo LGMH.

2.4.1 Pênfigo foliáceo endêmico no Brasil

O Brasil é o país com a maior incidência de PF do mundo, com 25 a 35 casos por milhão por ano (Revisado por CELERE et al., 2017). Em algumas comunidades indígenas, como os Xavantes do Mato Grosso e os Terena do Mato Grosso do Sul, as altíssimas taxas de prevalência chegam a 1,4 e 3,4%, respectivamente, sendo a maior taxa de prevalência já descrita para uma doença autoimune no mundo (HANS-FILHO et al., 1996; SCHMIDT; KASPERKIEWICZ; JOLY, 2019a).

A área endêmica inclui os estados brasileiros de Mato Grosso do Sul, Mato Grosso, Goiás, Minas Gerais, São Paulo e Paraná (FIGURA 9) (LOMBARDI et al., 1992). O PF endêmico no Brasil se diferencia do PF esporádico principalmente pela presença de casos em familiares relacionados geneticamente, em sua epidemiologia, distribuição geográfica e demográfica dos afetados, e pela sensação de dor e ardência nas lesões, motivo pelo qual é chamado de “fogo selvagem”. No entanto, as manifestações clínicas e as características imunopatológicas do PF endêmico são similares à forma esporádica da doença (DIAZ et al., 1989; HANS-FILHO et al., 1999; PALLER; MANCINI, 2011).

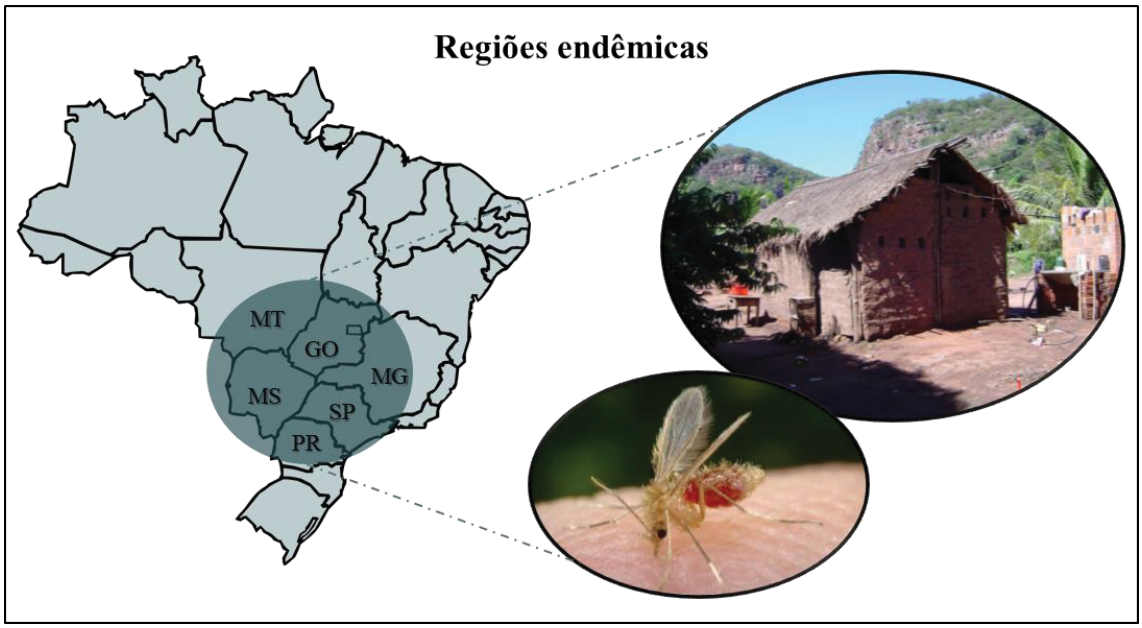


FIGURA 9: FOCOS ENDÊMICOS DE PÊNFIGO FOLIÁCEO (PF) NO BRASIL. Os focos endêmicos abrangem os estados de Mato Grosso (MT), Mato Grosso do Sul (MS), Goiás (GO), Minas Gerais (MG), São Paulo (SP) e Paraná (PR). Em destaque um exemplo das condições de vida de muitos indivíduos na localidade Limão Verde (MS), onde PF é endêmico. Pacientes dessa comunidade também relataram a presença de muitos insetos hematófagos, como pernilongos, barbeiros e percevejo-de-cama. Destaca-se também o inseto *Lutzomyia longipalpis*, espécie mais amplamente estudada como possível inoculador de partículas salivares que contém um fator desencadeador da resposta autoimune primária em pacientes com pênfigo foliáceo endêmico. As imagens foram extraídas de AOKI, et al., 2015 e WILSON, 2009.

2.4.2 Diagnóstico e tratamento do pênfigo foliáceo

O diagnóstico do PF é feito com base em três critérios: 1) história clínica e exame físico; 2) características histopatológicas das biópsias e 3) a presença de autoanticorpos anti-DSG1 detectados por imunofluorescência direta e imunofluorescência indireta. Dentro dos exames físicos o mais importante e específico no diagnóstico de pênfigo é o sinal de Nikolsky, o qual é realizado aplicando pressão na região periférica à lesão primária e secundária. O resultado é positivo se ocorre separação da camada superficial da epiderme (JAMES; CULTON; DIAZ, 2011). Tradicionalmente o tratamento é feito com glucocorticosteroides ou com anticorpos monoclonais contra receptores específicos dos linfócitos B, como o rituximab contra CD20. Porém, estes tratamentos podem trazer vários efeitos secundários indesejados

e potencialmente graves, além de não sempre serem efetivos na eliminação da doença (GRANDO, 2012a; SCHMIDT; KASPERKIEWICZ; JOLY, 2019b).

2.4.3 Fatores desencadeadores do PF

No PF, assim como na maioria das doenças autoimunes de herança complexa, múltiplos fatores genéticos e ambientais contribuem para a susceptibilidade (CULTON *et al.,* 2008; TRON *et al.,* 2005). A existência de regiões endêmicas da doença sugere que existem fatores ambientais específicos dessas áreas que possivelmente desencadeiam a doença. O fato de apenas uma pequena parcela daqueles que habitam as áreas endêmicas desenvolverem a doença indica a existência de fatores genéticos que resultam em uma susceptibilidade diferencial (MORAES et al., 1997). Como serão descritos em detalhes a seguir, existem diversos estudos que apontam que polimorfismos genéticos estão associados ao maior risco de desenvolvimento de PF.

2.4.3.1 Fatores ambientais

Dados epidemiológicos mostram que a doença ocorre em áreas silvestres colonizadas, e que desaparece quando estas áreas são urbanizadas. Em um estudo realizado a partir do ano 1994 na comunidade indígena Terena, em Limão Verde (Mato Grosso do Sul), foi observado que, em comparação a indivíduos controles, a maioria dos pacientes moravam em vivendas precárias, constituídas por chão de terra, paredes de adobe e tetos de palha, sem eletricidade ou encanamento de água apropriados, e relataram terem sido picados por insetos como borrachudos, barbeiros e percevejo-de-cama (AOKI et al., 2004). Adicionalmente, em um estudo na cidade de Goiânia (Goiás), foi relatado que pacientes de PF moram em áreas rurais e com grande exposição à picadas de borrachudos da família Simuliidae (LOMBARDI et al., 1992) .

Dessa forma, uma das hipóteses mais amplamente estudada é a existência de um mosquito como agente desencadeador da autorreatividade, mais especificamente um mosquito da família Simuliidae que poderia transmitir algum vírus ou bactéria, ou ainda inocular proteínas salivares que desencadeiam uma resposta cruzada. Assim, os pacientes que desenvolvem PF endêmico seriam indivíduos que geraram anticorpos contra o antígeno inoculado e, através do fenômeno de

espalhamento intramolecular de epítopos, produzem anticorpos patogênicos contra a DSG1 (LI et al., 2003). Neste fenômeno, anticorpos são gerados inicialmente contra um antígeno externo, que é capaz de reconhecer de forma cruzada epítopos em autoantígenos, produzindo danos aos tecidos próprios. Isto resulta na liberação e exposição de outros epítopos dessa molécula que por sua vez são reconhecidos por outros anticorpos, propagando a resposta autoimune (LI et al., 2003).

Esta hipótese é reforçada em um estudo recente com pacientes e indivíduos em estágios pré-clínicos de PF da comunidade Terena de Limão Verde. Nele foi identificado um epítopo no antígeno LJM11 da glândula salivar de mosquitos *Lutzomyia longipalpis* semelhante a um epítopo de DSG1 no domínio EC2 e que levaria à produção de anticorpos não patogênicos (FIGURA 10). Assim, através do espalhamento de epítopos, posteriormente são produzidos anticorpos patogênicos contra epítopos nos domínios EC1 e EC2 de DSG1, entretanto, também são produzidos anticorpos não patogênicos contra os domínios EC3, EC4 e EC5 de DSG1 (PENG et al., 2020). Desta forma, a reatividade cruzada seria o mecanismo pelo qual uma resposta imune contra um antígeno exógeno leva a produção de autoanticorpos patogênicos em indivíduos geneticamente susceptíveis (PENG et al., 2020).

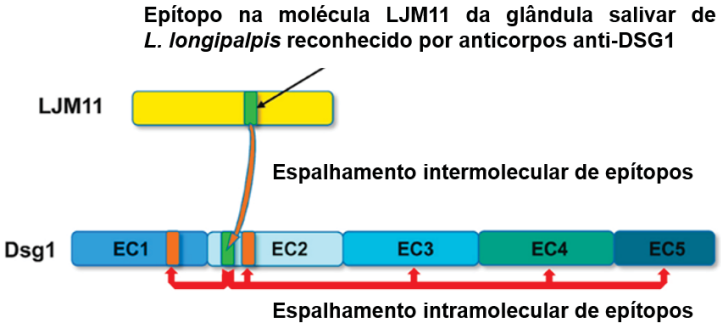


FIGURA 10: ESPALHAMENTO DE EPÍTOPO NO DESENVOLVIMENTO DE ANTICORPOS ANTI-DSG1. A partir de uma reação cruzada de anticorpos que reconhecem primeiramente um epítopo no antígeno LJM11 da glândula salivar de mosquitos *Lutzomyia longipalpis* e posteriormente um epítopo não patogênico em DSG1, outros anticorpos são gerados contra outros epítopos de DSG1 pelo fenômeno de espalhamento de epítopos. Estes anticorpos podem ser patogênicos quando reconhecem os domínios EC1 e EC2 de DSG1 e não patogênicos quando reconhecem os domínios EC3, EC4 e EC5 de DSG1 (PENG et al., 2020).

Outra evidência da existência de fatores ambientais agregados na área endêmica desencadeando a doença é a constatação de que indivíduos saudáveis também apresentam anticorpos anti-DSG1, cujas concentrações em sangue

aumentam conforme a proximidade da residência dessas pessoas com a área endêmica (WARREN et al., 2000).

Adicionalmente, em outro estudo recente realizado com soro total de pacientes e controles de PF de uma área endêmica no nordeste de São Paulo, foram observadas correlações entre os níveis de anti-DSG, anticorpos contra outras moléculas dos desmossomos, e anticorpos contra o antígeno maxadilan das glândulas salivares de *L. longipalpis*. Entretanto não foram encontradas correlações dos níveis de anti-DSG com os níveis de anticorpos anti-LJM11. Esses resultados sugerem que os anticorpos desses pacientes de PF não reconhecem LJM11, mas sim reconhecem o antígeno maxadilan, sugerindo que seria uma exposição constante a um pool de antígeno salivares o que estimularia o desenvolvimento de PF (VERNAL et al., 2020a). Desta forma, os pesquisadores não descartam que existam outros antígenos desencadeadores na área endêmica, principalmente porque os trabalhos focaram somente em antígenos com reação cruzadas com DSG1, sem incluir outros autoantígenos (PENG et al., 2020; VERNAL et al., 2020a).

Outro fator ambiental associado ao pênfigo são as infecções virais. Dos mecanismos descritos para explicar esta associação, a mais aceitável é um possível mimetismo molecular entre os vírus e as proteínas epidérmicas, e uma ativação do sistema imune como consequência do ataque viral. Este mecanismo envolve inflamação e citocinas que podem desencadear uma resposta imunopatológica, levando à proliferação de células T autorreativas pré-existentes (Revisado em RUOCCO et al., 2014). Adicionalmente, foi observado o surgimento de pênfigo após infecções virais (por exemplo os vírus do herpes simples (VHS), varicela-zoster (VVZ), Citomegalovirus (CMV), etc.) e tratamento com penicilina, cefalosporina, entre outros, em um fenômeno denominado de erupção “paraviral”, já que não resulta diretamente da infecção viral e sim da resposta do hospedeiro à presença do vírus dentro da pele. Desta forma, as células infectadas com os vírus liberariam moléculas endógenas que na presença de estímulos externos podem atuar como adjuvantes naturais iniciando uma resposta autoimune (Revisado em RUOCCO et al., 2014).

2.4.3.2 Fatores genéticos

A importância dos fatores genéticos é evidenciada pela agregação familial de indivíduos com a doença, já que foi descrito que 20% dos pacientes de pênfigo foliáceo endêmico tem familiares com a doença (ABRÉU-VÉLEZ et al., 2010). Ainda,

o envolvimento de fatores genéticos foram demonstrados pela vasta quantidade de estudos que mostram associação de variantes genéticas com o risco diferencial a desenvolver PF (AUGUSTO et al., 2021, Revisado em PETZL-ERLER, 2020).

Alelos e haplótipos de genes *HLA* classe II apresentam as associações mais fortes com PF. Foram associados com susceptibilidade ao pênfigo foliáceo endêmico do Brasil os alelos *DRB1*01:02, *04:04, *01:01, *01:03, *04:06, *14:06, *16:01* e o haplótipo *DRB1*01-DQA1*01:01-DQB1*05:01*; e com proteção os alelos *DRB1*07:01, *03:01, *14:02, *08:01, *08:02, *08:03, *08:04, *08:07, *11:01, *11:02, *11:03, *11:04, *15:01, *15:02, *15:03* e o haplótipo *DRB1*07:01-DQA1*02:01-DQB1*02:01* (BROCHADO et al., 2016; MORAES et al., 1997; PAVONI et al., 2003). Estas moléculas são importantes na apresentação de antígenos aos linfócitos T CD4+, podendo então determinar a especificidade dos antígenos reconhecidos por eles, e dos linfócitos B que serão posteriormente ativados. Também foram associados genes *HLA de classe I*, assim como outros pertencentes ao complexo MHC (AUGUSTO et al., 2021; BROCHADO et al., 2016; PIOVEZAN; PETZL-ERLER, 2013). As moléculas de HLA de classe I também apresentam antígenos aos linfócitos T (CD8+) e podem servir como ligantes de receptores KIR das células natural killers (NK), cujos polimorfismos também se encontram associados ao PF, sendo eles tanto alelos dos genes *KIR*, como variantes do tipo ausência/presença (AUGUSTO et al., 2012, 2015).

Polimorfismos em genes dos receptores LAIR1 e LAIR2, necessários para regulação de respostas imunes de linfócitos T e B, células NK, e outras células mononucleares do sangue também foram associados com PF (CAMARGO; AUGUSTO; PETZL-ERLER, 2016), assim como variantes do gene *KLRG1* expresso em células NK e linfócitos T (CIPOLLA et al., 2016).

Outros estudos com PF mostraram associação entre PF e o polimorfismo dos genes de citocinas necessárias para ativação, coestimulação e diferenciação de linfócitos T e B, como *PDCD1*, *CD86, CTLA4, CD40*, *CD40L*, *TNFSF13B*, *IL6*, e *IL4* (BRAUN-PRADO; PETZL-ERLER, 2007; DALLA-COSTA et al., 2010; MALHEIROS; PETZL-ERLER, 2009; PEREIRA et al., 2004). Da mesma maneira, também já foram associadas variantes em genes participantes das vias do sistema complemento do sistema imune inato: *C3*, *C5AR1*, *C8A*, *C9*, *CFH*, *CR2*, *ITGAM*, *ITGAX*, *MASP1*, *CR1* e *CD59* (BUMILLER-BINI et al., 2018; OLIVEIRA et al., 2019; SALVIANO-SILVA; PETZL-ERLER; BOLDT, 2017).

Adicionalmente, estudos realizados analisando os clonotipos de anticorpos anti-DSG em pacientes de pênfigo mostraram que alguns segmentos gênicos *IGHV* foram utilizados com maior frequência. Por exemplo, o segmento gênico *IGHV1-46* se encontra em alta frequência em autoanticorpos anti-DSG3 de pacientes com pênfigo vulgar e estudos de afinidade mostraram que este segmento gênico precisa de poucas ou nenhuma mutação somática para se ligar à DSG3 (CHO et al., 2014), e IGHV3-23 e IGHV3-30 foram encontrados em alta frequência em anticorpos anti-DSG1 em pacientes de pênfigo foliáceo (QIAN et al., 2009).

2.4.4 Imunopatologia do pênfigo

Como mencionado anteriormente, o pênfigo é caracterizado pelo aparecimento de bolhas na pele e/ou mucosas causadas principalmente por anticorpos patogênicos dirigidos principalmente contra moléculas de adesão celular das células epidérmicas (AMAGAI, 2002; TRON et al., 2005).

2.4.4.1 Autoantígenos no pênfigo

Em pacientes de pênfigo foliáceo, as lesões são encontradas nas camadas superficiais da pele. Já em pacientes com pênfigo vulgar, as lesões podem aparecer nas camadas mais profundas da mucosa, e pode também acometer a pele (GRANDO, 2012a; SCHMIDT; KASPERKIEWICZ; JOLY, 2019a). Os principais antígenos alvo no pênfigo são as desmogleínas 1 (DSG1) em PF e as DSG3 em PV (AMAGAI, 2002; TRON et al., 2005).

As DSG são glicoproteínas transmembrânicas de adesão dependentes de Ca^{2+}, localizadas nos desmossomos (FIGURA 11), estruturas de junção entre células epiteliais. O perfil de expressão das DSG1 e DSG3 difere entre as camadas da pele e mucosas: a DSG1 é expressa em toda a epiderme e mais frequentemente na camada subcórnea da pele (mais externa) e a DSG3 se encontra nas camadas profundas da epiderme, principalmente na camada basal e parabasal mais profundas das mucosas (FIGURA 12). Um dos argumentos para explicar a localização das lesões em PF e PV é a hipótese de compensação da expressão das DSG (FIGURA 12). Em função desse perfil de expressão, em pacientes de PF, os anticorpos anti-DSG1 levam ao desprendimento da camada mais externa da epiderme, e a formação de bolhas na superfície da pele. Em pacientes de PV no entanto, as bolhas são formadas a partir da ruptura da camada basal das mucosas. Quando pacientes de PV produzem tanto

anticorpos anti-DSG1 como anti-DSG3, podem acontecer rupturas suprabasais nas mucosas e pele (AMAGAI, 2002).

Apesar do papel aparentemente central das desmogleínas, colocar DSG1 e DSG3 no centro do processo patofisiológico de pênfigo tem sido questionado, devido a que DSG1 e DSG3 sozinhas não são capazes de manter a integridade da epiderme, e as observações histológicas não condizem com o esperado no caso de perda de função de desmogleínas (GRANDO, 2012a).

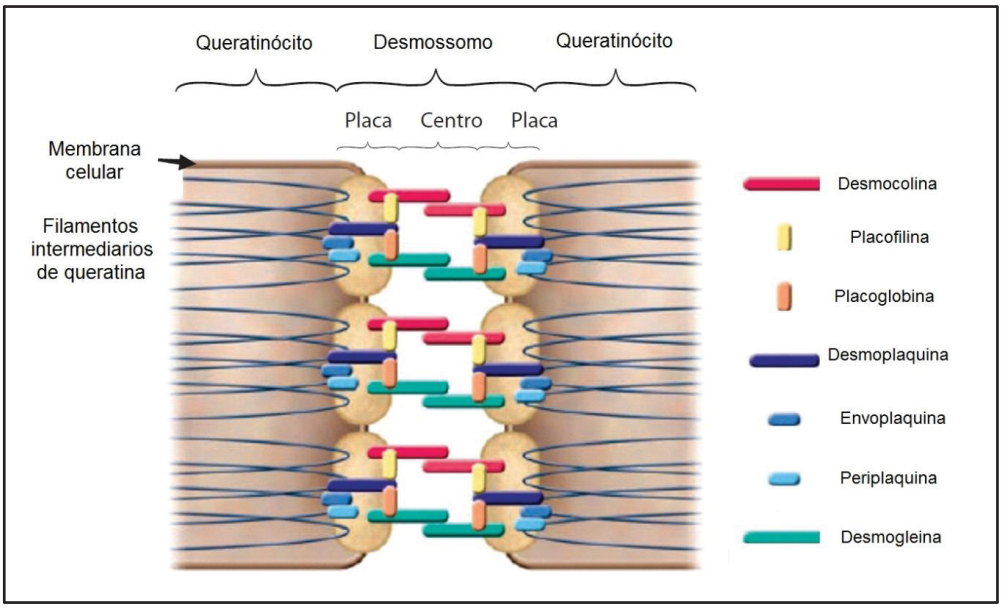


FIGURA 11: ESQUEMATIZAÇÃO DAS PROTEÍNAS QUE CONFORMAM OS DESMOSSOMOS, ESTRUTURA DE ADESÃO ENTRE QUERATINÓCITOS. Esta união é formada pelos filamentos intermediários de queratina, pelas placas de ancoramento de proteínas e as proteínas de adesão no meio extracelular desmossomal: desmoplaquina, placoglobina, placofilina, desmogleína e desmocolina. Algumas destas moléculas podem atuar como autoantígenos no pênfigo. Imagem adaptada de HERTL; EMING; VELDMAN, 2006.

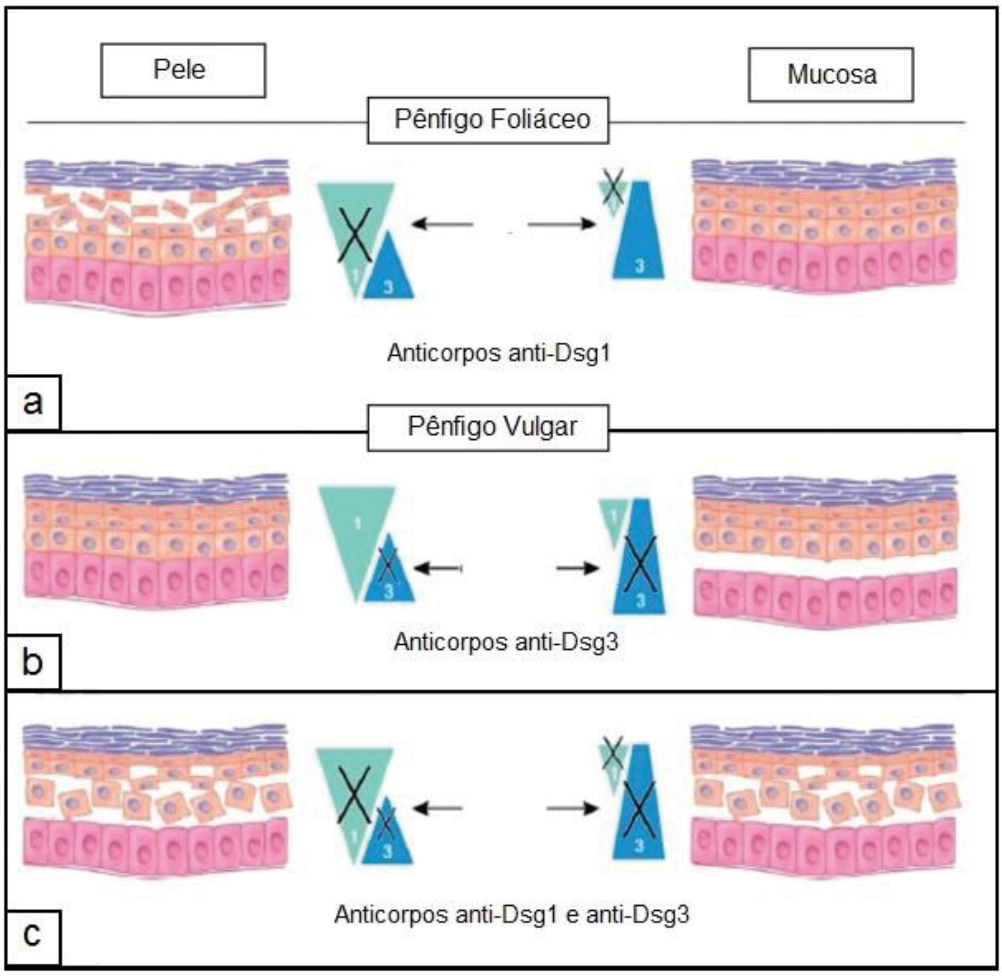


FIGURA 12: HIPÓTESE DE COMPENSAÇÃO E PERFIL DE EXPRESSÃO DAS DSG 1 E 3 NA PELE E MUCOSA. a) Os anticorpos anti-DSG1 provocam lesões na pele em PF, onde DSG3 é restrita, mas não na mucosa, onde DSG3 compensa a perda de função de DSG1. b) Anticorpos anti-DSG3 provocam lesões nas membranas mucosas em PV, onde DSG1 é restrita, mas não na pele já que a DSG1 compensa a função. c) Anticorpos anti-DSG1 e anti-DSG3 provocam lesões tanto na pele quanto nas mucosas em PV. Imagem adaptada de STANLEY (2008).

Além disso, mais de 50 proteínas humanas foram detectadas como autoantígenos reconhecidos por autoanticorpos de pacientes com pênfigo (GRANDO, 2012b; KAZEROUNIAN et al., 2000). Dentre eles, outras proteínas da placa desmossômica (FIGURA 11). Estas poderiam ser alvo de anticorpos devido a de reação cruzada com os anticorpos anti-DSG (PENG et al., 2020; VERNAL et al., 2020a).

Adicionalmente, em pacientes de PF endêmico da Colômbia foram encontrados autoanticorpos contra tecido nervoso periférico e central, e alterações nos tecidos nervosos, o que poderia explicar a sensação de ardência e dor nas lesões

manifestada na forma endêmica da doença mas ausente nas lesões de pacientes de PF esporádico (ABREU-VELEZ et al., 2011).

2.4.4.2 Papel dos autoanticorpos no processo acantolítico

A patogenicidade dos autoanticorpos anti-DSG já foi demonstrada por transferência passiva em camundongos de anticorpos anti-DSG1 de soro de pacientes de PF, com posterior desenvolvimento de bolhas nas camadas subcorneas da epiderme (ANHALT et al., 1982; ROSCOE et al., 1985). Também foi observado que mulheres grávidas com PV e PF com lesão ativa, podem dar à luz filhos com fenótipo temporário de pênfigo, que desapareceram algumas semanas após o nascimento, mostrando que a manifestação da doença é devida a transferência dos anticorpos patogênicos pela barreira placentária (AVALOS-DÍAZ et al., 2000; BIALYNICKI-BIRULA et al., 2011)

Análises imunohistoquímicas indicam que os anticorpos dirigidos contra DSG se ligam predominantemente ao domínio extracelular da proteína e que exercem sua propriedade patogênica inibindo as propriedades de adesão destas moléculas. Em consequência disso, formam-se as bolhas características da doença com desprendimento da camada superficial da epiderme, fenômeno conhecido como acantólise (TRON et al., 2005).

Além disso, os autoanticorpos podem ativar o sistema complemento no PF, levando à perda de coesão dos queratinócitos e ao desenvolvimento de bolhas intraepidérmicas. Componentes das vias clássicas e alternativas do sistema complemento, como C1q, C3, C4, C5, C7, C9 e o neoantigeno-MAC, são encontrados na substância intercelular em lesões de pacientes, assim como C3 e IgG (IgG1 e IgG4) entre os queratinócitos, sendo a IgG1 um potente ativador do sistema complemento. Dessa forma, a formação de MAC (complexo de ataque à membrana) nas células da epiderme, resultante da ativação do sistema complemento, teria um efeito tóxico e patogênico (Revisado em PANELIUS; MERI, 2015).

Outra hipótese para explicar a patogênese de pênfigo é o processo nomeado de apoptólise (apoptose + acantólise). Neste contexto, a ligação de autoanticorpos aos antígenos de PV ativaria receptores e sinais intracelulares como o fator EGF (do inglês *epidermal growth factor*) e quinases p38-MAP (do inglês *mitogen-activated protein kinase*) que induzem diversas vias de morte celular. Posteriormente, as

proteínas celulares são clivadas e estimulam a produção de autoanticorpos secundários, seguido de acantólise e morte celular (GRANDO, 2012a). Já em PF, foi demonstrada que a variabilidade em genes codificadores de diferentes vias de morte celular são importantes na susceptibilidade diferencial ao PF (BUMILLER-BINI et al., 2019).

Apesar das fortes evidências da patogenicidade dos anticorpos anti-DSG em pênfigo, estes autoanticorpos também são encontrados em indivíduos saudáveis. No caso de PF foram observados anticorpos IgG1 anti-DSG1 em indivíduos saudáveis da região endêmica, e em pacientes no estágio pré-clínico. E o aparecimento da doença está acompanhado da emergência de anticorpos IgG4 anti-DSG1. Também, controles de PV, ou seja, indivíduos saudáveis que possuem os alelos de *HLA* de susceptibilidade para o PV, também apresentam anticorpos anti-DSG3 da subclasse IgG1, em contraste aos anti-DSG3 IgG4 predominante dos pacientes (LI et al., 2003; SANTI; SOTTO, 2001).

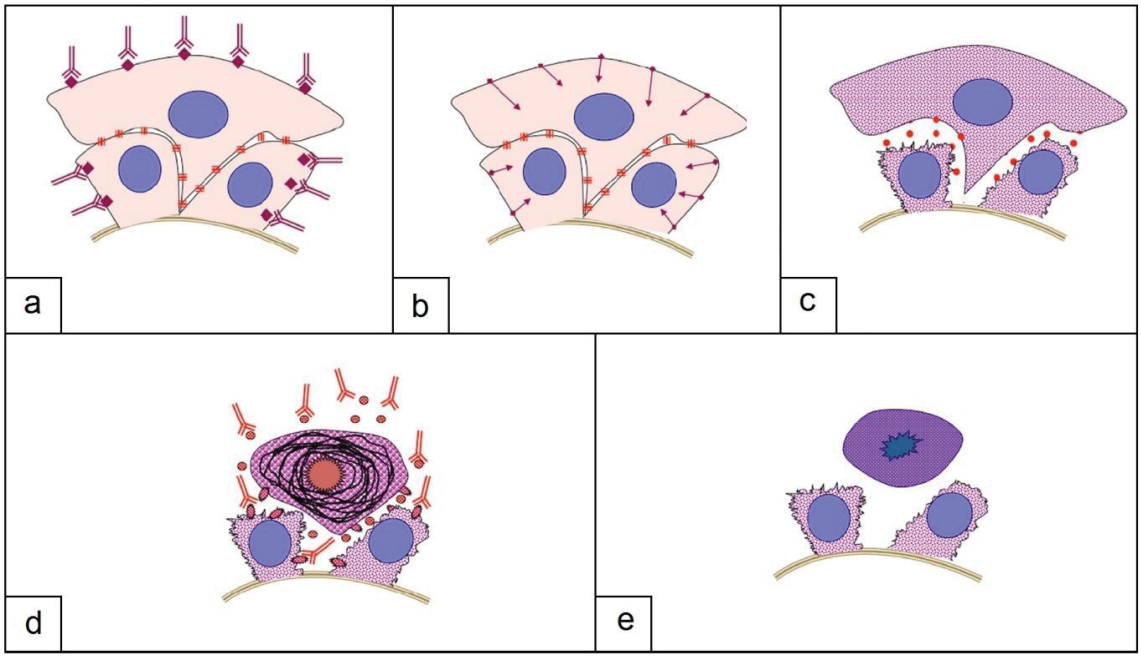


FIGURA 13: MODELO ESQUEMÁTICO DA ATIVAÇÃO DO PROCESSO DE APOPTÓLISE NO PÊNFIGO VULGAR. a) A ligação de autoanticorpos aos seus antígenos envia o sinal apoptolítico. b) A cascata de morte celular é ativada. c) Ocorre uma fosforilação das moléculas de adesão que resultaria numa contração das células basais. d) Uma clivagem massiva de proteínas leva ao colapso do citoesqueleto e liberação de desmossomos da membrana celular com a subsequente produção de autoanticorpos secundários. e) Morte das células suprabasais. Adaptado de Grando (2012).

2.4.4.3 Papel dos linfócitos B e linfócitos T no pênfigo

Apesar de ser uma doença caracterizada por anticorpos patogênicos e, portanto, mediada por linfócitos B, o desenvolvimento da autoimunidade em pênfigo requer também o envolvimento de linfócitos T. Assim, a perda da autotolerância, que resulta na produção de anticorpos anti-DSG, depende da desregulação tanto dos linfócitos T como dos linfócitos B (GRANDO, 2012a; PAN; ZHU; XU, 2015).

No pênfigo, a primeira etapa da produção de autoanticorpos é possivelmente desencadeada por um antígeno exógeno com epítopo similar à DSG, que levaria a uma reação cruzada (Revisado por PAN et al., 2015). Como mencionado anteriormente, em PF endêmico do Brasil, os antígenos salivares maxadilan e LJM11 seriam um dos desencadeadores da produção de anticorpos com reação cruzada com DSG1 (PENG et al., 2020; VERNAL et al., 2020b).

Em pênfigo foliáceo foi observado que a ativação de células B anti-DSG1 envolve células T CD4+, que reconhecem epítopos específicos do ectodomínio de DSG1 na presença de moléculas de HLA de alelos associados com a doença, induzindo a produção de citocinas características de resposta Th2, como IL-4, IL-5 e IL-6 (AOKI et al., 2004; HANS-FILHO et al., 1999). Também em PF foram encontrados níveis de expressão gênica diferencial entre pacientes e controles de PF em células T CD4+, incluindo genes relacionados à adesão e migração de linfócitos, apoptose, proliferação celular, citotoxicidade e apresentação de antígenos (MALHEIROS et al., 2014). Do mesmo modo, em pacientes com PV foram encontrados linfócitos T CD4+ que reconhecem DSG3 no contexto de moléculas de HLA de alelos associados com a doença, como HLA-DRB1*0402, DQB1*0503 (EMING et al., 2014).

Os linfócitos T auxiliares Th2 têm um papel crítico nesse processo. Em um estudo em PV foi observado que pacientes possuíam maiores concentrações de citocinas Th2 (IL-4, IL-10) que controles (SATYAM et al., 2009). Em outro estudo foi encontrada uma maior expressão de células Th2 anti-DSG3 em pacientes de PV, em contraste aos controles com alelos de HLA associados à PV, os quais possuem uma maior expressão de linfócitos Th1 anti-DSG3 (VELDMAN et al., 2003). Também foi observado que os níveis séricos de anticorpos anti-DSG3 se encontravam correlacionados com os níveis de atividade de Th2 anti-DSG3 em pacientes (RIZZO et al., 2005). O perfil de citocinas produzidas pelos linfócitos T determina o isotipo de anticorpo que será expresso pelos linfócitos B após o processo de mudança de classe por recombinação. Os sinais liberados pelos linfócitos Th2, como IL-4, levam à

mudança para IgG4 e IgE, e os sinais Th1 levam à expressão de IgG1 e IgG2 (FINKELMAN et al., 1990; KOTOWICZ; CALLARD, 1993).

Outra população de linfócitos T que parece participar na patogênese de pênfigo, e de outras doenças autoimunes, é a de linfócitos T reguladores (Treg) (PAN; ZHU; XU, 2015), os quais ajudam a manter a homeostasia do sistema imune adaptativo (O'GARRA; VIEIRA, 2004). Em pacientes de PV foi encontrado que tanto o número de células Treg, quanto os níveis de expressão de Foxp3, um fator de transcrição que controla o desenvolvimento destas células, se encontrava drasticamente diminuído em pacientes de PV (SUGIYAMA et al., 2007). Portanto, os autores concluem que pacientes de pênfigo apresentam uma desregulação na população de Treg, que seriam incapazes de inibir os linfócitos T e B autorreativos.

3 JUSTIFICATIVA E HIPÓTESE

O pênfigo foliáceo é uma doença autoimune na qual autoanticorpos têm um papel determinante na sua patogênese. Apesar disso, os polimorfismos que afetam as vias envolvidas nos rearranjos somáticos necessários para expressão dos anticorpos, incluindo os próprios genes de imunoglobulinas, foram pouco estudados no contexto de pênfigo.

Resultados anteriores que analisaram os genes de imunoglobulinas no contexto de pênfigo focaram exclusivamente em anticorpos anti-DSG. Até o momento, não existem dados do repertório total de imunoglobulinas em pacientes e controles. Por se tratar de uma doença autoimune e endêmica em certas regiões do mundo, o estudo do repertório de imunoglobulinas de indivíduos que residem na área endêmica é particularmente interessante pois pode informar o comportamento do repertório sob a pressão ambiental existente nessa região. Os fatores que determinam a patogenicidade dos autoanticorpos em pacientes com PF não são completamente compreendidos. O conhecimento de características específicas da região endêmica no repertório de imunoglobulinas tem o potencial de contribuir para o esclarecimento de fatores regionais que desencadeiam PF.

A hipótese deste trabalho é que polimorfismos dos genes que codificam moléculas envolvidas no desenvolvimento de anticorpos influenciam na susceptibilidade diferencial ao pênfigo foliáceo. Ainda, que o repertório de imunoglobulinas difere entre indivíduos com pênfigo foliáceo e indivíduos saudáveis.

O PF endêmico é um problema de saúde pública no Brasil. Trata-se de uma doença negligenciada de alta incidência em populações vulneráveis, que possui etiologia pouco compreendida e um tratamento que causa graves efeitos colaterais. Portanto, estudos que contribuem para a compressão da etiologia de PF são de extrema relevância e urgência, principalmente no Brasil, país em que existe o maior número de afetados no mundo. Apesar desse presente estudo não gerar aplicações diretas ao tratamento de pacientes, o estudo dos genes de imunoglobulinas e do repertório de células B tem o potencial de fornecer a base de estudos futuros que buscam melhorar a compreensão, o tratamento e prognóstico dessa doença.

4 OBJETIVOS

4.1 OBJETIVO GERAL

Verificar se polimorfismos em genes que codificam as moléculas envolvidas no desenvolvimento dos anticorpos impactam na susceptibilidade ao pênfigo foliáceo endêmico. Ainda, verificar se as características do repertório de imunoglobulinas diferem entre pacientes de pênfigo foliáceo endêmico e indivíduos saudáveis.

4.2 OBJETIVOS ESPECÍFICOS

I. Verificar se as variantes dos genes que codificam as cadeias pesadas (*IGH*), as cadeias leves (*IGK* e *IGL*) e daqueles que codificam as moléculas envolvidas no desenvolvimento e ativação dos linfócitos B, no rearranjo somático, na mudança de classe por recombinação e na hipermutação somática conferem um risco diferencial de desenvolver pênfigo foliáceo endêmico.

II. Caracterizar o repertório de IGHM e IGHG, resultantes da expressão dos genes de imunoglobulinas que codificam a cadeia pesada dos anticorpos IgM e IgG, em pacientes de pênfigo foliáceo endêmico e em indivíduos saudáveis da região endêmica.

III. Avaliar se as características do repertório de IGHM e IGHG diferem entre pacientes de pênfigo foliáceo endêmico e indivíduos saudáveis da região endêmica.

5 RESULTADOS

Os resultados dessa tese serão divididos em dois capítulos, cada um apresentado na forma de um manuscrito. O primeiro manuscrito, intitulado "*Variation in genes implicated in B-cell maturation and antibody production affects susceptibility to pemphigus*" foi publicado na revista *Immunology* (DOI: 10.1111/imm.13259). O segundo, intitulado "*The landscape of the immunoglobulin repertoire in endemic pemphigus foliaceus*" até presente data não foi submetido para publicação. Os materiais suplementares dos artigos dos capítulos 1 e 2 se encontram no apêndice 1 e 2, respectivamente.

5.1 CAPÍTULO 1 - VARIATION IN GENES IMPLICATED IN B-CELL MATURATION AND ANTIBODY PRODUCTION AFFECTS SUSCEPTIBILITY TO PEMPHIGUS

Immunology
The Journal of cells, molecules, systems and technologies
British Society for immunology

IMMUNOLOGY ORIGINAL ARTICLE

Variation in genes implicated in B-cell development and antibody production affects susceptibility to pemphigus

Verónica Calonga-Solís,[1] Leonardo M. Amorim,[1] Ticiana D. J. Farias,[1] Maria Luiza Petzl-Erler,[1] Danielle Malheiros[1] and Danillo G. Augusto[1,2]

[1]*Programa de Pós-Graduação em Genética, Departamento de Genética, Universidade Federal do Paraná, Curitiba, Brasil and* [2]*Department of Neurology, University of California San Francisco, San Francisco, CA,USA*

doi:10.1111/imm.13259
Received 8 June 2020; revised 23 August 2020; accepted 29 August 2020.
Correspondence: Danillo G. Augusto, Department of Neurology, University of California San Francisco, 675 Nelson Rising Lane, room 240. San Francisco, CA 94148, USA.
Email: danillo@augusto.bio.br
Senior author: Danillo G. Augusto

Abstract

Pemphigus foliaceus (PF) is an autoimmune blistering skin disease characterized by the presence of pathogenic autoantibodies against desmoglein 1, a component of intercellular desmosome junctions. PF occurs sporadically across the globe and is endemic in some Brazilian regions. Because PF is a B-cell-mediated disease, we aimed to study the impact of variants within genes encoding molecules involved in the different steps of B-cell development and antibody production on the susceptibility of endemic PF. We analysed 3,336 single nucleotide polymorphisms (SNPs) from 167 candidate genes genotyped with Illumina microarray in a cohort of 227 PF patients and 193 controls. After quality control and exclusion of non-informative and redundant SNPs, 607 variants in 149 genes remained in the logistic regression analysis, in which sex and ancestry were included as covariates. Our results revealed 10 SNPs within or nearby 11 genes that were associated with susceptibility to endemic PF (OR >1.56; $p < 0.005$): *rs6657275*G* (*TGFB2*); *rs1818545*A* (*RAG1/RAG2/IFTAP*); *rs10781530*A* (*PAXX*), *rs10870140*G* and *rs10781522*A* (*TRAF2*); *rs535068*A* (*TNFRSF1B*); *rs324011*A* (*STAT6*); *rs6432018*C* (*YWHAQ*); *rs17149161*C* (*YWHAG*); and *rs2070729*C* (*IRF1*). Interestingly, these SNPs have been previously associated with differential gene expression, mostly in peripheral blood, in publicly available databases. For the first time, we show that polymorphisms in genes involved in B-cell development and antibody production confer differential susceptibility to endemic PF, and therefore are candidates for possible functional studies to understand immunoglobulin gene rearrangement and its impact on diseases.

Keywords: genetics; immunogenetics; immunoglobulin; immunology; pemphigus foliaceus.

Keywords: pemphigus foliaceus, immunogenetics, immunology, genetics, immunoglobulin

Abbreviations:

Add	additive model
AID	activation-induced cytidine deaminase molecule
AICDA	activation-induced cytidine deaminase gene
AIMs	ancestry-informative markers
BMP	bone morphogenetic protein
CIITA	class II major histocompatibility complex trans activator
cpm	case per million
CSR	class switch recombination
DNA-PKcs	DNA-dependent protein kinase, catalytic subunit
Dom	dominant model
eQTL	expression quantitative trait loci
HGDP-CEPH	Human Genome Diversity Project - Centre d'Etude du Polymorphisme Humain
HLA	human leukocyte antigen
HSC	hematopoietic stem cells
IFTAP	intraflagellar transport associated protein
Ig	immunoglobulin
IGH	immunoglobulin heavy locus
IGK	immunoglobulin kappa locus
IGL	immunoglobulin lambda locus
IL	interleukin
IRF1	interferon regulatory factor 1
LD	linkage disequilibrium
MAF	minor allele frequency
NHEJ1	non-homologous end-joining factor 1 (XLF)
OR	odds ratio
PAXX	PAXX non-homologous end joining factor
PF	pemphigus foliaceus
RAG1	recombination-activating gene 1
RAG2	recombination-activating gene 2
Rec	recessive model
SHM	somatic hypermutation
SNP	single nucleotide polymorphism
sQTL	splicing quantitative trait loci
STAT6	signal transducer and activator of transcription 6
TGFB2	transforming growth factor beta 2

TNF	tumor necrosis factor
TNFRSF1B	TNF receptor superfamily member 1B
TRAF2	TNF receptor-associated factor 2
XRCC4	X-ray repair cross complementing 4
XRCC6	X-ray repair cross complementing 6 (Ku70)
XRCC5	X-ray repair cross complementing 5 (Ku80)
YWHAG	tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein gamma
YWHAQ	tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein theta

INTRODUCTION

Pemphigus foliaceus (PF) is a blistering skin disease characterized by epidermal cell detachment (acantholysis) in the upper layer of the epidermis. In PF, the loss of cell adhesion is a consequence of the presence of autoantibodies, mostly IgG1 and IgG4, against desmoglein 1 (DSG1), a desmosomal component of keratinocytes (Amagai and Stanley 2012; Kasperkiewicz et al. 2017). PF occurs sporadically across the globe, with an incidence of one case per million (cpm) (Bastuji-Garin et al. 1995; Marazza et al. 2009; Joly and Litrowski 2011; K. Kumar 2008). In Brazil, however, its incidence reaches 25 to 35 cpm in some endemic regions (Diaz et al. 1989). As a multifactorial disease, multiple genetic and environmental factors contribute to the risk of developing PF. The environmental factors that trigger the disease in the Brazilian endemic regions are not well established; however, they are possibly related to exposure to sunlight, mosquito bites, certain foods, and poor living conditions (Lombardi et al. 1992). In terms of susceptibility, several genetic variants have been identified as playing a role in the risk of developing PF (Maria Luiza Petzl-Erler 2020), including *HLA* (*human leukocyte antigen)* genes (Pavoni et al. 2003; M L Petzl-Erler and Santamaria 1989; Brochado et al. 2016; Piovezan and Petzl-Erler 2013), *KIR* (*killer-cell immunoglobulin-like receptor)* (Augusto et al. 2012, 2015; Farias et al. 2019) genes, genes of the complement system (V. Bumiller-Bini et al. 2018; Salviano-Silva, Petzl-Erler, and Boldt 2017; Oliveira et al. 2019), among others (Malheiros and Petzl-erler 2009; Cipolla et al. 2016; Lobo-Alves et al. 2019; Spadoni et al. 2019; Valéria Bumiller-Bini et al. 2019).

Antibodies are immunoglobulin (Ig) molecules produced by B cells after a series of somatic rearrangements in immunoglobulin genes. Structurally, antibodies can be divided into the variable domain, responsible for antigen binding, and the constant domain, which determines their effector function. Ig is a homoheterodimer composed by two identical heavy chains and two identical light chains; the heavy chain is encoded by the *immunoglobulin heavy locus* (*IGH*), and the light chains are encoded by *immunoglobulin lambda locus* (*IGL*) or *immunoglobulin kappa* (*IGK*) locus (Schroeder and Cavacini 2010). Each one of these genes is composed of multiple gene segments in a way that their germline configuration does not allow the transcription to a functional mRNA. To be transcribed, these genes first undergo a complex somatic DNA rearrangement process called V(D)J recombination during the

initial steps of B-cell development (Dudley et al. 2005), which results in a V(D)J exon that will encode the variable region (Jung et al. 2006; Bassing, Swat, and Alt 2002; Dudley et al. 2005). Following antigen encountering, a process called somatic hypermutation allows positive selection of B cells that exhibit Ig with increased antigen-binding affinity and results in a more effective immune response (Jacobs and Bross 2001). Finally, immunoglobulin genes undergo a process called class-switch recombination (CSR) that changes the isotype expressed by the B cells, from IgM and IgD to IgA, IgG, or IgE, (Figure 1), which changes the Ig effector function (Honjo 1983). All these processes involve molecules responsible for the recognition of target sequences, double-strand cleavage, non-homologous end joining, nucleotide deamination, excision, and addition. Because PF is an antibody-mediated disease, we hypothesized that germline variation in genes that affect the development of immunoglobulins are candidates for disease association. Here, we screened 3,336 variants in 167 genes directly involved in the antibody production and observed that ten single nucleotide polymorphisms (SNPs) were associated with PF.

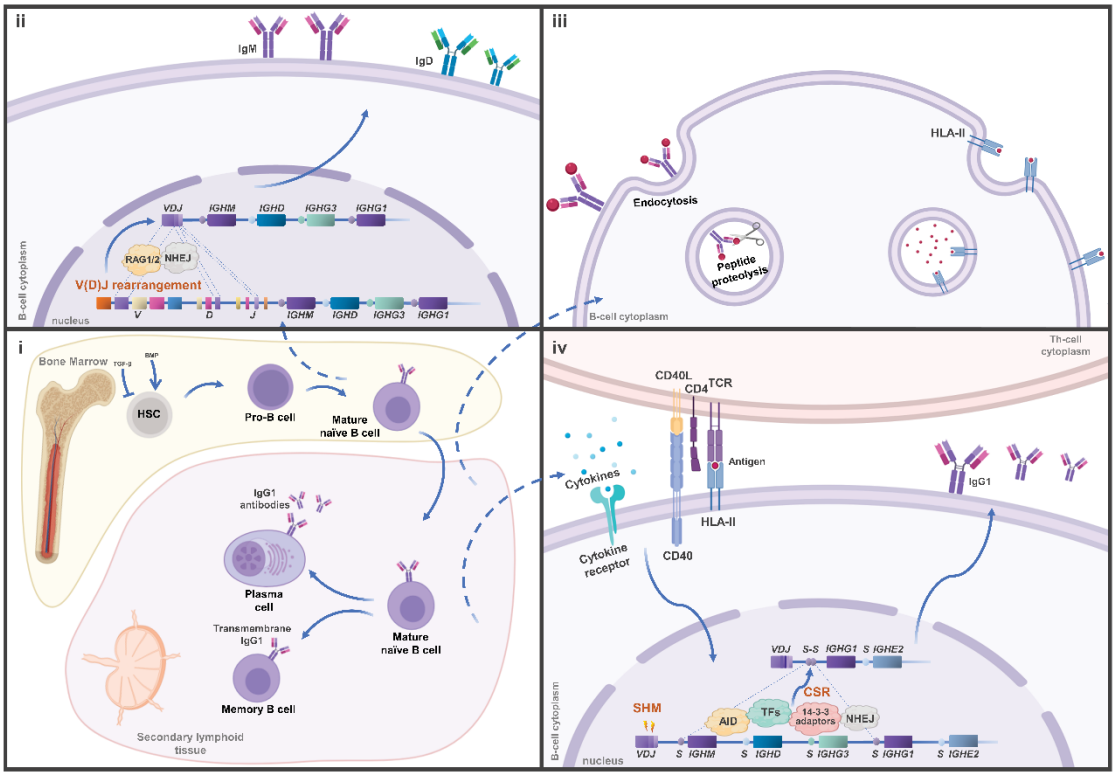


FIGURE 1: STAGES OF ANTIBODY PRODUCTION i) In the bone marrow, hematopoietic stem cells (HSC) can give rise to common lymphoid progenitor when stimulated by bone morphogenetic protein (BMP) signaling and by the suppression of transforming growth factor-beta (TGF-β) stimulation. ii) During B-cell development, in the pro-B cell stage, immunoglobulin heavy chain genes (*IGH*) undergo a somatic reassembly of their V, D, and J segments through DNA cleavage by RAG1 and RAG2 (V(D)J recombination-activating proteins 1 and 2), followed by non-homologous end joining (NHEJ) of the DNA.

These steps result in a VDJ exon that will encode the variable region of Ig heavy chain and will be expressed with the adjacent constant gene segments, *IGHM* and *IGHD* (IgM and IgD heavy chain). If this rearrangement is successful, V and J gene segments of immunoglobulin light chains (kappa or lambda) are also rearranged (not shown), and complete IgM and IgD molecules are expressed. The cells are now called mature naïve B cells. iii) In secondary lymphoid organs, immunoglobulin molecules on the surface of naïve B cells recognize and bind to their specific antigens, which are internalized by endocytosis and proteolyzed. iv) The resulting peptides are presented in the context of HLA class II to cognate (with same antigen specificity) CD4+ T helper cells through their T-cell receptor (TCR). Afterward, T cells express CD40L and cytokines, which bind to specific receptors of B cells, promoting B-cell activation, class-switch recombination (CSR), and somatic hypermutation (SHM) of the immunoglobulin gene. CSR requires precise orchestration of signaling molecules, transcription factors, and adaptor molecules (14-3-3) that recognized specific switch (S) regions of the DNA located upstream of each constant gene segment. 14-3-3 molecules are recognized by activation-induced cytidine deaminase (AID) that promotes DNA cleavage, which is followed by NHEJ that links the rearranged VDJ exon to the selected constant *IGH* gene segment. As a result, this process changes the isotype expressed by the B cells from IgM and IgD to IgA, IgG, or IgE. AID also plays a role in SHM by introducing single nucleotide substitutions in the V(D)J exon, which increases the diversity of immunoglobulins and allows positive selection of B cells with increased antigen-binding affinity. These steps result in the transcription of the rearranged genes into immunoglobulin molecules that can be secreted by B cell, which are called antibodies, with different effector functions and higher binding affinity for a specific antigen. This figure was made with biorender (https://biorender.com/).

METHODS

Study population

DNA was isolated from peripheral blood of 227 unrelated endemic PF patients and 193 unrelated controls without a history of any autoimmune disease (Table 1). To minimize possible population stratification, we included only individuals of predominantly European ancestry. We excluded individuals self-declared as non-Euro-descendants and those who reported a family history of miscegenation with non-European. The study participants were contacted at Hospital Adventista do Pênfigo (Campo Grande, Mato Grosso do Sul), Hospital das Clínicas da Faculdade de Medicina da USP (Ribeirão Preto, São Paulo), Lar da Caridade - Hospital do Fogo Selvagem (Uberaba, Minas Gerais), Hospital de Clínicas da UFPR, Hospital de Dermatologia Sanitária São Roque and Hospital Santa Casa de Misericórdia (Curitiba, Paraná). The diagnosis was based on clinical characterization by specialized dermatologists, immunological tests, histopathology, and immunohistochemistry of skin biopsies. All individuals voluntarily agreed to participate in this study and signed informed consent, according to the Declaration of Helsinki. This study was performed under Brazilian federal laws and approved by the Human Research Ethics Committee of the Federal University of Paraná under protocol number CAAE 02727412.4.0000.0096.

TABLE 1: CHARACTERIZATION OF PATIENTS AND CONTROLS

	Median age (years)	Sex	% of SSA	% of AMER	% of EUR
Patients	40.9 (6 – 83)	53% Female	15	12	73
Controls	44.8 (11 – 86)	52% Female	16	15	69

Percentages are the estimated proportion of Sub-Saharan African (SSA), Amerindian (AMER) and European (EUR) ancestry of the PF patients and control samples, based on the analysis of 71 ancestry-informative markers and using the HGDP-CEPH Database (CANN, 2002) as reference.

Selection of candidate genes and genotyping

We selected 167 genes knowingly implicated in antibody production, including those encoding Ig (heavy, kappa, and lambda chains). To select the candidate genes, we performed an extensive search in the literature databases (Google Scholar and PubMed) for review articles published in the last five years, using the terms “V(D)J rearrangement”, “class-switch recombination” (or “CSR”) and “somatic hypermutation”. We comprehensively searched all the references cited by the retrieved articles. The full list of candidate genes is given in Suppl. Table S1. Genotyping was performed by SNP microarrays using Infinium™ CoreExome-24 v1.1 BeadChip (Illumina, San Diego, USA).

Population structure analysis

Even though we carefully matched patients and controls for ancestry, the Brazilian population is admixed. To account for the possibility of population structure, we included another level of rigor by analyzing a panel of 71 previously validated (Soundararajan et al. 2016; Lao et al. 2006; Kosoy et al. 2009; Mychaleckyj et al. 2017) ancestry-informative markers (AIMs) (Suppl. Table S2). The allelic frequencies of these SNPs differ across the major continental population groups ($F_{ST} > 0.25$, $\delta > 0.40$), and the genotypes of HGDP-CEPH (Human Genome Diversity Project - Centre d'Etude du Polymorphisme Humain) populations are publicly available (Cann 2002). Pairwise F_{ST} was calculated using Arlequin v3.5.2 software (Excoffier and Lischer 2010) and δ was considered the difference between the frequencies of pairs of populations. We compared the study population to the HGDP-CEPH populations most closely related to the three major ancestral groups of the Brazilians (Salzano 2014): Sub-Saharan Africans (n = 120) – Biaka Pygmy, Mbuti Pygmy, Mandenka,

Yoruba, San, Bantu; Amerindians (n = 83) – Surui, Karitiana, Pima, Piapoco and Curripaco; and Europeans (n = 118) – French, French Basque, North Italian, Orcadian, Sardinian, Tuscan. For estimation of individual and populational ancestry, we used the software STRUCTURE v.2.2 (Falush, Stephens, and Pritchard 2007, 2003; Pritchard, Stephens, and Donnelly 2000) with a run length of 100,000 burn-in and 500,000 Markov chain Monte Carlo (MCMC) replications, the admixture model, and independent allele frequency model.

Association analysis

We used PLINK v1.9 (Chang et al. 2015) for all manipulation of SNP data. We extracted a total of 3,336 SNPs located within 3 Kbp upstream and downstream of each one of the 167 genes. From the total SNPs initially retrieved, we removed variants as follows: i) whose genotypes deviated from Hardy-Weinberg equilibrium in controls ($p < 0.05$); ii) or in strong linkage disequilibrium (LD) with any other variant ($r^2 \geq 0.80$). We established MAF ≥ 0.20 (minor allele frequency) to reach at least 80% power with a one-sided type I error rate $\alpha = 0.05$ to detect small to moderate effect sizes ($0.38 < d < 0.50$) (Gauderman 2002b, 2002a). We established the significance threshold as $p < 0.005$. After quality control and removal of redundant or non-informative SNPs, a total of 607 variants in 149 genes remained for association analysis (Suppl. Table S3). For association analysis, we performed logistic regression for additive, dominant and recessive models using sex and two principal components as covariates.

***In silico* analysis**

We used the online tool HaploReg (Ward and Kellis 2012) to evaluate if the associated SNPs could be implicated in structural and regulatory effects. To search for eQTL (expression quantitative trait loci) for variants that were associated with PF or in LD with them (r2 > 0.8), we used the tool Qtlizer (Munz et al. 2020), which compiles information from several databases (Carithers et al. 2015; Grimaldi-Bensouda et al. 2017; Zeller et al. 2010; Garnier et al. 2013; Westra et al. 2013; Gamazon et al. 2010; Ward and Kellis 2012; Franzén et al. 2016; Leslie, O'Donnell, and Johnson 2014; Buniello et al. 2019). We obtained splicing quantitative trait loci (sQTL) data from GTEx Portal (Carithers et al. 2015) and searched the transcript variants of the affected genes in Ensembl (Yates et al. 2019). We assessed LD between variants in the database

LDlink (Machiela and Chanock 2015) using European populations (CEU, TSI, FIN, GBR, IBS) of the 1000 Genomes project (1000 Genomes Project Consortium et al. 2015).

RESULTS

Lack of population stratification in patients and controls

Patients and controls were previously classified as predominantly Euro-descendants based on phenotypic characteristics and detailed assessment of family history. Here, our analysis using 71 AIMs showed that patients and controls were, in fact, homogenous regarding ancestry (p = 0.14). The proportions of Sub-Saharan African, Amerindian, and European compounds were 0.15, 0.12, and 0.73 for controls and 0.16, 0.15, and 0.69 for patients, respectively (Table 1, Figure 2).

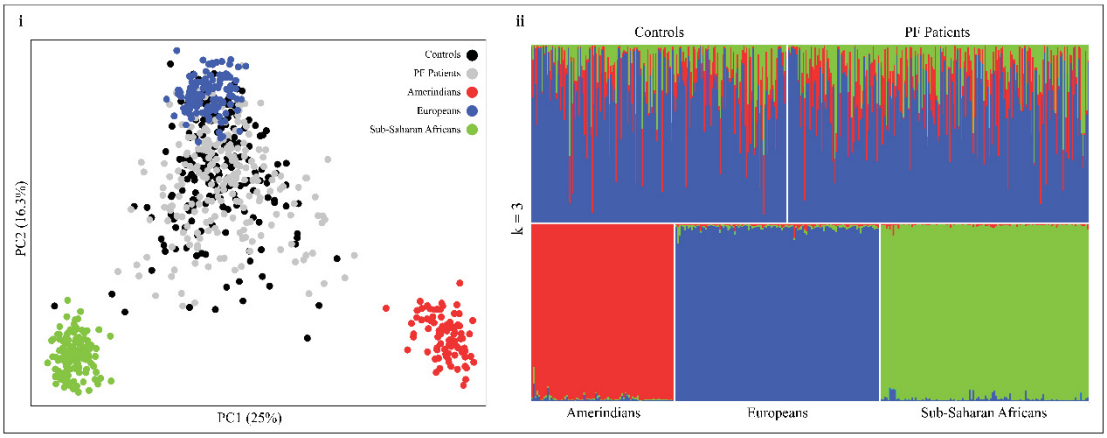


FIGURE 2: LACK OF POPULATION STRUCTURE IN THE STUDY POPULATION. (a) Principal component analysis and (b) Bar plot of inferred ancestry proportions performed with 71 ancestry-informative markers comparing PF patients and control samples with HGDP-CEPH samples from three regional populations: Sub-Saharan Africans (n = 120), Amerindians (n = 83), and Europeans (n = 118) (Cann 2002).

Genetic associations

We found 10 SNPs within 11 genes, or in their vicinity, related to B-cell development and antibody production associated with differential susceptibility to endemic PF either in the dominant (Dom), recessive (Rec), or additive (Add) models (Table 2). All variants associated with endemic PF were located in intronic or intergenic regions. One SNP is within a gene implicated in B-cell development, two in genes

involved in V(D)J rearrangement, and seven are close to or within genes that participate in antibody class-switch recombination.

Functional annotation

We performed a comprehensive in silico analysis using several publicly available databases and online tools. We found that all variants associated with endemic PF, or their proxy SNPs ($r^2 > 0.8$), had been previously associated with variable gene expression levels in different tissues, mainly in peripheral blood (Table 2) (Munz et al. 2020). Our in silico analysis also showed that these variants are located in sites containing chromatin regulatory histone marks and may be related to the regulation of promoters and enhancers in T and B cells (Ward and Kellis 2012) (Suppl. Table S4), as summarized below.

The intergenic variant *rs1818545*A* is in nearly absolute LD with nine other SNPs ($r^2 \approx 1$; $D' \approx 1$), including *rs7104753*C*, which is located in a predicted regulatory region of *RAG2* (*recombination-activating gene 2*) (Suppl. Table S4). We found three variants that have eQTL effects: the regulatory region variant *rs10781530*A*, located at 921 bp 5' of the *PAXX* (*PAXX non-homologous end-joining factor*) gene, which is involved in non-homologous end joining; *rs10870140*G* and *rs10781522*A*, two intronic variants of *TRAF2* (*TNF receptor-associated factor 2*), which have cis-eQTL effects on *TRAF2* in whole blood. Another SNP associated with endemic PF was the variant *rs2070729*C*, located in an intron of the gene *IRF1* (*interferon regulatory factor 1*). This variant has a cis-eQTL effect on *IRF1* and also a trans-eQTL effect on the gene *IL-13* (*interleukin-13*). We also observed that *rs6432018*C*, an intronic variant in the gene *YWHAQ* (*tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein, theta isoform*), is in high LD ($r^2 > 0.95$) with other 12 SNPs that have sQTL effects on *YWHAQ* (Table 2 and Suppl. Tables S5 and S6).

TABLE 2: VARIANTS ASSOCIATED WITH INCREASED RISK TO ENDEMIC PEMPHIGUS FOLIACEUS

Association analyses				Predicted eQTL of the associated SNPs and their proxy		
SNP **Genomic position***	**Location**	**Model**	**OR (95%CI)** **p-value**	**Affected gene**	**Cell, tissue or organ**	**effect** ***p*-value**
B-cell modulation						
rs6657275*G Chr1:218596461	intron of *TGFB2*	Rec	2.26 (1.33-3.84) $p = 2.6x10^{-3}$	***TGFB2***	Lung, testis, brain	eQTL<$8.2x10^{-3}$
				TGFB2-AS1	Whole blood, skeletal muscle	eQTL<$2.8x10^{-9}$
V(D)J rearrangement						
rs1818545*A Chr11:36612090	Intergenic between *RAG1*, *RAG2,* and *IFTAP*	Dom	1.85 (1.22-2.81) $p = 3.6 x10^{-3}$	***IFTAP***	Brain	eQTL<$7.1x10^{-3}$
rs10781530*A Chr9:139885948	921bp 5' of *PAXX*	Add	1.58 (1.16-2.15) $p = 3.6x10^{-3}$	***PAXX***	Whole blood, lung, heart, brain, artery, adipose	eQTL<$8.9x10^{-5}$
Class switch recombination and somatic hypermutation						
rs10870140*G Chr9:139796419	intron of *TRAF2*	Rec	1.76 (1.19-2.61) $p = 4.9x10^{-3}$	***TRAF2***	Whole blood	eQTL<$1.2x10^{-9}$
				PAXX	Brain	eQTL<$6.5 x10^{-6}$
rs10781522*A Chr9:139815053	intron of *TRAF2*	Add	1.61 (1.22-2.14) $p = 9x10^{-3}$	***TRAF2***	Whole blood, transverse colon, testis, lymphoblastoid cell, monocytes, skeletal muscle, skin, tibial nerve, dendritic cells	eQTL<$3.5 x10^{-4}$
		Rec	1.99 (1.30-3.05) $p = 1.5x10^{-3}$	***PAXX***	Blood, Brain	eQTL<$1.2 x10^{-3}$
rs535068*A Chr1:12189561	intron of *TNFRSF8*	Dom	3.11 (1.46-6.62) $p = 3.2x10^{-3}$	***TNFRSF1B***	Whole blood, skin, brain	eQTL<$3.1 x10^{-3}$
rs324011*A Chr12:57502182	intron of *STAT6*	Add	1.56 (1.16-2.10) $p = 3.3x10^{-3}$	***STAT6***	Whole blood, CD4+ lymphocytes, monocytes, brain, liver, colon sigmoid	eQTL<$1.4 x10^{-4}$
		Dom	1.94 (1.30-2.90) $p = 1.3x10^{-3}$			
rs2070729*C Chr5:131819921	intron of *IRF1*	Add	1.56 (1.17-2.09) $p = 2.8x10^{-3}$	***IRF1***	Whole blood, heart, monocytes	eQTL<$4.7 x10^{-4}$
		Rec	2.02 (1.25-3.27) $p = 4.4x10^{-3}$	***IRF1-AS1***	Whole blood, thyroid, spleen, skin, tibial nerve, Skeletal muscle, lung, heart, esophagus, colon, brain, artery, adipose	eQTL<$1.1 x10^{-4}$
				IL-13	Tibial nerve	eQTL<$4.2 x10^{-5}$
rs6432018*C Chr2:9721896	2.2kb 3' of *YWHAQ*	Add	1.69 (1.28-2.24) $p = 3x10^{-3}$	***YWHAQ***	Tibial artery	sQTL<$1.3 x10^{-6}$
		Dom	2.04 (1.32-3.14) $p = 1.2 x10^{-3}$			
rs17149161*C Chr7:75978229	intron of *YWHAG*	Add	1.63 (1.20-2.21) $p = 1.7 x10^{-3}$	***YWHAG***	Monocytes, adipose, lung	eQTL<$1.6 x10^{-4}$

SNP: rs ID of the single nucleotide polymorphism, Chr: chromosome, OR: odds ratio, CI: confidence interval. eQTL: expression quantitative trait loci. *Genomic position according to GRCh37.p13 primary assembly. The frequency of the associated alleles in patients and controls are given in supplementary table S3. The complete list of proxy SNPs is available in Supplementary Table S5.

DISCUSSION

Break of immune homeostasis and self-tolerance in some autoimmune disorders are directly related to the production of autoantibodies by B cells that differentiate into plasma cells (Tsubata 2017). In our study, we focused on genetic variants within genes related to B-cell development, activation, and maturation, and also immunoglobulin gene rearrangement, class-switch recombination, and somatic hypermutation. We aimed to contribute to a better understanding of these mechanisms in the context of endemic PF.

Our analysis with 71 previously validated ancestral-informative markers confirmed no ancestry bias between patients and controls from our study. Their conspicuous differences in genotypic frequencies among Europeans, Africans, and Amerindians indicate that these markers constitute a robust set to assess the continental ancestry of the study population. Nevertheless, we used two principal components as covariates in our logistic regression model to correct for possible minor differences that could contribute to spurious associations.

We used the *p*-value of 0.005 as a cut-off to identify genetic associations, as suggested by others (Di Leo and Sardanelli 2020; Ioannidis 2018; Benjamin et al. 2018), to achieve a low risk of false associations while not excluding possible real associations with smaller effects. Although the use of arbitrary cut-offs always bring the risk of type I error, we have taken several precautions to reduce the chances of false discoveries: a) we only analyzed variants for which our sample size allowed statistical power of at least 80% to detect low and moderate effects (MAF ≥ 0.20); b) we quantified the individual ancestry compound of each individual and adjusted our analysis using principal components as covariates even though our samples were carefully matched; c) we adjusted our analysis for sex and age. As we discuss in detail below, all our associations are provided with a plausible biological explanation, increasing the confidence of our results. Nevertheless, further replication in pemphigus or other autoimmune disease cohorts would be desirable to corroborate our findings.

We found 10 variants located within or in the vicinity of 11 genes related to B-cell development (*TGFB2*), V(D)J recombination (*RAG1*, *RAG2*, *PAXX*), B-cell activation, somatic hypermutation, and class-switch recombination (*TNFRSF1B*, *TRAF2*, *STAT6*, *IL-13*, *IRF1*, *YWHAG*, *YWHAQ*). The allele *rs6657275*G*, associated with increased risk to endemic PF, is located in the third intron of *TGFB2* (*transforming growth factor 2*) and has been previously associated with susceptibility to cerebral

malaria (Sambo et al. 2010). This SNP is in strong LD with other 39 variants, several of which have been predicted to have eQTL effects in different tissues on the gene *TGFB2* and its antisense long non-coding RNA (lncRNA) gene *TGFB2-AS1*. Some of the SNPs in LD with *rs6657275* are possibly related to regulatory histone marks H3K4me3 and H3K4me1 in T and B cells (Suppl. Table S4), which suggest that they may be involved in transcriptional regulation of *TGFB2*, and/or of the lncRNA genes *TGFB2-AS1* and *TGFB2-OT1*, which physically overlap (antisense gene overlap) the *TGFB2* gene. Interestingly, the lncRNA *TGFB2-AS1* has been associated with regulatory functions on TGF-β (product of *TGFB2* gene) and bone morphogenetic protein (BMP) signaling on keratinocytes (Papoutsoglou et al. 2019). The TGF-β and BMP also regulate the differentiation of hematopoietic stem cells (HSC) to the myeloid lineage or the lymphoid lineage, respectively (Naka and Hirao 2017). Additionally, TGF-β is secreted by B cells and may regulate B-cell proliferation (Kehrl et al. 1991). Dysregulation of TGF-β pathways is known to be implicated in antibody-mediated autoimmune disorders (Naka and Hirao 2017).

Two SNPs associated with endemic PF are located in genes whose products participate in the DNA cleavage and joining phases of the V(D)J recombination process: *rs1818545* (*RAG2*) and *rs10781530* (*PAXX*), respectively. The allele *rs1818545*A* has been previously associated with radiation-induced pneumonitis in lung cancer patients (Zhao et al. 2016). The impact of *rs1818545* on *RAG1* and *RAG2* expression can hardly be detected by the approaches applied for the investigation of gene expression because these genes are only expressed in developing lymphocytes (Oettinger et al. 1990), which were not included in the published studies. However, this variant and their proxy SNPs have cis-eQTL effects on the *IFTAP* (*intraflagellar transport-associated protein*) gene, whose expression is inversely correlated with *RAG1* and *RAG2* expression (Laszkiewicz et al. 2012). Moreover, *rs12283331* is in absolute LD ($r^2 = 1$; D' = 1) with *rs1818545* and is located in a predicted binding motif for the transcription factor Ik-2 in primary hematopoietic stem cells (Ward and Kellis 2012). Ik-2 promotes *RAG1* and *RAG2* transcription and downregulates *IFTAP* expression (Agnieszka et al. 2014). RAG1 is responsible for recognizing specific conserved recombination signal sequences adjacent to gene segments V, D, and J of immunoglobulins, responsible for the specificity of the double-strand cleavage of the DNA. RAG2, on the other hand, is necessary for the catalytic activity of RAG1 (Akamatsu and Oettinger 1998). Both RAG1 and RAG2 are part of the RAG complex,

which also mediates allelic exclusion of Ig, ensuring that each B cell expresses only one allele of the Ig genes (Vettermann and Schlissel 2010). Therefore, the association of *rs1818545*A* with increased risk to endemic PF could be related to the RAG1 and RAG2 role in the V(D)J regulation.

The variant *rs10781530*A* is located upstream of *PAXX* and is associated with a higher expression of this gene. PAXX molecules stabilize the enzymatic complex composed by XRCC4, XRCC6 (Ku70), XRCC5 (Ku80), DNA-PKcs, DNA ligase 4, and NHEJ1, which is required for the non-homologous end-joining pathway in V(D)J recombination (Kumar, Alt, and Frock 2016). This complex repairs the double-strand break in the first step of the rearrangement process. The two SNPs *rs10870140* and *rs10781522*, located within intronic regions of the gene *TRAF2* (73 kbp of the gene *PAXX*), are associated with endemic PF and have trans-eQTL effects on *PAXX.* Therefore, it is plausible that the associations observed for these SNPs are related to the DNA repair process during V(D)J rearrangement.

After antigen recognition, B-cell activation may be mediated by T cells. The variant *rs2070729*, located in the intron 9 of the *IRF1* gene, together with five other SNPs in strong LD with it, have eQTL effects on *IRF1* and *IL-13* gene expression. IRF1 is a transcription factor that promotes the expression of CIITA (class II major histocompatibility complex transactivator) molecules (Morris et al. 2002), a critical regulator of HLA class II expression in antigen-presenting cells, including B cells. HLA class II molecules are necessary for antigen presentation to T cells, T-cell activation, and consequently, activation of B cells with the same antigen specificity (Roche and Furuta 2015). Interestingly, a combination of certain *HLA-DRB1* alleles with a promoter variant in *CIITA* was strongly associated with endemic PF (OR = 14.05, $p < 10^{-6}$) (Piovezan and Petzl-Erler 2013).

T-cell activation also stimulates the production of TNF (tumor necrosis factor) by B cells, which activates the TNF receptor superfamily member 1B (TNFRSF1B or CD120b) in B cells. Activated TNFRSF1B interacts with TNFR-associated factor 2 (TRAF2), triggering the secretion of IgM (Hostager and Bishop 2002). Furthermore, Ig class switch is also regulated by another receptor on the surface of B cells, TNF receptor superfamily member 8 (TNFRSF8 or CD30), which prevents the class switching in B cells that are non-antigen- specific (Cerutti et al. 1998). Associated with differential endemic PF risk was the intronic SNP *rs535068* located in *TNFRSF8*. This SNP and other ten in strong LD with it ($r^2 > 0.8$) are located in a regulatory region of

TNFRSF8. This region, which appears to be relevant for the regulation of T and B cell lineages, includes enhancer and promoter segments that contain epigenetic chromatin marks such as H3K4me1, H3K4me3, H3K27ac, and H3K9ac (Suppl. Table S4). The SNP *rs535068* and six of the SNPs in strong LD with it also have trans-eQTL effects on the gene *TNFRSF1B* in blood and skin. Two other SNPs associated with endemic PF, *rs10870140* and *rs10781522*, were located in intronic regions of *TRAF2*, whose product also plays a critical role in B-cell activation. These variants have not only cis-eQTL effect on *TRAF2* but are also associated with differential expression of the gene *PAXX*, as mentioned above. The interaction between CD40 receptor and TRAF2 is essential for CD40-regulated class switch of Ig from IgM to IgG and B-cell activation (Jabara et al. 2002; Safran et al. 2010). Defects in these pathways have been suggested to cause the generation of pathogenic autoantibodies (Cerutti et al. 1998). The class switch to IgG and IgE in B cells activated by IL-4 and IL-13 is also mediated by the transcription factor STAT6 (signal transducer and activator of transcription 6) (Linehan et al. 1998; Turqueti-Neves et al. 2014). STAT6 acts as a regulator of several genes in IL-4- stimulated B cells (Schroder et al. 2002), which includes the *AICDA* (Zan and Casali 2013). The *AICDA* gene encodes the AID molecule (activation-induced cytidine deaminase), which plays a pivotal role during class-switch recombination by generating a double-strand break of the DNA (Staszewski et al. 2011). Moreover, AID action produces point mutations during somatic hypermutation (Liu et al. 2008). The intronic SNP *rs324011* in *STAT6*, also associated with endemic PF, is in LD with six other SNPs (Suppl. Table S5). All these seven variants have eQTL effects on *STAT6* expression in whole blood and other tissues. Additionally, four of them are located within enhancer or promoter regions and are predicted to participate in epigenetic chromatin modifications in several tissues, including T and B cells. Therefore, in the context of immunoglobulin gene rearrangement, it is plausible to suggest the association of *rs324011*A* with endemic PF may be related to differential expression of *STAT6*, which consequently affects the expression of *AICDA*.

The CSR process requires the precise recognition of the DNA cleavage sites on the switch (S) regions (Dunnick et al. 1993). The adaptor molecules 14-3-3 recognize these regions and function as scaffolds of the CSR machinery, interacting and stabilizing AID and other molecules, such as PKA-Ca (cAMP-dependent protein kinase catalytic subunit alpha) and Ung (uracil-DNA glycosylase) (Stavnezer and Schrader 2014). SNPs within two genes encoding 14-3-3 molecules (*YWHAG* and *YWHAQ*)

were associated with differential susceptibility to endemic PF. The variant *rs17149161*C* has an eQTL effect on *YWHAG*, and *rs6432018*C* has an sQTL effect on *YWHAQ*. sQTL are genetic variants that can change the splicing ratios of gene transcripts (Monlong et al. 2014). Interestingly, of the five transcript variants that have been described for *YWHAQ*, two of them do not encode the complete protein due to alternative splicing (Suppl. Table S6).

One limitation of our study is that some critical variants implicated in the disease may have been excluded due to three main reasons: i) some relevant variants may not have been genotyped in the microarray. A large number of immune-related genes, including immunoglobulin genes and others related to B-cell function, are overall poorly covered in microarray chips due to homology and structural variation that impose technical limitations for genotyping. ii) Other variants may have been removed from the analysis due to their low frequency in our cohort. iii) Some relevant genes for B-cell development are unknown or were missed. Therefore, some variants affecting PF pathogenesis may not have been uncovered by our study. However, we were able to screen over 3,000 SNPs and select the 607 most informative to our association analysis, which constitutes an unprecedented and comprehensive analysis of B-cell-related variants in this autoimmune disease. Our observations provide evidence that variation in B-cell-related genes has a pivotal effect on PF risk. Besides, we presented a comprehensive in silico analysis suggesting that regulation of molecules involved in B-cell activation, immunoglobulin isotype switching, and hypermutation may explain the associations that we observed for endemic PF. Therefore, our results justify further in-depth analysis of genes that were not well-covered in our study, applying different technologies that allow high-resolution characterization.

In summary, we carefully explored the variation in genes implicated in B-cell development and function in the context of the autoimmune B-cell-mediated nature of endemic PF. The associations that we observed can be explained by possible effects on the regulation of expression levels of molecules involved in the complex process of B-cell modulation, DNA sequence recognition, DNA cleavage and joining, and somatic hypermutation. For the first time, we show that polymorphisms in genes involved in autoantibody production might confer differential susceptibility to this disease. Therefore, we identified candidate genes for possible high-resolution and functional studies to understand Ig production and its impact on the etiology of endemic PF and other diseases.

AUTHOR CONTRIBUTIONS

VCS, DM and DA designed the study. DA performed microarray genotyping. VCS, LMA, TDJF analyzed the data. MLPE, DA and DM contributed with reagents. VCS and DA drafted the manuscript. All authors significantly contributed with ideas and critically reviewed this manuscript.

ACKNOWLEDGMENTS

This work was supported by grants from the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Fundação Araucária de Apoio ao Desenvolvimento Científico e Tecnológico do Paraná, PRONEX (Convênio 116/2018 – Protocolo 50530), and Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES).

CONFLICT OF INTEREST

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

REFERENCES

1000 Genomes Project Consortium, Adam Auton, Lisa D. Brooks, Richard M. Durbin, Erik P. Garrison, Hyun Min Kang, Jan O. Korbel, et al. 2015. “A Global Reference for Human Genetic Variation.” Nature 526 (7571): 68–74. https://doi.org/10.1038/nature15393.

Agnieszka, Łaszkiewicz, Bzdzion Łukasz, Kasztura Monika, Śniezewski Łukasz, Janik Sylwia, Kisielow Paweł, and Cebrat Małgorzata. 2014. “Ikaros and RAG-2-Mediated Antisense Transcription Are Responsible for Lymphocyte-Specific Inactivation of NWC Promoter.” PLoS ONE 9 (9). https://doi.org/10.1371/journal.pone.0106927.

Akamatsu, Yoshiko, and Marjorie A. Oettinger. 1998. “Distinct Roles of RAG1 and RAG2 in Binding the V(D)J Recombination Signal Sequences.” Molecular and Cellular Biology 18 (8): 4670–78. https://doi.org/10.1128/MCB.18.8.4670.

Amagai, Masayuki, and John R. Stanley. 2012. “Desmoglein as a Target in Skin Disease and Beyond.” Journal of Investigative Dermatology 132 (3 PART 2): 776–84. https://doi.org/10.1038/jid.2011.390.

Augusto, Danillo G., Sara C. Lobo-Alves, Marcia F. Melo, Noemi F. Pereira, and Maria Luiza Petzl-Erler. 2012. “Activating KIR and HLA Bw4 Ligands Are Associated to Decreased Susceptibility to Pemphigus Foliaceus, an Autoimmune Blistering Skin Disease.” PLoS ONE 7 (7): e39991. https://doi.org/10.1371/journal.pone.0039991.

Augusto, Danillo G., Geraldine M. O’Connor, Sara C. Lobo-Alves, Sara Bass, Maureen P. Martin, Mary Carrington, Daniel W. McVicar, and Maria Luiza Petzl-Erler. 2015. “Pemphigus

Is Associated with KIR3DL2 Expression Levels and Provides Evidence That KIR3DL2 May Bind HLA-A3 and A11 in Vivo." European Journal of Immunology 45 (7): 2052–60. https://doi.org/10.1002/eji.201445324.

Bassing, Craig H, Wojciech Swat, and Frederick W Alt. 2002. "The Mechanism and Regulation of Chromosomal V(D)J Recombination." Cell 109 (2): S45–55. https://doi.org/10.1016/S0092-8674(02)00675-X.

Bastuji-Garin, Sylvie, Rafika Souissi, Laurent Blum, Hamida Turki, Rafia Nouira, Bechir Jomaa, Abdelmajid Zahaf, et al. 1995. "Comparative Epidemiology of Pemphigus in Tunisia and France: Unusual Incidence of Pemphigus Foliaceus in Young Tunisian Women." Journal of Investigative Dermatology 104 (2): 302–5. https://doi.org/10.1111/1523-1747.ep12612836.

Benjamin, Daniel J., James O. Berger, Magnus Johannesson, Brian A. Nosek, E. J. Wagenmakers, Richard Berk, Kenneth A. Bollen, et al. 2018. "Redefine Statistical Significance." Nature Human Behaviour 2 (1): 6–10. https://doi.org/10.1038/s41562-017-0189-z.

Brochado, Maria José Franco, Daniela Francisca Nascimento, Wagner Campos, Neifi Hassan Saloum Deghaide, Eduardo Antonio Donadi, and Ana Maria Roselino. 2016. "Differential HLA Class I and Class II Associations in Pemphigus Foliaceus and Pemphigus Vulgaris Patients from a Prevalent Southeastern Brazilian Region." Journal of Autoimmunity 72 (August): 19–24. https://doi.org/10.1016/j.jaut.2016.04.007.

Bumiller-Bini, V., G.A. Cipolla, R.C. de Almeida, M.L. Petzl-Erler, D.G. Augusto, and A.B.W. Boldt. 2018. "Sparking Fire under the Skin? Answers from the Association of Complement Genes with Pemphigus Foliaceus." Frontiers in Immunology 9 (APR). https://doi.org/10.3389/fimmu.2018.00695.

Bumiller-Bini, Valéria, Gabriel Adelman Cipolla, Mariana Basso Spadoni, Danillo Gardenal Augusto, Maria Luiza Petzl-Erler, Marcia Holsbach Beltrame, and Angelica Beate Winter Boldt. 2019. "Condemned or Not to Die? Gene Polymorphisms Associated With Cell Death in Pemphigus Foliaceus." Frontiers in Immunology 10 (October). https://doi.org/10.3389/fimmu.2019.02416.

Buniello, Annalisa, Jacqueline A.L. Macarthur, Maria Cerezo, Laura W. Harris, James Hayhurst, Cinzia Malangone, Aoife McMahon, et al. 2019. "The NHGRI-EBI GWAS Catalog of Published Genome-Wide Association Studies, Targeted Arrays and Summary Statistics 2019." Nucleic Acids Research 47 (D1): D1005–12. https://doi.org/10.1093/nar/gky1120.

Cann, H. M. 2002. "A Human Genome Diversity Cell Line Panel." Science 296 (5566): 261b – 262. https://doi.org/10.1126/science.296.5566.261b.

Carithers, Latarsha J., Kristin Ardlie, Mary Barcus, Philip A. Branton, Angela Britton, Stephen A. Buia, Carolyn C. Compton, et al. 2015. "A Novel Approach to High-Quality Postmortem Tissue Procurement: The GTEx Project." Biopreservation and Biobanking 13 (5): 311–19. https://doi.org/10.1089/bio.2015.0032.

Cerutti, Andrea, Andrés Schaffer, Shefali Shah, Hong Zan, Hsiou Chi Liou, Raymond G. Goodwin, and Paolo Casali. 1998. "CD30 Is a CD40-Inducible Molecule That Negatively Regulates CD40- Mediated Immunoglobulin Class Switching in Non-Antigen-Selected Human B Cells." Immunity 9 (2): 247–56. https://doi.org/10.1016/S1074-7613(00)80607-X.

Chang, Christopher C., Carson C. Chow, Laurent C.A.M. Tellier, Shashaank Vattikuti, Shaun M. Purcell, and James J. Lee. 2015. "Second-Generation PLINK: Rising to the Challenge of Larger and Richer Datasets." GigaScience 4 (1): 7. https://doi.org/10.1186/s13742-015-0047-8.

Cipolla, Gabriel A., Jong Kook Park, Liana A. de Oliveira, Sara Cristina Lobo-Alves, Rodrigo C. de Almeida, Ticiana D.J. Farias, Débora de S. Lemos, Danielle Malheiros, Robert M. Lavker, and Maria Luiza Petzl-Erler. 2016. "A 3′UTR Polymorphism Marks Differential KLRG1 MRNA

Levels through Disruption of a MiR-584-5p Binding Site and Associates with Pemphigus Foliaceus Susceptibility." Biochimica et Biophysica Acta - Gene Regulatory Mechanisms 1859 (10): 1306–13. https://doi.org/10.1016/j.bbagrm.2016.07.006.

Diaz, L a, S a Sampaio, E a Rivitti, C R Martins, P R Cunha, C Lombardi, F a Almeida, R M Castro, M L Macca, and C Lavrado. 1989. "Endemic Pemphigus Foliaceus (Fogo Selvagem): II. Current and Historic Epidemiologic Studies." The Journal of Investigative Dermatology 92: 4–12. https://doi.org/10.1111/1523-1747.ep13070394.

Dudley, Darryll D, Jayanta Chaudhuri, Craig H Bassing, and Frederick W Alt. 2005. "Mechanism and Control of V(D)J Recombination versus Class Switch Recombination: Similarities and Differences." In Advances in Immunology, 86:43–112. https://doi.org/10.1016/S0065-2776(04)86002-4.

Dunnick, Wesley, Gerald Z. Hertz, Lori Scappino, and Christine Gritzmacher. 1993. "DNA Sequences at Immunoglobulin Switch Region Recombination Sites." Nucleic Acids Research 21 (3): 365–72. https://doi.org/10.1093/nar/21.3.365.

Excoffier, Laurent, and Heidi E L Lischer. 2010. "Arlequin Suite Ver 3.5: A New Series of Programs to Perform Population Genetics Analyses under Linux and Windows." Molecular Ecology Resources 10 (3): 564–67. https://doi.org/10.1111/j.1755-0998.2010.02847.x.

Falush, Daniel, Matthew Stephens, and Jonathan K. Pritchard. 2007. "Inference of Population Structure Using Multilocus Genotype Data: Dominant Markers and Null Alleles." Molecular Ecology Notes 7 (4): 574–78. https://doi.org/10.1111/j.1471-8286.2007.01758.x.

Falush, Daniel, Matthew Stephens, and Jonathan K Pritchard. 2003. "Inference of Population Structure Using Multilocus Genotype Data: Linked Loci and Correlated Allele Frequencies." Genetics 164 (4): 1567–87. https://doi.org/10.1001/jama.1987.03400040069013.

Farias, Ticiana Della Justina, Danillo G Augusto, Rodrigo Coutinho de Almeida, Danielle Malheiros, and Maria Luiza Petzl-Erler. 2019. "Screening the Full Leucocyte Receptor Complex Genomic Region Revealed Associations with Pemphigus That Might Be Explained by Gene Regulation." Immunology 156 (1): 86–93. https://doi.org/10.1111/imm.13003.

Franzén, Oscar, Raili Ermel, Ariella Cohain, Nicholas K. Akers, Antonio Di Narzo, Husain A. Talukdar, Hassan Foroughi-Asl, et al. 2016. "Cardiometabolic Risk Loci Share Downstream Cis- and Trans-Gene Regulation across Tissues and Diseases." Science 353 (6301): 827–30. https://doi.org/10.1126/science.aad6970.

Gamazon, Eric R., Wei Zhang, Anuar Konkashbaev, Shiwei Duan, Emily O. Kistner, Dan L. Nicolae, M. Eileen Dolan, and Nancy J. Cox. 2010. "SCAN: SNP and Copy Number Annotation." Bioinformatics 26 (2): 259–62. https://doi.org/10.1093/bioinformatics/btp644.

Garnier, Sophie, Vinh Truong, Jessy Brocheton, Tanja Zeller, Maxime Rovital, Philipp S. Wild, Andreas Ziegler, et al. 2013. "Genome-Wide Haplotype Analysis of Cis Expression Quantitative Trait Loci in Monocytes." PLoS Genetics 9 (1): 1–11. https://doi.org/10.1371/journal.pgen.1003240.

Gauderman, W. James. 2002a. "Sample Size Requirements for Association Studies of Gene-Gene Interaction." American Journal of Epidemiology 155 (5): 478–84. https://doi.org/10.1093/aje/155.5.478.

Grimaldi-Bensouda, Lamiae, Michel Rossignol, Isabelle Koné-Paut, Alain Krivitzky, Christine Lebrun-Frenay, Johanna Clet, David Brassat, et al. 2017. "Risk of Autoimmune Diseases and Human Papilloma Virus (HPV) Vaccines: Six Years of Case-Referent Surveillance." Journal of Autoimmunity 79: 84–90. https://doi.org/10.1016/j.jaut.2017.01.005.

Honjo, T. 1983. "Immunoglobulin Genes." Edited by Tasuku Honjo and Frederick Alt. Annual Review of Immunology 1 (1): 499–528. https://doi.org/10.1146/annurev.iy.01.040183.002435.

Hostager, Bruce S., and Gail A. Bishop. 2002. "Role of TNF Receptor-Associated Factor 2 in

the Activation of IgM Secretion by CD40 and CD120b." The Journal of Immunology 168 (7): 3318–22. https://doi.org/10.4049/jimmunol.168.7.3318.

Ioannidis, John P.A. 2018. "The Proposal to Lower P Value Thresholds to .005." JAMA - Journal of the American Medical Association 319 (14): 1429–30. https://doi.org/10.1001/jama.2018.1536.

Jabara, Haifa H., Dhafer Laouini, Erdyni Tsitsikov, Emiko Mizoguchi, Atul K. Bhan, Emanuela Castigli, Fatma Dedeoglu, Vadim Pivniouk, Scott R. Brodeur, and Raif S. Geha. 2002. "The Binding Site for TRAF2 and TRAF3 but Not for TRAF6 Is Essential for CD40-Mediated Immunoglobulin Class Switching." Immunity 17 (3): 265–76. https://doi.org/10.1016/S1074-7613(02)00394-1.

Jacobs, Heinz, and Linda Bross. 2001. "Towards an Understanding of Somatic Hypermutation." Current Opinion in Immunology 13 (2): 208–18. https://doi.org/10.1016/s0952-7915(00)00206-5.

Joly, Pascal, and Noémie Litrowski. 2011. "Pemphigus Group (Vulgaris, Vegetans, Foliaceus, Herpetiformis, Brasiliensis)." Clinics in Dermatology 29 (4): 432–36. https://doi.org/10.1016/j.clindermatol.2011.01.013.

Jung, David, Cosmas Giallourakis, Raul Mostoslavsky, and Frederick W Alt. 2006. "Mechanism and Control of V(D)J Recombination at the Immunoglobulin Heavy Chain Locus." Annual Review of Immunology 24 (1): 541–70. https://doi.org/10.1146/annurev.immunol.23.021704.115830.

Kasperkiewicz, Michael, Christoph T Ellebrecht, Hayato Takahashi, Jun Yamagami, Detlef Zillikens, Aimee S Payne, and Masayuki Amagai. 2017. "Pemphigus." Nature Reviews Disease Primers 3 (1): 17026. https://doi.org/10.1038/nrdp.2017.26.

Kehrl, John H., Alan Taylor, Seong-Jin Kim, and Anthony S. Fauci. 1991. "Transforming Growth Factor-Beta Is a Potent Negative Regulator of Human Lymphocytes." Annals of the New York Academy of Sciences 628 (1 Negative Regu): 345–53. https://doi.org/10.1111/j.1749-6632.1991.tb17267.x.

Kosoy, Roman, Rami Nassir, Chao Tian, Phoebe A. White, Lesley M. Butler, Gabriel Silva, Rick Kittles, et al. 2009. "Ancestry Informative Marker Sets for Determining Continental Origin and Admixture Proportions in Common Populations in America." Human Mutation 30 (1): 69–78. https://doi.org/10.1002/humu.20822.

Kumar, KidangazhiyathmanaAjith. 2008. "Incidence of Pemphigus in Thrissur District, South India." Indian Journal of Dermatology, Venereology and Leprology 74 (4): 349. https://doi.org/10.4103/0378-6323.42901.

Kumar, Vipul, Frederick W. Alt, and Richard L. Frock. 2016. "PAXX and XLF DNA Repair Factors Are Functionally Redundant in Joining DNA Breaks in a G1-Arrested Progenitor B-Cell Line." Proceedings of the National Academy of Sciences of the United States of America 113 (38): 10619–24. https://doi.org/10.1073/pnas.1611882113.

Lao, Oscar, Kate Van Duijn, Paula Kersbergen, Peter De Knijff, and Manfred Kayser. 2006. "Proportioning Whole-Genome Single-Nucleotide–Polymorphism Diversity for the Identification of Geographic Population Structure and Genetic Ancestry." The American Journal of Human Genetics 78 (4): 680–90. https://doi.org/10.1086/501531.

Laszkiewicz, Agnieszka, Lukasz Sniezewski, Monika Kasztura, Lukasz Bzdzion, Malgorzata Cebrat, and Pawel Kisielow. 2012. "Bidirectional Activity of the NWC Promoter Is Responsible for RAG-2 Transcription in Non-Lymphoid Cells." PLoS ONE 7 (9): 1–9. https://doi.org/10.1371/journal.pone.0044807.

Leo, Giovanni Di, and Francesco Sardanelli. 2020. "Statistical Significance: P Value, 0.05 Threshold, and Applications to Radiomics—Reasons for a Conservative Approach." European Radiology Experimental 4 (1). https://doi.org/10.1186/s41747-020-0145-y.

Leslie, Richard, Christopher J. O'Donnell, and Andrew D. Johnson. 2014. "GRASP: Analysis of Genotype-Phenotype Results from 1390 Genome-Wide Association Studies and Corresponding Open Access Database." Bioinformatics 30 (12): i185–94. https://doi.org/10.1093/bioinformatics/btu273.

Linehan, Leslie A., Wendy D. Warren, Patricia A. Thompson, Michael J. Grusby, and Michael T. Berton. 1998. "STAT6 Is Required for IL-4-Induced Germline Ig Gene Transcription and Switch Recombination." Journal of Immunology 161 (1): 302–10.

Liu, Man, Jamie L. Duke, Daniel J. Richter, Carola G. Vinuesa, Christopher C. Goodnow, Steven H. Kleinstein, and David G. Schatz. 2008. "Two Levels of Protection for the B Cell Genome during Somatic Hypermutation." Nature 451 (7180): 841–45. https://doi.org/10.1038/nature06547.

Lobo-Alves, S.C., D.G. Augusto, W.C.S. Magalhães, E. Tarazona-Santos, M.F. Lima-Costa, M.L. Barreto, B.L. Horta, R.C. Almeida, and M.L. Petzl-Erler. 2019. "Long Noncoding <scp>RNA</Scp> Polymorphisms Influence Susceptibility to Endemic Pemphigus Foliaceus." British Journal of Dermatology 181 (2): 324–31. https://doi.org/10.1111/bjd.17640.

Lombardi, C, P C Borges, A Chaul, S A Sampaio, Evandro A. Rivitti, H Friedman, C R Martins, J A Sanches Júnior, P R Cunha, and Raymond G. Hoffmann. 1992. "Environmental Risk Factors in Endemic Pemphigus Foliaceus (Fogo Selvagem). 'The Cooperative Group on Fogo Selvagem Research'." The Journal of Investigative Dermatology 98 (6): 847–50. https://doi.org/10.1111/1523-1747.ep12456932.

Machiela, Mitchell J., and Stephen J. Chanock. 2015. "LDlink: A Web-Based Application for Exploring Population-Specific Haplotype Structure and Linking Correlated Alleles of Possible Functional Variants: Fig. 1." Bioinformatics 31 (21): 3555–57. https://doi.org/10.1093/bioinformatics/btv402.

Malheiros, D, and M L Petzl-erler. 2009. "Individual and Epistatic Effects of Genetic Polymorphisms of B-Cell Co-Stimulatory Molecules on Susceptibility to Pemphigus Foliaceus." Genes and Immunity 10 (6): 547–58. https://doi.org/10.1038/gene.2009.36.

Marazza, G., H. C. Pham, L. Schärer, P. P. Pedrazzetti, T. Hunziker, R. M. Trüeb, D. Hohl, et al. 2009. "Incidence of Bullous Pemphigoid and Pemphigus in Switzerland: A 2-Year Prospective Study." British Journal of Dermatology 161 (4): 861–68. https://doi.org/10.1111/j.1365-2133.2009.09300.x.

Monlong, Jean, Miquel Calvo, Pedro G. Ferreira, and Roderic Guigó. 2014. "Identification of Genetic Variants Associated with Alternative Splicing Using SQTLseekeR." Nature Communications 5 (May). https://doi.org/10.1038/ncomms5698.

Morris, Ann C., Guy W. Beresford, Myesha R. Mooney, and Jeremy M. Boss. 2002. "Kinetics of a Gamma Interferon Response: Expression and Assembly of CIITA Promoter IV and Inhibition by Methylation." Molecular and Cellular Biology 22 (13): 4781–91. https://doi.org/10.1128/mcb.22.13.4781-4791.2002.

Munz, Matthias, Inken Wohlers, Eric Simon, Tobias Reinberger, Hauke Busch, Arne S. Schaefer, and Jeanette Erdmann. 2020. "Qtlizer: Comprehensive QTL Annotation of GWAS Results." BioRxiv. https://doi.org/https://doi.org/10.1101/495903.

Mychaleckyj, Josyf C., Alexandre Havt, Uma Nayak, Relana Pinkerton, Emily Farber, Patrick Concannon, Aldo A. Lima, and Richard L. Guerrant. 2017. "Genome-Wide Analysis in Brazilians Reveals Highly Differentiated Native American Genome Regions." Molecular Biology and Evolution 34 (3): msw249. https://doi.org/10.1093/molbev/msw249.

Naka, Kazuhito, and Atsushi Hirao. 2017. "Regulation of Hematopoiesis and Hematological Disease by TGF-β Family Signaling Molecules." Cold Spring Harbor Perspectives in Biology 9 (9): a027987. https://doi.org/10.1101/cshperspect.a027987.

Oettinger, M., D. Schatz, Carolyn Gorka, and David Baltimore. 1990. "RAG-1 and RAG-2,

Adjacent Genes That Synergistically Activate V(D)J Recombination." Science 248 (4962): 1517–23. https://doi.org/10.1126/science.2360047.

Oliveira, Luana Caroline, Gabriela Canalli Kretzschmar, Andressa Cristina Moraes dos Santos, Carolina Maciel Camargo, Renato Mitsunori Nisihara, Ticiana Della Justina Farias, Andre Franke, et al. 2019. "Complement Receptor 1 (CR1, CD35) Polymorphisms and Soluble CR1: A Proposed Anti-Inflammatory Role to Quench the Fire of 'Fogo Selvagem' Pemphigus Foliaceus." Frontiers in Immunology 10 (November): 1–15. https://doi.org/10.3389/fimmu.2019.02585.

Papoutsoglou, Panagiotis, Yutaro Tsubakihara, Laia Caja, Anita Morén, Paris Pallis, Adam Ameur, Carl-Henrik Heldin, and Aristidis Moustakas. 2019. "The TGFB2-AS1 LncRNA Regulates TGF-β Signaling by Modulating Corepressor Activity." Cell Reports 28 (12): 3182-3198.e11. https://doi.org/10.1016/j.celrep.2019.08.028.

Pavoni, D P, V M M S Roxo, a Marquart Filho, and M L Petzl-Erler. 2003. "Dissecting the Associations of Endemic Pemphigus Foliaceus (Fogo Selvagem) with HLA-DRB1 Alleles and Genotypes." Genes and Immunity 4 (2): 110–16. https://doi.org/10.1038/sj.gene.6363939.

Petzl-Erler, M L, and J Santamaria. 1989. "Are HLA Class II Genes Controlling Susceptibility and Resistance to Brazilian Pemphigus Foliaceus (Fogo Selvagem)?" Tissue Antigens 33: 408–14.

Petzl-Erler, Maria Luiza. 2020. "Beyond the HLA Polymorphism: A Complex Pattern of Genetic Susceptibility to Pemphigus." Genetics and Molecular Biology 43 (3): 1–26. https://doi.org/10.1590/1678-4685-gmb-2019-0369.

Piovezan, Bruno Zagonel, and Maria Luiza Petzl-Erler. 2013. "Both Qualitative and Quantitative Genetic Variation of MHC Class II Molecules May Influence Susceptibility to Autoimmune Diseases: The Case of Endemic Pemphigus Foliaceus." Human Immunology 74 (9): 1134–40. https://doi.org/10.1016/j.humimm.2013.06.008.

Pritchard, Jonathan K., Matthew Stephens, and Peter Donnelly. 2000. "Inference of Population Structure Using Multilocus Genotype Data." Genetics 155 (2): 945–59. http://www.ncbi.nlm.nih.gov/pubmed/10835412.

Roche, Paul A., and Kazuyuki Furuta. 2015. "The Ins and Outs of MHC Class II-Mediated Antigen Processing and Presentation." Nature Reviews Immunology 15 (4): 203–16. https://doi.org/10.1038/nri3818.

Safran, M., I. Dalah, J. Alexander, N. Rosen, T. Iny Stein, M. Shmoish, N. Nativ, et al. 2010. "GeneCards Version 3: The Human Gene Integrator." Database 2010 (August): baq020–baq020. https://doi.org/10.1093/database/baq020.

Salviano-Silva, Amanda, Maria Luiza Petzl-Erler, and Angelica Beate Winter Boldt. 2017. "CD59 Polymorphisms Are Associated with Gene Expression and Different Sexual Susceptibility to Pemphigus Foliaceus." Autoimmunity 6934 (June): 1–9. https://doi.org/10.1080/08916934.2017.1329830.

Salzano, Francisco M. 2014. "Interethnic Variability and Admixture in Latin America - Social Implications." Revista de Biología Tropical 1 (2): 405. https://doi.org/10.15517/rbt.v1i2.15273.

Sambo, Maria Rosário, Maria Jesus Trovoada, Carla Benchimol, Vatúsia Quinhentos, Lígia Gonçalves, Rute Velosa, Maria Isabel Marques, et al. 2010. "Transforming Growth Factor Beta 2 and Heme Oxygenase 1 Genes Are Risk Factors for the Cerebral Malaria Syndrome in Angolan Children." PLoS ONE 5 (6). https://doi.org/10.1371/journal.pone.0011141.

Schroder, Andreas J., Paul Pavlidis, Akinori Arimura, Danielle Capece, and Paul B. Rothman. 2002. "Cutting Edge: STAT6 Serves as a Positive and Negative Regulator of Gene Expression in IL-4-Stimulated B Lymphocytes." The Journal of Immunology 168 (3): 996–1000. https://doi.org/10.4049/jimmunol.168.3.996.

Schroeder, Harry W., and Lisa Cavacini. 2010. “Structure and Function of Immunoglobulins.” Journal of Allergy and Clinical Immunology 125 (2): S41–52. https://doi.org/10.1016/j.jaci.2009.09.046.

Soundararajan, Usha, Libing Yun, Meisen Shi, and Kenneth K. Kidd. 2016. “Minimal SNP Overlap among Multiple Panels of Ancestry Informative Markers Argues for More International Collaboration.” Forensic Science International: Genetics 23: 25–32. https://doi.org/10.1016/j.fsigen.2016.01.013.

Spadoni, Mariana Basso, Valéria Bumiller-Bini, Maria Luiza Petzl-Erler, Danillo Gardenal Augusto, and Angelica Beate Winter Boldt. 2019. “First Glimpse of Epigenetic Effects on Pemphigus Foliaceus.” Journal of Investigative Dermatology, July. https://doi.org/10.1016/j.jid.2019.07.691.

Staszewski, Ori, Richard E. Baker, Anna J. Ucher, Raygene Martier, Janet Stavnezer, and Jeroen E.J. Guikema. 2011. “Activation-Induced Cytidine Deaminase Induces Reproducible DNA Breaks at Many Non-Ig Loci in Activated B Cells.” Molecular Cell 41 (2): 232–42. https://doi.org/10.1016/j.molcel.2011.01.007.

Stavnezer, Janet, and Carol E. Schrader. 2014. “IgH Chain Class Switch Recombination: Mechanism and Regulation.” The Journal of Immunology 193 (11): 5370–78. https://doi.org/10.4049/jimmunol.1401849.

Tsubata, Takeshi. 2017. “B-Cell Tolerance and Autoimmunity.” F1000Research 6 (March): 391. https://doi.org/10.12688/f1000research.10583.1.

Turqueti-Neves, Adriana, Manuel Otte, Olivia Prazeres da Costa, Uta E. Höpken, Martin Lipp, Thorsten Buch, and David Voehringer. 2014. “B-Cell-Intrinsic STAT6 Signaling Controls Germinal Center Formation.” European Journal of Immunology 44 (7): 2130–38. https://doi.org/10.1002/eji.201344203.

Vettermann, Christian, and Mark S. Schlissel. 2010. “Allelic Exclusion of Immunoglobulin Genes: Models and Mechanisms.” Immunological Reviews 237 (1): 22–42. https://doi.org/10.1111/j.1600-065X.2010.00935.x.

Ward, Lucas D., and Manolis Kellis. 2012. “HaploReg: A Resource for Exploring Chromatin States, Conservation, and Regulatory Motif Alterations within Sets of Genetically Linked Variants.” Nucleic Acids Research 40 (D1): 930–34. https://doi.org/10.1093/nar/gkr917.

Westra, Harm-Jan, Marjolein J Peters, Tõnu Esko, Hanieh Yaghootkar, Claudia Schurmann, Johannes Kettunen, Mark W Christiansen, et al. 2013. “Systematic Identification of Trans EQTLs as Putative Drivers of Known Disease Associations.” Nature Genetics 45 (10): 1238–43. https://doi.org/10.1038/ng.2756.

Yates, Andrew D, Premanand Achuthan, Wasiu Akanni, James Allen, Jamie Allen, Jorge Alvarez-Jarreta, M Ridwan Amode, et al. 2019. “Ensembl 2020.” Nucleic Acids Research, November. https://doi.org/10.1093/nar/gkz966.

Zan, Hong, and Paolo Casali. 2013. “Regulation of Aicda Expression and AID Activity.” Autoimmunity 46 (2): 83–101. https://doi.org/10.3109/08916934.2012.749244.

Zeller, Tanja, Philipp Wild, Silke Szymczak, Maxime Rotival, Arne Schillert, Raphaele Castagne, Seraya Maouche, et al. 2010. “Genetics and beyond - the Transcriptome of Human Monocytes and Disease Susceptibility.” PLoS ONE 5 (5). https://doi.org/10.1371/journal.pone.0010693.

Zhao, Lina, Xia Pu, Yuanqing Ye, Charles Lu, Joe Y. Chang, and Xifeng Wu. 2016. “Association between Genetic Variants in DNA Double-Strand Break Repair Pathways and Risk of Radiation Therapy-Induced Pneumonitis and Esophagitis in Non-Small Cell Lung Cancer.” Cancers 8 (2): 17–20. https://doi.org/10.3390/cancers8020023.

5.2 CAPÍTULO 2 -THE LANDSCAPE OF THE IMMUNOGLOBULIN REPERTOIRE IN ENDEMIC PEMPHIGUS FOLIACEUS

Verónica Calonga-Solís[1,2], Michael Olbrich[2], Gabriel A. Cipolla[1], Danielle Malheiros[1], Axel Künstner[2], Ticiana Della Justina Farias[1], Carolina M. Camargo[1], Maria Luiza Petzl-Erler[1], Hauke Busch[2,5], Anke Fähnrich[2,5,*], Danillo G. Augusto[1,4,5,**]

[1] Human Molecular Genetics Laboratory, Postgraduate Program in Genetics, Department of Genetics, Federal University of Paraná (UFPR), Centro Politécnico, Jardim das Américas, 81531-990 Curitiba, Brazil.
[2] Medical Systems Biology Division, Lübeck Institute of Experimental Dermatology and Institute for Cardiogenetics, University of Lübeck, Ratzeburger Allee 160, 23562 Lübeck, Germany.
[3] Neuroscience Research Institute and Department of Molecular, Cellular and Developmental Biology, University of California, Santa Barbara, Santa Barbara, USA.
[4] Department of Neurology, University of California San Francisco, USA.
[5] These authors contributed equally

Corresponding authors:

* Anke Fähnrich:
Medical Systems Biology Division, Lübeck Institute of Experimental Dermatology and Institute for Cardiogenetics, University of Lübeck, Ratzeburger Allee 160, 23562 Lübeck, Germany

** Danillo G. Augusto, PhD, MSc, BSc:
Department of Neurology, University of California San Francisco
675 Nelson Rising Lane, room 240. San Francisco, CA 94148, USA
danillo@augusto.bio.br

ABSTRACT

Pemphigus foliaceus (PF) is a B-cell-mediated autoimmune disease of the skin that occurs sporadically across the globe. Impressively, its endemic form reaches in some Brazilian areas the highest prevalence ever reported for an autoimmune disease (3%), indicating that environmental factors trigger this disease. To provide novel relevant insights about PF pathogenesis, we performed an unprecedented B-cell repertoire characterization in endemic PF, analyzing different subgroups of patients (n = 16), controls from the endemic area (n = 6), and controls from a non-endemic area (n = 4). After isolating peripheral blood mononuclear cells, we extracted total RNA and performed high-throughput next-generation sequencing and deep characterization of the variable region (VDJ exon) of the immunoglobulin heavy chain of IgG and IgM. For the first time, we report that both patients and healthy individuals living in endemic areas exhibit a remarkably lower clonotype diversity than healthy controls from a non-endemic area, suggesting an intense environmental pressure shaping their repertoire. We observed longer CDR3 sequences in patients than in controls and identified disease-specific differences in the usage of variable segments. For example, we identified that gene segments IGHV3-23 and IGHV3-30 have reduced and increased expression, respectively, in patients with active disease than individuals without the disease ($p < 0.04$). We used a state-of-the-art bioinformatic analysis to identify two PF-specific clusters in our clonotype network constructed with the aggregation of CDR3 sequence similarities. Our results show that environmental factors possibly shape the entire immunoglobulin repertoire in PF endemic areas. We also identified clonotype differences in patients and controls that can be further explored towards understanding PF pathogenesis. Our findings can be analysed in further studies focusing on new therapies that can potentially reduce the global burden of this disease, particularly addressing the health disparities in endemic areas from Brazil and other countries.

Keywords: endemic pemphigus foliaceus, autoimmunity, immunoglobulin repertoire

INTRODUCTION

Pemphigus foliaceus (PF) is a B-cell meditated autoimmune skin blistering disease characterized by cell detachment between keratinocytes (acantholysis) in the upper layer of the skin. This process is caused by the presence of autoantibodies against desmoglein 1 (Dsg1), a member of the cell-cell adhesion proteins of the desmosomes of keratinocytes (Kasperkiewicz et al. 2017). Although PF occurs sporadically across the globe, with an incidence of one case per million (cpm) (Bastuji-Garin et al. 1995; Marazza et al. 2009; Joly and Litrowski 2011; Kumar 2008), this disease is endemic in Brazil (25-35 cpm) (Hans-Filho et al. 1996; L. a Diaz et al. 1989). In some Brazilian areas, endemic PF reaches an impressive prevalence of 3% in some rural areas (Schmidt, Kasperkiewicz, and Joly 2019), the highest ever reported for any autoimmune disease.

Several genetic variants have been strongly associated with PF risk, including variants within the major histocompatibility complex (MHC) and the gene *NOTCH4* (Petzl-Erler 2020, Augusto et al., 2021). Among the environmental factors associated with its endemicity, the hematophagous flies *Lutzomyia longipalpis* are among the most relevant. Studies have shown that anti-Dsg1 antibodies cross-react with LJM11 molecule, an antigen derived from *L. longipalpis* salivary glands (Vernal et al. 2020; Peng et al. 2020; Qian et al. 2012). Additionally, it has been shown that serum antibodies from PF patients and controls from the endemic area bind to Dsg1 and the antigen maxadilan, another component of the salivary gland of these flies (Vernal et al. 2020). These results suggest that environmental antigens can initiate auto-reactive antibodies response in PF, leading to the development of pathogenic autoantibodies in genetically susceptible individuals.

Despite their abnormally pathogenic role in autoimmune diseases, antibody-mediated responses are critical for identification and protection against pathogens, toxins, or allergens through specific antigen-binding followed by neutralization, opsonization, and complement activation or stimulation of other cells of the immune system (Chaplin 2010). The antibodies are encoded by the *immunoglobulin heavy locus* (*IGH*), *lambda locus* (*IGL*), and *kappa locus* (*IGK*) (Croce et al. 1979; McBride et al. 1982) and are secreted by a subset of B cells that differentiate in plasma cells. Before mRNA transcription, these genes undergo somatic recombination of their variable (V), diversity (D), and junction (J) gene segments (Jung et al. 2006). This V(D)J

rearrangement is followed by somatic hypermutation and class switch recombination in B cells that have been stimulated after encountering their cognate antigens (Dudley et al. 2005; Jacobs and Bross 2001). Altogether, these mechanisms generate a great diversity of antibodies capable of recognizing virtually any antigen (Nossal 2003), especially in their complementary-determining region three (CDR3), which is the most dominant determinant of antigen-binding specificity (D'Angelo et al. 2018).

Even though the rearranged immunoglobulin genes (clonotypes) of anti-Dsg1 autoantibodies have been previously described in patients with PF (Chen et al. 2017; Qian et al. 2009), there has never been an investigation of the immunoglobulin repertoire in patients and controls. We hypothesized that the repertoire diversity differs between healthy individuals from the endemic area and healthy individuals from the non-endemic area and that features of the repertoire, such as IGHV usage and CDR3 characteristics, are different between patients and controls.

Here, we fill this gap by performing a complete characterization of the immunoglobulin repertoire in patients with PF and controls from the endemic Brazilian region and non-endemic area. We identified contrasting features in their immune repertoires, which significantly advances understanding of the pathogenesis of this neglected autoimmune disease that affects thousands of individuals worldwide.

METHODS

Study population

We analysed 22 individuals from the PF endemic areas in Brazil, divided into four groups: i) non-treated patients (before hospital admission; n = 5); ii) patients under immunosuppressive treatment (~30 mg of prednisone/day; n = 5); patients in disease remission, without treatment (n = 6), healthy controls (n = 6). We also included another subset of healthy individuals from outside the endemic area (n = 4) recruited in Curitiba, Brazil. We included as controls individuals who reported no history of autoimmune disease, recent infection, or other known medical conditions. We excluded those who reported use of medications. Patients and controls were paired for ancestry. Demographics of the study population is shown in Table 1.

Patients were contacted at hospitals localized in the endemic area and were diagnosed by dermatologists specialized in PF based on immunological tests, histopathology, immunohistochemistry of skin biopsies, and autoimmune bullous skin disorder intensity score (ABSIS). This study was approved by the Human Research Ethics Committee of the Federal University of Parana under protocol number CAAE 02727412.4.0000.0096, according to Brazilian Federal laws and the Declaration of Helsinki.

Library preparation and sequencing of the immune repertoire

For total RNA isolation, peripheral blood mononuclear cells (PBMC) were lysed and stored in TRizol Reagent (Invitrogen, USA) at -80 °C. Total RNA was then isolated according to the manufacturer's instructions. We used the SMARTer Human BCR IgG IgM H/K/L Profiling Kit (Takara Bio, USA) for reverse transcription (RT) and library preparation. We transcribed cDNA from 100 ng of total RNA, applying a 5' RACE-like RT that incorporates unique molecular identifiers (UMIs) to each mRNA molecule to facilitate PCR error correction (Shugay et al. 2014).

We performed two rounds of PCR that selectively amplified the full-length VDJ region and a portion of the constant region of immunoglobulin IGHM and IGHG transcripts, whose final product was barcoded with Illumina adapters. We selected amplicons with 400–900 bp, purified with magnetic beads (MagSi-NGS). Libraries were analyzed on the Bioanalyser (Agilent) for quality control and then finally pooled at equimolar concentration, denatured, diluted to 12 pM, and sequenced with Illumina MiSeq using paired-end 2x300 bp protocol (Ravi, Walton, and Khosroheidari 2018).

Sequence analysis and clonotype identification

The generated fastq files were first filtered for high quality using the MIGEC software (Shugay et al. 2014) and then processed for PCR error correction by the aggregation of UMIs into molecular identifier groups (MIGs) (Figure 1). Each sample was processed with regard to its MIG size threshold using the default setting of the software. The annotation of CDR3 sequences and the V, D, J, and C genes were performed independently with the software MiXCR (Bolotin et al. 2015) and IMGT/HighV-QUEST available at the IMGT database (Alamyar et al. 2012; Li et al. 2013). We used the software MiXCR and CHANGE-O (Gupta et al. 2015) for clonotype aggregation, quantification, and filtering.

For novel allele discovery, we performed germline sequence inference with the software CHANGE-O, followed by novel allele determination with the R package TIgGER (Gadala-Maria et al. 2015). The annotation of somatic hypermutation per clonotype was performed with IMGT/HighV-QUEST and analyzed with the package Shazam from the Immcantation toolkit (Gupta et al. 2015). We removed low-frequency clonotypes ($f < 0.001$) from the analysis.

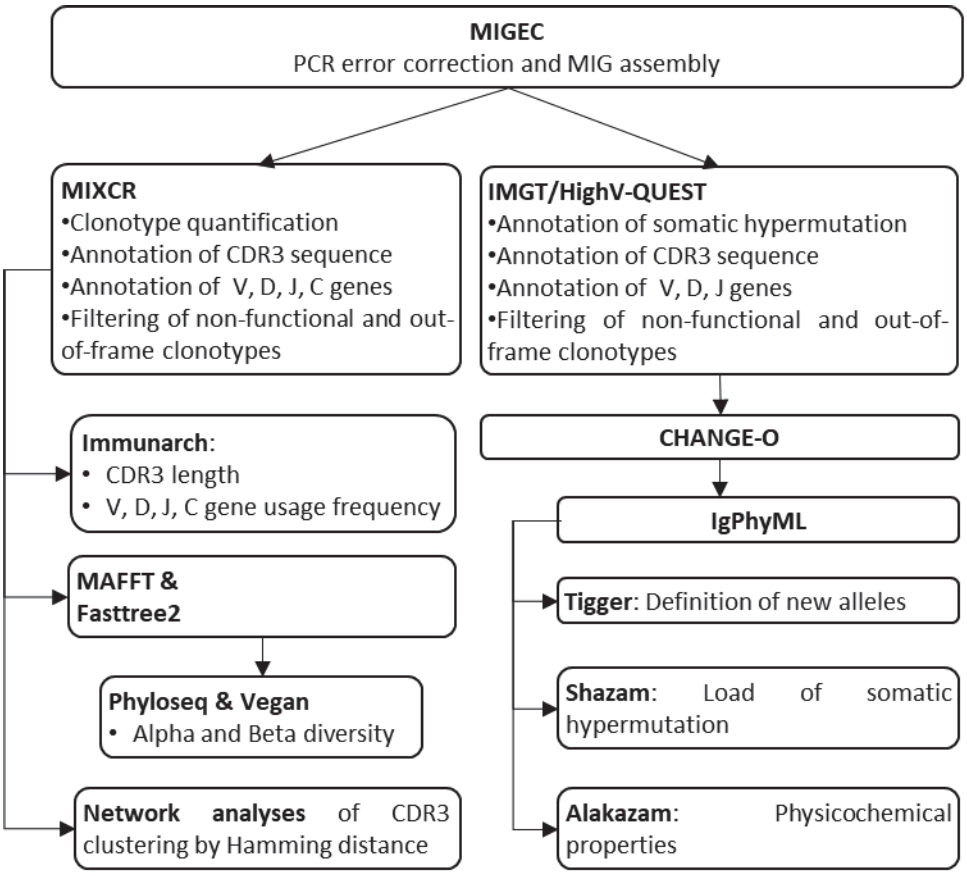


FIGURE 1: PIPELINE FOLLOWED FOR THE IMMUNE REPERTOIRE CHARACTERIZATION. MIG: MOLECULAR IDENTIFIER GROUP. CDR3: COMPLEMENTARITY DETERMINING REGION 3.

Data analysis

We estimated CDR3 length, IGHV, IGHJ, and IGHC gene usage frequencies with the R package Immunarch (Nazarov, Immunarch.bot, and Rumynskiy 2020). We applied the Shapiro-Wilk normality test (Shapiro and Wilk 1965) to analyze the CDR3 length

distribution and used the Kolmogorov–Smirnov test (Stephens 2006) to compare groups.

The frequency of each IGHV gene segment between groups with active disease and without the disease was compared with a Wilcoxon test (Rey and Neuhäuser 2011). The individual frequencies of the differentially expressed IGHV gene segments were included in the principal component analysis (PCA) to check for similarities among samples and clustering.

To evaluate clonotype physicochemical properties, we analyzed their CDR3 sequences with the package Alakazam (Gupta et al. 2015). To estimate CDR3 diversity, clonotype sequences were aligned using the software MAFFT, and we constructed a distance matrix from a phylogenetic tree with the software FastTree 2 (Price, Dehal, and Arkin 2010) applying the gamma and generalized time-reversible (GTR) model of nucleotide substitution. We used the R packages Phyloseq (McMurdie and Holmes 2013) and Vegan (Dixon 2003) to estimate alpha diversity using Faith's index of phylogenetic distance (PD) (Faith 1992) and Shannon index (Spellerberg and Fedor 2003), and UniFrac (Lozupone and Knight 2005) to evaluate beta diversity. For IGHG, we only considered medium-frequency and hyperexpanded clonotypes ($f>0.001$), while all clonotypes were considered for IGHM.

Furthermore, we investigated the similarities between the CDR3 sequences using network analysis as described previously (Pogorelyy and Shugay 2019; Pogorelyy et al. 2018; Madi et al. 2017). To construct a single network of clonotypes, we aggregated the clonotypes of all endemic samples into a single table maintaining the clonotype characteristics and sample of origin. For network assembly, we considered the CDR3 amino acid sequences as nodes, connected by edges according to their Hamming distance, i.e, the number of substitutions required to equalize sequences. We extracted all sub-networks that included clonotypes from at least four samples connected by a maximum Hamming distance of two for subsequent evaluation. The networks were processed using the R-package igraph (Csárdi and Nepusz 2006).

RESULTS

Clonotype frequency distribution differed among subgroups of patients and in controls

We analyzed the immunoglobulin repertoire in three groups of patients and one group of healthy individuals from the PF endemic region in Brazil, and another group of healthy individuals from a non-endemic region. We obtained the IGHM and IGHG clonotype repertoire, corresponding to the immunoglobulin heavy chain of IgM and IgG, and analyzed the entire variable region of the molecule (VDJ exon) and a fragment of the constant region to distinguish the isotypes and subclasses.

Although the raw read counts were similar among all groups, the mean number of MIGs and clonotypes of the non-endemic controls was up to ten times higher than the patient groups and controls from the endemic area (Table 1 and Supplementary Table S1).

TABLE 1: PATIENTS DIAGNOSTIC FEATURES AND REPERTOIRE SEQUENCING

	Non-treated patients	Patients under treatment	Patients in remission	Controls (endemic region)	Controls (non-endemic region)
Demographic characterization					
Mean Age	35.2	37	44.16	44	43.25
Sex	3F/2M	3F/2M	4F/2M	4F/2M	4F
Clinical characterization					
Median ABSIS	10	22,5	0	NA	NA
Mean Ab anti-Dsg1	206.98	216.41	35.57	NM	NM
IGHM repertoire					
Mean raw reads	1,246,283	1,203,399	1,211,195	1,296,708	1,241,521
Mean MIG	2,126.2	2,472.4	1,894.3	1,933.0	17,439.3
Mean clonotypes	1,359.8	1,346.4	1,109.3	1,224.2	12,268.5
IGHG repertoire					
Mean raw reads	2,099,649	1,818,216	1,878,301	1,900,283	1,362,850
Mean MIG	3,939.0	5,584.4	2,550.7	3,904.2	9,571.8
Mean clonotypes	1252.0	1,419.8	452.5	628.0	2,796.0

NA: Not applicable. NM: Not measured. F: Female. M: Male. ABSIS: Autoimmune Bullous Skin Disorder Intensity Score. MIG: molecular identifier group.

We analyzed the repertoire categorizing the clonotype frequencies into low (< 0.1%), medium (between 0.1 and 1%), and hyper-expanded (> 1%) (Figure 1). In

controls from the non-endemic area, the proportions of low-frequency clonotypes of IGHM and IGHG were 95% and 60%, respectively. Interestingly, controls from the endemic area exhibited larger proportions of medium-frequency and hyperexpanded clonotypes than those from the non-endemic area (Wilcoxon, $p < 0.04$). Large proportions of medium and high-frequency clonotypes were also observed in all group of patients (Figure 1)

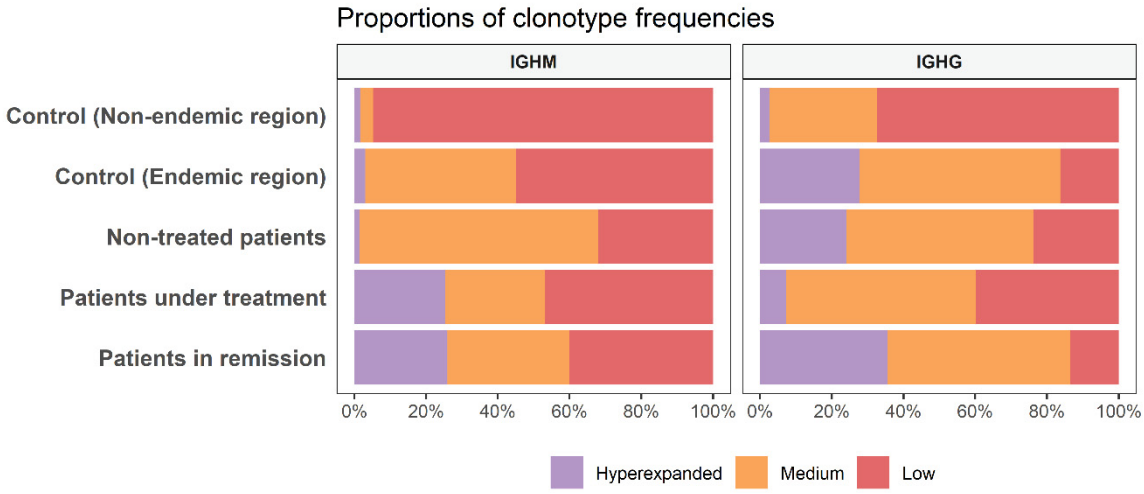


FIGURE 1: THE STUDY GROUPS DIFFER IN THE PROPORTION OF THEIR CLONOTYPE FREQUENCIES. Low (< 0.1%), Medium (between 0.1 and 1%) and Hyperexpanded (> 1%).

Both patients and controls from endemic areas present lower clonotype diversity than healthy individuals from the non-endemic area

We measured the phylogenetic distance between the CDR3 amino acid sequences of IGHM and IGHG clonotypes to evaluate the clonotype diversity within and among groups. We assessed the alpha diversity through Faith's phylogenetic diversity index (PD) (Faith 1992) by considering the total branch lengths of the phylogenetic tree within samples in each group. We observed no significant differences between the groups of patients and controls from the endemic area (Wilcoxon, $p > 0.05$). However, the controls from the non-endemic area exhibited a significantly higher PD than those from the endemic area (Wilcoxon, $p < 0.01$) for both IgM and IgG (Figure 2). This result was confirmed by the Shannon index, which considers the abundance

of clonotypes and the evenness of their frequencies (Supplementary Figure S1 – Apendice 2).

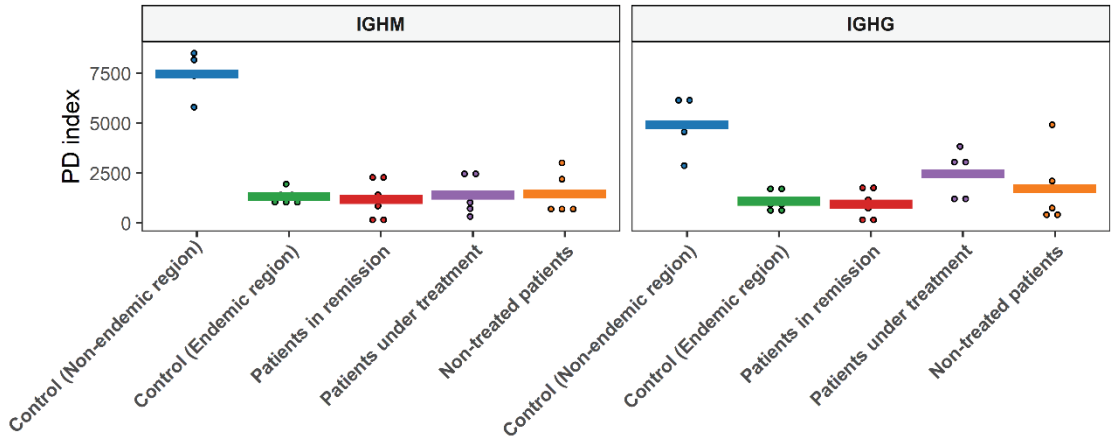


FIGURE 2: ALPHA DIVERSITY ASSESSED WITH PD (PHYLOGENETIC DIVERSITY INDEX). Non-endemic controls and endemic controls showed significant differences for both isotypes ($p < 0.01$). Comparisons between endemic samples were non-significant. The lines represent the mean PD value within the groups.

We further evaluated the similarities of the CDR3 sequences from different groups (beta diversity) with the UniFrac distances between the sequences in the phylogenetic tree. We performed a permutational multivariate analysis of variance (PERMANOVA) using the weighted UniFrac matrices of the phylogenetic distance to identify significant differences among the CDR3 sequences of the two geographic regions (endemic vs. non-endemic). For IGHM clonotypes, we only observed a tendency of separation between non-endemic (blue dots in Figure 3) and endemic regions (p = 0.11). However, a significant difference was observed for IGHG clonotypes between non-endemic (blue dots in Figure 3) and endemic regions (p = 0. 003; Figure 3). There was no differential clustering between controls and patients.

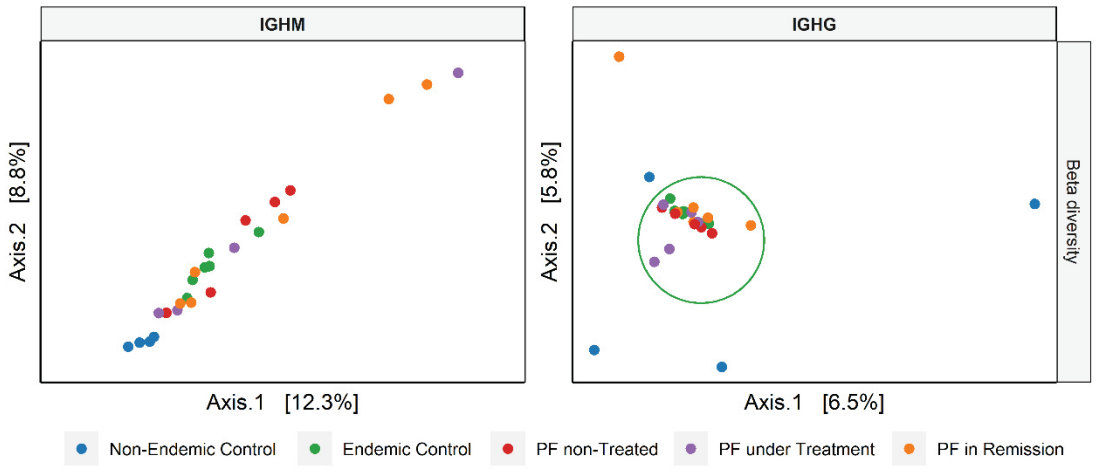


FIGURE 3: BETA DIVERSITY BASED PRINCIPAL COORDINATE ANALYSIS (PCOA) OF THE CDR3 AMINO ACID SEQUENCE CONSIDERING THE WHOLE REPERTOIRE FOR IGHM. No significant differentiation between regions (endemic vs. non-endemic, p = 0.11); and clonotypes with medium to high frequencies of IGHG, with significant clustering between regions (endemic vs. non-endemic, p = 0.003). One outlier data point was removed from IGHM PCoA.

IGHV2-70 (IGHM) and IGHV3-30 (IGHG) are differentially used in non-treated patients compared to controls from the endemic area

To evaluate if the IGHV gene segments are differently used in PF, we identified IGHV segments with significantly different frequencies between individuals with active disease (non-treated PF patients and PF patients under treatment) and those without the disease (controls from the endemic region and PF patients in remission) for both IGHM and IGHG isotypes (Table 2). The most frequent segments in patients with active disease were IGHV5-51 (IGHM) and IGHV3-30 (IGHG), while IGHV1-69 (IGHM) and IGHV3-23 (IGHG) were the most frequent in individuals without disease (Wilcoxon, $p < 0.05$; Table 2). Detailed IGHV gene usage frequency is given in Supplementary Table S2. We found no significant differential IGHJ gene usage (Supplementary Figure S2), nor differences in the IGHG isotype frequencies (Supplementary Figure S3). We performed a PCA using the IGHV segments that exhibited significantly different frequencies between individuals from the endemic region (Figure 4), and we found a separation between individuals with active disease (non-treated and under-treatment patients) in comparison to individuals without the disease (patients in remission and controls).

TABLE 2: DIFFERENTIAL IGHV USAGE IN INDIVIDUALS WITH ACTIVE PEMPHIGUS FOLIACEUS AND WITHOUT DISEASE

Isotype	Gene segment	Frequency		*p*-value*
		Without disease	Active disease	
IGHM	IGHV1-69	0.076	0.052	0.036
	IGHV2-5	0.008	0.014	0.038
	IGHV3-73	0.006	0.011	0.032
	IGHV5-51	0.026	0.037	0.002
IGHG	IGHV1-58	0.001	0.002	0.021
	IGHV3-23	0.098	0.080	0.017
	IGHV3-30	0.097	0.115	0.043

Active disease = patients with active PF lesions, with (n=5) and without treatment (n=5); without disease = healthy individuals from endemic areas (n=6) and patients in remission (n=6)

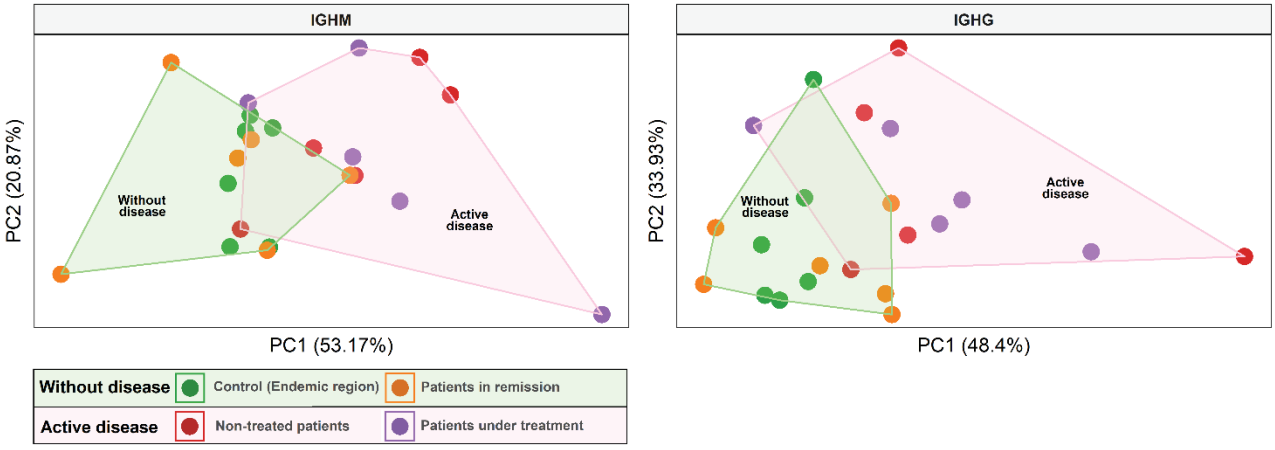


FIGURE 4: PRINCIPAL COMPONENT ANALYSIS WITH DIFFERENTIALLY USED IGHV GENE SEGMENTS IN ENDEMIC SAMPLES. The samples were grouped according to the disease state as active disease (PF non-treated and under treatment) and without disease (including endemic controls and PF in remission). Abbreviations: Abbreviations: PF: Patients with pemphigus foliaceus. ER: Endemic region.

IGHG clonotype of PF patients without immunosuppressive treatment have longer CDR3 amino acid sequences

We analyzed the distribution of the CDR3 length in all groups. Our Shapiro-Wilk normality test showed a normal distribution ($p > 0.05$) of IGHM and IGHG CDR3 length in the two control groups, while the distribution deviated from normality in all groups of patients, except the IGHM distribution in patients under treatment (Figure 5).

We found no statistical differences between the CDR3 length distribution of samples from the endemic region and controls from the non-endemic region (Kolmogorov-Smirnov test, $p > 0.05$). However, we observed an increased frequency of IGHG clonotypes with longer CDR3 sequences in patients who were not under the use of immunosuppressive treatment (arrows in Figure 5).

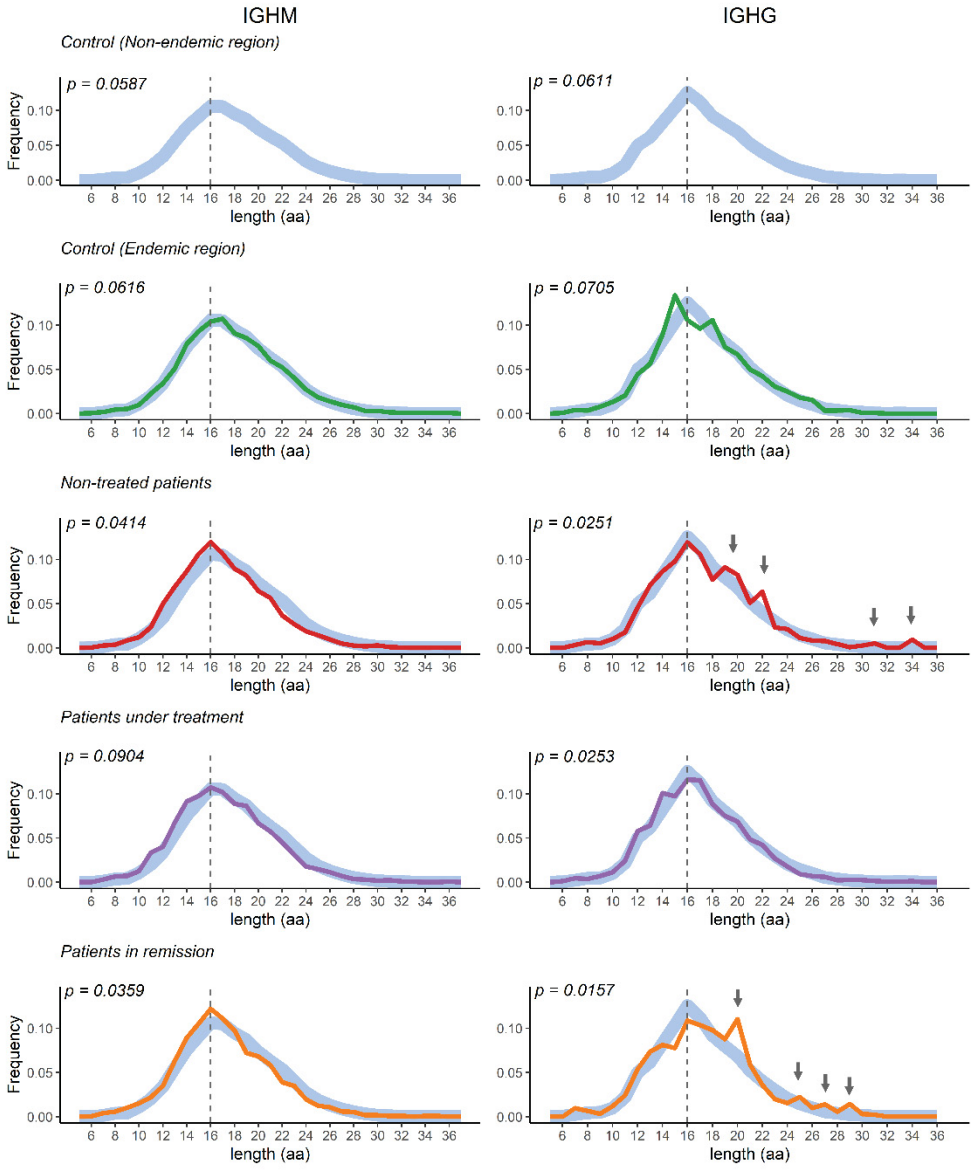


FIGURE 5: CDR3 LENGTH DISTRIBUTION IN THE STUDY POPULATION GROUPS. The curves represent the distribution of clonotypes with the different amino acid sequence lengths in their CDR3 region. Samples from the endemic region were compared with the controls from the non-endemic region as they represent a Gaussian distribution expected in healthy individuals (represented as thick light-blue curves). The vertical dotted line represents the median length. aa = amino acid. P values of the Shapiro-Wilk normality test for each group are represented in each graph.

Analyses of mutation load of the clonotypes and of their physicochemical properties showed no difference between groups.

Network analysis of CDR3 sequences reveals two clonotype clusters that might be implicated in pemphigus pathogenesis

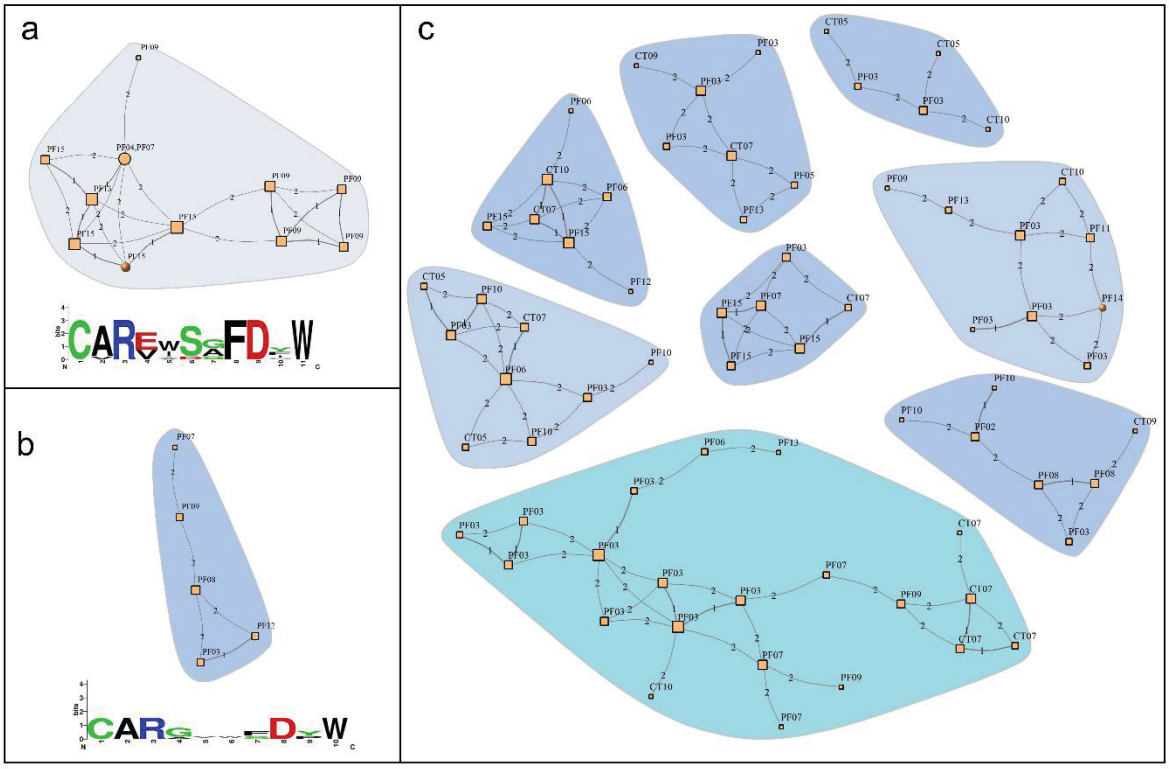


FIGURE 6: SIMILARITY NETWORKS AMONG CLONOTYPES OF PATIENTS AND CONTROLS FROM THE ENDEMIC AREA. (a-b) Networks consisting solely on clonotypes from patients (PF). (c) Networks consisting predominantly of clonotypes from patients with a few from controls (CT) from the endemic area. The networks were constructed based on the CDR3 nucleotide sequence of IGHG clonotypes. Each network connects at least four different samples. Nodes represent unique clonotypes and are labelled with the sample of origin. The numbers between nodes represent Hamming distances. Coloured letters represent the amino acid consensus sequence for each network; letter heights indicate the degree of conservation, and the most frequent amino acid for each position is placed on the top of each stack.

We evaluated the similarities between the CDR3 sequences of clonotypes from patients and controls from the endemic region. We considered a cluster relevant for PF if it connected clonotypes of at least four different samples (Figure 6), meaning that at least four individuals have similar clonotype sequences in each network. Two of these networks consist exclusively of clonotypes from patients (Figure 6a-b), and eight networks predominantly include clonotypes from patients, with a few clonotypes

found in controls from the endemic area (Figure 6c). Both PF clusters include clonotypes from five different samples, belonging to the three subgroups of patients. We found no networks only with clonotypes from controls of the endemic area. The CDR3 sequence of these clonotypes, and their associated IGHV, IGHJ, and IGHG segments are available in the Supplementary Table S3. The consensus sequence logo shows that the CDR3 sequences of the clonotypes in the network a (Figure 6a) are highly conserved compared to network b (Figure 6b). After comparing these clonotypes with the sequences available in the Pan immune repertoire database (PIRD) we found that several of them were previously associated with autoimmune disease, cancer, or pathogens.

DISCUSSION

We present the first study to analyze the IGHM and IGHG immune repertoire in patients with endemic PF, delivering the complete landscape of IgM and IgG in the peripheral blood of patients and controls. Because PF is the only known autoimmune disease that is also endemic, comparing healthy controls from the endemic areas with healthy individuals from a different region provides insights into the still uncovered environmental triggers, like pathogens and other antigens related to the already described precarious living conditions of PF patients (CULTON et al., 2008). In parallel, comparing healthy individuals with patients aids in distinguishing pathogenic from non-pathogenic repertoire characteristics.

Our first impactful observation was that those living in the endemic area exhibited a remarkably lower clonotype diversity than those healthy individuals living where PF is not endemic. This striking difference can be evidenced by different parameters, such as MIG and clonotype counts, the proportion of clonotype frequencies, and alpha diversity indexes.

The alpha diversity index represents the repertoire richness, in other words, the variability and abundance found in the CDR3 sequences within each group. Our study demonstrated a lower repertoire diversity in all samples from the endemic region than those from the non-endemic area. We show that most of the repertoire in individuals from the non-endemic area consists of low-frequency clonotypes, as would be expected for an immune repertoire that is not under stimuli by a small set of antigens (HERSHBERG; LUNING PRAK, 2015). In sharp contrast, medium frequency and

hyperexpanded clonotypes are more frequent in patients and controls from the endemic area, indicating an intense or constant environmental stimulation shaping their repertoire. We also observed that patients with non-treated PF exhibited a greater amount of medium frequency IGHM clonotypes, which could be a consequence of an early response against new antigens exposed in the skin lesions, as suggested by GRANDO, 2012. The high frequency of IGHG hyperexpanded clonotypes in patients is characteristic of an intense immune response against specific antigens (DOORENSPLEET et al., 2017). Besides, the reduction of hyperexpanded clonotypes in treated patients is consistent with the immunosuppression caused by the treatment. Therefore, our results indicate that the immune repertoire is shaped differently in the PF endemic area, possibly determined by environmental factors driving the expansion of specific clonotypes while decreasing the overall clonotype diversity.

In the beta-diversity based PCoA, the separation of individuals from endemic and non-endemic areas is a result of the lower UniFrac distance between clonotypes of the endemic samples due to similar CDR3 sequences, which could be the product of similar selective constraints or antigenic stimulation acting during immune repertoire development (CHAUDHARY; WESEMANN, 2018). Past studies corroborate this finding by showing that both patients and healthy individuals from the indigenous community of Limão Verde, located within the Brazilian endemic PF region, exhibited high IgG anti-Dsg1 levels, possibly due to constant exposition to an antigen that elicits immune responses generating autoantibodies. In contrast, anti-Dsg1 levels are lower in healthy individuals from other Brazilian localities and other countries (DIAZ et al., 2008).

IGHV utilization during immunoglobulin rearrangement is not entirely random (JACKSON et al., 2013). Some IGHV variants are preferentially used in the repertoire of healthy individuals, while others are overrepresented in some diseases (KITAURA et al., 2017). In the PCA based on the IGHV gene segment usage, we observed a clear separation between individuals with active disease from those without disease, indicating differential gene usage among these two groups. We identified seven gene segments with significantly different frequencies, of which IGHV3-30 (IGHG) is the most frequent in patients with active disease. IGHV3-30 has been described as one of the most used IGHV segments in antibodies anti-DSG1 from patients with PF from the endemic area (QIAN et al., 2009) and highly used in anti-DSG3 antibodies from a patient with pemphigus vulgaris (CHEN et al., 2017). High usage of this gene has also

been observed in patients with acute-on-chronic liver failure (YAN et al., 2019). Therefore, this gene may be more likely to lead to pathogenic antibodies. However, IGHV3-23, a gene segment that has also been found in higher frequency on anti-Dsg1 antibodies (QIAN et al., 2009), in our study was found in lower frequency in the groups with active disease than in those without disease, which could be a result of a proportional reduction due to the increase usage of IGHV3-30.

The study of IGHV gene usage is crucial because these genes encode the CDR1 and CDR2 regions that, together with CDR3 (SCHROEDER; CAVACINI, 2010), are critical for antigen binding and understanding differential antibody-antigen affinities (XU; DAVIS, 2000). The shift in IGHV usage in individuals with skin lesions may be implicated in the development of pathogenic antibodies, not only against anti-Dsg1 but also against other possible self-antigens. Ultimately, our detailed characterization of gene usage may be used in future studies to elucidate disease-associated mechanisms and test functional hypotheses.

An interesting feature in immunoglobulin repertoire studies is the length of the CDR3 region, as it can be an indicator of repertoire unbalance (MIQUEU et al., 2007). We found that patients with PF presented a deviation from normality in their distribution of CDR3 length with overrepresentation of clonotypes with longer CDR3 sequences, similar to what has been observed in other autoimmune diseases, like Immunoglobulin A Nephropathy, Crohn's disease and systemic lupus erythematosus (BASHFORD-ROGERS et al., 2019a; HUANG et al., 2019). Longer CDR3 sequences have been associated with more flexible and more polyreactive antibodies (LAFFY et al., 2017) and, therefore, with antibodies more prone to autoreactivity (Lange et al. 2014).

We analyzed the similarities among the clonotypes in the patients with PF to identify possible CDR3 sequences that can potentially lead to discovering new environmental antigens or self-antigens implicated in this disease. Although previous studies focusing on anti-Dsg1 antibodies showed no convergence in their clonotype sequences (Chen et al. 2017; Qian et al. 2009), we aimed to uncover additional frequent sequences in PF. We employed a state-of-the-art algorithm to construct similarity networks of clonotypes based on the Hamming distances of their CDR3 region sequences. This method computes the number of mismatching amino acids between sequences of equal length. The underlying assumption is that both sequence and length influence the binding properties of the antibodies to the antigens. Thus, similar sequences imply comparable epitope binding properties influencing the

clustering of sequences according to antigen specificity. Therefore, the CDR3 sequences clustering in the network analysis (Figure 6) could result from similar autoantigens or environmental antigens that are identified by PF-relevant antibodies, especially those in networks a and b which do not include clonotypes from controls. Utilizing the PIRD database, we found that the CDR3 sequence of the clonotypes in the network shown in figure 6a is similar to a clonotype previously found associated with the autoimmune disease IgA nephropathy (Huang et al. 2019).

It has been shown that the pathogenic anti-Dsg1 antibodies are mainly of the IgG4 isotype in Brazilian PF patients, whereas anti-Dsg1 IgG1 isotype is primarily present in healthy individuals and patients before the onset of disease. We intended to explore if this bias of IgG subclass is the entire repertoire; however, we found that a limitation of the library used for our sequencing is not allowing us to distinguish IGHG3 from IGHG4 gene segments. Nonetheless, we found no difference in the frequency of IGHG gene segments (IGHG1, IGHG2, and IGHG3/4) between groups (Supplementary Figure S3), a possible indication that the differential switching to IgG1 or IgG4 does not affect the whole repertoire.

CONCLUSIONS

In this study, we characterized the immunoglobulin repertoire of individuals from the endemic Brazilian region of pemphigus foliaceus and compared to healthy individuals from a non-endemic city in Brazil. We observed that the repertoire is deeply distinct in the two geographical locations, as we found that the endemic samples have lower clonotype counts, lower diversity, and a higher proportion of medium and hyperexpanded clonotypes. On the other hand, the repertoire of PF patients was characterized by a deviation of the normal distribution of the length of the CDR3 region, and a higher frequency of longer sequences, whis has been previously associated with autorreactive antibodies. Additionally, patients from the endemic area with active disease (PF non-treated and under treatment) have a different frequency of IGHV gene segment usage than individuals that have no lesions (endemic controls and PF in remission), being IGHV3-30 used with higher frequency in non-treated patients than in the endemic controls. In the network analysis, we found at least two clusters of clonotypes that can belong to antibodies involved in PF pathogenesis. In the future, these specific PF sequences could be used for the discovery of other environmental

PF triggers, unknown self-antigens, or pathogenic antibodies that could be used as target for therapeutical B-cells depletion.

ACKNOWLEDGEMENTS

The authors would like to express their gratitude to Fabian Ott for his kind assistance and help with the analyses performed in this study.

FUNDING

This work was supported by grants of Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq protocol 470483/2014-8), Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES/PROAP – Finance Code 001), Programa de Apoio a Núcleos de Excelência—Fundação Araucária de Apoio ao Desenvolvimento Científico e Tecnológico do Paraná (PRONEX-FA - Convênio 116/2018 - Protocol 50530), Fundação Araucária (FA protocol 39894.413.43926.1904/2013), Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy—EXC 22167-390884018, Deutscher Akademischer Austauschdienst (DAAD).

AUTHOR CONTRIBUTIONS

V.C.S., D.G.A. and A.F. conceived and designed the study. D.G.A., M.L.P.E. and H.B. financed the research. G.C., C.M.C. and T.D.J.F. collected the samples. V.C.S., D.G.A. and A.F. performed library preparation and sequencing. V.C.S., M.O., A.K. and A.F. performed data analysis. V.C.S. drafted the manuscript. M.O., D.G.A., A.F, H.B. and D.M. edited the manuscript. All authors read and approved the final manuscript.

CONFLICTS OF INTEREST:

The authors declare no competing interests.

REFERENCES

Alamyar, Eltaf, Véronique Giudicelli, Shuo Li, Patrice Duroux, and Marie Paule Lefranc. 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.

Bashford-Rogers, R. J.M., L. Bergamaschi, E. F. McKinney, D. C. Pombal, F. Mescia, J. C. Lee, D. C. Thomas, et al. 2019. “Analysis of the B Cell Receptor Repertoire in Six Immune-Mediated Diseases.” Nature 574 (7776): 122–26. https://doi.org/10.1038/s41586-019-1595-3.

Bashford-Rogers, Rachael J.M., Kenneth G.C. Smith, and David C. Thomas. 2018. “Antibody Repertoire Analysis in Polygenic Autoimmune Diseases.” Immunology 155 (1): 3–17. https://doi.org/10.1111/imm.12927.

Bastuji-Garin, Sylvie, Rafika Souissi, Laurent Blum, Hamida Turki, Rafia Nouira, Bechir Jomaa, Abdelmajid Zahaf, et al. 1995. “Comparative Epidemiology of Pemphigus in Tunisia and France: Unusual Incidence of Pemphigus Foliaceus in Young Tunisian Women.” Journal of Investigative Dermatology 104 (2): 302–5. https://doi.org/10.1111/1523-1747.ep12612836.

Bolotin, Dmitriy A, Stanislav Poslavsky, Igor Mitrophanov, Mikhail Shugay, Ilgar Z Mamedov, Ekaterina V Putintseva, and Dmitriy M Chudakov. 2015. “MiXCR: Software for Comprehensive Adaptive Immunity Profiling.” Nature Methods 12 (5): 380–81. https://doi.org/10.1038/nmeth.3364.

Chaplin, David D. 2010. “Overview of the Immune Response.” Journal of Allergy and Clinical Immunology 125 (2 SUPPL. 2): S3–23. https://doi.org/10.1016/j.jaci.2009.12.980.

Chaudhary, Neha, and Duane R. Wesemann. 2018. “Analyzing Immunoglobulin Repertoires.” Frontiers in Immunology 9 (MAR): 1–18. https://doi.org/10.3389/fimmu.2018.00462.

Chen, Jing, Qi Zheng, Christoph M. Hammers, Christoph T. Ellebrecht, Eric M. Mukherjee, Hsin Yao Tang, Chenyan Lin, et al. 2017. “Proteomic Analysis of Pemphigus Autoantibodies Indicates a Larger, More Diverse, and More Dynamic Repertoire than Determined by B Cell Genetics.” Cell Reports 18 (1): 237–47. https://doi.org/10.1016/j.celrep.2016.12.013.

Croce, C M, M Shander, J Martinis, L Cicurel, G G D’Ancona, T W Dolby, and H Koprowski. 1979. “Chromosomal Location of the Genes for Human Immunoglobulin Heavy Chains.” Proceedings of the National Academy of Sciences of the United States of America 76 (7): 3416–19. https://doi.org/10.1073/pnas.76.7.3416.

Csárdi, Gábor, and Tamás Nepusz. 2006. “The Igraph Software Package for Complex Network Research.” InterJournal Complex Sy. https://igraph.org.

Culton, Donna A., Ye Qian, Ning Li, David Rubenstein, Valeria Aoki, Gunter Hans Filhio, Evandro A. Rivitti, and Luis A. Diaz. 2008. “Advances in Pemphigus and Its Endemic Pemphigus Foliaceus (Fogo Selvagem) Phenotype: A Paradigm of Human Autoimmunity.” Journal of Autoimmunity 31 (4): 311–24. https://doi.org/10.1016/j.jaut.2008.08.003.

D’Angelo, Sara, Fortunato Ferrara, Leslie Naranjo, M. Frank Erasmus, Peter Hraber, and Andrew R.M. Bradbury. 2018. “Many Routes to an Antibody Heavy-Chain CDR3: Necessary, yet Insufficient, for Specific Binding.” Frontiers in Immunology 9 (MAR): 1–13. https://doi.org/10.3389/fimmu.2018.00395.

Diaz, L a, S a Sampaio, E a Rivitti, C R Martins, P R Cunha, C Lombardi, F a Almeida, R M Castro, M L Macca, and C Lavrado. 1989. “Endemic Pemphigus Foliaceus (Fogo Selvagem): II. Current and Historic Epidemiologic Studies.” The Journal of Investigative Dermatology 92: 4–12. https://doi.org/10.1111/1523-1747.ep13070394.

Diaz, Luis A., Phillip S. Prisayanh, David A. Dasher, Ning Li, Flor Evangelista, Valeria Aoki, Gunter Hans-Filho, Vandir Dos Santos, Bahjat F. Qaqish, and Evandro A. Rivitti. 2008. “The IgM Anti-Desmoglein 1 Response Distinguishes Brazilian Pemphigus Foliaceus (Fogo Selvagem) from Other Forms of Pemphigus.” Journal of Investigative Dermatology 128 (3): 667–75. https://doi.org/10.1038/sj.jid.5701121.

Dixon, Philip. 2003. “VEGAN, a Package of R Functions for Community Ecology.” Journal of Vegetation Science 14 (6): 927. https://doi.org/10.1658/1100-9233(2003)014[0927:vaporf]2.0.co;2.

Doorenspleet, M. E., L. Wester, C. P. Peters, T. B.M. Hakvoort, R. E. Esveldt, E. Vogels, A. H.C. van Kampen, et al. 2017. “Profoundly Expanded T-Cell Clones in the Inflamed and Uninflamed Intestine of Patients with Crohn’s Disease.” Journal of Crohn’s and Colitis 11 (7): 831–39. https://doi.org/10.1093/ecco-jcc/jjx012.

Dudley, Darryll D, Jayanta Chaudhuri, Craig H Bassing, and Frederick W Alt. 2005. “Mechanism and Control of V(D)J Recombination versus Class Switch Recombination: Similarities and Differences.” In Advances in Immunology, 86:43–112. https://doi.org/10.1016/S0065-2776(04)86002-4.

Faith, Daniel P. 1992. “Conservation Evaluation and Phylogenetic Diversity.” Biological Conservation 61 (1): 1–10. https://doi.org/10.1016/0006-3207(92)91201-3.

Gadala-Maria, Daniel, Gur Yaari, Mohamed Uduman, and Steven H. Kleinstein. 2015. “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 112 (8): E862–70. https://doi.org/10.1073/pnas.1417683112.

Grando, Sergei A. 2012. “Pemphigus Autoimmunity: Hypotheses and Realities.” Autoimmunity 45 (1): 7–35. https://doi.org/10.3109/08916934.2011.606444.

Gupta, Namita T., Jason A. Vander Heiden, Mohamed Uduman, Daniel Gadala-Maria, Gur Yaari, and Steven H. Kleinstein. 2015. “Change-O: A Toolkit for Analyzing Large-Scale B Cell Immunoglobulin Repertoire Sequencing Data.” Bioinformatics 31 (20): 3356–58. https://doi.org/10.1093/bioinformatics/btv359.

Hans-Filho, Gunter, Vandir Dos Santos, Joana H. Katayama, Valeria Aoki, Evandro A. Rivitti, Sebastiao A.P. Sampaio, Horacio Friedman, et al. 1996. “An Active Focus of High Prevalence of Fogo Selvagem on an Amerindian Reservation in Brazil.” Journal of Investigative Dermatology 107 (1): 68–75. https://doi.org/10.1111/1523-1747.ep12298213.

Hershberg, Uri, and Eline T. Luning Prak. 2015. “The Analysis of Clonal Expansions in Normal and Autoimmune B Cell Repertoires.” Philosophical Transactions of the Royal Society B: Biological Sciences 370 (1676). https://doi.org/10.1098/rstb.2014.0239.

Huang, Chen, Xuemei Li, Jinghua Wu, Wei Zhang, Shiren Sun, Liya Lin, Xie Wang, et al. 2019. “The Landscape and Diagnostic Potential of T and B Cell Repertoire in Immunoglobulin A Nephropathy.” Journal of Autoimmunity 97 (August 2018): 100–107. https://doi.org/10.1016/j.jaut.2018.10.018.

Jackson, Katherine J L, Marie J Kidd, Yan Wang, and Andrew M Collins. 2013. “The Shape of the Lymphocyte Receptor Repertoire: Lessons from the B Cell Receptor.” Frontiers in Immunology 4 (September): 263. https://doi.org/10.3389/fimmu.2013.00263.

Jacobs, Heinz, and Linda Bross. 2001. “Towards an Understanding of Somatic Hypermutation.” Current Opinion in Immunology 13 (2): 208–18. https://doi.org/10.1016/s0952-7915(00)00206-5.

Joly, Pascal, and Noémie Litrowski. 2011. “Pemphigus Group (Vulgaris, Vegetans, Foliaceus, Herpetiformis, Brasiliensis).” Clinics in Dermatology 29 (4): 432–36. https://doi.org/10.1016/j.clindermatol.2011.01.013.

Jung, David, Cosmas Giallourakis, Raul Mostoslavsky, and Frederick W Alt. 2006. “Mechanism and Control of V(D)J Recombination at the Immunoglobulin Heavy Chain Locus.” Annual Review of Immunology 24 (1): 541–70. https://doi.org/10.1146/annurev.immunol.23.021704.115830.

Kasperkiewicz, Michael, Christoph T Ellebrecht, Hayato Takahashi, Jun Yamagami, Detlef Zillikens, Aimee S Payne, and Masayuki Amagai. 2017. “Pemphigus.” Nature Reviews Disease Primers 3 (1): 17026. https://doi.org/10.1038/nrdp.2017.26.

Kitaura, Kazutaka, Hiroshi Yamashita, Hitomi Ayabe, Tadasu Shini, Takaji Matsutani, and Ryuji Suzuki. 2017. “Different Somatic Hypermutation Levels among Antibody Subclasses Disclosed by a New Next-Generation Sequencing-Based Antibody Repertoire Analysis.” Frontiers in Immunology 8 (MAY): 1–11. https://doi.org/10.3389/fimmu.2017.00389.

Kumar, KidangazhiyathmanaAjith. 2008. “Incidence of Pemphigus in Thrissur District, South India.” Indian Journal of Dermatology, Venereology and Leprology 74 (4): 349. https://doi.org/10.4103/0378-6323.42901.

Laffy, Julie M.J., Tihomir Dodev, Jamie A. Macpherson, Catherine Townsend, Hui Chun Lu, Deborah Dunn-Walters, and Franca Fraternali. 2017. “Promiscuous Antibodies Characterised by Their Physico-Chemical Properties: From Sequence to Structure and Back.” Progress in Biophysics and Molecular Biology 128: 47–56. https://doi.org/10.1016/j.pbiomolbio.2016.09.002.

Lange, Miles D., Lin Huang, Yangsheng Yu, Song Li, Hongyan Liao, Michael Zemlin, Kaihong Su, and Zhixin Zhang. 2014. “Accumulation of VH Replacement Products in IgH Genes Derived from Autoimmune Diseases and Anti-Viral Responses in Human.” Frontiers in Immunology 5 (JUL). https://doi.org/10.3389/fimmu.2014.00345.

Li, Shuo, Marie Paule Lefranc, John J. Miles, Eltaf Alamyar, Véronique Giudicelli, Patrice Duroux, J. Douglas Freeman, et al. 2013. “IMGT/HighV QUEST Paradigm for T Cell Receptor IMGT Clonotype Diversity and next Generation Repertoire Immunoprofiling.” Nature Communications 4 (May). https://doi.org/10.1038/ncomms3333.

Lozupone, Catherine, and Rob Knight. 2005. “UniFrac: A New Phylogenetic Method for Comparing Microbial Communities.” Applied and Environmental Microbiology 71 (12): 8228–35. https://doi.org/10.1128/AEM.71.12.8228-8235.2005.

Madi, Asaf, Asaf Poran, Eric Shifrut, Shlomit Reich-Zeliger, Erez Greenstein, Irena Zaretsky, Tomer Arnon, et al. 2017. “T Cell Receptor Repertoires of Mice and Humans Are Clustered in Similarity Networks around Conserved Public CDR3 Sequences.” ELife 6: 1–17. https://doi.org/10.7554/eLife.22057.

Marazza, G., H. C. Pham, L. Schärer, P. P. Pedrazzetti, T. Hunziker, R. M. Trüeb, D. Hohl, et al. 2009. “Incidence of Bullous Pemphigoid and Pemphigus in Switzerland: A 2-Year Prospective Study.” British Journal of Dermatology 161 (4): 861–68. https://doi.org/10.1111/j.1365-2133.2009.09300.x.

McBride, OW, PA Heiter, GF Hollis, D Swan, MC Otey, and P Leder. 1982. “Chromosomal Location of Human Kappa and Lambda Immunoglobulin Light Chain Constant Region Genes.” The Journal of Experimental Medicine 155 (5): 1480–90. https://doi.org/10.1084/jem.155.5.1480.

McMurdie, Paul J., and Susan Holmes. 2013. “Phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data.” PLoS ONE 8 (4). https://doi.org/10.1371/journal.pone.0061217.

Miqueu, Patrick, Marina Guillet, Nicolas Degauque, Jean Christophe Doré, Jean Paul Soulillou, and Sophie Brouard. 2007. “Statistical Analysis of CDR3 Length Distributions for the Assessment of T and B Cell Repertoire Biases.” Molecular Immunology 44 (6): 1057–64. https://doi.org/10.1016/j.molimm.2006.06.026.

Nazarov, Vadim, Immunarch.bot, and Eugene Rumynskiy. 2020. “Immunomind/Immunarch: 0.6.5: Basic Single-Cell Support.” Zenodo. https://doi.org/10.5281/ZENODO.3893991.

Nossal, Gustav J. V. 2003. “The Double Helix and Immunology.” Nature 421 (6921): 440–44. https://doi.org/10.1038/nature01409.

Peng, Bin, Brenda R. Temple, Jinsheng Yang, Songmei Geng, Donna A. Culton, and Ye Qian. 2020. “Identification of a Primary Antigenic Target of Epitope Spreading in Endemic

Pemphigus Foliaceus." Journal of Autoimmunity, no. August: 102561. https://doi.org/10.1016/j.jaut.2020.102561.

Pogorelyy, Mikhail V., Anastasia A. Minervina, Mikhail Shugay, Dmitriy M. Chudakov, Yuri B. Lebedev, Thierry Mora, and Aleksandra M. Walczak. 2018. "Detecting T-Cell Receptors Involved in Immune Responses from Single Repertoire Snapshots." BioRxiv, 1–13. https://doi.org/10.1101/375162.

Pogorelyy, Mikhail V., and Mikhail Shugay. 2019. "A Framework for Annotation of Antigen Specificities in High-Throughput T-Cell Repertoire Sequencing Studies." Frontiers in Immunology 10 (SEP): 1–9. https://doi.org/10.3389/fimmu.2019.02159.

Price, Morgan N., Paramvir S. Dehal, and Adam P. Arkin. 2010. "FastTree 2 - Approximately Maximum-Likelihood Trees for Large Alignments." PLoS ONE 5 (3). https://doi.org/10.1371/journal.pone.0009490.

Qian, Ye, Stephen H. Clarke, Valeria Aoki, Gunter Hans-Filhio, Evandro A. Rivitti, and Luis A. Diaz. 2009. "Antigen Selection of Anti-DSG1 Autoantibodies during and before the Onset of Endemic Pemphigus Foliaceus." Journal of Investigative Dermatology 129 (12): 2823–34. https://doi.org/10.1038/jid.2009.184.

Qian, Ye, Joseph S Jeong, Mike Maldonado, J. G. Valenzuela, Regis Gomes, Clarissa Teixeira, F. Evangelista, et al. 2012. "Cutting Edge: Brazilian Pemphigus Foliaceus Anti-Desmoglein 1 Autoantibodies Cross-React with Sand Fly Salivary LJM11 Antigen." The Journal of Immunology 189 (4): 1535–39. https://doi.org/10.4049/jimmunol.1200842.

Ravi, Rupesh Kanchi, Kendra Walton, and Mahdieh Khosroheidari. 2018. "MiSeq: A Next Generation Sequencing Platform for Genomic Analysis." In Disease Gene Identification: Methods and Protocols, 1706:223–32. https://doi.org/10.1007/978-1-4939-7471-9_12.

Rey, Denise, and Markus Neuhäuser. 2011. "Wilcoxon-Signed-Rank Test." In International Encyclopedia of Statistical Science, 1658–59. Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-04898-2_616.

Schmidt, Enno, Michael Kasperkiewicz, and Pascal Joly. 2019. "Pemphigus." Lancet 394 (10201): 882–94.

Schroeder, Harry W., and Lisa Cavacini. 2010. "Structure and Function of Immunoglobulins." Journal of Allergy and Clinical Immunology 125 (2): S41–52. https://doi.org/10.1016/j.jaci.2009.09.046.

Shapiro, S. S., and M. B. Wilk. 1965. "An Analysis of Variance Test for Normality (Complete Samples)." Biometrika 52 (3–4): 591–611. https://doi.org/10.1093/biomet/52.3-4.591.

Shugay, Mikhail, Olga V. Britanova, Ekaterina M. Merzlyak, Maria A. Turchaninova, Ilgar Z. Mamedov, Timur R. Tuganbaev, Dmitriy A. Bolotin, et al. 2014a. "Towards Error-Free Profiling of Immune Repertoires." Nature Methods 11 (6): 653–55. https://doi.org/10.1038/nmeth.2960.

Spellerberg, Ian F, and Peter J. Fedor. 2003. "A Tribute to Claude Shannon (1916-2001) and a Plea for More Rigorous Use of Species Richness, Species Diversity and the 'Shannon-Wiener' Index." Global Ecology and Biogeography 12 (3): 177–79. https://doi.org/10.1046/j.1466-822X.2003.00015.x.

Stephens, M. A. 2006. "Kolmogorov-Smirnov Statistics." In Encyclopedia of Statistical Sciences. Hoboken, NJ, USA: John Wiley & Sons, Inc. https://doi.org/10.1002/0471667196.ess1357.pub2.

Vernal, S., N. A. De Paula, V. R. Bollela, E. A. Lerner, and A. M. Roselino. 2020. "Pemphigus Foliaceus and Sand Fly Bites: Assessing the Humoral Immune Response to the Salivary Proteins Maxadilan and LJM11." British Journal of Dermatology, 10–12. https://doi.org/10.1111/bjd.19221.

Xu, John L., and Mark M. Davis. 2000. “Diversity in the CDR3 Region of V(H) Is Sufficient for Most Antibody Specificities.” Immunity 13 (1): 37–45. https://doi.org/10.1016/S1074-7613(00)00006-6.

Yan, Qiang, Lei Wang, Liusheng Lai, Song Liu, Huaizhou Chen, Jiaxing Zhang, Yong Dai, and Weiguo Sui. 2019. “Next Generation Sequencing Reveals Novel Alterations in B-Cell Heavy Chain Receptor Repertoires Associated with Acute-on-Chronic Liver Failure.” International Journal of Molecular Medicine 43 (1): 243–55. https://doi.org/10.3892/ijmm.2018.3946.

6 DISCUSSÃO GERAL

Na primeira parte desta tese foram considerados os polimorfismos em moléculas envolvidas na geração da diversidade dos anticorpos e sua associação com um risco aumentado de desenvolver pênfigo foliáceo endêmico. Apesar do papel importante dos anticorpos na patogênese da doença, esses genes não haviam sido previamente estudados nesse contexto.

O nosso estudo de associação se baseou em frequências alélicas em grupos de pacientes e controles. Como diferenças nas frequências podem ser resultado de diferenças de ancestralidade, o que poderia invalidar nossos resultados, um dos passos mais importantes do nosso trabalho foi a análise cuidadosa da ancestralidade da nossa amostra. Nós demonstramos que as nossas amostras de pacientes e controles apresentam proporções de ancestralidade comparáveis, de em média 71%, 15,5% e 13,5% de ancestralidade europeia, africana e indígena, respectivamente, e que poderíamos prosseguir com a análise de associação. A importância desse resultado é a validação de múltiplos estudos anteriores do nosso grupo de pesquisa (Revisado por PETZL-ERLER, 2020) que foram realizados em um momento em que não se tinha acesso a marcadores de ancestralidade e o pareamento de amostras era realizado de maneira subjetiva. Até recentemente, a identificação de ancestralidade dos indivíduos em estudos do nosso grupo era realizada com base nas características morfológicas faciais, tipo de cabelo e cor da pele, além de uma cuidadosa averiguação por parte do pesquisador, a fim de estimar a contribuição relativa das ascendências europeias, africanas e indígenas. Apesar de subjetiva, estudos do nosso grupo mostraram que esse método era eficaz para identificar principais grupos ancestrais dos indivíduos (BRAUN-PRADO et al., 2000; PROBST et al., 2000), mas nunca houve uma validação quantitativa e direta dessa abordagem. Portanto, a demonstração de que pacientes e controles pareados a partir de traços morfológicos apresentaram as mesmas proporções étnicas baseadas em uma análise de marcadores de ancestralidade vão além do nosso estudo de imunoglobulinas em PF, validando dezenas de estudos anteriores do nosso grupo de pesquisa.

É interessante notar que apesar de ser uma amostra constituída predominantemente de eurodescendentes, essa população da região endêmica é altamente miscigenada. Infelizmente, nossas análises limitaram-se a apenas esse grupo ancestral pois não haveria poder estatístico para analisar outros grupos menos representados. Também é importante salientar que as proporções de ancestralidade

encontradas em nosso estudo não refletem a população brasileira como um todo e tampouco a região endêmica, mas sim uma amostra de indivíduos predominantemente eurodescendentes que foram artificialmente selecionados dentro do conjunto total da população.

Apesar de verificarmos que os grupos de pacientes e controles são semelhantes em termos de ancestralidade, fomos extremamente rigorosos e utilizamos análise de componentes principais como covariáveis do nosso modelo de regressão. Dessa forma, a análise de associação pode ser ajustada para variações aparentemente imperceptíveis, mas que poderiam causar pequenas distorções dos nossos resultados. Encontramos associações significativas para vários SNPs em genes que codificam principalmente fatores de transcrição e de sinalização celular nas três grandes etapas estudadas: desenvolvimento de linfócitos B, rearranjo V(D)J e ativação dos linfócitos B, que leva a hipermutação somática e a mudança de classe por recombinação dos genes de imunoglobulinas. Todas as variantes foram encontradas em regiões intrónicas ou intergênicas, portanto, os seus efeitos no genótipo não seriam devido a mudanças nas sequências polipeptídicas ou na estrutura das moléculas, como tem sido o caso para a maioria das variantes associadas a doenças complexas (ZHANG; LUPSKI, 2015).

Um dos objetivos deste estudo foi analisar os polimorfismos nos genes de imunoglobulinas (*IGH*, *IGK* e *IGL*), uma vez que estudos prévios associaram variantes em suas sequências codificadoras, mais especificamente as que determinam os alótipos de Ig, com um risco aumentado a desenvolver doenças e outras características (TABELA 2). Entretanto, ao filtrar os dados obtidos com a genotipagem por microarranjo encontramos uma já antecipada baixa densidade de SNPs nesses genes, de 1 a 7 SNPs em cada 100 kb (em contraste a uma média de 88 SNPs/100kb encontrada nos demais genes candidatos), sendo nenhum deles marcador de alótipos de Ig. Isso reflete a dificuldade em incluir essas regiões em microarranjos de genotipagem e evidencia como os estudos de GWAS existentes não possuem capacidade de encontrar associações nessas regiões. Desta forma, essa limitação não permitiu que avaliássemos o polimorfismo desses genes específicos em PF de maneira aprofundada. Ainda, concluímos que a ausência de associação apontada pelo nosso estudo e também em todos os GWAS realizados a partir de microarranjos de genotipagem até o momento são provavelmente reflexo direto de limitações técnicas dessa abordagem. Isto demonstra a importância de estudos especificamente

desenhados para lidar com as características específicas dos genes de imunoglobulinas.

Estudos anteriores demonstraram associações de PF e polimorfismos de genes considerados na nossa análise, como por exemplo os genes *CD86, CD40*, *CD40L*, *TNFSF13B*, *IL4* (DALLA-COSTA et al., 2010; MALHEIROS; PETZL-ERLER, 2009; PEREIRA et al., 2004). Apesar disso, não encontramos associação com variantes desses genes no presente trabalho. No entanto, essa ausência de associação não implica que esses SNPs, ou mesmo, esses genes, não estejam envolvidos na susceptibilidade diferencial ao PF. Existem dois motivos principais para essa aparente discrepância: (1) as variantes previamente associadas podem não ter sido incluídas no microarranjo utilizado e (2) essas variantes podem ter sido excluídas da nossa análise devido à frequência inferior a MAF = 0,20. Para que pudéssemos obter poder estatísticos suficiente, fizemos cálculos que indicaram MAF = 0,20 como mínimo necessário para obter um poder estatístico de 0,80 no nosso tamanho amostral. Dessa forma, variáveis importantes para o contexto de PF podem ter sido excluídas devido ao corte de MAF= 0,20, que é resultado da limitação do tamanho amostral do nosso estudo. Apesar disso, nosso estudo encontrou variáveis associadas com PF nunca antes descritas, para as quais foi possível gerar hipóteses funcionais que podem ser testadas futuramente.

Na segunda metade deste trabalho, focamos no resultado da expressão dos genes das imunoglobulinas, que é a análise de repertório de imunoglobulinas nas células B. Para a caracterização do repertório, analisamos os segmentos gênicos utilizados, suas frequências relativas, a diversidade, e também características da região CDR3.

O kit de preparação da biblioteca utilizado no nosso trabalho é um dos mais modernos, e que permite os melhores resultados no sequenciamento de imunoglobulinas (SMARTer Human BCR; Takara Bio). Primeiramente, a incorporação de identificadores moleculares (UMI) em cada molécula de mRNA inicial na amostra, permite uma caracterização mais precisa do repertório, uma vez que permite identificar e eliminar mutações introduzidas nas etapas de amplificação (SHUGAY et al., 2014). Ainda, esse kit permite o enriquecimento de cada isotipo separadamente, permitindo o sequenciamento dos clonotipos de IGHM e IGHG (contendo IgG1, IgG2, IgG3 e IgG4). Dessa forma, a representatividade na amostra biológica (no nosso caso, PBMC) de um isotipo não afeta o sequenciamento dos outros. Se todos os isotipos

fossem sequenciados ao mesmo tempo, haveria uma possível perda das sequências com mais baixa representatividade.

Sabe-se que o repertório de imunoglobulinas muda conforme o microbioma dos indivíduos, ou a sua exposição a patógenos (HONDA; LITTMAN, 2016). Um resultado surpreendente do nosso estudo foi a grande discrepância entre indivíduos saudáveis da região endêmica e aqueles de uma região não-endêmica. Este trabalho é o primeiro estudo a realizar a caracterização do repertório de imunoglobulinas focando também nas diferenças da região endêmica de PF. Portanto, além das alterações resultantes da doença, também identificamos características que podem estar relacionadas ao ambiente da região onde PF tem alta prevalência. Mostramos que o repertório completo, expresso pelos indivíduos saudáveis que residem naquela região parece ser afetado por algum fator ambiental presente na área endêmica.

As redes de clonotipos que apresentamos foram um dos resultados mais relevantes da análise do repertório em PF. Essas redes foram geradas a partir das similaridades entre as sequências da região CDR3 dos clonotipos, comparando pacientes e controles da região endêmica. As duas redes compostas unicamente de pacientes podem identificar características únicas de PF, e incluir anticorpos relevantes para a patogênese. Existem diversos bancos de dados dedicados ao repertório de imunoglobulinas, entretanto, somente o banco de dados PIRD (ZHANG et al., 2020) nos permitiu descarregar seus dados para a avaliação da semelhança entre nossos clonotipos e os clonotipos do banco. A partir dele observamos que sequências do grupamento com maior semelhança entre si, de acordo com a sequência consenso obtida para o grupamento, são semelhantes a clonotipos associados com outra doença autoimune, a nefropatia por IgA (HUANG et al., 2019). Desta forma é possível que estes clonotipos identificados em nossas amostras, e que também foram identificados em outra doença autoimune, conformem autoanticorpos com potencial patogênico, e que estejam implicadas no desenvolvimento da doença.

As principais conclusões desse estudo foram:

- A análise de ancestralidade mostrou que os grupos amostrais de pacientes e controles foram pareados adequadamente para a realização de análise de associação.
- Dez SNPs, pertencentes a 11 genes codificadores de moléculas que participam no desenvolvimento e ativação dos linfócitos, no rearranjo V(D)J, na hipermutação

somática e na mudança de classe por recombinação, foram associados com susceptibilidade diferencial ao pênfigo foliáceo.

- As variantes associadas se encontram em regiões intergênicas ou intrônicas e as análises *in sílico* apontam um possível papel dessas variantes na regulação da expressão, da estrutura da cromatina ou *splicing* alternativo.
- O repertório de IG de indivíduos saudáveis da região endêmica de PF possui menor número de clonotipos, menor diversidade alfa, e diferente uso de segmentos gênicos IGHV em comparação aos indivíduos saudáveis da região não endêmica.
- O repertório de pacientes de PF se assemelha ao dos controles da região endêmica em relação aos níveis baixos de diversidade alfa e nas suas proporções da frequência dos clonotipos, com um enriquecimento de clonotipos com frequência média ou alta. Apesar disso, clonotipos de maior frequência foram ainda mais representados em pacientes.
- Os clonotipos IGHG de pacientes de PF apresentam desvios da distribuição Gaussiana da frequência do comprimento da região CDR3, apresentando uma maior representação de alguns CDR3 de maior tamanho.
- Os segmentos gênicos IGHV3-23 e IGHV3-30 têm expressão significativamente aumentada em indivíduos sem a doença e em pacientes com doença ativa, respectivamente.
- Identificamos dois conjuntos de clonotipos característicos de PF após os grupamentos baseados nas semelhanças das sequências de CDR3, que podem pertencer a anticorpos relevantes na patogênese da doença.

7 CONSIDERAÇÕES FINAIS

Anteriormente, durante a minha dissertação de mestrado, estudamos a diversidade contida nos segmentos gênicos gama nos genes da cadeia pesada das imunoglobulinas em populações brasileiras com um aspecto evolutivo (CALONGA-SOLÍS et al., 2019). Em contrapartida, nesse presente estudo nós expandimos o estudo desses genes no contexto do pênfigo foliáceo endêmico.

Nós mostramos que as variantes dos genes codificadores de moléculas envolvidas na produção de anticorpos podem conferir um risco aumentado ao PF, podendo afetar a expressão de moléculas envolvidas no desenvolvimento e ativação dos linfócitos B, e no reconhecimento, clivagem e união do DNA durante o rearranjo somático dos genes de imunoglobulinas, alterando assim o repertório de anticorpos dos indivíduos.

Também mostramos que o repertório de imunoglobulinas difere na utilização de segmentos gênicos IGHV e no comprimento da região CDR3 entre pacientes de pênfigo foliáceo e controles da área endêmica. Ao mesmo tempo, possuem características semelhantes, como as proporções de frequências de clonotipos e níveis de diversidade, que os distinguem dos controles da área não endêmica considerada neste estudo.

Para complementar o estudo do repertório de imunoglobulinas, está em andamento a análise do repertório dos receptores de linfócitos T (TCR). Essa próxima etapa tem como objetivo verificar se o repertório de linfócitos T também difere entre pacientes e controles, e entre indivíduos saudáveis das regiões endêmica e não endêmica. Identificaremos os clonotipos de TCR relevantes no PF para contrastá-los com aqueles existentes em bancos de dados de TCR e previamente associados com autoimunidade, alergia ou patógenos. Até esse momento, já concluímos a construção das bibliotecas a partir de mRNA de células mononucleadas de sangue periférico, e já foram sequenciadas as bibliotecas de 14 das 26 amostras. O sequenciamento havia sido inicialmente previsto para ser concluído em abril de 2020, mas teve que ser adiado devido ao fechamento da Universidade da Califórnia no período de pandemia. Dessa forma, temos como perspectiva de continuidade o preparo de um novo manuscrito com abordagem semelhante, mas que foca no repertório de células T.

Trabalhos futuros que explorem os clonotipos de células B e T são promissores para encontrar possíveis alvos terapêuticos para a doença, como por

exemplo promover a depleção específica de clones de linfócitos T e B patogênicos. Acreditamos que mais esforços são necessários para trazer novas perspectivas terapêuticas para essa doença que causa grande sofrimento a uma parcela altamente vulnerável da população brasileira. Esperamos que nossos resultados possam abrir novos caminhos no estudo dessa doença negligenciada e contribuir para a diminuição das disparidades sociais no acesso a tratamentos adequados na saúde pública brasileira.

8 REFERENCIAS BIBLIOGRÁFICAS

ABREU-VELEZ, A. M. et al. Neural system antigens are recognized by autoantibodies from patients affected by a new variant of endemic pemphigus foliaceus in Colombia. **Journal of Clinical Immunology**, v. 31, n. 3, p. 356–368, 2011.

ABRÉU-VÉLEZ, A. M. et al. Endemic pemphigus foliaceus over a century: Part I. **North American journal of medical sciences**, v. 2, n. 2, p. 51–9, 2010.

AMAGAI, M. Pemphigus as a paradigm of autoimmunity and cell adhesion. **Keio Journal of Medicine**, v. 51, n. 3, p. 133–139, set. 2002.

ANHALT, G. J. et al. Induction of Pemphigus in Neonatal Mice by Passive Transfer of IgG from Patients with the Disease. **New England Journal of Medicine**, v. 306, n. 20, p. 1189–1196, 20 maio 1982.

AOKI, V. et al. Environmental Risk Factors in Endemic Pemphigus Foliaceus (Fogo Selvagem). **Journal of Investigative Dermatology Symposium Proceedings**, v. 9, n. 1, p. 34–40, 2004.

AOKI, V. et al. Update on fogo selvagem , an endemic form of pemphigus foliaceus. **The Journal of Dermatology**, v. 42, n. September 2014, p. 18–26, 2015.

APLIN, B. D. et al. Tolerance through Indifference: Autoreactive B Cells to the Nuclear Antigen La Show No Evidence of Tolerance in a Transgenic Model. **The Journal of Immunology**, v. 171, n. 11, p. 5890–5900, 2003.

AUGUSTO, D. G. et al. Activating KIR and HLA Bw4 ligands are associated to decreased susceptibility to pemphigus foliaceus, an autoimmune blistering skin disease. **PLoS ONE**, v. 7, n. 7, p. e39991, jan. 2012.

AUGUSTO, D. G. et al. Pemphigus is associated with KIR3DL2 expression levels and provides evidence that KIR3DL2 may bind HLA-A3 and A11 in vivo. **European Journal of Immunology**, v. 45, n. 7, p. 2052–2060, jul. 2015.

AUGUSTO, D. G. et al. Unsuspected Associations with Variants within the Genes NOTCH4 and STEAP2-ASI Uncovered by a Genome-Wide Association Study in Endemic Pemphigus Foliaceus. **Journal of Investigative Dermatology**, maio 2021.

AVALOS-DÍAZ, E. et al. Transplacental passage of maternal pemphigus foliaceus autoantibodies induces neonatal pemphigus. **Journal of the American Academy of Dermatology**, v. 43, n. 6, p. 1130–1134, 2000.

BASHFORD-ROGERS, R. J. M. et al. Analysis of the B cell receptor repertoire in six immune-mediated diseases. **Nature**, v. 574, n. 7776, p. 122–126, 2019a.

BASHFORD-ROGERS, R. J. M. et al. Analysis of the B cell receptor repertoire in six immune-mediated diseases. **Nature**, v. 574, n. 7776, p. 122–126, 25 out. 2019b.

BASHFORD-ROGERS, R. J. M.; SMITH, K. G. C.; THOMAS, D. C. Antibody repertoire analysis in polygenic autoimmune diseases. **Immunology**, v. 155, n. 1, p. 3–17, 2018a.

BASHFORD-ROGERS, R. J. M.; SMITH, K. G. C.; THOMAS, D. C. Antibody repertoire analysis in polygenic autoimmune diseases. **Immunology**, v. 155, n. 1, p. 3–17, 2018b.

BASTUJI-GARIN, S. et al. Comparative Epidemiology of Pemphigus in Tunisia and France: Unusual Incidence of Pemphigus Foliaceus in Young Tunisian Women. **Journal of Investigative Dermatology**, v. 104, n. 2, p. 302–305, fev. 1995.

BIALYNICKI-BIRULA, R. et al. Pregnancy-triggered maternal pemphigus vulgaris with persistent gingival lesions. **Acta dermatovenerologica Croatica : ADC**, v. 19, n. 3, p. 170–5, jan. 2011.

BONILLA, F. A.; OETTGEN, H. C. Adaptive immunity. **Journal of Allergy and Clinical Immunology**, v. 125, n. 2 SUPPL. 2, p. S33–S40, 2010.

BRAUN-PRADO, K. et al. HLA class I polymorphism, as characterised by PCR-SSOP, in a Brazilian exogamic population. **Tissue antigens**, v. 56, n. 5, p. 417–27, nov. 2000.

BRAUN-PRADO, K.; PETZL-ERLER, M. L. Programmed cell death 1 gene (PDCD1) polymorphism and pemphigus foliaceus (fogo selvagem) disease susceptibility. **Genetics and Molecular Biology**, v. 30, n. 2, p. 314–321, 2007.

BROCHADO, M. J. F. et al. Differential HLA class I and class II associations in pemphigus foliaceus and pemphigus vulgaris patients from a prevalent Southeastern Brazilian region. **Journal of Autoimmunity**, v. 72, p. 19–24, ago. 2016.

BUCK, D. et al. Genetic variants in the immunoglobulin heavy chain locus are associated with the IgG index in multiple sclerosis. **Annals of Neurology**, v. 73, n. 1, p. 86–94, 2013.

BUMILLER-BINI, V. et al. Sparking fire under the skin? Answers from the association of complement genes with pemphigus foliaceus. **Frontiers in Immunology**, v. 9, n. APR, 2018.

BUMILLER-BINI, V. et al. Condemned or Not to Die? Gene Polymorphisms Associated With Cell Death in Pemphigus Foliaceus. **Frontiers in Immunology**, v. 10, n. October, 2019.

CALONGA-SOLÍS, V. et al. Unveiling the Diversity of Immunoglobulin Heavy Constant Gamma (IGHG) Gene Segments in Brazilian Populations Reveals 28 Novel Alleles and Evidence of Gene Conversion and Natural Selection. **Frontiers in Immunology**, v. 10, n. June, 4 jun. 2019.

CAMARGO, C. M.; AUGUSTO, D. G.; PETZL-ERLER, M. L. Differential gene expression levels might explain association of LAIR2 polymorphisms with pemphigus. **Human Genetics**, v. 135, n. 2, p. 233–244, 2016.

CANN, H. M. A Human Genome Diversity Cell Line Panel. **Science**, v. 296, n. 5566, p. 261b – 262, 12 abr. 2002.

CELERE, B. S. et al. Spatial distribution of pemphigus occurrence over five decades in Southeastern Brazil. **American Journal of Tropical Medicine and Hygiene**, v. 97, n. 6, p. 1737–1745, 2017.

CHAMS-DAVATCHI, C. et al. Pemphigus: Analysis of 1209 cases. **International Journal of Dermatology**, v. 44, n. 6, p. 470–476, 2005.

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

CHEN, J. et al. Proteomic Analysis of Pemphigus Autoantibodies Indicates a Larger, More Diverse, and More Dynamic Repertoire than Determined by B Cell Genetics. **Cell Reports**, v. 18, n. 1, p. 237–247, 2017.

CHO, M. J. et al. Shared VH1-46 gene usage by pemphigus vulgaris autoantibodies indicates common humoral immune responses among patients. **Nature Communications**, v. 5, n. May, 2014.

CHU, P. G.; ARBER, D. A. CD79: A Review. **Applied Immunohistochemistry & Molecular Morphology**, v. 9, n. 2, p. 97–106, jun. 2001.

CIPOLLA, G. A. et al. A 3′UTR polymorphism marks differential KLRG1 mRNA levels through disruption of a miR-584-5p binding site and associates with pemphigus foliaceus susceptibility. **Biochimica et Biophysica Acta - Gene Regulatory Mechanisms**, v. 1859, n. 10, p. 1306–1313, 2016.

CROCE, C. M. et al. Chromosomal location of the genes for human immunoglobulin heavy chains. **Proceedings of the National Academy of Sciences of the United States of America**, v. 76, n. 7, p. 3416–9, 1979.

CULTON, D. A. et al. Advances in pemphigus and its endemic pemphigus foliaceus (Fogo Selvagem) phenotype: A paradigm of human autoimmunity. **Journal of Autoimmunity**, v. 31, n. 4, p. 311–324, dez. 2008.

DALLA-COSTA, R. et al. Polymorphisms in the 2q33 and 3q21 chromosome regions including T-cell coreceptor and ligand genes may influence susceptibility to pemphigus foliaceus. **Human immunology**, v. 71, n. 8, p. 809–17, ago. 2010.

DIAZ, L. A. et al. The IgM anti-desmoglein 1 response distinguishes Brazilian pemphigus foliaceus (Fogo Selvagem) from other forms of pemphigus. **Journal of Investigative Dermatology**, v. 128, n. 3, p. 667–675, 2008.

DIAZ, L. A et al. Endemic pemphigus foliaceus (Fogo Selvagem): II. Current and historic epidemiologic studies. **The Journal of investigative dermatology**, v. 92, p. 4–12, 1989.

DOORENSPLEET, M. E. et al. Profoundly expanded T-cell clones in the inflamed and uninflamed intestine of patients with Crohn's disease. **Journal of Crohn's and Colitis**, v. 11, n. 7, p. 831–839, 2017.

DUDLEY, D. D. et al. Mechanism and Control of V(D)J Recombination versus Class Switch Recombination: Similarities and Differences. In: **Advances in immunology**. [s.l: s.n.]. v. 86p. 43–112.

EASTMAN, Q. M.; LEU, T. M. J.; SCHATZ, D. G. Initiation of V(D)J recombination in vitro obeying the 12/23 rule. **Nature**, v. 380, n. 6569, p. 85–88, 1996.

EMING, R. et al. Pathogenic IgG Antibodies against Desmoglein 3 in Pemphigus Vulgaris Are Regulated by HLA-DRB1*04:02–Restricted T Cells. **The Journal of Immunology**, v. 193, n. 9, p. 4391–4399, 1 nov. 2014.

FINKELMAN, F. D. et al. Lymphokine control of in vivo immunoglobulin isotype selection. **Annual Review of Immunology**, v. 8, p. 303–330, 1990.

FUXA, M.; SKOK, J. A. Transcriptional regulation in early B cell development. **Current Opinion in Immunology**, v. 19, n. 2, p. 129–136, 2007.

GEISBERGER, R.; LAMERS, M.; ACHATZ, G. The riddle of the dual expression of IgM and IgD. **Immunology**, v. 118, n. 4, p. 429–437, 2006.

GLANVILLE, J. et al. Deep sequencing in library selection projects: what insight does it bring? **Current Opinion in Structural Biology**, v. 33, p. 146–160, ago. 2015.

GRANDO, S. A. Pemphigus autoimmunity: Hypotheses and realities. **Autoimmunity**, v. 45, n. 1, p. 7–35, 2012a.

GRANDO, S. A. **Pemphigus autoimmunity: Hypotheses and realitiesAutoimmunity**, 2012b.

GRANOFF, D. M. et al. Interactive effect of genes associated with

immunoglobulin allotypes and HLA specificities on susceptibility to Haemophilus influenzae disease. **Journal of immunogenetics**, v. 11, n. 3–4, p. 181–8, jun. 1984.

GRAWUNDER, U. et al. Down-regulation of RAG1 and RAG2 gene expression in PreB cells after functional immunoglobulin heavy chain rearrangement. **Immunity**, v. 3, n. 5, p. 601–608, 1995.

HALVERSON, R.; TORRES, R. M.; PELANDA, R. Receptor editing is the main mechanism of B cell tolerance toward membrane antigens. **Nature Immunology**, v. 5, n. 6, p. 645–650, 23 jun. 2004.

HANS-FILHO, G. et al. An active focus of high prevalence of Fogo selvagem on an Amerindian reservation in Brazil. **Journal of Investigative Dermatology**, v. 107, n. 1, p. 68–75, 1996.

HANS-FILHO, G. et al. Endemic pemphigus foliaceus (fogo selvagem)--1998. The Cooperative Group on Fogo Selvagem Research. **Clinics in dermatology**, v. 17, n. 2, p. 225–35; discussion 105-6, 4 mar. 1999.

HERSHBERG, U. et al. Persistence and selection of an expanded B-cell clone in the setting of rituximab therapy for Sjögren's syndrome. **Arthritis Research and Therapy**, v. 16, n. 1, 2014.

HERSHBERG, U.; LUNING PRAK, E. T. The analysis of clonal expansions in normal and autoimmune B cell repertoires. **Philosophical Transactions of the Royal Society B: Biological Sciences**, v. 370, n. 1676, 2015.

HERTL, M.; EMING, R.; VELDMAN, C. T cell control in autoimmune bullous skin disorders. **Journal of Clinical Investigation**, v. 116, n. 5, p. 1159–1166, 2006.

HERVÉ, M. et al. CD40 ligand and MHC class II expression are essential for human peripheral B cell tolerance. **Journal of Experimental Medicine**, v. 204, n. 7, p. 1583–1593, 9 jul. 2007.

HOLLING, T. M.; SCHOOTEN, E.; VAN DEN ELSEN, P. J. Function and regulation of MHC class II molecules in T-lymphocytes: of mice and men. **Human Immunology**, v. 65, n. 4, p. 282–290, abr. 2004.

HONDA, K.; LITTMAN, D. R. The microbiota in adaptive immune homeostasis and disease. **Nature**, v. 535, n. 7610, p. 75–84, 2016.

HUANG, C. et al. The landscape and diagnostic potential of T and B cell repertoire in Immunoglobulin A Nephropathy. **Journal of Autoimmunity**, v. 97, n. August 2018, p. 100–107, 2019.

HÜBNER, F. et al. Prevalence and Age Distribution of Pemphigus and

Pemphigoid Diseases in Germany. **Journal of Investigative Dermatology**, v. 136, n. 12, p. 2495–2498, 2016.

HWANG, J. K.; ALT, F. W.; YEAP, L.-S. Related Mechanisms of Antibody Somatic Hypermutation and Class Switch Recombination. **Microbiology spectrum**, v. 3, n. 1, p. MDNA3- 0037–2014, 1 fev. 2015.

JACKSON, K. J. L. et al. The shape of the lymphocyte receptor repertoire: lessons from the B cell receptor. **Frontiers in immunology**, v. 4, n. September, p. 263, jan. 2013.

JAIRO, Ó.; OCAMPO, V.; LOPERA, M. M. V. Inmunopatogenia del pénfigo vulgar y el pénfigo foliáceo. v. 24, n. 3, p. 272–286, 2011.

JAMES, K. A.; CULTON, D. A.; DIAZ, L. A. Diagnosis and Clinical Features of Pemphigus Foliaceus. **Dermatologic Clinics**, v. 29, n. 3, p. 405–412, jul. 2011.

JEFFERIS, R.; LEFRANC, M. P. Human immunoglobulin allotypes: Possible implications for immunogenicity. **mAbs**, v. 1, n. 4, p. 332–338, 2009.

JOLY, P.; LITROWSKI, N. Pemphigus group (vulgaris, vegetans, foliaceus, herpetiformis, brasiliensis). **Clinics in Dermatology**, v. 29, n. 4, p. 432–436, jul. 2011.

JUNG, D. et al. Mechanism and control of V(D)J recombination at the immunoglobulin heavy chain locus. **Annual review of immunology**, v. 24, n. 1, p. 541–70, abr. 2006.

KAZEROUNIAN, S. et al. **Envoplakin and periplakin, the paraneoplastic pemphigus antigens, are also recognized by pemphigus foliaceus autoantibodies.The Journal of investigative dermatology**, 2000.

KINNUNEN, T. et al. Accumulation of peripheral autoreactive B cells in the absence of functional human regulatory T cells. **Blood**, v. 121, n. 9, p. 1595–1603, 2013.

KITAURA, K. et al. Different somatic hypermutation levels among antibody subclasses disclosed by a new next-generation sequencing-based antibody repertoire analysis. **Frontiers in Immunology**, v. 8, n. MAY, p. 1–11, 2017.

KOTOWICZ, K.; CALLARD, R. E. Human immunoglobulin class and IgG subclass regulation: Dual action of interleukin-4. **European Journal of Immunology**, v. 23, n. 9, p. 2250–2256, 1993.

KUMAR, K. Incidence of pemphigus in Thrissur district, south India. **Indian Journal of Dermatology, Venereology and Leprology**, v. 74, n. 4, p. 349, 2008.

LAFFY, J. M. J. et al. Promiscuous antibodies characterised by their physico-

chemical properties: From sequence to structure and back. **Progress in Biophysics and Molecular Biology**, v. 128, p. 47–56, 2017.

LANGE, M. D. et al. Accumulation of VH replacement products in IgH genes derived from autoimmune diseases and anti-viral responses in human. **Frontiers in Immunology**, v. 5, n. JUL, 2014.

LEBIEN, T. W.; TEDDER, T. F. B lymphocytes: How they develop and function. **Blood**, v. 112, n. 5, p. 1570–1580, 2008.

LEFRANC, M.-P. et al. IMGT®, the international ImMunoGeneTics information system® 25 years on. **Nucleic acids research**, v. 43, n. Database issue, p. D413-22, 28 jan. 2015.

LEFRANC, M.-P.; LEFRANC, G. **The Immunoglobulin FactsBook**. [s.l.] Academic Press, 2001.

LEFRANC, M.-P. P.; LEFRANC, G. G. Human Gm, Km, and Am allotypes and their molecular characterization: A remarkable demonstration of polymorphism. **Methods in Molecular Biology**, v. 882, p. 635–680, 2012.

LI, N. et al. The Role of Intramolecular Epitope Spreading in the Pathogenesis of Endemic Pemphigus Foliaceus (Fogo Selvagem). **The Journal of Experimental Medicine**, v. 197, n. 11, p. 1501–1510, 2 jun. 2003.

LOMBARDI, C. et al. Environmental risk factors in endemic pemphigus foliaceus (Fogo selvagem). “The Cooperative Group on Fogo Selvagem Research”. **The Journal of investigative dermatology**, v. 98, n. 6, p. 847–50, jun. 1992.

MAK, T. W.; SAUNDERS, M. E.; JETT, B. D. **Primer to the Immune Response**. [s.l.] Elsevier, 2014.

MAŁECKA, A. et al. Immunoglobulin heavy and light chain gene features are correlated with primary cold agglutinin disease onset and activity. **Haematologica**, v. 101, n. 9, p. e361-4, 1 set. 2016.

MALHEIROS, D. et al. Genome-wide gene expression profiling reveals unsuspected molecular alterations in pemphigus foliaceus. **Immunology**, p. n/a-n/a, 2014.

MALHEIROS, D.; PETZL-ERLER, M. L. Individual and epistatic effects of genetic polymorphisms of B-cell co-stimulatory molecules on susceptibility to pemphigus foliaceus. **Genes and immunity**, v. 10, n. 6, p. 547–558, 2009.

MALU, S. et al. Role of non-homologous end joining in V(D)J recombination. **Immunologic Research**, v. 54, n. 1–3, p. 233–246, 2012.

MARAZZA, G. et al. Incidence of bullous pemphigoid and pemphigus in Switzerland: A 2-year prospective study. **British Journal of Dermatology**, v. 161, n. 4, p. 861–868, 2009.

MATSUDA, F. et al. The complete nucleotide sequence of the human immunoglobulin heavy chain variable region locus. **Journal of Experimental Medicine**, v. 188, n. 11, p. 2151–2162, 1998.

MCBRIDE, O. et al. Chromosomal location of human kappa and lambda immunoglobulin light chain constant region genes. **The Journal of Experimental Medicine**, v. 155, n. 5, p. 1480–1490, 1 maio 1982.

MIQUEU, P. et al. Statistical analysis of CDR3 length distributions for the assessment of T and B cell repertoire biases. **Molecular Immunology**, v. 44, n. 6, p. 1057–1064, 2007.

MONTALBANO, A. et al. V(D)J Recombination Frequencies Can Be Profoundly Affected by Changes in the Spacer Sequence. **The Journal of Immunology**, v. 171, n. 10, p. 5296–5304, 2003.

MORAES, M. E. et al. An epitope in the third hypervariable region of the DRB1 gene is involved in the susceptibility to endemic pemphigus foliaceus (fogo selvagem) in three different Brazilian populations. **Tissue antigens**, v. 49, n. 1, p. 35–40, jan. 1997.

MURAMATSU, M. et al. **Class Switch Recombination and Hypermutation Require Activation-Induced Cytidine Deaminase (AID), a Potential RNA Editing EnzymeCell**. [s.l: s.n.].

MURPHY, K.; WEAVER, C. **Janeway's immunobiology**. 9th editio ed. [s.l.] Garland Science, 2016.

NAKA, K.; HIRAO, A. Regulation of hematopoiesis and hematological disease by TGF-β family signaling molecules. **Cold Spring Harbor Perspectives in Biology**, v. 9, n. 9, p. 25, 2017.

NAMBOODIRI, A. M.; PANDEY, J. P. The human cytomegalovirus TRL11/IRL11-encoded FcγR binds differentially to allelic variants of immunoglobulin G1. **Archives of Virology**, v. 156, n. 5, p. 907–910, 11 maio 2011.

NEMAZEE, D. Mechanisms of central tolerance for B cells. **Nature Reviews Immunology**, v. 17, n. 5, p. 281–294, 2017.

NOSSAL, G. J. V. The double helix and immunology. **Nature**, v. 421, n. 6921, p. 440–444, jan. 2003.

O'GARRA, A.; VIEIRA, P. Regulatory T cells and mechanisms of immune system control. **Nature Medicine**, v. 10, n. 8, p. 801–805, 2004.

OLIVEIRA, L. C. et al. Complement Receptor 1 (CR1, CD35) Polymorphisms and Soluble CR1: A Proposed Anti-inflammatory Role to Quench the Fire of "Fogo Selvagem" Pemphigus Foliaceus. **Frontiers in Immunology**, v. 10, n. November, p. 1–15, 22 nov. 2019.

OUTTERS, P. et al. Long-Range Control of V(D)J Recombination & Allelic Exclusion: Modeling Views. **Advances in Immunology**, v. 128, n. D, p. 363–413, 2015.

PALLER, A. S.; MANCINI, A. J. Bullous Disorders of Childhood. In: **Hurwitz Clinical Pediatric Dermatology**. [s.l.] Elsevier, 2011. p. 303–320.

PAN, M.; ZHU, H.; XU, R. Immune cellular regulation on autoantibody production in pemphigus. **The Journal of Dermatology**, v. 42, n. 1, p. 11–17, 2015.

PANDEY, J. P. et al. Association between immunoglobulin allotypes and immune responses to Haemophilus influenzae and Meningococcus polysaccharides. **Lancet (London, England)**, v. 1, n. 8109, p. 190–2, 27 jan. 1979.

PANDEY, J. P. et al. Immunoglobulin G heavy chain (Gm) allotypes in multiple sclerosis. **Journal of Clinical Investigation**, v. 67, n. 6, p. 1797–1800, 1981.

PANDEY, J. P. et al. Genetic markers of immunoglobulin G and susceptibility to breast cancer. **Human Immunology**, v. 73, n. 11, p. 1155–1158, nov. 2012.

PANDEY, J. P.; LI, Z. The forgotten tale of immunoglobulin allotypes in cancer risk and treatment. **Experimental Hematology & Oncology**, v. 2, n. 1, p. 1–7, 2013.

PANDEY, J. P.; NAMBOODIRI, A. M.; ELSTON, R. C. Immunoglobulin G genotypes and the risk of schizophrenia. **Human Genetics**, n. Pandey 2014, 8 jul. 2016.

PANELIUS, J.; MERI, S. Complement system in dermatological diseases - fire under the skin. **Frontiers in medicine**, v. 2, n. January, p. 3, 29 jan. 2015.

PARKER, D. C. T Cell-Dependent B Cell Activation. **Annual Review of Immunology**, v. 11, n. 1, p. 331–360, abr. 1993.

PAVONI, D. P. et al. Dissecting the associations of endemic pemphigus foliaceus (Fogo Selvagem) with HLA-DRB1 alleles and genotypes. **Genes and immunity**, v. 4, n. 2, p. 110–6, mar. 2003.

PELANDA, R.; TORRES, R. M. Central B-Cell tolerance: Where selection begins. **Cold Spring Harbor Perspectives in Biology**, v. 4, n. 4, p. 1–16, 2012.

PENG, B. et al. Identification of a primary antigenic target of epitope spreading in endemic pemphigus foliaceus. **Journal of Autoimmunity**, n. August, p. 102561, 2020.

PEREIRA, N. F. et al. Cytokine gene polymorphisms in endemic pemphigus foliaceus: A possible role for IL6 variants. **Cytokine**, v. 28, p. 233–241, 2004.

PETZL-ERLER, M. L. Beyond the HLA polymorphism: A complex pattern of genetic susceptibility to pemphigus. **Genetics and Molecular Biology**, v. 43, n. 3, p. 1–26, 2020.

PIOVEZAN, B. Z.; PETZL-ERLER, M. L. Both qualitative and quantitative genetic variation of MHC class II molecules may influence susceptibility to autoimmune diseases: the case of endemic pemphigus foliaceus. **Human immunology**, v. 74, n. 9, p. 1134–40, set. 2013.

PISANTI, S. et al. Pemphigus vulgaris: Incidence in Jews of different ethnic groups, according to age, sex, and initial lesion. **Oral Surgery, Oral Medicine, Oral Pathology**, v. 38, n. 3, p. 382–387, set. 1974.

PROBST, C. M. et al. HLA polymorphism and evaluation of European, African, and Amerindian contribution to the white and mulatto populations from Paraná, Brazil. **Human biology**, v. 72, n. 4, p. 597–617, ago. 2000.

QIAN, Y. et al. Antigen selection of Anti-DSG1 autoantibodies during and before the onset of endemic pemphigus foliaceus. **Journal of Investigative Dermatology**, v. 129, n. 12, p. 2823–2834, 2009.

RAJEWSKY, K.; FÖRSTER, I.; CUMANO, A. Evolutionary and somatic selection of the antibody repertoire in the mouse. **Science (New York, N.Y.)**, v. 238, n. 4830, p. 1088–94, 1987.

RIZZO, C. et al. Direct characterization of human T cells in pemphigus vulgaris reveals elevated autoantigen-specific Th2 activity in association with active disease. **Clinical and Experimental Dermatology**, v. 30, n. 5, p. 535–540, 2005.

ROGOZIN, I. B.; DIAZ, M. Deaminase-Triggered Process a Two-Step Activation-Induced Cytidine RGYW/WRCY Motif and Probably Reflects Hypermutation Than the Widely Accepted Predictor of Mutability at G:C Bases in Ig Cutting Edge: DGYW/WRCH Is a Better. 2004.

ROSCOE, J. T. et al. Brazilian pemphogus foliaceus autoantibodies are pathogenic to BALB/c mice by passive transfer. **Journal of Investigative Dermatology**, v. 85, n. 6, p. 538–541, 1985.

RUOCCO, E. et al. Viruses and pemphigus: An intriguing never-ending story. **Dermatology**, v. 229, n. 4, p. 310–315, 2014.

SALVIANO-SILVA, A.; PETZL-ERLER, M. L.; BOLDT, A. B. W. CD59 polymorphisms are associated with gene expression and different sexual susceptibility to pemphigus foliaceus. **Autoimmunity**, v. 6934, n. June, p. 1–9, 23 maio 2017.

SANTI, C. G.; SOTTO, M. N. Immunopathologic characterization of the tissue response in endemic pemphigus foliaceus (fogo selvagem). **Journal of the American Academy of Dermatology**, v. 44, n. 3, p. 446–450, 2001.

SATYAM, A. et al. Involvement of TH1/TH2 cytokines in the pathogenesis of autoimmune skin diseasepemphigus vulgaris. **Immunological Investigations**, v. 38, n. 6, p. 498–509, 2009.

SCHATZ, D. G.; JI, Y. Recombination centres and the orchestration of V(D)J recombination. **Nature Reviews Immunology**, v. 11, n. 4, p. 251–263, 11 abr. 2011.

SCHMIDT, E.; KASPERKIEWICZ, M.; JOLY, P. Pemphigus. **The Lancet**, v. 394, n. 10201, p. 882–894, 2019a.

SCHMIDT, E.; KASPERKIEWICZ, M.; JOLY, P. Pemphigus. **Lancet**, v. 394, n. 10201, p. 882–894, set. 2019b.

SCHROEDER, H. W.; CAVACINI, L. Structure and function of immunoglobulins. **Journal of Allergy and Clinical Immunology**, v. 125, n. 2, p. S41–S52, fev. 2010.

SFIKAKIS, P. P. et al. Clonal expansion of B-cells in human systemic lupus erythematosus: Evidence from studies before and after therapeutic B-cell depletion. **Clinical Immunology**, v. 132, n. 1, p. 19–31, 2009.

SHUGAY, M. et al. Towards error-free profiling of immune repertoires. **Nature Methods**, v. 11, n. 6, p. 653–655, 2014.

STAVNEZER, J.; GUIKEMA, J. E. J.; SCHRADER, C. E. Mechanism and Regulation of Class Switch Recombination. **Annual Review of Immunology**, v. 26, n. 1, p. 261–292, abr. 2008.

STEINBERG, A.; MORELL, A. The effect of Gm (23) on the concentration of IgG2 and IgG4 in normal human serum. **The Journal of …**, n. 23, 1973.

SUGIYAMA, H. et al. CD4+CD25high regulatory T cells are markedly decreased in blood of patients with pemphigus vulgaris. **Dermatology**, v. 214, n. 3, p. 210–220, 2007.

TRON, F. et al. Genetic factors in pemphigus. **Journal of autoimmunity**, v.

24, n. 4, p. 319–28, jun. 2005.

TURCHANINOVA, M. A. et al. High-quality full-length immunoglobulin profiling with unique molecular barcoding. **Nature Protocols**, v. 11, n. 9, p. 1599–1616, 4 set. 2016.

VÁZQUEZ BERNAT, N. et al. High-Quality Library Preparation for NGS-Based Immunoglobulin Germline Gene Inference and Repertoire Expression Analysis. **Frontiers in Immunology**, v. 10, n. 9, p. 1599–1616, 5 abr. 2019.

VELDMAN, C. et al. Dichotomy of Autoreactive Th1 and Th2 Cell Responses to Desmoglein 3 in Patients with Pemphigus Vulgaris (PV) and Healthy Carriers of PV-Associated HLA Class II Alleles. **The Journal of Immunology**, v. 170, n. 1, p. 635–642, 2003.

VERNAL, S. et al. Pemphigus foliaceus and sand fly bites: assessing the humoral immune response to the salivary proteins maxadilan and LJM11. **British Journal of Dermatology**, p. 10–12, 2020a.

VERNAL, S. et al. Pemphigus foliaceus and sand fly bites: assessing the humoral immune response to the salivary proteins maxadilan and LJM11. **British Journal of Dermatology**, v. 183, n. 5, p. 958–960, 13 nov. 2020b.

VOS, Q. et al. B-cell activation by T-cell-independent type 2 antigens as an integral part of the humoral immune response to pathogenic microorganisms. **Immunological Reviews**, v. 176, n. 1, p. 154–170, ago. 2000.

WARDEMANN, H.; BUSSE, C. E. Novel Approaches to Analyze Immunoglobulin Repertoires. **Trends in Immunology**, v. 38, n. 7, p. 471–482, 2017.

WARREN, S. J. et al. The prevalence of antibodies against desmoglein 1 in endemic pemphigus foliaceus in Brazil. Cooperative Group on Fogo Selvagem Research. **The New England journal of medicine**, v. 343, n. 1, p. 23–30, 2000.

WATSON, C.; BREDEN, F. The immunoglobulin heavy chain locus: genetic variation, missing data, and implications for human disease. **Genes and Immunity**, v. 1312, n. 10, p. 363–373, 2012.

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**. [s.l.] The American Society of Human Genetics, 2013. v. 92

WHITTINGHAM, S. et al. Interaction of HLA and Gm in autoimmune chronic active hepatitis. **Clinical and experimental immunology**, v. 43, n. 1, p. 80–6, 1981.

WILSON, R. PLoS Pathogens Issue Image | Vol. 5(8) August 2009. **PLoS Pathogens**, v. 5, n. 8, p. ev05.i08, 28 ago. 2009.

XU, J. L.; DAVIS, M. M. Diversity in the CDR3 region of V(H) is sufficient for most antibody specificities. **Immunity**, v. 13, n. 1, p. 37–45, 2000.

XU, Z. et al. Immunoglobulin class-switch DNA recombination: induction, targeting and beyond. **Nature Reviews Immunology**, v. 12, n. 7, p. 517–531, 25 jul. 2012.

YAN, Q. et al. Next generation sequencing reveals novel alterations in B-cell heavy chain receptor repertoires associated with acute-on-chronic liver failure. **International Journal of Molecular Medicine**, v. 43, n. 1, p. 243–255, 2019.

YANCOPOULOS, G. D.; ALT, F. W. Developmentally controlled and tissue-specific expression of unrearranged VH gene segments. **Cell**, v. 40, n. 2, p. 271–281, 1985.

ZHANG, F.; LUPSKI, J. R. Non-coding genetic variants in human disease. **Human Molecular Genetics**, v. 24, n. R1, p. R102–R110, 2015.

ZHANG, W. et al. PIRD: Pan Immune Repertoire Database. **Bioinformatics**, v. 36, n. 3, p. 897–903, 2020.

APENDICE 1

Table S1: List of candidate genes considered in the association analysis (continues)

Process involved	Gene symbol	Chromosomal location	Chromosome	Gene position (GRCh37.p13)	Gene size	SNP in microarray	SNPs filtering
Immunoglobulin genes	IGH	14q32.33	14	(106022614-107298051)	1275437	35	12
Immunoglobulin genes	IGK	2p11.2	2	(89156874-90274235)	1117361	11	0
Immunoglobulin genes	IGL	22q11.2	22	(22380474-23265085)	884611	61	25
V(D)J Recombination	ATM	11q22.3	11	(108093559-108239829)	146270	108	2
V(D)J Recombination	BRCA1	17q21.31	17	(41196312-41277500)	81188	78	2
V(D)J Recombination	CSNK2A2	16q21	16	(58191811-58231782)	39971	5	3
V(D)J Recombination	DCLRE1C	10p13	10	(14948870-14996106)	47236	19	0
V(D)J Recombination	DNTT	10q24.1	10	(98064085-98098321)	34236	22	2
V(D)J Recombination	H2AX	11q23.3	11	(118964580-118966177)	1597	7	1
V(D)J Recombination	LIG4	13q33.3	13	(108859790-108870716)	10926	28	2
V(D)J Recombination	MRE11	11q21	11	(94150466-94227040)	76574	30	4
V(D)J Recombination	NBN	8q21.3	8	(90945564-90996899)	51335	34	2
V(D)J Recombination	NHEJ1	2q35	2	(219940046-220025587)	85541	23	2
V(D)J Recombination	PAXX	9q34.3	9	(139886870-139888428)	1558	8	1
V(D)J Recombination	POLL	10q24.32	10	(103338639-103348027)	9388	20	1
V(D)J Recombination	POLM	7p13	7	(44111846-44122129)	10283	13	3
V(D)J Recombination	PRKDC	8q11.21	8	(48685669-48872743)	187074	70	1
V(D)J Recombination	RAD50	5q31.1	5	(131892616-131980313)	87697	43	0
V(D)J Recombination	RAD51	15q15.1	15	(40987327-41024356)	37029	8	1
V(D)J Recombination	RAG1	11p12	11	(36589563-36601312)	11749	37	1
V(D)J Recombination	RAG2	11p13	11	(36613493-36619829)	6336	21	1
V(D)J Recombination	XRCC4	5q14.2	5	(82373317-82649579)	276262	30	8
V(D)J Recombination	XRCC5	2q35	2	(216972381-217071016)	98635	23	8
V(D)J Recombination	XRCC6	22q13.2	22	(42017250-42060052)	42802	4	0
SHM and CSR	AICDA	12p13.31	12	(8754762-8765463)	10701	9	2
SHM and CSR	APEX1	14q11.2	14	(20923290-20925931)	2641	23	1
SHM and CSR	APEX2	Xp11.21	X	(55026756-55034306)	7550	14	1
SHM and CSR	APLF	2p13.3	2	(68694691-68807294)	112603	30	1
SHM and CSR	APOBEC1	12p13.31	12	(7801996-7818502)	16506	18	4
SHM and CSR	CHEK2	22q12.1	22	(29083731-29137822)	54091	25	4
SHM and CSR	ERCC1	19q13.32	19	(45910591-45927177)	16586	32	3
SHM and CSR	ERCC4	16p13.12	16	(14014014-14046205)	32191	34	2
SHM and CSR	MSH2	2p21-p16.3	2	(47630206-47710367)	80161	25	2
SHM and CSR	MSH6	2p16.3	2	(48010221-48034092)	23871	35	1
SHM and CSR	PAGR1	16p11.2	16	(29827528-29833816)	6288	6	1
SHM and CSR	PAXIP1	7q36.2	7	(154735400-154794682)	59282	18	5
SHM and CSR	POLB	8p11.21	8	(42195973-42229331)	33358	14	0
SHM and CSR	POLH	6p21.1	6	(43543878-43588260)	44382	29	0
SHM and CSR	REV1	2q11.2	2	(100016938-100106480)	89542	34	4
SHM and CSR	RPA1	17p13.3	17	(1733273-1802848)	69575	26	8
SHM and CSR	RPA2	1p35.3	1	(28218035-28241308)	23273	6	1
SHM and CSR	RPA3	7p21.3	7	(7676149-7758238)	82089	12	7
SHM and CSR	RPAIN	17p13.2	17	(5322961-5336340)	13379	25	3

Table S1: List of candidate genes considered in the association analysis (continued)

Process involved	Gene symbol	Chromosomal location	Chromosome	Gene position (GRCh37.p13)	Gene size	SNP in microarray	SNPs filtering
SHM and CSR	SUPT16H	14q11.2	14	(21819631-21852425)	32794	7	0
SHM and CSR	SUPT5H	19q13.2	19	(39936186-39967310)	31124	8	3
SHM and CSR	TP53	17p13.1	17	(7571720-7590868)	19148	24	2
SHM and CSR	UNG	12q24.11	12	(109535399-109548798)	13399	11	1
CSR regulation	ATF2	2q31.1	2	(175936978-176032934)	95956	9	2
CSR regulation	BACH2	6q15	6	(90636247-91006627)	370380	59	21
CSR regulation	BATF	14q24.3	14	(75988784-76013335)	24551	4	1
CSR regulation	BCL6	3q27.3	3	(187439165-187463513)	24348	15	1
CSR regulation	CD14	5q31.3	5	(140011313-140013286)	1973	7	2
CSR regulation	CD40	20q13.12	20	(44746893-44758384)	11491	7	0
CSR regulation	CD40LG	Xq26.3	X	(135730336-135742549)	12213	8	1
CSR regulation	CD79A	19q13.2	19	(42381190-42385439)	4249	7	1
CSR regulation	CD79B	17q23.3	17	(62006096-62009714)	3618	5	2
CSR regulation	CD80	3q13.33	3	(119243140-119278481)	35341	21	4
CSR regulation	CD86	3q13.33	3	(121774209-121839990)	65781	16	4
CSR regulation	CEBPB	20q13.13	20	(48801140-48808606)	7466	2	0
CSR regulation	CHUK	10q24.31	10	(101948055-101989367)	41312	12	3
CSR regulation	CREBBP	16p13.3	16	(3775055-3930121)	155066	38	1
CSR regulation	CUX1	7q22.1	7	(101459184-101927250)	468066	73	25
CSR regulation	E2F1	20q11.22	20	(32263292-32274210)	10918	11	1
CSR regulation	E2F2	1p36.12	1	(23832920-23857712)	24792	16	3
CSR regulation	E2F3	6p22.3	6	(20402137-20493945)	91808	16	6
CSR regulation	E2F4	16q22.1	16	(67226068-67232821)	6753	21	1
CSR regulation	E2F5	8q21.2	8	(86089619-86126753)	37134	8	1
CSR regulation	E2F6	2p25.1	2	(11584501-11606303)	21802	3	1
CSR regulation	E2F7	12q21.2	12	(77415026-77459360)	44334	30	2
CSR regulation	ELF1	13q14.11	13	(41506055-41593508)	87453	31	2
CSR regulation	EP300	22q13.2	22	(41488614-41576081)	87467	52	3
CSR regulation	ETS1	11q24.3	11	(128328656-128457453)	128797	27	13
CSR regulation	EXO1	1q43	1	(242011491-242053241)	41750	54	8
CSR regulation	FCER2	19p13.2	19	(7753643-7767032)	13389	21	4
CSR regulation	FOS	14q24.3	14	(75745481-75748937)	3456	5	0
CSR regulation	GAB1	4q31.21	4	(144257983-144395718)	137735	14	6
CSR regulation	HNRNPDL	4q21.22	4	(83343717-83351378)	7661	6	0
CSR regulation	HOXC4	12q13.13	12	(54388716-54449814)	61098	23	6
CSR regulation	ICOS	2q33.2	2	(204801471-204826300)	24829	7	2
CSR regulation	ICOSLG	21q22.3	21	(45642874-45660887)	18013	17	4
CSR regulation	ID1	20q11.21	20	(30193086-30194318)	1232	6	1
CSR regulation	ID2	2p25.1	2	(8822113-8824583)	2470	1	1
CSR regulation	ID3	1p36.12	1	(23884421-23886285)	1864	7	1
CSR regulation	IFNG	12q15	12	(68548550-68553521)	4971	2	0
CSR regulation	IKBKB	8p11.21	8	(42128820-42190171)	61351	17	1
CSR regulation	IKBKG	Xq28	X	(153770459-153793261)	22802	4	0

Table S1: List of candidate genes considered in the association analysis (continued)

Process involved	Gene symbol	Chromosomal location	Chromosome	Gene position (GRCh37.p13)	Gene size	SNP in microarray	SNPs filtering
CSR regulation	IKZF1	7p12.2	7	(50343679-50472799)	129120	20	8
CSR regulation	IKZF2	2q34	2	(213864408-214016333)	151925	22	8
CSR regulation	IKZF3	17q12-q21.1	17	(37913968-38020441)	106473	11	3
CSR regulation	IKZF4	12q13.2	12	(56411922-56432219)	20297	15	1
CSR regulation	IKZF5	10q26.13	10	(124750322-124768366)	18044	8	2
CSR regulation	IL10	1q32.1	1	(206940948-206945839)	4891	9	2
CSR regulation	IL13	5q31.1	5	(131993865-131996801)	2936	8	4
CSR regulation	IL15	4q31.21	4	(142557749-142655140)	97391	10	3
CSR regulation	IL4	5q31.1	5	(132009678-132018370)	8692	5	1
CSR regulation	IL5	5q31.1	5	(131877136-131892555)	15419	6	0
CSR regulation	INPP5D	2q37.1	2	(233924677-234116549)	191872	44	15
CSR regulation	IRF1	5q31.1	5	(131776132-131826465)	50333	15	9
CSR regulation	IRF4	6p25.3	6	(391739-411443)	19704	17	5
CSR regulation	JAK1	1p31.3	1	(65298906-65533429)	234523	36	14
CSR regulation	JAK3	19p13.11	19	(17935591-17958841)	23250	27	2
CSR regulation	JUN	1p32.1	1	(59246463-59249785)	3322	3	1
CSR regulation	MAFK	7p22.3	7	(1570368-1582679)	12311	6	4
CSR regulation	MAP3K14	17q21.31	17	(43340486-43394430)	53944	17	5
CSR regulation	MDC1	6p21.33	6	(30667584-30685458)	17874	62	2
CSR regulation	MLH1	3p22.2	3	(37034841-37092337)	57496	41	4
CSR regulation	MYB	6q23.3	6	(135502453-135540311)	37858	19	4
CSR regulation	NCL	2q37.1	2	(232319459-232329208)	9749	17	3
CSR regulation	NFKB1	4q24	4	(103422486-103538459)	115973	32	4
CSR regulation	PARP1	1q42.12	1	(226548392-226595801)	47409	28	3
CSR regulation	PAX5	9p13.2	9	(36833272-37034476)	201204	40	24
CSR regulation	PIK3R1	5q13.1	5	(67511584-67597649)	86065	24	11
CSR regulation	PMS2	7p22.1	7	(6012870-6048737)	35867	23	0
CSR regulation	POU2AF1	11q23.1	11	(111222981-111250157)	27176	11	3
CSR regulation	POU2F1	1q24.2	1	(167190066-167396582)	206516	16	3
CSR regulation	POU2F2	19q13.2	19	(42590262-42636625)	46363	4	1
CSR regulation	PRDM1	6q21	6	(106534195-106557814)	23619	22	2
CSR regulation	PRKACA	19p13.1	19	(14202500-14228559)	26059	6	0
CSR regulation	PTPRC	1q31.3-q32.1	1	(198608098-198726605)	118507	48	9
CSR regulation	RBBP8	18q11.2	18	(20513295-20606451)	93156	22	2
CSR regulation	RELA	11q13	11	(65421067-65430443)	9376	14	0
CSR regulation	RELB	19q13.32	19	(45504707-45541456)	36749	11	1
CSR regulation	RUNX1	21q22.12	21	(36160098 -36421631)	261533	31	20
CSR regulation	RUNX2	6p21.1	6	(45296054-45518819)	222765	23	10
CSR regulation	RUNX3	1p36.11	1	(25226002-25291648)	65646	8	4
CSR regulation	SFN	1p36.11	1	(27189787-27190449)	662	3	0
CSR regulation	SLC22A2	6q25.3	6	(160637794-160679963)	42169	43	3
CSR regulation	SMAD3	15q22.33	15	(67358036-67487533)	129497	21	14
CSR regulation	SMAD4	18q21.2	18	(48556583-48611412)	54829	7	2

Table S1: List of candidate genes considered in the association analysis (continued)

Process involved	Gene symbol	Chromosomal location	Chromosome	Gene position (GRCh37.p13)	Gene size	SNP in microarray	SNPs filtering
CSR regulation	SMAD7	18q21.1	18	(46446223-46477081)	30858	12	6
CSR regulation	SP1	12q13.13	12	(53773979-53810226)	36247	10	1
CSR regulation	SPI1	11p11.2	11	(47376409-47400127)	23718	3	1
CSR regulation	STAT1	2q32.2	2	(191833762-191878976)	45214	12	4
CSR regulation	STAT6	12q13.3	12	(57489187-57522883)	33696	14	2
CSR regulation	SWAP70	11p15.4	11	(9685624-9774538)	88914	12	6
CSR regulation	TCF3	19p13.3	19	(1609289-1652328)	43039	24	1
CSR regulation	TFCP2	12q13.12-q13.13	12	(51487539-51566926)	79387	14	4
CSR regulation	TGFB1	19q13.2	19	(41836812-41859831)	23019	7	2
CSR regulation	TGFB2	1q41	1	(218518676-218617961)	99285	16	9
CSR regulation	TGFB3	14q24	14	(76424442-76448092)	23650	7	1
CSR regulation	TGFB4	1q42.12	1	(226124298-226129083)	4785	5	1
CSR regulation	TGFBR1	9q22.33	9	(101866320-101916474)	50154	7	1
CSR regulation	TGFBR2	3p24.1	3	(30647994-30735634)	87640	38	14
CSR regulation	TLR1	4p14	4	(38797876-38806412)	8536	33	1
CSR regulation	TLR2	4q31.3	4	(154605441-154627243)	21802	36	3
CSR regulation	TLR3	4q35.1	4	(186990309-187006252)	15943	30	5
CSR regulation	TLR4	9q33.1	9	(120466453-120479769)	13316	30	1
CSR regulation	TLR5	1q41	1	(223282748-223316624)	33876	38	2
CSR regulation	TLR6	4p14	4	(38825325-38858438)	33113	30	2
CSR regulation	TLR7	Xp22.2	X	(12885202-12908480)	23278	11	1
CSR regulation	TLR8	Xp22.2	X	(12924739-12941288)	16549	10	3
CSR regulation	TLR9	3p21.2	3	(52255096-52260179)	5083	21	1
CSR regulation	TNFRSF13B	17p11.2	17	(16842398-16875402)	33004	29	4
CSR regulation	TNFRSF8	1p36.22	1	(12123434-12204264)	80830	24	8
CSR regulation	TNFSF13	17p13.1	17	(7461609-7464925)	3316	6	1
CSR regulation	TNFSF13B	13q33.3	13	(108921977-108960832)	38855	6	2
CSR regulation	TP53BP1	15q15.3	15	(43699412-43785354)	85942	35	1
CSR regulation	TRAF1	9q33.2	9	(123664671-123691451)	26780	15	2
CSR regulation	TRAF2	9q34.3	9	(139776385-139821067)	44682	14	3
CSR regulation	TRAF3	14q32.32	14	(103243816-103377837)	134021	14	7
CSR regulation	TRAF6	11p12	11	(36505317-36531863)	26546	6	1
CSR regulation	YWHAB	20q13.12	20	(43514240-43537173)	22933	5	2
CSR regulation	YWHAE	17p13.3	17	(1247834-1303556)	55722	6	5
CSR regulation	YWHAG	7q11.23	7	(75956108-75988342)	32234	7	3
CSR regulation	YWHAH	22q12.3	22	(32340479-32353590)	13111	3	2
CSR regulation	YWHAQ	2p25.1	2	(9724106-9771106)	47000	5	3
CSR regulation	YWHAZ	8q22.3	8	(101930804-101965221)	34417	2	2

Table S3: Single nucleotide polymorphism of the candidate genes selected analysed in the association study (Continues)

Chr	SNP	Ref	Alt	Gene	Frequency of reference allele		Chr	SNP	Ref	Alt	Gene	Frequency of reference allele	
					Patients	Controls						Patients	Controls
1	rs535068	A	G	TNFRSF8	0.7327	0.6641	1	rs4656538	G	A	POU2F1	0.4247	0.4237
1	rs6657275	G	A	TGFB2	0.4683	0.3854	1	rs6686725	A	C	PTPRC	0.4262	0.4211
2	rs6432018	C	A	YWHAQ	0.5181	0.3958	1	rs1326271	G	A	PTPRC	0.2068	0.2592
5	rs2070729	C	A	IRF1	0.5352	0.4418	1	exm2268809	A	G	PTPRC	0.475	0.4398
7	rs17149161	C	A	YWHAG	0.7569	0.6622	1	rs9803750	A	G	PTPRC	0.3909	0.3571
9	rs10870140	G	A	TRAF2	0.7579	0.6875	1	rs2359952	G	A	PTPRC	0.3773	0.3842
9	rs10781522	G	A	TRAF2	0.3721	0.4896	1	rs16843742	G	A	PTPRC	0.2156	0.1925
9	rs10781530	A	C	PAXX	0.7106	0.6223	1	rs12120762	G	A	PTPRC	0.4253	0.4115
11	rs1818545	A	G	RAG2	0.4124	0.3053	1	rs7540378	A	G	PTPRC	0.214	0.2367
12	rs324011	A	G	STAT6	0.3866	0.2932	1	rs11589894	G	A	PTPRC	0.3227	0.3307
1	rs12029016	A	G	TNFRSF8	0.3009	0.3263	1	rs1800871	A	G	IL10	0.3795	0.3691
1	rs6690493	A	G	TNFRSF8	0.4118	0.4193	1	rs1800896	G	A	IL10	0.3047	0.3325
1	rs12092682	A	C	TNFRSF8	0.325	0.2656	1	rs1418555	A	G	TGFB2	0.2626	0.3037
1	rs501525	G	A	TNFRSF8	0.3273	0.3333	1	rs1417488	A	G	TGFB2	0.2878	0.3398
1	rs17422381	A	G	TNFRSF8	0.3891	0.474	1	rs6662137	C	A	TGFB2	0.35	0.3717
1	rs1201122	G	A	TNFRSF8	0.2257	0.2888	1	rs2027566	C	A	TGFB2	0.4683	0.4267
1	rs630542	A	C	TNFRSF8	0.4205	0.4215	1	rs1539399	A	G	TGFB2	0.4163	0.362
1	rs2282720	G	A	E2F2	0.4386	0.5026	1	rs2799083	G	A	TGFB2	0.2683	0.3032
1	exm31057	A	C	E2F2	0.3945	0.445	1	rs2000220	A	G	TGFB2	0.4	0.466
1	rs2038027	A	G	E2F2	0.4562	0.4245	1	rs1342586	A	G	TGFB2	0.2773	0.2251
1	rs1050096	G	A	ID3	0.3073	0.3194	1	exm151500	G	A	TLR5	0.3167	0.3594
1	rs2236852	A	G	RUNX3	0.5115	0.474	1	rs2096142	A	G	TLR5	0.3688	0.3958
1	rs742230	G	A	RUNX3	0.4208	0.3776	1	rs2493163	A	G	TGFB4	0.2489	0.2292
1	rs11249206	G	A	RUNX3	0.4606	0.4921	1	rs3219110	G	A	PARP1	0.4201	0.4297
1	rs1848185	A	G	RUNX3	0.2727	0.2786	1	rs3219090	A	G	PARP1	0.3945	0.3594
1	rs2474470	A	G	RPA2	0.4525	0.4529	1	rs907190	A	C	PARP1	0.2466	0.1927
1	exm2268894	C	A	JUN	0.2511	0.2135	1	rs1776139	C	A	EXO1	0.45	0.5
1	rs1048007	A	G	JAK1	0.4502	0.5393	1	rs735943	A	G	EXO1	0.3535	0.328
1	rs2780815	A	C	JAK1	0.3348	0.4089	1	exm164070	A	G	EXO1	0.1843	0.2188
1	rs310219	A	G	JAK1	0.3535	0.2989	1	rs1635501	G	A	EXO1	0.3756	0.3646
1	rs310209	A	C	JAK1	0.4208	0.3802	1	exm164096	A	G	EXO1	0.4072	0.401
1	rs310202	G	A	JAK1	0.3932	0.3037	1	exm164114	A	G	EXO1	0.3122	0.3099
1	exm-rs310199	G	A	JAK1	0.509	0.4219	1	rs4150005	G	A	EXO1	0.217	0.2027
1	rs10889503	A	C	JAK1	0.4299	0.3842	1	exm2273350	A	G	EXO1	0.2903	0.254
1	rs7519042	A	G	JAK1	0.2534	0.2132	2	rs4669330	A	G	ID2	0.3982	0.3246
1	rs4916014	G	A	JAK1	0.4593	0.4084	2	rs7608161	G	A	YWHAQ	0.2398	0.2943
1	rs4915675	G	A	JAK1	0.3922	0.3325	2	rs6432025	A	G	YWHAQ	0.314	0.3825
1	rs7545743	G	A	JAK1	0.2959	0.2801	2	rs4292073	C	A	E2F6	0.4658	0.4136
1	rs535857	G	A	JAK1	0.2851	0.2292	2	rs7602094	G	A	MSH2	0.362	0.3333
1	rs11208574	A	G	JAK1	0.3182	0.2578	2	rs3136228	C	A	MSH2	0.2805	0.2839
1	rs6588109	G	C	JAK1	0.4521	0.3663	2	rs1800935	G	A	MSH6	0.2182	0.2161
1	rs12068411	A	G	POU2F1	0.2557	0.2708	2	rs6546420	A	G	APLF	0.2188	0.1854
1	rs10800299	G	A	POU2F1	0.3394	0.3403	2	rs1451245	A	G	REV1	0.2254	0.2184

Table S3: Single nucleotide polymorphism of the candidate genes selected analysed in the association study (Continued)

Chr	SNP	Ref	Alt	Gene	Frequency of reference allele		Chr	SNP	Ref	Alt	Gene	Frequency of reference allele	
					Patients	Controls						Patients	Controls
2	rs3087403	A	G	REV1	0.2919	0.2943	2	rs7569837	G	A	INPP5D	0.2886	0.3385
2	rs769105	G	A	REV1	0.4887	0.4948	2	rs6740918	A	G	INPP5D	0.2262	0.1979
2	rs28369942	D	I	REV1	0.3825	0.377	2	rs10203185	A	G	INPP5D	0.3614	0.3099
2	rs212347	G	A	ATF2	0.2023	0.2005	2	rs9247	A	G	INPP5D	0.2295	0.1719
2	rs13388308	A	G	ATF2	0.2546	0.2592	3	rs6550005	A	G	TGFBR2	0.2421	0.1979
2	rs3771300	A	C	STAT1	0.4615	0.4505	3	rs1835538	A	G	TGFBR2	0.3995	0.4545
2	rs11894425	G	A	STAT1	0.3507	0.3464	3	rs9850060	G	A	TGFBR2	0.2671	0.3307
2	rs34997637	A	G	STAT1	0.2182	0.2031	3	rs4522809	A	G	TGFBR2	0.4955	0.4505
2	rs2030171	A	G	STAT1	0.4977	0.4634	3	rs11924422	C	A	TGFBR2	0.3795	0.401
2	exm-rs4675374	A	G	ICOS	0.31	0.3646	3	rs13075948	A	G	TGFBR2	0.2647	0.2891
2	rs10183087	C	A	ICOS	0.3568	0.3698	3	rs1155705	G	A	TGFBR2	0.3318	0.3228
2	rs17277107	A	G	IKZF2	0.2466	0.2448	3	rs2082224	A	G	TGFBR2	0.2844	0.2487
2	rs7607184	A	G	IKZF2	0.4514	0.4632	3	rs995435	A	G	TGFBR2	0.3386	0.2995
2	rs7585103	C	A	IKZF2	0.363	0.3842	3	rs749794	G	A	TGFBR2	0.3532	0.3613
2	rs1482579	G	A	IKZF2	0.4636	0.4791	3	rs11466511	C	A	TGFBR2	0.2432	0.2057
2	rs10203838	G	A	IKZF2	0.4369	0.4397	3	rs3773649	A	G	TGFBR2	0.3493	0.3594
2	rs10497991	A	G	IKZF2	0.3659	0.3553	3	rs1346907	A	G	TGFBR2	0.4186	0.3984
2	rs6435757	G	A	IKZF2	0.2949	0.3168	3	rs744751	A	G	TGFBR2	0.1918	0.2251
2	exm2261164	G	A	IKZF2	0.2545	0.2079	3	rs1800734	A	G	MLH1	0.2133	0.2289
2	rs828907	A	C	XRCC5	0.3226	0.3386	3	rs1799977	G	A	MLH1	0.242	0.2827
2	rs828910	G	A	XRCC5	0.4005	0.4557	3	rs1558529	G	A	MLH1	0.3932	0.3958
2	rs207878	G	A	XRCC5	0.325	0.3403	3	rs10849	G	A	MLH1	0.4887	0.4635
2	rs2160981	G	A	XRCC5	0.4886	0.4005	3	exm2269449	A	G	TLR9	0.457	0.4167
2	rs3821107	G	A	XRCC5	0.2262	0.2526	3	rs7628626	A	C	CD80	0.2315	0.2
2	rs207922	A	G	XRCC5	0.2324	0.2527	3	rs4688014	A	G	CD80	0.4159	0.3947
2	rs207946	G	A	XRCC5	0.4045	0.4297	3	rs491407	G	A	CD80	0.2059	0.233
2	rs9288518	G	A	XRCC5	0.4864	0.4555	3	rs7648642	C	A	CD80	0.4292	0.4555
2	rs897477	G	A	NHEJ1	0.2896	0.263	3	rs17203397	G	A	CD86	0.4295	0.4058
2	rs897476	G	A	NHEJ1	0.4321	0.4476	3	exm2264034	C	A	CD86	0.375	0.401
2	rs10180468	A	G	NCL	0.2215	0.2552	3	rs2681420	G	A	CD86	0.1925	0.2196
2	rs7598759	A	G	NCL	0.4521	0.4818	3	exm342618	A	G	CD86	0.2146	0.2605
2	rs4973410	G	A	NCL	0.4548	0.4476	3	rs1005099	C	A	BCL6	0.5045	0.4948
2	rs4973063	A	G	INPP5D	0.4587	0.4505	4	rs5743551	A	G	TLR1	0.5023	0.4766
2	rs4439944	G	A	INPP5D	0.3484	0.3429	4	rs6833914	A	G	TLR6	0.3091	0.3542
2	rs10933428	G	A	INPP5D	0.3659	0.388	4	rs6531673	A	G	TLR6	0.4171	0.3927
2	rs6437089	G	A	INPP5D	0.2591	0.2552	4	rs1599961	A	G	NFKB1	0.4429	0.4267
2	exm2269379	A	G	INPP5D	0.2172	0.2266	4	rs1598856	A	G	NFKB1	0.395	0.4089
2	rs7570061	A	G	INPP5D	0.1972	0.2184	4	rs230535	A	C	NFKB1	0.3539	0.3429
2	rs10193128	G	A	INPP5D	0.3529	0.3542	4	rs230529	A	G	NFKB1	0.4498	0.4289
2	rs11693862	G	A	INPP5D	0.4954	0.45	4	rs1519551	G	A	IL15	0.5182	0.4738
2	rs7419666	G	A	INPP5D	0.3145	0.3395	4	rs6850492	A	G	IL15	0.3447	0.3333
2	exm276259	A	G	INPP5D	0.3157	0.3272	4	exm2256380	G	A	IL15	0.4358	0.4684
2	exm276263	G	C	INPP5D	0.3477	0.3958	4	rs300913	G	A	GAB1	0.3529	0.3255

Table S3: Single nucleotide polymorphism of the candidate genes selected analysed in the association study (Continued)

Chr	SNP	Ref	Alt	Gene	Frequency of reference allele		Chr	SNP	Ref	Alt	Gene	Frequency of reference allele	
					Patients	Controls						Patients	Controls
4	rs300912	A	G	GAB1	0.457	0.4844	5	rs20541	A	G	IL13	0.2557	0.2734
4	rs1472873	G	A	GAB1	0.3507	0.2865	5	rs2243270	G	A	IL4	0.3219	0.3564
4	rs3805253	G	A	GAB1	0.5158	0.4635	5	rs2569190	A	G	CD14	0.4184	0.4755
4	rs3805236	A	G	GAB1	0.2715	0.3281	5	rs5744455	A	G	CD14	0.2308	0.1927
4	rs3795246	G	A	GAB1	0.3796	0.3989	6	rs2671422	A	G	IRF4	0.2489	0.2526
4	rs7696323	A	G	TLR2	0.2311	0.246	6	rs2001508	C	A	IRF4	0.2545	0.2552
4	rs11938228	A	C	TLR2	0.4055	0.3883	6	rs1877175	A	G	IRF4	0.2136	0.2214
4	rs7656411	C	A	TLR2	0.3382	0.3575	6	exm-rs872071	G	A	IRF4	0.3977	0.3906
4	rs7657186	A	G	TLR3	0.2477	0.2031	6	rs7757906	A	G	IRF4	0.3371	0.3325
4	rs13126816	A	G	TLR3	0.2306	0.2618	6	rs9368188	A	G	E2F3	0.3303	0.3333
4	rs3775291	A	G	TLR3	0.2568	0.2943	6	rs9295464	A	C	E2F3	0.2523	0.2656
4	exm2277278	A	C	TLR3	0.2886	0.276	6	rs2328488	A	G	E2F3	0.2202	0.2266
4	rs10025405	G	A	TLR3	0.3643	0.3568	6	rs6926668	G	A	E2F3	0.2431	0.2812
5	rs7701498	A	G	PIK3R1	0.2172	0.1911	6	rs12205236	A	C	E2F3	0.3096	0.349
5	rs10940157	A	G	PIK3R1	0.2477	0.276	6	rs1042317	A	G	E2F3	0.25	0.25
5	rs831227	A	C	PIK3R1	0.4931	0.4922	6	exm-rs2894043	G	C	MDC1	0.21	0.2263
5	rs7709243	G	A	PIK3R1	0.4295	0.4401	6	rs3132584	A	C	MDC1	0.1963	0.2249
5	rs34303	G	A	PIK3R1	0.4457	0.3901	6	rs7757282	G	A	RUNX2	0.3986	0.4371
5	rs34309	A	G	PIK3R1	0.2991	0.3194	6	rs1321081	A	G	RUNX2	0.4	0.3796
5	rs2302975	G	A	PIK3R1	0.4749	0.4392	6	rs1321080	A	C	RUNX2	0.2052	0.235
5	rs7716675	G	A	PIK3R1	0.2159	0.2199	6	rs9296459	A	G	RUNX2	0.242	0.2763
5	rs3815701	G	A	PIK3R1	0.2511	0.1927	6	rs1321075	A	C	RUNX2	0.2104	0.2068
5	rs3730089	A	G	PIK3R1	0.2217	0.2042	6	rs2819854	A	G	RUNX2	0.4932	0.4922
5	rs3756668	G	A	PIK3R1	0.3955	0.487	6	rs910586	A	G	RUNX2	0.3122	0.3073
5	rs1478486	A	G	XRCC4	0.4037	0.4372	6	rs6914610	C	A	RUNX2	0.2373	0.2353
5	rs3901654	A	C	XRCC4	0.4174	0.4263	6	rs7748231	G	A	RUNX2	0.4545	0.445
5	exm-rs6452524	A	G	XRCC4	0.4683	0.474	6	rs12209785	G	A	RUNX2	0.2317	0.2539
5	rs1382367	A	G	XRCC4	0.3791	0.4113	6	rs292255	A	G	BACH2	0.4747	0.4453
5	rs10514249	G	A	XRCC4	0.4706	0.4688	6	rs292251	G	A	BACH2	0.4384	0.4241
5	rs7711825	A	C	XRCC4	0.2661	0.2016	6	exm2270359	G	A	BACH2	0.4429	0.4557
5	rs301289	A	G	XRCC4	0.3182	0.2812	6	rs12208074	A	G	BACH2	0.2854	0.2775
5	rs1805377	A	G	XRCC4	0.2833	0.2596	6	rs10806414	A	G	BACH2	0.1864	0.2161
5	rs7730247	C	A	IRF1	0.2604	0.2292	6	rs6936055	A	G	BACH2	0.3295	0.3099
5	exm-rs12521868	A	C	IRF1	0.2851	0.2578	6	rs4707586	A	G	BACH2	0.3439	0.3359
5	rs7703230	A	C	IRF1	0.4318	0.4245	6	rs9451316	A	G	BACH2	0.4286	0.4453
5	rs11745587	A	G	IRF1	0.3432	0.4127	6	rs12663434	A	C	BACH2	0.3273	0.3351
5	rs2522051	A	G	IRF1	0.4725	0.4634	6	rs9342217	A	G	BACH2	0.3532	0.3037
5	exm-rs2522056	A	G	IRF1	0.2591	0.2135	6	rs9342219	A	G	BACH2	0.2682	0.2225
5	rs2548993	G	A	IRF1	0.3009	0.3411	6	rs4053608	A	G	BACH2	0.2854	0.263
5	rs13165038	G	A	IRF1	0.3219	0.401	6	rs207269	A	G	BACH2	0.4068	0.4089
5	rs3798134	A	G	IL13	0.2283	0.2304	6	exm2262189	A	G	BACH2	0.4389	0.4427
5	exm-rs2040704	G	A	IL13	0.2811	0.2906	6	rs12209546	A	G	BACH2	0.4909	0.5
5	exm-rs1295686	A	G	IL13	0.3403	0.3672	6	rs3734660	G	A	BACH2	0.3721	0.3776

Table S3: Single nucleotide polymorphism of the candidate genes selected analysed in the association study (Continued)

Chr	SNP	Ref	Alt	Gene	Frequency of reference allele		Chr	SNP	Ref	Alt	Gene	Frequency of reference allele	
					Patients	Controls						Patients	Controls
4	rs300912	A	G	GAB1	0.457	0.4844	5	rs20541	A	G	IL13	0.2557	0.2734
4	rs1472873	G	A	GAB1	0.3507	0.2865	5	rs2243270	G	A	IL4	0.3219	0.3564
4	rs3805253	G	A	GAB1	0.5158	0.4635	5	rs2569190	A	G	CD14	0.4184	0.4755
4	rs3805236	A	G	GAB1	0.2715	0.3281	5	rs5744455	A	G	CD14	0.2308	0.1927
4	rs3795246	G	A	GAB1	0.3796	0.3989	6	rs2671422	A	G	IRF4	0.2489	0.2526
4	rs7696323	A	G	TLR2	0.2311	0.246	6	rs2001508	C	A	IRF4	0.2545	0.2552
4	rs11938228	A	C	TLR2	0.4055	0.3883	6	rs1877175	A	G	IRF4	0.2136	0.2214
4	rs7656411	C	A	TLR2	0.3382	0.3575	6	exm-rs872071	G	A	IRF4	0.3977	0.3906
4	rs7657186	A	G	TLR3	0.2477	0.2031	6	rs7757906	A	G	IRF4	0.3371	0.3325
4	rs13126816	A	G	TLR3	0.2306	0.2618	6	rs9368188	A	G	E2F3	0.3303	0.3333
4	rs3775291	A	G	TLR3	0.2568	0.2943	6	rs9295464	A	C	E2F3	0.2523	0.2656
4	exm2277278	A	C	TLR3	0.2886	0.276	6	rs2328488	A	G	E2F3	0.2202	0.2266
4	rs10025405	G	A	TLR3	0.3643	0.3568	6	rs6926668	G	A	E2F3	0.2431	0.2812
5	rs7701498	A	G	PIK3R1	0.2172	0.1911	6	rs12205236	A	C	E2F3	0.3096	0.349
5	rs10940157	A	G	PIK3R1	0.2477	0.276	6	rs1042317	A	G	E2F3	0.25	0.25
5	rs831227	A	C	PIK3R1	0.4931	0.4922	6	exm-rs2894043	G	C	MDC1	0.21	0.2263
5	rs7709243	G	A	PIK3R1	0.4295	0.4401	6	rs3132584	A	C	MDC1	0.1963	0.2249
5	rs34303	G	A	PIK3R1	0.4457	0.3901	6	rs7757282	G	A	RUNX2	0.3986	0.4371
5	rs34309	A	G	PIK3R1	0.2991	0.3194	6	rs1321081	A	G	RUNX2	0.4	0.3796
5	rs2302975	G	A	PIK3R1	0.4749	0.4392	6	rs1321080	A	C	RUNX2	0.2052	0.235
5	rs7716675	G	A	PIK3R1	0.2159	0.2199	6	rs9296459	A	G	RUNX2	0.242	0.2763
5	rs3815701	G	A	PIK3R1	0.2511	0.1927	6	rs1321075	A	C	RUNX2	0.2104	0.2068
5	rs3730089	A	G	PIK3R1	0.2217	0.2042	6	rs2819854	A	G	RUNX2	0.4932	0.4922
5	rs3756668	G	A	PIK3R1	0.3955	0.487	6	rs910586	A	G	RUNX2	0.3122	0.3073
5	rs1478486	A	G	XRCC4	0.4037	0.4372	6	rs6914610	C	A	RUNX2	0.2373	0.2353
5	rs3901654	A	C	XRCC4	0.4174	0.4263	6	rs7748231	G	A	RUNX2	0.4545	0.445
5	exm-rs6452524	A	G	XRCC4	0.4683	0.474	6	rs12209785	G	A	RUNX2	0.2317	0.2539
5	rs1382367	A	G	XRCC4	0.3791	0.4113	6	rs292255	A	G	BACH2	0.4747	0.4453
5	rs10514249	G	A	XRCC4	0.4706	0.4688	6	rs292251	G	A	BACH2	0.4384	0.4241
5	rs7711825	A	C	XRCC4	0.2661	0.2016	6	exm2270359	G	A	BACH2	0.4429	0.4557
5	rs301289	A	G	XRCC4	0.3182	0.2812	6	rs12208074	A	G	BACH2	0.2854	0.2775
5	rs1805377	A	G	XRCC4	0.2833	0.2596	6	rs10806414	A	G	BACH2	0.1864	0.2161
5	rs7730247	C	A	IRF1	0.2604	0.2292	6	rs6936055	A	G	BACH2	0.3295	0.3099
5	exm-rs12521868	A	C	IRF1	0.2851	0.2578	6	rs4707586	A	G	BACH2	0.3439	0.3359
5	rs7703230	A	C	IRF1	0.4318	0.4245	6	rs9451316	A	G	BACH2	0.4286	0.4453
5	rs11745587	A	G	IRF1	0.3432	0.4127	6	rs12663434	A	C	BACH2	0.3273	0.3351
5	rs2522051	A	G	IRF1	0.4725	0.4634	6	rs9342217	A	G	BACH2	0.3532	0.3037
5	exm-rs2522056	A	G	IRF1	0.2591	0.2135	6	rs9342219	A	G	BACH2	0.2682	0.2225
5	rs2548993	G	A	IRF1	0.3009	0.3411	6	rs4053608	A	G	BACH2	0.2854	0.263
5	rs13165038	G	A	IRF1	0.3219	0.401	6	rs207269	A	G	BACH2	0.4068	0.4089
5	rs3798134	A	G	IL13	0.2283	0.2304	6	exm2262189	A	G	BACH2	0.4389	0.4427
5	exm-rs2040704	G	A	IL13	0.2811	0.2906	6	rs12209546	A	G	BACH2	0.4909	0.5
5	exm-rs1295686	A	G	IL13	0.3403	0.3672	6	rs3734660	G	A	BACH2	0.3721	0.3776

Table S3: Single nucleotide polymorphism of the candidate genes selected analysed in the association study (Continued)

Chr	SNP	Ref	Alt	Gene	Frequency of reference allele		Chr	SNP	Ref	Alt	Gene	Frequency of reference allele	
					Patients	Controls						Patients	Controls
6	exm-rs3757247	A	G	BACH2	0.4227	0.401	7	rs2906655	A	G	CUX1	0.2629	0.2486
6	exm-rs1847472	A	C	BACH2	0.2705	0.2526	7	rs434564	G	A	CUX1	0.3869	0.3594
6	rs661713	G	A	BACH2	0.3379	0.3047	7	rs202159	A	G	CUX1	0.2248	0.2318
6	rs9111	G	A	BACH2	0.2133	0.1979	7	rs12668172	G	A	CUX1	0.4299	0.4896
6	rs12212193	G	A	BACH2	0.3348	0.375	7	rs10259603	G	A	CUX1	0.3402	0.3691
6	rs535780	G	A	PRDM1	0.4182	0.3906	7	rs11769397	A	G	CUX1	0.2841	0.3053
6	rs573869	A	G	PRDM1	0.2729	0.276	7	rs12671456	A	C	CUX1	0.2056	0.2169
6	exm2270409	A	C	MYB	0.4796	0.4843	7	rs12672026	A	G	CUX1	0.4217	0.4062
6	rs210943	A	G	MYB	0.4225	0.3936	7	rs201492	G	A	CUX1	0.4358	0.4372
6	rs210941	A	G	MYB	0.422	0.3828	7	rs11975153	A	G	CUX1	0.2169	0.2211
6	rs210937	G	A	MYB	0.4521	0.4503	7	rs201523	G	A	CUX1	0.4901	0.4751
6	rs2450975	A	C	SLC22A2	0.2581	0.2053	7	rs201525	A	G	CUX1	0.2159	0.199
6	rs532482	G	A	SLC22A2	0.2945	0.2199	7	rs11760331	A	G	CUX1	0.2783	0.2734
6	rs624249	A	C	SLC22A2	0.3182	0.3516	7	rs6465852	G	A	CUX1	0.4498	0.4476
7	rs3814477	G	A	MAFK	0.5046	0.4424	7	rs409224	C	A	CUX1	0.371	0.3854
7	rs4720833	A	G	MAFK	0.3653	0.3307	7	rs2694158	A	C	CUX1	0.3744	0.3932
7	rs3735656	G	A	MAFK	0.5094	0.4521	7	rs803064	G	A	CUX1	0.461	0.4583
7	rs892523	A	G	MAFK	0.4163	0.3895	7	rs712839	A	G	CUX1	0.2824	0.2684
7	rs6945447	A	G	RPA3	0.4426	0.4801	7	rs803073	A	G	CUX1	0.4144	0.4188
7	rs1859604	A	C	RPA3	0.3819	0.3482	7	exm646305	C	G	CUX1	0.4163	0.3906
7	rs10230651	A	G	RPA3	0.2805	0.3047	7	rs751840	G	A	PAXIP1	0.2603	0.2421
7	rs12668835	A	G	RPA3	0.3258	0.3203	7	exm674990	G	A	PAXIP1	0.2682	0.2969
7	rs1557997	G	A	RPA3	0.3227	0.3613	7	rs2272174	G	A	PAXIP1	0.4475	0.4531
7	rs4720750	A	G	RPA3	0.2091	0.2161	7	rs2293261	A	G	PAXIP1	0.3858	0.3828
7	rs7791564	G	A	RPA3	0.4091	0.4245	7	rs306283	A	G	PAXIP1	0.3387	0.3455
7	rs7778745	G	A	POLM	0.2919	0.2592	8	rs2272733	A	G	IKBKB	0.2227	0.1901
7	rs1139270	A	G	POLM	0.2964	0.2943	8	rs1551655	C	A	PRKDC	0.2466	0.237
7	rs4640970	A	G	POLM	0.3257	0.3586	8	rs4150869	G	A	E2F5	0.3525	0.3466
7	rs6583437	A	G	IKZF1	0.4521	0.4346	8	rs1063053	A	G	NBN	0.2603	0.2658
7	rs6583440	A	G	IKZF1	0.4654	0.4237	8	rs1805818	A	C	NBN	0.3196	0.3385
7	rs7800411	G	A	IKZF1	0.275	0.2816	8	rs4734497	G	A	YWHAZ	0.2489	0.2839
7	rs7790846	A	G	IKZF1	0.2295	0.2891	8	rs964917	A	G	YWHAZ	0.4421	0.4712
7	rs12718598	A	G	IKZF1	0.4306	0.4733	9	exm750947	G	A	PAX5	0.2146	0.1979
7	rs12669559	C	A	IKZF1	0.3624	0.3822	9	rs7044118	G	A	PAX5	0.3	0.3037
7	rs10230385	G	A	IKZF1	0.3091	0.3125	9	rs4074255	A	G	PAX5	0.3665	0.3411
7	rs11980379	G	A	IKZF1	0.2352	0.267	9	rs10973120	G	A	PAX5	0.425	0.4661
7	rs2109857	A	G	YWHAG	0.3914	0.3229	9	rs7034991	C	A	PAX5	0.2636	0.2526
7	rs917424	A	G	YWHAG	0.2489	0.2316	9	rs10123925	G	A	PAX5	0.3219	0.276
7	rs2970471	A	G	CUX1	0.4493	0.4497	9	rs7868521	A	C	PAX5	0.3379	0.3646
7	rs10245008	G	A	CUX1	0.4386	0.4193	9	rs3780151	G	A	PAX5	0.4295	0.474
7	rs11765661	A	G	CUX1	0.21	0.2	9	rs7032626	G	A	PAX5	0.2386	0.2644
7	exm-rs4729759	A	G	CUX1	0.2602	0.25	9	rs13288123	G	A	PAX5	0.2215	0.2656
7	rs6465836	A	G	CUX1	0.3739	0.4271	9	rs7853360	G	A	PAX5	0.2443	0.2708

Table S3: Single nucleotide polymorphism of the candidate genes selected analysed in the association study (Continued)

Chr	SNP	Ref	Alt	Gene	Frequency of reference allele		Chr	SNP	Ref	Alt	Gene	Frequency of reference allele	
					Patients	Controls						Patients	Controls
9	rs4407980	A	G	PAX5	0.3056	0.2968	11	rs4283016	A	G	POU2AF1	0.3932	0.401
9	rs10121384	A	G	PAX5	0.2955	0.2801	11	rs643788	G	A	H2AX	0.3958	0.375
9	rs10814496	G	A	PAX5	0.4269	0.3984	11	rs7111236	G	A	ETS1	0.4018	0.4319
9	rs3780169	A	G	PAX5	0.3507	0.3639	11	rs4937334	A	G	ETS1	0.1932	0.224
9	rs7045954	A	G	PAX5	0.2283	0.2304	11	rs10790956	A	G	ETS1	0.3597	0.4053
9	rs10814498	G	A	PAX5	0.4055	0.3613	11	rs4937339	A	C	ETS1	0.2126	0.2005
9	rs10125775	A	G	PAX5	0.4614	0.474	11	rs7125574	A	G	ETS1	0.3174	0.2513
9	rs3758171	A	G	PAX5	0.2814	0.2513	11	rs4254089	G	A	ETS1	0.2805	0.2995
9	rs3824344	A	G	PAX5	0.4124	0.3474	11	rs7115613	A	G	ETS1	0.2795	0.3151
9	rs7020413	G	A	PAX5	0.3143	0.2926	11	exm2267240	G	A	ETS1	0.3982	0.3783
9	rs6476606	A	G	PAX5	0.3682	0.3646	11	exm-rs11221332	A	G	ETS1	0.2159	0.2057
9	rs959396	C	A	PAX5	0.4124	0.4091	11	rs11819995	A	G	ETS1	0.2091	0.2474
9	exm2264344	C	A	PAX5	0.4543	0.4505	11	rs4245079	A	C	ETS1	0.4795	0.4424
9	rs334349	A	G	TGFBR1	0.3136	0.3185	11	rs7117932	A	G	ETS1	0.3462	0.3776
9	rs1554973	G	A	TLR4	0.2682	0.3016	11	rs1944854	A	G	ETS1	0.4406	0.3947
9	rs10435844	C	A	TRAF1	0.2767	0.2853	12	rs10431309	A	G	APOBEC1	0.1881	0.224
9	rs1014530	A	G	TRAF1	0.4205	0.4479	12	rs10772596	A	G	APOBEC1	0.4247	0.4818
9	rs4880073	A	G	TRAF2	0.4336	0.4144	12	rs9651863	A	G	APOBEC1	0.4771	0.4401
10	exm2271427	A	G	DNTT	0.2647	0.2786	12	rs7309127	G	A	APOBEC1	0.3841	0.3073
10	rs2154183	A	G	DNTT	0.2233	0.2407	12	exm2251234	G	A	AICDA	0.4118	0.4427
10	exm849056	G	A	CHUK	0.4562	0.4286	12	rs2580874	A	G	AICDA	0.4518	0.4583
10	rs12570957	A	C	CHUK	0.3647	0.3438	12	exm2267349	A	G	TFCP2	0.4455	0.4211
10	rs12247992	A	G	CHUK	0.21	0.1979	12	rs2730648	G	A	TFCP2	0.2894	0.2766
10	rs1055362	G	A	POLL	0.2453	0.2593	12	rs2640518	G	A	TFCP2	0.2083	0.1878
10	rs11190	C	A	IKZF5	0.3455	0.3184	12	rs6580799	A	G	TFCP2	0.4095	0.3802
10	rs9423240	G	A	IKZF5	0.3977	0.4	12	rs7131938	A	G	SP1	0.2023	0.2109
11	exm2271582	C	A	SWAP70	0.4568	0.4605	12	rs2241820	G	A	HOXC4	0.477	0.4424
11	rs4548612	A	G	SWAP70	0.3659	0.3333	12	rs736825	G	C	HOXC4	0.3158	0.3641
11	rs397250	G	A	SWAP70	0.2828	0.3021	12	rs10876528	A	C	HOXC4	0.2896	0.3289
11	rs360157	G	A	SWAP70	0.4409	0.3953	12	rs7136889	C	A	HOXC4	0.2763	0.3281
11	exm890034	G	C	SWAP70	0.3914	0.3281	12	rs10747691	A	G	HOXC4	0.3484	0.3385
11	rs360140	C	A	SWAP70	0.2834	0.2604	12	rs2171216	A	G	HOXC4	0.4341	0.4844
11	rs5030411	A	G	TRAF6	0.4523	0.4688	12	rs2456973	C	A	IKZF4	0.2591	0.2461
11	rs3740955	A	G	RAG1	0.4266	0.5078	12	rs12298170	G	A	STAT6	0.456	0.3797
11	rs10838698	G	A	SPI1	0.3699	0.3359	12	rs2279575	A	G	E2F7	0.3773	0.4531
11	rs2508678	A	G	MRE11	0.4266	0.3828	12	rs11116780	G	A	E2F7	0.3545	0.3438
11	rs663530	A	G	MRE11	0.2673	0.254	12	rs7978946	A	G	UNG	0.2421	0.2135
11	rs654718	G	A	MRE11	0.3562	0.3568	13	exm2251437	A	G	ELF1	0.4182	0.4036
11	rs476137	C	A	MRE11	0.47	0.5079	13	rs7334701	A	G	ELF1	0.3201	0.3571
11	rs619972	G	A	ATM	0.4429	0.4711	13	rs1151403	G	A	LIG4	0.3795	0.3698
11	rs425538	C	A	ATM	0.3239	0.3614	13	rs7322498	G	A	LIG4	0.3122	0.3516
11	rs1815950	G	A	POU2AF1	0.2273	0.2578	13	rs16972217	A	G	TNFSF13B	0.2986	0.2899
11	rs7116862	A	G	POU2AF1	0.2752	0.3073	13	rs9520835	A	G	TNFSF13B	0.2676	0.2737

Table S3: Single nucleotide polymorphism of the candidate genes selected analysed in the association study (Continued)

Chr	SNP	Ref	Alt	Gene	Frequency of reference allele		Chr	SNP	Ref	Alt	Gene	Frequency of reference allele	
					Patients	Controls						Patients	Controls
14	exm1085011	C	A	APEX1	0.4231	0.4036	16	rs37358	A	G	CSNK2A2	0.3773	0.3802
14	rs1024775	A	C	BATF	0.3561	0.3069	16	rs1035549	A	G	CSNK2A2	0.2841	0.2891
14	rs2268625	G	A	TGFB3	0.2059	0.224	16	exm1248113	A	C	E2F4	0.4404	0.4319
14	rs2075771	A	G	TRAF3	0.3864	0.4162	17	rs7219354	A	G	YWHAE	0.293	0.3184
14	rs8022180	A	G	TRAF3	0.4247	0.4686	17	rs11650689	A	G	YWHAE	0.2685	0.2646
14	rs12147254	A	G	TRAF3	0.1986	0.2304	17	rs8078073	G	A	YWHAE	0.1895	0.209
14	rs2144826	A	G	TRAF3	0.242	0.2277	17	rs4790084	A	G	YWHAE	0.4406	0.4974
14	rs7143963	A	G	TRAF3	0.3477	0.3281	17	rs11655548	G	A	YWHAE	0.3824	0.3516
14	exm1129231	G	A	TRAF3	0.4773	0.4375	17	rs8065937	C	A	RPA1	0.3539	0.3482
14	exm-rs10133111	A	G	TRAF3	0.2795	0.276	17	rs2287321	G	A	RPA1	0.475	0.4398
14	exm2267741	A	G	IGH	0.2382	0.1596	17	rs11867830	G	A	RPA1	0.214	0.2326
14	rs1024350	G	A	IGH	0.4659	0.4712	17	rs3786136	A	G	RPA1	0.2857	0.224
14	rs8009638	A	G	IGH	0.3136	0.2917	17	rs17292175	A	G	RPA1	0.2919	0.3203
14	rs4280141	A	G	IGH	0.2308	0.2422	17	rs7406953	G	A	RPA1	0.2682	0.2094
14	rs6576233	A	G	IGH	0.3356	0.3211	17	rs17734	G	A	RPA1	0.5208	0.4707
14	rs7157975	G	A	IGH	0.4886	0.4766	17	rs9914073	G	A	RPA1	0.2099	0.2184
14	rs7142108	G	A	IGH	0.3045	0.2827	17	exm1283761	A	G	RPAIN	0.2986	0.3203
14	rs17112078	G	A	IGH	0.3869	0.3516	17	rs2189338	A	G	RPAIN	0.4367	0.4505
14	rs2078693	G	A	IGH	0.2727	0.2737	17	rs1050390	A	G	RPAIN	0.25	0.2593
14	rs885883	G	A	IGH	0.2977	0.3395	17	exm1288075	A	G	TNFSF13	0.367	0.3021
14	exm2267742	G	A	IGH	0.4518	0.4346	17	rs8073498	C	A	TP53	0.3341	0.349
14	rs10149476	C	A	IGH	0.491	0.4634	17	exm1288658	G	C	TP53	0.2941	0.2344
15	rs4924501	C	A	RAD51	0.4182	0.4609	17	exm2272495	G	A	TNFRSF13B	0.4729	0.4349
15	rs689647	A	G	TP53BP1	0.2455	0.2656	17	rs12938061	A	G	TNFRSF13B	0.4593	0.5184
15	rs7162912	A	C	SMAD3	0.2851	0.2995	17	rs12603708	A	G	TNFRSF13B	0.2919	0.2656
15	rs10518707	G	A	SMAD3	0.35	0.3974	17	rs2274892	C	A	TNFRSF13B	0.3773	0.3482
15	rs4776881	G	A	SMAD3	0.4409	0.4349	17	exm-rs907092	A	G	IKZF3	0.3493	0.3203
15	rs4776888	G	A	SMAD3	0.3402	0.3385	17	rs9303277	A	G	IKZF3	0.4931	0.4342
15	rs2118612	G	A	SMAD3	0.3664	0.3534	17	rs9635726	A	G	IKZF3	0.2318	0.2565
15	rs1992215	G	A	SMAD3	0.3205	0.3194	17	exm1326586	G	A	BRCA1	0.3213	0.3168
15	rs12102171	A	G	SMAD3	0.2127	0.2031	17	exm1326613	A	G	BRCA1	0.4523	0.445
15	rs4147358	A	C	SMAD3	0.3562	0.3021	17	rs708563	G	A	MAP3K14	0.4705	0.4896
15	rs2118610	A	G	SMAD3	0.3409	0.401	17	exm1330395	G	C	MAP3K14	0.2636	0.263
15	rs1866316	G	A	SMAD3	0.3659	0.356	17	rs7213493	A	G	MAP3K14	0.2098	0.2027
15	rs744910	A	G	SMAD3	0.4523	0.4424	17	rs2867316	A	G	MAP3K14	0.3486	0.3796
15	rs4601989	A	G	SMAD3	0.2671	0.237	17	rs12449740	G	A	MAP3K14	0.4909	0.4688
15	rs17228212	G	A	SMAD3	0.2033	0.2207	17	rs7209608	A	C	CD79B	0.4163	0.4505
15	rs7183244	A	G	SMAD3	0.2694	0.2895	17	rs2320125	A	G	CD79B	0.395	0.3594
16	exm-rs886528	G	A	CREBBP	0.4276	0.4476	18	rs4517886	A	C	RBBP8	0.2362	0.1885
16	exm-rs3136202	A	G	ERCC4	0.4197	0.377	18	rs4800141	G	A	RBBP8	0.4628	0.3844
16	rs9646271	A	G	ERCC4	0.2636	0.2604	18	rs4939827	A	G	SMAD7	0.4282	0.466
16	rs1057451	A	C	PAGR1	0.1918	0.2266	18	rs12953717	A	G	SMAD7	0.3791	0.4162
16	rs4784901	G	A	CSNK2A2	0.225	0.2292	18	rs4464148	G	A	SMAD7	0.2936	0.2708

Table S3: Single nucleotide polymorphism of the candidate genes selected analysed in the association study (Continued)

Chr	SNP	Ref	Alt	Gene	Frequency of reference allele		Chr	SNP	Ref	Alt	Gene	Frequency of reference allele	
					Patients	Controls						Patients	Controls
18	rs7238442	A	G	SMAD7	0.4614	0.4896	21	rs2834737	G	A	RUNX1	0.4083	0.4031
18	rs4939832	G	A	SMAD7	0.25	0.2604	21	rs7283760	A	G	ICOSLG	0.3032	0.3229
18	rs3736242	A	G	SMAD7	0.1991	0.224	21	rs15927	A	G	ICOSLG	0.2791	0.2816
18	rs12455792	A	G	SMAD4	0.4205	0.3854	21	exm1576212	A	G	ICOSLG	0.29	0.2552
18	rs12456284	G	A	SMAD4	0.2386	0.2318	21	rs2070561	G	A	ICOSLG	0.3295	0.2917
19	rs10415670	G	A	TCF3	0.2237	0.199	22	rs4820285	G	A	IGL	0.2397	0.2396
19	rs2277991	G	A	FCER2	0.294	0.2816	22	rs735455	A	G	IGL	0.2738	0.25
19	rs12971845	G	A	FCER2	0.2805	0.2891	22	rs762470	A	G	IGL	0.4045	0.3984
19	rs11260013	G	A	FCER2	0.4114	0.4219	22	rs5750465	A	G	IGL	0.3403	0.3763
19	rs7246264	A	G	FCER2	0.2421	0.2448	22	exm2261183	A	C	IGL	0.4289	0.3542
19	rs3212713	A	G	JAK3	0.375	0.3776	22	rs5750496	A	G	IGL	0.475	0.4948
19	rs3212701	A	G	JAK3	0.3059	0.3115	22	rs5756895	G	A	IGL	0.2104	0.2356
19	rs1865093	A	G	SUPT5H	0.1837	0.2289	22	rs11704841	C	A	IGL	0.338	0.3141
19	rs1529733	A	G	SUPT5H	0.4116	0.4297	22	rs4820327	A	C	IGL	0.3116	0.2619
19	rs2060271	A	G	SUPT5H	0.4495	0.4115	22	rs9622968	G	A	IGL	0.3032	0.2789
19	rs4803455	A	C	TGFB1	0.4839	0.4553	22	rs9306334	A	G	IGL	0.3235	0.3542
19	rs1800469	A	G	TGFB1	0.3456	0.3684	22	rs2073454	A	C	IGL	0.4726	0.4401
19	rs1056995	A	G	CD79A	0.3227	0.3229	22	rs2105815	A	G	IGL	0.2964	0.2891
19	rs4803790	G	A	POU2F2	0.2909	0.3325	22	rs5757258	A	G	IGL	0.422	0.4031
19	rs10412761	A	G	RELB	0.4772	0.4868	22	rs6001219	G	A	IGL	0.3778	0.3385
19	rs1005165	A	G	ERCC1	0.2091	0.1823	22	rs2073446	A	G	IGL	0.3151	0.2995
19	rs3212986	A	C	ERCC1	0.2817	0.3005	22	rs9611215	A	C	IGL	0.406	0.4346
19	rs11615	A	G	ERCC1	0.4253	0.474	22	rs2330020	A	C	IGL	0.4656	0.4948
20	rs6060260	A	G	ID1	0.2136	0.1901	22	rs8137866	G	A	IGL	0.4568	0.4505
20	exm2268288	G	A	E2F1	0.4344	0.4141	22	rs1573600	A	C	IGL	0.2624	0.237
20	rs2239535	A	G	YWHAB	0.2851	0.3021	22	rs4821948	A	C	IGL	0.2376	0.2356
20	rs2425672	A	G	YWHAB	0.3955	0.3984	22	rs1475930	G	A	IGL	0.3881	0.3665
21	rs2070370	A	G	RUNX1	0.3914	0.3822	22	rs11090195	A	G	IGL	0.347	0.3492
21	exm2254596	A	C	RUNX1	0.4658	0.445	22	rs737885	C	A	IGL	0.2831	0.2891
21	rs2268278	A	C	RUNX1	0.3886	0.3822	22	rs394409	A	G	IGL	0.4862	0.4297
21	rs2409535	G	A	RUNX1	0.3643	0.3639	22	rs1547014	A	G	CHEK2	0.3356	0.3255
21	rs2268288	G	A	RUNX1	0.1939	0.2487	22	rs2073327	G	A	CHEK2	0.3043	0.3182
21	rs2834655	A	G	RUNX1	0.2851	0.301	22	rs5762764	G	A	CHEK2	0.2805	0.2969
21	exm2273008	A	G	RUNX1	0.3462	0.2801	22	rs5762766	G	A	CHEK2	0.2593	0.2447
21	rs2300399	A	G	RUNX1	0.2159	0.1745	22	rs2301415	G	A	YWHAH	0.21	0.263
21	rs8130963	G	A	RUNX1	0.3813	0.3411	22	rs5998196	G	A	YWHAH	0.4412	0.4136
21	rs4817699	A	G	RUNX1	0.1968	0.2173	22	rs9611497	A	G	EP300	0.3077	0.3246
21	rs2834683	G	A	RUNX1	0.4748	0.3974	22	exm1611785	G	A	EP300	0.3562	0.3151
21	rs2834698	G	A	RUNX1	0.3676	0.4583	22	rs2076577	A	G	EP300	0.3848	0.4005
21	rs2834709	A	G	RUNX1	0.4009	0.4062	23	rs3764880	G	A	TLR8	0.3576	0.316
21	rs2834714	G	A	RUNX1	0.3348	0.3482	23	rs4830807	A	C	TLR8	0.4506	0.4772
21	exm-rs2014300	A	G	RUNX1	0.2466	0.2135	23	rs2159377	A	G	TLR8	0.2072	0.2414
21	rs2834719	A	G	RUNX1	0.2659	0.1927	23	rs704145	A	G	APEX2	0.3526	0.3253
21	rs2834725	G	A	RUNX1	0.2805	0.237	23	rs1126535	G	A	CD40LG	0.3241	0.2812
21	rs881386	A	G	RUNX1	0.2511	0.263	23	exm2273302	A	G	TLR7	0.2229	0.2076

Table S4: Predicted regulatory effects according to Haploreg database of variants rs6657275, rs1473526, rs407960, rs1818545, rs10781530 and of those in strong linkage disequilibrium with them ($r^2 \geq 0.8$).

Query SNP: rs6657275 and variants with r^2 >= 0.8

chr	pos (hg38)	Distance to associated SNP in kb	LD (r^2)	LD (D')	variant	Ref	Alt	AFR freq	AMR freq	ASN freq	EUR freq	SiPhy cons	Promoter histone marks	Enhancer histone marks	DNAse
1	218418281	-4.8	1	1	rs11466399	G	A	0.51	0.44	0.75	0.28				
1	218421578	-1.5	1	1	rs1934852	G	T	0.51	0.45	0.75	0.28			6 tissues	4 tissues
1	218423119	0.0	1	1	**rs6657275**	A	G	0.51	0.45	0.75	0.28			BRN	
1	218423299	0.2	1	1	rs6668651	G	T	0.51	0.45	0.75	0.28			BRN	
1	218423702	0.6	1	1	rs6671370	G	A	0.55	0.45	0.75	0.28			BRN	
1	218423955	0.8	1	1	rs12048049	C	G	0.51	0.45	0.75	0.28			BRN, GI, LNG	
1	218424986	1.9	0.97	1	rs4846478	C	G	0.35	0.43	0.73	0.27			8 tissues	
1	218425068	1.9	0.97	1	rs4846479	G	T	0.35	0.43	0.74	0.27			8 tissues	
1	218425127	2.0	0.97	1	rs4846480	A	T	0.35	0.43	0.74	0.27			8 tissues	
1	218425142	2.0	1	1	rs4846481	C	T	0.5	0.44	0.75	0.28			8 tissues	
1	218427077	4.0	0.97	1	rs6658473	C	T	0.36	0.43	0.74	0.27			ESDR, FAT, BRST	
1	218431336	8.2	0.97	1	rs1890995	G	A	0.35	0.43	0.75	0.27			5 tissues	ESDR,OVRY
1	218432119	9.0	0.97	1	rs10482792	G	A	0.35	0.43	0.75	0.27				SKIN,PLCNT
1	218432267	9.1	1	1	rs10482795	G	A	0.51	0.44	0.75	0.28				
1	218436360	13.2	1	1	rs6684205	A	G	0.51	0.44	0.75	0.28			MUS	
1	218436912	13.8	1	1	rs1418553	C	T	0.51	0.44	0.75	0.28			FAT, BRST, BRN	
1	218441563	18.4	0.96	0.99	rs900	A	T	0.35	0.43	0.73	0.27				
1	218442109	19.0	0.99	0.99	rs991967	A	C	0.51	0.44	0.75	0.28				
1	218442671	19.6	0.92	0.99	rs201094302	TC	T	0.45	0.42	0.73	0.26				
1	218443793	20.7	0.97	0.99	rs1046017	C	G	0.51	0.44	0.75	0.28				
1	218444201	21.1	0.98	0.99	rs6704255	G	A	0.42	0.44	0.74	0.28			MUS, LNG	
1	218444558	21.4	0.98	0.99	rs6683598	C	T	0.42	0.44	0.74	0.28			MUS, LNG	
1	218446304	23.2	0.98	0.99	rs1342590	C	T	0.42	0.44	0.74	0.28			FAT, MUS	IPSC
1	218446857	23.7	0.99	0.99	rs12032375	C	T	0.51	0.44	0.75	0.28			FAT, MUS	
1	218447405	24.3	0.96	0.99	rs1473526	T	C	0.35	0.43	0.74	0.27		FAT	MUS	MUS,MUS
1	218448759	25.6	0.88	0.97	rs2015850	A	G	0.33	0.41	0.75	0.27				
1	218448977	25.9	0.97	0.99	rs1473527	G	A	0.63	0.47	0.78	0.28				
1	218449933	26.8	0.94	0.99	rs6657698	G	A	0.36	0.43	0.77	0.27				
1	218451191	28.1	0.92	0.96	rs10429950	T	C	0.36	0.42	0.76	0.28				
1	218453278	30.2	0.94	0.99	rs7515360	T	A	0.36	0.43	0.77	0.27			MUS	
1	218453979	30.9	0.94	0.99	rs7549303	C	G	0.36	0.43	0.77	0.27				
1	218454682	31.6	0.97	0.99	rs6666381	T	A	0.64	0.47	0.78	0.28				
1	218455687	32.6	0.97	0.99	rs1318580	T	C	0.65	0.47	0.78	0.28				
1	218455991	32.9	0.9	0.97	rs1797069	A	G	0.34	0.4	0.74	0.27				
1	218456859	33.7	0.94	0.99	rs1797070	G	A	0.38	0.43	0.77	0.27				
1	218458110	35.0	0.95	0.98	rs6604614	C	G	0.63	0.46	0.78	0.28				
1	218459026	35.9	0.94	0.99	rs4846483	C	A	0.38	0.43	0.76	0.27			5 tissues	GI
1	218461020	37.9	0.9	0.99	rs6604615	T	G	0.52	0.44	0.76	0.26			13 tissues	
1	218461445	38.3	0.92	0.99	rs1108548	A	G	0.65	0.47	0.77	0.27			10 tissues	LNG
1	218461737	38.6	0.9	0.99	rs7512679	C	T	0.52	0.44	0.76	0.26			8 tissues	

Table S4: Predicted regulatory effects according to Haploreg database of variants rs6657275, rs1473526, rs407960, rs1818545, rs10781530 and of those in strong linkage disequilibrium with them ($r^2 \geq 0.8$).

Query SNP: rs6657275 and variants with r^2 >= 0.8

chr	pos (hg38)	Proteins bound	Motifs changed	NHGRI/EBI GWAS hits	GRASP QTL hits	Selected eQTL hits	GENCODE genes	dbSNP func annot
1	218418281						TGFB2	intronic
1	218421578		Pou2f2				TGFB2	intronic
1	218423119		GR		1 hit	1 hit	TGFB2	intronic
1	218423299		7 altered motifs				TGFB2	intronic
1	218423702		4 altered motifs				TGFB2	intronic
1	218423955		4 altered motifs				TGFB2	intronic
1	218424986		8 altered motifs				TGFB2	intronic
1	218425068		CTCF		1 hit		TGFB2	intronic
1	218425127		Hoxa5,PRDM1	1 hit			TGFB2	intronic
1	218425142		AP-1,Egr-1				TGFB2	intronic
1	218427077		TBX5				TGFB2	intronic
1	218431336		6 altered motifs	1 hit	1 hit		TGFB2	intronic
1	218432119	GATA3	18 altered motifs		1 hit		TGFB2	intronic
1	218432267		GR		1 hit		TGFB2	intronic
1	218436360		HNF1,Pax-4,ZBTB7A	1 hit	1 hit	1 hit	TGFB2	intronic
1	218436912		4 altered motifs		3 hits		TGFB2	intronic
1	218441563		Maf				TGFB2	3'-UTR
1	218442109		Foxa				TGFB2	3'-UTR
1	218442671		8 altered motifs				TGFB2	3'-UTR
1	218443793		9 altered motifs				TGFB2	3'-UTR
1	218444201		GR,Pou3f1,YY1				TGFB2	3'-UTR
1	218444558						TGFB2	3'-UTR
1	218446304		LUN-1		2 hits		1.7kb 3' of TGFB2	
1	218446857		AP-1				2.2kb 3' of TGFB2	
1	218447405		5 altered motifs				2.8kb 3' of TGFB2	
1	218448759		Hdx,Myc				4.1kb 3' of TGFB2	
1	218448977						4.4kb 3' of TGFB2	
1	218449933		Pou5f1		1 hit		5.3kb 3' of TGFB2	
1	218451191		4 altered motifs		2 hits		6.6kb 3' of TGFB2	
1	218453278		Irf				8.7kb 3' of TGFB2	
1	218453979						9.4kb 3' of TGFB2	
1	218454682		8 altered motifs				10kb 3' of TGFB2	
1	218455687		8 altered motifs				11kb 3' of TGFB2	
1	218455991		BDP1,Brachyury,LUN-1				11kb 3' of TGFB2	
1	218456859		5 altered motifs				12kb 3' of TGFB2	
1	218458110		Rad21				13kb 3' of TGFB2	
1	218459026		Foxa,Foxd3,Sox		1 hit		14kb 3' of TGFB2	
1	218461020		6 altered motifs				16kb 3' of TGFB2	
1	218461445		Arid5b				17kb 3' of TGFB2	
1	218461737		6 altered motifs				17kb 3' of TGFB2	

Table S4: Predicted regulatory effects according to Haploreg database of variants rs6657275, rs1473526, rs407960, rs1818545, rs10781530 and of those in strong linkage disequilibrium with them ($r^2 \geq 0.8$).

Query SNP: rs1818545 and variants with r^2 >= 0.8

chr	pos (hg38)	Distance to associated SNP in kb	LD (r^2)	LD (D')	variant	Ref	Alt	AFR freq	AMR freq	ASN freq	EUR freq	SiPhy cons	Promoter histone marks	Enhancer histone marks	DNAse
11	36590540	0.0	1	1	**rs1818545**	C	T	0.59	0.34	0.64	0.23		THYM	BLD	THYM
11	36595317	4.8	1	1	rs7104753	A	G	0.57	0.33	0.65	0.23		19 tissues	6 tissues	BLD,THYM
11	36598542	8.0	1	1	rs12283331	G	A	0.56	0.33	0.65	0.23		BLD, THYM		THYM
11	36615883	25.3	0.97	0.99	rs11827987	T	C,G	0.61	0.33	0.65	0.23			BLD, BRN	
11	36620823	30.3	0.97	0.99	rs1105593	C	T	0.61	0.33	0.65	0.23		THYM	BLD	
11	36622088	31.5	0.93	0.98	rs201432972	TTA	T	0.54	0.33	0.65	0.23		THYM, BLD		
11	36622089	31.5	0.93	0.98	rs147783549	TA	T	0.54	0.33	0.65	0.23		THYM, BLD		
11	36634339	43.8	0.93	0.97	rs7942300	A	G	0.61	0.33	0.65	0.23			GI	
11	36638037	47.5	0.93	0.97	rs12804142	T	C	0.61	0.33	0.65	0.23				
11	36639852	49.3	0.93	0.97	rs16926234	G	A	0.56	0.33	0.65	0.23			GI	
11	36647580	57.0	0.91	0.96	rs10836583	C	T	0.55	0.33	0.65	0.23				
11	36657260	66.7	0.93	0.97	rs7113441	G	C	0.56	0.33	0.65	0.23		BLD	5 tissues	

Query SNP: rs1818545 and variants with r^2 >= 0.8

chr	pos (hg38)	Proteins bound	Motifs changed	NHGRI/EBI GWAS hits	GRASP QTL hits	Selected eQTL hits	GENCODE genes	dbSNP func annot
11	36590540	GATA2	GR,TATA				RAG2	
11	36595317		5 altered motifs				RAG2	intronic
11	36598542		Ik-2				C11orf74	intronic
11	36615883						C11orf74	intronic
11	36620823		4 altered motifs				C11orf74	intronic
11	36622088		12 altered motifs				C11orf74	intronic
11	36622089		12 altered motifs				C11orf74	intronic
11	36634339		MZF1::1-4,Pax-4				C11orf74	intronic
11	36638037		Pax-5,TCF12				C11orf74	intronic
11	36639852		4 altered motifs				C11orf74	intronic
11	36647580		9 altered motifs				C11orf74	intronic
11	36657260		Nkx2,Nkx3				C11orf74	intronic

Table S4: Predicted regulatory effects according to Haploreg database of variants rs6657275, rs1473526, rs407960, rs1818545, rs10781530 and of those in strong linkage disequilibrium with them ($r^2 \geq 0.8$).

Query SNP: rs10781530 and variants with r^2 >= 0.8

chr	pos (hg38)	Distance to associated SNP in kb	LD (r^2)	LD (D')	variant	Ref	Alt	AFR freq	AMR freq	ASN freq	EUR freq	SiPhy cons	Promoter histone marks	Enhancer histone marks	DNAse
9	136975156	-16.3	0.9	0.98	rs7849705	C	T	0.58	0.58	0.69	0.68		BLD, CRVX, SKIN	16 tissues	7 tissues
9	136975462	-16.0	0.91	0.99	rs3814499	G	A	0.57	0.57	0.81	0.69		4 tissues	19 tissues	9 tissues
9	136975638	-15.9	0.91	0.99	rs3814500	C	T	0.65	0.58	0.81	0.69		4 tissues	18 tissues	7 tissues
9	136975661	-15.8	0.9	0.98	rs3814501	C	T	0.58	0.57	0.81	0.68		4 tissues	18 tissues	11 tissues
9	136981689	-9.8	0.91	0.99	rs6926	A	C	0.71	0.59	0.83	0.69		SKIN, CRVX	11 tissues	7 tissues
9	136982031	-9.5	0.91	0.99	rs4880180	G	T	0.71	0.59	0.83	0.69		CRVX	11 tissues	PLCNT,CRVX
9	136982677	-8.8	0.91	0.99	rs2049043	A	C	0.62	0.58	0.83	0.69		SKIN, CRVX	11 tissues	ESDR,CRVX
9	136983865	-7.6	0.91	0.99	rs7029565	C	T	0.61	0.58	0.83	0.69		CRVX	4 tissues	BLD,SKIN
9	136984376	-7.1	0.91	0.99	rs4880082	T	C	0.61	0.58	0.83	0.69		SKIN	7 tissues	OVRY
9	136984661	-6.8	0.91	0.99	rs2049041	G	A	0.57	0.58	0.83	0.69		SKIN	9 tissues	
9	136986675	-4.8	0.84	0.92	rs10870154	G	A	0.56	0.56	0.81	0.67				
9	136988288	-3.2	0.9	0.98	rs10870155	A	C	0.57	0.57	0.82	0.68				
9	136988961	-2.5	0.92	1	rs4880182	G	C	0.6	0.58	0.83	0.69			4 tissues	PLCNT
9	136989496	-2.0	0.96	0.99	rs10781529	T	C	0.53	0.56	0.8	0.66				
9	136989995	-1.5	0.99	1	rs7860430	C	T	0.54	0.57	0.82	0.67			4 tissues	
9	136991087	-0.4	0.99	1	rs10870156	A	G	0.73	0.59	0.83	0.67				8 tissues
9	136991310	-0.2	1	1	rs10870157	C	T	0.73	0.59	0.83	0.67				4 tissues
9	136991368	-0.1	1	1	rs10870158	C	A	0.54	0.57	0.83	0.67				7 tissues
9	136991496	0.0	1	1	**rs10781530**	C	A	0.54	0.57	0.83	0.67				15 tissues
9	136992590	1.1	0.94	0.98	rs10870159	G	C	0.59	0.57	0.83	0.66				39 tissues
9	136998671	7.2	0.99	1	rs7470867	C	T	0.54	0.56	0.83	0.67		CRVX	12 tissues	14 tissues
9	136999551	8.1	0.99	1	rs6560653	A	G	0.75	0.58	0.83	0.67			11 tissues	9 tissues
		-136991.5	0.89	1	rs80052925	T	C	0.71	0.57	0.79	0.65			12 tissues	BLD
9	136999880	8.4	0.83	1	rs7022347	C	A	0.59	0.54	0.76	0.63			12 tissues	BLD
9	137000977	9.5	0.97	0.99	rs12340777	T	C	0.59	0.56	0.82	0.67			5 tissues	PLCNT
9	137001062	9.6	0.92	0.98	rs7854838	C	T	0.53	0.57	0.81	0.66			5 tissues	
9	137001546	10.1	0.97	0.99	rs7388671	G	A	0.74	0.57	0.82	0.67			4 tissues	
9	137002219	10.7	0.93	0.99	rs199897466	G	16-mer	0.59	0.56	0.82	0.66		SKIN, LIV	4 tissues	PLCNT,LIV
9	137002712	11.2	0.98	0.99	rs2049040	C	T	0.74	0.58	0.82	0.67			PLCNT, LIV	
9	137004813	13.3	0.96	0.99	rs4355883	T	C	0.6	0.57	0.82	0.66			BLD	
9	137004900	13.4	0.82	0.99	rs4576515	C	T	0.51	0.54	0.78	0.63			BLD	
9	137009585	18.1	0.97	0.99	rs7048567	A	G	0.68	0.57	0.82	0.67			4 tissues	OVRY
9	137011907	20.4	0.89	0.96	rs2271862	G	A	0.61	0.56	0.82	0.66			IPSC, BRN	IPSC
9	137012061	20.6	0.84	0.93	rs2271863	T	C	0.67	0.57	0.82	0.68			5 tissues	MUS
9	137013390	21.9	0.82	0.91	rs4880188	T	C	0.66	0.56	0.82	0.67			6 tissues	4 tissues
9	137013687	22.2	0.83	0.93	rs10870162	A	G	0.62	0.57	0.82	0.68				

Table S4: Predicted regulatory effects according to Haploreg database of variants rs6657275, rs1473526, rs407960, rs1818545, rs10781530 and of those in strong linkage disequilibrium with them ($r^2 \geq 0.8$).

Query SNP: rs10781530 and variants with r^2 >= 0.8

chr	pos (hg38)	Proteins bound	Motifs changed	NHGRI/EBI GWAS hits	GRASP QTL hits	Selected eQTL hits	GENCODE genes	dbSNP func annot
9	136975156	POL2	8 altered motifs			37 hits	RP11-229P13.19	
9	136975462	PU1				39 hits	RP11-229P13.19	
9	136975638	6 bound proteins	ATF3			39 hits	RP11-229P13.19	
9	136975661	5 bound proteins	MAZ,NERF1a			37 hits	RP11-229P13.19	
9	136981689	CEBPB	6 altered motifs			39 hits	PTGDS	3'-UTR
9	136982031	GATA2,JUND	26 altered motifs			41 hits	PTGDS	
9	136982677		AP-2,SETDB1,Znf143		14 hits	38 hits	PTGDS	
9	136983865		RXRA,Rad21			36 hits	PTGDS	intronic
9	136984376					39 hits	PTGDS	intronic
9	136984661		4 altered motifs			38 hits	PTGDS	intronic
9	136986675		6 altered motifs			29 hits	264bp 3' of LCNL1	
9	136988288		LXR,TFII-I,TR4			33 hits	1.9kb 3' of LCNL1	
9	136988961	ZNF263				32 hits	2.6kb 3' of LCNL1	
9	136989496		16 altered motifs			31 hits	2.9kb 5' of C9orf142	
9	136989995		EBF,Pax-5,RFX5			36 hits	2.4kb 5' of C9orf142	
9	136991087	ZNF263	CACD,PU.1,STAT			38 hits	1.3kb 5' of C9orf142	
9	136991310	ZNF263	BDP1,ELF1,YY1			39 hits	1.1kb 5' of C9orf142	
9	136991368	ZNF263	4 altered motifs		5 hits	41 hits	1kb 5' of C9orf142	
9	136991496	POL2	CTCF			41 hits	921bp 5' of C9orf142	
9	136992590	12 bound proteins	7 altered motifs			39 hits	C9orf142	intronic
9	136998671	4 bound proteins	PPAR			39 hits	1.9kb 5' of CLIC3	
9	136999551		Brachyury,Foxp3,Sin3Ak-20			44 hits	2.7kb 5' of CLIC3	
			16 altered motifs				3.1kb 5' of CLIC3	
9	136999880		9 altered motifs			27 hits	3.1kb 5' of CLIC3	
9	137000977	ERALPHA_A,EGR1				37 hits	4.2kb 5' of CLIC3	
9	137001062	ERALPHA_A,EGR1	14 altered motifs			30 hits	4.3kb 5' of CLIC3	
9	137001546					39 hits	4.7kb 5' of CLIC3	
9	137002219		8 altered motifs			34 hits	5kb 3' of ABCA2	
9	137002712				5 hits	46 hits	4.5kb 3' of ABCA2	
9	137004813		Nanog,Pou3f3,Sox			36 hits	2.4kb 3' of ABCA2	
9	137004900		9 altered motifs			38 hits	2.3kb 3' of ABCA2	
9	137009585		Duxl,HNF4,RXRA		13 hits	44 hits	ABCA2	synonymous
9	137011907		4 altered motifs			43 hits	ABCA2	synonymous
9	137012061		17 altered motifs			37 hits	ABCA2	intronic
9	137013390	EGR1,IRF1	8 altered motifs			33 hits	ABCA2	intronic
9	137013687		ERalpha-a,Pax-4			31 hits	ABCA2	intronic

Table S4: Predicted regulatory effects according to Haploreg database of variants rs6657275, rs1473526, rs407960, rs1818545, rs10781530 and of those in strong linkage disequilibrium with them ($r^2 \geq 0.8$).

Query SNP: rs535068 and variants with $r^2 >= 0.8$

chr	pos (hg38)	Distance to associated SNP in kb	LD (r^2)	LD (D')	variant	Ref	Alt	AFR freq	AMR freq	ASN freq	EUR freq	SiPhy cons	Promoter histone marks	Enhancer histone marks	DNAse
1	12126876	-2.6	0.91	0.99	rs1148475	A	G	0.96	0.8	0.7	0.68		BLD, FAT	6 tissues	
1	12128037	-1.5	0.97	1	rs1148474	A	G	0.82	0.79	0.76	0.67		BLD	5 tissues	
1	12128425	-1.1	0.98	0.99	rs1201122	G	A	0.75	0.79	0.62	0.67		BLD	4 tissues	BLD
1	12128594	-0.9	0.98	1	rs1201123	C	G	0.78	0.79	0.62	0.67			BLD	
1	12128715	-0.8	0.98	0.99	rs1201124	G	C	0.78	0.79	0.62	0.66			BLD	
1	12129176	-0.3	0.99	1	rs1201126	T	A	0.78	0.79	0.62	0.67			BLD	
1	12129504	0.0	1	1	**rs535068**	G	A	0.52	0.76	0.31	0.66			STRM, BLD	
1	12130684	1.2	0.96	0.99	rs685332	G	A	0.62	0.78	0.32	0.67		FAT, BLD	12 tissues	
1	12131143	1.6	0.93	0.96	rs482170	G	T	0.74	0.81	0.6	0.66		FAT, BLD	6 tissues	
1	12131502	2.0	0.94	0.99	rs112380046	8-mer	A	0.86	0.81	0.74	0.67			6 tissues	
1	12132213	2.7	0.92	0.99	rs1201171	A	G	0.87	0.82	0.74	0.68		BLD, CRVX	10 tissues	13 tissues

Query SNP: rs535068 and variants with $r^2 >= 0.8$

chr	pos (hg38)	Proteins bound	Motifs changed	NHGRI/EBI GWAS hits	GRASP QTL hits	Selected eQTL hits	GENCODE genes	dbSNP func annot
1	12126876		CTCF,p300			1 hit	TNFRSF8	intronic
1	12128037		4 altered motifs				TNFRSF8	intronic
1	12128425	POL24H8	Ets,Nkx2,p53			1 hit	TNFRSF8	intronic
1	12128594		Egr-1,NF-E2				TNFRSF8	intronic
1	12128715					1 hit	TNFRSF8	intronic
1	12129176		9 altered motifs				TNFRSF8	intronic
1	12129504		7 altered motifs			1 hit	TNFRSF8	intronic
1	12130684		Egr-1				TNFRSF8	intronic
1	12131143		10 altered motifs				TNFRSF8	intronic
1	12131502		8 altered motifs				TNFRSF8	intronic
1	12132213	7 bound proteins	IRC900814,Sp100		1 hit		TNFRSF8	intronic

Table S4: Predicted regulatory effects according to Haploreg database of variants rs6657275, rs1473526, rs407960, rs1818545, rs10781530 and of those in strong linkage disequilibrium with them ($r^2 \geq 0.8$).

Query SNP: rs2070729 and variants with r^2 >= 0.8

chr	pos (hg38)	Distance to associated SNP in kb	LD (r^2)	LD (D')	variant	Ref	Alt	AFR freq	AMR freq	ASN freq	EUR freq	SiPhy cons	Promoter histone marks	Enhancer histone marks	DNAse
5	132477527	-6.7	0.96	1	rs4705862	A	T	0.67	0.56	0.68	0.47		LIV	12 tissues	11 tissues
5	132483136	-1.1	0.99	-1	rs11347983	CT	C	0.24	0.44	0.31	0.52			4 tissues	BLD
5	132484229	0.0	1	1	**rs2070729**	C	A	0.66	0.56	0.68	0.48			10 tissues	BLD,SKIN
5	132490150	5.9	0.98	0.99	rs2070721	T	G	0.66	0.56	0.69	0.48		24 tissues	BRN	23 tissues
5	132492083	7.9	0.98	-0.99	rs41525648	A	G	0.24	0.44	0.32	0.52		5 tissues	20 tissues	7 tissues
5	132496822	12.6	0.98	-0.99	rs2548998	G	A	0.24	0.44	0.31	0.52		24 tissues		42 tissues

Query SNP: rs2070729 and variants with r^2 >= 0.8

chr	pos (hg38)	Proteins bound	Motifs changed	NHGRI/EBI GWAS hits	GRASP QTL hits	Selected eQTL hits	GENCODE genes	dbSNP func annot
5	132477527	CTCF,RAD21	12 altered motifs			16 hits	1.5kb 3' of C5orf56	intronic
5	132483136		5 altered motifs			13 hits	IRF1	3'-UTR
5	132484229	POL24H8		1 hit	10 hits	18 hits	IRF1	intronic
5	132490150	10 bound proteins	5 altered motifs			17 hits	IRF1	intronic
5	132492083	STAT1	Ik-2			12 hits	1.3kb 5' of IRF1	
5	132496822	6 bound proteins	6 altered motifs			12 hits	6kb 5' of IRF1	

Table S4: Predicted regulatory effects according to Haploreg database of variants rs6657275, rs1473526, rs407960, rs1818545, rs10781530 and of those in strong linkage disequilibrium with them ($r^2 \geq 0.8$).

Query SNP: rs6432018 and variants with r^2 >= 0.8

chr	pos (hg38)	Distance to associated SNP in kb	LD (r^2)	LD (D')	variant	Ref	Alt	AFR freq	AMR freq	ASN freq	EUR freq	SiPhy cons	Promoter histone marks	Enhancer histone marks	DNAse
2	9576484	-5.3	0.94	0.98	rs10169269	A	G	0.84	0.44	0.06	0.57				
2	9577516	-4.3	0.97	0.98	rs7596929	A	T	0.84	0.44	0.06	0.58			HRT, GI, BLD	HRT,BLD
2	9577979	-3.8	0.96	0.98	rs7556757	G	T	0.84	0.44	0.06	0.58			BRST, FAT	
2	9579671	-2.1	0.91	0.98	rs10929590	C	G	0.85	0.44	0.06	0.56				
2	9581486	-0.3	0.98	0.99	rs147266645	C	CTCA	0.86	0.44	0.06	0.58			MUS	
			0.97	0.99	rs199878165	T	TCAC	0.85	0.44	0.06	0.57			MUS	
2	9581767	0.0	1	1	**rs6432018**	C	A	0.86	0.44	0.06	0.58			MUS	GI
2	9581897	0.1	0.99	1	rs6432019	T	C	0.85	0.44	0.06	0.58				
2	9585791	4.0	0.89	0.98	rs6432020	C	T	0.85	0.44	0.06	0.56			LNG	
2	9588750	7.0	0.97	0.99	rs7565646	A	G	0.86	0.44	0.06	0.58				
2	9589579	7.8	0.97	0.99	rs13415136	A	G	0.86	0.44	0.06	0.58				
2	9589686	7.9	0.97	0.99	rs3791748	C	G	0.86	0.44	0.06	0.58				
2	9594149	12.4	0.91	0.99	rs11674769	T	C	0.86	0.43	0.06	0.56			14 tissues	7 tissues
2	9594221	12.5	0.95	0.99	rs11674777	T	A	0.86	0.44	0.06	0.57			14 tissues	
2	9595522	13.8	0.93	0.99	rs6706227	C	T	0.86	0.43	0.06	0.57			16 tissues	5 tissues
2	9595771	14.0	0.93	0.99	rs6734469	G	A	0.86	0.43	0.06	0.57			14 tissues	BLD
2	9596745	15.0	0.97	0.99	rs6432023	C	T	0.86	0.44	0.06	0.58		BLD	16 tissues	BLD
2	9598645	16.9	0.93	0.98	rs7569485	C	T	0.86	0.44	0.05	0.57			BLD	
2	9599096	17.3	0.95	0.98	rs10210225	A	G	0.86	0.45	0.06	0.58				
2	9599300	17.5	0.92	0.97	rs11887863	G	T	0.85	0.45	0.06	0.57				
2	9599426	17.7	0.94	0.98	rs1111976	T	G	0.86	0.45	0.06	0.59				
2	9600432	18.7	0.96	0.98	rs3951115	T	C	0.86	0.45	0.11	0.58				
2	9601417	19.7	0.89	0.97	rs6745213	A	G	0.86	0.44	0.06	0.56				
2	9601939	20.2	0.94	0.97	rs62119427	T	C	0.86	0.45	0.11	0.58			ADRL, GI	
2	9603005	21.2	0.84	0.93	rs6432024	A	G	0.85	0.47	0.13	0.57				ADRL
2	9607351	25.6	0.93	0.97	rs4669400	A	G	0.86	0.47	0.13	0.57			BLD	
2	9610601	28.8	0.92	0.98	rs7558086	T	C	0.86	0.48	0.13	0.59			9 tissues	
2	9613603	31.8	0.86	0.96	rs11890807	C	A	0.86	0.44	0.06	0.56		STRM, SKIN	16 tissues	11 tissues
2	9613957	32.2	0.83	0.94	rs11902264	G	C	0.86	0.44	0.06	0.56		4 tissues	17 tissues	SKIN,SKIN,SKIN
2	9614340	32.6	0.85	0.96	rs12692388	C	T	0.86	0.43	0.06	0.56			14 tissues	
2	9615237	33.5	0.89	0.96	rs7556678	T	C	0.86	0.44	0.06	0.57			15 tissues	

Table S4: Predicted regulatory effects according to Haploreg database of variants rs6657275, rs1473526, rs407960, rs1818545, rs10781530 and of those in strong linkage disequilibrium with them ($r^2 \geq 0.8$).

Query SNP: rs6432018 and variants with r^2 >= 0.8

chr	pos (hg38)	Proteins bound	Motifs changed	NHGRI/EBI GWAS hits	GRASP QTL hits	Selected eQTL hits	GENCODE genes	dbSNP func annot
2	9576484		E2A,Ets,SIRT6				7.5kb 3' of YWHAQ	
2	9577516		9 altered motifs				6.5kb 3' of YWHAQ	
2	9577979		4 altered motifs			1 hit	6kb 3' of YWHAQ	
2	9579671						4.3kb 3' of YWHAQ	
2	9581486		5 altered motifs				2.5kb 3' of YWHAQ	
			4 altered motifs				2.5kb 3' of YWHAQ	
2	9581767		AP-1,HNF4,PPAR	1 hit	3 hits	1 hit	2.2kb 3' of YWHAQ	
2	9581897		Nrf-2,TCF11::MafG,ZID				2.1kb 3' of YWHAQ	
2	9585791						YWHAQ	intronic
2	9588750		BRCA1				YWHAQ	intronic
2	9589579		6 altered motifs				YWHAQ	intronic
2	9589686		Pou2f2,STAT			1 hit	YWHAQ	intronic
2	9594149		GATA,Klf7			1 hit	YWHAQ	intronic
2	9594221		Gfi1,Gfi1b			1 hit	YWHAQ	intronic
2	9595522	4 bound proteins	Sin3Ak-20,Zic				YWHAQ	intronic
2	9595771		GATA			1 hit	YWHAQ	intronic
2	9596745		Hic1			1 hit	YWHAQ	intronic
2	9598645		9 altered motifs				YWHAQ	intronic
2	9599096		Hand1				YWHAQ	intronic
2	9599300		Hmx,Hoxa5				YWHAQ	intronic
2	9599426	SETDB1	CTCF,Nkx2,Nkx3			1 hit	YWHAQ	intronic
2	9600432		ATF3,CACD,SRF				YWHAQ	intronic
2	9601417		5 altered motifs				YWHAQ	intronic
2	9601939		Pax-5,TCF12				YWHAQ	intronic
2	9603005					1 hit	YWHAQ	intronic
2	9607351		ERalpha-a,SIRT6				YWHAQ	intronic
2	9610601		11 altered motifs				YWHAQ	intronic
2	9613603	P300,STAT3				1 hit	YWHAQ	intronic
2	9613957		5 altered motifs			1 hit	YWHAQ	intronic
2	9614340		Foxl1,HDAC2,STAT		1 hit	1 hit	YWHAQ	intronic
2	9615237		Hoxa10				YWHAQ	intronic

Table S4: Predicted regulatory effects according to Haploreg database of variants rs6657275, rs1473526, rs407960, rs1818545, rs10781530 and of those in strong linkage disequilibrium with them ($r^2 \geq 0.8$).

Query SNP: rs17149161 and variants with r^2 >= 0.8

chr	pos (hg38)	Distance to associated SNP in kb	LD (r^2)	LD (D')	variant	Ref	Alt	AFR freq	AMR freq	ASN freq	EUR freq	SiPhy cons	Promoter histone marks	Enhancer histone marks	DNAse
7	76305323	-43.6	0.86	0.97	rs7459185	G	C	0.25	0.28	0.41	0.27		SKIN, GI	16 tissues	10 tissues
7	76306199	-42.7	0.83	0.97	rs12539175	C	G	0.24	0.27	0.41	0.27			11 tissues	SKIN
7	76306258	-42.7	0.85	0.97	rs12533832	G	A	0.25	0.28	0.41	0.27			11 tissues	SKIN,BRST
7	76306305	-42.6	0.8	0.91	rs12533836	G	T	0.33	0.33	0.57	0.3		MUS	16 tissues	4 tissues
7	76307874	-41.0	0.86	0.97	rs12531348	G	C	0.25	0.28	0.41	0.27		MUS, LIV	13 tissues	PANC,MUS
7	76307895	-41.0	0.86	0.97	rs61294900	A	T	0.25	0.28	0.41	0.27		MUS	6 tissues	MUS
7	76308203	-40.7	0.85	0.97	rs56919252	T	C	0.25	0.28	0.45	0.27			4 tissues	PLCNT,THYM
7	76309280	-39.6	0.86	0.97	rs2080010	T	C	0.36	0.28	0.44	0.27			MUS, THYM	
7	76310040	-38.9	0.85	0.98	rs12532872	G	A	0.55	0.29	0.44	0.27			5 tissues	
7	76310381	-38.5	0.86	0.98	rs12532935	G	C	0.48	0.3	0.44	0.27		5 tissues	15 tissues	7 tissues
7	76314035	-34.9	0.89	0.98	rs12533978	T	C	0.25	0.28	0.45	0.28		IPSC	18 tissues	11 tissues
7	76317626	-31.3	0.89	0.98	rs6953281	C	T	0.25	0.28	0.45	0.28		19 tissues	15 tissues	19 tissues
7	76318236	-30.7	0.89	0.98	rs7787526	C	T	0.26	0.28	0.45	0.28		24 tissues	BRN	53 tissues
7	76320735	-28.2	0.86	0.94	rs199713430	A	AAC	0.33	0.31	0.45	0.29			MUS	
7	76320736	-28.2	0.91	0.97	rs201656159	A	AC	0.38	0.32	0.45	0.28			MUS	
7	76321913	-27.0	0.95	0.97	rs7789940	A	G	0.31	0.31	0.45	0.29			FAT, MUS, SKIN	IPSC
7	76323980	-24.9	0.95	0.97	rs758944	C	A	0.31	0.31	0.45	0.29			4 tissues	
7	76329871	-19.0	0.99	0.99	rs2072435	G	A	0.31	0.31	0.45	0.29			9 tissues	
7	76332330	-16.6	1	1	rs6948661	T	C	0.31	0.31	0.45	0.29			16 tissues	12 tissues
7	76335899	-13.0	0.98	0.99	rs73140051	G	A	0.31	0.3	0.45	0.29			8 tissues	
7	76339221	-9.7	0.93	1	rs142158671	G	12-mer	0.3	0.29	0.43	0.28			9 tissues	
7	76341312	-7.6	0.99	1	rs57172088	G	A	0.32	0.3	0.45	0.29			15 tissues	
7	76345796	-3.1	0.99	1	rs7796797	A	C	0.31	0.31	0.45	0.29			6 tissues	
7	76346269	-2.6	1	1	rs7779014	C	T	0.31	0.31	0.45	0.29			15 tissues	13 tissues
7	76346460	-2.5	1	1	rs73140055	A	T	0.21	0.31	0.45	0.29			15 tissues	7 tissues
7	76348912	0.0	1	1	**rs17149161**	C	A	0.31	0.31	0.45	0.29			5 tissues	SKIN
7	76356056	7.1	0.99	1	rs11765693	A	G	0.31	0.31	0.45	0.29		BRN, MUS, HRT	15 tissues	BLD,SKIN
7	76357473	8.6	1	1	rs73140069	T	A	0.31	0.31	0.45	0.29		23 tissues	4 tissues	9 tissues
7	76365704	16.8	1	1	rs76024966	C	T	0.3	0.31	0.45	0.29		FAT	17 tissues	

Table S4: Predicted regulatory effects according to Haploreg database of variants rs6657275, rs1473526, rs407960, rs1818545, rs10781530 and of those in strong linkage disequilibrium with them ($r^2 \geq 0.8$).

Query SNP: rs17149161 and variants with r^2 >= 0.8

chr	pos (hg38)	Proteins bound	Motifs changed	NHGRI/EBI GWAS hits	GRASP QTL hits	Selected eQTL hits	GENCODE genes	dbSNP func annot
7	76305323	POL2	PRDM1		2 hits	2 hits	1kb 3' of HSPB1	
7	76306199		ATF3,CDP,SREBP			2 hits	1.9kb 3' of HSPB1	
7	76306258		7 altered motifs			2 hits	2kb 3' of HSPB1	
7	76306305		COMP1,GCM			1 hit	2kb 3' of HSPB1	
7	76307874		GATA,GCM,p53			2 hits	3.6kb 3' of HSPB1	
7	76307895					2 hits	3.6kb 3' of HSPB1	
7	76308203					2 hits	3.9kb 3' of HSPB1	
7	76309280		Maf			1 hit	5kb 3' of HSPB1	
7	76310040		4 altered motifs				5.7kb 3' of HSPB1	
7	76310381	ERALPHA_A,CMYC,HAE2F1	CTCF,EBF		1 hit	1 hit	6.1kb 3' of HSPB1	
7	76314035	PU1	Ets,SRF,STAT			5 hits	9.7kb 3' of HSPB1	
7	76317626	HMGN3,POL2	TFII-I			4 hits	9.2kb 3' of YWHAG	
7	76318236	33 bound proteins	AP-1,ZBTB33			4 hits	8.6kb 3' of YWHAG	
7	76320735		18 altered motifs				6.1kb 3' of YWHAG	
7	76320736		20 altered motifs			1 hit	6.1kb 3' of YWHAG	
7	76321913		Nkx2	1 hit		7 hits	4.9kb 3' of YWHAG	
7	76323980		SIX5,Sox,Znf143	1 hit	4 hits	7 hits	2.8kb 3' of YWHAG	
7	76329871		BCL,STAT			7 hits	YWHAG	synonymous
7	76332330	POL2,CMYC	4 altered motifs			7 hits	YWHAG	intronic
7	76335899		5 altered motifs			7 hits	YWHAG	intronic
7	76339221		5 altered motifs			1 hit	YWHAG	intronic
7	76341312					4 hits	YWHAG	intronic
7	76345796		HES1,NRSF,SMC3			6 hits	YWHAG	intronic
7	76346269	POL2,POL24H8,POL2B	Nkx3	1 hit	1 hit	6 hits	YWHAG	intronic
7	76346460		BarhI1			4 hits	YWHAG	intronic
7	76348912	STAT3	PLZF,TATA	1 hit	2 hits	6 hits	YWHAG	intronic
7	76356056	POL2B	HDAC2		1 hit	6 hits	YWHAG	intronic
7	76357473	CEBPB	5 altered motifs			6 hits	YWHAG	intronic
7	76365704		5 altered motifs			5 hits	6.7kb 5' of YWHAG	

Table S4: Predicted regulatory effects according to Haploreg database of variants rs6657275, rs1473526, rs407960, rs1818545, rs10781530 and of those in strong linkage disequilibrium with them ($r^2 \geq 0.8$).

Query SNP: rs10870140 and variants with r^2 >= 0.8

chr	pos (hg38)	Distance to associated SNP in kb	LD (r^2)	LD (D')	variant	Ref	Alt	AFR freq	AMR freq	ASN freq	EUR freq	SiPhy cons	Promoter histone marks	Enhancer histone marks	DNAse
9	136850355	-51.6	0.84	0.95	rs872463	A	G	0.84	0.59	0.89	0.76		14 tissues	15 tissues	8 tissues
9	136852540	-49.4	0.84	0.95	rs10123375	G	A	0.82	0.59	0.88	0.76				6 tissues
9	136854615	-47.4	0.82	0.95	rs7869864	C	T	0.87	0.6	0.88	0.76			10 tissues	GI,LIV
9	136859658	-42.3	0.86	0.96	rs2236544	G	A	0.82	0.59	0.89	0.75		4 tissues	18 tissues	10 tissues
9	136860365	-41.6	0.83	0.94	rs1806500	G	C	0.87	0.59	0.87	0.75		GI	8 tissues	
9	136860922	-41.0	0.85	0.95	rs11301266	AG	A	0.81	0.58	0.89	0.75			11 tissues	MUS
9	136862025	-39.9	0.86	0.96	rs945385	G	A	0.81	0.58	0.88	0.75			5 tissues	4 tissues
9	136865589	-36.4	0.84	0.95	rs10781517	T	C	0.94	0.6	0.89	0.76				6 tissues
9	136865807	-36.2	0.86	0.96	rs7037205	C	G	0.82	0.58	0.88	0.75		23 tissues	HRT, SPLN	7 tissues
9	136866378	-35.6	0.85	0.95	rs76229681	CG	C	0.82	0.59	0.88	0.75		24 tissues		53 tissues
9	136866465	-35.5	0.83	0.92	rs10870133	A	G	0.87	0.58	0.88	0.75				53 tissues
9	136866722	-35.2	0.87	0.93	rs4880147	G	C	0.83	0.57	0.89	0.74		24 tissues		22 tissues
9	136866812	-35.2	0.86	0.94	rs4880148	A	G	0.9	0.59	0.88	0.75		23 tissues	SPLN	7 tissues
9	136869615	-32.4	0.91	0.96	rs10747036	T	G	0.81	0.59	0.88	0.75				
9	136875889	-26.1	0.84	0.94	rs4880150	C	T	0.76	0.58	0.87	0.73			PLCNT, BLD	
9	136876385	-25.6	0.91	0.97	rs10116850	A	G	0.91	0.6	0.88	0.75				PLCNT
9	136877386	-24.6	0.83	0.94	rs7032070	C	T	0.8	0.57	0.86	0.73				
9	136877771	-24.2	0.9	0.96	rs4880071	T	G	0.91	0.59	0.89	0.75				
9	136877801	-24.2	0.9	0.96	rs4880152	G	A	0.87	0.59	0.89	0.75				
9	136879037	-22.9	0.86	0.94	rs35186197	AG	A	0.85	0.59	0.87	0.74				
9	136879524	-22.4	0.93	0.97	rs139265603	AAAC	A	0.84	0.59	0.89	0.75				
9	136879929	-22.0	0.93	0.97	rs10747037	T	A	0.82	0.59	0.89	0.74				
9	136880120	-21.8	0.92	0.97	rs10781519	T	C	0.91	0.6	0.89	0.75				
9	136880401	-21.6	0.93	0.97	rs11145913	C	T	0.86	0.58	0.89	0.75			GI, BLD	
9	136880571	-21.4	0.92	0.97	rs7048334	C	G	0.87	0.59	0.89	0.75			7 tissues	MUS
9	136880591	-21.4	0.92	0.97	rs7023607	A	G	0.91	0.6	0.89	0.75			7 tissues	
9	136880694	-21.3	0.92	0.97	rs7048473	C	T	0.87	0.59	0.89	0.75			7 tissues	BLD
9	136881482	-20.5	0.92	0.97	rs10747038	C	T	0.9	0.6	0.89	0.75				5 tissues
9	136881742	-20.2	0.91	0.96	rs3739943	G	C	0.81	0.58	0.88	0.74		6 tissues	18 tissues	12 tissues
9	136881948	-20.0	0.93	0.97	rs3739942	C	T	0.82	0.59	0.89	0.75		8 tissues	18 tissues	35 tissues
9	136882197	-19.8	0.92	0.97	rs4880153	T	C	0.91	0.61	0.89	0.75		12 tissues	17 tissues	16 tissues
9	136883586	-18.4	0.9	0.95	rs4880154	G	A	0.87	0.59	0.88	0.74			7 tissues	PLCNT,CRVX
9	136885010	-17.0	0.92	0.97	rs10781520	G	A	0.87	0.59	0.88	0.75				15 tissues
9	136886047	-15.9	0.93	0.97	rs10747039	G	A	0.81	0.59	0.88	0.74				22 tissues
9	136888347	-13.6	0.94	0.99	rs11145918	C	G	0.86	0.6	0.88	0.75		LNG, CRVX	7 tissues	
9	136898474	-3.5	0.81	1	rs2784075	T	C	0.89	0.61	0.89	0.78		BLD	8 tissues	MUS
9	136899393	-2.6	0.81	1	rs1800631	C	A	0.85	0.62	0.89	0.78		BLD	4 tissues	7 tissues
9	136899879	-2.1	0.8	1	rs2784074	C	T	0.9	0.62	0.89	0.78			5 tissues	
9	136901052	-0.9	0.81	1	rs2811757	A	G	0.89	0.62	0.89	0.78			18 tissues	6 tissues
9	136901967	0.0	1	1	**rs10870140**	T	C	0.79	0.59	0.89	0.74				25 tissues
9	136902187	0.2	0.99	1	rs10870141	G	A	0.81	0.59	0.89	0.74				12 tissues
9	136902989	1.0	0.95	0.98	rs200086429	G	GAC	0.91	0.59	0.87	0.74		BLD, GI	17 tissues	

Table S4: Predicted regulatory effects according to Haploreg database of variants rs6657275, rs1473526, rs407960, rs1818545, rs10781530 and of those in strong linkage disequilibrium with them ($r^2 \geq 0.8$).

Query SNP: rs10870140 and variants with r^2 >= 0.8

chr	pos (hg38)	Proteins bound	Motifs changed	NHGRI/EBI GWAS hits	GRASP QTL hits	Selected eQTL hits	GENCODE genes	dbSNP func annot
9	136850355	INI1	8 altered motifs			12 hits	PHPT1	intronic
9	136852540	POL2S2	ATF3		1 hit	14 hits	MAMDC4	intronic
9	136854615	POL2			2 hits	14 hits	MAMDC4	synonymous
9	136859658	HNF4A	GATA		2 hits	14 hits	MAMDC4	intronic
9	136860365		Ets,Evi-1,SIX5			11 hits	MAMDC4	intronic
9	136860922		GR,SEF-1,THAP1			5 hits	124bp 3' of MAMDC4	
9	136862025		CTCF		3 hits	15 hits	93bp 3' of EDF1	
9	136865589		9 altered motifs			9 hits	EDF1	intronic
9	136865807		MZF1::1-4,Myb			11 hits	EDF1	intronic
9	136866378	29 bound proteins	24 altered motifs			12 hits	91bp 5' of EDF1	
9	136866465	28 bound proteins	MAZ			9 hits	178bp 5' of EDF1	
9	136866722		NRSF,Sin3Ak-20,ZID			12 hits	435bp 5' of EDF1	
9	136866812					8 hits	525bp 5' of EDF1	
9	136869615		DEC,EBF,Nkx2			11 hits	3.3kb 5' of EDF1	
9	136875889		Pax-5			12 hits	6kb 5' of TRAF2	
9	136876385		4 altered motifs			9 hits	5.5kb 5' of TRAF2	
9	136877386		Mef2			10 hits	4.5kb 5' of TRAF2	
9	136877771		AP-3,HMG-IY			9 hits	4.1kb 5' of TRAF2	
9	136877801					9 hits	4.1kb 5' of TRAF2	
9	136879037		10 altered motifs			10 hits	2.9kb 5' of TRAF2	
9	136879524		4 altered motifs			3 hits	2.4kb 5' of TRAF2	
9	136879929		8 altered motifs		3 hits	14 hits	2kb 5' of TRAF2	
9	136880120		E2A,Lmo2-complex,ZEB1			9 hits	1.8kb 5' of TRAF2	
9	136880401		7 altered motifs			11 hits	1.5kb 5' of TRAF2	
9	136880571		E2A,HEY1			9 hits	1.3kb 5' of TRAF2	
9	136880591		EBF,Nanog,TLX1::NFIC			9 hits	1.3kb 5' of TRAF2	
9	136880694		SREBP,Sin3Ak-20			12 hits	1.2kb 5' of TRAF2	
9	136881482		GATA			9 hits	429bp 5' of TRAF2	
9	136881742	ELF1,USF2,NFKB	10 altered motifs		1 hit	15 hits	169bp 5' of TRAF2	
9	136881948	11 bound proteins	EBF		3 hits	14 hits	TRAF2	
9	136882197	CTCF	Hic1,Pax-4			12 hits	TRAF2	
9	136883586		ERalpha-a,Gfi1			10 hits	TRAF2	
9	136885010	4 bound proteins	5 altered motifs		4 hits	12 hits	TRAF2	
9	136886047	5 bound proteins	4 altered motifs			12 hits	TRAF2	
9	136888347		7 altered motifs			10 hits	TRAF2	intronic
9	136898474		Myf,RREB-1		1 hit	11 hits	TRAF2	intronic
9	136899393	4 bound proteins			2 hits	12 hits	TRAF2	intronic
9	136899879		7 altered motifs			11 hits	TRAF2	intronic
9	136901052		4 altered motifs			11 hits	TRAF2	intronic
9	136901967	11 bound proteins	Irf,STAT		1 hit	15 hits	TRAF2	intronic
9	136902187	4 bound proteins	AP-1,Myc			12 hits	TRAF2	intronic
9	136902989					11 hits	TRAF2	intronic

Table S4: Predicted regulatory effects according to Haploreg database of variants rs6657275, rs1473526, rs407960, rs1818545, rs10781530 and of those in strong linkage disequilibrium with them ($r^2 \geq 0.8$).

Query SNP: rs10870140 and variants with r^2 >= 0.8

chr	pos (hg38)	Distance to associated SNP in kb	LD (r^2)	LD (D')	variant	Ref	Alt	AFR freq	AMR freq	ASN freq	EUR freq	SiPhy cons	Promoter histone marks	Enhancer histone marks	DNAse
9	136903454	1.5	0.99	1	rs7047374	T	C	0.86	0.6	0.89	0.75		BLD, GI	16 tissues	10 tissues
9	136903549	1.6	0.93	0.99	rs66521080	TA	T,TT	0.83	0.59	0.86	0.73			9 tissues	5 tissues
9	136903690	1.7	0.98	1	rs7028806	G	T	0.88	0.6	0.89	0.75			9 tissues	GI,LNG
9	136904210	2.2	0.97	1	rs7868660	C	T	0.88	0.6	0.89	0.75			4 tissues	
9	136904234	2.3	0.99	1	rs7868671	C	T	0.88	0.6	0.89	0.75			4 tissues	
9	136904365	2.4	0.99	1	rs10735659	C	T	0.86	0.6	0.89	0.75				
9	136904500	2.5	0.93	0.98	rs10781521	C	T	0.85	0.59	0.87	0.74				
9	136904863	2.9	0.99	1	rs7868124	G	A	0.8	0.59	0.89	0.75			6 tissues	BLD
9	136906021	4.1	0.97	0.99	rs7854924	C	T	0.79	0.59	0.88	0.74			LNG	
9	136906798	4.8	0.97	0.99	rs4880157	C	T	0.86	0.6	0.89	0.74			4 tissues	BLD
9	136907480	5.5	0.98	0.99	rs10119096	T	C	0.86	0.6	0.89	0.74				
9	136909024	7.1	0.88	0.99	rs6560650	G	C	0.83	0.58	0.85	0.72			IPSC	
9	136909181	7.2	0.98	1	rs7388748	T	C	0.88	0.6	0.89	0.75			IPSC, SKIN, PANC	
9	136909506	7.5	0.99	1	rs4636298	G	A	0.79	0.59	0.89	0.74			4 tissues	PLCNT
9	136909822	7.9	0.99	1	rs4567138	G	A	0.81	0.59	0.89	0.74			4 tissues	
9	136909894	7.9	0.99	1	rs4448378	C	T	0.81	0.59	0.89	0.74			4 tissues	
9	136910227	8.3	0.96	1	rs4880158	T	C	0.91	0.61	0.89	0.75			IPSC, SKIN, LNG	BLD
9	136910716	8.7	0.96	1	rs4880159	T	C	0.94	0.61	0.89	0.75			IPSC, SKIN, LNG	
9	136912559	10.6	0.96	1	rs17250448	A	G	0.91	0.61	0.89	0.75			BLD	
9	136912751	10.8	0.97	1	rs17250462	TG	T	0.85	0.61	0.89	0.75			BLD	IPSC
9	136913171	11.2	0.97	1	rs13301872	A	C	0.79	0.6	0.89	0.75			BLD	
9	136913262	11.3	0.96	1	rs7872595	G	A	0.91	0.61	0.89	0.75			BLD	
9	136914187	12.2	0.95	0.99	rs4880160	T	C	0.94	0.61	0.89	0.75			BLD, SKIN, LNG	
9	136915981	14.0	0.96	0.99	rs4880161	A	G	0.89	0.61	0.89	0.75			BLD, LNG	
9	136916143	14.2	0.96	0.99	rs4880162	C	T	0.89	0.61	0.89	0.75			BLD, LNG	
9	136916200	14.2	0.96	0.99	rs4880163	G	A	0.85	0.61	0.89	0.75			BLD, BRN, LNG	
9	136917030	15.1	0.96	0.99	rs7039663	G	A	0.81	0.6	0.89	0.75			BLD, SKIN	
9	136917133	15.2	0.96	1	rs7027246	A	G	0.91	0.61	0.89	0.75			BLD, SKIN	OVRY
9	136917539	15.6	0.96	1	rs7034762	T	C	0.91	0.61	0.89	0.75			BLD, SKIN	
9	136918679	16.7	0.96	1	rs4880164	G	C	0.91	0.61	0.88	0.75			BLD	
9	136919242	17.3	0.97	0.99	rs7030079	C	G	0.91	0.6	0.89	0.74				
9	136919355	17.4	0.92	0.97	rs7048838	G	A	0.88	0.6	0.88	0.74				
9	136919403	17.4	0.95	0.98	rs7048940	G	A	0.91	0.6	0.88	0.74				
9	136923717	21.8	0.99	1	rs17250631	CTT	C	0.83	0.6	0.88	0.75			4 tissues	
9	136923720	21.8	0.9	0.99	rs201838550	TTC	T	0.81	0.58	0.85	0.73			4 tissues	
9	136931231	29.3	0.81	0.93	rs7028413	C	T	0.9	0.6	0.89	0.75			FAT	
9	136931594	29.6	0.81	0.93	rs10747040	T	C	0.9	0.6	0.89	0.75				
9	136932121	30.2	0.81	0.93	rs4880166	T	G	0.9	0.6	0.88	0.75			FAT, STRM, GI	BLD
9	136932455	30.5	0.81	0.93	rs7039147	A	G	0.9	0.6	0.88	0.75			FAT	BLD
9	136932658	30.7	0.81	0.93	rs10747041	T	C	0.9	0.6	0.88	0.75				
9	136933202	31.2	0.81	0.92	rs35603124	TA	T	0.9	0.6	0.88	0.75				
9	136934657	32.7	0.8	0.93	rs4880168	A	G	0.9	0.6	0.88	0.76			4 tissues	IPSC

Table S4: Predicted regulatory effects according to Haploreg database of variants rs6657275, rs1473526, rs407960, rs1818545, rs10781530 and of those in strong linkage disequilibrium with them ($r^2 \geq 0.8$).

Query SNP: rs10870140 and variants with r^2 >= 0.8

chr	pos (hg38)	Proteins bound	Motifs changed	NHGRI/EBI GWAS hits	GRASP QTL hits	Selected eQTL hits	GENCODE genes	dbSNP func annot
9	136903454	GR,POL2				14 hits	TRAF2	intronic
9	136903549	GR				7 hits	TRAF2	intronic
9	136903690		19 altered motifs			9 hits	TRAF2	intronic
9	136904210		12 altered motifs			10 hits	TRAF2	intronic
9	136904234		Pax-5			11 hits	TRAF2	intronic
9	136904365		4 altered motifs			11 hits	TRAF2	intronic
9	136904500		GLI,SP1			11 hits	TRAF2	intronic
9	136904863					15 hits	TRAF2	intronic
9	136906021		7 altered motifs			12 hits	TRAF2	intronic
9	136906798		SMC3,SP1			13 hits	TRAF2	intronic
9	136907480		GZF1,NRSF			11 hits	TRAF2	intronic
9	136909024		4 altered motifs			11 hits	TRAF2	intronic
9	136909181		ATF3,HNF4			9 hits	TRAF2	intronic
9	136909506		RP58,SREBP		2 hits	12 hits	TRAF2	intronic
9	136909822		6 altered motifs			12 hits	TRAF2	intronic
9	136909894		p300			12 hits	TRAF2	intronic
9	136910227		5 altered motifs		1 hit	13 hits	TRAF2	intronic
9	136910716		11 altered motifs			9 hits	TRAF2	intronic
9	136912559		6 altered motifs			10 hits	TRAF2	intronic
9	136912751		9 altered motifs			11 hits	TRAF2	intronic
9	136913171		5 altered motifs			15 hits	TRAF2	intronic
9	136913262		Glis2,HNF4			10 hits	TRAF2	intronic
9	136914187				3 hits	12 hits	TRAF2	intronic
9	136915981		7 altered motifs			10 hits	TRAF2	intronic
9	136916143		DMRT5,FAC1			13 hits	TRAF2	intronic
9	136916200					14 hits	TRAF2	intronic
9	136917030		9 altered motifs			12 hits	TRAF2	intronic
9	136917133		ATF3,ATF6,XBP-1			13 hits	TRAF2	intronic
9	136917539		4 altered motifs			10 hits	TRAF2	intronic
9	136918679		HNF4,Pax-4			10 hits	TRAF2	intronic
9	136919242					9 hits	TRAF2	intronic
9	136919355		ATF3,Sox,YY1			10 hits	TRAF2	intronic
9	136919403		GR,HNF4,VDR			8 hits	TRAF2	intronic
9	136923717		GR			3 hits	TRAF2	intronic
9	136923720		GR			10 hits	TRAF2	intronic
9	136931231		Mtf1,Zbtb3			9 hits	4.6kb 3' of TRAF2	
9	136931594		Zbtb3			10 hits	5kb 3' of TRAF2	
9	136932121		Nkx6-1,Pax-6,Pou4f3			11 hits	5kb 3' of RP11-229P13.25	
9	136932455		Crx,Pitx2			9 hits	4.7kb 3' of RP11-229P13.25	
9	136932658		ERalpha-a,Esr2,RXRA			9 hits	4.5kb 3' of RP11-229P13.25	
9	136933202		4 altered motifs			9 hits	4kb 3' of RP11-229P13.25	
9	136934657		5 altered motifs			11 hits	2.5kb 3' of RP11-229P13.25	

Table S4: Predicted regulatory effects according to Haploreg database of variants rs6657275, rs1473526, rs407960, rs1818545, rs10781530 and of those in strong linkage disequilibrium with them ($r^2 \geq 0.8$)

Query SNP: rs10781522 and variants with r^2 >= 0.8

chr	pos (hg38)	Distance to associated SNP in kb	LD (r^2)	LD (D')	variant	Ref	Alt	AFR freq	AMR freq	ASN freq	EUR freq	SiPhy cons	Promoter histone marks	Enhancer histone marks	DNAse
9	136920601	0.0	1	1	**rs10781522**	G	A	0.52	0.46	0.75	0.62				
9	136926616	6.0	0.82	0.93	rs3750512	G	A	0.66	0.48	0.86	0.63			5 tissues	SKIN
9	136927069	6.5	0.82	0.93	rs10870143	C	T	0.68	0.49	0.86	0.63				
9	136927584	7.0	0.82	0.93	rs1040246	G	C	0.67	0.49	0.86	0.63				SKIN
9	136927852	7.3	0.81	0.92	rs7045531	C	T	0.67	0.49	0.86	0.63		SKIN, MUS	12 tissues	6 tissues
9	136928090	7.5	0.82	0.93	rs908831	C	T	0.67	0.49	0.86	0.63		MUS	11 tissues	16 tissues
9	136931236	10.6	0.8	0.92	rs7047011	G	A	0.69	0.5	0.86	0.63			FAT	
9	136932498	11.9	0.8	0.92	rs7042327	A	G	0.67	0.5	0.85	0.63			FAT	
9	136933347	12.7	0.8	0.92	rs10781524	G	A	0.67	0.49	0.86	0.63				
9	136933872	13.3	0.8	0.92	rs7028946	G	A	0.67	0.49	0.86	0.63			8 tissues	LNG

Table S4: Predicted regulatory effects according to Haploreg database of variants rs6657275, rs1473526, rs407960, rs1818545, rs10781530 and of those in strong linkage disequilibrium with them ($r^2 \geq 0.8$)

Query SNP: rs10781522 and variants with r^2 >= 0.8

chr	pos (hg38)	Proteins bound	Motifs changed	NHGRI/EBI GWAS hits	GRASP QTL hits	Selected eQTL hits	GENCODE genes	dbSNP func annot
9	136920601		HDAC2,Pax-6			7 hits	TRAF2	intronic
9	136926616		Gfi1,PEBP		3 hits	5 hits	8bp 3' of TRAF2	
9	136927069		Myf			3 hits	461bp 3' of TRAF2	
9	136927584		5 altered motifs		2 hits	3 hits	976bp 3' of TRAF2	
9	136927852	MAX,USF1				3 hits	1.2kb 3' of TRAF2	
9	136928090		13 altered motifs		3 hits	5 hits	1.5kb 3' of TRAF2	
9	136931236		Mtf1			1 hit	4.6kb 3' of TRAF2	
9	136932498		LXR,NF-E2			2 hits	4.7kb 3' of RP11-229P13.25	
9	136933347					3 hits	3.8kb 3' of RP11-229P13.25	
9	136933872		GLI,Irf,NRSF			3 hits	3.3kb 3' of RP11-229P13.25	

Table S4: Predicted regulatory effects according to Haploreg database of variants rs6657275, rs1473526, rs407960, rs1818545, rs10781530 and of those in strong linkage disequilibrium with them ($r^2 \geq 0.8$)

Query SNP: rs324011 and variants with r^2 >= 0.8

chr	pos (hg38)	Distance to associated SNP in kb	LD (r^2)	LD (D')	variant	Ref	Alt	AFR freq	AMR freq	ASN freq	EUR freq	SiPhy cons	Promoter histone marks	Enhancer histone marks	DNAse
12	57103222	-5.2	0.86	0.95	rs703816	T	C	0.12	0.37	0.25	0.37			4 tissues	
12	57108399	0.0	1	1	**rs324011**	C	T	0.14	0.39	0.26	0.36			12 tissues	BLD,PANC
12	57109992	1.6	0.96	1	rs167769	C	T	0.13	0.39	0.26	0.35		23 tissues	9 tissues	13 tissues
12	57114866	6.5	0.9	0.96	rs3001427	G	T	0.18	0.42	0.3	0.35				
12	57115319	6.9	0.9	0.96	rs3122929	C	T	0.18	0.42	0.3	0.35				
12	57115786	7.4	0.89	0.95	rs3001425	C	T	0.43	0.44	0.34	0.35				
12	57118687	10.3	0.88	0.94	rs12368672	C	G	0.15	0.43	0.28	0.35		SKIN	14 tissues	SKIN,SKIN

Table S4: Predicted regulatory effects according to Haploreg database of variants rs6657275, rs1473526, rs407960, rs1818545, rs10781530 and of those in strong linkage disequilibrium with them ($r^2 \geq 0.8$)

Query SNP: rs324011 and variants with r^2 >= 0.8

chr	pos (hg38)	Proteins bound	Motifs changed	NHGRI/EBI GWAS hits	GRASP QTL hits	Selected eQTL hits	GENCODE genes	dbSNP func annot
12	57103222		BCL,Zbtb3			1 hit	STAT6	intronic
12	57108399		BCL,NF-kappaB		1 hit	3 hits	STAT6	intronic
12	57109992		HDAC2,Hsf,Pax-5	2 hits	8 hits	4 hits	STAT6	intronic
12	57114866		Irf,Pax-6			1 hit	STAT6	
12	57115319		GR			1 hit	STAT6	
12	57115786		5 altered motifs			2 hits	STAT6	
12	57118687		Zbtb12		3 hits	4 hits	STAT6	

Abbreviations for tables S4

Position: base-pair localization for each SNP on chromosome 19 (Genome version: GRCh38)
eQTL: expression Quantitative trait loci
bp: base pair
chr: chromosome
Populations from 1000 genomes project: AFR: African. AMR: American. ASN: Asian. EUR: European.
SiPhy: http://portals.broadinstitute.org/genome_bio/siphy/index.html
NHGRI/EBI: GWAS Catalog. https://www.ebi.ac.uk/gwas/
GRASP QTL: Genome-Wide Repository of Associations Between SNPs and Phenotypes. https://grasp.nhlbi.nih.gov
GENCODE: Encyclopædia of genes and gene variants. https://www.gencodegenes.org/
dbSNP: SNP data base from NCBI. https://www.ncbi.nlm.nih.gov/snp/
LD: Linkage of disequilibrium

Table S5: Variants associated with increased risk to endemic pemphigus foliaceus, their proxy SNPs and eQTL and sQTL effect of these variants

Associated SNP - Genomic position*	Location	Model	OR (95%CI) p-value	eQTL affected gene	Tissue	Proxy eQTL SNP	effect *p*-value
B-cell modulation							
rs6657275*G Chr1:218596461	intron of *TGFB2*	REC	2.26 (1.33-3.84) p = 2.6x10-3	*TGFB2*	Lung,Testis,Brain - Cerebellum,Brain - Parietal lobe	rs6684205, rs6604615, rs1108548, rs1342590, rs1418553, rs4846479, rs1890995, rs10482792, rs10482795, rs6657698, rs10429950, rs4846483	eQTL<$8.2x10^{-3}$
				TGFB2-AS1	Whole blood,Muscle skeletal	rs6604615	eQTL:$2.8x10^{-9}$
V(D)J rearrangement							
rs1818545*A Chr11:36612090	Intergenic between *RAG1*, *RAG2* and *HEPIS*	DOM	1.85 (1.22-2.81) p = 3.6 x10-3	*HEPIS*	Brain	rs7104753, rs11827987, rs7942300, rs12804142, rs7113441, rs1105593, rs12283331, rs16926234, rs10836583	eQTL<$7.1x10^{-3}$
rs10781530*A Chr9:139885948	921bp 5' of *PAXX*	ADD	1.58 (1.16-2.15) p = 3.6x10-3	*PAXX*	Whole blood, Lung, Heart, Brain, Artery, Adipose	rs10781529, rs10870156, rs10870157, rs10870158, rs7860430	eQTL<$8.9x10^{-5}$
				PAXX	Whole Blood, Tissue, Thyroid, Mammary Tissue, Lung, Esophagus, Cultured fibroblasts, Brain, Adipose	-	sQTL<$1.1x10^{-6}$

Table S5: Variants associated with increased risk to endemic pemphigus foliaceus, their proxy SNPs and eQTL and sQTL effect of these variants

Associated SNP - Genomic position*	Location	Model	OR (95%CI) p-value	eQTL affected gene	Tissue	Proxy eQTL SNP	effect *p*-value
Class switch recombiantion and somatic hypermutation							
rs10870140*G Chr9:139796419	intron of *TRAF2*	REC	1.76 (1.19-2.61) p = 4.9x10-3	*TRAF2*	Whole blood	rs10870141, rs4567138, rs4448378	eQTL<1.2x10^{-9}
				PAXX	Brain	rs10870141, rs4567138, rs4448378	eQTL<6.5 x10^{-6}
				TRAF2	Brain	-	sQTL:4.1x10^{-10}
				PAXX	Brain	-	sQTL<3.6x10^{-7}
rs10781522*A Chr9:139815053	intron of *TRAF2*	ADD	1.61 (1.22-2.14) p = 9x10-3	*TRAF2*	Whole blood, Colon Transverse, Testis, Lymphoblastoid cell, Monocytes, Muscle skeletal, Skin - Not sun exposed , Tibial Nerve, Dendritic cells	rs7045531, rs10870143, rs7042327, rs1040246, rs908831, rs10781524, rs7028946, rs7047011, rs2784069, rs3750512, rs17250694, rs7852970	eQTL<3.5 x10^{-4}
		REC	1.99 (1.30-3.05) p = 1.5x10-3	*PAXX*	Blood, Brain - Cerebellum	rs10781524, rs908831, rs3750512, rs7028946, rs3750512	eQTL<1.2 x10^{-3}
rs535068*A Chr1:12189561	intron of *TNFRSF8*	DOM	3.11 (1.46-6.62) p = 3.2x10-3	*TNFRSF1B*	Whole blood, Skin - Sun exposed (Lower leg), Brain - Cerebellum	rs1148475, rs1201171, rs685332, rs482170, rs1201124, rs1201122	eQTL<3.1 x10^{-3}
rs324011*A Chr12:57502182	intron of *STAT6*	ADD	1.56 (1.16-2.10) p = 3.3x10-3	*STAT6*	Whole blood, CD4+ lymphocytes, Monocytes, Brain, Liver, Colon Sigmoid	rs12368672, rs167769, rs3001425, rs3001427, rs3122929, rs703816	eQTL<1.4 x10^{-4}
		DOM	1.94 (1.30-2.90) p = 1.3x10-3	*STAT6*	Whole Blood, Thyroid, Stomach, Small Intestine, Skin - Sun Exposed, Skin - Not Sun Exposed, Prostate, Nerve - Tibial, Mammary Tissue, Lung, Esophagus, Cultured fibroblasts, Colon, Artery, Adipose	-	sQTL<1.3x10^{-7}

Table S5: Variants associated with increased risk to endemic pemphigus foliaceus, their proxy SNPs and eQTL and sQTL effect of these variants

Associated SNP - Genomic position*	Location	Model	OR (95%CI) p-value	eQTL affected gene	Tissue	Proxy eQTL SNP	effect *p*-value
rs2070729*C Chr5:131819921	intron of *IRF1*	ADD	1.56 (1.17-2.09) p = 2.8x10-3	*IRF1*	Whole blood, Heart - Left ventricle, Monocytes	rs4705862, rs2070721, rs11347983, rs41525648, rs2548998	eQTL<4.7 x10^{-4}
				IRF1-AS1	Whole blood, Thyroid, Spleen, Skin - Sun exposed, Skin - Not sun exposed, Nerve - Tibial, Muscle skeletal, Lung, Heart, Esophagus, Colon, Brain, Artery, Adipose	rs4705862, rs2070721, rs11347983, rs41525648, rs2548999	eQTL<1.1 x10^{-4}
		REC	2.02 (1.25-3.27) p = 4.4x10-3	*IL13*	Nerve - Tibial	rs4705862, rs2070721, rs11347983, rs41525648, rs2549000	eQTL<4.2 x10^{-5}
rs6432018*C Chr2:9721896	2.2kb 3' of *YWHAQ*	ADD	1.69 (1.28-2.24) p = 3x10-3	*YWHAQ*	Artery - Tibial	rs6432018, rs6432019, rs147266645, rs7596929, rs7565646, rs13415136, rs3791748, rs6432023, rs7556757, rs3951115, rs11674777, rs10210225	sQTL<1.3 x10^{-6}
		DOM	2.04 (1.32-3.14) p = 1.2 x10-3				
rs17149161*C Chr7:75978229	intron of *YWHAG*	ADD	1.63 (1.20-2.21) p = 1.7 x10-3	*YWHAG*	Monocytes, Adipose, Lung	rs76024966, rs6948661, rs7789940, rs758944, rs2072435, rs142158671, rs57172088, rs7796797, rs7779014, rs11765693, rs73140069, rs73140051, rs73140055	eQTL<1.6 x10^{-4}

Table S6: Transcript variants of genes with associated SNP with sQTL effect (Continues)

Query SNP: rs10781530 (*PAXX*), rs10870140 (*TRAF2*)

Name	Transcript ID	bp	Protein	Biotype	CCDS	UniProt	RefSeq Match	Flags
PAXX-201	ENST00000371620.4	808	204aa	Protein coding	CCDS7020	Q9BUH6	NM_183241.3	TSL:1GENCODE basicAPPRIS P1MANE Select v0.8
PAXX-208	ENST00000498095.4	1077	No protein	Retained intron	-	-	-	TSL:5
PAXX-207	ENST00000493968.5	958	No protein	Retained intron	-	-	-	TSL:1
PAXX-204	ENST00000481187.5	794	No protein	Retained intron	-	-	-	TSL:3
PAXX-203	ENST00000467845.5	744	No protein	Retained intron	-	-	-	TSL:2
PAXX-205	ENST00000483807.5	705	No protein	Retained intron	-	-	-	TSL:5
PAXX-206	ENST00000492564.2	701	No protein	Retained intron	-	-	-	TSL:3
PAXX-202	ENST00000463765.1	496	No protein	Retained intron	-	-	-	TSL:2

Table S6: Transcript variants of genes with associated SNP with sQTL effect (Continued)

Query SNP: rs10870140 (*TRAF2*)

Name	Transcript ID	bp	Protein	Translation ID	Biotype	CCDS	UniProt	RefSeq Match	Flags
TRAF2-201	ENST00000247668.7	2266	501aa	ENSP00000247668.2	Protein coding	CCDS7013	A0A024R8H5 Q12933	NM_021138.4	TSL:1GENCODE basicAPPRIS P1MANE Select v0.8
TRAF2-204	ENST00000429509.5	823	193aa	ENSP00000406524.1	Protein coding	-	B1AMX8	-	CDS 3' incompleteTSL:3
TRAF2-203	ENST00000419057.5	709	171aa	ENSP00000405860.1	Protein coding	-	B1AMX7	-	CDS 3' incompleteTSL:3
TRAF2-202	ENST00000414589.1	680	121aa	ENSP00000397653.1	Protein coding	-	B1AMY1	-	CDS 3' incompleteTSL:3
TRAF2-208	ENST00000482854.5	852	No protein	-	Processed transcript	-	-	-	TSL:5
TRAF2-207	ENST00000474950.5	380	No protein	-	Processed transcript	-	-	-	TSL:5
TRAF2-206	ENST00000469701.1	308	No protein	-	Processed transcript	-	-	-	TSL:3
TRAF2-205	ENST00000466107.1	233	No protein	-	Processed transcript	-	-	-	TSL:3

Table S6: Transcript variants of genes with associated SNP with sQTL effect (Continued)
Query SNP: rs324011 (*STAT6*)

Name	Transcript ID	bp	Protein	Translation ID	Biotype	CCDS	UniProt	RefSeq Match	Flags
STAT6-201	ENST00000300134.8	3963	847aa	ENSP00000300134.3	Protein coding	CCDS8931	P42226	NM_003153.5	TSL:1GENCODE basicAPPRIS P1
STAT6-205	ENST00000543873.6	3119	847aa	ENSP00000438451.2	Protein coding	CCDS8931	P42226	-	TSL:2GENCODE basicAPPRIS P1
STAT6-202	ENST00000454075.7	3037	847aa	ENSP00000401486.3	Protein coding	CCDS8931	P42226	-	TSL:2GENCODE basicAPPRIS P1
STAT6-219	ENST00000556155.5	2924	847aa	ENSP00000451742.1	Protein coding	CCDS8931	P42226	-	TSL:1GENCODE basicAPPRIS P1
STAT6-203	ENST00000537215.6	2886	737aa	ENSP00000444530.2	Protein coding	CCDS53804	P42226	-	TSL:2GENCODE basic
STAT6-204	ENST00000538913.6	2879	737aa	ENSP00000445409.2	Protein coding	CCDS53804	P42226	-	TSL:2GENCODE basic
STAT6-209	ENST00000553533.2	4018	865aa	ENSP00000451546.2	Protein coding	-	H0YJH6	-	TSL:3GENCODE basic
STAT6-224	ENST00000640254.2	2728	797aa	ENSP00000491116.2	Protein coding	-	A0A1W2PNW1	-	CDS 3' incompleteTSL:5
STAT6-206	ENST00000553275.1	670	64aa	ENSP00000450732.1	Protein coding	-	G3V2L2	-	CDS 3' incompleteTSL:3
STAT6-218	ENST00000555849.5	661	159aa	ENSP00000452394.1	Protein coding	-	G3V5K5	-	CDS 3' incompleteTSL:4
STAT6-222	ENST00000557635.5	644	74aa	ENSP00000450747.1	Protein coding	-	G3V2M3	-	CDS 3' incompleteTSL:4
STAT6-208	ENST00000553499.5	603	154aa	ENSP00000451074.1	Protein coding	-	G3V370	-	CDS 3' incompleteTSL:4
STAT6-211	ENST00000554663.5	565	104aa	ENSP00000450665.1	Protein coding	-	G3V2H4	-	CDS 3' incompleteTSL:4
STAT6-207	ENST00000553397.5	553	144aa	ENSP00000452203.1	Protein coding	-	G3V568	-	CDS 3' incompleteTSL:4
STAT6-220	ENST00000556259.5	548	141aa	ENSP00000452373.1	Protein coding	-	G3V5I8	-	CDS 3' incompleteTSL:3
STAT6-215	ENST00000555318.1	465	155aa	ENSP00000450428.1	Protein coding	-	H0YIY2	-	CDS 5' and 3' incompleteTSL:3
STAT6-225	ENST00000651176.1	3869	720aa	ENSP00000498693.1	Nonsense mediated decay	-	A0A494C0T4	-	-
STAT6-212	ENST00000554764.6	3380	38aa	ENSP00000451909.1	Nonsense mediated decay	-	Q5FBW6	-	TSL:2
STAT6-213	ENST00000555104.5	516	38aa	ENSP00000450510.1	Nonsense mediated decay	-	Q5FBW6	-	TSL:4
STAT6-214	ENST00000555222.5	2185	No protein	-	Retained intron	-	-	-	TSL:2
STAT6-217	ENST00000555641.1	1804	No protein	-	Retained intron	-	-	-	TSL:5
STAT6-223	ENST00000557781.5	1569	No protein	-	Retained intron	-	-	-	TSL:2
STAT6-221	ENST00000557563.5	871	No protein	-	Retained intron	-	-	-	TSL:2
STAT6-216	ENST00000555375.5	623	No protein	-	Retained intron	-	-	-	TSL:3
STAT6-210	ENST00000554202.1	248	No protein	-	Retained intron	-	-	-	TSL:2

Table S6: Transcript variants of genes with associated SNP with sQTL effect (Continued)

Query SNP: rs6432018
(*YWHAQ*)

Name	Transcript ID	bp	Protein	Translation ID	Biotype	CCDS	UniProt	RefSeq Match	Flags
YWHAQ-202	ENST00000381844.8	2216	245aa	ENSP00000371267.4	Protein coding	CCDS1666	P27348	-	TSL:1GENCODE basicAPPRIS P1
YWHAQ-201	ENST00000238081.8	2196	245aa	ENSP00000238081.3	Protein coding	CCDS1666	P27348	NM_006826.4	TSL:1GENCODE basicAPPRIS P1MANE Select v0.8
YWHAQ-203	ENST00000446619.1	665	149aa	ENSP00000398990.1	Protein coding	-	E9PG15	-	CDS 3' incompleteTSL:3
YWHAQ-205	ENST00000474715.1	709	No protein	-	Processed transcript	-	-	-	TSL:3
YWHAQ-204	ENST00000460093.1	345	No protein	-	Processed transcript	-	-	-	TSL:2

APENDICE 2

Table S1: Demographic, clinical and sample characteristics of the study individuals (Continues)

Status	Sample Code	Area of residence	Hospital of sample colection	Sex	Age	Daily prednisone dosage	ABSIS score	Antibody anti-Dsg1	Antibody anti-Dsg3
Non-Endemic Control	CT01	Curitiba	-	F	46	-	NA	NA	NA
	CT02	Curitiba	-	F	41	-	NA	NA	NA
	CT03	Curitiba	-	F	46	-	NA	NA	NA
	CT04	Curitiba	-	F	40	-	NA	NA	NA
Endemic Control	CT05	MS	-	F	37	-	NA	0	0
	CT06	MS, PR	-	M	66	-	NA	0	0
	CT07	MG	-	F	22	-	NA	NA	NA
	CT08	MG	-	F	58	-	NA	NA	NA
	CT09	MG	-	M	40	-	NA	NA	NA
	CT10	MS	-	F	41	-	NA	NA	NA
PF non-Treated	PF01	PR	Hospital Adventista do Pênfigo	M	47	-	10	241.05	0
	PF02	MG	Lar da Caridade - Hospital do Fogo Selvagem de Uberaba	F	54	-	10	180.3	0
	PF03	MG, SP	Lar da Caridade - Hospital do Fogo Selvagem de Uberaba	F	15	-	100	264.35	177.22
	PF04	SP	Hospital das Clínicas da Faculdade de Medicina de Ribeirão Preto da USP	F	28	-	1	130.91	0
	PF05	PR	Hospital de Clínicas da UFPR	M	32	-	NA	218.27	NA
PF under Treatment	PF06	MG	Lar da Caridade - Hospital do Fogo Selvagem de Uberaba	F	32	20 mg	4	159.11	0
	PF07	MG, MS, SP	Lar da Caridade - Hospital do Fogo Selvagem de Uberaba	F	49	10 mg	22.5	194.31	0
	PF08	MG, SP	Hospital Adventista do Pênfigo	M	33	20 mg	31.5	256.49	0
	PF09	GO	Hospital Adventista do Pênfigo	F	15	40 mg	65	129.89	NA
	PF10	PR	Hospital Adventista do Pênfigo	M	56	70 mg	3	342.25	NA
PF in Remission	PF11	MS	Hospital Adventista do Pênfigo	F	62	-	0	0	NA
	PF12	MS	Hospital Adventista do Pênfigo	M	70	-	0	NA	NA
	PF13	MS	Hospital Adventista do Pênfigo	F	31	-	0	0	NA
	PF14	MG	Lar da Caridade - Hospital do Fogo Selvagem de Uberaba	F	17	-	0	116.41	NA
	PF15	MG	Lar da Caridade - Hospital do Fogo Selvagem de Uberaba	F	46	-	0	0	NA
	PF16	SP	Hospital das Clínicas da Faculdade de Medicina de Ribeirão Preto da USP	M	39	-	0	61.44	NA

NA: Not applicable
NM: Not measured

Table S1: Demographic, clinical and sample characteristics of the study individuals (Continued)

Status	Sample Code	Blood volume	Plasma Volume	PBMC count	IGHM raw reads	IGHM MIGs	IGHM clonotypes	IGHG raw reads	IGHG MIGs	IGHG clonotypes
Non-Endemic Control	CT01	-	-	-	1,119,116	17,001	11,580	1,396,521	6,847	2,363
	CT02	-	-	-	1,255,130	20,106	14,346	862,058	6,859	1,510
	CT03	-	-	9,900,000	1,295,721	19,002	14,852	1,612,387	12,109	3,731
	CT04	-	-	9,225,000	1,296,115	13,648	8,296	1,580,433	12,472	3,580
Endemic Control	CT05	55	27	19,750,000	1,303,399	3,020	1,962	2,424,624	1,354	383
	CT06	39	22	6,262,500	1,299,510	1,864	1,207	1,291,335	1,851	300
	CT07	42	22	30,875,000	1,280,659	1,603	961	1,750,312	5,682	897
	CT08	41	22	36,875,000	1,298,264	2,153	1,421	2,413,344	893	243
	CT09	51	30	34,500,000	1,300,704	1,667	841	1,673,636	12,420	1,514
	CT10	51	27	33,375,000	1,297,710	1,291	953	1,848,444	1,225	431
PF non-Treated	PF01	42	22	29,375,000	1,082,058	947	656	2,052,441	579	158
	PF02	52	27	50,125,000	1,294,362	656	452	2,407,735	540	273
	PF03	42	25	51,875,000	1,294,966	4,996	3,136	2,191,654	14,151	4,531
	PF04	35	20	24,800,000	1,298,695	765	493	2,181,214	752	214
	PF05	50	25	36,750,000	1,261,336	3,267	2,062	1,665,200	3,673	1,084
PF under Treatment	PF06	52	27	27,875,000	1,306,158	1,473	876	2,344,641	6,104	1,694
	PF07	50	35	46,300,000	1,286,116	3,832	2,724	1,420,364	5,301	1,621
	PF08	41	25	14,450,000	1,299,033	3,375	2,362	1,837,497	1,928	495
	PF09	40	25	71,250,000	819,976	2,812	560	1,605,544	9,795	2,586
	PF10	40	22	48,750,000	1,305,712	870	210	1,883,036	4,794	703
PF in Remission	PF11	55	30	54,125,000	1,024,730	175	88	1,017,242	207	59
	PF12	55	27	42,375,000	1,306,435	2,125	721	1,635,061	2,023	519
	PF13	44	22	37,500,000	1,294,578	3,790	2,133	1,547,168	3,805	901
	PF14	39	22	30,375,000	1,064,137	1,932	1,329	1,814,455	1,097	320
	PF15	50	27	45,750,000	1,295,071	3,099	2,272	1,633,565	2,616	845
	PF16	35	22	11,025,000	1,282,218	245	113	3,622,314	5,556	71

MIGs: Molecular identifier groups. Reflects each immunoglobulin RNA molecule present in the sample.

Table S2: Percentage of usage of IGHV gene segments per group (continues)

Isotype	Gene Segment	Non-treated patients	Patients under treatment	Patients in Remission	Controls (endemic region)	Controls (non-endemic region)
IGHM	IGHV1-17	0.02%	0.00%	0.01%	0.03%	0.00%
	IGHV1-18	**2.44%**	**2.16%**	**2.03%**	**2.33%**	**2.99%**
	IGHV1-2	**3.81%**	**3.38%**	**3.76%**	**3.29%**	**4.25%**
	IGHV1-24	**1.00%**	**1.18%**	**1.27%**	**1.24%**	**1.26%**
	IGHV1-3	**1.01%**	0.97%	0.91%	0.71%	**1.17%**
	IGHV1-45	0.05%	0.01%	0.00%	0.03%	0.01%
	IGHV1-46	**1.31%**	**1.37%**	0.99%	**1.36%**	**1.38%**
	IGHV1-58	0.28%	0.05%	0.10%	0.21%	0.20%
	IGHV1-67	0.01%	0.00%	0.00%	0.00%	0.00%
	IGHV1-69	**5.10%**	**5.23%**	**7.00%**	**7.18%**	**9.19%**
	IGHV1-69-2	0.08%	0.01%	0.11%	0.07%	0.03%
	IGHV1-69D	0.12%	0.08%	0.07%	0.17%	0.12%
	IGHV1-8	**1.77%**	**1.54%**	**2.38%**	**1.74%**	**1.22%**
	IGHV2-26	0.21%	0.24%	0.06%	0.19%	0.22%
	IGHV2-5	**1.24%**	**1.55%**	0.68%	0.66%	**1.11%**
	IGHV2-70	0.10%	0.24%	0.13%	0.25%	0.32%
	IGHV3-11	**2.05%**	**1.91%**	**1.76%**	**2.56%**	**2.85%**
	IGHV3-13	0.44%	0.22%	0.50%	0.23%	0.32%
	IGHV3-15	**2.08%**	**1.94%**	**1.44%**	**1.66%**	**2.64%**
	IGHV3-16	0.00%	0.01%	0.00%	0.00%	0.00%
	IGHV3-20	0.48%	0.40%	0.26%	0.18%	0.22%
	IGHV3-21	**4.47%**	**4.00%**	**4.23%**	**4.25%**	**4.72%**
	IGHV3-22	0.00%	0.01%	0.00%	0.00%	0.02%
	IGHV3-23	**7.40%**	**8.02%**	**9.11%**	**8.39%**	**8.21%**
	IGHV3-30	**8.29%**	**8.80%**	**10.21%**	**8.77%**	**11.91%**
	IGHV3-33	0.23%	0.28%	0.54%	0.31%	0.34%
	IGHV3-38	0.01%	0.02%	0.00%	0.02%	0.02%
	IGHV3-43	0.41%	0.32%	0.17%	0.53%	0.20%
	IGHV3-48	**1.44%**	**1.58%**	**1.23%**	**1.28%**	**1.40%**
	IGHV3-49	0.85%	0.66%	**1.41%**	0.70%	**1.03%**
	IGHV3-52	0.00%	0.00%	0.00%	0.00%	0.01%
	IGHV3-53	**3.28%**	**3.59%**	**2.75%**	**2.94%**	**3.05%**

Frequencies higher than 1% are marked in bold

Table S2: Percentage of usage of IGHV gene segments per group (continued)

Isotype	Gene Segment	Non-treated patients	Patients under treatment	Patients in Remission	Controls (endemic region)	Controls (non-endemic region)
IGHM	IGHV3-64	0.30%	0.31%	0.09%	0.40%	0.35%
	IGHV3-65	0.00%	0.00%	0.00%	0.03%	0.00%
	IGHV3-66	0.58%	0.27%	0.52%	0.46%	0.28%
	IGHV3-7	**5.36%**	**4.70%**	**5.37%**	**3.89%**	**3.90%**
	IGHV3-72	0.48%	0.23%	0.21%	0.25%	0.26%
	IGHV3-73	0.96%	**1.20%**	0.47%	0.70%	0.59%
	IGHV3-74	**2.86%**	**2.76%**	**2.52%**	**2.31%**	**2.44%**
	IGHV3-9	**1.67%**	**1.34%**	**1.88%**	**1.32%**	**1.77%**
	IGHV4-28	**5.13%**	**3.49%**	**3.48%**	**3.97%**	**3.53%**
	IGHV4-30-2	**1.78%**	**1.33%**	**1.73%**	**1.34%**	**1.67%**
	IGHV4-31	**2.71%**	**3.60%**	**1.14%**	**3.54%**	**1.57%**
	IGHV4-34	**8.05%**	**10.91%**	**10.30%**	**11.43%**	**7.92%**
	IGHV4-39	**6.40%**	**6.96%**	**7.72%**	**6.91%**	**5.15%**
	IGHV4-4	**7.07%**	**5.79%**	**6.41%**	**6.02%**	**4.71%**
	IGHV4-55	0.01%	0.01%	0.05%	0.05%	0.01%
	IGHV4-59	0.56%	0.66%	0.56%	0.81%	0.45%
	IGHV4-61	0.01%	0.00%	0.01%	0.01%	0.00%
	IGHV5-10-1	0.13%	**1.07%**	0.26%	0.64%	0.53%
	IGHV5-51	**3.66%**	**3.76%**	**2.45%**	**2.77%**	**2.64%**
	IGHV6-1	**1.70%**	**1.53%**	**1.42%**	**1.61%**	**1.49%**
	IGHV7-27	0.00%	0.00%	0.00%	0.00%	0.00%
	IGHV7-4-1	0.62%	0.32%	0.31%	0.24%	0.35%
IGHG	IGHV1-17	0.00%	0.00%	0.00%	0.00%	0.01%
	IGHV1-18	**3.20%**	**3.75%**	**2.34%**	**3.21%**	**3.81%**
	IGHV1-2	**2.84%**	**4.56%**	**3.21%**	**4.08%**	**4.18%**
	IGHV1-24	**1.33%**	**1.90%**	**1.43%**	**1.48%**	**1.22%**
	IGHV1-3	0.96%	**1.42%**	0.74%	0.70%	**1.38%**
	IGHV1-45	0.04%	0.06%	0.00%	0.01%	0.03%
	IGHV1-46	**1.68%**	**2.17%**	**1.16%**	**2.02%**	**1.51%**
	IGHV1-58	0.21%	0.18%	0.06%	0.01%	0.10%
	IGHV1-67	0.00%	0.04%	0.04%	0.02%	0.00%

Frequencies higher than 1% are marked in bold

Table S2: Percentage of usage of IGHV gene segments per group (continued)

Isotype	Gene Segment	Non-treated patients	Patients under treatment	Patients in Remission	Controls (endemic region)	Controls (non-endemic region)
IGHG	IGHV1-69	**4.65%**	**4.79%**	**5.37%**	**5.75%**	**6.15%**
	IGHV1-69-2	0.06%	0.03%	0.06%	0.00%	0.04%
	IGHV1-69D	0.19%	0.13%	0.00%	0.09%	0.06%
	IGHV1-8	**1.40%**	0.97%	**1.59%**	**1.15%**	0.85%
	IGHV2-26	0.25%	0.20%	0.21%	0.30%	0.23%
	IGHV2-5	**1.68%**	**1.38%**	0.81%	**1.21%**	**1.23%**
	IGHV2-70	**1.20%**	0.26%	0.33%	0.36%	0.41%
	IGHV3-11	**2.39%**	**2.70%**	**2.22%**	**2.03%**	**2.56%**
	IGHV3-13	0.37%	0.42%	0.57%	0.25%	0.36%
	IGHV3-15	**2.26%**	**1.94%**	**3.52%**	**3.35%**	**3.27%**
	IGHV3-16	0.00%	0.01%	0.05%	0.00%	0.00%
	IGHV3-20	0.54%	0.93%	0.67%	0.20%	0.29%
	IGHV3-21	**3.90%**	**4.94%**	**4.34%**	**4.35%**	**4.87%**
	IGHV3-22	0.02%	0.01%	0.02%	0.01%	0.00%
	IGHV3-23	**7.94%**	**8.03%**	**10.35%**	**9.46%**	**9.34%**
	IGHV3-30	**11.84%**	**11.25%**	**9.15%**	**8.68%**	**12.02%**
	IGHV3-33	**1.59%**	**1.00%**	**1.02%**	**1.43%**	**1.08%**
	IGHV3-38	0.00%	0.03%	0.23%	0.00%	0.01%
	IGHV3-43	0.84%	0.61%	0.28%	0.68%	0.34%
	IGHV3-48	**1.90%**	**2.14%**	**2.34%**	**2.03%**	**2.56%**
	IGHV3-49	**2.36%**	0.81%	**1.72%**	**1.05%**	**1.15%**
	IGHV3-52	0.00%	0.00%	0.00%	0.02%	0.00%
	IGHV3-53	**1.20%**	**2.55%**	**2.55%**	**2.45%**	**3.94%**
	IGHV3-64	0.18%	0.36%	0.22%	0.24%	0.20%
	IGHV3-65	0.00%	0.00%	0.00%	0.00%	0.02%
	IGHV3-66	0.52%	0.18%	0.38%	0.26%	0.38%
	IGHV3-7	**2.95%**	**5.06%**	**4.89%**	**4.43%**	**5.60%**
	IGHV3-72	0.65%	0.39%	0.38%	0.80%	0.71%
	IGHV3-73	0.85%	0.85%	0.55%	0.61%	0.57%
	IGHV3-74	**3.35%**	**4.57%**	**2.37%**	**3.32%**	**3.10%**
	IGHV3-9	**1.44%**	**1.25%**	**2.24%**	**1.34%**	**1.66%**

Frequencies higher than 1% are marked in bold

Table S2: Percentage of usage of IGHV gene segments per group (continued)

Isotype	Gene Segment	Non-treated patients	Patients under treatment	Patients in Remission	Controls (endemic region)	Controls (non-endemic region)
IGHG	**IGHV4-28**	**6.20%**	**3.68%**	**2.45%**	**4.51%**	**3.00%**
	IGHV4-30-2	**1.71%**	**1.89%**	**1.75%**	**1.38%**	**1.49%**
	IGHV4-31	**3.18%**	**2.46%**	**1.73%**	**2.90%**	**1.59%**
	IGHV4-34	**3.43%**	**3.33%**	**5.82%**	**4.58%**	**3.59%**
	IGHV4-39	**5.41%**	**5.52%**	**4.39%**	**4.96%**	**4.63%**
	IGHV4-4	**6.89%**	**5.25%**	**9.76%**	**6.76%**	**5.14%**
	IGHV4-55	0.09%	0.04%	0.00%	0.01%	0.04%
	IGHV4-59	**2.75%**	**1.08%**	**2.60%**	**3.22%**	**1.65%**
	IGHV4-61	0.00%	0.02%	0.70%	0.02%	0.01%
	IGHV5-10-1	0.40%	0.86%	0.35%	0.64%	0.51%
	IGHV5-51	**1.58%**	**3.06%**	**1.79%**	**2.69%**	**2.20%**
	IGHV6-1	**1.22%**	0.75%	0.83%	0.76%	0.71%
	IGHV7-27	0.00%	0.00%	0.03%	0.00%	0.00%
	IGHV7-4-1	0.28%	0.15%	0.37%	0.14%	0.20%

Frequencies higher than 1% are marked in bold

Table S3: Clusters of clonotypes identified after network analyses (Continues)

Cluster ID	Clonotype ID	Sample	Group	Clone Count in Sample	Clone Fraction in Sample	nSeqCDR3	aaSeqCDR3
a	Cl-01	PF04,PF07	PF non-Treated , PF under Treatment	7	0.002	TGTGCGAGAGAGTGGAGTGCTTTTGATATCTGG	CAREWSAFDIW
a	Cl-02	PF09	PF under Treatment	68	0.007	TGTGCGAGAGTGATTTCCGGATTTGACTACTGG	CARVISGFDYW
a	Cl-03	PF09	PF under Treatment	22	0.002	TGTGCGAGAGTGATTTCCGGATTTGACTGCTGG	CARVISGFDCW
a	Cl-04	PF09	PF under Treatment	12	0.001	TGTGCGCGAGTGATTTCCGGATTTGACCAGTGG	CARVISGFDQW
a	Cl-05	PF09	PF under Treatment	4	0.000	TGTGCGAGAGTGGTTTCCGGATTTGACTACTGG	CARVVSGFDYW
a	Cl-06	PF09	PF under Treatment	2	0.000	TGTGCGAGAGAAACGGATGCTTTTGATATCTGG	CARETDAFDIW
a	Cl-07	PF13	PF in Remission	2	0.001	TGTGCGAGAGAGTGGAGTGCTTTTGATTTCTGG	CAREWSAFDFW
a	Cl-08	PF15	PF in Remission	100	0.038	TGTGCGCGAGAGTGGAGTAGTTTTGACTACTGG	CAREWSSFDYW
a	Cl-09	PF15	PF in Remission	21	0.008	TGTGCGCGAGAGTGGAGTGGTTTTGACTACTGG	CAREWSGFDYW
a	Cl-10	PF15	PF in Remission	19	0.007	TGTGTGAGAGAGTGGAGCGCCTTTGACTTCTGG	CVREWSAFDFW
a	Cl-11	PF15	PF in Remission	3	0.001	TGTGCGAGAGAGTGGAGCAGTTTTGACTTCTGG	CAREWSSFDFW
b	Cl-12	PF03	PF non-Treated	2	0.000	TGTGCGAGGGGCCATTGGTTCGACCCCTGG	CARGHWFDPW
b	Cl-13	PF07	PF under Treatment	1	0.000	TGTGCGAGAGCTGTGACTTCGGACTACTGG	CARAVTSDYW
b	Cl-14	PF08	PF under Treatment	5	0.003	TGTGCGAGAGGGGTGTGGTTTGACTACTGG	CARGVWFDYW
b	Cl-15	PF09	PF under Treatment	7	0.001	TGCGCGAGAGGAGTGCGCTCTGACTACTGG	CARGVRSDYW
b	Cl-16	PF12	PF in Remission	12	0.006	TGTGCGCGAGGGCGGTGGTTCGACCCCTGG	CARGRWFDPW
c1	Cl-17	PF14	PF in Remission	39	0.035	TGTACTAATAGGGATTACTGG	CTNRDYW
c1	Cl-18	PF03	PF non-Treated	3	0.000	TGTACCTTGAAAGACTACTGG	CTLKDYW
c1	Cl-19	PF03	PF non-Treated	3	0.000	TGTACACGCATGGACTACTGG	CTRMDYW
c1	Cl-20	PF03	PF non-Treated	3	0.000	TGTGCGAGAGGGGACTACTGG	CARGDYW
c1	Cl-21	PF03	PF non-Treated	2	0.000	TGCACACGGATGGACTTCTGG	CTRMDFW
c1	Cl-22	PF09	PF under Treatment	2	0.000	TGTGTGAGACCTTCAAACTGG	CVRPSNW
c1	Cl-23	CT10	Endemic Control	1	0.001	TGTGCGACTCTTGACTACTGG	CATLDYW
c1	Cl-24	PF11	PF in Remission	1	0.005	TGTGCGATTAGAGACTACTGG	CAIRDYW

Table S3: Clusters of clonotypes identified after network analyses (Continued)

Cluster ID	Sample	Clonotype ID	IGHV gene segment (Score)	IGHD gene segment (Score)	IGHJ gene segment (Score)	IGHC gene segment (Score)
a	PF04,PF07	Cl-01	IGHV3-11*00(4072)	IGHD3-3*00(25),IGHD5-12*00(25),IGHD5-18*00(25)	IGHJ3*00(470)	
a	PF09	Cl-02	IGHV3-7*00(3955.7)	IGHD2-21*00(35),IGHD2-15*00(30),IGHD4-23*00(30)	IGHJ4*00(397.2)	IGHG1*00(69.2),IGHG2*00(55.7),IGHGP*00(55.7)
a	PF09	Cl-03	IGHV3-7*00(3776.6)	IGHD2-21*00(35),IGHD2-15*00(30),IGHD4-23*00(30)	IGHJ4*00(353.9)	IGHG1*00(76.4),IGHG2*00(74.6),IGHGP*00(74.1)
a	PF09	Cl-04	IGHV3-7*00(3956.6)	IGHD2-21*00(35),IGHD2-15*00(30),IGHD4-23*00(30)	IGHJ4*00(314),IGHJ5*00(285)	IGHG1*00(74.3)
a	PF09	Cl-05	IGHV3-7*00(3883)	IGHD2-21*00(35),IGHD1-14*00(33),IGHD2-15*00(30)	IGHJ4*00(401)	IGHG1*00(65.5),IGHG2*00(65.5),IGHGP*00(65.5)
a	PF09	Cl-06	IGHV1-18*00(1146)		IGHJ3*00(490)	
a	PF13	Cl-07	IGHV4-59*00(411),IGHV4-61*00(411),IGHV4-4*00(382)	IGHD3-3*00(25),IGHD5-12*00(25),IGHD5-18*00(25)	IGHJ3*00(412)	IGHG1*00(71.5),IGHG2*00(70.5),IGHGP*00(66.5)
a	PF15	Cl-08	IGHV4-59*00(2964.6),IGHV4-61*00(2782.6),IGHV4-4*00(2618.5)	IGHD2-15*00(35),IGHD2-2*00(31),IGHD3-3*00(31)	IGHJ4*00(430)	IGHG1*00(65.6),IGHGP*00(61.5),IGHG2*00(59.8)
a	PF15	Cl-09	IGHV4-59*00(3011.1),IGHV4-61*00(2825.2)	IGHD3-3*00(45)	IGHJ4*00(401)	IGHG1*00(67),IGHG2*00(66.4),IGHGP*00(65.3)
a	PF15	Cl-10	IGHV4-59*00(2615.1),IGHV4-61*00(2538.1)	IGHD1-26*00(25),IGHD3-3*00(25),IGHD5-18*00(25)	IGHJ4*00(324)	IGHG1*00(63.1),IGHGP*00(51.5)
a	PF15	Cl-11	IGHV4-59*00(2957.7),IGHV4-61*00(2804.7),IGHV4-4*00(2580.7)	IGHD2-21*00(30),IGHD6-19*00(30)	IGHJ4*00(343),IGHJ5*00(292)	IGHG1*00(55.3),IGHG2*00(54.7),IGHGP*00(54.7)
b	PF03	Cl-12	IGHV5-51*00(3478)	IGHD5-24*00(25)	IGHJ5*00(431)	IGHG1*00(71),IGHG2*00(68),IGHGP*00(68)
b	PF07	Cl-13	IGHV3-23*00(3701)	IGHD4-17*00(31),IGHD2-21*00(30)	IGHJ4*00(400),IGHJ5*00(350)	IGHG1*00(53)
b	PF08	Cl-14	IGHV4-34*00(1638.7)	IGHD2-21*00(25),IGHD2-8*00(25)	IGHJ4*00(430)	IGHG3*00(66.3),IGHG4*00(66.3)
b	PF09	Cl-15	IGHV1-3*00(2596)	IGHD3-3*00(25)	IGHJ4*00(324)	IGHG1*00(81.2),IGHG2*00(78),IGHGP*00(75.2)
b	PF12	Cl-16	IGHV1-8*00(3058)		IGHJ5*00(431)	IGHG1*00(80.1),IGHG2*00(79),IGHGP*00(74.6)
c1	PF14	Cl-17	IGHV3-53*00(3505.5),IGHV3-66*00(3309.7)	IGHD2-8*00(30)	IGHJ4*00(371),IGHJ5*00(350)	IGHG1*00(63.9)
c1	PF03	Cl-18	IGHV3-30*00(3949),IGHV3-33*00(3833)		IGHJ4*00(400),IGHJ5*00(350)	IGHG1*00(101.5),IGHG2*00(92.5)
c1	PF03	Cl-19	IGHV2-70*00(2548)		IGHJ6*00(295)	IGHG3*00(79),IGHG4*00(79)
c1	PF03	Cl-20	IGHV1-69*00(3143),IGHV1-69D*00(2990)		IGHJ4*00(342),IGHJ5*00(292)	IGHG1*00(67),IGHG2*00(67),IGHGP*00(67)
c1	PF03	Cl-21	IGHV2-70*00(850),IGHV2-26*00(750)		IGHJ6*00(314)	
c1	PF09	Cl-22	IGHV1-69*00(2906),IGHV1-69D*00(2719)		IGHJ4*00(331),IGHJ5*00(321),IGHJ1*00(302)	IGHG1*00(69.5),IGHG2*00(69.5),IGHGP*00(62.5)
c1	CT10	Cl-23	IGHV3-23*00(3631)		IGHJ4*00(391)	IGHG3*00(69),IGHG4*00(69)
c1	PF11	Cl-24	IGHV3-23*00(4048)		IGHJ4*00(400),IGHJ5*00(350)	IGHG1*00(60),IGHG2*00(58),IGHGP*00(58)

Table S3: Clusters of clonotypes identified after network analyses (Continued)

Cluster ID	Clonotype ID	Sample	Group	Clone Count in Sample	Clone Fraction in Sample	nSeqCDR3	aaSeqCDR3
c1	Cl-25	PF13	PF in Remission	1	0.000	TGTGTTCGCGGGGACAACTGG	CVRGDNW
c2	Cl-26	CT07	Endemic Control	33	0.006	TGTGCGAGGACTAATGCTTTTGATATCTGG	CARTNAFDIW
c2	Cl-27	CT07	Endemic Control	32	0.006	TGTGCGAGAATGAATGCGTTCGACCCCTGG	CARMNAFDPW
c2	Cl-28	PF07	PF under Treatment	12	0.002	TGTGCAAGAGGTTGGGCTTTTGATGTCTGG	CARGWAFDVW
c2	Cl-29	PF06	PF under Treatment	11	0.002	TGTGCGAGAGGTCGAAGTATGGACGTCTGG	CARGRSMDVW
c2	Cl-30	PF07	PF under Treatment	10	0.002	TGTGCGAGAGACGACTACGTCGACCACTGG	CARDDYVDHW
c2	Cl-31	PF03	PF non-Treated	7	0.000	TGTACTAGGGACTACGGTATGGACGTCTGG	CTRDYGMDVW
c2	Cl-32	PF03	PF non-Treated	6	0.000	TGTGCGAGAGATTGGGGGTTTGACTCCTGG	CARDWGFDSW
c2	Cl-33	PF09	PF under Treatment	5	0.001	TGTGCGAGACGTGACTACTTTGACCTCTGG	CARRDYFDLW
c2	Cl-34	PF13	PF in Remission	4	0.001	TGTGTAAGAGGAAAAAGTATGGACGTCTGG	CVRGKSMDVW
c2	Cl-35	CT07	Endemic Control	3	0.001	TGTGCGAGAATGAATGCTTTTGATATCTGG	CARMNAFDIW
c2	Cl-36	PF03	PF non-Treated	3	0.000	TGTGTTAGGGACTATGCCATGGACGTCTGG	CVRDYAMDVW
c2	Cl-37	PF03	PF non-Treated	3	0.000	TGTGCGAGAGATTGGAGTATGGACGTCTGG	CARDWSMDVW
c2	Cl-38	PF03	PF non-Treated	2	0.000	TGTGTAAGGGACTATGGTATGGACGTCTGG	CVRDYGMDVW
c2	Cl-39	PF03	PF non-Treated	2	0.000	TGTGCGAGAGATTGGGGATGTGACTGTTGG	CARDWGCDCW
c2	Cl-40	PF03	PF non-Treated	2	0.000	TGTGCGAGAGATTGGGGGTTTGACTACTGG	CARDWGFDYW
c2	Cl-41	CT07	Endemic Control	1	0.000	TGTGGGAGGACTAATGCTTTTGATGTCTGG	CGRTNAFDVW
c2	Cl-42	CT10	Endemic Control	1	0.001	TGTGCGAAAGATTCGGGTTTTGACTACTGG	CAKDSGFDYW
c2	Cl-43	PF03	PF non-Treated	1	0.000	TGTGCGAGAGATTGGGCCTTTGACTACTGG	CARDWAFDYW
c2	Cl-44	PF03	PF non-Treated	1	0.000	TGTGCGAGAGATTGGGGCATGGACGTCTGG	CARDWGMDVW
c2	Cl-45	PF07	PF under Treatment	1	0.000	TGTGCGAGAGATGACTACTTTGACTACTGG	CARDDYFDYW
c2	Cl-46	PF09	PF under Treatment	1	0.000	TGTGCGAGAGGCCCGGCTTTTGATATCTGG	CARGPAFDIW
c3	Cl-47	PF10	PF under Treatment	20	0.004	TGTGCGAGAATATATTTTGACTACTGG	CARIYFDYW
c3	Cl-48	CT05	Endemic Control	4	0.003	TGTGCGAGAGGTCCCTTTGAATACTGG	CARGPFEYW

Table S3: Clusters of clonotypes identified after network analyses (Continued)

Cluster ID	Sample	Clonotype ID	IGHV gene segment (Score)	IGHD gene segment (Score)	IGHJ gene segment (Score)	IGHC gene segment (Score)
c1	PF13	Cl-25	IGHV3-15*00(3470)	IGHD3-10*00(28)	IGHJ4*00(313),IGHJ5*00(292),IGHJ1*00(273)	IGHG1*00(67),IGHG2*00(67),IGHGP*00(67)
c2	CT07	Cl-26	IGHV4-28*00(2898.2),IGHV4-39*00(2851.6),IGHV4-61*00(2625.8)	IGHD4-11*00(25),IGHD4-23*00(25),IGHD4-4*00(25)	IGHJ3*00(480)	IGHG1*00(67),IGHGP*00(60.5),IGHG2*00(57.5)
c2	CT07	Cl-27	IGHV4-28*00(2824.6),IGHV4-39*00(2792.9),IGHV4-61*00(2665.6)		IGHJ5*00(440)	IGHG1*00(66.3),IGHG2*00(60.7),IGHGP*00(57.1)
c2	PF07	Cl-28	IGHV6-1*00(3562.5)		IGHJ3*00(373)	IGHG1*00(70.8),IGHG2*00(61.9),IGHGP*00(61.9)
c2	PF06	Cl-29	IGHV3-7*00(3857.3)		IGHJ6*00(372)	IGHG1*00(63.6),IGHG2*00(62.9),IGHGP*00(61.5)
c2	PF07	Cl-30	IGHV3-7*00(3451.4)		IGHJ4*00(364),IGHJ5*00(362)	IGHG1*00(77.7),IGHG2*00(76.6),IGHGP*00(75.1)
c2	PF03	Cl-31	IGHV3-7*00(3391.5)	IGHD4-17*00(45),IGHD4-23*00(45)	IGHJ6*00(372)	IGHG1*00(82),IGHG2*00(81.5),IGHGP*00(81.5)
c2	PF03	Cl-32	IGHV3-7*00(3682)	IGHD3-16*00(30)	IGHJ4*00(401),IGHJ5*00(392)	IGHG1*00(69.3),IGHG2*00(68.8),IGHGP*00(68.8)
c2	PF09	Cl-33	IGHV3-21*00(3682)		IGHJ4*00(422),IGHJ5*00(372)	IGHGP*00(97),IGHG2*00(87)
c2	PF13	Cl-34	IGHV3-13*00(3275)	IGHD3-3*00(25)	IGHJ6*00(401)	IGHG1*00(63.5),IGHG2*00(62.5),IGHGP*00(62.5)
c2	CT07	Cl-35	IGHV4-28*00(3010),IGHV4-39*00(2947.5),IGHV4-61*00(2851)		IGHJ3*00(451)	IGHG1*00(69.5),IGHG2*00(69.5),IGHGP*00(62.5)
c2	PF03	Cl-36	IGHV3-7*00(3220)	IGHD2-2*00(30),IGHD3-10*00(30),IGHD3-22*00(30)	IGHJ6*00(353)	IGHG1*00(81),IGHG2*00(81),IGHGP*00(81)
c2	PF03	Cl-37	IGHV3-7*00(3633)	IGHD3-3*00(30)	IGHJ6*00(401)	IGHG1*00(89),IGHG2*00(89),IGHGP*00(88)
c2	PF03	Cl-38	IGHV3-7*00(3455)	IGHD3-10*00(40),IGHD5-18*00(35),IGHD5-5*00(35)	IGHJ6*00(343)	IGHG1*00(71),IGHG2*00(67),IGHGP*00(67)
c2	PF03	Cl-39	IGHV1-46*00(3522)	IGHD7-27*00(35)	IGHJ4*00(352),IGHJ5*00(340),IGHJ1*00(311)	IGHG2*00(69),IGHGP*00(69),IGHG1*00(63)
c2	PF03	Cl-40	IGHV3-7*00(3924)	IGHD3-16*00(30)	IGHJ4*00(372)	IGHG2*00(52),IGHG1*00(49)
c2	CT07	Cl-41	IGHV4-28*00(2835),IGHV4-61*00(2773),IGHV4-39*00(2763)	IGHD4-11*00(25),IGHD4-23*00(25),IGHD4-4*00(25)	IGHJ3*00(393)	IGHG1*00(67),IGHG2*00(67),IGHGP*00(67)
c2	CT10	Cl-42	IGHV3-30*00(4133),IGHV3-33*00(4017)	IGHD3-10*00(25),IGHD6-19*00(25),IGHD6-25*00(25)	IGHJ4*00(430)	IGHG2*00(63),IGHGP*00(63),IGHG1*00(57)
c2	PF03	Cl-43	IGHV3-7*00(3866)		IGHJ4*00(440)	IGHG1*00(67),IGHG2*00(67),IGHGP*00(67)
c2	PF03	Cl-44	IGHV1-46*00(3273)	IGHD3-16*00(25),IGHD7-27*00(25)	IGHJ6*00(353)	IGHG1*00(81),IGHG2*00(81),IGHGP*00(81)
c2	PF07	Cl-45	IGHV3-21*00(4213),IGHV3-48*00(3604)		IGHJ4*00(480)	IGHG3*00(67),IGHG4*00(67)
c2	PF09	Cl-46	IGHV3-11*00(1397),IGHV3-21*00(1397),IGHV3-48*00(1339)		IGHJ3*00(460)	IGHG2*00(56)
c3	PF10	Cl-47	IGHV3-66*00(3207.7),IGHV3-53*00(3120.7)		IGHJ4*00(377.8)	IGHG1*00(71.3)
c3	CT05	Cl-48	IGHV3-7*00(3688)		IGHJ4*00(411),IGHJ5*00(350)	IGHG1*00(56),IGHG2*00(56)

Table S3: Clusters of clonotypes identified after network analyses (Continued)

Cluster ID	Clonotype ID	Sample	Group	Clone Count in Sample	Clone Fraction in Sample	nSeqCDR3	aaSeqCDR3
c3	Cl-49	PF03	PF non-Treated	3	0.000	TGTGCGAGGCGATATTTTGACTACTGG	CARRYFDYW
c3	Cl-50	PF06	PF under Treatment	3	0.000	TGTGCGAGGGGTGTCTTTGACTATTGG	CARGVFDYW
c3	Cl-51	PF10	PF under Treatment	3	0.001	TGTGCGAAAGGTACGGAAGACTACTGG	CAKGTEDYW
c3	Cl-52	PF03	PF non-Treated	2	0.000	TGTGCGAGAGGCACCCATGACTACTGG	CARGTHDYW
c3	Cl-53	CT05	Endemic Control	1	0.001	TGTGCGAGACGCTACTTTGACTCCTGG	CARRYFDSW
c3	Cl-54	CT07	Endemic Control	1	0.000	TGTGCGAGATTAGTTTTTGATTATTGG	CARLVFDYW
c3	Cl-55	PF10	PF under Treatment	1	0.000	TGTGCGAGGGGGCCTAGTGACTACTGG	CARGPSDYW
c4	Cl-56	PF15	PF in Remission	14	0.005	TGTGTAAAAGATATGTCCCCGGGCGGTGCGGACGTCTGG	CVKDMSPGGADVW
c4	Cl-57	PF07	PF under Treatment	11	0.002	TGTGTAAAAGATTTGAGACCTGGGGGTGCGGACGTCTGG	CVKDLRPGGADVW
c4	Cl-58	PF15	PF in Remission	7	0.003	TGTGTAAAAGATATTACCCCGGGTGGTGCGGACGTCTGG	CVKDITPGGADVW
c4	Cl-59	CT07	Endemic Control	4	0.001	TGTACAAAAGATATAACCCCGGGGGGTGCGGACGTCTGG	CTKDITPGGADVW
c4	Cl-60	PF03	PF non-Treated	1	0.000	TGTACAAAAGATCTGTTACCTGGAGGTGCAGACGTCTGG	CTKDLLPGGADVW
c4	Cl-61	PF15	PF in Remission	1	0.000	TGTGTAAAAGATCTGTCCCCGGGCGGTGCGGACGTCTGG	CVKDLSPGGADVW
c5	Cl-62	PF10	PF under Treatment	9	0.002	TGTGCAAGAGATTTGAGTGGGCCTGACGACTACTGG	CARDLSGPDDYW
c5	Cl-63	PF08	PF under Treatment	6	0.003	TGTGCGAGACCCCTTAGTGGGAGTTTTGACTACTGG	CARPLSGSFDYW
c5	Cl-64	CT09	Endemic Control	1	0.000	TGTGCGAGAGAACTTGGGGGGGTCTTTGACTACTGG	CARELGGVFDYW
c5	Cl-65	PF02	PF non-Treated	1	0.002	TGTGCAAGAGACCTTAGTGGGAGCCGAGACTACTGG	CARDLSGSRDYW
c5	Cl-66	PF03	PF non-Treated	1	0.000	TGTGCAAGAGGATTGGAGGGGAGCTTTGACTACTGG	CARGLEGSFDYW
c5	Cl-67	PF08	PF under Treatment	1	0.001	TGTGCGAGACCCCTTGGTGGGAGTTTTGACTACTGG	CARPLGGSFDYW
c5	Cl-68	PF10	PF under Treatment	1	0.000	TGTTCAAGAGATCTGTCCGGGAGTAGGGACTACTGG	CSRDLSGSRDYW
c6	Cl-69	PF06	PF under Treatment	7	0.001	TGTGCGAGGCCTTTTGACTGG	CARPFDW
c6	Cl-70	CT10	Endemic Control	6	0.005	TGTGCGAGAGATGTGGACTGG	CARDVDW
c6	Cl-71	PF12	PF in Remission	6	0.003	TGTGCAAGAGATATCCGAATA	CARDIRI

Table S3: Clusters of clonotypes identified after network analyses (Continued)

Cluster ID	Sample	Clonotype ID	IGHV gene segment (Score)	IGHD gene segment (Score)	IGHJ gene segment (Score)	IGHC gene segment (Score)
c3	PF03	Cl-49	IGHV3-7*00(3979.7)		IGHJ4*00(431)	IGHG1*00(74),IGHG2*00(69.3)
c3	PF06	Cl-50	IGHV3-21*00(3222.5)	IGHD2-8*00(25)	IGHJ4*00(324)	IGHG1*00(65),IGHGP*00(63)
c3	PF10	Cl-51	IGHV3-7*00(3139)	IGHD2-21*00(25),IGHD4-23*00(25),IGHD6-19*00(25)	IGHJ4*00(400),IGHJ5*00(350),IGHJ1*00(331)	IGHG1*00(57),IGHG2*00(57),IGHGP*00(57)
c3	PF03CT0	Cl-52	IGHV1-8*00(4087)		IGHJ4*00(410),IGHJ5*00(350)	IGHG1*00(70),IGHG2*00(68),IGHGP*00(68)
c3	5CT0	Cl-53	IGHV3-7*00(3923)		IGHJ4*00(441)	IGHG1*00(67),IGHG2*00(67),IGHGP*00(67)
c3	7	Cl-54	IGHV3-66*00(3574),IGHV3-53*00(3429)	IGHD1-7*00(25),IGHD2-8*00(25),IGHD4-4*00(25)	IGHJ4*00(343),IGHJ5*00(311),IGHJ1*00(282)	IGHG1*00(85),IGHG2*00(83),IGHGP*00(83),IGHG3*00(69),IGHG4*00(69)
c3	PF10	Cl-55	IGHV1-3*00(349),IGHV3-30*00(323),IGHV3-66*00(323)	IGHD2-15*00(25),IGHD2-2*00(25),IGHD2-8*00(25)	IGHJ4*00(381),IGHJ5*00(321)	IGHG1*00(60),IGHG2*00(58),IGHGP*00(58)
c4	PF15	Cl-56	IGHV3-9*00(3408.7)	IGHD6-13*00(30),IGHD6-19*00(30),IGHD6-25*00(30)	IGHJ6*00(352)	IGHG1*00(64.8),IGHG2*00(60),IGHGP*00(54)
c4	PF07	Cl-57	IGHV3-9*00(3542.3)	IGHD3-16*00(30),IGHD7-27*00(30),IGHD2-21*00(26)	IGHJ6*00(381)	IGHG1*00(65.5)
c4	PF15CT0	Cl-58	IGHV3-9*00(3434)	IGHD2-21*00(35),IGHD6-19*00(35),IGHD6-25*00(35)	IGHJ6*00(294)	IGHG1*00(78),IGHGP*00(62.6)
c4	7	Cl-59	IGHV3-9*00(3558)	IGHD6-13*00(30),IGHD6-19*00(30),IGHD6-25*00(30)	IGHJ6*00(323)	IGHG1*00(73.8),IGHG2*00(72),IGHGP*00(66)
c4	PF03	Cl-60	IGHV3-9*00(3488)	IGHD4-23*00(30),IGHD1-20*00(26),IGHD1-7*00(26)	IGHJ6*00(371)	IGHG1*00(67),IGHG2*00(67),IGHGP*00(67)
c4	PF15	Cl-61	IGHV3-9*00(3303)	IGHD3-10*00(30),IGHD6-19*00(30),IGHD6-25*00(30)	IGHJ6*00(294)	IGHG1*00(67),IGHG2*00(67),IGHGP*00(67)
c5	PF10	Cl-62	IGHV3-74*00(3662.9)	IGHD1-26*00(30),IGHD3-3*00(30),IGHD2-15*00(28)	IGHJ4*00(371),IGHJ5*00(323)	IGHG1*00(72),IGHG2*00(71),IGHGP*00(69.8)
c5	PF08CT0	Cl-63	IGHV4-34*00(3580)	IGHD1-26*00(45)	IGHJ4*00(343)	IGHG1*00(64)
c5	9	Cl-64	IGHV3-23*00(3411)	IGHD3-16*00(26)	IGHJ4*00(382)	IGHG1*00(53),IGHG2*00(53),IGHGP*00(53)
c5	PF02	Cl-65	IGHV3-74*00(3486)	IGHD1-26*00(50)	IGHJ4*00(400),IGHJ5*00(350),IGHJ1*00(331)	IGHGP*00(41)
c5	PF03	Cl-66	IGHV3-74*00(4074)	IGHD3-3*00(31),IGHD3-10*00(30),IGHD3-16*00(30)	IGHJ4*00(411)	IGHG1*00(71),IGHG2*00(67),IGHGP*00(67)
c5	PF08	Cl-67	IGHV4-34*00(826)	IGHD3-16*00(38),IGHD1-26*00(35)	IGHJ4*00(343)	IGHG1*00(66),IGHGP*00(57)
c5	PF10	Cl-68	IGHV3-74*00(633),IGHV3-48*00(547)	IGHD2-15*00(38)	IGHJ4*00(400),IGHJ5*00(350),IGHJ1*00(331)	IGHG1*00(72),IGHG2*00(72),IGHGP*00(58)
c6	PF06CT1	Cl-69	IGHV4-39*00(2975.3)	IGHD3-3*00(25),IGHD3-9*00(25)	IGHJ4*00(244),IGHJ5*00(234)	IGHG1*00(65),IGHG2*00(64.1),IGHGP*00(62.1)
c6	0	Cl-70	IGHV1-46*00(2835.3)		IGHJ4*00(360),IGHJ5*00(350)	IGHG1*00(75.3),IGHG2*00(74),IGHGP*00(74)
c6	PF12	Cl-71	IGHV3-74*00(3817.5)	IGHD2-21*00(25)	IGHJ4*00(310),IGHJ5*00(310)	IGHG1*00(72.5)

Table S3: Clusters of clonotypes identified after network analyses (Continued)

Cluster ID	Clonotype ID	Sample	Group	Clone Count in Sample	Clone Fraction in Sample	nSeqCDR3	aaSeqCDR3
c6	Cl-72	CT07	Endemic Control	5	0.001	TGTGCGAGAGACGTGGTGTGG	CARDVVW
c6	Cl-73	PF06	PF under Treatment	4	0.001	TGTGCGGAAGACGTCTACTGG	CAEDVYW
c6	Cl-74	PF15	PF in Remission	2	0.001	TGTGCGAGAGATGTGCGGTGG	CARDVRW
c6	Cl-75	PF15	PF in Remission	2	0.001	TGTGCGAGAAACGTTATCTGG	CARNVIW
c7	Cl-76	PF13	PF in Remission	7	0.002	TGTGCAAGAGGGGATCAAGCTTTTGATATCTGG	CARGDQAFDIW
c7	Cl-77	PF03	PF non-Treated	5	0.000	TGTGCTAGGGACCACATGGCCTTCGACATCTGG	CARDHMAFDIW
c7	Cl-78	PF03	PF non-Treated	2	0.000	TGTGCGAGAGTCTTGGGGGCTTTTGATATCTGG	CARVLGAFDIW
c7	Cl-79	CT07	Endemic Control	1	0.000	TGTGCGAGAGGGTCTGGTGCTTTTGATATCTGG	CARGSGAFDIW
c7	Cl-80	CT09	Endemic Control	1	0.000	TGTGCGAGAGATCGAGAGGCTTTTGATTATTGG	CARDREAFDYW
c7	Cl-81	PF03	PF non-Treated	1	0.000	TGTGCGAGAGATAGGGGTGCTTTTGATATCTGG	CARDRGAFDIW
c7	Cl-82	PF05	PF non-Treated	1	0.000	TGTGCCAGAGGGCCGTATGCTTTTGATATCTGG	CARGPYAFDIW
c8	Cl-83	CT05	Endemic Control	4	0.003	TGTGTGAGAGAGGTTGCGGCGGCTGGCTCTTTCTGG	CVREVAAAGSFW
c8	Cl-84	PF03	PF non-Treated	4	0.000	TGTGTGAGAGCTATAGCAGCAGCTGGTTCTTTCTGG	CVRAIAAAGSFW
c8	Cl-85	PF03	PF non-Treated	2	0.000	TGTGTGAGAGCGATAGCAGCGGCTGCTAGCTATTGG	CVRAIAAAASYW
c8	Cl-86	CT05	Endemic Control	1	0.001	TGCGTTAGAGCTATAGCAGCTGCTGAAGGCTACTGG	CVRAIAAAEGYW
c8	Cl-87	CT10	Endemic Control	1	0.001	TGTGTGAGGGCCATAGCCTTGGCTGACAGCTACTGG	CVRAIALADSYW

Table S3: Clusters of clonotypes identified after network analyses (Continued)

Cluster ID	Sample	Clonotype ID	IGHV gene segment (Score)	IGHD gene segment (Score)	IGHJ gene segment (Score)	IGHC gene segment (Score)
c6	CT07	Cl-72	IGHV3-7*00(3966)	IGHD1-14*00(30),IGHD2-15*00(30),IGHD2-21*00(30)	IGHJ4*00(311),IGHJ5*00(311)	IGHG1*00(69.8),IGHG2*00(69.8),IGHGP*00(69.8)
c6	PF06	Cl-73	IGHV3-23*00(3162.7)		IGHJ4*00(380),IGHJ5*00(350)	IGHG1*00(68.3),IGHG2*00(67.7),IGHGP*00(67.7)
c6	PF15	Cl-74	IGHV3-7*00(3750)		IGHJ4*00(340),IGHJ5*00(340)	IGHG1*00(67),IGHG2*00(67),IGHGP*00(67)
c6	PF15	Cl-75	IGHV3-7*00(3865)	IGHD3-10*00(25),IGHD3-16*00(25),IGHD3-9*00(25)	IGHJ3*00(322)	IGHG1*00(67),IGHG2*00(67),IGHGP*00(67)
c7	PF13	Cl-76	IGHV6-1*00(3794)	IGHD7-27*00(30)	IGHJ3*00(431)	IGHG1*00(58.6)
c7	PF03	Cl-77	IGHV3-73*00(2956)	IGHD2-21*00(30),IGHD1-14*00(26)	IGHJ3*00(286)	IGHG1*00(89.4),IGHG2*00(88.6),IGHGP*00(87.4)
c7	PF03	Cl-78	IGHV4-39*00(3165.5),IGHV4-28*00(3082.5),IGHV4-61*00(3059.5)	IGHD3-16*00(30)	IGHJ3*00(460)	IGHG1*00(77),IGHG2*00(69)
c7	CT07	Cl-79	IGHV3-11*00(3595)		IGHJ3*00(384)	IGHGP*00(72),IGHG1*00(69),IGHG2*00(69)
c7	CT09	Cl-80	IGHV1-46*00(3270)		IGHJ3*00(345)	IGHG1*00(69),IGHG2*00(69),IGHGP*00(67)
c7	PF03	Cl-81	IGHV1-18*00(3848)		IGHJ3*00(441)	IGHG1*00(67),IGHG2*00(67),IGHGP*00(67)
c7	PF05	Cl-82	IGHV4-59*00(3260),IGHV4-61*00(3078),IGHV4-4*00(2941)		IGHJ3*00(451)	IGHG2*00(52),IGHG1*00(43)
c8	CT05	Cl-83	IGHV3-7*00(3852.7),IGHV3-48*00(3130.7)	IGHD6-25*00(31),IGHD6-19*00(30),IGHD6-13*00(27)	IGHJ4*00(350),IGHJ5*00(350)	IGHG1*00(60),IGHGP*00(55.3)
c8	PF03	Cl-84	IGHV3-7*00(3962)	IGHD6-13*00(80)	IGHJ4*00(350),IGHJ5*00(350),IGHJ1*00(321)	IGHG1*00(74),IGHG2*00(74),IGHGP*00(74)
c8	PF03	Cl-85	IGHV3-7*00(3991)	IGHD6-25*00(60),IGHD6-13*00(52),IGHD6-19*00(52)	IGHJ4*00(351),IGHJ5*00(340),IGHJ1*00(311)	IGHG1*00(57),IGHG2*00(53),IGHGP*00(53)
c8	CT05	Cl-86	IGHV3-7*00(3162),IGHV3-48*00(2611),IGHV3-11*00(2561)	IGHD6-13*00(56),IGHD6-6*00(53),IGHD6-25*00(51)	IGHJ4*00(351),IGHJ5*00(321),IGHJ1*00(302)	IGHGP*00(65)
c8	CT10	Cl-87	IGHV3-7*00(3884)	IGHD6-19*00(37),IGHD5-18*00(35),IGHD5-5*00(35)	IGHJ4*00(351),IGHJ5*00(321)	

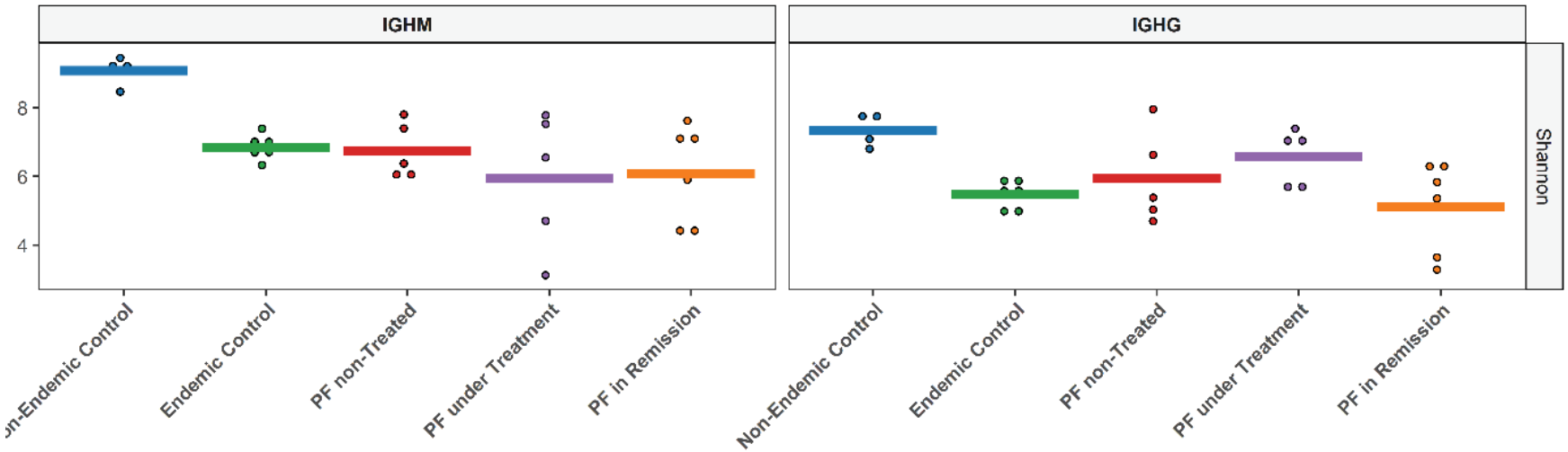


Figure S1: Shannon diversity: Alpha diversity assessed with PD (phylogenetic diversity index). Non-endemic controls and endemic controls showed significant differences for both isotypes ($p < 0.01$). Comparisons between endemic samples were non-significant. The lines represent the mean Shannon diversity value within the groups.

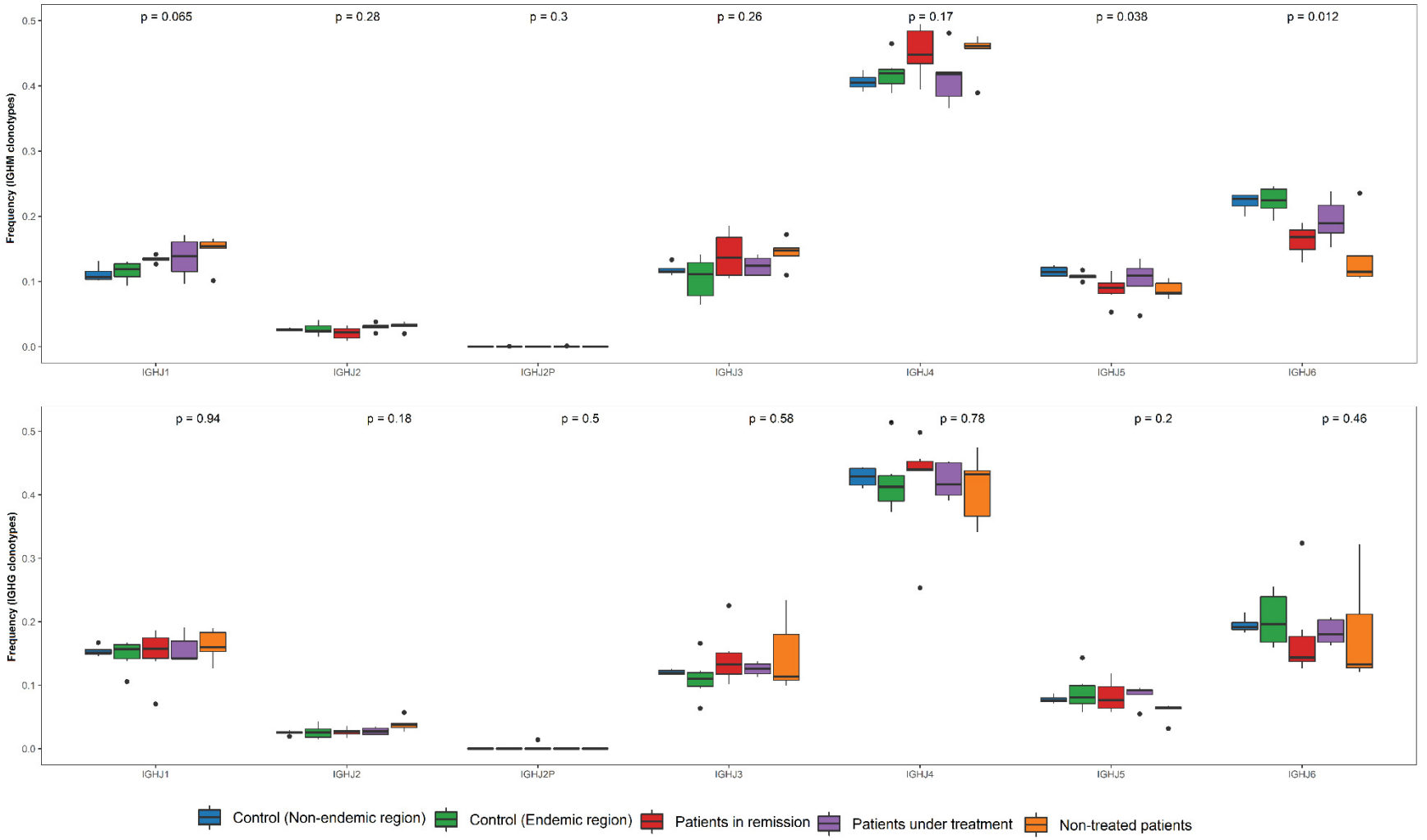


Figure S2: IGHJ gene segment usage frequency among groups. Most comparisons showed no significant differences between groups.

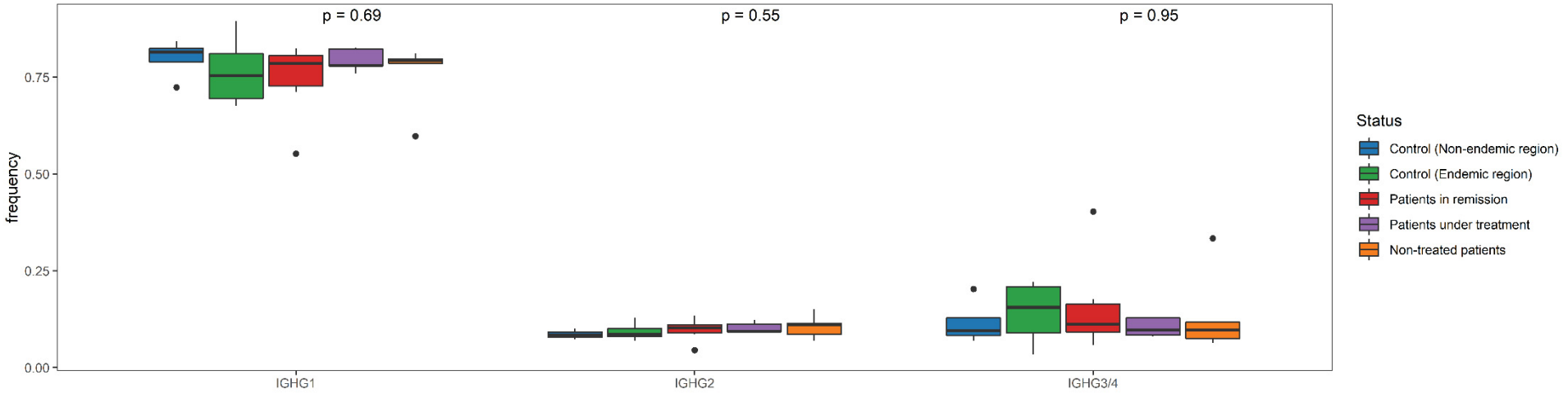


Figure S3: IGHG isotype frequency among groups showed no significant difference.