UNIVERSIDADE DE SÃO PAULO
INSTITUTO DE MATEMÁTICA E ESTATÍSTICA
PROGRAMA INTERUNIDADES DE PÓS-GRADUAÇÃO EM BIOINFORMÁTICA

SUZANE DE ANDRADE BARBOZA

Genômica comparativa aplicada no estudo da invasividade de *Streptococcus pyogenes*

São Paulo
2019

SUZANE DE ANDRADE BARBOZA

Genômica comparativa aplicada no estudo da invasividade de *Streptococcus pyogenes*

Versão corrigida contendo as alterações solicitadas pela comissão julgadora em 21 de Outubro de 2019. A versão original encontra-se em acervo reservado na Biblioteca do Instituto de Matemática e Estatística e na Biblioteca Digital de Teses e Dissertações da USP (BDTD), de acordo com a Resolução CoPGr 6018, de 13 de outubro de 2011.

Área de Concentração: Bioinformática

Orientador: Prof. Dr. Luciano Antonio Digiampietri

São Paulo

2019

FICHA CATALOGRÁFICA

Barboza, Suzane de Andrade
B239 Genômica comparativa aplicada no estudo da invasividade de streptococcus pyogenes / Suzane de Andrade Barboza, [orient.] Luciano Antonio Digiampietri. São Paulo : 2019.
85 p.

Dissertação (Mestrado) - Universidade de São Paulo
Orientador: Prof. Dr. Luciano Antonio Digiampietri
Data de Defesa: 21/10/2019

Programa Interunidades de Pós-Graduação em Bioinformática
Área de concentração: Bioinformática

1. Streptococcus pyogenes. 2. Genômica comparativa. 3. Fatores de virulência. 4. Invasividade. 5. Rede gênica. 6. Streptococcus pyogenes database. I. Digiampietri, Luciano Antônio, orientador. II. Universidade de São Paulo. III.Título.

CDD: 572.8

Elaborada pelo Serviço de Informação e Biblioteca Carlos Benjamin de Lyra do IME-USP, pela Bibliotecária Maria Lucia Ribeiro CRB-8/2766

Dissertação de autoria de Suzane de Andrade Barboza, sob o título **"Genômica comparativa aplicada no estudo da invasividade de *Streptococcus pyogenes*"**, apresentado às Unidades de Pós-graduação em Bioinformática da Universidade de São Paulo, para obtenção do título de Mestre em Ciências pelo Programa de Pós-graduação em Bioinformática, na área de concentração Bioinformática, aprovada em 21 de outubro de 2019 pela comissão julgadora constituída pelos doutores:

Prof. Dr. Luciano Antonio Digiampietri
Universidade de São Paulo
Presidente

Prof. Dr. Ariane Machado Lima
Universidade de São Paulo

Prof. Dr. Júlio Cezar Franco de Oliveira
Universidade Federal de São Paulo (Unifesp)

Prof. Dr. Fábio Sarubbi Raposo do Amaral
Universidade Federal de São Paulo (Unifesp)

Dedico este trabalho aos amores mais importantes da minha vida, Fabio, Ivana, Jorge e Rodrigo, por toda a ajuda, apoio, carinho, força (e ocasionais broncas) que me deram sempre que eu precisei.

Agradecimentos

Agradeço antes de tudo aos meus pais, Ivana e Jorge, por todo o apoio, torcida e incentivo que me dão em todas as etapas da minha vida, por me levantarem quando algo me derruba e por me ensinarem que não há obstáculo que eu não consiga vencer.

Agradeço especialmente ao meu irmão, Fabio, por ser meu melhor amigo, por entender todas as vezes que fico maluca e por demonstrar interesse em todas as etapas do meu trabalho, mesmo detestando estudar biologia.

Agradeço ao meu orientador, Prof. Dr. Luciano Antonio Digiampietri, por toda a paciência e orientação que recebi, por sempre incentivar minhas ideias, pela disponibilidade de fazer reuniões fora do horário comercial e responder meus e-mails o mais rápido possível, inclusive aos domingos e feriados.

Agradeço ao doutorando Caio Rafael do Nascimento Santiago, por todo o auxílio que me deu na parte das ferramentas de análise, pelo maravilhoso trabalho no portal e também por sua disponibilidade e rapidez em responder todas as minhas dúvidas (que não foram poucas).

Agradeço a todos os professores que contribuíram com a minha formação, principalmente aos professores do programa de bioinformática.

Agradeço ainda às minhas amigas, Amanda e Ana Paula, por compartilharem as experiências que tiveram durantes seus projetos, por todas os momentos de risadas que tivemos e por todo o apoio que me deram durante meu crescimento, tanto pessoal quanto profissional.

Por fim agradeço ao meu namorado, Rodrigo, pelo carinho, paciência e apoio que me deu durante os períodos mais difíceis, pela ajuda que me deu neste projeto e por sempre me ajudar a ver meu caminho quando muitas vezes eu mesma não consigo ver.

“Importante não é ver o que ninguém nunca viu,
mas pensar o que ninguém nunca pensou sobre algo que todo mundo vê”

(Arthur Schopenhauer)

Resumo

BARBOZA, Suzane de Andrade. **Genômica comparativa aplicada no estudo da invasividade de *Streptococcus pyogenes.*** 2019, 77 f. Dissertação (Mestrado em Ciências) – Programa Interunidades de Pós-Graduação em Bioinformática, Universidade de São Paulo, São Paulo, 2019.

Streptococcus pyogenes é um patógeno Gram positivo estritamente humano, associado a uma vasta gama de infecções invasivas e não-invasivas, ocupando o quarto lugar entre os patógenos mais associados a óbitos. A diversidade dos resultados clínicos dessas infecções pode ser explicada pela aquisição de material genético exógeno, majoritariamente composto por fatores de virulência como toxinas ou adesinas. O principal fator de virulência de *S. pyogenes* é a proteína M, cuja região hiper-variável é utilizada como chave de classificação do estreptococo. Estudos de epidemiologia molecular demonstraram a existência de uma relação entre genótipo *emm* e patogenicidade, que vem sendo extensamente investigada por meio de comparações genômicas. No entanto, pouco se sabe sobre a origem do comportamento invasivo de algumas cepas desse estreptococo. Apesar dos avanços do sequenciamento de nova geração, a grande capacidade desse patógeno para o rearranjo do genoma requer um grande número de isolados a serem comparados, o que dificulta a realização de comparações de genomas completos. Com isso em mente, foram utilizadas ferramentas baseadas na construção de uma rede gênica como uma nova abordagem para a comparação genômica de 55 isolados de *S. pyogenes*. A reanotação de todos os genomas foi essencial para a atualização e padronização dos dados, evitando que genes ortólogos fossem agrupados em diferentes famílias. Na rede gênica, essas famílias foram representadas visualmente como *clusters*. Essa nova estratégia facilitou a identificação de genes pertencentes ao *pan* e ao *core* genoma, genes exclusivos e hipotéticos que poderiam estar relacionados à virulência de *S. pyogenes*. As comparações indicaram que não há um subconjunto de genes cuja presença influenciasse unicamente e diretamente o comportamento invasivo de certos isolados, fator que deve estar relacionado à regulação dos fatores de virulência. Apesar disso, vários fatores de virulência e proteínas hipotéticas foram exclusivamente associados a cepas relacionas à mesma doença ou genótipo *emm*. O estudo dessas proteínas deve auxiliar no entendimento das relações entre genótipo e alterações fenotípicas. Por fim, 14 proteínas foram destacadas como potenciais genes-alvo para o desenvolvimento de uma vacina anti-estreptocócica não-orientada à proteína M, visando evitar o surgimento de reações imunológicas cruzadas que podem levar ao surgimento de doenças autoimunes. Todos os resultados gerados nesse estudo foram disponibilizados em uma nova base de dados (http://143.107.58.250/reportStrep2/), desenvolvida para a centralização e padronização dos dados genômicos de *S. pyogenes*. A incorporação de novos dados e implementação de novas ferramentas de genômica comparativa continuarão a ser realizadas pela equipe.

Palavras-chave: *Streptococcus pyogenes*; Genômica comparativa; Fatores de virulência; Invasividade; Rede Gênica; base de dados de *Streptococcus pyogenes*.

Abstract

BARBOZA, Suzane de Andrade. **Comparative genomics applied in the study on the invasiveness of *Streptococcus pyogenes.*** 2019, 77 p. Dissertation (Master of Science) – Bioinformatics Graduate Program, University of São Paulo, São Paulo, 2019.

Streptococcus pyogenes is a uniquely human Gram positive pathogen related to a wide range of invasive and non-invasive diseases, having the fourth highest mortality rate among bacterial pathogens. The diversity of clinical outcomes of these infections can be explained by the acquisition of exogenous genetic material, mostly composed of virulence factors such as toxins or adhesins. The major virulence factor of *S. pyogenes* is the M protein, which hypervariable region is used for its classification. Molecular epidemiology studies showed a relation between M genotype and pathogenicity, which is being intensively investigated by genomic comparisons. However, little is known about the origin of the invasive behavior of some strains of this pathogen. Despite the advances of NGS sequencing, the great capacity of this pathogen for genomic rearrangements requires an elevated number of strains to be compared, which makes it difficult to perform whole-genome comparisons. With this in mind, gene network-based tools were as a new approach to perform a comparative analysis of 55 *S. pyogenes* strains. The re-annotation of all genomes was essential to update and standardize data, avoiding the separation of true orthologs into different gene families. In the network, gene families were visually represented as clusters. This new strategy facilitated the identification of pan and core genomes, exclusive and hypothetical genes that could be related to the virulence of *S. pyogenes*. The comparisons indicated that there is not a set of genes whose presence influences uniquely and directly the invasive behavior of some strains, which is probably related to the regulation of virulence factors. Despite this, several factors and hypothetical proteins were exclusively associated to strains related to the same disease or sharing the same M genotype. The study of these proteins can bring to light new connections between genotype and phenotypical divergences. Finally, 14 proteins were highlighted as potential targets for an alternative non-M protein oriented vaccine, aiming for the avoidance of cross-reactive reactions that can leave to the development of autoimmune diseases. All results created in this study are available at a new database (http://143.107.58.250/reportStrep2/), developed for the standardization and centralization of *S. pyogenes* genomic data. The incorporation of new data and implementation of new comparative genomic tools will continue to be driven by our team.

Keywords: *Streptococcus pyogenes*; Comparative genomics; Virulence factors; Invasiveness; Gene network; *Streptococcus pyogenes* database.

Lista de figuras

Lista de gráficos

Lista de quadros

Lista de tabelas

Sumário

1. Introdução

Streptococcus pyogenes (ou estreptococo do grupo A) é um patógeno Gram positivo associado a uma vasta gama de infeções, tanto invasivas quanto não-invasivas, que podem variar de uma simples faringite a infecções graves como fasciíte necrosante ou bacteriemia (BREIMAN et al., 1993; CUNNINGHAM, 2000; LAMAGNI et al., 2008). Registros epidemiológicos sobre este microrganismo evidenciaram variações nos índices de mortalidade e morbidade ao longo do tempo (HOPKINS; MACLEAN, 2006; KATZ; MORENS, 1992), indicando divergências no nível de virulência de clones de *S. pyogenes.* Esse evento pode depender de dois fatores: [1] características do hospedeiro, como estado de saúde (incluindo coinfecções), idade, sexo, contexto geográfico, etc. (EFSTRATIOU; LAMAGNI, 2017; LAMAGNI et al., 2008), e [2] aquisição e expressão dos chamados fatores de virulência: adesinas, enzimas, toxinas e produtos do metabolismo ligados diretamente ao processo infeccioso (BIDET; BONACORSI, 2014).

Com o barateamento dos processos de sequenciamento e o avanço de ferramentas voltadas à genômica comparativa, muitos genomas foram inteiramente sequenciados, buscando-se entender a relação entre esses fatores de virulência e o nível de invasividade dos isolados (BARBOZA et al., 2015; FERNANDES et al., 2017; IBRAHIM et al., 2016; MEYGRET et al., 2016; SORIANO et al., 2014). Através desses estudos foi delineada a plasticidade genômica deste estreptococo, ou seja, grande capacidade de reorganização genômica (BESSEN et al., 2015; NAKAGAWA et al., 2003). O principal evento observado é a aquisição de material genético exógeno, sendo por meio da incorporação de profagos ou ainda por meio da integração de elementos conjugativos durante coinfecção do hospedeiro (BAO et al., 2014; GREEN et al., 2005; SUMBY et al., 2005).

Embora essa característica tenha justificado a ligação de algumas cepas com certas patologias (FLORES et al., 2015; GREEN et al., 2005; MEISAL et al., 2010; RASMUSSEN; EDÉN; BJORCK, 2000), pouco ainda se sabe sobre a origem da diversidade clínica piogênica e das diferenças no nível de invasividade observadas em clones de mesmo genótipo (MEYGRET et al., 2016). Isso porque poucos estudos de genômica comparativa focaram na diferenciação de isolados relacionados a infecções invasivas e não-invasivas (FERNANDES et al., 2017; NASSER et al., 2014), raramente comparando isolados de diferentes genótipos (BAO et al., 2017).

1.1 Definição do problema

A utilização de ferramentas de genômica comparativa aplicadas no estudo de genomas completos de *S. pyogenes* pode ajudar a identificar relações diretas entre a presença de fatores de virulência e a variação no nível de invasividade observada em diversos isolados, ou ainda na procura de genes altamente conservados que possam ser usados como alvos para o desenvolvimento de uma vacina antiestreptocóccica. No entanto, este patógeno possui um grande número de genótipos, além de estar associado a uma vasta gama de doenças. Somando-se estes fatores com a grande capacidade de aquisição de material extragenômico e rearranjo do genoma que esse organismo possui, faz-se necessário conduzir este estudo com um alto número de genomas para se obter um resultado acurado.

Entretanto, quanto maior o número de isolados, mais difícil se faz o uso dessas ferramentas comparativas. Entre as principais dificuldades encontradas ao se realizar estudos que abranjam todas essas variáveis estão: [1] limitação das atuais ferramentas de comparação de genomas evolutivamente próximos, principalmente na identificação de genes exclusivamente associados a um grande número de variáveis, e [2] baixo número de estratégias eficientes na representação visual dessa grande quantidade de dados, com o intuito de facilitar a análise e a identificação de padrões.

Além disso, a falta de padronização e centralização de dados dificulta a recuperação e análise das informações necessárias. Um bom exemplo é o grande número de genomas que foram anotados por diferentes ferramentas baseadas em abordagens distintas, o que resulta na incompatibilidade de parte das informações disponíveis. Outro exemplo é o grande esforço e tempo dispendido na agregação de dados específicos sobre um grande número de genomas, como por exemplo no momento de criação de arquivos FASTA contendo os genes de virulência de todos genomas completos disponíveis na base do NCBI.

1.2 Objetivos

O principal objetivo deste estudo é a aplicação de novas ferramentas de genômica comparativa em um estudo sobre *S. pyogenes*, buscando-se compreender a origem da invasividade desse patógeno por meio da identificação de genes exclusivos.

A iniciativa do projeto se concentra em três eixos principais: [1] aplicação das ferramentas desenvolvidas pela equipe do grupo de pesquisa contextualizado no NUBIC[1] (Núcleo de Apoio à Pesquisa em Biologia Computacional e Genômica), comparando os resultados obtidos com modelos filogenéticos já publicados para parte desses genomas, [2] identificação de genes de interesse, como fatores de virulência exclusivamente relacionados à invasividade ou ainda genes alvo para o desenvolvimento de uma vacina, e [3] centralização e padronização de dados genômicos de *S. pyogenes*, por meio da criação de um portal online que reúna dados e ferramentas utilizadas no estudo deste estreptococo.

Desta forma, o projeto pretende aplicar novas ferramentas desenvolvidas especialmente para a comparação de isolados evolutivamente próximos buscando a identificação de genes potencialmente ligados a importantes características de certos clones de *S. pyogenes,* além de disponibilizar um acesso rápido às sequências e ferramentas computacionais utilizadas por meio da disponibilização de um portal exclusivo para este estreptococo.

1.3 Hipóteses iniciais

Com a utilização de novas abordagens no estudo comparativo das cepas de *S. pyogenes* espera-se determinar a origem da invasividade de certas cepas deste patógeno, bem como identificar fatores relacionados à doença e/ou ao genótipo. Entretanto, apesar de haver correlações entre alguns fatores de virulência com certas diferenças comportamentais observadas em alguns isolados, não há evidências de que a origem da invasividade de *S. pyogenes* esteja obrigatoriamente ligada a algum fator genético. Desta forma, foram elaboradas as seguintes hipóteses sobre a invasividade deste estreptococo:

[1] Pode ser explicada por um gene único, obrigatoriamente presente em todas as cepas invasivas e ausente em todas as cepas não-invasivas;

[2] Pode ser resultado de combinações de conjuntos de fatores de virulência que facilitam a adesão, invasão celular e sobrevivência do patógeno, não estando todos os fatores necessariamente presentes em todas as cepas invasivas;

[3] Não pode ser explicada pela simples presença/ausência de fatores de virulência específicos, mas está relacionada a regulação do nível de expressão destes;

[1] http://lbi.usp.br/nubic/

[4] Pode ser resultante da presença de polimorfismos que alterem a composição, expressão ou regulação de fatores de virulência, liberando maiores quantidades de toxinas ou ainda estimulando a superexpressão de efeitos inflamatórios no hospedeiro;

[5] Não tem qualquer explicação genética, sendo resultado de uma combinação das condições clínicas e imunológicas do hospedeiro com o sítio e modo de infecção. Em resumo, todas as cepas teriam igual capacidade de invasão e invasividade.

1.4 Organização do documento

Este documento está organizado da seguinte forma: o capítulo 2 foi dedicado à apresentação de conceitos fundamentais relacionados à *S. pyogenes*. A metodologia proposta para a execução deste projeto está descrita no capítulo 3. A apresentação e discussão dos resultados é feita no capítulo 4, enquanto o capítulo 5 é destinado à conclusão e definição das perspectivas deste projeto.

2. Conceitos fundamentais

Estreptococos são bactérias Gram positivas com formato esférico que não se separam completamente após a divisão celular, formando cadeias lineares de tamanho variável e imóveis. São anaeróbios facultativos, catalase negativos e separados em três grupos de acordo com perfil hemolítico: α-hemolíticos quando há lise parcial das hemácias, β-hemolíticos quando ocorre lise completa das hemácias, ou γ-hemolíticos quando não há produção de hemolisinas (MURRAY; ROSENTHAL; PFALLER, 2014).

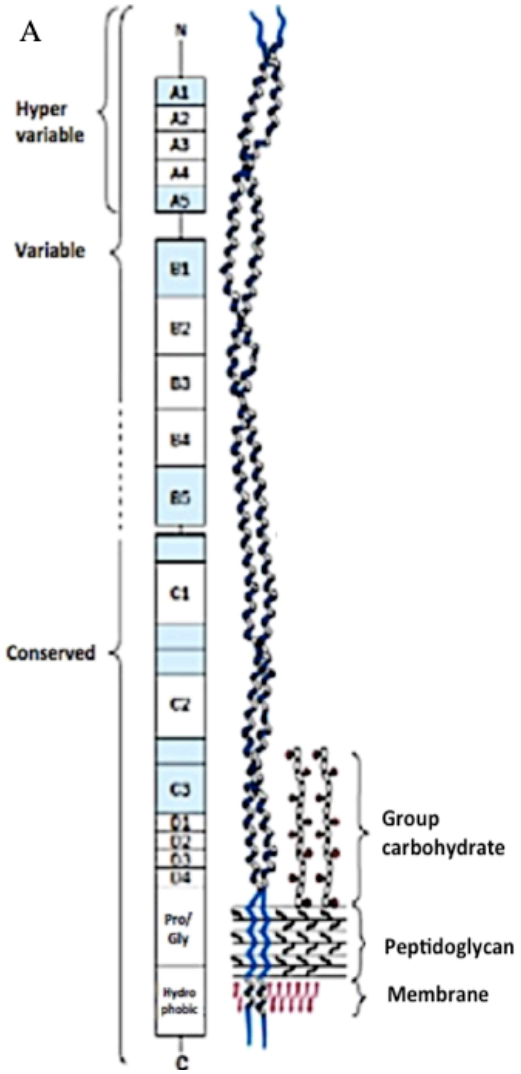
Lancefield propõe, em 1933, uma maneira eficiente de classificação dos estreptococos hemolíticos, baseada nas características antigênicas de um polissacarídeo chamado carboidrato C, presente na parede celular bacteriana (LANCEFIELD, 1933). O método consistia na realização de reações de precipitação com antissoros, que culminou na criação de 6 grupos: A, B C, D, F e N. Posteriormente mais grupos foram descritos, somando 20 grupos (A à H, e K à U), determinados pelo reconhecimento de um amino-açúcar do carboidrato C (BROOKS et al., 2014).

2.1 Estreptococos do grupo A

O *Streptococcus pyogenes* é o único membro do grupo A de estreptococos, caracterizado pela observação de suas colônias (brancas, com diâmetro inferior à 2mm), beta-hemólise, presença do amino-açúcar ramnose-N-acetilglicosamina no carboidrato C e identificação de um antígeno tipo-específico na parede celular, a proteína M (MURRAY; ROSENTHAL; PFALLER, 2014).

Com a extremidade C terminal ancorada na membrana, a proteína M atravessa a parede celular e tem sua extremidade N terminal exposta à superfície celular (Figura 1A). Sua estrutura primária é formada por quatro domínios (de A a D), divergentes em tamanho e sequência nucleotídica (MACHEBOEUF et al., 2011). A estrutura terciária compreende o dobramento de duas cadeias em alfa-hélice espiral sugerindo um mimetismo com a tropomiosina (Figura 1B), uma das proteínas regulatórias da contração do músculo estriado esquelético (BISNO; BRITO; COLLINS, 2003; MARTINS et al., 2008; MCNAMARA et al., 2008). Atualmente, essa similaridade estrutural é a única explicação para sequelas pós-infecção como o reumatismo articular agudo (MARTINS et al., 2008).

Figura 1 - Representação da estrutura terciária das proteínas M (A) e tropomiosina (B)



Fonte: (BESSEN et al., 1996)

Fonte: (BRENCIANI et al., 2011)

A proteína M é um dos principais fatores de virulência de *S. pyogenes,* tendo participação de todas as etapas do processo infeccioso (colonização, internalização celular/invasão de tecidos e difusão da infecção) por meio da ativação de mecanismos que interagem ou reprimem as defesas do hospedeiro. Sua carga negativa é responsável por uma repulsão eletrostática que reprime a fagocitose, bloqueando o componente C3b do sistema complemento (BIDET; BONACORSI, 2014; MURRAY; ROSENTHAL; PFALLER, 2014).

A proteína apresenta uma grande variabilidade, principalmente nos últimos 11 aminoácidos da extremidade N terminal, ao contrário da extremidade C terminal que se mostra relativamente conservada (BEACHEY et al., 1981). A imunidade do hospedeiro é específica e depende das características antigênicas da proteína M (BEACHEY et al., 1981),

que são a chave da classificação serológica de *S. pyogenes* desenvolvida por Lancefield. Quando criada, algumas dificuldades na tipagem de certos genótipos (FACKLAM et al., 1999) foram um impedimento na popularização desta abordagem, incentivando a criação de métodos com menor especificidade e maior praticidade (STOLLERMAN, 1997). Entretanto, com o avanço da tecnologia a genotipagem da região hipervariável por meio de sequenciamento permitiu a padronização do protocolo e identificação de mais de 200 genótipos, usados em estudos epidemiológicos e filogenéticos (MCMILLAN et al., 2013). Em publicações mais recentes essa classificação é complementada ainda com o uso de outros métodos, como agrupamento (clusterização) ou sequenciamento multi-locos (ENRIGHT et al., 2001; MCGREGOR et al., 2004; MCMILLAN et al., 2013; SANDERSON-SMITH et al., 2014), porém a genotipagem da proteína M ainda é o método de classificação mais adotado nos estudos sobre *S. pyogenes*.

2.2 A importância de S. pyogenes à saúde pública

Epidemias relacionadas ao *Streptococcus pyogenes* têm sido observadas desde o século 19, período em que a febre escarlate e sepse puerperal eram as principais causas de morbidade e mortalidade piogênicas (PHILLIPS, 1938). Foi somente em 1867 que práticas antissépticas e higiênicas foram adotadas, diminuindo então a incidência de casos em países industrializados (HENNINGHAM et al., 2012).

Em 1993 as infecções causadas por *S. pyogenes* foram classificadas entre invasivas e não invasivas (BREIMAN et al., 1993). As infecções chamadas não invasivas, principalmente cutâneo-mucosas (impetigo, equidema, angina, dermatite estreptocócica perianal, vulvovaginite, etc.) são normalmente consideradas benignas, se desprezando as eventuais complicações pós-infecção que causam sequelas ao nível imunológico, como o reumatismo articular agudo e a glomerulonefrite aguda (BOMBACI et al., 2009; CUNNINGHAM, 2000, 2008; DALE et al., 2001). Já as infecções chamadas invasivas (ou severas), podem ser supurativas (bacteriemia, febre puerperal, dermohipodermite necrosante, etc) ou não supurativas (choque tóxico estreptocócico), e são responsáveis por mais de 500.000 mortes/ano (OLIVEIRA et al., 2017), deixando *S. pyogenes* atrás apenas de *Mycobacterium tuberculosis, Streptococcus pneumoniae* e *Haemophilus influenzae* nos índices de mortalidade por infecções bacterianas (CARAPETIS et al., 2005). A maior parte dos casos fatais são resultado do desenvolvimento pós-infeccioso de cardiopatia reumática crônica (EFSTRATIOU; LAMAGNI, 2017), uma das principais sequelas imunológicas que

impossibilitaram, até hoje, o desenvolvimento de uma vacina anti-estreptocóccica (BARROS et al., 2015).

2.2.1 Epidemiologia molecular de *S. pyogenes*

A nível global, nota-se que o número de infecções invasivas vem aumentando desde os anos 80, sem que alguma explicação epidemiológica ou fisiológica tenha sido atribuída a esse fenômeno preocupante (CUNNINGHAM, 2000; SANYAHUMBI et al., 2016). Entre os tipos M fortemente associados a infecções invasivas destacam-se os tipos 1, 3, 12, 28 e 89 (CARAPETIS et al., 2005; CUNNINGHAM, 2008; STEER et al., 2009b), embora nem todas as cepas com esses genótipos estejam associadas a um quadro infeccioso invasivo (BARBOZA et al., 2015; BERES et al., 2006, 2016).

Os estudos conduzidos na última década (BEN ZAKOUR et al., 2012; BESSEN et al., 2015; D'HUMIÈRES et al., 2015; EFSTRATIOU; LAMAGNI, 2017; MEISAL et al., 2010; SMEESTERS et al., 2006; STEER et al., 2009b) evidenciaram uma grande diversidade genotípica na distribuição global, principalmente entre países emergentes e desenvolvidos, sendo que nos primeiros observa-se um perfil epidemiológico distinto e mais diversificado que o perfil observado em países desenvolvidos, como representado na tabela 1, que agrupa a frequência de isolados *S. pyogenes* reportada na literatura entre 1999 e 2009 de acordo com a distribuição de genótipos observados em três regiões: (A) regiões desenvolvidas (Europa, América do Norte, Austrália, Nova Zelândia e Japão), (B) África e (C) regiões do Pacífico (América central, América do Sul Caribe e áreas indígenas australianas).

Tabela 1 - Frequência genotípica de *S. pyogenes* por região. Os genótipos são representados do mais frequente (à esquerda) ao menos frequente. Os genótipos mais associados a infecções invasivas se encontram em destaque

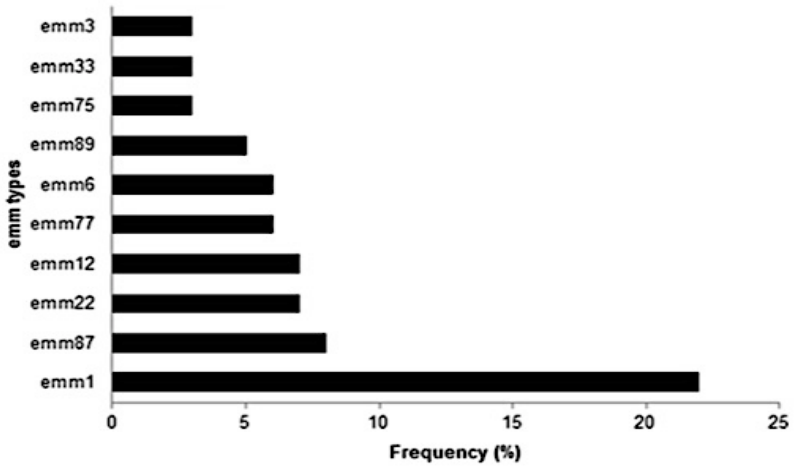
Região	Genótipos mais frequentes																							
A	1	12	28	3	4	89	6	75	77	2	11	22	81	87	5	49	53	58	44	82	73	83	18	92
B	12	75	74	1	3	63	25	18	28	9	5	71	22	77	99	8	81	76	103	44	114	43		
C	55	11	70	33	25	44	53	80	103	76	89	100	1	71	22	93	52	91	73	69	56	58	78	14

Adaptado de: (STEER et al., 2009b).

Já no Brasil, poucos são os estudos epidemiológicos realizados, apesar da alta incidência de infecções por *S. pyogenes* (BARROS et al., 2015; SMEESTERS et al., 2006; TARTOF et al., 2010; TEIXEIRA et al., 2001). O único estudo realizado em São Paulo (BARROS et al., 2015) mostra algumas semelhanças com a distribuição da Europa e Ásia,

tendo o *emm*1 como o mais frequente e um grande número de isolados *emm*12 e *emm*87 (Gráfico 1).

Gráfico 1 - Epidemiologia molecular de *S. pyogenes* em São Paulo entre 2008-2011



Fonte: (BARROS et al., 2015)

Apesar de sua importância, nota-se a carência de estudos epidemiológicos realizados principalmente nos países subdesenvolvidos, onde a incidência de casos é maior, dificultando o desenvolvimento de uma vacina de importância global baseada apenas nesta proteína (STEER et al., 2009b). Considerando-se ainda a grande variação na distribuição de diferentes genótipos observada nas diferentes regiões analisadas, além do risco de desenvolvimento de sequelas pós-infecciosas causadas pelo mimetismo da proteína M com a tropomiosina, faz-se cada vez mais importantes os estudos baseados na identificação de outros possíveis genes-alvo.

2.2.2 Fatores de virulência de *S. pyogenes*

A diversidade clínica observada em *S. pyogenes é explicada*, além de outros fatores, pela aquisição e regulação dos fatores de virulência. Assim como a proteína M, este estreptococo é capaz de sintetizar outros fatores ao longo do processo infeccioso. A invasão de tecidos, por exemplo, é mediada por adesinas como SclA e SclB, e facilitada pela produção de invasinas e enzimas extracelulares como hialuronidase, estreptoquinase, estreptolisinas O e S, a cisteíno protease (SPeB), DNase, NADase, etc. Essas enzimas normalmente são tóxicas,

podendo provocar necroses tissulares, como as observadas em dermatites hipodérmicas necrosantes (OLSEN; MUSSER, 2010).

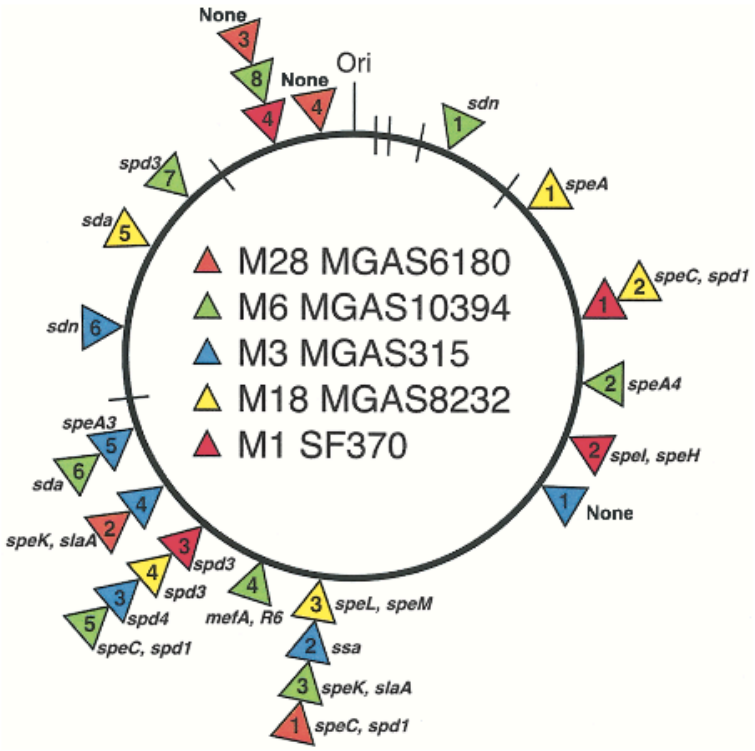
Já as exotoxinas com característica antigênica (SpeA, SpeC, SpeD, Ssa, SmeZ, etc) são, em sua maioria, codificadas por genes de origem fágica. O perfil superantigênico, no entanto, se mostra bem diversificado e normalmente relacionado ao genótipo *emm* (Tabela 2). É possível perceber nesse estreptococo uma grande diversidade de fagos encontrados em diferentes cepas, bem como o grande número de sítios de inserção e sua localização no genoma (Figura 2).

Tabela 2 – Perfil superantigênico dos genótipos mais frequentes de *S. pyogenes* em São Paulo

emm type	Number of samples	Superantigens number (%)							
		spG	smeZ	speC	ssa	speH	speI	speJ	speA
emm 1	48	45	48	6	10	1	3	39	30
emm 53	4	4	3	3	1	1	1	1	-
emm 33	7	7	6	4	5	2	1	6	-
emm 22	16	14	16	13	11	2	1	-	-
emm 12	15	13	15	14	1	8	8	-	-
emm 78	5	4	5	3	1	2	-	1	-
emm 6	12	12	12	83	2	-	1	1	1
emm 87	18	13	18	15	13	-	13	13	-
emm 77	13	6	11	5	2	-	1	1	-
emm 89	11	11	11	8	2	-	4	4	-
emm 3	6	6	5	1	3	-	-	-	6
emm 183	5	5	5	1	-	4	-	-	-
emm 75	6	6	6	2	3	-	-	-	-

Fonte: (BARROS et al., 2015)

Figura 2 – Representação da diversidade fágica encontrada em genomas de *S. pyogenes* de diferentes genótipos. Na figura, cada cor representa um genoma diferente, seguindo a legenda no centro da imagem (genótipo *emm* seguido do nome da cepa). Os triângulos representam os profagos encontrados e sua posição de inserção no genoma, sendo numerados em ordem acompanhando o sentido horário. As toxinas superantigênicas portadas pelos fagos são descritas ao lado dos triângulos.



Fonte: (GREEN et al., 2005)

Além de profagos também foram identificados elementos conjugativos e integrativos (ICEs) portando genes de resistência a antibióticos (SORIANO et al., 2014), adesinas e exotoxinas (GREEN et al., 2005) que podem ser transferidos lateralmente entre isolados de diferentes genótipos *emm*, assim como a outros estreptococos (SITKIEWICZ et al., 2011). Em algumas cepas, a soma das bases de todo o DNA exógeno pode somar até 10% do genoma completo (BANKS et al., 2004).

A caracterização do nível de invasividade das cepas pode depender, além de outros fatores, da combinação de fatores de virulência (BIDET; BONACORSI, 2014). Com isso em mente, a utilização da técnica de sequenciamento de genoma completo para estudos de comparações genômicas aumenta (BANKS et al., 2004; BEN ZAKOUR et al., 2012; BESSEN et al., 2015; GREEN et al., 2005; LONGO et al., 2015; NASSER et al., 2014), buscando-se entender a origem da invasividade e patogenicidade de certas cepas, bem como a

procura de genes alternativos para a fabricação de vacinas. A relação de genomas completos de *S. pyogenes* na plataforma NCBI já atingiu mais de 50 isolados (fev.2017).

2.3 Importância da centralização dos dados genômicos de S. pyogenes

Pelos artigos analisados (ARNAUD et al., 2012; CHERRY et al., 2012; CUI et al., 2006; HENRY et al., 2014; INGLIS et al., 2012; SKRZYPEK et al., 2010; WINSOR et al., 2016), percebe-se que as principais informações relacionadas em uma comparação genômica são: genótipo *emm* (para *S. pyogenes*), cepa, invasividade, patologia, ano de isolamento, região de isolamento, informações clínicas do hospedeiro, perfil superantigênico das cepas (para *S. pyogenes*) e filogenia. Na ausência de um estudo global organizado, essas informações são relacionadas em extensas revisões bibliográficas, que acabam por desprezar parte dos resultados produzidos devido à falta de padronização na obtenção e inferência estatística dos dados (STEER et al., 2009b), principalmente em relação às condições do hospedeiro e contexto geo-temporal. A centralização dos dados também pode garantir a atualização da relação dos fatores de virulência identificados em *S. pyogenes,* que são abordados em estudos comparativos utilizando-se revisões (CHEN et al., 2005; NAKAGAWA et al., 2003), além de facilitar a recuperação das sequências desses fatores relacionando as anotações dos isolados disponíveis. Essa atualização deve se estender às anotações dos genomas depositados, onde se percebe uma variância no nome e identificação dos códons iniciadores em sequências antigas e recentes (WINSOR et al., 2016).

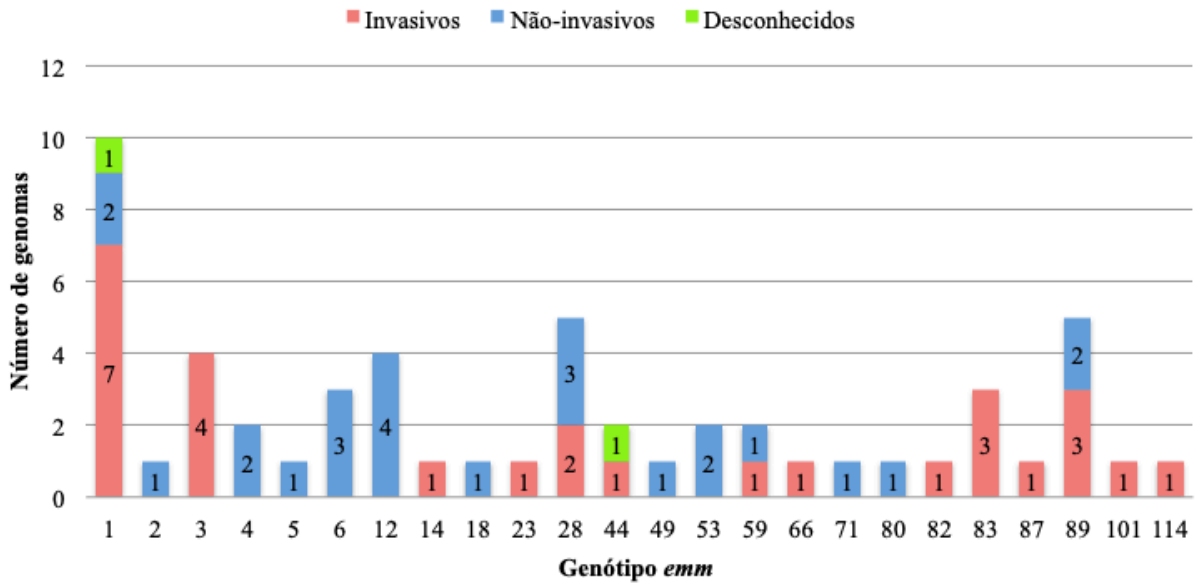
Devido a sua importância, existem diversas plataformas similares que são específicas para um organismo e que disponibilizam ferramentas de padronização e comparação de sequências genômicas. Relações atualizadas de fatores de virulência, elementos extragenômicos (ICEs, Virus, transposons, etc.) estão presentes na maioria das plataformas. Entre as principais ferramentas encontradas estão algoritmos de identificação de fatores de virulência, predição e anotação automática de genes, identificação de elementos extragenômicos, ferramentas para alinhamento e tradução de sequências e de representação gráfica dos dados. Às bases mais antigas foram incorporadas ferramentas mais sofisticadas, como visualização e comparação em larga escala, modelização tridimensional de proteínas, representação ou simulação de vias metabólicas, representação do *core* genoma (conjunto de genes que são encontrados em todos os isolados), ferramentas para *design* de *primers* e análises filogenéticas.

3. Materiais e métodos

3.1 Caracterização e anotação dos genomas de S. pyogenes para o estudo

Os genomas incluídos neste estudo foram obtidos diretamente da plataforma do *National Center for Biotechnology Information* (NCBI), usando a seguinte especificação de busca: "(*streptococcus pyogenes) AND "Streptococcus pyogenes"[porgn:__txid1314] complete genome"*. Foram aceitos apenas genomas completos e montados, excluindo-se genomas que não tivessem a informação de seu genótipo *emm* contida na descrição. No total, foram identificados 55 genomas (setembro, 2016). Destes, 28 genomas foram sequenciados à partir de isolados associados a infeções invasivas, 25 de isolados associados a infecções não-invasivas e 2 cujo perfil invasivo não foi descrito ou não pôde ser inferido pelas informações fornecidas. No total, 24 genótipos foram representados, como mostra o gráfico 2. Uma relação contendo mais informações sobre estes genomas está disponível no apêndice A.

Gráfico 2 – Distribuição dos genótipos *emm* das cepas inseridas no estudo



Fonte: Suzane de Andrade Barboza, 2019

A classificação de uma infecção como sendo invasiva ou não invasiva não depende exclusivamente da gravidade da infecção, mas do sítio onde é encontrado: de maneira geral, uma infecção localizada em uma região estéril (como sangue, meninge, coração, cérebro, etc.) é classificada como invasiva, enquanto infecções situadas em locais não-estéreis (pele,

estômago, intestino, etc.) são ditas não-invasivas (LANCEFIELD, 1933). Entretanto, sequelas pós-infecciosas como glomerulonefrite (que atinge o rim) e febre reumática (que atinge coração, cérebro e articulações) não seguem esta classificação, visto que são doenças autoimunes desenvolvidas no período de uma a duas semanas após uma infecção da faringe, não-invasiva. A relação de infecções invasivas e não invasivas causadas pelas cepas que compõe o conjunto de dados é descrita na tabela 3.

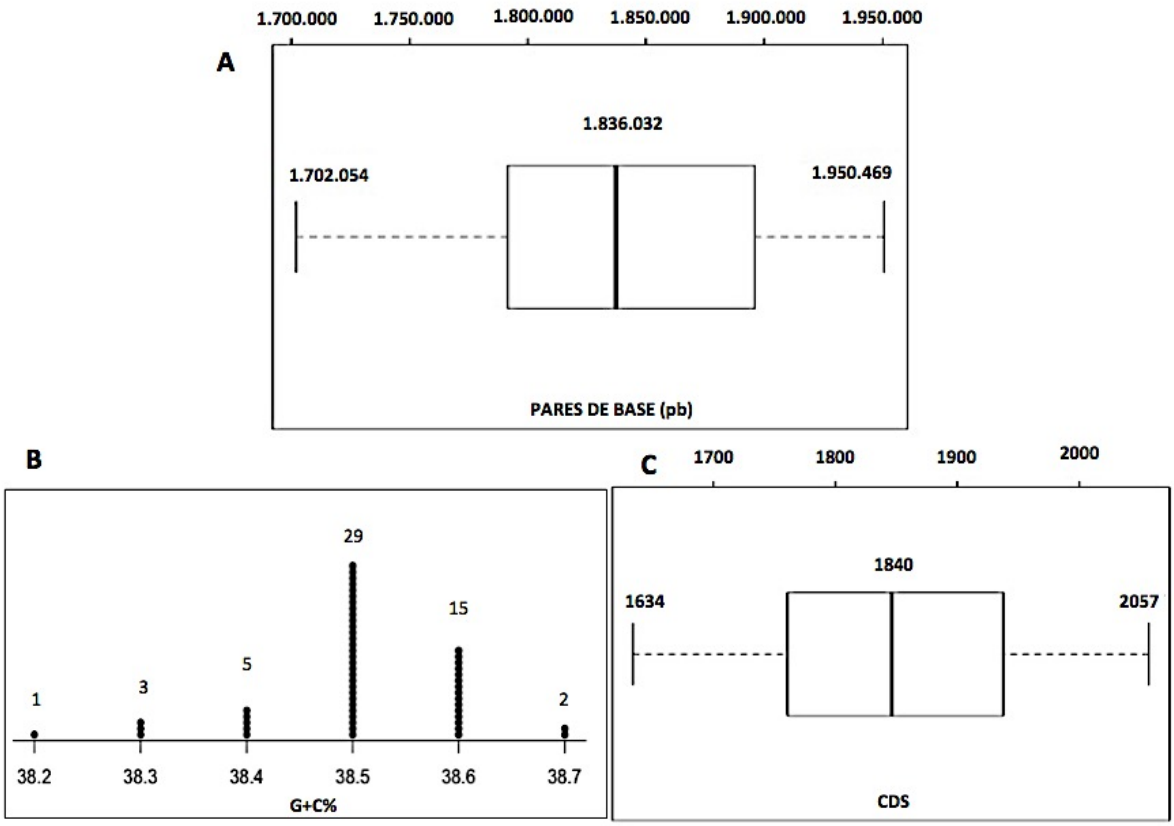
Tabela 3 – Infecções causadas pelas cepas selecionadas para este projeto. Isolados não associados a alguma doença foram reunidos como "Desconhecido". Na segunda coluna, I representa infeções invasivas, N indica não-invasivas

Doença	Invasividade	Quantidade de isolados	Genótipos associados
Fasciite necrosante	I	7	1, 23, 83, 87, 89
Síndrome de choque tóxico estreptocócico	I	4	1, 3
Infeção do fluido cerebrospinal	I	1	1
Meningite	I	2	1
Febre escarlate	N	4	1, 12
Bacteriemia	I	4	82, 83, 101, 114
Glomerulonefrite pós-estreptocócica aguda	N	3	12, 49, 59
Faringite	N	9	3, 4, 6, 12, 28, 80, 89
Febre reumática aguda	N	4	5, 6, 18
Dermatite superficial	N	2	2, 71
Sepse puerperal	I	1	28
Dermite estreptocócica perianal	N	2	28
Endometrite	I	2	28, 44
Impetigo	N	2	53
Infecção de tecido mole	I	1	59
Abcesso subcutâneo	I	1	66
Gastrite fleimonosa	I	1	89
Desconhecido	-	5	1, 14, 44, 89

Fonte: Suzane de Andrade Barboza, 2019

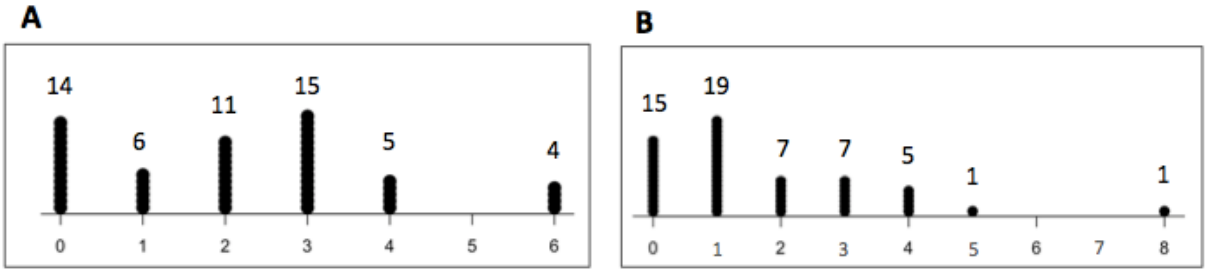
A fim de padronizar a anotação de todos os genomas, estes foram reanotados com a ferramenta RASTtk (WATTAM et al., 2014), disponível na plataforma PATRIC (WATTAM et al., 2017). Após a condução da anotação, verificou-se que os genomas possuem uma grande variação em relação ao tamanho e quantidade de CDSs, embora a porcentagem G+C tenha se mantido praticamente constante (Figura 3). Essas variações observadas são causadas, principalmente, pela capacidade de aquisição de material extragenômico, composto principalmente pela inserção de profagos (Figura 4).

Figura 3 – Caracterização dos genomas de *S. pyogenes*. O *boxplot* A representa a distribuição do tamanho dos genomas em pares de base. O *dotplot* B indica a variação da porcentagem G+C nos genomas, que são representados por pontos. O *boxplot* C mostra a diferença no número de CDS dos isolados



Fonte: Suzane de Andrade Barboza, 2019

Figura 4 – Quantidade de profagos inteiros (A) e parciais presentes nos genomas (B), sendo cada genoma um ponto no gráfico



Fonte: Suzane de Andrade Barboza, 2019

3.2 Anotação e genômica comparativa

Para a análise, foram usadas as ferramentas desenvolvidas pelo grupo de pesquisa, utilizadas previamente na comparação de 15 genomas de *Xanthomonas* (DIGIAMPIETRI et al., 2019). Estas ferramentas se encontram integradas em um arcabouço que permite a análise em três etapas principais: [1] identificação de genes homólogos, [2] comparação de genomas completos e [3] análise de redes gênicas. Uma lista completa com as ferramentas e parâmetros utilizados é dada no apêndice B.

3.2.1 Identificação de genes homólogos

A identificação de genes homólogos foi conduzida utilizando uma abordagem de similaridade baseada em modelagem de grafos (SANTIAGO; PEREIRA; DIGIAMPIETRI, 2018). Para a primeira etapa, o alinhamento local de todas as CDSs contra todas as CDSs é calculado. Dois limiares são utilizados para verificar se um alinhamento atingiu ou não uma qualidade mínima para ser utilizado nas próximas etapas do algoritmo: o valor máximo de *e-value* e a percentagem mínima de alinhamento. Neste caso, a percentagem de alinhamento foi fixada em 45%, após serem feitos testes que buscassem o melhor valor a produzir a menor quantidade possível de grupos parálogos, não sendo, ao mesmo tempo, muito restritivo a ponto de separar famílias de componentes que possuam a mesma função anotada. O *e-value* foi determinado automaticamente, buscando maximizar o coeficiente de clusterização da rede de genes homólogos (rede na qual cada gene corresponde a um elemento ou nó e dois genes são ligados caso o alinhamento entre suas sequências satisfaça as restrições apresentadas). O coeficiente de clusterização mede a probabilidade de dois elementos ligados a um elemento central (no caso os genes ligados por uma relação de homologia) também se liguem entre si,

sendo assim a maximização do coeficiente de clusterização produz relações de homologia mais homogêneas.

Após a identificação das relações de homologia entre os genes, os clusters de genes foram novamente divididos utilizando análises filogenéticas, as árvores com braços mais longos que um limiar (*threshold*) tiveram seus genes novamente divididos em grupos menores (DING; BAUMDICKER; NEHER, 2018). Por último, uma etapa adicional foi realizada a fim de identificar proteínas multi-domínio, levando em conta a assimetria do alinhamento disposto no grafo de cada família identificada.

3.2.2 Comparação de genomas completos

Afim de construir os cladogramas e a rede gênica (de forma a sumarizar as relações entre todos os genes e todos os genomas), a estratégia utilizada foi a comparação de todos os genomas contra todos os genomas. A primeira abordagem para a construção do cladograma utilizou uma matriz baseada na presença (valor 1) ou ausência (valor 0) dos genes em cada uma das famílias de genes homólogos definidas na etapa anterior. Uma matriz quadrada foi então criada a fim de calcular a distância entre os genomas, utilizando a distância Euclidiana entre os genomas da matriz de presença/ausência. Essas distâncias foram posteriormente utilizadas para a criação de um cladograma, utilizando o método *neighbor-join* presente na biblioteca Phylip (FELSENSTEIN, 2005). A segunda abordagem produziu uma *supertree* (CREEVEY; MCINERNEY, 2005), baseada nas filogenias de cada família de genes, resultando em uma árvore que reflete simplificadamente todas as relações identificadas, o método de inferência escolhido foi o *Quartet Fit* com *Neighbor Interchange* disponível no programa Clann (CREEVEY; MCINERNEY, 2005).

Em relação aos fatores de virulência, foi utilizado o método de inteligência artificial *Random Tree* para auxiliar na identificação dos principais genes potencialmente relacionados a características clinicamente relevantes (como invasividade ou doença). Este método é baseado na construção de uma árvore de decisão binária, sendo os atributos de cada nó escolhidos aleatoriamente. A árvore resultante (após passar pelo processo de poda automático) pode ser facilmente compreendida, sendo então muito utilizada para explicar ou caracterizar fenômenos reais (ROKACH; MAIMON, 2007).

3.2.3 Análise de redes gênicas

Com o objetivo de melhor compreender as relações intergênicas, foi criada uma rede gráfica em que cada família de genes homólogos é um *cluster* (ou componente conexo) composto de nós que representam os genes de cada genoma. As arestas representam um alinhamento entre os genes conectados que satisfaz os parâmetros definidos. Posteriormente, um algoritmo foi aplicado a fim e aproximar ou separar graficamente os genes de acordo com suas arestas.

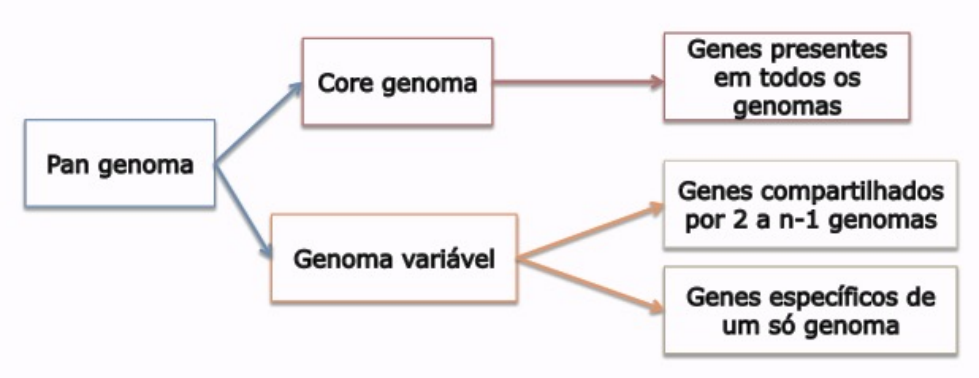
3.2.4 Apresentação dos resultados

Todas as informações apresentadas neste estudo foram disponibilizadas através da criação da *Streptococcus pyogenes Database* (SPD), desenvolvida em outros projetos que colaboraram com o presente estudo com o objetivo de uniformizar e centralizar os dados genômicos de *S. pyogenes*, acessível pelo *link* http://143.107.58.250/reportStrep2/. Os resultados obtidos são dispostos em vários níveis dentro do portal: o primeiro nível compreende a relação de presença/ausência genomas nas famílias, utilizada para facilitar a identificação de genes exclusivos a isolados que compartilhem um certo conjunto de características; o segundo nível dispõe os alinhamentos das sequências que compõe as famílias de genes. Para esses alinhamentos foi incluído um valor de dissimilaridade, que mede a diversidade das bases em um alinhamento a fim de determinar pares de bases mais correlatos a determinados grupos de organismos, como SNPs, por exemplo. Por fim, o último nível compreende as representações filogenéticas, que utilizam a métrica MIST (*Most Isolated Subtree*) a fim de determinar o quão isolados são os grupos em sub-árvores exclusivas.

4. Resultados e discussão

Após a realização de todos os alinhamentos, 4.466 grupos de genes foram identificados, configurando assim o *pan* genoma, que corresponde ao conjunto total de genes de todas as cepas de um grupo, composto pelo *core* genoma (conjunto de genes presentes em todos os isolados) e pelo o genoma variável, que é o conjunto de genes presentes em apenas um subconjunto das cepas (TETTELIN et al., 2005), como representado pela figura 5.

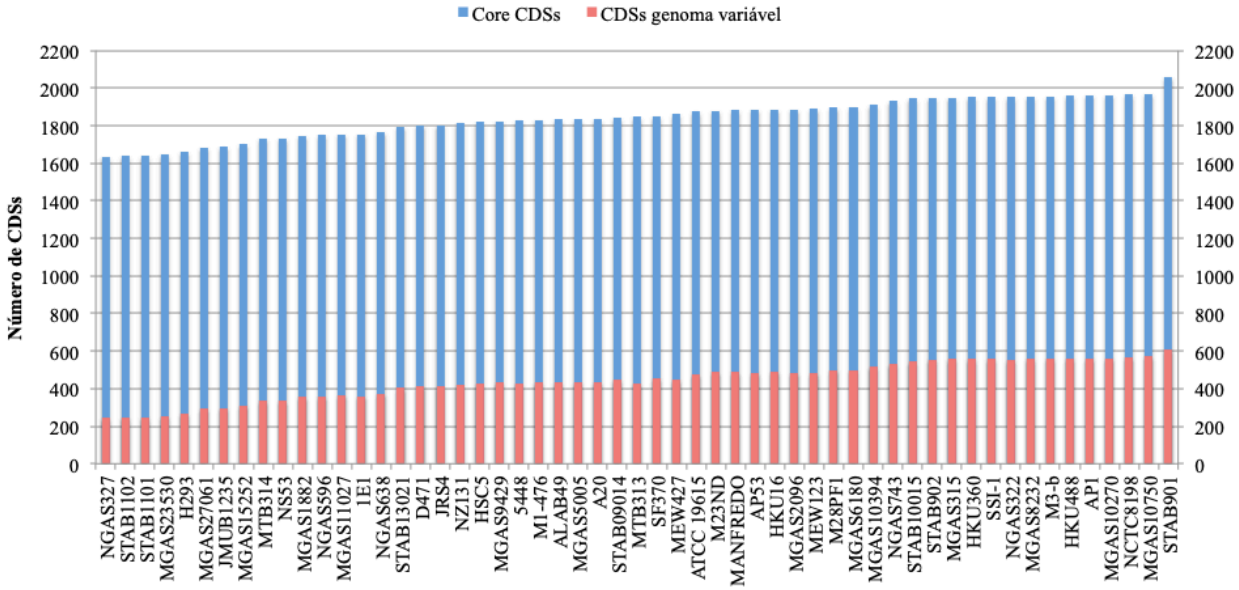
Figura 5 – Representação da composição do *pan* genoma



Fonte: Suzane de Andrade Barboza, 2019

Do total de grupos de genes, 2.636 famílias (59%) agregam dois ou mais genes, agrupando no total 99.390 CDSs. Entretanto, apenas 1.271 grupos são compostos por pelo menos um gene de cada um dos 55 genomas selecionados, indicando que, considerando os parâmetros definidos, o *core* genoma é composto por aproximadamente 28,4% do total de famílias. Em relação às cepas, a quantidade de genes que compõe o genoma variável é de aproximadamente 24% do total de CDSs, como representado no gráfico 3.

Gráfico 3 – Total de CDSs por cepa, composto por *core* CDSs (azul) e CDSs do genoma variável (vermelho)

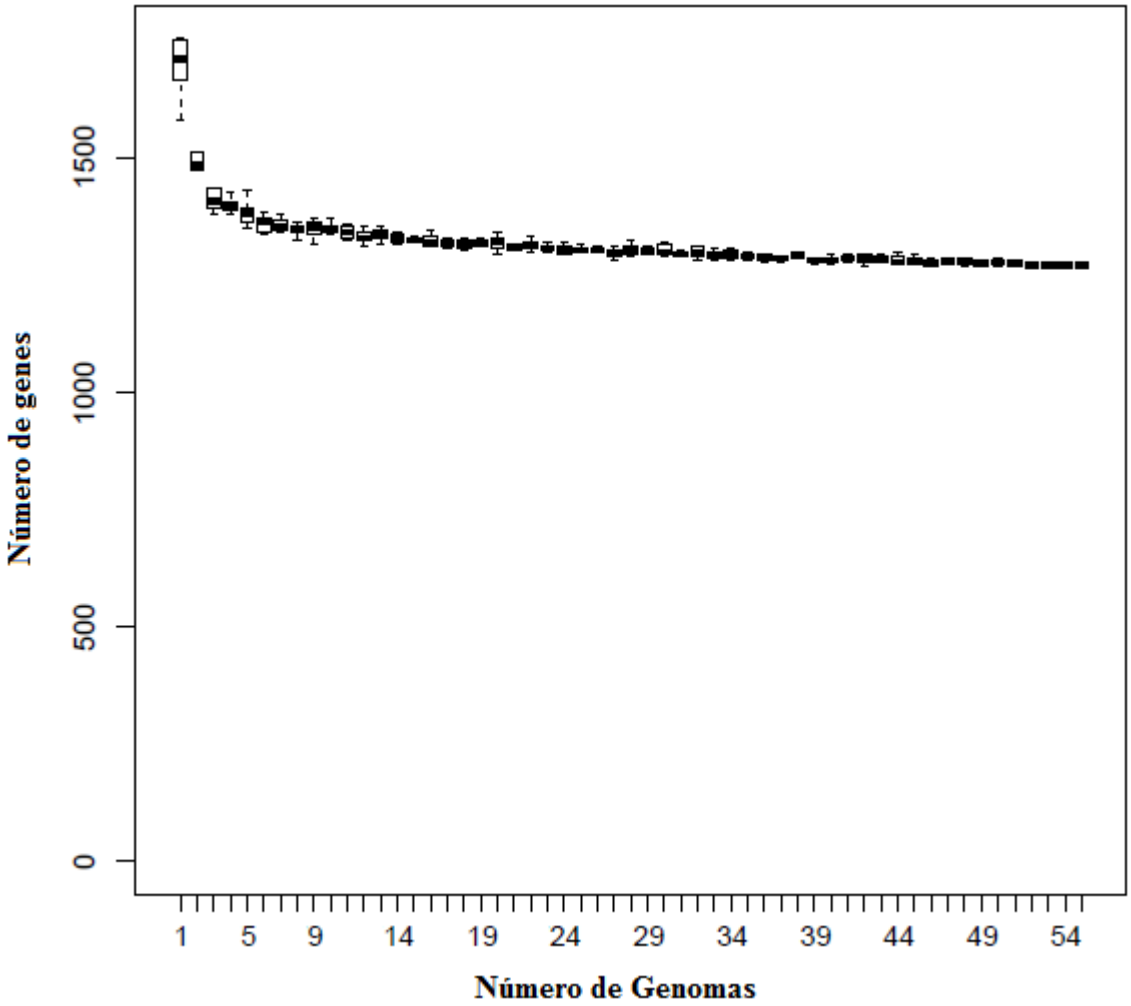


Fonte: Suzane de Andrade Barboza, 2019

Um estudo realizado em 2007 com 11 genomas estabeleceu um *core* genoma composto por 1.297 genes (LEFÉBURE; STANHOPE, 2007). Mais recentemente, em 2016, a análise de 19 genomas resultou na elevação deste número para 1.342 genes (MARUYAMA; WATANABE; NAKAGAWA, 2016). Além do diferente número de genomas estudados, a diferença observada entre o *core* genoma definido neste estudo (1.271 genes) e o de estudos anteriores pode ser explicada pela implantação de uma etapa adicional orientada à padronização das anotações por meio da realização de alinhamentos do tipo todos os genes *vs* todos os genes, o que pode ter prevenido a criação de diferentes famílias formadas à partir de um mesmo gene.

Observa-se pela figura 6 que o *core* genoma encontra-se próximo a um equilíbrio, indicando que os genomas analisados de *Streptococcus pyogenes* conseguem representar de maneira bastante satisfatória o *core* genoma para esta espécie. Isto é, mesmo com a adição de novos genomas da espécie não é esperada mudança significativa no tamanho do *core* genoma.

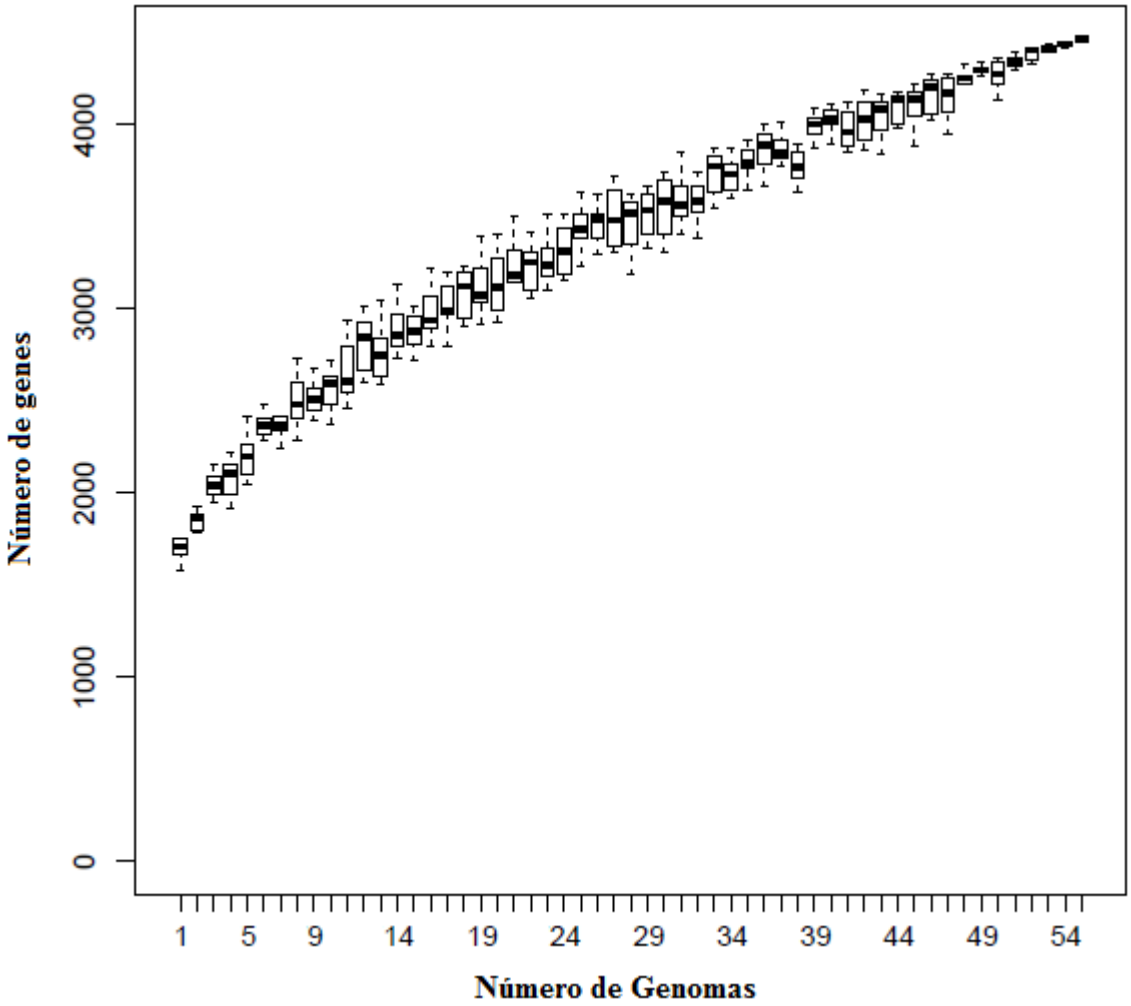
Figura 6 – Número de genes do *core* genoma em relação ao número de genomas



Fonte: Suzane de Andrade Barboza, 2019

A figura 7 apresenta a variação do *pan* genoma em relação à quantidade de genomas. Diferentemente do que foi observado para o *core* genoma, não é possível dizer que a curva de variação do *pan* genoma está próxima de se estabilizar. Uma das possíveis razões para isto é a plasticidade genômica deste estreptococo, ou seja, sua grande capacidade de reorganização genômica (BESSEN et al., 2015; NAKAGAWA et al., 2003).

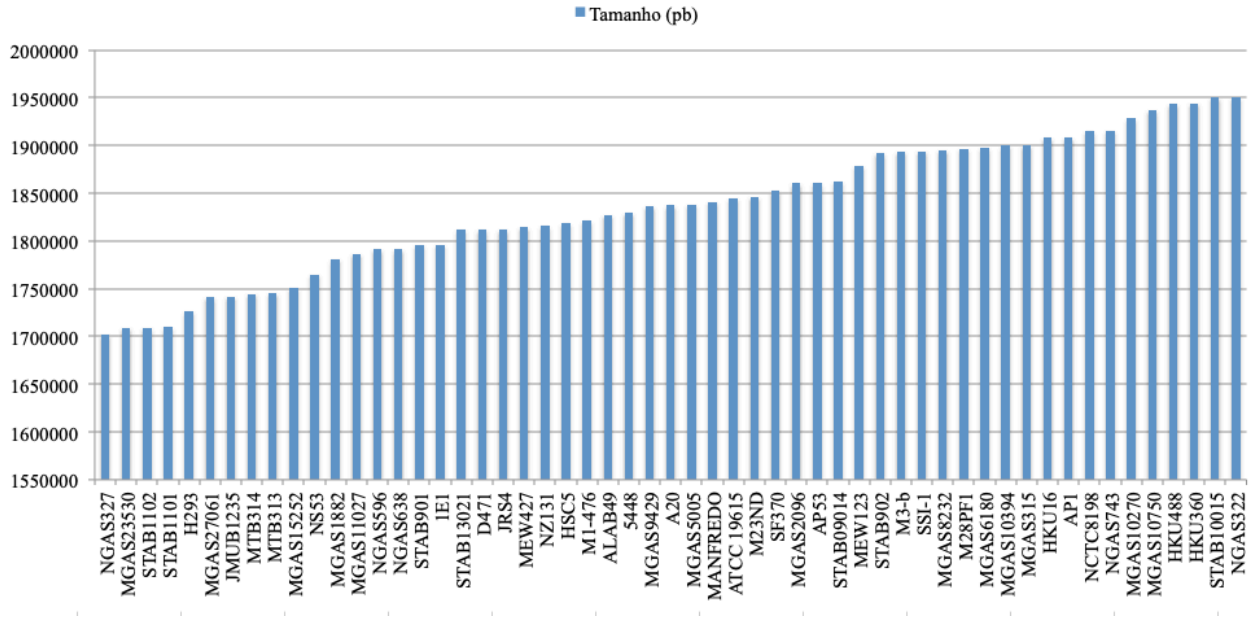
Figura 7 – Número de genes do *pan* genoma em relação ao número de genomas



Fonte: Suzane de Andrade Barboza, 2019

Interessante notar que as cepas com o maior número de CDSs não necessariamente possuem os maiores genomas (Gráfico 4). Embora a ordem das cepas nos gráficos 3 e 4 não seja muito diferente, a cepa com o maior número de CDSs (STAB901) aparece apenas na 16ª posição no gráfico 4, o que indica que esta cepa abriga uma proporção incomum entre regiões codificantes e não codificantes em seu genoma. Essa peculiaridade já havia sido previamente descrita (CHAUDHARI; GUPTA; DUTTA, 2016). Entretanto, esse estudo também relata a observação deste evento na cepa M1-476, fato que não é percebido no presente estudo. Essa diferença pode estar relacionada com o diferente número de genomas analisados.

Gráfico 4 – Distribuição do tamanho do genoma das cepas de *S. pyogenes*



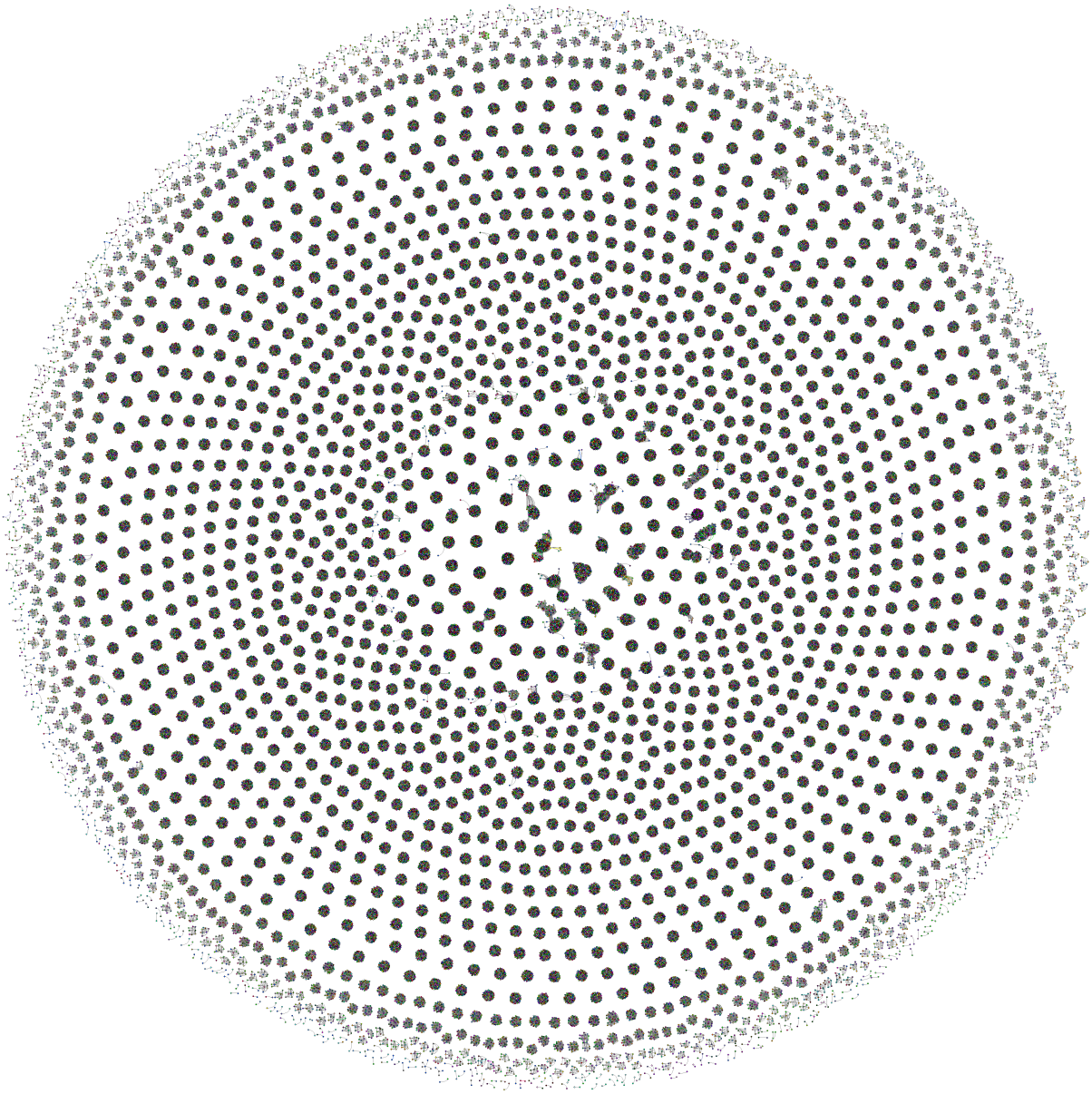
Fonte: Suzane de Andrade Barboza, 2019

É possível notar uma diferença de 105 CDSs entre a cepa STAB901 e o segundo isolado com o maior número de CDSs (MGAS10750), e de 302 CDSs entre STAB901 e 1E1, as únicas cepas relacionadas ao genótipo *emm*44, indicando que STAB901 possui um perfil diferente dos outros isolados. Tendo em mente que o número de CDSs e o tamanho do genoma não são diretamente proporcionais, esse evento observado pode ser um resultado de um processo de redução do genoma, como já foi proposto anteriormente (CHAUDHARI; GUPTA; DUTTA, 2016), visto que processos como duplicação ou aquisição genômica deveriam ter aumentado o tamanho do genoma. Entretanto, é interessante notar que de acordo com este estudo (CHAUDHARI; GUPTA; DUTTA, 2016) STAB901 apresenta um número maior de genes exclusivamente ausentes quando comparado com outros isolados (n=257), enquanto neste projeto foram identificados apenas 11. Da mesma forma, o estudo previamente mencionado relata como 11 o número de genes exclusivamente encontrados no genoma do STAB901, contra 215 genes encontrados neste estudo. Essas diferenças podem ser estar relacionadas com a anotação dos genomas, sendo que, com uma anotação mais recente conduzida com a ferramenta PATRIC (WATTAM et al., 2014) foram anotados 2.057 genes para o isolado STAB901, enquanto apenas 1.358 foram relatados com a ferramenta PGAP, utilizada na base de dados NCBI (TATUSOVA et al., 2016). A possibilidade de ter havido a anotação de pseudogenes com a anotação PATRIC não pode ser descartada, porém vale

lembrar que essa desproporção entre o tamanho do genoma e o número de CDSs não é encontrada em outros isolados anotados com a ferramenta.

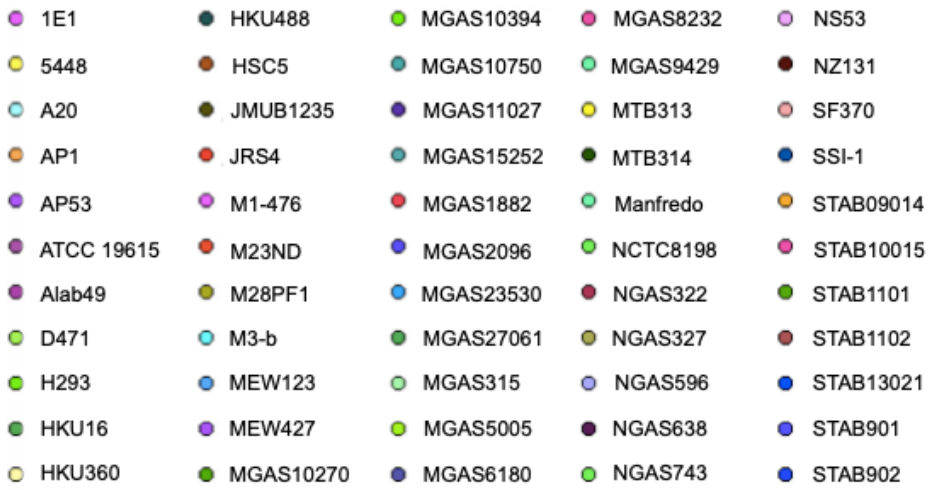
O resultado gráfico de todos os alinhamentos é uma rede de genes onde cada *cluster* (ou componente conexo) é composto por um conjunto de genes homólogos, que são representados por pontos (Figura 8). Uma cor foi designada a cada genoma (Figura 9), e as relações de homologia são representadas por traços (ou arestas). Quando a relação de homologia se dá entre dois genes do mesmo genoma, o traço é desenhado com a cor do genoma.

Figura 8 – Rede de genes homólogos criada com 55 genomas de *S. pyogenes*



Fonte: Suzane de Andrade Barboza, 2019

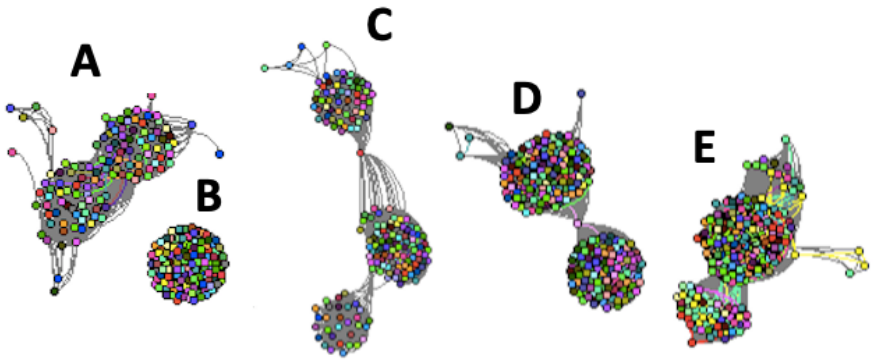
Figura 9 – Legenda de cores atribuídas aos genomas de *S. pyogenes*



Fonte: Suzane de Andrade Barboza, 2019

É possível notar na figura 8 que a grande maioria das famílias foram agrupadas em componentes com um grande coeficiente de clusterização, resultado da boa combinação dos parâmetros definidos. No centro da imagem percebe-se a presença de componentes maiores, tendo alguns um coeficiente de clusterização mais baixo. A figura 10 dispõe um recorte de alguns desses *clusters* para melhor visualização:

Figura 10 – Recorte de alguns *clusters* da rede de genes homólogos



Fonte: Suzane de Andrade Barboza, 2019

No componente A da figura é possível perceber dois agrupamentos distintos, conectados entre si. Trata-se das proteínas SclA e SclB, muito semelhantes entre si. O *cluster* B (*peptide chain release factor* 3) foi incluído na figura como um modelo de agrupamento

com alto coeficiente de clusterização. O componente C agrega as proteínas M e M-*like*, enquanto os *clusters* D e E são compostos por genes anotados como hyaluronidase e *mobile element protein*, respectivamente.

É fácil notar que os componentes C e D possuem *clusters* conectados entre si por um ou poucos genes. Essa configuração pode indicar a ligação entre estes genes e um ancestral comum, ou que uma combinação diferente dos parâmetros escolhidos poderia ter separado esses agrupamentos não-ortólogos, ou ainda representar um processo de rearranjo que levou ao surgimento de proteínas multi-domínio (VOGEL et al., 2004). É interessante notar que, no componente D, os genes que promovem a ligação entre os *clusters* pertencem ao mesmo genoma (NS53), sugerindo a ocorrência de um processo de duplicação. A representação visual desses alinhamentos se faz bastante útil na análise de um grande volume de alinhamentos decorrentes de comparações genômicas.

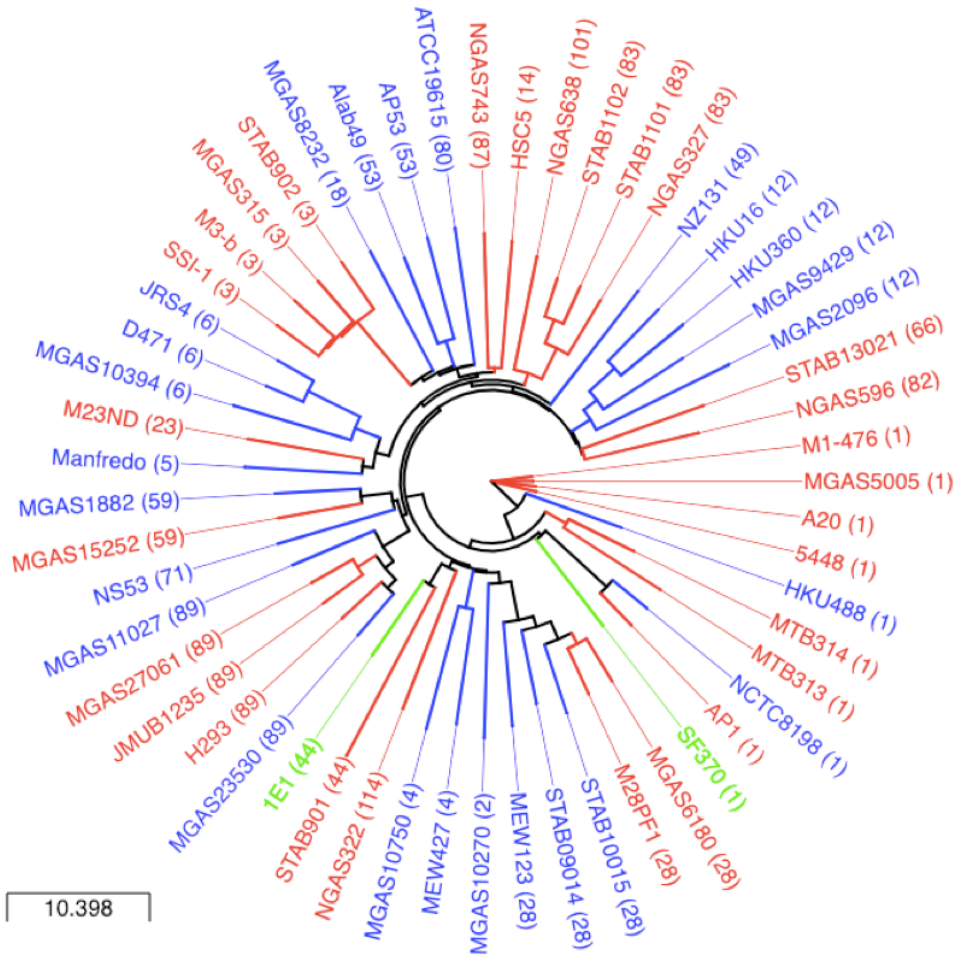
Partindo da análise da figura 8 é possível perceber que não há um gene específico encontrado exclusivamente em cepas que compartilhem as mesmas características, como perfil invasivo ou doença. Essas características podem então ser relacionadas à expressão dos fatores de virulência (BESSEN et al., 2015; FLORES et al., 2015; IKEBE et al., 2010; RIBARDO; LAMBERT; MCIVER, 2004; VEGA; MALKE; MCIVER, 2016), não se excluindo uma possível relação com o estado de saúde do hospedeiro no momento de infecção ou o modo de contágio. Pode-se afirmar, contudo, que mesmo que a presença de genes não seja diretamente (ou exclusivamente) determinante na origem da invasividade de cepas de *S. pyogenes*, sua aquisição tem um papel importante na sobrevivência e colonização, auxiliando na invasão de diferentes nichos (FERNANDES et al., 2017).

4.1 Análises filogenéticas

Sabe-se que S. *pyogenes* possui uma grande capacidade de integrar DNA extragenômico, incorporando desta forma diversos fatores de virulência (BESSEN et al., 2015; NAKAGAWA et al., 2003), especialmente fragmentos originados em profagos (FERRETTI et al., 2001; GREEN et al., 2005), que desempenham papéis importantes na recombinação à partir de transferência horizontal de genes (REID et al., 2001). Com isso em mente, a maior parte dos estudos optou pelo uso de árvores construídas apenas com genes do *core* genoma concatenados (BERES et al., 2006; BESSEN et al., 2015; MARUYAMA; WATANABE; NAKAGAWA, 2016), já que os sinais filogenéticos são extremamente afetados pela perda/aquisição de genes (WOLF et al., 2001). Entretanto, a sensibilidade de

árvores baseadas na presença/ausência de genes para a recombinação à nível genômico pode representar mais fielmente a evolução de um patógeno com grande plasticidade genômica, como é o caso de *S. pyogenes,* já que barreiras temporais e geográficas terminam por isolar algumas cepas ou genótipos e dificultar a transferência horizontal (BESSEN et al., 2015). Espera-se dessa forma que fragmentos inseridos mais recentemente nos genomas estejam conservados na maioria dos isolados (MARUYAMA; WATANABE; NAKAGAWA, 2016). À partir dessas considerações foi criado um cladograma baseado na presença e ausência de genes em cada um dos isolados (Figura 11).

Figura 11 – Cladograma construído baseado na presença/ausência de todos os genes dos 55 isolados de *S. pyogenes*. As linhas espessas representam os braços da árvore, enquanto os traços mais finos foram utilizados para facilitar a visualização dos dados. Isolados invasivos são representados em vermelho, não-invasivos em azul e isolados com perfil desconhecido foram representados em verde. O genótipo *emm* dos isolados é apresentado entre parêntesis.



Fonte: Suzane de Andrade Barboza, 2019

Todos os isolados foram agrupados pelos genótipos *emm*, uma tendência que é percebida em outros estudos (BESSEN et al., 2015; MARUYAMA; WATANABE; NAKAGAWA, 2016). O cladograma da figura 11 segue o mesmo padrão da única árvore baseada em *pan* genoma que foi publicada recentemente (CHAUDHARI; GUPTA; DUTTA, 2016), ambas apresentando algumas divergências em relação às árvores baseadas em *core* genoma (BESSEN et al., 2015; MARUYAMA; WATANABE; NAKAGAWA, 2016), como a aproximação dos isolados *emm*28, *emm*2 e *emm*4, enquanto nas árvores de *core* genoma tendem a aproximar as cepas *emm*28 e *emm*12. É válido notar que, embora à primeira vista pareça que o isolado *emm*12 (MGAS1882) tenha sido agrupado com um isolado *emm*59, uma análise mais aprofundada revelou que MGAS1882 também possui o genótipo *emm*59 (FITTIPALDI et al., 2012). Da mesma forma, a cepa A20 deveria ter sido classificada como *emm*1 (ZHENG et al., 2013), implicando que a matriz binária construída à partir do *pan* genoma conseguiu separar de forma eficiente todos os isolados *emm*12 e *emm*1.

Como esperado, os isolados SF370 e MGAS5005 se encontram em ramos separados no cladograma, um evento que provavelmente ocorreu por volta de 1980 (NASSER et al., 2014; SUMBY et al., 2005). Entretanto, a incorporação de um elemento conjugativo e estudos de SNPs determinaram que a cepa MGAS5005 deveria ter sido plotada mais distante da raiz da árvore do que a SF370 (SUMBY et al., 2005). Além disso, percebe-se que as cepas *emm*1 estão posicionadas mais próximas à origem da árvore do que os isolados com o menor número de CDSs. Esses eventos indicam que os isolados *emm*1, especialmente MGAS5005, parecem ter incorporado um conjunto de genes mais heterogêneo e exclusivo, acabando por receber uma pontuação menor na matriz. Da mesma forma, NGAS327 (isolado com o menor número de CDSs) provavelmente incorporou um conjunto de genes que é comumente encontrado em outros genomas. Já a posição mais distante da raiz é ocupada pelo STAB901, isolado com o maior número de CDSs.

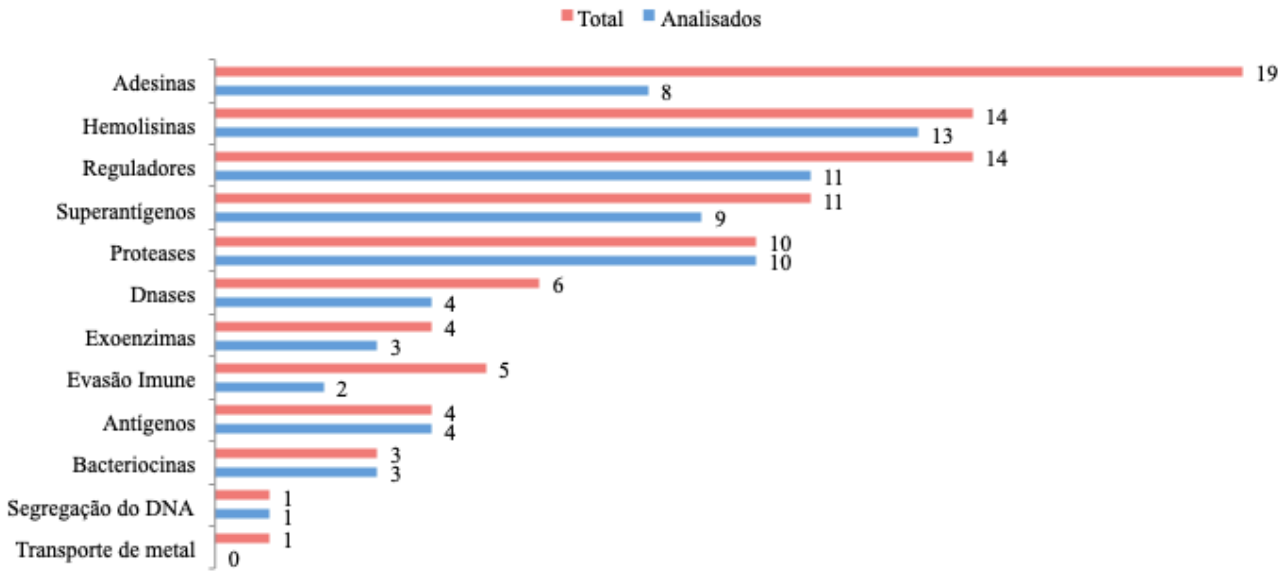
Observando o cladograma, não se percebe uma conexão clara entre a aquisição/perda de genes e o perfil invasivo das cepas, mesmo entre cepas de mesmo genótipo, o que impede a dedução do perfil invasivo dos isolados 1E1 e SF370. Em relação às cepas *emm*28, nota-se que isolados invasivos foram separados dos não invasivos. Esse evento ocorre devido à presença de três proteínas hipotéticas exclusivamente nos isolados não-invasivos (proteínas anotadas como fig|1314.408.peg.1042, fig|1314.373.peg.4 e fig|1314.381.peg.401 na base de dados disponibilizada de *S. pyogenes*). Essas proteínas poderiam ter uma relação na regulação de algum fator de virulência, levando à queda do poder invasivo dos isolados. Interessante

notar também que os isolados MTB313 e MTB314 (únicos isolados *emm*1 associados à meningite), foram agrupados no mesmo ramo da árvore.

4.2 Análise comparativa de genes

Devido ao grande número de genes que compõe o *pan* genoma de *S. pyogenes* (4.466), a análise comparativa de genes foi restringida apenas aos fatores de virulência já conhecidos, que são registrados na *Virulence Factors of Pathogenic Bacteria Database* (CHEN et al., 2005). A lista obtida nesse portal foi ainda completada por meio de uma revisão da literatura atual, totalizando 92 genes (C). Alguns fatores de virulência (n = 24), porém, não foram anotados ou podem ter sido anotados incorretamente, e por esse motivo não foram analisados (Figura 12).

Figura 12 – Distribuição dos fatore s de virulência de *S. pyogenes* por função



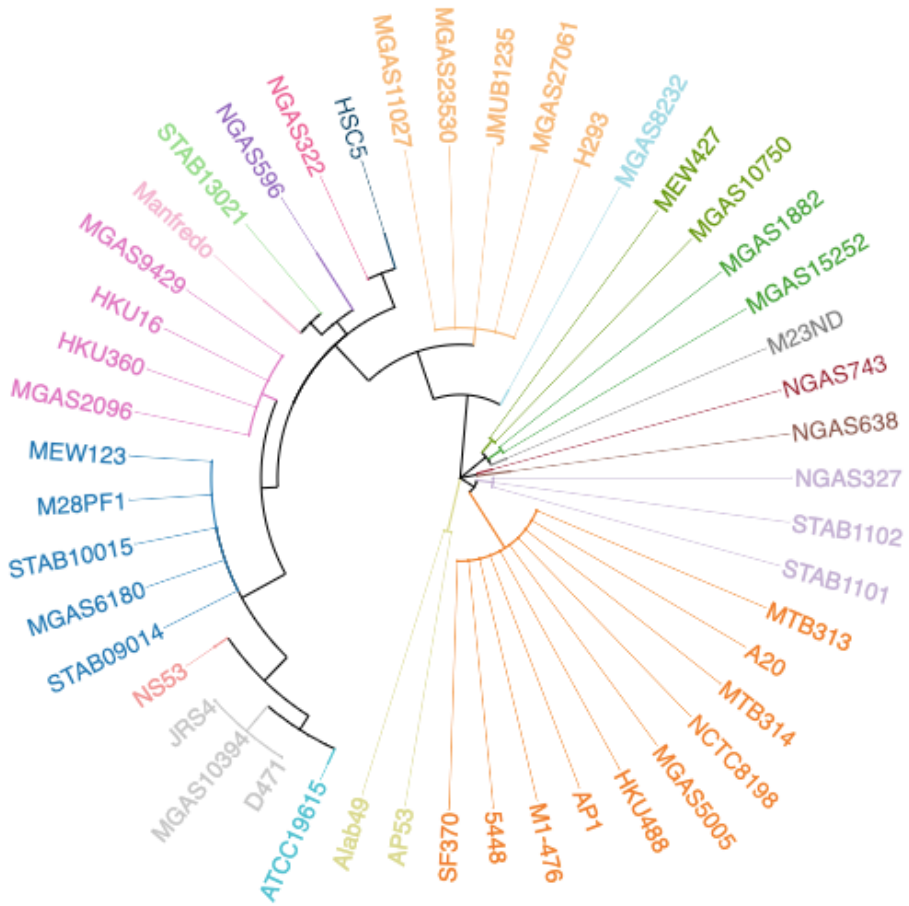
Fonte: Suzane de Andrade Barboza, 2019

Em alguns casos, genes anotados com o mesmo produto foram agrupados em famílias diferentes, algumas delas compostas por apenas um ou dois isolados. Este evento ocorreu quando duas proteínas muito parecidas, como adesinas, foram anotadas com o mesmo produto, mas diferenciadas no momento do alinhamento. Em alguns casos, como ocorreu com as proteínas M e M-like, não foi possível separar corretamente as duas proteínas. Nesses casos, as sequências tiveram que ser baixadas manualmente da base NCBI para a análise. Esses exemplos ilustram as dificuldades enfrentadas no momento da definição dos parâmetros

de alinhamento. Na maioria dos casos, porém, todos os genes foram corretamente separados em clusters distintos e com alto coeficiente de clusterização, como visto anteriormente na figura 8, indicando que, apesar de problemas pontuais, os parâmetros escolhidos foram uma boa combinação para o caso desse estreptococo.

Para todos os fatores de virulência foi feito um alinhamento múltiplo, que foi usado para a construção de cladogramas. Nota-se nos cladogramas que quase todos os isolados com o mesmo genótipo também foram agrupados no mesmo ramo da árvore (Figura 13), indicando que os fatores de virulência podem ter sofrido a mesma pressão imunológica que levou à diferenciação da proteína M (PANCHAUD et al., 2009).

Figura 13 – Cladograma criado à partir da proteína SmeZ. As linhas espessas representam os braços da árvore, enquanto os traços mais finos foram utilizados para facilitar a visualização dos dados. Os isolados foram coloridos de acordo com seus genótipos.



Fonte: Suzane de Andrade Barboza, 2019

De todos as cepas deste estudo, apenas duas estão associadas com impetigo, ambas *emm*53. Nesses isolados, foi percebido dois eventos exclusivos: uma substituição e uma deleção de bases em uma Salivaricina A (fig|1314.373.peg.1578) e a ausência de uma proteína anotada como uma salivaricina putativa (fig|1314.362.peg.1722), que está presente em todos os outros isolados. Embora a salivaricina seja induzida apenas na presença de saliva, a conservação desta proteína em quase todos os isolados *S. pyogenes* sugere a possibilidade dessa proteína estar ligada a algum processo ainda desconhecido (VEGA; MALKE; MCIVER, 2016).

Em relação aos isolados *emm*1 associados à meningite, uma substituição de um aminoácido foi encontrada na proteína CapA (fig|1314.400.peg.630). Da mesma forma, percebe-se pequenas diferenças nas proteínas C3 (fig|1314.400.peg.341) dos isolados *emm*59 invasivos e não-invasivos. Interessante notar que já foi relatado anteriormente que proteínas da família C3 têm ligação com o potencial invasivo de *S. pyogenes* (COYE; COLLINS, 2004; HOFF et al., 2011). Já em relação à exotoxina G (fig|1314.400.peg.178) percebe-se maiores diferenças entre os dois isolados, onde a proteína do isolado *emm*59 invasivo é idêntica à proteína do isolado *emm*66 invasivo, sendo os dois isolados associados a doenças na pele. A exotoxina G é classificada como superantígeno, ligada diretamente à expressão da resposta imune do hospedeiro (SRISKANDAN; FAULKNER; HOPKINS, 2007), o que significa que a semelhança observada nessas proteínas pode estar relacionada à invasividade dos isolados *emm*59 e *emm*66.

Uma proteína semelhante à SmeZ (fig|1314.358.peg.1662) foi encontrada em todas as cepas *emm*3, *emm*4 e *emm*114. Entretanto, nos isolados *emm*3 (todos associados à síndrome do choque tóxico) esta proteína é idêntica e diferente da encontrada nos isolados *emm*4 e *emm*114. É interessante ressaltar que já foi descrita uma deleção na proteína SmeZ de cepas *emm*3 ligadas à essa síndrome que origina um códon de parada (*stop codon*) prematuro, que levou à diminuição do potencial invasivo dessas cepas (TURNER et al., 2012). Em relação à exotoxina J (fig|1314.362.peg.784), cuja atividade é induzida durante infeção na faringe (VIRTANEVA et al., 2005), nota-se a substituição de um aminoácido exclusivamente nos isolados *emm*4 associados à faringite.

É interessante destacar que uma adesina ligada ao colágeno (fig|1314.356.peg.107) e uma proteína da família C5 (fig|1314.370.peg.1870) são idênticas nos isolados não invasivos e menores do que as presentes nos isolados invasivos. O fenômeno oposto é percebido em uma permease (fig|1314.404.peg.607) e em uma toxina exfoliativa (fig|1314.363.peg.778), nas quais as proteínas dos isolados não invasivos são maiores do que as presentes nos isolados

invasivos. Essa diferença pode ter um impacto na invasividade de *S. pyogenes* de forma semelhante ao que ocorre com a proteína SclA (FLORES et al., 2015), onde uma deleção presente nos isolados invasivos acaba gerando um stop códon prematuro, gerando uma proteína mais invasiva do que a original.

A fim de complementar as observações realizadas à partir dos cladogramas foi criada uma relação entre o número de cópias de cada um dos fatores de virulência, a fim de destacar os fatores que mais estivessem relacionados ao perfil invasivo, doença ou genótipo, chamados de classificadores. O subconjunto de classificadores mais acurado contém 17 proteínas e conseguiu separar corretamente todos os genomas de acordo com seu genótipo em 100% dos casos, indicando que genomas com diferentes genótipos possuem combinações distintas de fatores de virulência (Figura 14). A relação dos atributos e famílias utilizados neste conjunto de classificadores é dada no apêndice D. Percebe-se, pela figura que alguns padrões de comportamento que parecem ser associados a alguns genótipos poderiam ser explicados pela diferente combinação desses 17 fatores de virulência. É interessante notar que apenas os genomas de isolados *emm*6 não foram agrupados em um único ramo dessa árvore. Diferentemente dos isolados associados com febre reumática, a cepa MGAS10394 (faringite) é a única *emm*6 que possui uma cópia do fator mitogênico 3. Não há uma conexão clara entre esse fator e faringite, por isso essa diferença poderia ser explicada pela barreira temporal existente entre os isolados, já que a cepa ligada à faringite foi isolada em 2004, enquanto as cepas associadas à febre reumática foram isoladas apenas em 2015.

Figura 14 – Árvore de decisão dos classificadores (ou atributos) para o genótipo *emm*. Cada galho da árvore representa um fator de virulência, para o qual é designado um valor F (falso) ou T (verdadeiro) indicando a presença ou ausência deste fator naquele grupo. Dois valores são designados na frente de cada grupo: o total de genomas classificados no grupo / o número de genomas classificados incorretamente.

```
RandomTree - M Genotype
==========

76_SpeJ = T
|  5_Exfoliative toxin-like = F
|  |  59_Hyaluronidase = F
|  |  |  102_MutR = T
|  |  |  |  79_SmeZ = F : M44 (2/0)
|  |  |  |  79_SmeZ = T : M12 (4/0)
|  |  |  102_MutR = F
|  |  |  |  109_Mga = T
|  |  |  |  |  65_Mitogenic factor 3 = F : M82 (1/0)
|  |  |  |  |  65_Mitogenic factor 3 = T : M2 (1/0)
|  |  |  |  109_Mga = F : M83 (3/0)
|  |  59_Hyaluronidase = T
|  |  |  62_Mitogenic factor 1 = F
|  |  |  |  72_SpeH = T : M23 (1/0)
|  |  |  |  72_SpeH = F
|  |  |  |  |  109_Mga = T : M114 (1/0)
|  |  |  |  |  109_Mga = F : M4 (2/0)
|  |  |  62_Mitogenic factor 1 = T
|  |  |  |  65_Mitogenic factor 3 = F : M6 (2/0)
|  |  |  |  65_Mitogenic factor 3 = T
|  |  |  |  |  12_IdeS = F : M6 (1/0)
|  |  |  |  |  12_IdeS = T : M18 (1/0)
|  5_Exfoliative toxin-like = T
|  |  125_Conjugal protein = F
|  |  |  24_GRAB = T : M89 (5/0)
|  |  |  24_GRAB = F : M14 (1/0)
|  |  125_Conjugal protein = T : M28 (5/0)
76_SpeJ = F
|  118_RalP = T
|  |  105_MutR = T
|  |  |  77_SpeK = F : M1 (10/0)
|  |  |  77_SpeK = T : M87 (1/0)
|  |  105_MutR = F
|  |  |  5_Exfoliative toxin-like = F
|  |  |  |  72_SpeH = T : M49 (1/0)
|  |  |  |  72_SpeH = F : M59 (2/0)
|  |  |  5_Exfoliative toxin-like = T
|  |  |  |  67_SpeA = F : M66 (1/0)
|  |  |  |  67_SpeA = T : M71 (1/0)
|  118_RalP = F
|  |  12_IdeS = F
|  |  |  79_SmeZ = F : M3 (4/0)
|  |  |  79_SmeZ = T
|  |  |  |  8_Exfoliative toxin-like = F
|  |  |  |  |  67_SpeA = F : M101 (1/0)
|  |  |  |  |  67_SpeA = T : M80 (1/0)
|  |  |  |  8_Exfoliative toxin-like = T : M53 (2/0)
|  |  12_IdeS = T : M5 (1/0)
```

Fonte: Suzane de Andrade Barboza, 2019

A fim de auxiliar o estudo da invasividade de *S. pyogenes*, um segundo conjunto de classificadores foi criado, dessa vez em relação ao perfil invasivo das cepas. Utilizando 15 genes. 44 dos 55 genomas (83,0189%) foram corretamente classificados (Figura 15). A relação dos atributos e famílias utilizados neste conjunto de classificadores é dada no apêndice E.

Figura 15 – Árvore de decisão dos classificadores (ou atributos) para o perfil invasivo. Cada galho da árvore representa um fator de virulência, para o qual é designado um valor F (falso) ou T (verdadeiro) indicando a presença ou ausência deste fator naquele grupo. Dois valores são designados na frente de cada grupo: o total de genomas classificados no grupo / o número de genomas classificados incorretamente.

```
RandomTree - Invasiveness
==========

28_Collagen like proteins = F
|  69_SpeC = F
|  |  53_Hemolysin = F
|  |  |  71_SpeG = F
|  |  |  |  80_SmeZ = F
|  |  |  |  |  38_SagB = F
|  |  |  |  |  |  7_Exfoliative toxin-like = F : Invasive (17/4)
|  |  |  |  |  |  7_Exfoliative toxin-like = T : Invasive (1/0)
|  |  |  |  |  38_SagB = T : Invasive (2/0)
|  |  |  |  80_SmeZ = T : Invasive (1/0)
|  |  |  71_SpeG = T : Non-Invasive (1/0)
|  |  53_Hemolysin = T : Non-Invasive (1/0)
|  69_SpeC = T
|  |  70_SpeG = T
|  |  |  7_Exfoliative toxin-like = F
|  |  |  |  12_IdeS = F
|  |  |  |  |  88_Isp = F
|  |  |  |  |  |  63_Mitogenic factor 1 = F
|  |  |  |  |  |  |  32_Internalin-like protein = F
|  |  |  |  |  |  |  |  74_SpeI = F
|  |  |  |  |  |  |  |  |  80_SmeZ = F
|  |  |  |  |  |  |  |  |  |  8_Exfoliative toxin-like = F : Non-Invasive (12/5)
|  |  |  |  |  |  |  |  |  |  8_Exfoliative toxin-like = T : Non-Invasive (2/0)
|  |  |  |  |  |  |  |  |  80_SmeZ = T : Invasive (1/0)
|  |  |  |  |  |  |  |  74_SpeI = T : Non-Invasive (1/0)
|  |  |  |  |  |  |  32_Internalin-like protein = T : Non-Invasive (1/0)
|  |  |  |  |  |  63_Mitogenic factor = T : Non-Invasive (1/0)
|  |  |  |  |  88_Isp = T : Non-Invasive (1/0)
|  |  |  |  12_IdeS = T : Non-Invasive (2/0)
|  |  |  7_Exfoliative toxin-like = T : Invasive (1/0)
|  |  70_SpeG = F : Non-Invasive (4/0)
28_Collagen like proteins = T : Invasive (4/0)
```

Fonte: Suzane de Andrade Barboza, 2019

É possível perceber na figura 15 que não existe um único fator de virulência que seja responsável diretamente pela invasividade dos isolados, mas que a diferente combinação destes 15 fatores leva a padrões que poderiam explicar diferentes comportamentos observados. Embora esse conjunto de classificadores tenha classificado incorretamente 11 cepas, é possível que um conjunto maior de genes (considerando, por exemplo, proteínas hipotéticas) possa melhorar o número de acertos e ser usado, dessa forma, como um modelo de classificação complementar ao do genótipo *emm*.

Adicionalmente, foi criado um terceiro conjunto de classificadores, com o objetivo de auxiliar na compreensão da associação da presença de fatores de virulência com as doenças. Representado pela figura 16, o subconjunto de 23 genes conseguiu separar corretamente 48 genomas (87,2727%). Duas das 7 cepas classificadas incorretamente não continham em suas informações a doença à qual eram relacionados e por esse motivo foram, neste estudo, classificadas como perfil desconhecido. Essa classificação pode gerar erros, já que no momento da montagem da árvore, o programa obrigatoriamente considera as duas cepas como relacionadas à mesma doença, o que não necessariamente é verdade. Por esse motivo, considera-se que apenas 5 cepas foram realmente classificadas erroneamente. A relação dos atributos e famílias utilizados neste conjunto de classificadores é dada no F.

Figura 16 - Árvore de decisão dos classificadores (ou atributos) para doenças. Cada galho da árvore representa um fator de virulência, para o qual é designado um valor F (falso) ou T (verdadeiro) indicando a presença ou ausência deste fator naquele grupo. Dois valores são designados na frente de cada grupo: o total de genomas classificados no grupo / o número de genomas classificados incorretamente. Abreviações utilizadas: ARF (febre reumática), APG (gastrite aguda), BAC (bacteriemia), CFI (infecção do fluido cerebrospinal), ENDO (endometrite), IMP (impetigo), MENIN (meningite), NF (fasciíte necrosante), PANDAS (desordens pediátricas neuropsiquiátricas associadas à sequelas autoimunes), PH (faringite), PS (sepse puerperal), PSC (celulite perianal) SA (abcesso subcutâneo), SD (dermatite superficial), SF (febre escarlate), SI (infeção da pele), STI (infeção do tecido superficial), STSS (síndrome do choque tóxico estreptocócico), U (desconhecido) e WI (infecção pulmonar).

```
RandomTree - Disease
==========

58_Hyaluronidase = F
|  65_Mitogenic factor = F
|  |  64_Mitogenic factor = F
|  |  |  79_SmeZ = F
|  |  |  |  24_GRAB = T : U (1/0)
|  |  |  |  24_GRAB = F : ENDO (1/0)
|  |  |  79_SmeZ = T
|  |  |  |  5_Exfoliative toxin-like = F
|  |  |  |  |  118_Ralp = T : STI (1/0)
|  |  |  |  |  118_Ralp = F : NF (2/0)
|  |  |  |  5_Exfoliative toxin-like = T
|  |  |  |  |  67_SpeA = F
|  |  |  |  |  |  109_Mga = T : NF (1/0)
|  |  |  |  |  |  109_Mga = F : PH (1/0)
|  |  |  |  |  67_SpeA = T : SI (1/0)
|  |  64_Mitogenic factor 2 = T
|  |  |  62_Mitogenic factor 1 = F
|  |  |  |  56_Streptococcal phospholipase = F : PH (2/1)
|  |  |  |  56_Streptococcal phospholipase = T : ENDO (3/2)
|  |  |  62_Mitogenic factor 1 = T : APG (1/0)
|  65_Mitogenic factor 3 = T
|  |  64_Mitogenic factor 2 = F
|  |  |  59_Hyaluronidase = F : MENIN (2/0)
|  |  |  59_Hyaluronidase = T : NF (4/2)
|  |  64_Mitogenic factor = T : SF (1/0)
58_Hyaluronidase = T
|  59_Hyaluronidase = F
|  |  79_SmeZ = F
|  |  |  24_GRAB = T : APG (1/0)
|  |  |  24_GRAB = F : SD (1/0)
|  |  79_SmeZ = T
|  |  |  62_Mitogenic factor 1 = F
|  |  |  |  76_SpeJ = T
|  |  |  |  |  5_Exfoliative toxin-like  = F : BAC (2/0)
|  |  |  |  |  5_Exfoliative toxin-like  = T
|  |  |  |  |  |  105_MutR = T : U (2/1)
|  |  |  |  |  |  105_MutR = F : PH (1/0)
|  |  |  |  76_SpeJ = F
|  |  |  |  |  102_MutR = T : PH (1/0)
|  |  |  |  |  102_MutR = F : APG (1/0)
|  |  |  62_Mitogenic factor 1 = T
|  |  |  |  123_Salivaricin = F
|  |  |  |  |  68_Phage exotoxin  = F : PH (1/0)
|  |  |  |  |  68_Phage exotoxin  = T : SF (2/0)
|  |  |  |  123_Salivaricin = T : BAC (1/0)
|  59_Hyaluronidase (hyl) = T
|  |  102_MutR = T
|  |  |  69_SpeC = F
|  |  |  |  5_Exfoliative toxin-like  = F : STSS (4/0)
|  |  |  |  5_Exfoliative toxin-like  = T
|  |  |  |  |  123_Salivaricin = F : SA (1/0)
|  |  |  |  |  123_Salivaricin = T : U (1/0)
|  |  |  69_SpeC = T
|  |  |  |  68_Phage exotoxin  = F
|  |  |  |  |  5_Exfoliative toxin-like  = F : BAC (1/0)
|  |  |  |  |  5_Exfoliative toxin-like  = T : IMP (2/0)
|  |  |  |  68_Phage exotoxin  = T
|  |  |  |  |  5_Exfoliative toxin-like  = F
|  |  |  |  |  |  17_C5a peptidase  = F : PH (1/0)
|  |  |  |  |  |  17_C5a peptidase  = T : PANDAS (1/0)
|  |  |  |  5_Exfoliative toxin-like  = T : NF (1/0)
|  |  102_MutR = F
|  |  |  118_Ralp = T
|  |  |  |  67_SpeA = F : WI (1/0)
|  |  |  |  67_SpeA = T : U (2/1)
|  |  |  118_Ralp = F
|  |  |  |  56_Streptococcal phospholipase  = F
|  |  |  |  |  68_Phage exotoxin  = F : ARF (4/0)
|  |  |  |  |  68_Phage exotoxin  = T : NF (1/0)
|  |  |  |  56_Streptococcal phospholipase  = T : PH (1/0)
```

Fonte: Suzane de Andrade Barboza, 2019

É possível notar que cepas relacionadas às mesmas doenças podem apresentar combinações muito diferentes desses fatores, indicando que a expressão/regulação desses fatores deve ser mais determinante no desenvolvimento de quadros infecciosos mais invasivos do que a simples presença/ausência desses fatores no genoma (KANSAL et al., 2000; SUMBY et al., 2005, 2006).

4.3 Identificação de genes exclusivos

São chamados genes exclusivos os genes que estão presentes em genomas que compartilham um certo conjunto de características e ausentes em todos os outros isolados. A identificação de genes exclusivos pode auxiliar no estudo do comportamento de *S. pyogenes,* por isso, os isolados foram agrupados de acordo com o perfil de invasividade, genótipo e associação a doenças. Grupos compostos por apenas um isolado foram desconsiderados, e cepas com características desconhecidas foram analisadas independentemente.

Em relação à invasividade, não há nenhum gene cuja presença ou ausência seja exclusiva às cepas de acordo com o perfil invasivo ou não-invasivo. Entretanto, quando o perfil de invasividade foi analisado dentro dos grupos de genótipos, foi destacado um gene presente apenas em cepas *emm*89: fig|1314.381.peg.164 (*mobile element protein*).

A relação de genes exclusivamente presentes em cepas agrupadas por seus genótipos é dada no quadro 1, enquanto a relação dos genes exclusivos presentes nos isolados agrupados em relação às doenças associadas é dada no quadro 2. Essas relações podem auxiliar futuros experimentos que visem estabelecer uma relação entre fatores de virulência, genótipos ou doenças (CHAUSSEE et al., 2007).

Quadro 1 – Relação de genes exclusivos por genótipo. O número da anotação refere-se à chave de identificação na base de dados de *S. pyogenes* disponibilizada neste estudo.

Genótipo	Número da anotação	Produto (anotação)
emm 1	fig|1314.415.peg.72	4-diphosphocytidyl-2-C-methyl-D-erythritol kinase
emm 1	fig|1314.383.peg.372	Hypothetical protein
emm 1	fig|1314.415.peg.944	Short-chain dehydrogenase
emm 1	fig|1314.383.peg.1955	Transposase
emm 1	fig|1314.415.peg.1628	Transport protein SgaT, putative
emm 1	fig|1314.408.peg.1070	5'-nucleotidase YjjG (EC 3.1.3.5)

emm 1	fig|1314.408.peg.1071	Acetyltransferase
emm 1	fig|1314.408.peg.1072	Hypothetical protein
emm 1	fig|1314.408.peg.1678	Streptococcal inhibitor of complement (SIC)
emm 12	fig|1314.359.peg.192	Mobile element protein
emm 12	fig|1314.359.peg.1561	Hypothetical protein
emm 12	fig|1314.403.peg.1043	Uridine phosphorylase (EC 2.4.2.3)
emm 12	fig|1314.404.peg.1674	Drs12.02 gene
emm 12	fig|1314.404.peg.1110	Uridine phosphorylase (EC 2.4.2.3)
emm 28	fig|1314.374.peg.1191	Late competence protein ComEC, DNA transport
emm 28	fig|1314.401.peg.769	Hypothetical protein
emm 28	fig|1314.401.peg.771	Hypothetical protein
emm 28	fig|1314.401.peg.773	Transcriptional regulator, MarR family
emm 28	fig|1314.401.peg.774	Hypothetical protein
emm 28	fig|1314.401.peg.775	Hypothetical protein
emm 28	fig|1314.401.peg.777	Hypothetical protein
emm 28	fig|1314.401.peg.1506	Hypothetical protein
emm 28	fig|1314.389.peg.765	Phage integrase
emm 28	fig|1314.401.peg.772	Hypothetical protein
emm 28	fig|1314.401.peg.776	Probable mobilization protein MobA,
emm 28	fig|1314.382.peg.352	Hypothetical protein
emm 28	fig|1314.401.peg.770	Hypothetical protein
emm 28	fig|1314.374.peg.1076	Hypothetical protein
emm 28	fig|1314.401.peg.768	DUF1706 domain-containing protein
emm 28	fig|1314.365.peg.342	Hypothetical protein
emm 28	fig|1314.374.peg.141	Hypothetical protein
emm 28	fig|1314.401.peg.780	Hypothetical protein
emm 3	fig|1314.407.peg.168	Transcriptional regulator, RofA-like Protein (RALP)
emm 3	fig|1314.360.peg.1066	Mobile element protein
emm 3	fig|1314.371.peg.194	ABC transporter, N-acetylneuraminate-binding protein
emm 3	fig|1314.360.peg.310	Hypothetical protein
emm 3	fig|1314.371.peg.1623	Hypothetical protein
emm 3	fig|1314.371.peg.948	Phage protein
emm 3	fig|1314.371.peg.555	Hypothetical protein
emm 3	fig|1314.371.peg.958	Phage protein
emm 3	fig|1314.407.peg.567	Phage protein
emm 3	fig|1314.371.peg.1778	Collagen-like surface protein
emm 3	fig|1314.371.peg.562	ORF009
emm 3	fig|1314.371.peg.1790	Exotoxin, Streptococcal mitogenic exotoxin Z (SmeZ)
emm 3	fig|1314.391.peg.137	Hypothetical protein
emm 3	fig|1314.391.peg.533	Hypothetical protein
emm 4	fig|1314.366.peg.27	Putative choline binding protein
emm 4	fig|1314.366.peg.28	Putative choline binding protein
emm 4	fig|1314.366.peg.118	Sortase A, LPXTG specific
emm 4	fig|1314.366.peg.350	Hypothetical protein
emm 4	fig|1314.390.peg.1121	Mobile element protein

emm 4	fig|1314.366.peg.1165	Outer surface protein, cellobiose operon
emm 4	fig|1314.366.peg.1248	Kup system potassium uptake protein
emm 4	fig|1314.390.peg.1571	Transcriptional regulator, MerR family
emm 4	fig|1314.366.peg.79	Hypothetical protein
emm 4	fig|1314.390.peg.1639	Hypothetical protein
emm 4	fig|1314.390.peg.32	Hypothetical protein
emm 4	fig|1314.390.peg.33	Possible EA31 gene protein, phage lambda
emm 4	fig|1314.390.peg.117	Cell wall surface anchor family protein, LPXTG motif
emm 4	fig|1314.390.peg.176	Hypothetical protein
emm 4	fig|1314.390.peg.262	Isochorismatase (EC 3.3.2.1)
emm 4	fig|1314.390.peg.1795	Hypothetical protein
emm 44	fig|1314.400.peg.351	Hypothetical protein
emm 44	fig|1314.370.peg.396	Hypothetical protein
emm 44	fig|1314.400.peg.144	Cystathionine beta-lyase (EC 4.4.1.8)
emm 44	fig|1314.400.peg.329	Putative oxidase STM1620
emm 44	fig|1314.400.peg.1587	Integrase
emm 44	fig|1314.400.peg.742	Hypothetical protein
emm 44	fig|1314.400.peg.829	Hypothetical protein
emm 44	fig|1314.400.peg.1028	Hypothetical protein
emm 44	fig|1314.400.peg.1539	Hypothetical protein
emm 44	fig|1314.400.peg.1588	Integrase
emm 44	fig|1314.400.peg.1589	Hypothetical protein
emm 44	fig|1314.400.peg.1590	Hypothetical protein
emm 44	fig|1314.400.peg.1591	Hypothetical protein
emm 44	fig|1314.400.peg.1592	Hypothetical cytosolic protein
emm 44	fig|1314.400.peg.1686	Hypothetical protein
emm 53	fig|1314.386.peg.341	Hypothetical protein
emm 53	fig|1314.386.peg.858	Exfoliative toxin
emm 59	fig|1314.397.peg.83	Late competence protein ComGA
emm 59	fig|1314.399.peg.351	Hypothetical protein
emm 59	fig|1314.397.peg.343	Hypothetical protein
emm 59	fig|1314.397.peg.344	Hypothetical protein
emm 59	fig|1314.399.peg.352	Hypothetical protein
emm 59	fig|1314.397.peg.401	Multidrug resistance efflux pump PmrA
emm 59	fig|1314.397.peg.1487	Putative SalK homologue
emm 59	fig|1314.397.peg.989	Exfoliative toxin
emm 59	fig|1314.399.peg.1307	Hypothetical protein
emm 59	fig|1314.399.peg.1506	Hypothetical protein
emm 59	fig|1314.399.peg.1521	Hypothetical protein
emm 6	fig|1314.356.peg.774	Mobile element protein
emm 6	fig|1314.356.peg.324	Hypothetical protein
emm 6	fig|1314.406.peg.432	ATPase
emm 6	fig|1314.406.peg.1776	Dipeptidase (EC 3.4.-.-)
emm 6	fig|1314.406.peg.394	Mobile element protein
emm 6	fig|1314.406.peg.395	Mobile element protein

emm 6	fig|1314.406.peg.1327	Hypothetical protein
emm 6	fig|1314.356.peg.349	Hypothetical protein
emm 6	fig|1314.356.peg.727	Hyaluronate lyase precursor (EC 4.2.2.1)
emm 6	fig|1314.356.peg.1033	Hypothetical protein
emm 6	fig|1314.356.peg.1034	Site-specific recombinase
emm 6	fig|1314.356.peg.1035	Phage integrase
emm 6	fig|1314.356.peg.1036	Hypothetical protein
emm 6	fig|1314.356.peg.1038	LPXTG anchored putative adhesin
emm 6	fig|1314.356.peg.1040	Hypothetical protein
emm 83	fig|1314.394.peg.718	Hypothetical protein
emm 83	fig|1314.395.peg.941	Hypothetical protein
emm 89	fig|1314.392.peg.769	Subtilin transport ATP-binding protein spaT
emm 89	fig|1314.367.peg.1224	Butyrate-acetoacetate CoA-transferase subunit A
emm 89	fig|1314.367.peg.334	Hypothetical protein
emm 89	fig|1314.368.peg.349	Hypothetical protein
emm 89	fig|1314.367.peg.367	Hypothetical protein
emm 89	fig|1314.367.peg.890	Hypothetical protein
emm 89	fig|1314.367.peg.349	Phage integrase: site-specific recombinase
emm 89	fig|1314.367.peg.346	Hypothetical protein
emm 89	fig|1314.381.peg.341	Mobile element protein
emm 89	fig|1314.381.peg.342	Mobile element protein
emm 89	fig|1314.381.peg.343	Mobile element protein
emm 89	fig|1314.381.peg.347	Mobile element protein
emm 89	fig|1314.381.peg.1656	Hypothetical protein

Fonte: Suzane de Andrade Barboza, 2019

Quadro 2 – Relação de genes exclusivos por doenças. O número da anotação refere-se à chave de identificação na base de dados de *S. pyogenes* disponibilizada neste estudo.

Doença	Número da anotação	Produto (anotação)
Impetigo	fig|1314.386.peg.341	Hypothetical protein
Impetigo	fig|1314.386.peg.858	Exfoliative toxin
Meningite	fig|1314.409.peg.117	Transcriptional regulator, LysR family
Meningite	fig|1314.409.peg.1176	Cell division protein FtsQ
Meningite	fig|1314.409.peg.1120	Phage-associated protein
Meningite	fig|1314.409.peg.1119	Phage portal protein; Phage capsid and scaffold

Fonte: Suzane de Andrade Barboza, 2019

Em ambas as tabelas, pode-se perceber a presença de alguns fatores de virulência de grande importância para os processos de colonização e interação com o sistema imunológico

do hospedeiro, como as proteínas SIC e SmeZ, relacionadas com a defesa e ativação da resposta imune humana (FERNIE-KING et al., 2001; GERLACH; SCHMIDT; FLEISCHER, 2001), reguladores, adesinas e toxinas. Vale notar que a associação de proteínas exfoliativas com impetigo já foi descrita anteriormente (NISHIFUJI et al., 2010), e alguns membros da família de reguladores LysR possuem um importante papel na resistência contra as células de defesa do hospedeiro (VEGA et al., 2017). Assim sendo, futuras investigações sobre a expressão dessas proteínas (principalmente em relação às proteínas hipotéticas destacadas neste estudo) e sua relação com o metabolismo de *S. pyogenes*, afim de melhor compreendermos seu comportamento.

4.4 Identificação de genes-alvo para desenvolvimento de vacina anti-estreptocócica

Durante a análise de genes, foram encontrados 16 genes bem conservados que estão presentes em todos os isolados com apenas uma cópia no genoma. Esses genes foram alinhados contra o genoma humano GRCh38 (GenBank GCA_000001405.27) utilizando os parâmetros *default* do BLAST e -e = 0.00001. No total, 14 genes retornaram sem nenhum resultado de alinhamento, se tornando assim grandes candidatos para o desenvolvimento de uma vacina não-orientada à proteína M, com baixo risco para o surgimento de reações imunológicas cruzadas contra proteínas humanas (Quadro 3).

Quadro 3 – Caracterização dos genes-alvo para desenvolvimento de vacina anti-estreptocócica. O número da anotação refere-se à chave de identificação na base de dados de *S. pyogenes* disponibilizada neste estudo. O índice de diversidade compreende valores entre 0 e 1 baseado a diversidade das bases no alinhamento entre os componentes do *cluster*.

Produto (anotação)	Tamanho (aminoácidos)	Diversidade	Número da anotação
Streptolysin S precursor	53	0	fig|1314.400.peg.558
Streptolysin S biosynthesis protein C	303-352	0.0003212256	fig|1314.400.peg.560
Streptolysin S biosynthesis protein D	452	0.0000236070	fig|1314.400.peg.561
Streptolysin S biosynthesis protein E	201-223	0.0006057946	fig|1314.400.peg.562
Streptolysin S biosynthesis protein F	226-227	0.0005293627	fig|1314.400.peg.563

Streptolysin S export ATP-binding protein (SagG)	307	0.0000111977	fig|1314.359.peg.1365
Streptolysin S export transmembrane permease (SagH)	245-375	0.0006162584	fig|1314.400.peg.565
Streptolysin S export transmembrane permease (SagI)	261-372	0.0005594531	fig|1314.400.peg.566
Hemolysins and related proteins containing CBS domains	360-444	0.0001575032	fig|1314.400.peg.303
Immunodominant antigen A	204	0.0006287646	fig|1314.400.peg.1731
Immunoreactive protein Se23.5 (Fragment)	201-204	0.0001978877	fig|1314.400.peg.1058
RopB, Rgg-like transcription regulator	274-280	0.0001612609	fig|1314.395.peg.1510
Response regulator CsrR	228	0.0642524485	fig|1314.400.peg.270 (ortólogo 1)
Two-component response regulator SA14-24	230-236	0.0642524485	fig|1314.400.peg.270 (ortólogo 2)

Fonte: Suzane de Andrade Barboza, 2019

Sabe-se que o desenvolvimento da uma vacina anti-estreptocócica tem enfrentado vários desafios, como a minimização do risco de desenvolvimento de reações cruzadas que levam ao surgimento de sequelas autoimunes (MASSELL; HONIKMAN; AMEZCUA, 1969). A maior parte dos estudos orientados a este desenvolvimento focam nas regiões variável ou hipervariável da proteína M, já que é o principal fator de virulência de *S. pyogenes*, presente em todos os isolados. Entretanto, a despeito de todas as vantagens da utilização da proteína M como alvo, estão presentes também sérias limitações: quando baseada na região conservada da proteína, a extensão e eficiência da resposta imunológica ainda é incerta (MCARTHUR; WALKER, 2006; TSOI et al., 2015), e quando baseada na região variável, o desenvolvimento de respostas imunes de tipo-específico começam a se tornar uma barreira para o desenvolvimento de uma vacina de amplitude global (STEER et al., 2009a, 2016). Com isso em mente, diversos estudos começaram a se orientar no desenvolvimento de uma vacina não-orientada à proteína M. Entretanto, nenhuma vacina desse tipo chegou à fase de testes até o momento (VEKEMANS et al., 2019). Esta é a principal razão pela qual a identificação de possíveis alvos, como os 14 genes identificados neste estudo, se faz tão importante para a orientação de futuros estudos visados à proteção global contra *S. pyogenes*. Entre estes, merecem destaque as proteínas transmembranas como SagH e SagI que, devido à sua exposição na superfície membranar, facilitam o reconhecimento pelos anticorpos.

5. Conclusão e perspectiva

Este projeto teve como principal objetivo o estudo de genômica comparativa com o intuito de melhor entender a grande diferença no nível de invasividade que se observa em certos clones de *S. pyogenes*. Um dos maiores desafios para a comparação de genomas completos é a análise de um grande volume de dados, especialmente em casos em que o objeto de estudo se apresenta associado a uma vasta gama de genótipos e/ou doenças, e possui uma grande capacidade de rearranjo de DNA, como *S. pyogenes*.

O primeiro problema enfrentado na análise foi a recuperação de dados com anotações não-padronizadas. A reanotação dos genomas foi um passo essencial para se obter uma qualidade na análise dos alinhamentos realizados, garantindo a atualização das anotações mais antigas, já que o ano de isolamento das cepas selecionadas para este estudo varia de 2006 a 2016. À partir desses dados, foram realizados alinhamentos que buscaram agrupar os genes definidos como ortólogos em famílias. Durante a análise, foram identificadas duas situações-problema: (1) formação de famílias que agregaram genes diferentes porém anotados com o mesmo produto e (2) a aglomeração de genes anotados com mais de um produto.

A primeira situação tem origem já no momento da anotação, que não conseguiu diferenciar duas proteínas que agregam os mesmos motivos. Em outros momentos da análise, percebeu-se que certos fatores de virulência descritos na literatura não haviam sido reconhecidos ou corretamente anotados pela ferramenta escolhida, especialmente em relação às adesinas. Sobre esse fator, deve ser levado em consideração que a ferramenta escolhida foi desenvolvida especialmente para a anotação de genomas bacterianos, tendo assim dificuldade para reconhecer proteínas que mimetizam as proteínas humanas. Entretanto, o fato de a maioria das famílias de genes ter sido formada por genes que compartilhavam a mesma anotação indica que, apesar de suas limitações, a ferramenta foi consistente, o que foi confirmado com a análise dos alinhamentos de cada *cluster*. Ainda assim, pode ser de grande interesse o desenvolvimento de uma ferramenta de anotação mais voltada a esse estreptococo, ou a patógenos que possuam como hospedeiros os seres humanos.

A segunda situação problema é causada pela combinação de parâmetros definida após as anotações, que impediu a diferenciação de proteínas muito semelhantes como as proteína M e M-like. Nesta etapa, a representação das famílias de genes como *clusters* em uma rede gênica facilitou a avaliação dos parâmetros selecionados para a definição dos grupos de ortólogos. Analisando a formação de famílias agrupadas com um alto coeficiente de clusterização, percebeu-se que um aumento na sensibilidade dos alinhamentos poderia

acarretar a separação de grupos ortólogos que possuíssem uma grande variabilidade em algumas de suas regiões. Assim, foi escolhido o conjunto de parâmetros que resultasse no maior coeficiente de clusterização, preservando agrupamentos com divergências maiores.

Além da avaliação dos parâmetros, a criação de uma rede gênica auxiliou na identificação de presença/ausência de genes em grupos de isolados que compartilhassem as mesmas características, como invasividade. É correto afirmar que, apesar de a presença de uma vasta gama de fatores de virulência facilitar a adesão e sobrevivência de *S. pyogenes,* não há um gene ou conjunto de genes cuja presença está unicamente e diretamente relacionada comportamento invasivo desse estreptococo. Essa observação foi confirmada ainda com outras ferramentas em diferentes etapas da análise, como por exemplo durante a identificação de genes exclusivos ou genes classificadores. Entretanto, a análise dos cladogramas associada à leitura dos alinhamentos permitiu a identificação de certas diferenças entre as sequências de aminoácidos de isolados invasivos e não-invasivos em proteínas já descritas anteriormente como sendo ligadas ao aumento ou diminuição do comportamento invasivo deste patógeno. Partindo destas observações percebe-se que a regulação desses genes é a causa mais provável da origem da invasividade de *S. pyogenes,* possivelmente aliada ao modo de contágio e saúde imunológica do hospedeiro.

Apesar de não haver genes exclusivamente ligados às cepas invasivas, foram encontrados fatores relacionados ao impetigo e meningite, além de algumas adesinas, toxinas e reguladores relacionados a certos genótipos, o que deve ajudar a entender melhor a conexão entre alguns genótipos e as características fenotípicas observadas. Além disso, foram identificadas diversas proteínas hipotéticas ligadas ao genótipo *emm*. Estudos futuros focados na expressão dessas proteínas podem determinar se há uma relação entre essas proteínas e o comportamento de certos isolados.

A relação entre certos subconjuntos de genes e os genótipos *emm* também foi observada no cladograma de presença/ausência de genes, que conseguiu separar corretamente todos os isolados pelo genótipo, e permitiu a observação da separação dos isolados *emm*28 invasivos e não-invasivos em ramos diferentes da árvore, devido à presença de três proteínas nos isolados não-invasivos que se mostraram ausentes nos isolados invasivos. Apesar do cladograma ser afetado pela perda/incorporação de genes, conforme demonstrado pela disposição incorreta das cepas *emm*1 na árvore, a sensibilidade desse tipo de cladograma para a recombinação genômica deve ajudar a refletir melhor a evolução de *S. pyogenes*, levando em consideração o modo com que as barreiras temporal e geográfica afetam a aquisição de

DNA extragenômico. Um bom exemplo é a separação do isolado MGAS10394 dos outros isolados *emm*6, provavelmente causada pela barreira temporal existente entre os isolados.

A identificação de genes classificadores se fez importante na definição de um subconjunto de genes que separou corretamente todas as cepas de acordo com seu genótipo, demonstrando a diversidade de fatores de virulência entre cepas de mesmo genótipo. Já os classificadores voltados ao perfil invasivo e associação a doenças tiveram uma alta taxa de sucesso na predição do perfil dos isolados. Embora não tenham conseguido realizar corretamente todas as classificações, as árvores auxiliam na identificação de genes potencialmente relacionados às características mencionadas. O desempenho desses classificadores pode ser melhorado por meio da inclusão de dados sobre as proteínas hipotéticas. Além de muito úteis na definição do comportamento dos isolados, esses classificadores podem auxiliar ainda na identificação das proteínas mais provavelmente relacionadas com as características fenotípicas dos isolados, orientando dessa forma futuros estudos laboratoriais focados na compreensão do comportamento de *S. pyogenes*

Muito importante também é a análise de genes conservados em todas as cepas, especialmente no que concerne a identificação de possíveis genes alvo para o desenvolvimento de uma vacina. Considerando os desafios encontrados durante a confecção de vacinas baseadas na proteína M, 14 genes bem conservados são apresentados como uma alternativa para a produção de uma vacina com baixo risco para o desenvolvimento de sequelas imunológicas pós-infecciosas.

Por fim, a disponibilização de todos os resultados em uma base de dados exclusiva para *S. pyogenes* deve ajudar futuros estudos sobre genômica comparativa, facilitando o acesso aos dados e manipulação de dados de *S. pyogenes*.

5.1 Principais contribuições

As principais contribuições obtidas por meio deste trabalho foram:

- Determinação de que a origem da invasividade dos isolados de *S. pyogenes* não está ligada à simples presença ou ausência de fatores de virulência, apontando que o perfil invasivo desse estreptococo está provavelmente relacionada à regulação desses fatores;
- Identificação de 14 genes-alvo potenciais para o desenvolvimento de uma vacina não-orientada à proteína M;
- Identificação de genes potencialmente relacionados ao comportamento de *S. pyogenes*

que poderão ser usados como classificadores para certas características de relevância clínica;

- Identificação das principais diferenças presentes na sequência de aminoácidos de proteínas anteriormente descritas como relacionadas ao perfil invasivo desse patógeno;
- Identificação de genes exclusivamente relacionados ao genótipo *emm* ou à doença causada pelo isolado;
- Disponibilização de um portal exclusivamente dedicado a este patógeno, criado pelo aluno de doutorado Caio Rafael do Nascimento Santiago, unindo os resultados obtidos neste estudo com as ferramentas utilizadas e/ou criadas durante a realização da comparação genômica destes isolados, buscando assim centralizar e facilitar o acesso e processamento dos dados genômicos de *S. pyogenes.*

Espera-se que essas contribuições possam orientar e facilitar os futuros trabalhos de genômica comparativa e ensaios experimentais voltados à compreensão e combate a esse patógeno de grande relevância à saúde pública.

5.2 Trabalhos futuros

Entre as principais perspectivas desse projeto estão a manutenção e o desenvolvimento do portal disponibilizado, com o objetivo de permitir a constante atualização das anotações e a inserção de novos genomas completos, além da incorporação de novas ferramentas que auxiliem a recuperação e processamento de dados genômicos de *S. pyogenes.*

Em relação aos classificadores, se faz importante a inclusão de dados de proteínas hipotéticas aos fatores de virulência, buscando assim um aumento no número de classificações corretas. Dessa forma, os classificadores poderão ser utilizados como métodos complementares para a determinação do perfil dos isolados, além de contribuir na identificação de novos fatores de virulência.

Futuros trabalhos poderiam realizar uma análise mais profunda dos genes que foram incorretamente colocados no mesmo cluster e fatores de virulência que não foram anotados pela ferramenta, além do estudo das proteínas hipotéticas e reguladores. Além disso, faz-se importante o desenvolvimento de uma revisão bibliográfica que reúna os dados de epidemiologia molecular a nível global, para fins de monitoramento.

Já os pesquisadores da parte laboratorial poderão aproveitar os dados produzidos por este estudo a fim de determinar se as diferenças apontadas entre as proteínas de isolados

invasivos e não-invasivos têm um papel na expressão das mesmas, interferindo no comportamento dos isolados. Faz-se importante também determinar o papel das proteínas exclusivamente associadas ao genótipo ou doenças, principalmente as anotadas como hipotéticas. Além disso, a identificação potenciais genes-alvo para o desenvolvimento de uma vacina não-orientada à proteína M deve orientar futuros trabalhos na área de imunologia.

Por fim, pretende-se publicar os resultados aqui obtidos em um periódico da área.

Referências[1]

ARNDT, D. et al. PHASTER: a better, faster version of the PHAST phage search tool. **Nucleic acids research**, v. 44, n. W1, p. W16-21, 2016.

ARNAUD, M. B. et al. The Aspergillus Genome Database (AspGD): recent developments in comprehensive multispecies curation, comparative genomics and community resources. **Nucleic Acids Research**, v. 40, n. D1, p. D653–D659, 1 jan. 2012.

BANKS, D. J. et al. Progress toward Characterization of the Group A *Streptococcus* Metagenome: Complete Genome Sequence of a Macrolide ‐ Resistant Serotype M6 Strain. **The Journal of Infectious Diseases**, v. 190, n. 4, p. 727–738, 15 ago. 2004.

BAO, Y.-J. et al. Comparative pathogenomic characterization of a non-invasive serotype M71 strain Streptococcus pyogenes NS53 reveals incongruent phenotypic implications from distinct genotypic markers. **Pathogens and Disease**, v. 75, n. 5, 31 jul. 2017.

BAO, Y. et al. Unique genomic arrangements in an invasive serotype M23 strain of Streptococcus pyogenes identify genes that induce hypervirulence. **Journal of Bacteriology**, v. 196, n. 23, p. 4089–4102, 2014.

BARBOZA, S. A. et al. Complete genome sequence of noninvasive Streptococcus pyogenes M/emm28 strain STAB10015, isolated from a child with perianal dermatitis in French Brittany. **Genome Announcements**, v. 3, n. 4, 2015.

BARROS, S. F. DE et al. Streptococcus pyogenes strains in Sao Paulo, Brazil: molecular characterization as a basis for StreptInCor coverage capacity analysis. **BMC infectious diseases**, v. 15, p. 308, 5 ago. 2015.

BEACHEY, E. H. et al. Type-specific protective immunity evoked by synthetic peptide of Streptococcus pyogenes M protein. **Nature**, 1981.

BEN ZAKOUR, N. L. et al. Analysis of a Streptococcus pyogenes Puerperal Sepsis Cluster by Use of Whole-Genome Sequencing. **Journal of Clinical Microbiology**, v. 50, n. 7, p. 2224–2228, 2012.

BERES, S. B. et al. Molecular genetic anatomy of inter- and intraserotype variation in the human bacterial pathogen group A Streptococcus. **Proceedings of the National Academy of Sciences of the United States of America**, v. 103, n. 18, p. 7059–64, 2 maio 2006.

BERES, S. B. et al. Transcriptome Remodeling Contributes to Epidemic Disease Caused by the Human Pathogen Streptococcus pyogenes. **mBio**, v. 7, n. 3, 2016.

BESSEN, D. E. et al. **Genetic Correlates of Throat and Skin Isolates of Group A StreptococciThe Journal of Infectious Diseases**. [s.l: s.n.].

BESSEN, D. E. et al. Molecular epidemiology and genomics of group A Streptococcus. **Infection, genetics and evolution : journal of molecular epidemiology and evolutionary genetics in infectious diseases**, v. 33, p. 393–418, jul. 2015.

[1] De acordo com a Associação Brasileira de Normas Técnicas. NBR 6023.

BIDET, P.; BONACORSI, S. Facteurs de pathogénicité de Streptococcus pyogenes. . 1 nov. 2014, p. S54–S61.

BISNO, A.; BRITO, M.; COLLINS, C. Molecular basis of group A streptococcal virulence. **The Lancet Infectious Diseases**, v. 3, n. 4, p. 191–200, 1 abr. 2003.

BOMBACI, M. et al. Protein Array Profiling of Tic Patient Sera Reveals a Broad Range and Enhanced Immune Response against Group A Streptococcus Antigens. **PLoS ONE**, v. 4, n. 7, p. e6332, 22 jul. 2009.

BREIMAN, R. F. et al. Defining the Group A Streptococcal Toxic Shock Syndrome: Rationale and Consensus Definition. **JAMA: The Journal of the American Medical Association**, v. 269, n. 3, p. 390–391, 20 jan. 1993.

BRENCIANI, A. et al. Two distinct genetic elements are responsible for erm(TR)-mediated erythromycin resistance in tetracycline-susceptible and tetracycline-resistant strains of Streptococcus pyogenes. **Antimicrobial agents and chemotherapy**, v. 55, n. 5, p. 2106–12, maio 2011.

BRESSMANN, T. Self-inflicted cosmetic tongue split: A case report. **Journal of the Canadian Dental Association**, v. 70, n. 3, p. 156–157, 2004.

BROOKS, G. F. et al. **Microbiologia Médica de Jawetz, Melnick & Adelberg**. 26. ed. [s.l.] {AMGH} Editora, 2014.

CARAPETIS, J. R. et al. The global burden of group A streptococcal diseases. **The Lancet Infectious Diseases**, v. 5, n. 11, p. 685–694, 1 nov. 2005.

CHAUDHARI, N. M.; GUPTA, V. K.; DUTTA, C. BPGA- an ultra-fast pan-genome analysis pipeline. **Scientific Reports**, v. 6, n. 1, p. 24373, 13 jul. 2016.

CHAUSSEE, M. A. et al. Growth phase-associated changes in the transcriptome and proteome of Streptococcus pyogenes. **Archives of Microbiology**, v. 189, n. 1, p. 27–41, 23 nov. 2007.

CHEN, L. et al. VFDB: a reference database for bacterial virulence factors. **Nucleic acids research**, v. 33, n. Database issue, p. D325-8, 1 jan. 2005.

CHERRY, J. M. et al. Saccharomyces Genome Database: the genomics resource of budding yeast. **Nucleic Acids Research**, v. 40, n. D1, p. D700–D705, 1 jan. 2012.

COYE, L. H.; COLLINS, C. M. Identification of SpyA, a novel ADP-ribosyltransferase of Streptococcus pyogenes. **Molecular Microbiology**, v. 54, n. 1, p. 89–98, 11 ago. 2004.

CREEVEY, C. J.; MCINERNEY, J. O. Clann: investigating phylogenetic information through supertree analyses. **Bioinformatics**, v. 21, n. 3, p. 390–392, 1 fev. 2005.

CUI, L. et al. ChloroplastDB: the Chloroplast Genome Database. **Nucleic acids research**, v. 34, n. Database issue, p. D692-6, 1 jan. 2006.

CUNNINGHAM, M. W. Pathogenesis of Group A Streptococcal Infections. **Clinical microbiology reviews**, v. 13, n. 3, p. 470–511, 1 jul. 2000.

CUNNINGHAM, M. W. Pathogenesis of group A streptococcal infections and their sequelae. In: **Hot Topics in Infection and Immunity in Children IV**. [s.l.] Springer, 2008. p. 29–42.

D'HUMIÈRES, C. et al. Comparative epidemiology of Streptococcus pyogenes emm-types causing invasive and noninvasive infections in French children by use of high-resolution melting-polymerase chain reaction. **The Pediatric infectious disease journal**, v. 34, n. 6, p. 557–61, 1 jun. 2015.

DALE, R. C. et al. Poststreptococcal acute disseminated encephalomyelitis with basal ganglia involvement and auto-reactive antibasal ganglia antibodies. **Annals of Neurology**, v. 50, n. 5, p. 588–595, 1 nov. 2001.

DIGIAMPIETRI, L. et al. A gene based bacterial whole genome comparison toolkit. **REVISTA DE INFORMÁTICA TEÓRICA E APLICADA: RITA**, v. 26, n. 1, p. 36–46, 2019.

DING, W.; BAUMDICKER, F.; NEHER, R. A. panX: pan-genome analysis and exploration. **Nucleic Acids Research**, v. 46, n. 1, p. 1–12, 9 jan. 2018.

EFSTRATIOU, A.; LAMAGNI, T. Epidemiology of Streptococcus pyogenes. In: FERRETTI, J. J.; STEVENS, D. L.; FISCHETTI, V. A. (Eds.). . **Streptococcus pyogenes: Basic Biology to Clinical Manifestations**. [s.l.] University of Oklahoma Health Sciences Center, 2017. p. 1–44.

ENRIGHT, M. C. et al. Multilocus sequence typing of Streptococcus pyogenes and the relationships between emm type and clone. **Infection and immunity**, v. 69, n. 4, p. 2416–27, 1 abr. 2001.

FACKLAM, R. et al. emm typing and validation of provisional M types for group A streptococci. **Emerging infectious diseases**, v. 5, n. 2, p. 247–253, 1999.

FELSENSTEIN, J. **PHYLIP (Phylogeny Inference Package)**, 2005.

FERNANDES, G. R. et al. Genomic Comparison among Lethal Invasive Strains of Streptococcus pyogenes Serotype M1. **Frontiers in Microbiology**, v. 8, p. 1–10, 23 out. 2017.

FERNIE-KING, B. A. et al. Streptococcal inhibitor of complement (SIC) inhibits the membrane attack complex by preventing uptake of C567 onto cell membranes. **Immunology**, v. 103, n. 3, p. 390–8, jul. 2001.

FERRETTI, J. J. et al. Complete genome sequence of an M1 strain of Streptococcus pyogenes. . 10 abr. 2001, p. 4658–63.

FITTIPALDI, N. et al. Full-genome dissection of an epidemic of severe invasive disease caused by a hypervirulent, recently emerged clone of group A Streptococcus. **The American journal of pathology**, v. 180, n. 4, p. 1522–34, 1 abr. 2012.

FLORES, A. R. et al. Natural variant of collagen-like protein a in serotype M3 group a Streptococcus increases adherence and decreases invasive potential. **Infection and immunity**, v. 83, n. 3, p. 1122–9, mar. 2015.

GERLACH, D.; SCHMIDT, K. H.; FLEISCHER, B. Basic streptococcal superantigens (SPEX/SMEZ or SPEC) are responsible for the mitogenic activity of the so-called mitogenic factor (MF). **FEMS Immunology and Medical Microbiology**, v. 30, n. 3, p. 209–216, 2001.

GREEN, N. M. M. et al. Genome sequence of a serotype M28 strain of group A Streptococcus: potential new insights into puerperal sepsis and bacterial disease specificity. **The Journal of Infectious Diseases**, v. 192, n. 5, p. 760–770, 1 set. 2005.

HENNINGHAM, A. et al. Pathogenesis of Group A Streptococcal Infections. **Discov med.**, v. 13, n. 72, p. 329–342, 2012.

HENRY, V. J. et al. OMICtools: an informative directory for multi-omic data analysis. **Database : the journal of biological databases and curation**, v. 2014, 2014.

HOFF, J. S. et al. SpyA, a C3-like ADP-ribosyltransferase, contributes to virulence in a mouse subcutaneous model of Streptococcus pyogenes infection. **Infection and immunity**, v. 79, n. 6, p. 2404–11, 1 jun. 2011.

HOPKINS, S.; MACLEAN, A. B. Group A streptococcus: A continued threat. **Journal of Obstetrics and Gynaecology**, v. 26, n. 7, p. 593–595, 2006.

IBRAHIM, J. et al. Genome analysis of streptococcus pyogenes associated with pharyngitis and skin infections. **PLoS ONE**, v. 11, n. 12, 2016.

IKEBE, T. et al. Highly Frequent Mutations in Negative Regulators of Multiple Virulence Genes in Group A Streptococcal Toxic Shock Syndrome Isolates. **PLoS Pathogens**, v. 6, n. 4, 2010.

INGLIS, D. O. et al. The Candida genome database incorporates multiple Candida species: multispecies search and analysis tools with curated gene and protein information for Candida albicans and Candida glabrata. **Nucleic Acids Research**, v. 40, n. Database issue, p. D667--D674, 2012.

KANSAL, R. G. et al. Inverse Relation between Disease Severity and Expression of the Streptococcal Cysteine Protease, SpeB, among Clonal M1T1 Isolates Recovered from Invasive Group A Streptococcal Infection Cases. **Infection and immunity**, v. 68, n. 11, nov. 2000.

KATZ, A. R.; MORENS, D. M. Severe Streptococcal Infections in Historical Perspective. **Clinical Infectious Diseases**, v. 14, n. 1, p. 298–307, 1 jan. 1992.

LAMAGNI, T. L. et al. Epidemiology of severe Streptococcus pyogenes disease in Europe. **Journal of Clinical Microbiology**, v. 46, n. 7, p. 2359–2367, jul. 2008.

LANCEFIELD, R. A SEROLOGICAL DIFFERENTIATION OF HUMAN AND OTHER GROUPS OF HEMOLYTIC STREPTOCOCCI. **The Journal of experimental medicine**, v. 57, n. 4, p. 571–595, 1933.

LEFÉBURE, T.; STANHOPE, M. J. Evolution of the core and pan-genome of Streptococcus: Positive selection, recombination, and genome composition. **Genome Biology**, v. 8, n. 5, 2007.

LONGO, M. et al. Complete Genome Sequence of Streptococcus pyogenes emm28 Clinical Isolate M28PF1, Responsible for a Puerperal Fever. **Genome announcements**, v. 3, n. 4, 16 jul. 2015.

MACHEBOEUF, P. et al. Streptococcal M1 protein constructs a pathological host fibrinogen network Atomic coordinates and structure factors for M1 BC1-FgD (2XNX) and M1 A-FgD (2XNY) have been deposited with the Protein Data Bank. HHS Public Access. **Nature**, v. 472, n. 7341, p. 64–68, 2011.

MARTINS, T. B. et al. Comprehensive analysis of antibody responses to streptococcal and tissue antigens in patients with acute rheumatic fever. **International Immunology**, v. 20, n. 3, p. 445–452, 1 mar. 2008.

MARUYAMA, F.; WATANABE, T.; NAKAGAWA, I. Streptococcus pyogenes Genomics. In: FERRETTI, J. J.; STEVENS, D. L.; FISCHETTI, V. A. (Eds.). . **Streptococcus pyogenes: Basic Biology to Clinical Manifestations**. [s.l.] University of Oklahoma Health Sciences Center, 2016.

MASSELL, B. F.; HONIKMAN, L. H.; AMEZCUA, J. Rheumatic Fever Following Streptococcal Vaccination. **JAMA**, v. 207, n. 6, p. 1115, 10 fev. 1969.

MCARTHUR, J. D.; WALKER, M. J. Domains of group A streptococcal M protein that confer resistance to phagocytosis, opsonization and protection: implications for vaccine development. **Molecular Microbiology**, v. 59, n. 1, p. 1–4, 1 jan. 2006.

MCGREGOR, K. F. et al. Multilocus Sequence Typing of Streptococcus pyogenes Representing Most Known emm Types and Distinctions among Subpopulation Genetic Structures. **Journal of Bacteriology**, v. 186, n. 13, p. 4285–94, 1 jul. 2004.

MCMILLAN, D. J. et al. Updated model of group A Streptococcus M proteins based on a comprehensive worldwide study. **Clinical microbiology and infection**, v. 19, n. 5, p. E222--E229, maio 2013.

MCNAMARA, C. et al. **Coiled-Coil Irregularities and Instabilities in Group A Streptococcus M1 Are Required for VirulenceScience**. [s.l: s.n.].

MEISAL, R. et al. Molecular characteristics of pharyngeal and invasive emm3 Streptococcus pyogenes strains from Norway, 1988–2003. **European Journal of Clinical Microbiology & Infectious Diseases**, v. 29, n. 1, p. 31–43, 6 jan. 2010.

MEYGRET, A. et al. Genome Sequence of the Uncommon Streptococcus pyogenes M/emm66 Strain STAB13021, Isolated from Clonal Clustered Cases in French Brittany. **Genome announcements**, v. 4, n. 4, p. 1–2, 2016.

MURRAY, P.; ROSENTHAL, K. S.; PFALLER, M. A. **Microbiologia Médica**. 7. ed. [s.l.] Elsevier Brasil, 2014.

NAKAGAWA, I. et al. Genome sequence of an M3 strain of streptococcus pyogenes reveals a large-scale genomic rearrangement in invasive strains and new insights into phage evolution. **Genome Research**, v. 13, n. 6 A, p. 1042–1055, 2003.

NASSER, W. et al. Evolutionary pathway to increased virulence and epidemic group A Streptococcus disease derived from 3,615 genome sequences. **Proceedings of the National Academy of Sciences of the United States of America**, v. 111, n. 17, p. E1768-76, 29 abr. 2014.

NISHIFUJI, K. et al. Removal of amino-terminal extracellular domains of desmoglein 1 by staphylococcal exfoliative toxin is sufficient to initiate epidermal blister formation. **Journal of Dermatological Science**, v. 59, n. 3, p. 184–191, 1 set. 2010.

OLIVEIRA, D. M. P. DE et al. Blood Group Antigen Recognition via the Group A Streptococcal M Protein Mediates Host Colonization. **mBio**, v. 8, n. 1, p. e02237--16, 2017.

OLSEN, R. J.; MUSSER, J. M. Molecular Pathogenesis of Necrotizing Fasciitis. **Annual Review of Pathology: Mechanisms of Disease**, v. 5, n. 1, p. 1–31, 15 jan. 2010.

PANCHAUD, A. et al. M-protein and other intrinsic virulence factors of Streptococcus pyogenes are encoded on an ancient pathogenicity island. **BMC Genomics**, v. 10, p. 198, 2009.

PHILLIPS, M. H. History of the Prevention of Puerperal Fever. **British medical journal**, v. 1, n. 4017, p. 1–7, 1938.

RASMUSSEN, M.; EDÉN, A.; BJORCK, L. SclA, a novel collagen-like surface protein of Streptococcus pyogenes. **Infection and immunity**, v. 68, n. 11, nov. 2000.

REID, S. D. et al. Multilocus analysis of extracellular putative virulence proteins made by group A Streptococcus: Population genetics, human serologic response, and gene transcription. **PNAS**, v. 98, 2001.

RIBARDO, D. A.; LAMBERT, T. J.; MCIVER, K. S. Role of Streptococcus pyogenes two-component response regulators in the temporal control of Mga and the Mga-regulated virulence gene emm. **Infection and Immunity**, v. 72, n. 6, p. 3668–73, 2004.

ROKACH, L.; MAIMON, O. **Data Mining with Decision Trees - Theory and Applications**. [s.l.] World Scientific, 2007. v. 69

SANDERSON-SMITH, M. et al. A systematic and functional classification of Streptococcus

pyogenes that serves as a new tool for molecular typing and vaccine development. **Journal of Infectious Diseases**, v. 210, n. 8, p. 1325–1338, 2014.

SANTIAGO, C.; PEREIRA, V.; DIGIAMPIETRI, L. Homology Detection Using Multilayer Maximum Clustering Coefficient. **Journal of Computational Biology**, v. 25, n. 12, p. 1328–1338, 13 ago. 2018.

SANYAHUMBI, A. S. et al. Global Disease Burden of Group A Streptococcus. In: FERRETTI, J. J.; STEVENS, D. L.; FISCHETTI, V. A. (Eds.). . **Streptococcus pyogenes: Basic Biology to Clinical Manifestations**. [s.l.] University of Oklahoma Health Sciences Center, 2016.

SITKIEWICZ, I. et al. Lateral gene transfer of streptococcal ICE element RD2 (region of difference 2) encoding secreted proteins. **BMC Microbiology**, v. 11, n. 1, p. 65, 1 abr. 2011.

SKRZYPEK, M. S. et al. New tools at the Candida Genome Database: biochemical pathways and full-text literature search. **Nucleic Acids Research**, v. 38, n. suppl_1, p. D428–D432, 1 jan. 2010.

SMEESTERS, P. R. et al. Differences between Belgian and Brazilian Group A Streptococcus Epidemiologic Landscape. **PLoS ONE**, v. 1, n. 1, p. e10, 20 dez. 2006.

SORIANO, N. et al. Complete Genome Sequence of Streptococcus pyogenes M/emm44 Strain STAB901, Isolated in a Clonal Outbreak in French Brittany. **Genome announcements**, v. 2, n. 6, p. e01174-14, 20 nov. 2014.

SRISKANDAN, S.; FAULKNER, L.; HOPKINS, P. Streptococcus pyogenes: Insight into the function of the streptococcal superantigens. **The International Journal of Biochemistry & Cell Biology**, v. 39, n. 1, p. 12–19, 1 jan. 2007.

STEER, A. C. et al. Group A streptococcal vaccines: facts versus fantasy. **Current Opinion in Infectious Diseases**, v. 22, n. 6, p. 544–552, 2009a.

STEER, A. C. et al. Global emm type distribution of group A streptococci: systematic review and implications for vaccine development. **The Lancet Infectious Diseases**, v. 9, n. 10, p. 611–616, 1 out. 2009b.

STEER, A. C. et al. Status of research and development of vaccines for Streptococcus pyogenes. **Vaccine**, v. 34, n. 26, p. 2953–2958, 3 jun. 2016.

STOLLERMAN, G. H. Changing Streptococci and Prospects For The Global Eradication of Rheumatic Fever. **Perspectives in Biology and Medicine**, v. 40, n. 2, p. 165–189, 1997.

SUMBY, P. et al. Evolutionary Origin and Emergence of a Highly Successful Clone of Serotype M1 Group A *Streptococcus* Involved Multiple Horizontal Gene Transfer Events. **The Journal of Infectious Diseases**, v. 192, n. 5, p. 771–782, 1 set. 2005.

SUMBY, P. et al. Genome-Wide Analysis of Group A Streptococci Reveals a Mutation That Modulates Global Phenotype and Disease Specificity. **PLoS Pathogens**, v. 2, n. 1, p. 41–49, 2006.

TARTOF, S. Y. et al. **Factors associated with Group A Streptococcus emm type diversification in a large urban setting in Brazil: a cross-sectional studyBMC Infectious Diseases**. [s.l: s.n.].

TATUSOVA, T. et al. NCBI prokaryotic genome annotation pipeline. **Nucleic acids research**, v. 44, n. 14, p. 6614–24, 2016.

TEIXEIRA, L. M. et al. Genetic and Phenotypic Features of Streptococcus pyogenes Strains Isolated in Brazil That Harbor New emm Sequences. **Journal of clinical microbiology**, v. 39, n. 9, p. 3290–3295, 2001.

TETTELIN, H. et al. Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: Implications for the microbial "pan-genome". **Proceedings of the National Academy of Sciences**, v. 102, n. 39, p. 13950–13955, 27 set. 2005.

TSOI, S. K. et al. Correlates of Protection for M Protein-Based Vaccines against Group A Streptococcus. **Journal of immunology research**, v. 2015, p. 167089, 2015.

TURNER, C. E. et al. Superantigenic Activity of emm3 Streptococcus pyogenes Is Abrogated by a Conserved, Naturally Occurring smeZ Mutation. **PLoS ONE**, v. 7, n. 10, p. e46376, 2012.

VEGA, L. A. et al. The Transcriptional Regulator CpsY Is Important for Innate Immune Evasion in Streptococcus pyogenes. **Infection and immunity**, v. 85, n. 3, 2017.

VEGA, L. A.; MALKE, H.; MCIVER, K. S. **Virulence-Related Transcriptional Regulators of Streptococcus pyogenes**. [s.l.] University of Oklahoma Health Sciences Center, 2016.

VEKEMANS, J. et al. The Path to Group A Streptococcus Vaccines: World Health Organization Research and Development Technology Roadmap and Preferred Product Characteristics. **Clinical Infectious Diseases**, v. ciy1143, p. 1–7, 8 jan. 2019.

VIRTANEVA, K. et al. Longitudinal analysis of the group A Streptococcus transcriptome in experimental pharyngitis in cynomolgus macaques. **Proceedings of the National Academy of Sciences of the United States of America**, v. 102, n. 25, p. 9014–9, 21 jun. 2005.

VOGEL, C. et al. Structure, function and evolution of multidomain proteins. **Current Opinion in Structural Biology**, v. 14, n. 2, p. 208–216, 1 abr. 2004.

WATTAM, A. R. et al. PATRIC, the bacterial bioinformatics database and analysis resource. **Nucleic acids research**, v. 42, n. Database issue, p. D581--D591, jan. 2014.

WATTAM, A. R. et al. Improvements to PATRIC, the all-bacterial Bioinformatics Database and Analysis Resource Center. **Nucleic acids research**, v. 45, n. D1, p. D535–D542, 2017.

WINSOR, G. L. et al. Enhanced annotations and features for comparing thousands of Pseudomonas genomes in the Pseudomonas genome database. **Nucleic acids research**, v. 44, n. D1, p. D646-53, 4 jan. 2016.

WOLF, Y. I. et al. Genome trees constructed using five different approaches suggest new major bacterial clades. **BMC Evolutionary Biology**, v. 1, n. 1, p. 8, 23 out. 2001.

ZHENG, P.-X. et al. Complete Genome Sequence of emm1 Streptococcus pyogenes A20, a Strain with an Intact Two-Component System, CovRS, Isolated from a Patient with Necrotizing Fasciitis. **Genome announcements**, v. 1, n. 1, jan. 2013.

ZHOU, Y. et al. PHAST: A Fast Phage Search Tool. **Nucleic Acids Research**, v. 39, n. suppl, p. W347–W352, 1 jul. 2011.

Apêndice A - Características dos genomas completos de *S. pyogenes* selecionados

Os genótipos das cepas estão descritos na coluna M. A coluna perfil indica a invasividade da cepa, sendo invasivo (I), não invasivo (N) ou desconhecido (D). O tamanho é dado em pares de bases e o número de CDS foi indicado após anotação com a ferramenta *RASTtk*. Finalmente a data de incorporação da versão do genoma utilizada neste projeto no banco do *NCBI* é dada na última coluna.

Isolado	M	Perfil	Acesso NCBI	Tamanho (pb)	GC%	CDS	Data
5448	1	I	CP008776.1	1.829.516	38,5	1826	2014
A20	1	I	CP003901.1	1.837.281	38,5	1833	2012
AP1	1	I	CP007537.1	1.908.294	38,5	1963	2014
M1-476	1	I	AP012491.2	1.821.128	38,5	1830	2012
MGAS5005	1	I	NC_007297.2	1.838.554	38,5	1832	2004
MTB313	1	I	AP014572.1	1.745.332	38,5	1846	2014
MTB314	1	I	AP014585.1	1.744.827	38,5	1731	2014
NCTC8198	1	N	LN831034.1	1.914.862	38,5	1967	2015
SF370	1	D	AE004092.2	1.852.433	38,5	1849	2014
HKU488	1	N	CP012045.1	1.943.415	38,5	1959	2015
NGAS638	101	I	CP010450.1	1.791.401	38,6	1767	2015
NGAS322	114	I	CP010449.1	1.950.469	38,3	1952	2015
HKU16	12	N	AFRY01000001.1	1.908.100	38,5	1886	2012
HKU360	12	N	CP009612.1	1.944.537	38,5	1951	2014
MGAS2096	12	N	NC_008023.1	1.860.355	38,7	1886	2006
MGAS9429	12	N	CP000259.1	1.836.467	38,5	1823	2006
HSC5	14	I	CP006366.1	1.818.351	38,5	1818	2013
MGAS8232	18	N	NC_003485.1	1.895.017	38,5	1953	2002
MGAS10270	2	N	NC_008022.1	1.928.252	38,4	1963	2006
M23ND	23	I	CP008695.1	1.846.477	38,6	1879	2014
MGAS6180	28	I	NC_007296	1.897.573	38,4	1897	2005
STAB10015	28	N	CP011068.1	1.950.454	38,2	1945	2015
M28PF1	28	I	CP011535.2	1.896.976	38,4	1895	2016
MEW123	28	N	CP014139.1	1.878.699	38,3	1889	2016
STAB09014	28	N	CP011069.1	1.862.487	38,4	1842	2015
MGAS315	3	I	NC_004070.1	1.900.521	38,6	1948	2002
SSI-1	3	I	BA000034.2	1.894.275	38,6	1951	2002
STAB902	3	I	NZ_CP007041.1	1.892.124	38,5	1947	2013
M3-b	3	I	AP014596.1	1.893.821	38,5	1954	2014
MGAS10750	4	N	NC_008024.1	1.937.111	38,3	1969	2006
MEW427	4	N	CP014138.1	1.814.455	38,5	1861	2016
STAB901	44	I	CP007024.1	1.795.609	38,5	2057	2013
1E1	44	D	CP007241.1	1.796.152	38,5	1755	2014
NZ131	49	N	CP000829.1	1.815.785	38,6	1812	2007
MANFREDO	5	N	NC_009332.1	1.841.271	38,6	1882	2006

ALAB49	53	N	NC_017596.1	1.827.308	38,6	1832	2011
AP53	53	N	NZ_CP013672.1	1.860.554	38,6	1885	2015
MGAS15252	59	I	NC_017040.1	1.750.832	38,5	1700	2011
MGAS1882	59	N	CP003121.1	1.781.029	38,5	1745	2011
D471	6	N	NZ_CP011415.1	1.811.968	38,6	1802	2015
JRS4	6	N	CP011414.1	1.811.968	38,6	1803	2015
MGAS10394	6	N	CP000003.1	1.899.877	38,7	1910	2004
STAB13021	66	I	CP014278.2	1.811.936	38,5	1795	2016
NS53	71	N	CP015238.1	1.765.123	38,4	1732	2016
ATCC 19615	80	N	NZ_CP008926.1	1.844.804	38,5	1875	2014
NGAS596	82	I	CP007561.1	1.791.306	38,5	1749	2014
STAB1101	83	I	NZ_CP007240.1	1.709.790	38,6	1640	2014
NGAS327	83	I	CP007562.1	1.702.054	38,6	1634	2014
STAB1102	83	I	CP007023.1	1.709.442	38,6	1639	2013
NGAS743	87	I	CP007560.1	1.915.554	38,5	1930	2014
MGAS11027	89	N	NZ_CP013838.1	1.786.874	38,6	1754	2016
MGAS27061	89	I	NZ_CP013840.1	1.741.348	38,5	1683	2016
H293	89	I	NZ_HG316453.1	1.726.248	38,6	1661	2013
JMUB1235	89	I	AP017629.1	1.741.982	38,5	1686	2016
MGAS23530	89	N	CP013839.1	1.709.394	38,5	1647	2016

Fonte: Suzane de Andrade Barboza, 2019

Apêndice B – Relação de fatores das ferramentas e parâmetros utilizados

1 Caracterização dos genomas

1.1 Obtenção dos genomas

Os genomas incluídos neste estudo foram obtidos diretamente da plataforma do *National Center for Biotechnology Information* (NCBI) usando a seguinte especificação de busca: "(*streptococcus pyogenes) AND "Streptococcus pyogenes"[porgn:__txid1314] complete genome"*.

1.2 Ferramenta de anotação

A fim de padronizar a anotação de todos os genomas, estes foram reanotados pela ferramenta RASTk (WATTAM et al., 2014) disponível na plataforma PATRIC (WATTAM et al., 2017).

1.3 Identificação de profagos inteiros ou parciais

A identificação de fagos completos ou parciais nos genomas foi conduzida com a ferramenta PHASTER (ARNDT et al., 2016; ZHOU et al., 2011).

2 Identificação de genes homólogos

2.1 Identificação de ortólogos

A fim de identificar os genes ortólogos para a criação de clusters (ou famílias gênicas), foi conduzido o alinhamento local de todas as CDSs contra todas as CDSs, utilizando percentagem mínima de alinhamento = 41% e *e-value* máximo = 10^{-10}. Cada família teve sua filogenia construída utilizando o Clustal Omega e o FastTree (com seus parâmetros padrões). Os ramos das filogenias que foram maiores que 0,4 era removidos, causando dessa forma a subdivisão das famílias.

2.2 Determinação do melhor e-value/coeficiente de clusterização

O *e-value* foi determinado automaticamente, buscando maximizar o coeficiente de clusterização da rede de genes homólogos, utilizando a ferramenta Multilayer Clustering do arcabouço GTACG. Foram obtidos vários valores diferentes para o *e-value*, cada um resultando em restrições mais rigorosas para a formação das famílias. São eles: 10^{-14}, 10^{-27}, 10^{-44}, 10^{-46}, 10^{-47}, 10^{-52} e 10^{-59}.

2.3 Identificação de proteínas multi-domínio

Por último, uma etapa adicional foi realizada a fim de identificar proteínas multi-domínio, levando em conta a assimetria do alinhamento disposto no grafo de cada família identificada, utilizando a ferramenta Multilayer Clustering do arcabouço GTACG. Para ser considerada multi-domínio a proteína precisa ter um coeficiente de clusterização menor que a média de sua homólogas, além disso a diferença entre a proteína multi-domínio para as de domínio único deve ter mais do que 30% de diferença no alinhamento e a diferença precisa de no mínimo 100 aminoácidos.

3 Comparação de genomas completos

3.1 Cladograma de presença/ausência

A primeira abordagem para a construção do cladograma utilizou uma matriz baseada na presença (valor 1) ou ausência (valor 0), utilizando a distância Euclidiana. Essas distâncias foram posteriormente utilizadas para a criação de um cladograma. por meio do método *neighbor-join* presente na biblioteca Phylip (FELSENSTEIN, 2005).

3.2 Criação da supertree

A segunda abordagem produziu uma *supertree* (CREEVEY; MCINERNEY, 2005), baseada nas filogenias de cada família de genes, resultando em uma árvore que reflete simplificadamente todas as relações identificadas. Foi utilizada a ferramenta Clann, utilizando

os parâmetros “Samples=1, N-Reps=1 e nbis = 1”, e o método de inferência *Quartet Fit* com *Neighbor Interchange* (CREEVEY; MCINERNEY, 2005).

4 Rede gênica

4.1.1 Criação da rede gênica de clusterização

Com o objetivo de melhor compreender as relações intergênicas, foi criada uma rede gráfica em que cada família de genes homólogos é um *cluster* (ou componente conexo) composto de nós que representam os genes de cada genoma. Foi para isso utilizada a ferramenta Multilaeyer Clustering do arcabouço GTACG, com percentagem de alinhamento mínimo = 41% e *e-value* máximo = 10^{-10}.

As arestas representam um alinhamento entre os genes conectados que satisfaz os parâmetros definidos. Posteriormente, um algoritmo do tipo *force-directed* foi desenvolvido e aplicado a fim de aproximar ou separar graficamente os genes de acordo com suas arestas.

5 Portal

5.1. Criação do portal

Todas as informações apresentadas neste estudo foram disponibilizadas através da criação da *Streptococcus pyogenes Database* (SPD), desenvolvida com o uso do arcabouço GTACG (*Gene Tag Assessment by Comparative Genomics*), com o objetivo de uniformizar e centralizar os dados genômicos de *S. pyogenes*, acessível pelo *link* http://143.107.58.250/reportStrep2/.

5.2. Criação de um valor de dissimilaridade

O segundo nível de informações do portal dispõe os alinhamentos das sequências que compõe as famílias de genes. Para esses alinhamentos foi incluído um valor de dissimilaridade, que mede a diversidade das bases em um alinhamento a fim de determinar pares de bases mais correlatos a determinados grupos de organismos, como SNPs, por

exemplo. Este valor foi criado na ferramenta GTACG (*Gene Tag Assessment by Comparative Genomics*).

5.3. Determinação dos grupos em sub-árvores

Por fim, o último nível compreende as representações filogenéticas, a fim de determinar o quão isolados são os grupos em sub-árvores exclusivas, também criado na ferramenta GTACG (*Gene Tag Assessment by Comparative Genomics*), utilizando a métrica MIST (*Most Isolated Subtree*).

6 Determinação de genes exclusivos

6.1 Determinação de genes exclusivos

A identificação de genes exclusivamente relacionados a invasividade, genótipo e associação a doenças foi realizada à partir do portal desenvolvido neste projeto, selecionando por meio do primeiro nível do portal em quais genomas as famílias gênicas deveriam estar obrigatoriamente presentes/ausentes de acordo com as características a serem estudadas.

6.2 Determinação de genes-alvo para vacinas

Durante a análise de genes, foram encontrados 16 genes bem conservados que estão presentes em todos os isolados com apenas uma cópia no genoma. Esses genes então foram alinhados contra o genoma humano, a fim de excluir genes que possuíssem alguma similaridade, diminuindo assim a possibilidade de desenvolvimento de uma doença autoimune induzida pela vacinação. Para o alinhamento, foi utilizada a ferramenta BLAST com os *default* do BLAST e *e-value* máximo = 0,00001, contra o genoma humano de referência GRCh38 (GenBank GCA_000001405.27).

Apêndice C – Relação de fatores de virulência de *S. pyogenes*

Fator de virulência	Função	Produto (Anotação)
GRAB	Adesina	Protein G-related alpha 2 macroglobulin-binding protein (GRAB)
Collagen like proteins	Adesina	Collagen-like surface protein Tail-specific protease
Collagen like proteins	Adesina	Collagen-like surface protein
Collagen like proteins	Adesina	Collagen adhesion protein
Enn M-like protein (*enn)*	Adesina	*not analyzed*
Fibronectin-binding (*fbp54*)	Adesina	*not analyzed*
Fibronectin-binding (*fbpA*)	Adesina	*not analyzed*
Fibronectin-binding F2-like protein (prtF2)	Adesina	*not analyzed*
Fibronecting binding protein (*sfbX*)	Adesina	*not analyzed*
Fibronecting binding protein 1 (sbf1*)*	Adesina	*not analyzed*
GAPDH (*plr/gapA*)	Adesina	*not analyzed*
Internalin-like protein	Adesina	Internalin, putative
Laminin-binding protein	Adesina	Laminin-binding protein
M protein	Adesina	Antiphagocytic M protein
Mrp M-like protein (*mrp)*	Adesina	*not analyzed*
R28 protein (*spr28*)	Adesina	*not analyzed*
R6 surface protein	Adesina	*not analyzed*
Serum Opacity Factor SOF (*sfbII/sof*)	Adesina	*not analyzed*
Substrate-binding lipoprotein AdcA	Adesina	Zinc ABC transporter, substrate-binding lipoprotein AdcA
Immunogenic secreted protein	Antígeno	Immunogenic secreted protein
Immunogenic surface protein	Antígeno	Group B streptococcal surface immunogenic protein
Immunoreactive protein	Antígeno	Immunoreactive protein Se23.5 (Fragment)
Myosin antigen	Antígeno	67kDa Myosin-cross-reactive streptococcal antigen
Salivaricin-related Proteins	Bacteriocina	Lantibiotic salivaricin A
Salivaricin-related Proteins	Bacteriocina	Putative salivaricin A ABC transporter (ATP-binding protein)
Salivaricin-related Proteins	Bacteriocina	Salivaricin A modification enzyme
Mitogenic Factors	Dnase	Streptodornase B; Mitogenic factor 1
Mitogenic Factors	Dnase	Streptococcal extracellular nuclease 2;

		Mitogenic factor 2
Mitogenic Factors	Dnase	Streptococcal extracellular nuclease 3; Mitogenic factor 3
Mitogenic Factors (*mf4/spd4*)	Dnase	*not analyzed*
Streptodornase	Dnase	Streptodornase D
Streptodornase (*sda)*	Dnase	*not analyzed*
Capsule	Evasão imune	Capsule biosynthesis protein capA
Capsule (*hasA*)	Evasão imune	*not analyzed*
Capsule (*hasB*)	Evasão imune	*not analyzed*
Capsule (*hasC*)	Evasão imune	*not analyzed*
SIC	Evasão imune	Streptococcal inhibitor of complement (SIC)
ADP-ribosyltransferase (*spyA*)	Exoenzima	*not analyzed*
EndoS	Exoenzima	Secreted Endo-beta-N-acetylglucosaminidase (EndoS)
Hyaluronidase	Exoenzima	Phage hyaluronidase
Streptococcal phopholipase	Exoenzima	Streptococcal phospholipase A2; _Toximoron (Other)
CAMP	Hemolisina	CAMP
Hemolysin	Hemolisina	Hemolysins and related proteins containing CBS domains
Hemolysin	Hemolisina	Hemolysin hyl III
Streptolysin O (*slo)*	Hemolisina	*not analyzed*
Streptolysin S (Operon)	Hemolisina	Streptolysin S precursor (SagA)
Streptolysin S (Operon)	Hemolisina	Secreted antigen GbpB/SagA/PcsB, putative peptidoglycan hydrolase
Streptolysin S (Operon)	Hemolisina	Streptolysin S biosynthesis protein B (SagB)
Streptolysin S (Operon)	Hemolisina	Streptolysin S biosynthesis protein C (SagC)
Streptolysin S (Operon)	Hemolisina	Streptolysin S biosynthesis protein D (SagD)
Streptolysin S (Operon)	Hemolisina	Streptolysin S biosynthesis protein E (SagE)
Streptolysin S (Operon)	Hemolisina	Streptolysin S biosynthesis protein F (SagF)
Streptolysin S (Operon)	Hemolisina	Streptolysin S export ATP-binding protein (SagG)
Streptolysin S (Operon)	Hemolisina	Streptolysin S export transmembrane permease (SagH)
Streptolysin S (Operon)	Hemolisina	Streptolysin S export transmembrane permease (SagI)
C3-degrading proteinase	Protease	C3-degrading proteinase

C3-degrading proteinase	Protease	C3 family ADP-ribosyltransferase (EC 2.4.2.-)
C5a peptidase	Protease	C5a peptidase
Exfoliating toxin-like	Protease	Exfoliative toxin A
IdeS	Protease	IdeS/Mac/Sib38 IgG-degrading protease
Serine endopeptidase	Protease	Serine endopeptidase
Serine protease	Protease	Serine protease DegP/HtrA
Serine protease	Protease	Rhomboid family serine protease
SpeB/cysteine proteinase	Protease	SpeB/Cysteine protease
Streptokinase A	Protease	Streptokinase
comR	Regulador	*not analyzed*
COV R/S	Regulador	Response regulator CsrR
COV R/S	Regulador	Two-component response regulator SA14-24
COV R/S	Regulador	Transmembrane histidine kinase CsrS
MGA	Regulador	M protein trans-acting positive regulator (Mga)
MGA	Regulador	Mga-associated protein
MutR	Regulador	Positive transcriptional regulator, MutR family
PRSA	Regulador	Foldase protein prsA 1 precursor
RALP	Regulador	RofA-like protein (RALP)
rgg3	Regulador	*not analyzed*
ROF A	Regulador	Transcriptional regulator, RofA
RopA	Regulador	Proteinase maturation protein (RopA)
RopB	Regulador	RopB; Rgg-like transcription regulator
SALR (*salR)*	Regulador	*not analyzed*
Conjugal transfer protein (Ctp)	Segregação do DNA	Conjugal transfer protein, putative
Mitogenic exotoxin Z	Superantígeno	Streptococcal mitogenic exotoxin Z (SmeZ)
Streptococcal pyrogenic exotoxin	Superantígeno	Streptococcal pyrogenic exotoxin A (SpeA);
Streptococcal pyrogenic exotoxin	Superantígeno	Exotoxin, phage associated
Streptococcal pyrogenic exotoxin	Superantígeno	Streptococcal pyrogenic exotoxin C (SpeC)
Streptococcal pyrogenic exotoxin	Superantígeno	Streptococcal pyrogenic exotoxin G (SpeG)
Streptococcal pyrogenic exotoxin	Superantígeno	Streptococcal pyrogenic exotoxin H (SpeH)
Streptococcal pyrogenic	Superantígeno	Streptococcal pyrogenic exotoxin J

exotoxin		(SpeJ)
Streptococcal pyrogenic exotoxin	Superantígeno	Streptococcal pyrogenic exotoxin K (SpeK)
Streptococcal pyrogenic exotoxin (*speL*)	Superantígeno	*not analyzed*
Streptococcal pyrogenic exotoxin (*speM*)	Superantígeno	*not analyzed*
Streptococcal superantigen	Superantígeno	Immunodominant antigen A
Metal binding protein SlOC (*mtsA*)	Transporte de metal	*not analyzed*

Fonte: Suzane de Andrade Barboza, 2019

Apêndice D – Relação dos atributos e famílias utilizados no conjunto de classificadores para o genótipo *emm*

Attribute	Produto (Anotação)
5_Exfoliating toxin-like	(fig|1314.358.peg353, 361.peg.1576, 365.peg.1301, 367.peg.1177/1442, 368.peg.1247/1510, 373.peg.333/529, 374.peg.1353, 380.peg.1181/1446, 381.peg.1151/1471, 382.peg.1249, 383.peg.1727, 387.peg.1263, 388.peg.289, 389.peg.567, 392.peg.1139/1405, 393.peg.1422, 396.peg.1396, 398.peg.598, 401.peg.553, 408.peg.265, 409.peg.1473, 410.peg.1578, 413.peg.1345/1578, 414.peg.1598, 415.peg.1326)
6_Exfoliating toxin-like	(fig|1314.397.peg.989)
8_Exfoliating toxin-like	(fig|1314.386.peg.858)
12_IdeS	(fig|1314.357.peg.1156, 362.peg.785, 370.peg.362, 400.peg.312)
24_GRAB	(fig|1314.400.peg.1109)
59_Hyaluronidase	(fig|1314.400.peg.788)
62_Mitogenic factor 1	(fig|1314.403.peg.1435)
64_Mitogenic factor 2	(fig|1314.406.peg.1561)
65_Mitogenic factor 3	(fig|1314.406.peg.1561)
67_SpeA	(fig|1314.414.peg.9920)
68_Phage exotoxin	(fig|1314.414.peg.992)
72_SpeH	(fig|1314.357.peg.1033)
75_SpeJ	(fig|1314.393.peg.1103)
76_SpeJ	(fig|1314.362.peg.784)
77_SpeK	(fig|1314.385.peg.1227)
79_SmeZ	(fig|1314.408.peg.1663)
102_MutR	(fig|1314.374.peg.1789)
104_MutR	(fig|1314.398.peg.1736)
105_MutR	(fig|1314.361.peg.1724, 365.peg.1739, 367.peg.1575, 370.peg.1928, 373.peg.1702, 374.peg.1788,380.peg.1578, 381.peg.1549, 382.peg.1684, 383.peg.1851, 384.peg.1640, 386.peg.1755,389.peg.1755, 392.peg.1536, 393.peg.1816, 401.1738, 408.peg.1719, 409.peg.1621, 410.peg.1734, 411.peg.1860, 412.peg.1856, 413.peg.1726, 414.peg.1723, 415.peg.1721)
109_Mga	(fig|1314.400.peg.160)
118_Ralp	(fig|1314.400.peg.101)
125_Conjugal Protein	(fig|1314.401.peg.544)

Fonte: Suzane de Andrade Barboza, 2019

Apêndice E – Relação dos atributos e famílias utilizados no conjunto de classificadores para invasividade

Attribute	Produto (Anotação)
7_Exfoliating toxin-like	(fig|1314.361.peg.373)
8_Exfoliating toxin-like	(fig|1314.386.peg.858)
12_IdeS	(fig|1314.357.peg.1156, 362.peg.785, 370.peg.362, 400.peg.312)
28_Collagen like proteins	(fig|1314.371.peg.1778)
32_Internalin-like protein	(fig|1314.382.peg.1283)
38_SagB	(fig|1314.370.peg.647)
53_Hemolysin	(fig|1314.369.peg.914)
63_Mitogenic factor 1	(fig|1314.363.peg.1757)
64_Mitogenic factor 2	(fig|1314.406.peg.1561)
69_SpeC	(fig|1314.393.peg.1103)
70_SpeG	(fig|1314.400.peg.178)
71_SpeG	(fig|1314.405.peg.163, 406.peg.215, 387.peg.1210)
74_SpeI	(fig|1314.405.peg.685, 406.peg.739, 356.peg.686)
80_SmeZ	(fig|1314.358.peg.1662)
88_Isp	(fig|1314.390.peg.1706)

Fonte: Suzane de Andrade Barboza, 2019

Apêndice F – Relação dos atributos e famílias utilizados no conjunto de classificadores para doença

Attribute	Produto (Anotação)
5_Exfoliating toxin-like	(fig|1314.358.peg353, 361.peg.1576, 365.peg.1301, 367.peg.1177/1442, 368.peg.1247/1510, 373.peg.333/529, 374.peg.1353, 380.peg.1181/1446, 388.peg.289, 389.peg.567, 392.peg.1139/1405, 393.peg.1422, 396.peg.1396, 398.peg.598, 401.peg.553, 408.peg.265, 409.peg.1473, 410.peg.1578, 413.peg.1345/1578, 414.peg.1598, 415.peg.1326)
17_C5a peptidase	(fig|1314.370.peg.1870)
24_GRAB	(fig|1314.400.peg.1109)
56_Streptococcal phospholipase	(fig|1314.386.peg.1010)
57_Hyaluronidase	(fig|1314.400.peg.788)
58_Hyaluronidase	(fig|1314.400.peg.788)
59_Hyaluronidase	(fig|1314.400.peg.788)
62_Mitogenic factor 1	(fig|1314.403.peg.1435)
64_Mitogenic factor 2	(fig|1314.406.peg.1561)
65_Mitogenic factor 3	(fig|1314.406.peg.1561)
67_SpeA	(fig|1314.414.peg.9920)
68_Phage exotoxin	(fig|1314.414.peg.992)
69_SpeC	(fig|1314.393.peg.1103)
76_SpeJ	(fig|1314.362.peg.784)
79_SmeZ	(fig|1314.408.peg.1663)
102_MutR	(fig|1314.374.peg.1789)
103_MutR	(fig|1314.402.peg.1770)
104_MutR	(fig|1314.398.peg.1736)
105_MutR	(fig|1314.361.peg.1724, 365.peg.1739, 367.peg.1575, 370.peg.1928, 373.peg.1702, 374.peg.1788,380.peg.1578, 381.peg.1549, 382.peg.1684, 411.peg.1860, 412.peg.1856, 413.peg.1726, 414.peg.1723, 415.peg.1721)
109_Mga	(fig|1314.400.peg.160)
118_Ralp	(fig|1314.400.peg.101)
123_Salivaricinivaricin A	(fig|1314.373.peg.1578)
125_Conjugal Protein	(fig|1314.401.peg.544)

Fonte: Suzane de Andrade Barboza, 2019