Tahila Andrighetti

Influência do microbioma intestinal no desenvolvimento de doença de Crohn

Tese apresentada no Instituto de Biociências da Universidade Estadual Paulista "Júlio de Mesquita Filho" para a obtenção do título de Doutora em Ciências Biológicas (Genética).

Orientador: Prof. Ney Lemke

Botucatu

Julho de 2019

FICHA CATALOGRÁFICA ELABORADA PELA SEÇÃO TÉC. AQUIS. TRATAMENTO DA INFORM.
DIVISÃO TÉCNICA DE BIBLIOTECA E DOCUMENTAÇÃO - CÂMPUS DE BOTUCATU - UNESP
BIBLIOTECÁRIA RESPONSÁVEL: ROSANGELA APARECIDA LOBO-CRB 8/7500

Andrighetti, Tahila.
Influência do microbioma intestinal no desenvolvimento de doença de Crohn / Tahila Andrighetti. - Botucatu, 2019

Tese (doutorado) - Universidade Estadual Paulista "Júlio de Mesquita Filho", Instituto de Biociências de Botucatu
Orientador: Ney Lemke
Capes: 20202008

1. Crohn, Doença de. 2. Biologia de sistemas. 3. Bioinformática. 4. Autofagia. 5. Microbioma gastrointestinal. 6. Interação proteína-proteína.

Palavras-chave: autofagia; bioinformática; biologia de sistemas; doença de Crohn; interação microbioma-hospedeiro.

Agradecimentos

- Ao professor Ney Lemke pela orientação, suporte, conhecimento, por estar sempre disponível para ajudar e por me acompanhar durante esse processo com dedicação e paciência.
- Aos orientadores do Instituto Earlham, Tamas Korcsmaros e Padhmanand Sudhakar por terem me acolhido no seu grupo de pesquisa, me acompanhado, auxiliado durante todo o projeto e me transmitido conhecimento com humildade, motivação e positividade.
- Aos colegas de trabalho do LBBC Luis, Rafael e Rodrigo com quem pude discutir ciência, filosofia e aleatoriedades.
- Aos colegas do Instituto Earlham, que me acolheram em Norwich com muito carinho e com quem pude contar inúmeras vezes, tanto em relação ao projeto de pesquisa quanto na vida pessoal.
- À toda a minha família, aos meus pais Eloi e Jaqueline e à minha irmã Giovana que sempre me ofereceram suporte, compreensão e amor e que, mesmo estando longe, me davam a certeza de que não estou sozinha.
- Aos meus amigos, com quem sempre pude contar nos momentos difíceis e nos de descontração. Me ajudaram a tornar o processo mais leve.
- À todas as pessoas que me fizeram mal de alguma forma durante minha jornada em Botucatu. Me ajudaram a me conhecer cada vez mais, a me fortalecer, evoluir e também me mostraram os caminhos aos quais não tenho a intenção de seguir.
- À CNPq pelo apoio financeiro com a bolsa de doutorado e à CAPES pelo apoio financeiro com a bolsa de doutorado sanduíche.

"Eu não sei qual é o motivo dessa supervalorização da racionalidade. Os pássaros só são livres porque podem voar. A liberdade é, justamente, a incapacidade de se perceber as limitações."

Frida Kahlo

Resumo

A doença de Crohn é um subtipo de doença inflamatória intestinal caracterizada pela inflamação intestinal e disbiose da microbiota. A doença é causada uma manifestação atípica da resposta imune à presença de proteínas microbianas alteradas na interface da mucosa intestinal. Várias análises meta-ômicas têm sido executadas para caracterizar a microbiota intestinal em casos de doença de Crohn. Entretanto, devido à limitações das tecnologias e desvantagens inerentes aos métodos experimentais existentes, os mecanismos mediados pelas proteínas do microbioma alterado de doença de Crohn ainda não foram exploradas. Para analisar essa interação a nível molecular, executamos uma abordagem computacional que combina conjuntos de dados de metaproteômica publicados de um estudo clínico de pares de gêmeos com predições de interação entre proteínas microbianas e humanas. Essa predição baseia-se em características estruturais, inferências de regiões desordenadas, vias de sinalização e redes de interação para determinar as possíveis funções mediadas pelas proteínas microbianas. Como resultado, foi obtida uma rede de interação microbioma-hospedeiro que inicia sua sinalização com proteínas bacterianas que interagem com proteínas humanas de superfície celular. A expansão desse sinal foi modelada em direção ao núcleo da célula, onde observamos potenciais pontos de modulação de genes de autofagia: um dos processos celulares mais modificados em células de doença de Crohn em comparação às saudáveis. Esse modelo revelou diferenças em potencial entre as condições saudáveis e com doença de Crohn, onde podemos observar modulações de diversos processos biológicos relacionados à autofagia e ao desenvolvimento de doença de Crohn, como mitofagia, apoptose, diferenciação e proliferação celular.

Abstract

Crohn's disease (CD) is a subtype of inflammatory bowel disease (IBD) characterized by intestinal inflammation and microbiome dysbiosis. The disease is caused by an atypical manifestation of immune response to altered microbial proteins located in the intestinal mucosa. Various meta-omic based analyses have been carried out to profile and characterize the gut microbiota upon onset of CD. However, the molecular mechanisms mediated by the altered CD microbiome wasn't explored yet due to the technology limitations and disavantages of the experimental methods. In order to analyse this interactio in a molecular level, we performed a computational approach which uses metaproteomic datasets from a twin-pair CD cohort study and prediction of the interaction between human and bacterial proteins. This prediction relies on structural feature based interaction prediction between microbial and host receptor proteins, disordered region inferences, signalling pathways and interaction networks to determine the possible functions mediated by the microbial proteins. As a result, we obtained a host-microbiome interaction network which starts the molecular signal with bacterial proteins interacting with human receptor proteins interactions. The expansion of this signal was modeled in direction to the cellular nucleus, where reaches autophagy genes modulation spots, hence autophagy is one of the key cellular processes to CD development. This model revealed potential differences between CD and healthy conditions and that there are different cellular processes related to autophagy and CD being modulated, as mitophagy, apoptosis, cellular differentiation and cellular proliferation.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

BCR	Receptores de células B (*B cells receptors*)
EROs	Espécies reativas de oxigênio
ERNs	Espécies reativas de nitrogênio
GO	Ontologia de gene (*Gene ontology*)
GWAS	Estudos de associação ampla do genoma (*Genome wide association studies*)
IAP	Inibidoras de proteínas apoptóticas (*Inhibitor of apoptotic proteins*)
IBD	Doenças inflamatórias intestinais (*Inflammatory Bowel Diseases*)
IL	Interleucina
MHC	Complexo principal de histocompatibilidade (*Major histocompatibility complex*)
NK	*Natural Killer*
NLR	*Nucleotide-binding oligomerization domain protein-like receptors*
PAMPs	Padrões Moleculares Associados a Patógenos (*Pathogen-associated molecular pattern*)
PLC	Progenitor linfóide comum
PMC	Progenitor mielóide comum
PPI	Interações proteína-proteína (*Proteína-protein interactions*)
PRR	Receptores de reconhecimento de padrões (*Pattern Recognition Receptor*)
RU	Retocolite ulcerativa
SCFA	Ácidos graxos de cadeia curta (*Short Fatty Chain Acids*)

TCR	Receptores de células T (*T cells receptors*)
TLR	*Toll-like receptors*
Tc	Células T citotóxicas
Th	Células T auxiliares (*T helper*)
TNF	*Tumor necroses factor*
TRI	Interações transcricionais (*Transcriptional regulatory interactions*)
WGS	*Whole genome shotgun*

Sumário

1 Introdução

1.1 Doenças inflamatórias intestinais

Doença de Crohn (DC) e retocolite ulcerativa (RU) são doenças inflamatórias intestinais (IBDs, do inglês *Inflammatory Bowel Diseases*) caracterizadas pela inflamação crônica da mucosa do trato gastrointestinal, principalmente no íleo terminal, porção distal do intestino delgado e cólon intestinal (HENDERSON; STEVENS, 2012). Os principais sintomas manifestados pelos pacientes com IBD são diarréia, dor abdominal, sangramento gastrointestinal e perda de peso, apresentando períodos de relapso e remissão (ANBAZHAGAN et al., 2018). As IBDs não apresentam cura e os recursos terapêuticos da doença estão relacionados ao controle da inflamação, restauração dos déficits nutricionais e tratamento dos sintomas, como a diarréia. Esses tratamentos interferem na qualidade de vida e frequentemente requerem intervenção cirúrgica (MICHAIL; BULTRON; DEPAOLO, 2013; ANBAZHAGAN et al., 2018).

O desenvolvimento das IBDs é atribuído à inflamação crônica das células intestinais resultante da soma de fatores genéticos, imunológicos e ambientais (BAUMGART; SANDBORN, 2012). Nas primeiras décadas de pesquisa, os fatores genéticos foram considerados os principais responsáveis pelas IBD; no entanto eles só explicam uma pequena fração da hereditariedade da doença (MONDAL; KUGATHASAN, 2017), ressaltando a importância dos fatores extrínsecos em sua etiologia. Não obstante, dependendo do grau de predisposição genética caracterizada pela ocorrência de mutações e da profundeza dos fatores extrínsecos, a severidade das doenças evolui para uma forma extrema.

Fatores extrínsecos como hábitos alimentares, drogas, poluição, radiação, estresse, exposição a antibióticos e higiene modulam a funcionalidade do microbioma intestinal, influenciando na suscetibilidade do indivíduo de desenvolver alguma das IBDs e em sua intensidade (BAUMGART; SANDBORN, 2012; SHEEHAN; SHANAHAN, 2017). Alguns fatos corroboram que os protagonistas no desenvolvimento das IBDs são os micro-organismos: primeiramente, estudos que mostram que ratos livres de micro-organismos não desenvolvem nenhuma das IBDs

(STANGE; WEHKAMP, 2016). Além disso, há uma notória diferença entre a microbiota intestinal de pacientes saudáveis e de pacientes com com DC ou RU (PASCAL et al., 2017). A microbiota alterada desencadeia uma resposta imune agressiva, resultando na inflamação crônica. Deste modo, podemos considerar que a disbiose intestinal é o elo entre os elementos extrínsecos e a resposta imune do hospedeiro (MANICHANH et al., 2012).

1.2 Microbioma intestinal

Estima-se que o número de microrganismos presentes no corpo humano seja de aproximadamente 100 trilhões de células, dez vezes mais do que o número de células humanas presentes no corpo de um indivíduo (BELLA et al., 2013). Somente no intestino delgado, o número de células varía de 10^4 a 10^7 células por grama, enquanto no colo o número chega a 10^{12} células por grama (SEKIROV et al., 2010). A maior parte desses microrganismos do intestino são comensais e essenciais para garantir o funcionamento apropriado dos processos biológicos (AVIELLO; KNAUS, 2017).

Microrganismos comensais são aqueles que co-evoluíram com o hospedeiro e dificilmente são encontrados em vida livre. Como consequência dessa co-evolução, as bactérias comensais passaram a apresentar efeitos modulatórios ou estimuladores no sistema imunológico do hospedeiro. A interação dos micro-organismos com o sistema imune ativa funções protetivas do epitélio, como secreção de peptídeos antimicrobianos e produção de muco. Esse estímulo também incentiva o recrutamento de células imunológicas e maturação de tecidos linfóides (IVANOV; HONDA, 2012).

As bactérias imunomodulatórias produzem compostos que modulam processos homeostáticos do hospedeiro como o desenvolvimento, diferenciação ou função efetora de células imunológicas e epiteliais (IVANOV; HONDA, 2012). Um dos exemplos mais notórios de microrganismos imunomodulatórios são as bactérias fibrolíticas como *Faecalibacterium prausnitzii*, *Roseburia* e *Oridobacter* (SHEEHAN; SHANAHAN, 2017; MORGAN et al., 2012). Essas espécies produzem ácidos graxos de cadeia curta (SCFAs, em inglês *short fatty chain acids*) a partir da fermentação de polissacarídeos. SCFAs são utilizadas pelo epitélio do cólon como

principal fonte de energia e modulam diversos processos biológicos responsáveis por manter a homeostase do intestino como proliferação celular, diferenciação de células T, acetilação de histonas, resposta imune e expressão de genes. Em pacientes com doença de Crohn, observa-se uma quantidade diminuída de bactérias fibrolíticas. Amostras de fezes de adultos com IBDs apresentam menos SCFAs em comparação aos saudáveis, o que demonstra um potencial protetivo dos SCFAs em relação ao desenvolvimento das IBDs (LANE; ZISMAN; SUSKIND, 2017).

Diferente composição microbiótica dos indivíduos com IBDs comparado aos saudáveis é uma das características mais marcantes na doença. A microbiota intestinal de indivíduos saudáveis é composta por quatro filos principais: Firmicutes, Bacteroides, Proteobacteria e Actinobacteria (SHEEHAN; SHANAHAN, 2017). Indivíduos com IBDs apresentam menor biodiversidade e menor proporção de micro-organismos do filo Firmicutes, principalmente bactérias das famílias Lachnospiraceae e Ruminococcaceae (WILLING et al., 2010). Também apresentam aumento na concentração de microrganismos da família Enterobacteriaceae (LANE; ZISMAN; SUSKIND, 2017), como as bactérias patogênicas *Shigella*, *Salmonella* e linhagens aderentes e invasivas de *Escherichia coli* (AIEC). Essas bactérias apresentam mecanismos que aumentam a inflamação e que combatem as bactérias comensais. Deste modo, adquirem vantagem competitiva na presença de estresse oxidativo ocasionado pela mucosa inflamada e não são combatidas pelo sistema imune, sendo uma das causas os defeitos genéticos característicos de indivíduos com IBDs (MORGAN et al., 2012; SHEEHAN; SHANAHAN, 2017; KOSTIC; XAVIER; GEVERS, 2014).

1.3 Sistema imunológico

O sistema imunológico é constituído de um conjunto de tecidos, células e moléculas que atuam em resposta à invasores externos ao corpo do indivíduo, como por exemplo patógenos e substâncias tóxicas (CRUVINEL et al., 2010). Esse sistema é dividido em inato e adquirido: o primeiro representa a primeira barreira de imunidade, incitando uma resposta rápida ao agente externo; o sistema adquirido é uma resposta mais complexa, que

é ativado no caso do sistema inato não ser o suficiente para combater o agente externo. O sistema adquirido também é responsável por gerar e armazenar memórias imunológicas de longo prazo (YATIM; LAKKIS, 2015). As células do sistema imunológico são derivadas de células tronco pluripotentes hematopoiéticas presentes na medula óssea, que se diferenciam em dois tipos: progenitor linfóide comum (PLC) e progenitor mielóide comum (PMC). As PLC originam as células NK (do inglês, *Natural Killer*) e leucócitos: células do sistema imunológico adaptativo, que posteriormente se diferenciam entre leucócitos T e B. As PMC originam as células fagocíticas do sistema imunológico inato: neutrófilos, macrófagos e células dendríticas (STEPHEN; HAJJAR, 2017) (Figura 1).

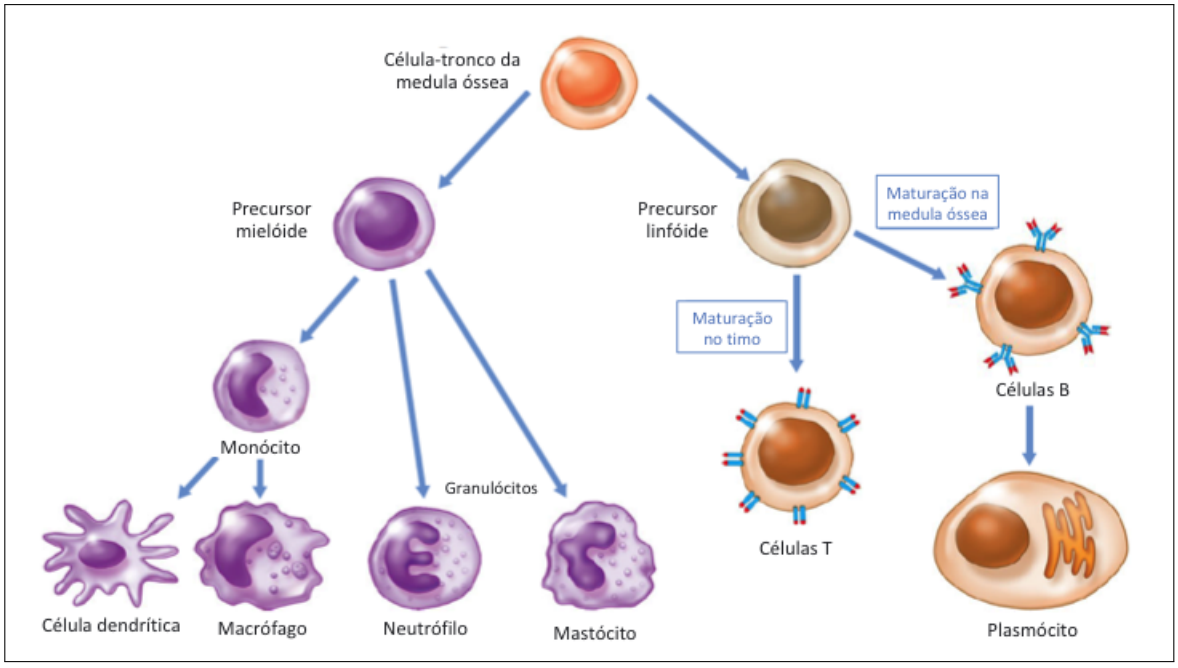

Figura 1: Diferenciação das células imunológicas. Figura modificada de (MADIGAN, 2012).

Sistema Imunológico Inato

Quando o corpo entra em contato com agentes externos, a primeira barreira a ser ativada é o sistema imune inato. As primeiras células a serem ativadas são células fagocíticas como neutrófilos, macrófagos, células NK e células dendríticas (MADIGAN et al., 2016). Essas células reconhecem os patógenos a partir dos Receptores de Reconhecimento de Padrões (PRR) presentes em sua superfície. Os PRR reconhecem partículas chamadas Padrões Moleculares

Associados a Patógenos (PAMPs): moléculas características de microrganismos, que não ocorrem em células humanas. Esses padrões reconhecidos pelos PRR têm origem genética, e não são modificados ao longo da vida do indivíduo, portanto têm a capacidade de reconhecer o agente infectante sem exposição prévia (CRUVINEL et al., 2010).

Depois de reconhecer os agentes externos, as células fagocíticas tentam eliminá-los por fagocitose, processo do qual as células invasoras são englobadas, mortas e digeridas pelas células fagocíticas. Se essa reação primária não for o suficiente para combater a infecção, os peptídeos resultantes da degradação dos patógenos são transformados em antígenos e são apresentados em proteínas de superfície chamadas complexo principal de histocompatibilidade (MHC). Existem dois tipos de MHC: o MHC-I está presente em todas as células do hospedeiro, enquanto o MHC-II está presente somente nas células fagocíticas (MADIGAN et al., 2016; STEPHEN; HAJJAR, 2017). A apresentação dos antígenos no MHC é um dos sinais de ativação do sistema imune adaptativo (STEPHEN; HAJJAR, 2017).

Sistema Imunológico Adaptativo

Quando a resposta do sistema inato não é suficiente para combater a infecção, o sistema imunológico adaptativo é acionado pela apresentação de antígenos no MHC das células fagocíticas. O reconhecimento do complexo MHC-peptídeo é feito por receptores de superfície localizados em células chamadas linfócitos. Há dois tipos de linfócitos: as células T e as células B. Cada linfócito apresenta um tipo de receptor de superfície específico chamados TCR (receptores de células T) e BCR (receptores de células B) (STEPHEN; HAJJAR, 2017). Além disso, os linfócitos apresentam mecanismos de armazenamento de memória do antígeno do qual entram em contato sendo, portanto, capazes de ocasionar respostas imunes específicas (STEPHEN; HAJJAR, 2017).

O primeiro contato dos linfócitos com o complexo MHC-peptídeo desencadeia a resposta imune primária. Essa resposta primeiramente estimula a propagação de linfócitos com a reatividade aos mesmos antígenos reconhecidos. Essa reatividade é específica e os linfócitos reagem somente àquele antígeno. Quando os linfócitos entram em contato com o mesmo patógeno ou antígeno novamente, desencadeiam a resposta imune secundária, que é

mais rápida e intensa do que a primária. Essa resposta promove a marcação dos patógenos para que sejam alvo de destruição (MADIGAN et al., 2016).

Existem vários tipos de linfócitos, eles estão descritos abaixo:

- **Células T auxiliares:** As células T auxiliares (Th) são as que interagem com os MHC-II. Depois da interação, elas promovem a propagação das Th e diferenciação entre Th1 e Th2. As Th1 antígeno-específicas interagem com os MHC-II de macrófagos, os ativando e estimulando a produção de citocinas pelas próprias Th1. Citocinas são classes de proteínas solúveis que estimulam outras células imunológicas para que iniciem ou intensifiquem a resposta imune e fagocitose e promovam o início do processo inflamatório. A função das células Th2 é produzir citocinas que estimulam a produção de anticorpos pelas células B (MADIGAN et al., 2016).

- **Células T citotóxicas:** As células T citotóxicas (Tc) interagem com os complexos MHC-I/peptídeo de uma célula infectada. Elas desenvolvem especificidade para aquele antígeno e secretam toxinas que matam as células portadoras dele (MADIGAN et al., 2016).

- **Células B:** Os linfócitos B são células que possuem anticorpos em sua superfície. Anticorpos são proteínas que interagem com antígenos específicos, reconhecendo os patógenos. Cada uma das células B reconhecem especificamente um tipo de antígeno e atuam em meios extracelulares como no sangue e em secreções mucosas (MADIGAN et al., 2016).

  Quando uma célula B tem a sua primeira interação com um antígeno, ela fagocita o patógeno, o digere e apresenta os peptídeos resultantes em MHC-II da sua superfície. Esses MHC-II/peptídeo são reconhecidos pelas células Th2, estimulando a produção de citocinas. As citocinas promovem a clonagem dessas células B antígeno-reativas. Alguns clones permanecem no sangue como células de memória para induzir uma resposta secundária no caso de uma nova infecção do mesmo antígeno. Outros clones diferenciam-se em plasmócitos produtores de anticorpos, cuja função será bloquear a

interação do patógeno ou de seus produtos com a célula hospedeira ou marcá-los para a destruição (MADIGAN et al., 2016).

Inflamação

Inflamação é uma resposta imunológica ativada não somente em caso de infecções, mas também de injúrias e danos contra tecidos e modificações bioquímicas que podem comprometer a homeostase do organismo. O processo inflamatório controlado é considerado benéfico, uma vez que seu intuito é principalmente combater agentes infecciosos e restaurar a homeostase do organismo (MEDZHITOV, 2008). Entretanto, quando descontrolada, a inflamação pode levar o desenvolvimento de doenças que incluem as IBDs (LANE; ZISMAN; SUSKIND, 2017).

O processo inflamatório inicia-se com o reconhecimento do tecido injuriado ou infeccionado pelos receptores *toll-like receptors* (TLRs) e *nucleotide-binding oligomerization domain protein-like receptors* (NLRs) que estão presentes nos macrófagos e mastócitos. Posteriormente, inicia-se a produção de proteínas mediadoras inflamatórias como citocinas e chemoquinas que recrutam neutrófilos e outros leucócitos ao local em questão. Os mediadores inflamatórios também são responsáveis por ativar o endotélio dos vasos sanguíneos, aumentando sua permeabilidade e permitindo que os neutrófilos cheguem às veias pós-capilares e tecidos extravasculares para acessar o tecido inflamado. Como consequência, o organismo manifesta vermelhidão, temperatura elevada e inchaço no local, que são sintomas característicos das inflamações (MEDZHITOV, 2008; KUPRASH; NEDOSPASOV, 2016). Chegando ao tecido alvo, os neutrófilos são ativados tanto pelo contato com os patógenos quanto pelas citocinas produzidas e combatem os patógenos a partir da produção de moléculas tóxicas como elastase, proteinase 3, catepsina G e espécies reativas de oxigênio (EROs) e de nitrogênio (ERNs) (MEDZHITOV, 2008; KUPRASH; NEDOSPASOV, 2016).

Citocinas são classes de polipeptídeos e glicoproteínas produzidas pelas células imunes (principalmente células T, neutrófilos e macrófagos) para atuar na sinalização dos processos imunológicos. Diferentes citocinas desempenham papéis específicos na ativação, diferenciação, proliferação ou recrutamento de células do sistema imunológico (FERREIRA et al., 2019). As principais citocinas estão descritas abaixo, de acordo com (ALBERTS et al., 2004).

- **IL-2:**

  Função: estímulo da proliferação e diferenciação

  Células produtoras: células T auxiliares, algumas células T citotóxicas e mastócitos ativados

  Células alvo: células T ativadas e células B

- **IL-4:**

  Função: estímulo da proliferação, maturação de anticorpos nas células B e inibição do desenvolvimento das células Th1.

  Células produtoras: células Th2 e mastócitos

  Células alvo: células Th e células B

- **IL-5:**

  Função: estímulo da proliferação e maturação

  Células produtoras: células Th2 e mastócitos

  Células alvo: eosinófilos e células B

- **IL-10:**

  Função: inibição do desenvolvimento de células Th1 e de macrófagos

  Células produtoras: células Th2, macrófagos e células dendríticas

  Células alvo: eosinófilos e células B

- **IL-12:**

  Função: indução do desenvolvimento de células Th2 e inibição do desenvolvimento de células Th1

  Células produtoras: células B, macrófagos e células dendríticas

  Células alvo: células T virgens

- **IFN-γ:**

Função: ativação de genes de macrófago e MHC, aumento da expressão de MHC em vários tipos celulares

Células produtoras: células Th1

Células alvo: células B, macrófagos e células endoteliais

- **TNF-α:**

  Função: ativação celular

  Células produtoras: células Th1 e macrófagos

  Células alvo: células endoteliais

Quando a resposta inflamatória é bem sucedida eliminando os agentes infecciosos, inicia-se a fase de reparação do tecido e resolução da inflamação. Para a iniciação dessa fase, é crucial a transição entre a produção de mediadores pró-inflamatórios para os anti-inflamatórios, como lipoxinas, que dá-se a partir de mecanismos moleculares e celulares ativos, incluindo a entrada dos neutrófilos no estado apoptótico. As lipoxinas permitem o início da fase de resolução inibindo o recrutamento de neutrófilos e promovendo o recrutamento de monócitos, que removem as células mortas. Outros lipídios mediadores como resolvinas e protectinas reagem com os fatores de crescimento produzidos pelos macrófagos e, juntamente aos processos dos monócitos, iniciam o remodelamento do tecido inflamado (MEDZHITOV, 2008; ONALI; FAVALE; FANTINI, 2019).

Quando a resposta inflamatória falha em eliminar os agentes infecciosos, o quadro é promovido a um estado de inflamação crônica. Neste caso, os neutrófilos são substituídos por macrófagos e células T. Como consequência, podem ocorrer danos nos tecidos em decorrência de efeitos colaterais como, por exemplo, as respostas auto-imune (MEDZHITOV, 2008).

A ocorrência das IBDs são exemplos de consequência do descontrole do processo inflamatório. Nesse caso, a inflamação descontrolada é efeito de uma resposta imunológica anormal contra os antígenos da própria microbiota comensal de indivíduos suscetíveis geneticamente. Esse processo causador da inflamação crônica ainda no caso das IBDs não é completamente explicado pela literatura (ONALI; FAVALE; FANTINI, 2019). (ONALI; FAVALE; FANTINI, 2019) apresenta duas hipóteses a respeito da origem inflamatória das

IBD: a primeira é de que há perda na tolerância que o sistema imunológico apresenta com os antígenos dos microrganismos comensais. Nesse caso, a inflamação evolui para o quadro crônico por causa da exposição contínua aos antígenos, que estão constantemente presentes em contato com o tecido. De acordo com a segunda hipótese, os sistemas de alternância entre os mecanismos pró- e anti-inflamatórios apresentam defeitos que levam à perturbação do processo inflamatório (ONALI; FAVALE; FANTINI, 2019).

Além dessas hipóteses, outros processos biológicos relacionados à inflamação apresentam defeitos em células de indivíduos com DC e também estão relacionados à indução ou exacerbação do processo inflamatório excessivo. Como exemplos temos a autofagia e apoptose, processos são ativados para a resolução da inflamação. Dependendo da condição, a resolução pode ser levada a um destino pró-sobrevivência celular, onde a via de autofagia será ativada, ou pró-morte celular, onde a apoptose será ativada (MESSER, 2017). Além disso, a apoptose e autofagia estão relacionadas à homeostase e regeneração da barreira intestinal que, quando prejudicada, oportuniza a penetração do epitélio por células imunológicas e bacterianas, também desencadeando o processo inflamatório (BLANDER, 2016).

1.4 Homeostase do tecido intestinal

A homeostase intestinal é fundamental para o funcionamento correto do intestino, bem como a manutenção da comunidade microbiana. O tecido epitelial intestinal desempenha os principais papéis para manter a homeostase, pois é a barreira do lumen intestinal, onde habita a microbiota. Sendo assim, apresenta uma estrutura complexa de diferentes tipos celulares cada qual com uma função específica, que passa por uma constante renovação orquestrada por diversos processos biológicos como diferenciação, apoptose e autofagia, que necessitam atuar em sincronia para que o intestino mantenha a homeostase (DELGADO; GRABINGER; BRUNNER, 2016).

O epitélio intestinal é constituído por uma camada única de células epiteliais que formam uma barreira de proteção entre a luz do intestino e a lâmina própria e estão dispostos no intestino em estrutura de vilosidades. O tecido interno das vilosidades intestinais é formado

basicamente por mioblastos, tecido fibroso e vasos sanguíneos, assim sendo responsável pela circulação sanguínea das criptas intestinais (POWELL et al., 2011) (Figura 2).

As células presentes no epitélio intestinal podem ser dos tipos absortivas, secretoras ou células-tronco. As células-tronco estão localizadas no fundo das criptas intestinais e diferenciam-se em enterócitos absortivos ou células secretoras (células caliciformes, células de Paneth ou células enteroendócrinas) de acordo com a sinalização recebida (OBATA et al., 2012). No fundo da cripta, também estão localizadas as células de Paneth, responsáveis por proteger o epitélio contra a entrada de microrganismos produzindo α-defensina HD5, um peptídeo com propriedades antibióticas e α-defensina HD6, que forma redes na cripta do intestino para restringir a mobilidade das bactérias (STANGE; WEHKAMP, 2016; SHI, 2007). Em direção ao topo da cripta encontram-se as células caliciformes, que secretam muco em direção ao lúmen e as células epiteliais absortivas (DELGADO; GRABINGER; BRUNNER, 2016) (Figura 2).

A quantidade das células intestinais é balanceada em indivíduos saudáveis, entretanto, pacientes com IBDs apresentam anormalidades em sua composição. Em indivíduos com doença de Crohn é observado um aumento nas células caliciformes, diminuição nas células de Paneth e muco intestinal mais espesso. Em indivíduos com retocolite ulcerativa há redução na produção de muco e defeitos de diferenciação das células caliciformes (HEAZLEWOOD et al., 2008). Deste modo, o comprometimento da barreira de defesa da mucosa intestinal possibilita o contato dos micro-organismos com o epitélio, ocasionando a inflamação característica das IBDs (STANGE; WEHKAMP, 2016).

A estrutura saudável do epitélio é mantida por uma constante renovação que mantém a integridade do epitélio, impedindo a invasão de microrganismos e permitindo a secreção apropriada de peptídeos antibacterianos e absorção de nutrientes. A renovação acontece com a ocorrência simultânea entre a extrusão de células epiteliais senescentes no lúmen e a diferenciação das células tronco, que posteriormente migram em direção ao topo da cripta substituindo as células dispersas. Quando a célula senescente é sinalizada para ser liberada no lúmen, ela estará entrando em um estado apoptótico e será degradada por células dendríticas e macrófagos que realizam a autofagia da célula. Em condições saudáveis, a proporção de

novas células que se diferenciam e migram para o topo da cripta compensa o número de células extrudadas, mantendo a barreira intacta. Além disso, os macrófagos e células dendríticas são capazes de degradar as células soltas no lúmen antes que seu número seja suficiente para desencadear um processo inflamatório (BLANDER, 2016; DELGADO; GRABINGER; BRUNNER, 2016) (Figura 2).

Em células de indivíduos com IBDs, observa-se um nível aumentado de apoptose das células epiteliais, que passa a ser responsável pela permeabilidade da barreira intestinal: as células tronco não são capazes de se diferenciar em tempo suficiente para substituir as células mortas do epitélio, portanto a integridade da barreira passa a ser comprometida permitindo a invasão microbiana na lâmina própria e desencadeando processo inflamatório. Além disso, observa-se que a autofagia também está comprometida em pacientes de IBDs, portanto o alto número de células exiladas no lúmen em decorrência da apoptose aumentada não é degradado eficientemente, também induzindo processos inflamatórios (BLANDER, 2016).

Autofagia

Autofagia é um processo homeostático que tem o intuito de reciclar componentes citosólicos não mais funcionais como organelas, proteínas e microorganismos invasores decompondo-os para que suas moléculas sejam reutilizadas pelas células. Em condições normais a autofagia atua em nível basal, mas é intensificada em casos de estresse como privação de nutrientes, hipóxia, estresse oxidativo e infecção (MIZUSHIMA, 2007; KUBISCH et al., 2013).

Existem três tipos distintos de autofagia em células de mamíferos: autofagia mediada por chaperonas, microautofagia e macroautofagia. A autofagia mediada por chaperonas é responsável pela degradação de proteínas que apresentam motifs específicos em sua sequência de aminoácidos. No processo de microfagia, ocorre o engolfamento dos componentes citoplasmáticos diretamente pelo lisossomo. No processo de macroautofagia, há vesículas de membrana dupla especializadas em sequestrar os componentes celulares a serem degradados, chamadas autofagossomo (HENDERSON; STEVENS, 2012). Nesta tese, utilizamos o termo “autofagia” nos referindo à macroautofagia, uma vez que é o processo envolvido no estudo.

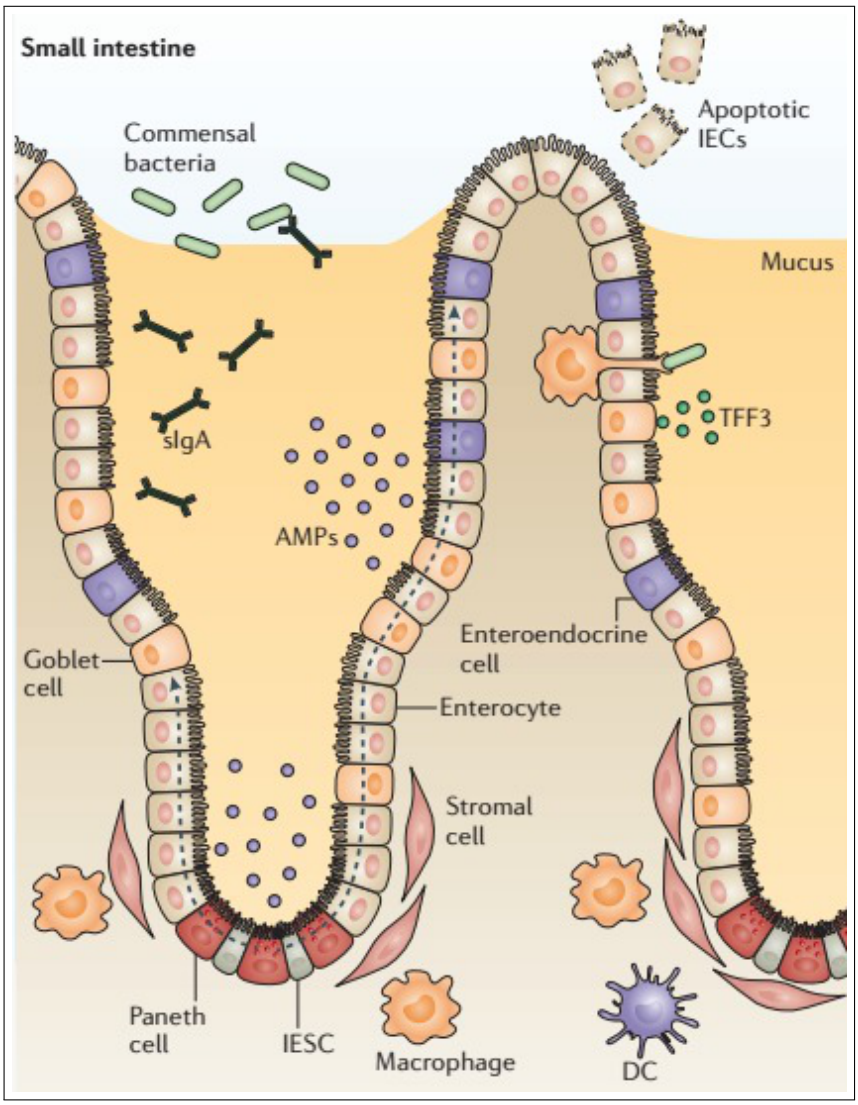

Figura 2: O epitélio do intestino é formado por células caliciformes (*Goblet cell*), células de Paneth (*Paneth cell*), células-tronco (IESC), enterócitos (*Enterocyte*) e células enteroendócrinas (*Enteroendocrine cell*). As células epiteliais senescentes entram em apoptose (*Apoptotic IECs*) e são liberadas no lúmen intestinal. Em um epitélio saudável, as células-tronco se proliferam e diferenciam-se nos outros tipos de células epiteliais e as células diferenciadas migram em direção ao topo das vilosidades para substituir as células que entraram em apoptose. (Figura modificada de (PETERSON; ARTIS, 2014))

Sumariamente, a autofagia inicia-se com a formação de uma vesícula de membrana bilipídica, o autofagossomo, que compõe-se ao redor do elemento realizando seu sequestro. O autofagossomo fusiona-se então com lisossomos que contêm enzimas de degradação que realizam a fragmentação do componente, disponibilizando suas moléculas para serem reutilizadas pela célula ou como fonte de energia (GREEN; GALLUZZI; KROEMER, 2011) (Figura 4). Um total de 38 genes estão envolvidos no processo de autofagia, cada qual desempenhando

diferentes papéis ao longo das etapas (Figura 3) (KUBISCH et al., 2013; MIZUSHIMA, 2007).

Alguns genes envolvidos no processo de autofagia são considerados como genes de susceptibilidade da doença de Crohn, incluindo ATG16L1, NOD2, ULK1 e IRGM. ATG16L1 é um dos genes responsáveis pela biogênese do autofagossomo, enquanto os outros genes estão envolvidos na indução da autofagia (HENDERSON; STEVENS, 2012; YANO; KURATA, 2009; KUBISCH et al., 2013; MICHAIL; BULTRON; DEPAOLO, 2013).

Um dos principais sinais de indução por esses genes é a presença de patógenos no intestino, entretanto, mesmo que o indivíduo não apresente mutações nesses genes, muitas bactérias como *Salmonella* Typhimurium podem interferir no processo de autofagia. Na doença de Crohn, observamos a presença aumentada de *Salmonella* e outros microrganismos patógenos. Uma vez que a autofagia é um modulador importante da maquinaria de inflamação, podemos sugerir que a presença desses patógenos seja um meio indireto de interferir no sistema imune e, desse modo, aumentando a susceptibilidade à doença de Crohn (SUDHAKAR et al., 2019).

Mitofagia

A maioria dos processos de autofagia realizam a degradação não seletiva de componentes celulares. Entretanto, existem tipos específicos de autofagia que realizam a reciclagem de tipos selecionados de componentes alvo (CHEONG et al., 2008) como é o caso da mitofagia, cujo alvo de degradação é a organela mitocôndria.

A mitocôndria é uma organela importante o mantimento da homeostase intestinal, pois é a principal reguladora de espécies reativas de oxigênio (EROs) celular. Em um nível controlado, EROs regula processos importantes como diferenciação e proliferação celular e defesa imunológica contra microrganismos invasores. Entretanto, em níveis muito elevados é extremamente prejudicial para o tecido, aumentando os níveis de inflamação e induzindo à morte celular. Quando danificadas, as mitocôndrias produzem um alto nível de EROS e são recicladas por meio de autofagia, se a célula estiver em condições saudáveis (HAMACHER-BRADY; BRADY, 2016; MOTTAWEA et al., 2016; CORRIDONI et al., 2018).

Consequentemente, a disfunção da mitocôndria está relacionada à patogênese da DC por meio dos danos causados pelo excesso de EROs. Não obstante, estudos mostram que indivíduos com DC apresentam deficiência na produção de energia pela mitocôndria, mitocôndrias morfologicamente danificadas e altos níveis de EROs (CORRIDONI et al., 2018; MOTTAWEA et al., 2016).

Apoptose

A apoptose é o principal tipo de morte celular programada que ocorre nos tecidos. Ocorre durante a embriogênese ou em adultos para a renovação periódica do tecido para o mantimento da homeostase. Também é um mecanismo de defesa contra patógenos, danos celulares ou estímulos externos como algumas drogas ou irradiação (ELMORE, 2007).

Existem duas vias de sinalização de apoptose: a intrínseca e a extrínseca. A via extrínseca é estimulada por marcadores presentes na superfície da célula que ligam-se a moléculas específicas e propagam sinais apoptóticos para o interior da célula. Esses receptores incluem proteínas como as da família *tumor necrosis factor* (TNF) que trimerizam-se e ligam-se a caspases iniciadoras como a caspase-8 e -10, que sinalizam para as caspases-3, -6 e -7 que são responsáveis por promover a degradação celular e últimos estágios da apoptose (MUKHOPADHYAY et al., 2014).

O estímulo da via intrínseca da apoptose ocorre na mitocôndria em resposta a vários estímulos de estresse como EROs, danos no DNA, privação de fatores de transcrição e hipóxia e inicia com a permeabilização da membrana mitocondrial externa. Essa via é controlada por proteínas pró- e anti-apoptóticas da família Bcl-2 que atuam em conjunto para controlar as etapas do processo. A fase inicial é estimulada quando as proteínas pró-apoptóticas Bax e Bak dimerizam-se e penetram a mitocôndria, permitindo a liberação do citocromo C no citosol. O citocromo C, por sua vez, liga-se à proteína Apaf-1 (*apoptotic protein activating factor-1*) que inicia a formação do apoptossomo. O apoptossomo é um complexo de sete proteínas responsável por ativar a caspase-9, que ativa caspase-3 iniciando a apoptose (MUKHOPADHYAY et al., 2014). Um dos controles da apoptose é feito por proteínas chamadas inibidoras chamadas inibidoras de proteínas apoptóticas (IAP, do ingês

inhibitor of apoptotic proteins). As IAP permanecem ligadas às caspases, que permanecem inativas, até que os complexos Smac/DIABLO ou HtrA2/Omi sejam liberados da mitocôndria e desassociem as IAP das caspases (MUKHOPADHYAY et al., 2014). Durante o processo de morte celular, as células apoptóticas são sinalizadas por marcadores que permitem que sejam encontradas e fagocitadas por macrófagos, evitando uma resposta imunológica contra as células mortas (NAGATA, 2018).

Em indivíduos com doença de Crohn, a apoptose pode estar em nível aumentado ou diminuído dependendo do tipo de célula. Por exemplo, linfócitos T saudáveis entram em estado apoptótico após cumprirem seu papel de identificar e eliminar estímulos patológicos. Entretanto, linfócitos T da lâmina própria de indivíduos com DC são resistentes à apoptose, deste modo, induzindo a propagação da inflamação (EDER et al., 2015). Por outro lado, observou-se um aumento no nível de apoptose nas células do epitélio intestinal do tipo enterócitos em condições de inflamação, o que promove um aumento na permeabilidade da barreira epitelial. Essa permeabilidade permite a invasão de células imunológicas e bactérias na lâmina própria, aumentando o nível da inflamação (SABATINO et al., 2003).

1.5 Susceptibilidade genética às doenças inflamatórias intestinais

Defeitos nos processos biológicos que mantêm a homeostase dos tecidos relacionados ao desenvolvimento de IBDs estão associados a variações genéticas presentes nos indivíduos susceptíveis às doenças. Um exemplo de gene de risco é Nod2. Esse gene reconhece estímulos bacterianos e em resposta estimula a secreção de peptídeos antimicrobianos e é altamente expresso pelas células de Paneth, cuja secreção de α-defensina é comprometida em pacientes com mutação em Nod2 (SHI, 2007; YANO; KURATA, 2009). Variantes desse gene representam aumento no risco de desenvolvimento das IBDs (SHI, 2007; STANGE; WEHKAMP, 2016).

Outro gene importante para o aumento da susceptibilidade de IBDs é o Atg16L1, proteína relacionada à autofagia (SHI, 2007; WANG et al., 2018). Ratos com mutações no gene Atg16L1 apresentam anormalidades nas células de Paneth incluindo diminuição na produção de lisozimas, grânulos reduzidos e desorganizados, mitocôndria degenerada e falta de

microvilosidades apicais (YANO; KURATA, 2009; HENDERSON; STEVENS, 2012). Outros estudos mostram que células com Atg16L1 deficiente não realizam autofagia nem no nível mais basal, desbalanceando a composição da microbiota intestinal (YANO; KURATA, 2009). Além disso, macrófagos deficientes em Atg16L1 produziram altas quantidades das interleucinas pró-inflamatórias interleucina 1-beta (IL-1β) e interleucina 18 (IL-18), contribuindo para o aumento da inflamação do tecido intestinal (SAITOH et al., 2008). Além dos variantes citados, existem mais de 163 genes de risco para IBDs identificados. Alguns dos mais relevantes clinicamente estão apresentados na tabela 1.

Tabela 1: Lista de genes de risco para doenças inflamatórias intestinais (WANG et al., 2018).

Gene	Função
Nod2	Resposta do sistema imune inato
Atg16L1	Homeostase da mucosa e degradação de componentes celulares (autofagia)
IL23R	Receptor de citocina tipo 1
IBD5	Citocina associada ao risco de desenvolvimento de doença de Crohn
TLR4	Receptor *Toll-like*. Detecta lipopolissacarídeos de bactérias gram-negativas e ativa o sistema imune
OCTN1	Proteína presente na membrana plasmática responsável pelo cotransporte de íons de sódio e ergotioneína
IRGM	Regula a autofagia em resposta a patógenos intracelulares
DLG5	Proteína importante em sítios de contato intercelular
LRRk2	Substrato para autofagia mediada por chaperonas
PTPN22	Influencia na resposta dos receptores de celulas T e B
IL10R	Intermedeia o sinal imunosupressor da interleucina 10, inibindo a síntese de citocinas pró-inflamatórias
TNF	Citocina envolvida na inflamação sistêmica

A informação sobre os genes de susceptibilidade às doenças inflamatórias intestinais têm sido obtidos por estudos de associação ampla do genoma (GWAS) (HENDERSON; STEVENS, 2012). GWAS é a análise de como variantes genéticos estão distribuídos entre indivíduos de diferentes populações. A partir dessas análises, é possível determinar associações

entre os variantes e diferentes doenças (NORRGARD, 2008), como no caso dos genes de susceptibilidade às IBDs descritos na seção anterior.

Enquanto as informações sobre a susceptibilidade genética dos indivíduos com IBDs são obtidas a partir de GWAS, informações do microbioma intestinal são obtidas pela exploração de tecnologias como 16S rRNA, metagenômica *whole genome shotgun* (WGS), metatranscriptômica e metaproteômica. Estudos de 16S rRNA e metagenômica permitem a identificação das características relacionadas à microbiota de indivíduos com IBDs e são úteis para inferências primárias associadas à doença. Entretanto, essas análises excluem informações essenciais sobre a presença ou ausência de proteínas microbianas que são importantes para o estudo completo da interação microbioma-hospedeiro. Os estudos de metatranscriptômica fornecem um panorama mais completo a respeito das proteínas expressas pela comunidade microbiana do indivíduo, como por exemplo informações sobre a diferença entre os transcritos microbianos de indivíduos saudáveis e indivíduos com IBDs (ERICKSON et al., 2012; PRESLEY et al., 2012; JUSTE et al., 2014). Entretanto, apesar desses avanços tecnológicos, não há esclarecimento suficiente dos mecanismos moleculares e vias de sinalização do hospedeiro que são mediadas por proteínas microbianas, seja de indivíduos saudáveis ou doentes.

1.6 Redes biológicas

As informações obtidas por dados ômicos resultam em dados moleculares de larga escala que necessitam de uma análise computacional para a interpretação de seu contexto biológico (REDESTIG et al., 2018). Essa etapa representa um dos maiores desafios da ciência atualmente, uma vez que a quantidade de dados gerados ultrapassa a lei de Moore. Isso significa que a quantidade de dados biológicos aumenta exponencialmente mais rápido do que a capacidade computacional de armazenar e processar os dados obtidos com as tecnologias disponíveis atualmente (OLIVEIRA, 2019). Deste modo, o desenvolvimento de novos métodos para análise computacional de dados biológicos de larga escala é imprescindível para o avanço das ciências biológicas.

Outro desafio está na conjuntura de como as análises biológicas são tradicionalmente

feitas: assume-se que os sistemas biológicos podem ser subdivididos em módulos e busca-se entendimento de suas partes separadamente. Essa abordagem é chamada de reducionista. Entretanto, uma análise fragmentada do sistema nem sempre encontra as resultados satisfatórios, pois os organismos são sistemas complexos e interconectados em diversos níveis. Deste modo, a biologia de sistemas emergiu com a necessidade de uma abordagem holística que pudesse levar em conta o cenário biológico como um todo (GREEN et al., 2017; JUNKER; SCHREIBER, 2008).

Uma das estratégias utilizadas na biologia de sistemas para uma investigação global é a utilização de redes para a análise dos dados ômicos (JUNKER; SCHREIBER, 2008). A partir da geração de redes de dados ômicos é possível inferir modelos que são utilizados para a predição de padrões presentes em diferentes condições, como por exemplo saudável e patológica (BADER; KüHNER; GAVIN, 2008). Também é possível gerar diversos tipos de redes biológicas com diferentes níveis de interação molecular e dados ômicos, como por exemplo redes regulatórias, redes de interação de proteínas e redes metabólicas (JUNKER; SCHREIBER, 2008).

As redes de interações proteína-proteína (PPIs) apresentam alto potencial em representar o funcionamento de processos biológicos e também ser utilizadas para traçar modelos de vias moleculares, pois grande parte dos processos biológicos da célula são controlados por interações físicas entre proteínas, bem como muitas doenças são atribuídas a alterações dessas interações (JUNKER; SCHREIBER, 2008). Há diversos métodos de averiguar a interação entre duas proteínas, sendo os mais usados o sistema Y2H (do inglês, *yeast two-hybrid system*) e AP-MS (do inglês, *Affinity Purifications/Mass Spectrometry*) (BADER; KüHNER; GAVIN, 2008). Entretanto, os métodos experimentais apresentam limitações como baixa cobertura e vieses de localização celular e de tipos de proteínas, além do custo financeiro. Deste modo, a utilização de métodos computacionais de predição de interação entre proteínas mostra-se uma alternativa viável para simulações prévias de candidatos à interação e também pode conceder informações a respeito dos detalhes das interações que não é possível observar pelos métodos experimentais (SHOEMAKER; PANCHENKO, 2007).

Dentre os diversos métodos computacionais de predição de PPIs estão incluídas abor-

dagens que utilizam co-localização dos genes em um cluster, padrões de co-evolução entre proteínas, co-expressão de genes, perfis filogenéticos e também a presença de domínios e motifs nas proteínas (SHOEMAKER; PANCHENKO, 2007). A maior parte das interações entre as proteínas ocorrem entre regiões específicas chamadas domínios e motifs. Domínios são regiões da estrutura da proteína que reconhecem alvos e ligam-se a domínios ou motifs de outras proteínas (BADER; KüHNER; GAVIN, 2008). Motifs, por sua vez, são pequenas sequências conservadas de aminoácidos localizadas no C-término da proteína e é uma região de baixa afinidade de ligação (KORCSMAROS et al., 2013). Bancos de dados como DOMINE (YELLABOINA et al., 2011), PFAM (EL-GEBALI et al., 2019) e ELM (GOUW et al., 2018) contém informações preditas e/ou experimentalmente validadas sobre quais domínios e motifs que interagem entre si.

Redes de Interação Microbioma-Hospedeiro

A grande maioria das ferramentas computacionais e bancos de dados de interação entre proteínas é direcionada a moléculas pertencentes a uma única espécie e técnicas experimentais para a obtenção dessas informações são demoradas, financeiramente e computacionalmente custosas (NOURANI; KHUNJUSH; DURMUŞ, 2015; DYER; MURALI; SOBRAL, 2008; DYER et al., 2010; GOULD et al., 2018; WEIMER et al., 2018; FOLEY et al., 2013). Entretanto, o estudo das interações microbioma-hospedeiro são importantes para o avanço de diversos campos como agricultura, pecuária, biotecnologia e saúde. Independente do organismo, a interação microbioma-hospedeiro ocorre a partir da interação entre moléculas como proteínas, metabólitos e pequenos RNAs secretados por ambos microbioma e pelo hospedeiro (HEINKEN et al., 2013; NOURANI; KHUNJUSH; DURMUŞ, 2015; KATIYAR-AGARWAL; JIN, 2010). As moléculas dos micro-organismos têm o potencial de interagir com as moléculas do hospedeiro, que podem lançar cascatas intracelulares capazes de afetar a expressão de genes chave e interagir com proteínas, assim modulando processos biológicos do organismos hospedeiro. Deste modo, redes de interação microbioma-hospedeiro fornecem informações de como as interações entre suas moléculas interferem na fisiologia e metabolismo do hospedeiro (GUVEN-MAIOROV; TSAI; NUSSINOV, 2017), tendo o potencial de lançar

luz em como os micro-organismos podem estar relacionados no desenvolvimento de doenças, como as IBDs.

Gene name	Protein name	UniProt ID	Phases of autophagy: Induction	Cargo recognition and packaging	Atg protein cycling	Vesicle nucleation	Vesicle expansion and completion	Transport of autophagosome	Fusion with lysosome
ULK1	Serine/threonine-protein kinase ULK1	O75385	■		■				
ULK2	Serine/threonine-protein kinase ULK2	Q8IYT8	■		■				
RB1CC1	RB1-inducible coiled-coil protein 1	Q8TDY2	■						
ATG13	Autophagy-related protein 13	O75143	■		■				
ATG101	Autophagy-related protein 101	Q9BSB4	■						
p62/SQSTM1	Sequestosome-1	Q13501		■					
ATG2A	Autophagy-related protein 2 homolog A	Q2TAZ0			■				
ATG2B	Autophagy-related protein 2 homolog B	Q96BY7			■				
WIPI1	WD repeat domain phosphoinositide-interacting protein 1	Q5MNZ9			■				
WIPI2	WD repeat domain phosphoinositide-interacting protein 2	Q9Y4P8			■				
BECN1	Beclin-1	Q14457				■			
ATG14	Beclin 1-associated autophagy-related key regulator	Q6ZNE5				■			
PIK3R4	Phosphoinositide 3-kinase regulatory subunit 4	Q99570				■			
PIK3C3	Phosphatidylinositol 3-kinase catalytic subunit type 3	Q8NEB9				■			
UVRAG	UV radiation resistance-associated gene protein	Q9P2Y5				■			
AMBRA1	Activating molecule in BECN1-regulated autophagy protein 1	Q9C0C7				■			
ATG3	Ubiquitin-like-conjugating enzyme ATG3	Q9NT62					■		
ATG4A	Cysteine protease ATG4A	Q8WYN0					■		
ATG4B	Cysteine protease ATG4B	Q9Y4P1					■		
ATG4C	Cysteine protease ATG4C	Q96DT6					■		
ATG4D	Cysteine protease ATG4D	Q86TL0					■		
ATG5	Autophagy protein 5	Q9H1Y0					■		
ATG7	Ubiquitin-like modifier-activating enzyme ATG7	O95352					■		
MAP1LC3A	Microtubule-associated proteins 1A/1B light chain 3A	Q9H492					■		
MAP1LC3B	Microtubule-associated proteins 1A/1B light chain 3B	Q9GZQ8					■		
MAP1LC3C	Microtubule-associated proteins 1A/1B light chain 3C	Q9BXW4					■		
MAP1LC3B2	Microtubule-associated proteins 1A/1B light chain 3 beta 2	A6NCE7					■		
GABARAP	Gamma-aminobutyric acid receptor-associated protein	O95166					■		
GABARAPL1	Gamma-aminobutyric acid receptor-associated protein-like 1	Q9H0R8					■		
GABARAPL2	Gamma-aminobutyric acid receptor-associated protein-like 2	P60520					■		
GABARAPL3	Gamma-aminobutyric acid receptor-associated protein-like 3	Q9BY60					■		
ATG9A	Autophagy-related protein 9A	Q7Z3C6	■				■		
ATG10	Ubiquitin-like-conjugating enzyme ATG10	Q9H0Y0					■		
ATG12	Ubiquitin-like protein ATG12	O94817					■		
ATG16L1	Autophagy-related protein 16-1	Q676U5					■		
ATG16L2	Autophagy-related protein 16-2	Q8NAA4					■		
FYCO1	FYVE and coiled-coil domain-containing protein 1	Q9BQS8						■	
TECPR1	Tectonin beta-propeller repeat-containing protein 1	Q7Z6L1							■

Figura 3: Tabela contendo os genes que compõem a maquinaria de autofagia e as etapas a que estão relacionados (KUBISCH et al., 2013)

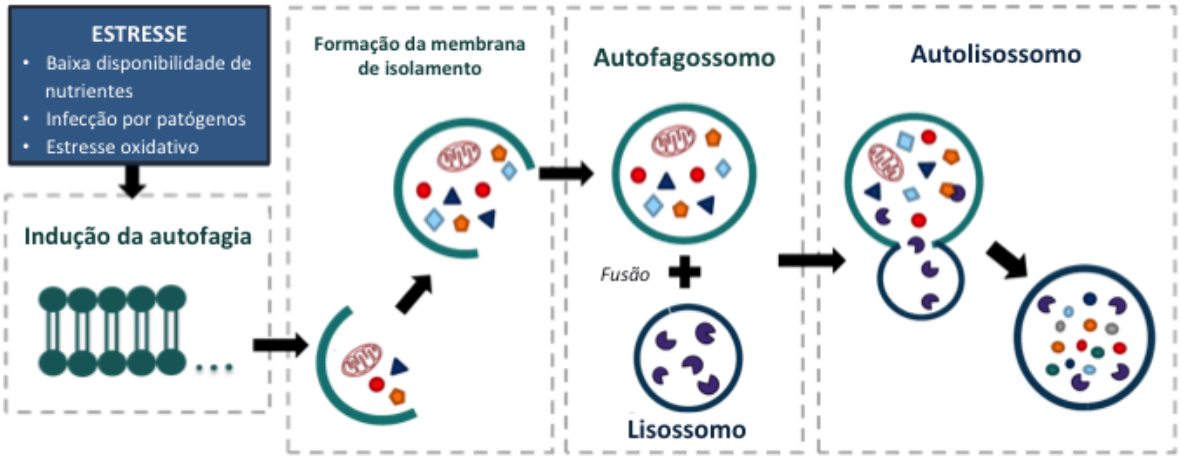

Figura 4: Etapas do processo de autofagia. Primeiramente, a indução da autofagia por condições de estresse induz a formação da membrana de isolamento do autofagossomo. Quando a formação do autofagossomo se completa, ele fusiona-se com o lisossomo, que contém enzimas de degradação, formando o autolisosssomo. No autolisossomo, os componentes são degradados e suas moléculas são liberadas para a utilização na síntese de novos componentes celulares.

2 Justificativa e Proposta

Na última década, a incidência das doenças inflamatórias intestinais têm aumentado progressivamente. Embora apresente maior predominância em países desenvolvidos, países da América Latina como o Brasil, têm apresentado aumentos significativos na manifestação das doenças: do ano de 1988 a 2012 houve um aumento percentual anual de 11,1% nos casos de doença de Crohn e 14,9% nos casos de retocolite ulcerativa, com tendência de aumento nos anos seguintes (NG et al., 2018).

Projetos que visam estudar IBDs com parcerias internacionais são de extrema relevância, pois países como a Inglaterra - onde a doutoranda realizou doutorado sanduíche - possuem mais recursos, maior quantidade de dados e maiores avanços nas pesquisas, pois já vem enfrentando problemas com essas doenças em maior escala e anteriormente aos países da América Latina. Trazer esse conhecimento com potencial para novas tecnologias relacionadas à cura, tratamento e profilaxia de IBDs para o Brasil passa a ser uma questão de saúde pública visto as estatísticas de crescimento de pacientes com IBDs e o prejuízo que os pacientes apresentam em sua qualidade de vida. Além disso, a metodologia proposta tem o potencial de ser aplicada em outras problemáticas, outras doenças e outros organismos.

Nesse estudo propomos a utilização de uma abordagem computacional integrada que acessa o impacto potencial de proteínas microbianas diferencialmente presentes em pacientes com doença de Crohn e indivíduos saudáveis. Utilizamos assinaturas de interação baseadas nas características estruturais das proteínas para predizer a interação entre receptores de proteínas humanas e proteínas microbianas expressadas unicamente em indivíduos com DC ou saudáveis. O conjunto de dados de proteínas microbianas foi obtido de um estudo sueco (ERICKSON et al., 2012) que identifica diferenças metaproteômicas entre duplas de gêmeos dos quais um deles fora diagnosticado com DC e o outro não. Comparando os pares de gêmeos entre si, é possível desprezar a variabilidade genética como principal fator determinístico contribuindo para a doença. Também determinamos vias de sinalização e processos biológicos em potencial pelos quais genes de autofagia, um dos principais processos relacionados com o desenvolvimento

de DC, podem ser afetados pela ligação das proteínas microbianas aos receptores humanos. Utilizando testes de enriquecimento e informação *a priori*, fomos capazes de identificar cadeias sinalizadoras canônicas e não canônicas relacionadas a processos modulados unicamente pelas proteínas encontradas em indivíduos com DC.

Hipótese

"*A autofagia é modulada pelo microbioma intestinal. Essa modulação ocorre de maneiras diferentes entre indivíduos com doença de Crohn e indivíduos saudáveis.*"

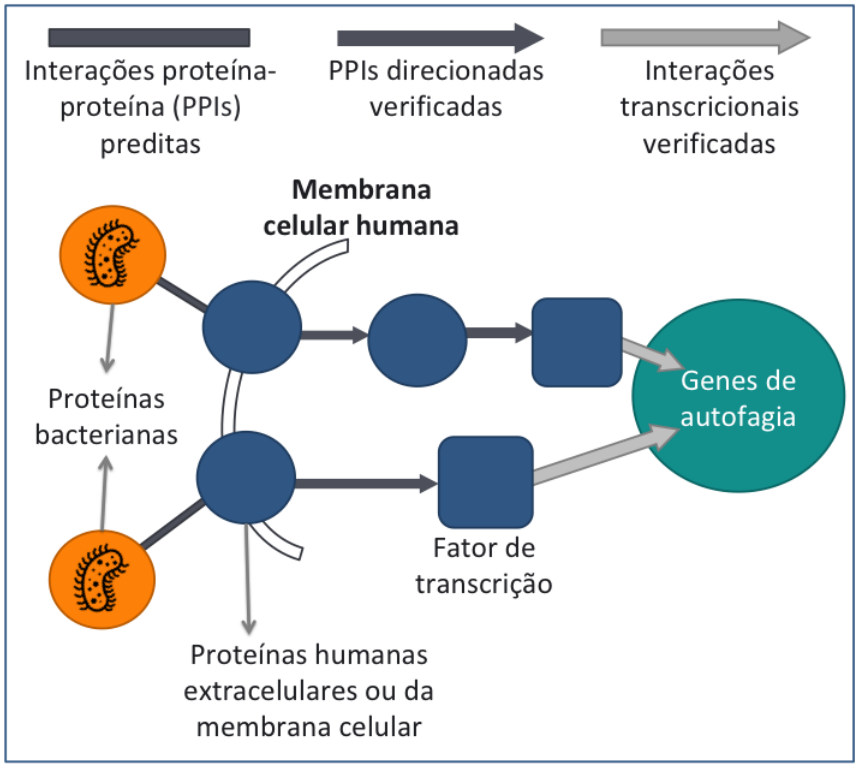

Figura 5: Sugerimos que as proteínas bacterianas interagem com proteínas receptoras humanas que provoca cascatas de sinalização pelas quais a expressão dos genes de autofagia são modulados.

"**Quais são as diferenças nos mecanismos moleculares pelos quais as proteínas dos indivíduos saudáveis e com doença de Crohn modulam a autofagia?**"

3 Objetivos

O principal objetivo dessa tese é identificar e caracterizar os possíveis mecanismos moleculares mediados pela microbiota disbiótica na Doença de Crohn. Para esse fim utilizamos vias de sinalização e redes de interação moduladas a partir de dados de metagenômica e metaproteômica de indivíduos com DC.

Objetivos

- analisar se e como as proteínas bacterianas de pacientes com DC podem modular a autofagia;
- identificar proteínas microbianas chaves que podem induzir ou impedir o desenvolvimento das DC;
- identificar microrganismos chave para a modulação diferencial da autofagia e de outros mecanismos moleculares associados à doença;
- estabelecer um protocolo computacional para análise de interação microbioma-hospedeiro a nível molecular utilizando uma abordagem de biologia de sistemas

4 Metodologia

A metodologia utilizada nessa tese foi desenvolvida pela autora com o auxílio de seus colaboradores do Instituto Earlham (Norwich, Inglaterra). A figura 6 apresenta um esquema geral do método desenvolvido e será citada ao longo do capítulo.

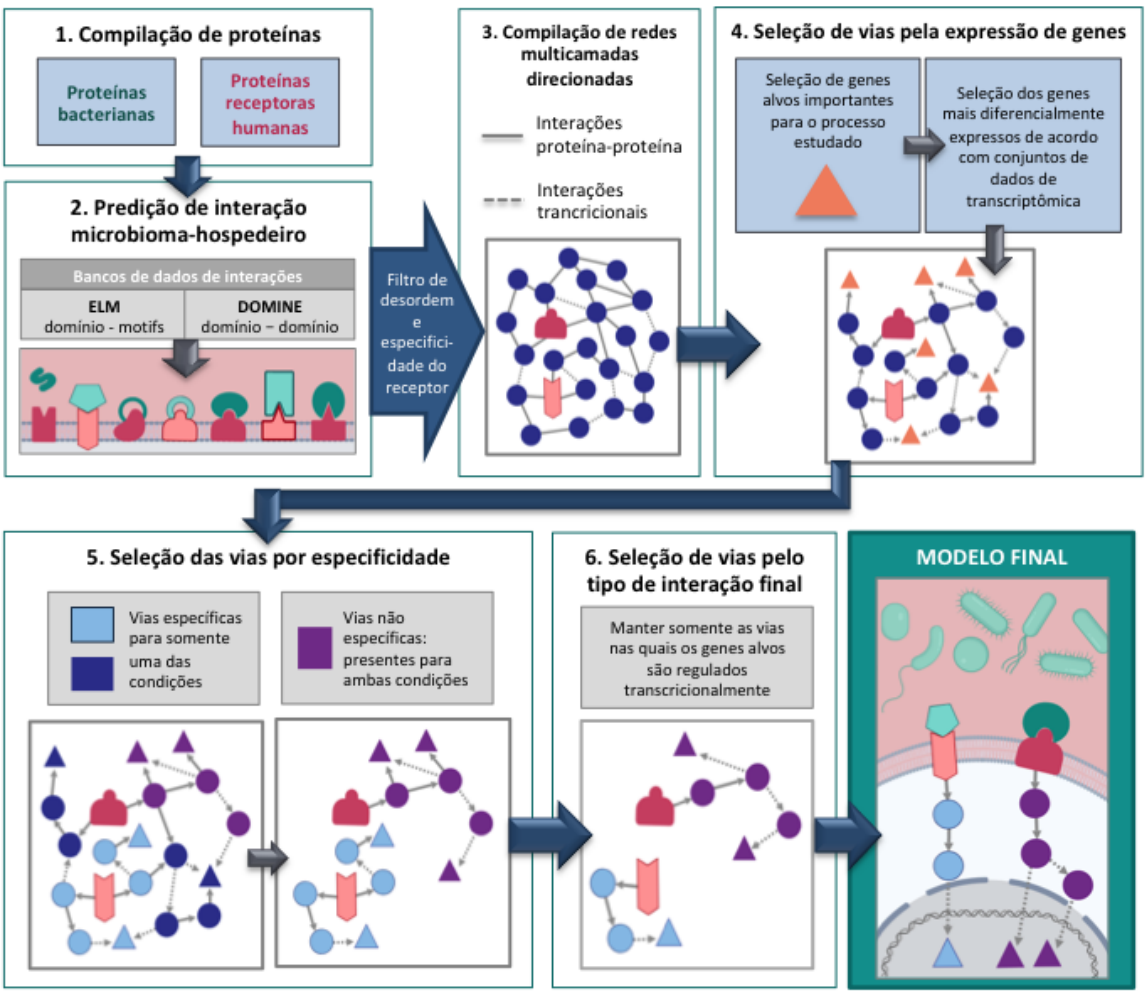

Figura 6: Esquema ilustrando as etapas do método desenvolvido. O método inicia com a compilação de proteínas bacterianas e receptores humanos, seguido da predição das interações entre eles. Em seguida, foi feito um filtro das interações, e, a partir do resultado obtido, foram compiladas redes com camadas de interações transcricionais e PPIs. Essas redes foram posteriormente filtradas pela seleção dos genes diferencialmente expressos, especificidade da rede e, finalmente, as últimas interações das vias para a obtenção do modelo final.

4.1 Interação microbioma-hospedeiro

A primeira parte do método consiste na busca de quais proteínas microbianas e proteínas humanas potencialmente interagem entre si. Para isso, primeiramente foram compiladas as proteínas bacterianas presentes unicamente em indivíduos com DC ou saudáveis e também as proteínas humanas mais prováveis de interagir com as bacterianas, de acordo com sua localização (Figura 6.1). Depois da compilação das proteínas, foi realizada a predição de quais proteínas bacterianas têm o potencial de interagir com quais proteínas humanas utilizando informações de interações domínio-motif disponíveis em bancos de dados (Figura 6.2).

Compilação de dados

- Proteínas bacterianas

A lista de proteínas bacterianas foi obtida do conjunto de dados disponibilizado por (ERICKSON et al., 2012).

Nesse estudo, Erickson e colaboradores extraíram o metaproteoma de amostras de fezes de seis pares de gêmeos monozigóticos com doença de Crohn ileal (IDC) ou colônica (DCC)[1]. Mais detalhes sobre as amostras do estudo estão apresentados na tabela 2.

As informações das proteínas bacterianas disponibilizadas por (ERICKSON et al., 2012) são anotações de códigos de domínios Pfam e de outras bases de dados de acordo com a similaridade de cada uma com proteínas disponíveis.

- Proteínas humanas

As proteínas humanas utilizadas para a predição de interação com as bacterianas foram as localizadas na matriz extracelular e membrana plasmática de acordo com os seguintes bancos de dados (Figura 7.2):

[1] Para a simpificação da análise, não consideramos as localizações das amostras, considerando ambas IDC e DCC como DC.

Tabela 2: Tabela contendo os detalhes dos pacientes dos quais foram obtitas as amostras do estudo de (ERICKSON et al., 2012).

Ano de nascimento	**Fenótipo**	**Sexo**	**Ano do diagnóstico**	**Cirurgia (ano)**
1951	Saudável	F	-	-
1951	Saudável	F	-	-
1947	CCD	M	41	-
1947	CCD	M	40	-
1962	ICD	F	23	1985
1962	ICD	F	24	1986
1953	ICD	M	23	1980
1953	ICD	M	23	1976
1954	ICD	F	20	1974
1954	Saudável	F	-	-
1953	ICD	M	20	1973
1953	Saudável	M	-	-

- ComPPI: contém informações da localização das proteínas na célula (VERES et al., 2015);
- MatrixDB: contém interações moleculares de componentes extracelulares (LAUNAY et al., 2015);
- Human Protein Atlas (HPA): contém informações das localizações das proteínas em órgãos, tecidos e células (UHLéN et al., 2015);

Essas as localizações foram selecionadas porque são onde as interações entre as proteínas bacterianas e as humanas são mais prováveis de acontecer.

Predição de interação microbioma-hospedeiro

Depois de compilar os conjuntos de proteínas humanas e bacterianas, foi feita a predição de quais estão propensas a interagir. Para essa predição, foram utilizadas informações referência de interações domínio-domínio e domínio-motifs obtidas nos bancos de dados DOMINE (YELLABOINA et al., 2011) e ELM (GOUW et al., 2018), respectivamente (Figura 6.2 e Figura 7.3). Esses bancos de dados possuem listas de interações entre domínio-domínio e domínio-motif já conhecidas na literatura e são considerados o repositório de padrão ouro

para esse tipo de dados. A comparação dos domínios e motifs coletados com os de referência foi feita por códigos desenvolvidos *in-house* na linguagem de programação Python.

Depois de obter os pares putativos, foram aplicados dois filtros para selecionar as interações de maior relevância. O primeiro trata-se de um filtro estrutural e visa excluir interações das quais o motif localiza-se dentro de domínios globulares ou dentro de regiões desordenadas com o intuito de reduzir o número de predições falso-positivas. Os bancos de dados utilizados para a obtenção dessas informações foram IUPRED (DOSZTáNYI et al., 2005), PFAM (EL-GEBALI et al., 2019) e InterPro (MITCHELL et al., 2019) (Figura 7.4.1). O segundo filtro refere-se à especificidade do receptor: foram mantidos somente os receptores que interagem com somente uma das condições (saudável ou DC) (Figura 7.4.2).

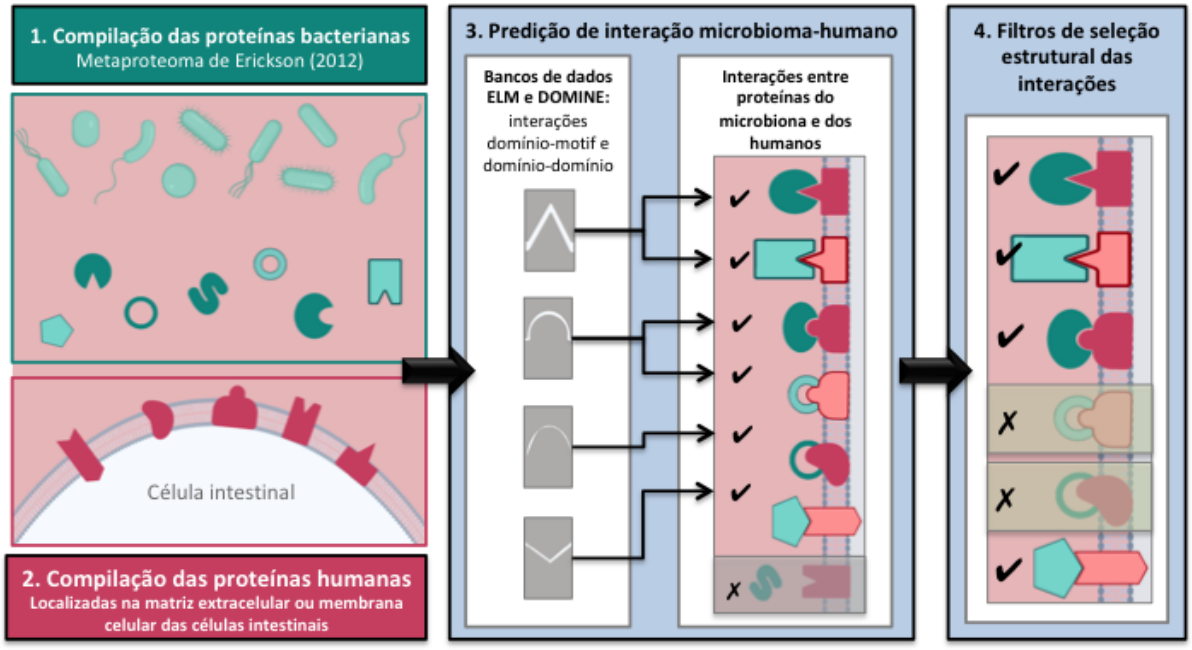

Figura 7: Figura ilustrativa do método de predição de interação entre proteínas. Depois da predição das proteínas bacterianas e receptores humanos, utilizou-se os bancos de dados ELM e DOMINE para consultar possíveis interações domínio-motif e domínio-domínio que podem estar presentes nas proteínas compiladas. Posterior à predição, realizou-se um filtro que excluiu interações pouco prováveis de ocorrer de acordo com sua estrutura.

4.2 Compilação da rede do hospedeiro

Nesse ponto do processo temos uma lista de possíveis PPIs entre as proteínas bacterianas e as receptoras humanas das condições de doença de Crohn e saudáveis. Para avaliar como essas proteínas bacterianas podem potencialmente modular os processos biológicos, foram compiladas redes diretas[2] de interação proteína-proteínas (PPIs) e interações transcricionais (TRIs) a partir dos receptores humanos presentes nessa lista (Figura 6.3). As PPIs e TRIs foram obtidas dos bancos de dados OmniPath (TüREI; KORCSMáROS; SAEZ-RODRIGUEZ, 2016), HTRI (BOVOLENTA; ACENCIO; LEMKE, 2012) e TRRUST (HAN et al., 2018). Esses bancos de dados contém vias de sinalização com PPIs e/ou TRIs manualmente curadas.

Depois de obtidas as redes, foram selecionadas vias de sinalização que iniciam nos receptores humanos e chegam até uma das 38 proteínas de autofagia (Figura 3) (KUBISCH et al., 2013) (Figura 6.4). A autofagia foi utilizada como principal processo biológico da rede por ser um dos processos mais modificados na doença de Crohn, por ter alto impacto em outros processos relacionados à doença e porque apresenta diversos genes de risco característicos da DC.

Foram mantidas somente vias de sinalização com até quatro proteínas de comprimento. A utilização desse critério foi importante para que possamos selecionar vias biologicamente relevantes, ou seja que contenham interações que causem impacto no funcionamento do organismo. As vias consideradas iniciam com os receptores na superfície ou fora da célula, inclui genes no citoplasma interagindo fisicamente e finaliza com genes nucleares. Um número inferior de interações não seria biologicamente plausível.

Para selecionar as vias mais biologicamente relevantes no contexto da doença de Crohn, a rede foi submetida a algumas etapas de filtragem.

[2] Uma rede direta é composta de interações diretas. São consideradas interações diretas quando a direção da interação entre dois nós é conhecida (JUNKER; SCHREIBER, 2008)

Genes diferencialmente expressos

Primeiramente, foram buscados quais genes de autofagia eram mais diferencialmente expressos entre indivíduos com DC e saudáveis. Para essa análise, foram selecionados três conjuntos de dados de transcriptômica que contém dados de expressão diferencial entre genes de indivíduos saudáveis *versus* indivíduos com DC no banco de dados de transcriptômica GEO (BARRETT et al., 2013). Para o processamento dos dados e cálculo do *fold change* (logFC) para a obtenção de valores de expressão diferencial, foi utilizado o programa GEO2R (BARRETT et al., 2013) com os parâmetros pré-definidos.

A partir dos valores obtidos pelo GEO2R, foram selecionados os genes de autofagia mais diferencialmente expressos de acordo com os seguintes critérios:

- *adjacent P value* maior do que 0.05;
- logFC maior do que 0.2 ou menor do que -0.2;
- devem ser diferencialmente expressos (de acordo com o critério 2) e com a mesma tendência em pelo menos 2 conjuntos de dados

Depois de selecionados os genes mais diferencialmente expressos, foram mantidas na rede somente as vias que atingiam esses genes (Figura 6.4).

Especificidade da rede

Na próxima etapa, manteve-se as vias específicas no máximo de camadas possível. Ou seja, vias que tinham nós compartilhados entre as duas condições saudável e com DC foram excluídas (Figura 6.5).

Genes regulados transcricionalmente

Na próxima etapa de filtragem, manteve-se somente as vias cujos genes de autofagia eram regulados transcricionalmente. Para este intuito, observou-se quais eram os tipos de

interação entre os dois últimos elementos das vias. Foram mantidos somente as vias que finalizaram com TRIs e excluídas as que finalizaram com PPIs (Figura 6.6). Utilizamos esse critério porque o intuito da análise é obter a informação a respeito da regulação dos genes de autofagia, que estão na última camada, não obstante, utilizamos dados de transcriptômica para a seleção desses genes.

4.3 Análise de ontologia de genes da rede

Neste ponto da metodologia, já temos a rede de sinalização que inicia com as proteínas de bactérias e finaliza com a regulação transcricional dos genes de autofagia. No intuito de observar quais eram os principais processos biológicos envolvidos na rede além da autofagia, foi feita uma análise de ontologia de gene (GO) de processos biológicos da última rede obtida. O aplicativo do ClueGO v. 2.5.2 (BINDEA et al., 2009) do programa Cytoscape foi utilizado para executar essa etapa. Foi selecionado um corte de p *Value* maior do que 0.05 e foram executadas simulações com três níveis de especificidade de rede: médio, detalhado e entre médio/detalhado. Os outros parâmetros foram mantidos como padrão do programa.

4.4 Análise taxonômica putativa

Para obter informações sobre a potencial origem taxonômica das proteínas bacterianas presentes na rede final, foram utilizadas informações de (WILLING et al., 2010), que executou uma análise taxonômica por 16S rRNA das mesmas amostras de (ERICKSON et al., 2012). Os dados deste estudo fornecem informações de quais espécies estavam presentes nas amostras e quais eram únicas de indivíduos saudáveis e com DC. As informações taxonômicas de (WILLING et al., 2010) foram então cruzadas com as informações funcionais de (ERICKSON et al., 2012) para verificar quais domínios das proteínas bacterianas que podem ocorrer em quais táxons presentes nas amostras. Os dados sobre quais domínios são encontrados em quais táxons foram obtidos no banco de dados PFAM (EL-GEBALI et al., 2019), que tem a informação sobre a ocorrência de domínios e famílias de proteínas em espécies e táxons

superiores.

5 Resultados

5.1 Predição da interação de proteínas microbioma-hospedeiro

Primeiramente foram obtidas proteínas bacterianas diferencialmente expressas presentes em indivíduos saudáveis e com DC do estudo (ERICKSON et al., 2012). Foram selecionadas somente as proteínas diferencialmente expressas entre as duas condições. No total, foram obtidas 1221 proteínas microbianas exclusivas de amostras saudáveis e 865 proteínas microbianas encontradas em amostras de pacientes com DC.

A seguinte etapa foi a compilação de proteínas humanas localizadas na matriz extracelular e membrana celular. Um total de 3601 proteínas humanas foram compiladas.

Depois, foi executada a predição de interações entre proteínas humanas e bacterianas compiladas nas etapas anteriores. Um total de 1921 interações foram preditas: 433 de proteínas bacterianas de indivíduos saudáveis e 1488 de indivíduos com DC. A informação sobre essas interações e seus motifs respectivos podem ser encontradas nos Resultados Suplementares, aba n^o 1.

5.2 Rede de sinalização

A próxima etapa foi a compilação de redes de sinalização de bancos de dados com informação de PPIs e TRIs. Foram mantidas somente as interações direcionadas. No total, foram obtidas 61756 interações PPIs do banco de dados OmniPath, e 60322 TRIs dos bancos de dados HTRI e TRRUST. Das redes compiladas, foram selecionadas vias que iniciam com os receptores que interagem com proteínas microbianas e finaliza com as proteínas de autofagia selecionadas. Isso resultou em 125506 cadeias de proteínas (com no máximo 4 proteínas), 13221 interações e 2212 nós no total (Figura 8).

Como as vias apresentam tamanho máximo de até 4 proteínas, pode-se observar a

organização da rede em uma estrutura de 5 camadas que inicia nas proteínas bacterianas e chega até os genes de autofagia. A primeira camada é constituída de proteínas bacterianas que interagem com os receptores, presentes na segunda camada. Os receptores estão conectados a proteínas intermediárias (3ª camada e 4ª camada) que os conectam aos genes de autofagia, na 5ª camada, como ilustrado na Figura 8.

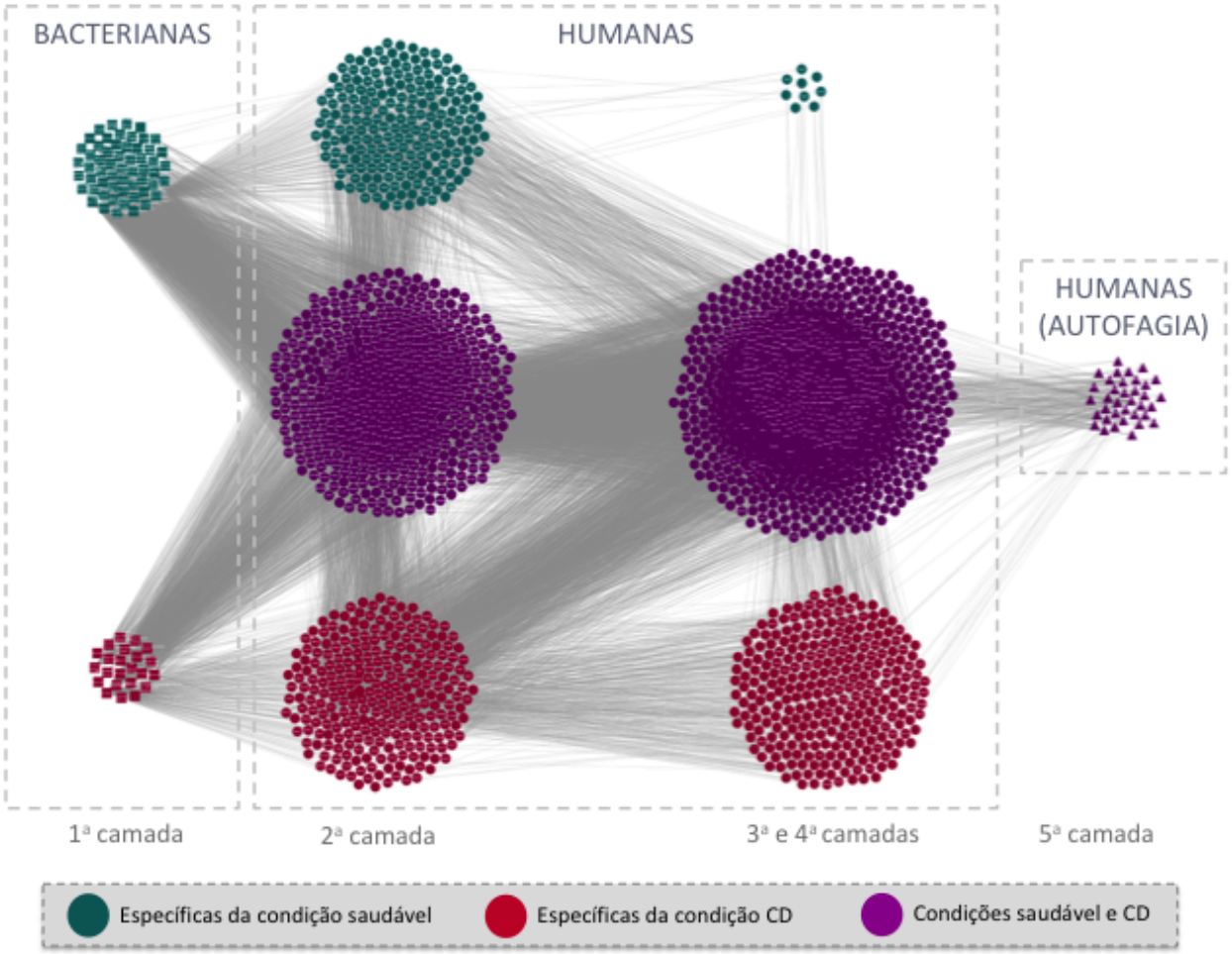

Figura 8: Primeira rede de interação obtida. Cada nó em forma de círculo e quadrado (1^a até 4^a camadas) representam proteínas. Os nós em formato de triângulo (5^a camada) representam genes de autofagia.

Filtro por expressão dos genes de autofagia

Para manter somente as vias biologicamente relevantes no contexto do desenvolvimento das DC, foram selecionados 3 conjuntos de dados de transcriptômica que contêm informações sobre genes de autofagia diferencialmente expressos entre pacientes saudáveis e com DC. Esses conjuntos de dados foram obtidos pelo banco de dados GEO, e apresentam código GEO GSE36807 (MONTERO-MELéNDEZ et al., 2013), GSE75214 (VANCAMELBEKE et al., 2017) e GSE9686 (CAREY et al., 2008). Os detalhes sobre as amostras utilizadas nesses estudos está apresentado na tabela 3. A seleção desses conjuntos de dados foi feita levando em conta qualidade dos dados presentes. Para essa avaliação foram gerados gráficos de distribuição de logFC *versus* valor P para avaliar a distribuição dos dados (Figura 9).

Tabela 3: Tabela contendo os detalhes dos pacientes dos quais foram obtitas as amostras dos conjuntos de dados obtidos pelo GEO.

		CD	**Controle**
GSE36807	Feminino	4	7 (F+M)
	Masculino	9	
	Faixa de idades	31-60 (n=10), <=30 (n=1), >60 (n=2)	–
GSE7214	Feminino	31 (ICD), 6 (CCD)	11
	Masculino	20 (ICD), 2 (CCD)	11
	Faixa de idades	29-54 (ICD), 34-44 (CCD)	52-73
GSE9686	Feminino	8	9
	Masculino	9	11
	Faixa de idades	5,4-17,3	5,7-18,1

Os genes de autofagia e seus respectivos valores em cada conjunto de dados estão apresentados na Tabela 4. De acordo com os critérios descritos na seção de Metodologia, foram selecionados 5 genes de autofagia do total de 38: ATG4D, ATG7, LC3A, LC3B e WIPI1.

Filtro por especificidade das vias

Observa-se na rede apresentada na figura 8 que muitos nós de proteínas humanas compartilham conexão de proteínas bacterianas de ambas as condições DC e saudáveis. Para

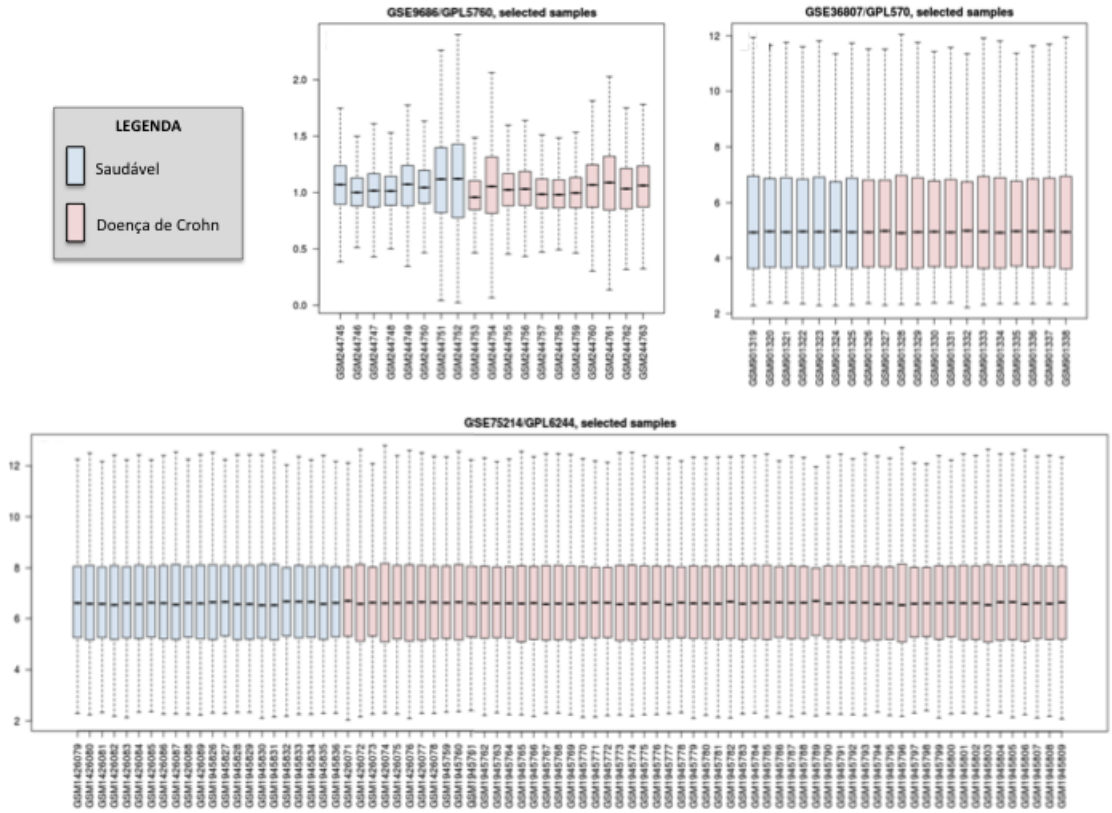

Figura 9: Gráficos de distribuição dos valores de expressão de cada amostra dos transcriptomas obtidos.

que a rede apresente maior especificidade possível, ela foi submetida à etapa de filtragem pela especificidade dos seus componentes. Primeiramente, filtramos os receptores. A figura 10 apresenta a rede depois da exclusão das vias com receptores que continham interações com ambas proteínas bacterianas específicas de DC e de saudável. Essa também apresenta somente as vias que finalizam com os 5 genes selecionados na etapa anterior.

Em seguida, foi realizada a filtragem das proteínas das camadas intermediárias (3ª, 4ª e 5ª). Foram filtradas as vias que continham proteínas da 3ª camada conectadas a ambos receptores de DC e de saudável. Entretanto, na 4ª e 5ª camada da rede não há nós específicos, portanto, foram mantidos os nós compartilhados. Depois dessa filtragem, a rede totalizou 65 nós e 102 conexões (59 PPIs, 3 PPIs + TRIs e 40 TRIs) (Figura 12).

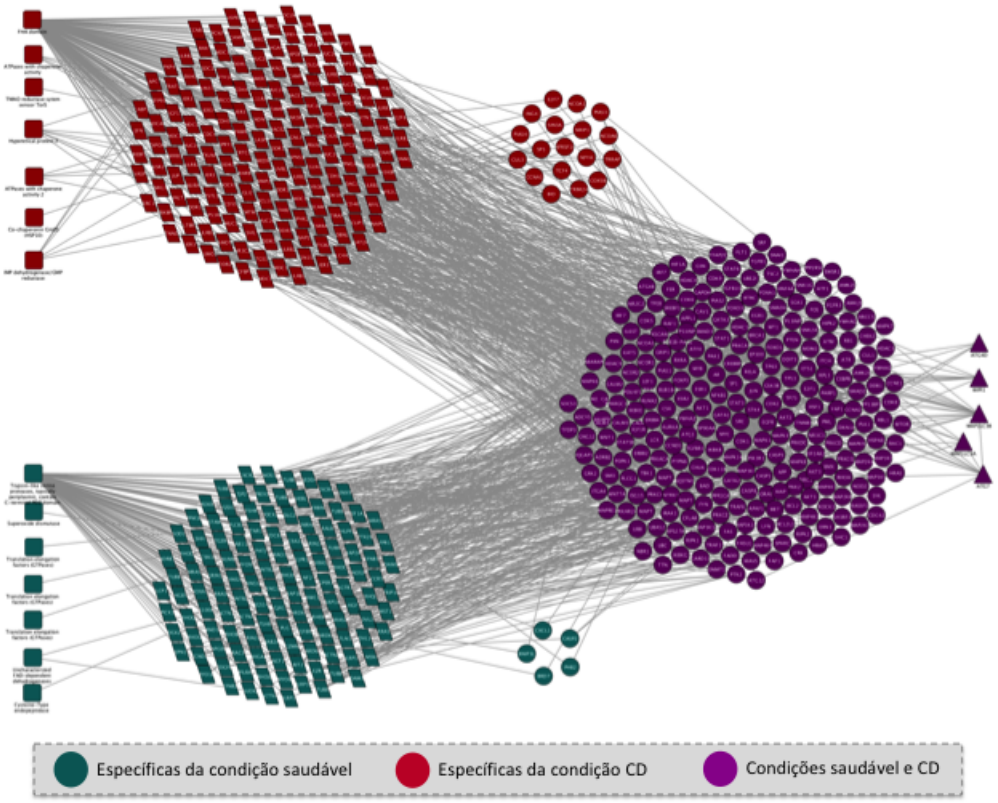

Figura 10: Imagem da segunda rede obtida depois da filtragem de especificidade dos nós de proteínas receptoras humanas e de proteínas de autofagia diferencialmente expressas.

Filtro por regulação transcricional dos genes de autofagia

Em seguida, foram excluídas as vias que apresentavam interações PPIs entre a 4ª e 5ª camada. Deste modo, a 4ª camada passou a ser composta de fatores de transcrição.

Dos 65 nós da rede da figura **??**), 6 são específicos da condição saudável, incluindo uma proteína bacteriana: *Tripsin-like serine protease*. 9 nós localizados na 4ª e 5ª camadas não são específicos. Eles incluem os fatores de treanscrição GATA1, AR, ATF1, GATA3 e FOXP3 e os genes de autofagia ATG4D, ATG7, LC3B and WIPI1. Os outros 50 nós são proteínas específicas de DC. Dentre essas, estão inclusas as proteínas bacterianas ATPases com atividade de chaperona, proteínas *Cold Shock*, IMP desidrogenase/GMP reductase e domínio FHA.

5.3 Análise de ontologia de genes

Nós utilizamos o aplicativo do Cytoscape ClueGO para executar a análise de ontologia de genes da rede nos níveis Médio, Detalhado e entre os dois na opção de especificidade do

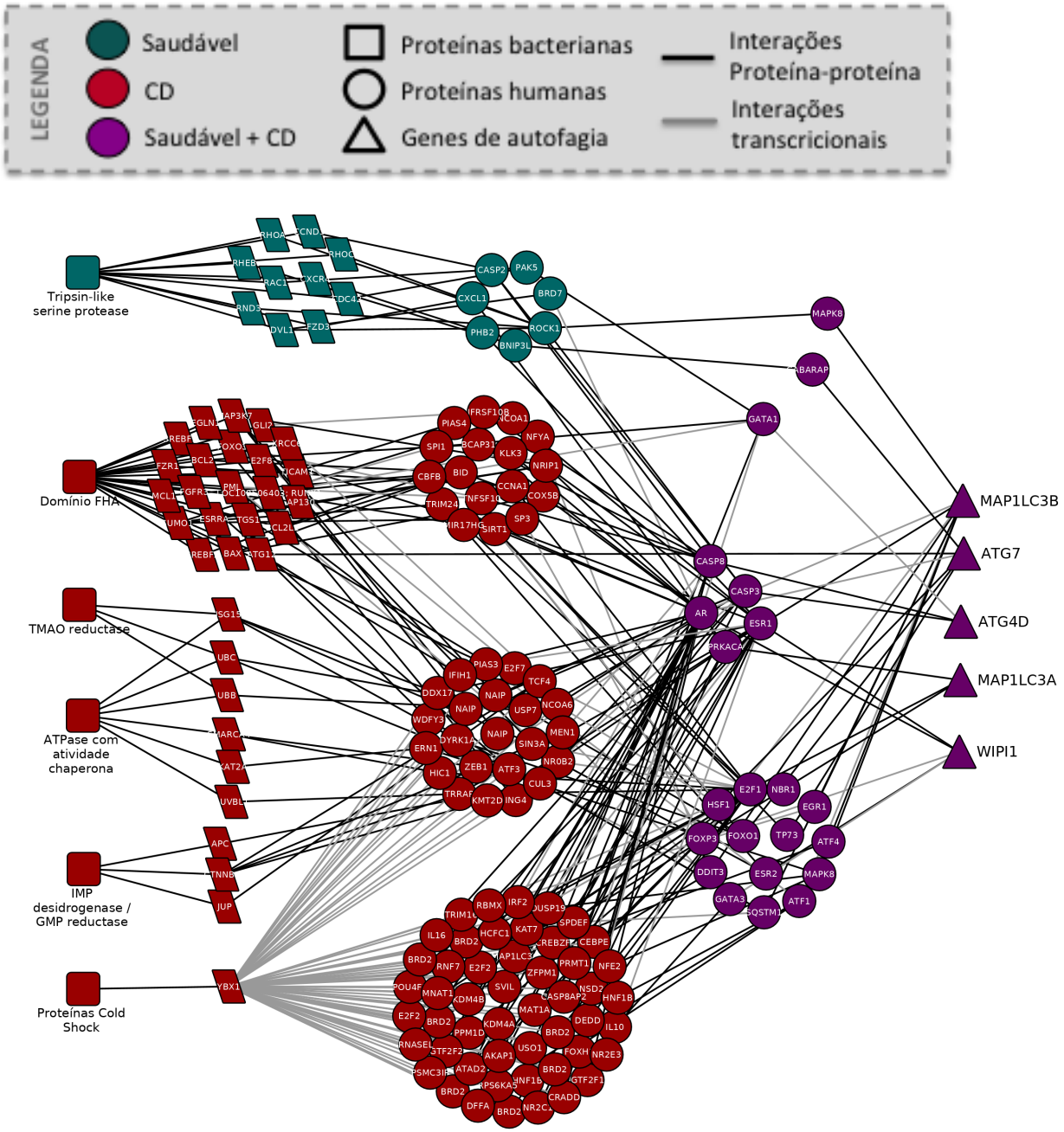

Figura 11: Imagem da rede obtida depois da filtragem de especificidade dos nós de proteínas da 3ª camada.

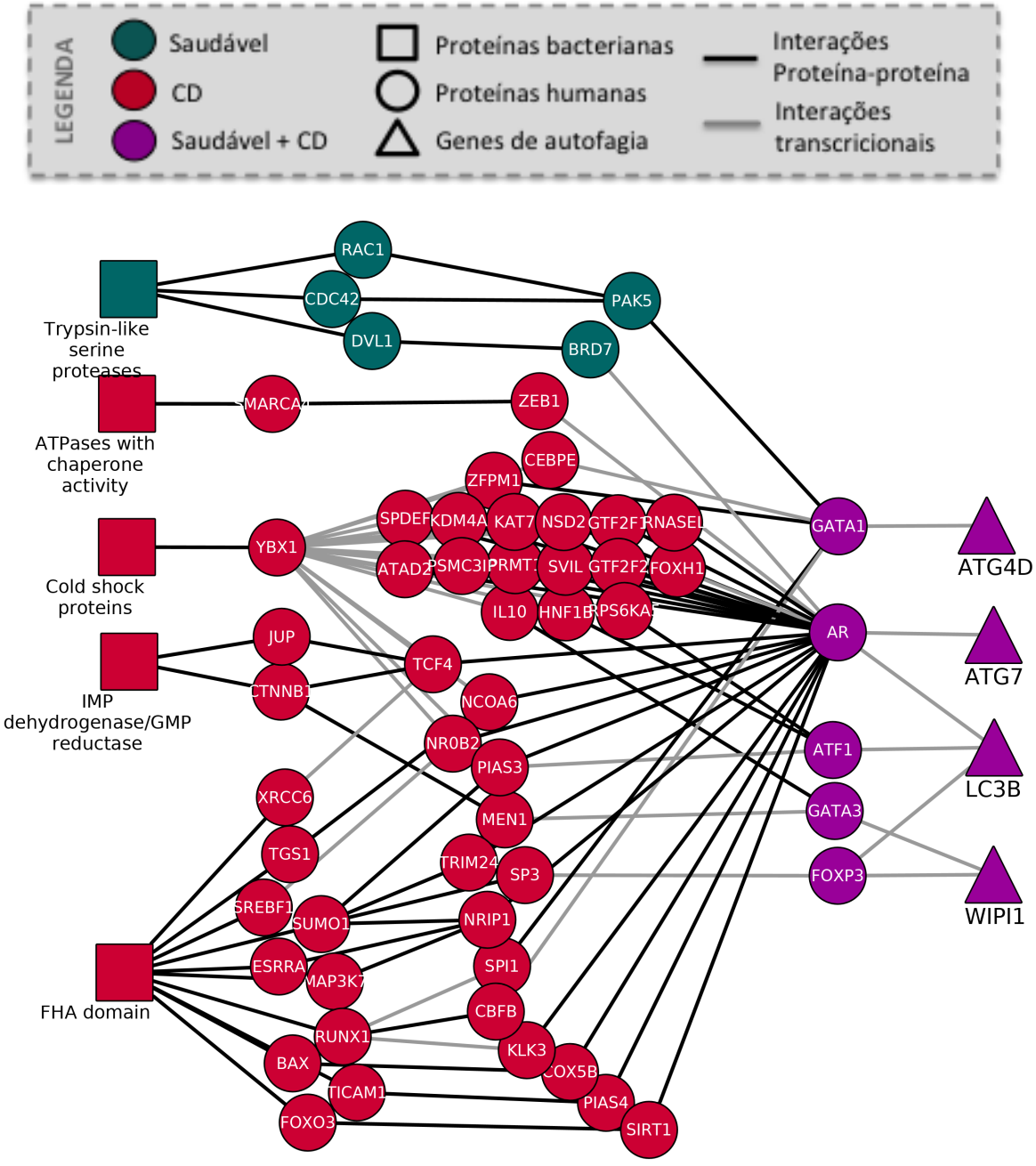

Figura 12: Imagem da rede obtida depois da filtragem de interações entre a entre a 4ª e 5ª camada.

aplicativo. Os três níveis geraram um total de 131 processos biológicos que estão mostrados na tabela 5. 40 processos foram classificados como importantes para o desenvolvimento de DC. A tabela completa com todos os processos identificados está nos Resultados Suplementares, aba n^o 2.

5.4 Análise taxonômica

Foi utilizado o banco de dados PFAM para obter informações sobre quais proteínas bacterianas encontradas em amostras DC e saudáveis por (ERICKSON et al., 2012) podem ser originadas de quais bactérias encontradas por (WILLING et al., 2010) nas mesmas amostras. Willing encontrou 16 gêneros bacterianos únicos de saudáveis e DC, incluindo 6 gêneros não identificados de 5 filos. Destes, 7 eram únicos de saudáveis e 9 de DC (Tabela 6 e Resultados Suplementares, aba n^o 3).

Quando foi consultado quais domínios bacterianos ocorrem em quais bactérias presentes nas amostras, encontramos que *Trypsin-like serine proteases*, a proteína bacteriana única de amostras saudáveis, pode ser originada de gêneros não identificados das famílias Lachnospiraceae e Porphyromonadaceae e de famílias não identificadas das ordens Clostridiales and Bacteroidales. A respeito das proteínas bacterianas únicas de DC, as bactérias *Aeromonas* e um gênero não identificado de Enterobacteriaceae são origens potenciais de todas as 4 proteínas bacterianas selecionadas: ATPases com atividade chaperona, proteínas *Cold shock*, IMP desidrogenase/GMP reductase e proteínas com domínio FHA. Além desses gêneros, *Lactobacillus*, *Citrobacter* e *Shigella* são origens putativas de ATPases com atividade chaperona, proteínas *Cold Shock* e MP desidrogenase/GMP reductase. Bactérias do gênero *Fusobacterium* podem ser a origem em potencial de proteínas de domínio FHA (Tabela 6).

Tabela 4: Tabela com os valores de expressão diferencial de cada gene de autofagia nos conjuntos de dados de transcriptoma de indivíduos com doença de Crohn e saudáveis. Valores acima de 0.2 representam super-expressão em DC, e valores abaixo de 0.2 representam sub-expressão em DC. Os valores selecionados em verde representam os genes que cumpriram todos os critérios citados na sessão de métodos, portanto foram selecionados.

Gene	logFC do GSE36807	logFC do GSE75214	logFC do GSE9686
AMBRA1	0.109	-0.12205879	-0.06580779
ATG10	0.03799999999999999	-0.25827022	0.25732727
ATG101	0.115	0.12410138	0.02396701
ATG12	-0.017	0.06408043	-0.13592089
ATG13	NA	NA	0.1329875
ATG14	-0.241	NA	-0.08465198
ATG16L1	NA	-0.1141205	-0.20621658
ATG16L2	-0.120	NA	-0.21900987
ATG2A	0.213	0.01977794	-0.03119577
ATG2B	0.1372	0.11136332000000002	-0.14026669
ATG3	NA	-0.28323893	0.24402129
ATG4A	0.293	-0.09962228	-0.02781534
ATG4B	-0.237	0.075645625	NA
ATG4C	-0.15	0.0637247	NA
ATG4D	NA	-0.260442	-0.38448015
ATG5	-0.09944999999999998	NA	-0.00696115
ATG7	0.16	0.21075753	0.21176386
ATG9A	0.361	-0.1623623	-0.15990263
BECN1	NA	-0.13805106	-0.01729438
FYCO1	0.006500000000000006	0.04576568	0.19988549
GABARAP	NA	NA	0.12573248
GABARAPL1	NA	-0.2310241	0.01513399
GABARAPL2	-0.148	NA	-0.10288038
LC3A	-0.212	-0.19826496	-0.235831285
LC3B	0.248	0.066244585	0.29480738
LC3B2	NA	0.16618842	NA
LC3C	NA	-0.10603317	-0.24714824
PIK3C3	-0.21700000000000003	-0.1878576	0.16484507
PIK3R4	NA	-0.10192457	0.12888458
RB1CC1	-0.10300000000000001	NA	0.08459826
SQSTM1	0.107	-0.09995885	-0.10989285
TECPR1	0.181	NA	0.14393926
ULK1	NA	0.05942704	-0.07635967
ULK2	-0.26949999999999996	-0.0575119	0.13351061
UVRAG	NA	0.04312765	0.2856167
WIPI1	0.279	0.2463333	0.23778814
WIPI2	-0.173	-0.00323564	-0.11067068
LEGENDA:	Valores -0.2 >0.2	Valores -0.2 <0.2	Genes selecionados

Tabela 5: Tabela com os processos biológicos identificados na rede pela análise de ontologia de genes.

GO ID	Termo GO (em inglês)	Genes associados	Nível de especificidade ClueGO
GO:0000422	autophagy of mitochondrion	ATG7, MAP1LC3B, SREBF1	Medium, Between Medium/Detailed
GO:0061726	mitochondrion disassembly	ATG7, MAP1LC3B, SREBF1	Medium
GO:0039519	modulation by virus of host autophagy	ATG7	Detailed
GO:0039521	suppression by virus of host autophagy	ATG7	Detailed
GO:0008635	activation of cysteine-type endopeptidase activity involved in apoptotic process by cytochrome c	BAX	Detailed
GO:0001783	B cell apoptotic process	BAX, IL10	Between Medium/Detailed
GO:1990117	B cell receptor apoptotic signaling pathway	BAX	Detailed
GO:2001237	negative regulation of extrinsic apoptotic signaling pathway	AR, GATA1, MEN1	Medium
GO:0051402	neuron apoptotic process	BAX, CTNNB1, FOXO3, NCOA6	Medium
GO:0002904	positive regulation of B cell apoptotic process	IL10	Detailed
GO:0043525	positive regulation of neuron apoptotic process	BAX, CTNNB1, FOXO3, NCOA6	Medium, Between Medium/Detailed
GO:0043523	regulation of neuron apoptotic process	BAX, CTNNB1, FOXO3, NCOA6	Medium, Between Medium/Detailed
GO:2000320	negative regulation of T-helper 17 cell differentiation	FOXP3	Detailed
GO:2000317	negative regulation of T-helper 17 type immune response	FOXP3	Detailed
GO:0002709	regulation of T cell mediated immunity	FOXP3, MAP3K7	Between Medium/Detailed
GO:2000319	regulation of T-helper 17 cell differentiation	FOXP3	Detailed
GO:2000316	regulation of T-helper 17 type immune response	FOXP3	Detailed
GO:0002456	T cell mediated immunity	FOXP3, GTF2F1, GTF2F2, MAP3K7	Medium
GO:0002250	adaptive immune response	FOXP3, GTF2F1, GTF2F2, MAP3K7, SIRT1	Medium
GO:0002460	adaptive immune response based on somatic recombination of immune receptors built from immunoglobulin superfamily domains	FOXP3, GTF2F1, GTF2F2, MAP3K7	Medium
GO:0050663	cytokine secretion	FOXP3, GATA3, GTF2F1, GTF2F2, IL10, NCOA6	Medium
GO:0032623	interleukin-2 production	FOXP3, MAP3K7, RUNX1	Medium
GO:0060302	negative regulation of cytokine activity	IL10	Detailed
GO:0045085	negative regulation of interleukin-2 biosynthetic process	FOXP3	Detailed
GO:0045403	negative regulation of interleukin-4 biosynthetic process	ZFPM1	Detailed
GO:0032713	negative regulation of interleukin-4 production	FOXP3, ZFPM1	Between Medium/Detailed
GO:0060300	regulation of cytokine activity	IL10	Detailed
GO:0002718	regulation of cytokine production involved in immune response	FOXP3, IL10, MAP3K7	Medium, Between Medium/Detailed
GO:0032663	regulation of interleukin-2 production	FOXP3, MAP3K7, RUNX1	Medium
GO:0060318	definitive erythrocyte differentiation	GATA1, ZFPM1	Between Medium/Detailed, Detailed
GO:0030218	erythrocyte differentiation	FOXO3, GATA1, PRMT1, SPI1, ZFPM1	Medium, Between Medium/Detailed
GO:0034101	erythrocyte homeostasis	FOXO3, GATA1, PRMT1, SPI1, ZFPM1	Medium, Between Medium/Detailed
GO:0030851	granulocyte differentiation	CEBPE, GATA1, RUNX1	Medium, Between Medium/Detailed
GO:0030099	myeloid cell differentiation	CEBPE, FOXO3, GATA1, NCOA6, PRMT1, RUNX1, SPI1, ZFPM1	Medium
GO:0002262	myeloid cell homeostasis	FOXO3, GATA1, PRMT1, SPI1, ZFPM1	Medium
GO:0045648	positive regulation of erythrocyte differentiation	FOXO3, GATA1, PRMT1	Medium, Between Medium/Detailed
GO:0045639	positive regulation of myeloid cell differentiation	FOXO3, GATA1, PRMT1, RUNX1	Medium, Between Medium/Detailed
GO:0010724	regulation of definitive erythrocyte differentiation	GATA1, ZFPM1	Between Medium/Detailed, Detailed
GO:0045646	regulation of erythrocyte differentiation	FOXO3, GATA1, PRMT1, SPI1, ZFPM1	Medium, Between Medium/Detailed
GO:0045637	regulation of myeloid cell differentiation	FOXO3, GATA1, PRMT1, RUNX1, SPI1, ZFPM1	Medium
GO:0048872	homeostasis of number of cells	FOXO3, GATA1, PRMT1, SPI1, ZFPM1	Medium
GO:0010811	positive regulation of cell-substrate adhesion	CDC42, JUP, RAC1	Medium

Grupos de processos	Genes
Mitocôndria e autofagia	ATG7, MAP1LC3B, SREBF1
Apoptose	BAX, IL10, GATA1, MEN1, CTNNB1, FOXO3, NCOA6
Células T	FOXP3, MAP3K7, GTF2F1, GTF2F2
Interleucinas e citocinas	FOXP3, MAP3K7, GTF2F1, GTF2F2, SIRT1, GATA3, IL10, NCOA6, RUNX1, ZFPM
Células imunológicas	GATA1, ZFPM1, FOXO3, PRMT1, SPI1, CEBPE
Outros	FOXO3, GATA1, PRMT1, SPI1, ZFPM1, CDC42, JUP, RAC1

Tabela 6: Tabela com os genes bacterianos e suas respectivas origens putativas.

Nome da proteína (Erickson et al. 2012)	**Nome e código PFAM**	**Condição**	**Táxons de origem potencial**	**Nível taxonômico**
Trypsin-like serine proteases	PDZ domain (PF00595)	Saudável	unidentified Porphyromonadaceae unidentified Lachnospiraceae unidentified Clostridiales unidentified Bacteroidales	Gênero Gênero Família Família
ATPases with chaperone activity	Clp_N (PF02861)	Doença de Crohn	*Lactobacillus* *Aeromonas* *Citrobacter* *Shigella* unidentified Enterobacteriaceae	Gênero Gênero Gênero Gênero Gênero
Cold Shock proteins	Cold shock domain (PF00313)	Doença de Crohn	*Lactobacillus* *Aeromonas* *Citrobacter* *Shigella* unidentified Enterobacteriaceae	Gênero Gênero Gênero Gênero Gênero
IMP dehydrogenase / GMP reductase	CBS domain (PF00571)	Doença de Crohn	*Lactobacillus* *Aeromonas* *Citrobacter* *Shigella* unidentified Enterobacteriaceae	Gênero Gênero Gênero Gênero Gênero
FHA domain	(PF00498)	Doença de Crohn	*Fusobacterium* *Aeromonas* unidentified Enterobacteriaceae	Gênero Gênero Gênero

6 Discussão

6.1 Análise da rede

Pode-se observar na rede da Figura **??**, vias de sinalização únicas para as condições saudável e DC por onde os genes de autofagia podem ser modulados. Um dos genes que mais se destacam é ATG4D: de acordo com a Tabela 4, observamos que é o único gene de autofagia sub-regulado na condição de DC. É esperado que os genes da maquinaria de autofagia estejam super-regulados nas amostras de DC, como é o caso de ATG7, LC3B e WIPI1, pois os tecidos de DC estão em condições de estresse, como a presença de bactérias patógenas e estresse oxidativo. Deste modo supõe-se que a expressão não esperada de ATG4D esteja sendo modulada diferencialmente pelas proteínas bacterianas das diferentes condições.

A família ATG4 é formada por proteínas que vão de ATG4A-D que desempenham papel importante na formação do autofagossomo, sendo responsáveis por primar e delipidar proteínas da família ATG8 (KAUFFMAN et al., 2018). Especificamente, ATG4D desempenha papéis importantes em processos mitocondriais, como produção de espécies reativas de oxigênio (EROS) e mitofagia (BETIN et al., 2012). A mitofagia pode ser induzida em casos de dano mitocondrial, principalmente quando os níveis de EROS está alto, e permite que as mitocôndrias danificadas sejam degradadas (HAMACHER-BRADY; BRADY, 2016). Na doença de Crohn, há evidências sobre a presença de mitocôndrias modificadas, EROs elevados e deficiência de produção de energia em células de pacientes com DC (AVIELLO; KNAUS, 2017), sugerindo que a mitofagia pode não estar funcionando apropriadamente nessas células e a sub-reulação da ATG4D na condição de DC pode explicar a disfunção da mitofagia.

A falha na reciclagem da mitocôndria gera mais EROS que aumentam o estado inflamatório da DC (AVIELLO; KNAUS, 2017). Durante a inflamação, os mecanismos de sobrevivência e morte celular são regulados de acordo com a severidade do caso. Quando somente a degradação de micro-organismos ou organelas podem solucionar o problema, a sinalização de sobrevivência é ativada e a autofagia é induzida. Entretanto, em um estado

inflamado onde as células apresentam condições extremas e irreversíveis, a autofagia é inibida, a apoptose é ativada e a célula é levada à morte. Uma das proteínas responsáveis pela regulação entre autofagia e apoptose é ATG4D (BETIN; LANE, 2009b; MESSER, 2017). Na forma clivada, ATG4D apresenta função pró-autofágica ativando a proteína GABARAP-L1 (*γ-aminobutyric acid receptor-associated protein-like 1*), que participa na formação do autofagossomo. Entretanto, quando a ATG4D clivada está super-expressa, ela induz a apoptose. Essa indução acontece a partir da interação entre o domínio BH3 (*Bcl-2 homology domain 3*) de ATG4D e membros da família BCL2 (BETIN; LANE, 2009a). Essa família de proteínas contém proteínas anti-apoptóticas, como BCL-2 e BCL-XL e pró-apoptóticas como BAX, BAD e PUMA que interagem umas com as outras pra suprimir ou ativar a morte celular apoptótica (KALE; OSTERLUND; ANDREWS, 2018).

De acordo com a nossa predição, a proteína bacteriana domínio FHA presente na amostra DC pode interagir com BAX por seu motif LIG_FHA_2 (TABELA SUPLEMENTAR DAS INTERAÇÕES). BAX é uma proteína cujas funções mais bem descritas ocorrem no citoplasma e na mitocôndria. Entretando, observa-se pelo banco de dados ComPPI (VERES et al., 2015) que BAX pode estar localizada na membrama celular e na matriz extracelular, com um *score* de localização maior do que 0.8 e é nessa localização onde ela tem possibilidade de interagir com a proteína bacteriana. Deste modo, sugere-se que essa interação interfira na interação da BAX com outras proteínas como Bcl-2, afetando ambas apoptose e autofagia. A proteína Bcl-2 também tem função de intermediação entre esses processos, pois atua como proteína pró-autofágica quando interage com Beclin1 e anti-apoptótica quando interage com BAX.

Outra proteína importante na regulação da apoptose é β-catenina (decodificada pelo gene CTNNB1, como aparece na rede da Figura **??**). Essa proteína é um componente da via canônica de sinalização Wnt, sendo essencial para a adesão entre as células e inibição da apoptose em células epiteliais (DONMEZ; DEMIREZEN; BEKSAC, 2016). Essa inibição acontece porque β-catenina diminui a expressão da pró-apoptótica DAPK-2 por meio do fator de transcrição TCF4 (LI et al., 2009), também presente na rede da Figura **??**. Não obstante, pacientes de DC apresentam defeitos de permeabilidade das células e em proteínas de junção (GASSLER et al., 2001) , além também nível de apoptose aumentada em células

epiteliais (SABATINO et al., 2003). Deste modo, pode-se sugerir que a interação da proteína bacteriana com β-catenina pode estar prejudicando sua função de inibição da apoptose e de mantimento da junção celular. Outro importante papel desempenhado pela β-catenina é na proliferação celular. Quando ativada, β-catenina transloca-se ao núcleo e liga-se ao fator de transcricao TCF. Esse gene regula a transcrição de genes como c-Myc e Cyclin-D1, consequentemente aumentando a proliferação celular. TCF4 é a proteína da família TCF mais expressa no epitélio intestinal, e também está presente na rede da Figura **??**.

A sinalização entre β-catenina e TCF também é conhecida por seu papel na diferenciação e rearranjo de células epiteliais. Quando essa via é suprimida, a diferenciação das células tronco em células de Paneth e a migração de células diferenciadas é prejudicada (BATLLE et al., 2002). Deste modo, a interação entre a proteína da bactéria e a β-catenina tem potencial de estar impedindo também a recomposição do epitélio.

Em células saudáveis, a β-catenina regula a adesão entre as células. Quando superexpressa e não segregada por E-caderina na membrana celular, é rapidamente fosforilada pelo complexo de destruição (APC)/axin/GSK-3β e é rapidamente degradada. Em tumores, a inativação do (APC)/axin/GSK-3β leva ao acúmulo da β-catenina no citosol, promovendo seu translocamento ao núcleo. Chegando ao núcleo, promove a transcrição de genes de proliferação, sobrevivência e diferenciação, que inibem da apoptose (SERAFINO et al., 2014).

Em células epiteliais com DC, a apoptose é aumentada e a proliferação e diferenciação estão diminuídas. Na rede da Figura **??**, observa-se a interação da proteína bacteriana IMP desidrogenase/GMP reductase com β-catenina. Essa interação é única da condição de DC, portanto sugere-se que a proteína bacteriana possa estar reprimindo a β-catenina de forma que a impeça de translocar-se ao núcleo. Isso impediria que genes de proliferação e diferenciação sejam expressos pela indução de β-catenina, prejudicando os processos de renovação da barreira epitelial e impedindo a diminuição da apoptose.

Outra proteína relacionada à genes de proliferação e diferenciação que interage com proteína bacteriana nas condições de DC é YBX1. Essa proteína é secretada pelas células de diferentes linhagens em resposta à sinais inflamatórios. Quando extracelular, ativa o receptor NOTCH3, que libera seu domínio intracelular. Esse domínio migra para o núcleo e ativa genes

relacionados à proliferação e diferenciação celular (RAUEN et al., 2009; INDER et al., 2017). Devido ao seu papel na proliferação de tumores, NOTCH3 é proposta ser um alvo terapêutico, pois limita a metástase quando desativado (INDER et al., 2017). Embora seu papel nas células epiteliais intestinais especificamente ainda não tenha sido encontrado na literatura, supõe-se que o sequestro de YBX1 pela proteína bacteriana na matriz extracelular possa impedir a ativação da NOTCH3, evitando expressão de genes de diferenciação e proliferação celular por essa via.

Além do seu papel na diferenciação e proliferação celular, NOTCH3 também é responsável pela expressão das proteínas anti-inflamatórias IL-4 e IL-10 (ANASTASI et al., 2003). (BERNHARDT et al., 2017) demonstrou o papel de YBX1 na progressão da inflamação no fígado. Eles observaram que macrófagos de ratos deficientes de YBX1 apresentam produção defeituosa de IL-10, consequentemente apresentando alto nível de infiltração de células inflamatórias no tecido e aumento da produção de citocinas e chemoquinas como TNF-α, IL-6, IL-8 e CCL5. Na rede apresentada na Figura **??**, a proteína bacteriana Cold Shock presente unicamente nas condições de DC interage com YBX1, que está conectada com IL-10. Se a interação entre a proteína bacteriana e YBX1 interfere nas funções intracelulares de YBX1, pode estar contribuindo para o aumento de inflamação das células DC por impedir que a expressão de IL-10 seja ativada por essa via. A deficiência de IL-10 é amplamente relatada no desenvolvimento de IBDs. Essa proteína desempenha importantes papéis para o controle de inflamação, como a modulação de síntese de proteínas, proliferação e diferenciação de macrófagos e controle de EROS pela ativação da mitofagia (IP et al., 2017; RATH; MOSCHETTA; HALLER, 2018; SHIM, 2019).

Outra proteína envolvida na mitofagia que interage com proteínas bacterianas unicamente presente em amostras de DC é FOXO3 (SONG et al., 2017). Observou-se que FOXO3 pode ser inativada por infecção bacteriana, aumentando o nível de citocinas pró-inflamatórias e, desse modo, a inflamação (SNOEKS et al., 2009). FOXO3 também é envolvida em alguns processos que também fazem parte da patogênese de DC, como inflamação, proliferação celular, autofagia e, mais uma vez, mitofagia (SNOEKS et al., 2009; HAGENBUCHNER; AUSSERLECHNER, 2013). Deste modo, a interação da proteína bacteriana com FOXO3 pode ser outro meio de facilitar a indução de DC por meio da interferência nesses processos.

6.2 Análise taxonômica

Para deduzir a origem putativa das proteínas bacterianas presentes na rede da Figura **??**, realizou-se a análise taxonômica das proteínas bacterianas. Esse resultado pode ser observado na Tabela 6.

Como observado na tabela, *Trypsin-like serine proteases* é a única proteína bacteriana presente unicamente em condições saudáveis da rede. Essa proteína que desempenha papel em diversos processos como resposta a choque térmico, lise de parede celular bacteriana, regulação da transcrição e outros. Ela é expressada em plantas e animais, portanto supõe-se que sua presença nos procariotos possa ser a consequência de sua adaptação no ambiente hospedeiro (TRIPATHI; SOWDHAMINI, 2008).

De acordo com a análise, observou-se que *Trypsin-like serine proteases* pode ser originada de i) gêneros não identificados das famílias Porphyromonadaceae e Lachnospiraceae ii) famílias não identificadas das ordens Bacteroidales e Clostridiales. Essas bactérias foram encontradas em amostras de indivíduos saudáveis, mas não em DC na análise de (WILLING et al., 2010) e também em outros estudos (LOH; BLAUT, 2012; MONDOT et al., 2016). Bactérias pertencentes a estes filos estimulam as células T regulatórias a diminuir a inflamação intestinal e também são produtoras de SCFA (YILMAZ et al., 2019; DONG et al., 2016). Deste modo, sugere-se que essas bactérias são importantes para um microbioma intestinal saudável (ZUO; NG, 2018) também por causa de outros processos importantes além da modulação da autofagia.

Em relação à origem em potencial das proteínas bacterianas específicas de DC, também observamos gêneros característicos da microbiota disbiótica da doença. Primeiramente, o gênero *Aeromonas*, pertencente à família Aeromonadaceae, tem o potencial de ser a origem de quatro proteínas bacterianas específicas de DC presentes na rede final: ATPases com atividades de chaperona, proteínas Cold Shock, IMP desidrogenase / GMP reductase e domínio FHA. Infecção por *Aeromonas*, juntamente com outros fatores, é conhecida por levar à IBD e ao seu relapso (LOBATóN et al., 2015).

Outra família característica da microbiota de DC que é a potencial origem da maioria das proteínas bacterianas únicas de DC é Enterobacteriacreae. Essa família já é conhecida por ter ocorrência aumentada em DC comparado com saudáveis, sendo considerada um marcador de DC (LANE; ZISMAN; SUSKIND, 2017; WRIGHT et al., 2015). A família Enterobacteriaceae é parte da classe Gammaproteobacteria e muitos patógenos intestinais pertencem à essa família incluindo *Escherichia*, *Salmonella*, *Shigella* e *Citrobacter*, que são causas de doenças enterógenas (JENKINS et al., 2017). O crescimento aumentado de Enterobacteriaceae em indivíduos DC é promovido pelo nitrato produzido pelas células hospedeiras em condições inflamatórias (BUTTó; SCHAUBECK; HALLER, 2015). Como notamos na Tabela 6, *Shigella*, *Citrobacter* e um gênero não identificado de Enterobacteriaceae estão inclusas como origens putativas das proteínas ATPases com atividade chaperona, proteínas Cold Shock, IMP desidrogenase / GMP reductase e domínio FHA. Primeiramente, *Citrobacter* pode ser encontrada em indivíduos com DC e é conhecida por desempenhar papéis em processos importantes para o desenvolvimento da doença, como prejudicar a cicatrização da mucosa e a integridade da barreira epitelial, interferir na composição da microbiota comensal e induzir a inflamação (KOROLEVA et al., 2015; MATTHEWS et al., 1980). *Shigella*, também é conhecida por estar presente em grande quantidade em amostras de DC (WRIGHT et al., 2015). Essa bactéria não somente coloniza o intestino, mas também escapa da maquinaria de autofagia quando invade as células (OGAWA et al., 2005; SUDHAKAR et al., 2019). Além disso, *Shigella* também é conhecida por aumentar a inflamação do tecido hospedeiro como uma estratégia de obtenção de vantagens sobre o microbioma comensal (LIU; PILLA; TANG, 2019).

A bactéria *Fusobacterium* é uma das origens putativa da proteína domínio FHA, também é considerada uma bactéria marcadora para doença de Crohn (PASCAL et al., 2017) e é associada com o aumento da expressão de citocinas pró-inflamatórias como IL-6, IL-12, IL-17 e $TNF - \alpha$ (BASHIR et al., 2016).

O outro gênero de bactéria é *Lactobacillus*. Ao contrário das outras bactérias identificadas como origem putativa das proteínas bacterianas de DC, *Lactobacillus* é normalmente conhecido como uma bactéria probiótica, mas foi identificado como possível origem das proteínas ATPases com atividade chaperona, proteínas *Cold Shock* e IMP desidrogenase/GMP

reductase. Não obstante, essa foi uma das bactérias testadas para uso como probiótico no tratamento de DC, mas mostrou-se pouco efetivo na prevenção de relapsos ou mantimento de remissão da doença (SCHULTZ et al., 2004; LICHTENSTEIN; AVNI-BIRON; BEN-BASSAT, 2016).

Estudo de probióticos para a prevenção de relapsos ou indução da remissão de DC têm executados, mas não tem-se obtido muito sucesso. *Saccharomyces boulardii*, uma espécie de fungo é a única espécies que mostrou efeitos benéficos para manter a remissão de DC nos pacientes (LICHTENSTEIN; AVNI-BIRON; BEN-BASSAT, 2016). Entretanto, nesse estudo não foram incluídos dados de fungos, portanto não podemos apresentar conclusões a nível molecular a respeito. Os gêneros de bactérias testadas para probióticos contra DC em estudos clínicos além de *Lactobacillus* encontradas na literatura foram *Bifidobacterium* e *Streptococcus* (LICHTENSTEIN; AVNI-BIRON; BEN-BASSAT, 2016). Nenhum desses gêneros pertencem às famílias Bacteroidales ou Crostridales, famílias que, de acordo com a nossa análise, podem ser a origem putativa da proteína *Trypsin-like serine proteases*, presente na nossa rede como moduladora dos genes de autofagia unicamente presente em amostras saudáveis.

Deste modo, pode-se sugerir que bactérias dos táxons encontrados como origem putativa de *Trypsin-like serine proteases* sejam testados como probióticos de tratamento de DC. Nos Resultados Suplementares, aba n^o 4, estão apresentados os táxons abaixo das ordens Clostridiales e Bacteroiadales que expressam essa proteína (de acordo com o PFAM). De acordo com essas informações, algumas bactérias podem ser sugeridas como futuros testes probióticos como *Coprococcus* (família Lachnospiraceae), que é um produtor de SCFA, moléculas já descritas como benéficas contra a DC.

Esses dados foram sugeridos de acordo com os dados que temos disponíveis até agora. Entretanto, muitas das espécies citadas ainda não foram muito bem descritas, como por exemplo o gênero *Anaerocolumna*, da família Lachnospiraceae. Outra dificuldade dessa análise inclui a dificuldade dos pesquisadores do estudo de WILLING et al. de identificar as bactérias com mais especificidade, principalmente levando em conta que o estudo foi realizado há nove anos atrás. A dificuldade da análise funcional também é inerente: poucas proteínas da análise

de ERICKSON et al. tiveram sua origem identificadas, bem como sua sequência completa.

Os estudos de metaproteômica ainda estão em sua infância, e mais análises são necessárias para conclusões mais exatas sobre os processos microbióticos. Enquanto aguarda-se avanços tecnológicos, a biologia de sistemas e predições computacionais com os dados por ora obtidos são de suma importância para que análises mais direcionadas para o tratamento das IBDs e outras doenças relacionadas com a microbiota sejam realizadas.

7 Conclusão

A doença de Crohn é uma doença de etiologia complexa e causas pouco claras. Somente há poucos anos o microbioma foi incluído como um dos fatores causadores da doença, deste modo, as investigações sobre como ele influencia no desenvolvimento da doença ainda estão em seus primórdios. Além disso, as tecnologias que poderiam ajudar a esclarecer as interações moleculares entre microbioma e hospedeiro, como a metaproteômica, também ainda são pouco utilizadas devido à dificuldade de execução e análise. Deste modo, a utilização de ferramentas computacionais e dados já obtidos para a formação de teorias que podem servir de atalho para análises experimentais são bem vindas.

Nesta tese, foi apresentada uma análise da interação microbioma-hospdeiro de indivíduos com doença de Crohn feita com abordagem de biologia de sistemas com a utilização de dados públicos. Primeiramente, observou-se algumas vias presentes unicamente em doença de Crohn que têm o potencial de modular a autofagia.

Um dos genes chaves que mostrou-se potencialmente modulado, foi ATG4D, que desempenha um papel especial na autofagia direcionada à mitocôndria: a mitofagia. Ao longo da rede, outras proteínas com influência na mitofagia corroboraram com o fato de que esse pode ser um processo importantemente modulado e merece atenção em análises posteriores.

Outros processos potencialmente modulados que também estão presentes na rede são relacionados à integridade e homeostase da barreira epitelial, como a diferenciação e proliferação celular e apoptose. Defeitos nesses processos também são descritos na DC, e estão relacionados com o aumento da inflamação 13. Eles também estão direta ou indiretamente relacionados à autofagia, o processo utilizado como foco inicial da análise, portanto, seus defeitos podem ser meios pelos quais a autofagia é modulada pelas proteínas bacterianas.

Ademais, analisamos as origens putativas das proteínas bacterianas presentes na rede. As bactérias identificadas corroboraram com a literatura em relação à composição microbiótica diferencial entre as condições. A partir desses dados podemos inferir sugestões de futuras

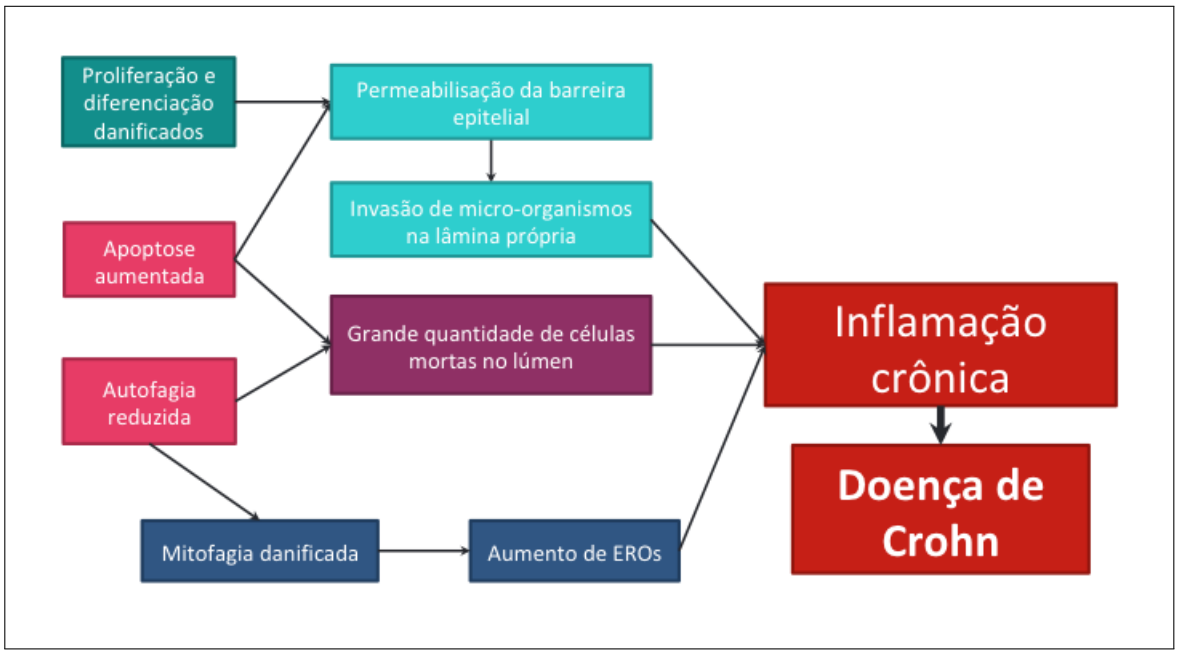

Figura 13: Esquema de como os processos biológicos afetados pelas proteínas bacterianas da rede obtida podem levar à inflamação crônica e doença de Crohn.

bactérias a serem usadas como tratamento probiótico para a DC, como *Coproccocus*.

Além das conclusões sobre os processos biológicos envolvidos nessa interação, essa análise rendeu a elaboração de um protocolo de análise computacional da interação microbioma-hospedeiro que poderá ser aplicado nas mais diversas problemáticas e espécies de hospedeiro.

Referências

ALBERTS, B. et al. *JD Biologia Molecular da Célula*. Brasil: Ed. Artes Médicas. 4 Edicão, 2004. Citado na página 20.

ANASTASI, E. et al. Expression of activated notch3 in transgenic mice enhances generation of t regulatory cells and protects against experimental autoimmune diabetes. *Journal of Immunology*, v. 171, n. 9, p. 4504–4511, nov 2003. ISSN 0022-1767. Disponível em: <http://dx.doi.org/10.4049/jimmunol.171.9.4504>. Citado na página 62.

ANBAZHAGAN, A. N. et al. Pathophysiology of IBD associated diarrhea. *Tissue barriers*, v. 6, n. 2, p. e1463897, may 2018. Disponível em: <http://dx.doi.org/10.1080/21688370.2018.1463897>. Citado na página 14.

AVIELLO, G.; KNAUS, U. G. ROS in gastrointestinal inflammation: Rescue or sabotage? *British Journal of Pharmacology*, v. 174, n. 12, p. 1704–1718, 2017. Disponível em: <http://dx.doi.org/10.1111/bph.13428>. Citado 2 vezes nas páginas 15 e 59.

BADER, S.; KüHNER, S.; GAVIN, A.-C. Interaction networks for systems biology. *FEBS Letters*, v. 582, n. 8, p. 1220–1224, apr 2008. Disponível em: <http://dx.doi.org/10.1016/j.febslet.2008.02.015>. Citado 2 vezes nas páginas 32 e 33.

BARRETT, T. et al. NCBI GEO: archive for functional genomics data sets–update. *Nucleic Acids Research*, v. 41, n. Database issue, p. D991–5, jan 2013. Disponível em: <http://dx.doi.org/10.1093/nar/gks1193>. Citado na página 45.

BASHIR, A. et al. Fusobacterium nucleatum, inflammation, and immunity: the fire within human gut. *Tumour Biology*, v. 37, n. 3, p. 2805–2810, mar 2016. Disponível em: <http://dx.doi.org/10.1007/s13277-015-4724-0>. Citado na página 64.

BATLLE, E. et al. Beta-catenin and TCF mediate cell positioning in the intestinal epithelium by controlling the expression of EphB/ephrinB. *Cell*, v. 111, n. 2, p. 251–263, oct 2002. ISSN 0092-8674. Disponível em: <https://www.ncbi.nlm.nih.gov/pubmed/12408869>. Citado na página 61.

BAUMGART, D. C.; SANDBORN, W. J. Crohn's disease. *The Lancet*, v. 380, n. 9853, p. 1590–1605, nov 2012. Disponível em: <http://dx.doi.org/10.1016/S0140-6736(12)60026-9>. Citado na página 14.

BELLA, J. M. D. et al. High throughput sequencing methods and analysis for microbiome research. *Journal of Microbiological Methods*, v. 95, n. 3, p. 401–414, dec 2013. Disponível em: <http://dx.doi.org/10.1016/j.mimet.2013.08.011>. Citado na página 15.

BERNHARDT, A. et al. Inflammatory cell infiltration and resolution of kidney inflammation is orchestrated by the cold-shock protein y-box binding protein-1. *Kidney International*, v. 92, n. 5, p. 1157–1177, jun 2017. Disponível em: <http://dx.doi.org/10.1016/j.kint.2017.03.035>. Citado na página 62.

BETIN, V. M. S.; LANE, J. D. Atg4D at the interface between autophagy and apoptosis. *Autophagy*, v. 5, n. 7, p. 1057–1059, oct 2009. Disponível em: <http://dx.doi.org/10.4161/auto.5.7.9684>. Citado na página 60.

BETIN, V. M. S.; LANE, J. D. Caspase cleavage of Atg4D stimulates GABARAP-l1 processing and triggers mitochondrial targeting and apoptosis. *Journal of Cell Science*, v. 122, n. Pt 14, p. 2554–2566, jul 2009. Disponível em: <http://dx.doi.org/10.1242/jcs.046250>. Citado na página 60.

BETIN, V. M. S. et al. A cryptic mitochondrial targeting motif in Atg4D links caspase cleavage with mitochondrial import and oxidative stress. *Autophagy*, v. 8, n. 4, p. 664–676, apr 2012. Disponível em: <http://dx.doi.org/10.4161/auto.19227>. Citado na página 59.

BINDEA, G. et al. ClueGO: a cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. *Bioinformatics*, v. 25, n. 8, p. 1091–1093, apr 2009. Disponível em: <http://dx.doi.org/10.1093/bioinformatics/btp101>. Citado na página 46.

BLANDER, J. M. Death in the intestinal epithelium-basic biology and implications for inflammatory bowel disease. *The FEBS Journal*, v. 283, n. 14, p. 2720–2730, jun 2016. Disponível em: <http://dx.doi.org/10.1111/febs.13771>. Citado 2 vezes nas páginas 23 e 25.

BOVOLENTA, L. A.; ACENCIO, M. L.; LEMKE, N. HTRIdb: an open-access database for experimentally verified human transcriptional regulation interactions. *BMC Genomics*, v. 13, p. 405, aug 2012. Disponível em: <http://dx.doi.org/10.1186/1471-2164-13-405>. Citado na página 44.

BUTTó, L. F.; SCHAUBECK, M.; HALLER, D. Mechanisms of microbe-host interaction in crohn's disease: Dysbiosis vs. pathobiont selection. *Frontiers in immunology*, v. 6, p. 555, nov 2015. Disponível em: <http://dx.doi.org/10.3389/fimmu.2015.00555>. Citado na página 64.

CAREY, R. et al. Activation of an IL-6:STAT3-dependent transcriptome in pediatric-onset inflammatory bowel disease. *Inflammatory Bowel Diseases*, v. 14, n. 4, p. 446–457, apr 2008. Disponível em: <http://dx.doi.org/10.1002/ibd.20342>. Citado na página 50.

CHEONG, H. et al. The atg1 kinase complex is involved in the regulation of protein recruitment to initiate sequestering vesicle formation for nonspecific autophagy in saccharomyces cerevisiae. *Molecular Biology of the Cell*, v. 19, n. 2, p. 668–681, feb 2008. Disponível em: <http://dx.doi.org/10.1091/mbc.E07-08-0826>. Citado na página 27.

CORRIDONI, D. et al. Emerging mechanisms of innate immunity and their translational potential in inflammatory bowel disease. *Frontiers in medicine*, v. 5, p. 32, feb 2018. Disponível em: <http://dx.doi.org/10.3389/fmed.2018.00032>. Citado 2 vezes nas páginas 27 e 28.

CRUVINEL, W. d. M. et al. Sistema imunitário: Parte i. fundamentos da imunidade inata com ênfase nos mecanismos moleculares e celulares da resposta inflamatória. *Revista brasileira de reumatologia*, v. 50, n. 4, p. 434–447, aug 2010. ISSN 1809-4570. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0482-50042010000400008&lng=pt&nrm=iso&tlng=en>. Citado 2 vezes nas páginas 16 e 18.

DELGADO, M. E.; GRABINGER, T.; BRUNNER, T. Cell death at the intestinal epithelial front line. *The FEBS Journal*, v. 283, n. 14, p. 2701–2719, 2016. Disponível em: <http://dx.doi.org/10.1111/febs.13575>. Citado 3 vezes nas páginas 23, 24 e 25.

DONG, B. et al. A new process to improve short-chain fatty acids and bio-methane generation from waste activated sludge. *Journal of environmental sciences (China)*, v. 43, p. 159–168, may 2016. Disponível em: <http://dx.doi.org/10.1016/j.jes.2015.10.004>. Citado na página 63.

DONMEZ, H. G.; DEMIREZEN, S.; BEKSAC, M. S. The relationship between beta-catenin and apoptosis: A cytological and immunocytochemical examination. *Tissue & cell*, v. 48, n. 3, p. 160–167, jun 2016. Disponível em: <http://dx.doi.org/10.1016/j.tice.2016.04.001>. Citado na página 60.

DOSZTáNYI, Z. et al. IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics*, v. 21, n. 16, p. 3433–3434, aug 2005. Disponível em: <http://dx.doi.org/10.1093/bioinformatics/bti541>. Citado na página 43.

DYER, M. D.; MURALI, T. M.; SOBRAL, B. W. The landscape of human proteins interacting with viruses and other pathogens. *PLoS Pathogens*, v. 4, n. 2, p. e32, feb 2008. Disponível em: <http://dx.doi.org/10.1371/journal.ppat.0040032>. Citado na página 33.

DYER, M. D. et al. The human-bacterial pathogen protein interaction networks of bacillus anthracis, francisella tularensis, and yersinia pestis. *Plos One*, v. 5, n. 8, p. e12089, aug 2010. Disponível em: <http://dx.doi.org/10.1371/journal.pone.0012089>. Citado na página 33.

EDER, P. et al. Disturbances in apoptosis of lamina propria lymphocytes in crohn's disease. *Archives of medical science : AMS*, v. 11, n. 6, p. 1279–1285, dec 2015. Disponível em: <http://dx.doi.org/10.5114/aoms.2015.54203>. Citado na página 29.

EL-GEBALI, S. et al. The pfam protein families database in 2019. *Nucleic Acids Research*, v. 47, n. D1, p. D427–D432, jan 2019. Disponível em: <http://dx.doi.org/10.1093/nar/gky995>. Citado 3 vezes nas páginas 33, 43 e 46.

ELMORE, S. Apoptosis: a review of programmed cell death. *Toxicologic Pathology*, v. 35, n. 4, p. 495–516, jun 2007. Disponível em: <http://dx.doi.org/10.1080/01926230701320337>. Citado na página 28.

ERICKSON, A. R. et al. Integrated metagenomics/metaproteomics reveals human host-microbiota signatures of crohn's disease. *Plos One*, v. 7, n. 11, p. e49138, nov 2012. Disponível em: <http://dx.doi.org/10.1371/journal.pone.0049138>. Citado 9 vezes nas páginas 9, 31, 37, 41, 42, 46, 48, 55 e 66.

FERREIRA, V. L. et al. Cytokines and interferons: types and functions. In: KHAN, W. A. (Ed.). *Autoantibodies and Cytokines*. IntechOpen, 2019. ISBN 978-1-78984-852-6. Disponível em: <https://www.intechopen.com/books/autoantibodies-and-cytokines/cytokines-and-interferons-types-and-functions>. Citado na página 20.

FOLEY, S. et al. Salmonella pathogenicity and host adaptation in chicken-associated serovars. *Microbiology and Molecular Biology Reviews*, v. 77, n. 4, p. 582–607, dec 2013. Disponível em: <http://dx.doi.org/10.1128/{MMBR}.00015->. Citado na página 33.

GASSLER, N. et al. Inflammatory bowel disease is associated with changes of enterocytic junctions. *American Journal of Physiology. Gastrointestinal and Liver Physiology*, v. 281, n. 1, p. G216–28, jul 2001. Disponível em: <http://dx.doi.org/10.1152/ajpgi.2001.281.1.G216>. Citado na página 60.

GOULD, A. L. et al. Microbiome interactions shape host fitness. *Proceedings of the National Academy of Sciences of the United States of America*, v. 115, n. 51, p. E11951–E11960, dec 2018. Disponível em: <http://dx.doi.org/10.1073/pnas.1809349115>. Citado na página 33.

GOUW, M. et al. The eukaryotic linear motif resource - 2018 update. *Nucleic Acids Research*, v. 46, n. D1, p. D428–D434, jan 2018. ISSN 0305-1048. Disponível em: <http://academic.oup.com/nar/article/doi/10.1093/nar/gkx1077/4612965>. Citado 2 vezes nas páginas 33 e 42.

GREEN, D. R.; GALLUZZI, L.; KROEMER, G. Mitochondria and the autophagy-inflammation-cell death axis in organismal aging. *Science*, v. 333, n. 6046, p. 1109–1112, aug 2011. ISSN 1095-9203. Disponível em: <http://dx.doi.org/10.1126/science.1201940>. Citado na página 26.

GREEN, S. et al. Network analyses in systems biology: new strategies for dealing with biological complexity. *Synthese*, v. 195, n. 4, p. 1–27, jan 2017. ISSN 0039-7857. Disponível em: <http://link.springer.com/10.1007/s11229-016-1307-6>. Citado na página 32.

GUVEN-MAIOROV, E.; TSAI, C.-J.; NUSSINOV, R. Structural host-microbiota interaction networks. *PLoS Computational Biology*, v. 13, n. 10, p. e1005579, oct 2017. Disponível em: <http://dx.doi.org/10.1371/journal.pcbi.1005579>. Citado na página 33.

HAGENBUCHNER, J.; AUSSERLECHNER, M. J. Mitochondria and FOXO3: breath or die. *Frontiers in physiology*, v. 4, p. 147, jun 2013. Disponível em: <http://dx.doi.org/10.3389/fphys.2013.00147>. Citado na página 62.

HAMACHER-BRADY, A.; BRADY, N. R. Mitophagy programs: mechanisms and physiological implications of mitochondrial targeting by autophagy. *Cellular and Molecular Life Sciences*, v. 73, n. 4, p. 775–795, feb 2016. Disponível em: <http://dx.doi.org/10.1007/s00018-015-2087-8>. Citado 2 vezes nas páginas 27 e 59.

HAN, H. et al. TRRUST v2: an expanded reference database of human and mouse transcriptional regulatory interactions. *Nucleic Acids Research*, v. 46, n. D1, p. D380–D386, jan 2018. Disponível em: <http://dx.doi.org/10.1093/nar/gkx1013>. Citado na página 44.

HEAZLEWOOD, C. K. et al. Aberrant mucin assembly in mice causes endoplasmic reticulum stress and spontaneous inflammation resembling ulcerative colitis. *PLoS Medicine*, v. 5, n. 3, p. e54, mar 2008. ISSN 1549-1676. Disponível em: <http://dx.doi.org/10.1371/journal.pmed.0050054>. Citado na página 24.

HEINKEN, A. et al. Systems-level characterization of a host-microbe metabolic symbiosis in the mammalian gut. *Gut microbes*, v. 4, n. 1, p. 28–40, feb 2013. Disponível em: <http://dx.doi.org/10.4161/gmic.22370>. Citado na página 33.

HENDERSON, P.; STEVENS, C. The role of autophagy in crohn's disease. *Cells*, v. 1, n. 3, p. 492–519, aug 2012. Disponível em: <http://dx.doi.org/10.3390/cells1030492>. Citado 4 vezes nas páginas 14, 25, 27 e 30.

INDER, S. et al. The notch-3 receptor: A molecular switch to tumorigenesis? *Cancer Treatment Reviews*, v. 60, p. 69–76, nov 2017. Disponível em: <http://dx.doi.org/10.1016/j.ctrv.2017.08.011>. Citado na página 62.

IP, W. K. E. et al. Anti-inflammatory effect of IL-10 mediated by metabolic reprogramming of macrophages. *Science*, v. 356, n. 6337, p. 513–519, may 2017. ISSN 0036-8075. Disponível em: <http://www.sciencemag.org/lookup/doi/10.1126/science.aal3535>. Citado na página 62.

IVANOV, I. I.; HONDA, K. Intestinal commensal microbes as immune modulators. *Cell Host & Microbe*, v. 12, n. 4, p. 496–508, oct 2012. Disponível em: <http://dx.doi.org/10.1016/j.chom.2012.09.009>. Citado na página 15.

JENKINS, C. et al. Enterobacteriaceae. In: *Infectious Diseases*. Elsevier, 2017. p. 1565–1578.e2. ISBN 9780702062858. Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/B9780702062858001805>. Citado na página 64.

JUNKER, B. H.; SCHREIBER, F. (Ed.). *Analysis of biological networks*. Hoboken, NJ, USA: John Wiley & Sons, Inc., 2008. ISBN 9780470041444. Disponível em: <http://doi.wiley.com/10.1002/9780470253489>. Citado 2 vezes nas páginas 32 e 44.

JUSTE, C. et al. Bacterial protein signals are associated with crohn's disease. *Gut*, v. 63, n. 10, p. 1566–1577, oct 2014. Disponível em: <http://dx.doi.org/10.1136/gutjnl-2012-303786>. Citado na página 31.

KALE, J.; OSTERLUND, E. J.; ANDREWS, D. W. BCL-2 family proteins: changing partners in the dance towards death. *Cell Death and Differentiation*, v. 25, n. 1, p. 65–80, 2018. Disponível em: <http://dx.doi.org/10.1038/cdd.2017.186>. Citado na página 60.

KATIYAR-AGARWAL, S.; JIN, H. Role of small RNAs in host-microbe interactions. *Annual Review of Phytopathology*, v. 48, p. 225–246, 2010. Disponível em: <http://dx.doi.org/10.1146/annurev-phyto-073009-114457>. Citado na página 33.

KAUFFMAN, K. J. et al. Delipidation of mammalian atg8-family proteins by each of the four ATG4 proteases. *Autophagy*, v. 14, n. 6, p. 992–1010, apr 2018. Disponível em: <http://dx.doi.org/10.1080/15548627.2018.1437341>. Citado na página 59.

KORCSMAROS, T. et al. Teaching the bioinformatics of signaling networks: an integrated approach to facilitate multi-disciplinary learning. *Briefings in Bioinformatics*, v. 14, n. 5, p. 618–632, sep 2013. Disponível em: <http://dx.doi.org/10.1093/bib/bbt024>. Citado na página 33.

KOROLEVA, E. P. et al. Citrobacter rodentium-induced colitis: A robust model to study mucosal immune responses in the gut. *Journal of Immunological Methods*, v. 421, p. 61–72, jun 2015. Disponível em: <http://dx.doi.org/10.1016/j.jim.2015.02.003>. Citado na página 64.

KOSTIC, A. D.; XAVIER, R. J.; GEVERS, D. The microbiome in inflammatory bowel disease: current status and the future ahead. *Gastroenterology*, v. 146, n. 6, p. 1489–1499, may 2014. Disponível em: <http://dx.doi.org/10.1053/j.gastro.2014.02.009>. Citado na página 16.

KUBISCH, J. et al. Complex regulation of autophagy in cancer - integrated approaches to discover the networks that hold a double-edged sword. *Seminars in Cancer Biology*, v. 23, n. 4, p. 252–261, aug 2013. Disponível em: <http://dx.doi.org/10.1016/j.semcancer.2013.06.009>. Citado 5 vezes nas páginas 7, 25, 27, 35 e 44.

KUPRASH, D. V.; NEDOSPASOV, S. A. Molecular and cellular mechanisms of inflammation. *Biochemistry. Biokhimiia*, v. 81, n. 11, p. 1237–1239, nov 2016. Disponível em: <http://dx.doi.org/10.1134/S0006297916110018>. Citado na página 20.

LANE, E. R.; ZISMAN, T. L.; SUSKIND, D. L. The microbiota in inflammatory bowel disease: current and therapeutic insights. *Journal of inflammation research*, v. 10, p. 63–73, jun 2017. Disponível em: <http://dx.doi.org/10.2147/{JIR}.S1160>. Citado 3 vezes nas páginas 16, 20 e 64.

LAUNAY, G. et al. MatrixDB, the extracellular matrix interaction database: updated content, a new navigator and expanded functionalities. *Nucleic Acids Research*, v. 43, n. Database issue, p. D321–7, jan 2015. Disponível em: <http://dx.doi.org/10.1093/nar/gku1091>. Citado na página 42.

LI, H. et al. Down-regulation of death-associated protein kinase-2 is required for beta-catenin-induced anoikis resistance of malignant epithelial cells. *The Journal of Biological Chemistry*, v. 284, n. 4, p. 2012–2022, jan 2009. Disponível em: <http://dx.doi.org/10.1074/jbc.M805612200>. Citado na página 60.

LICHTENSTEIN, L.; AVNI-BIRON, I.; BEN-BASSAT, O. Probiotics and prebiotics in crohn's disease therapies. *Best Practice & Research. Clinical Gastroenterology*, v. 30, n. 1, p. 81–88, feb 2016. Disponível em: <http://dx.doi.org/10.1016/j.bpg.2016.02.002>. Citado na página 65.

LIU, G.; PILLA, G.; TANG, C. M. Shigella host:pathogen interactions: keeping bacteria in the loop. *Cellular Microbiology*, p. e13062, may 2019. Disponível em: <http://dx.doi.org/10.1111/cmi.13062>. Citado na página 64.

LOBATóN, T. et al. Aeromonas species: an opportunistic enteropathogen in patients with inflammatory bowel diseases? a single center cohort study. *Inflammatory Bowel Diseases*, v. 21, n. 1, p. 71–78, jan 2015. Disponível em: <http://dx.doi.org/10.1097/{MIB}.0000000000000002>. Citado na página 63.

LOH, G.; BLAUT, M. Role of commensal gut bacteria in inflammatory bowel diseases. *Gut microbes*, v. 3, n. 6, p. 544–555, dec 2012. Disponível em: <http://dx.doi.org/10.4161/gmic.22156>. Citado na página 63.

MADIGAN, M. *Brock biology of microorganisms*. San Francisco: Benjamin Cummings, 2012. ISBN 032164963X. Citado 2 vezes nas páginas 7 e 17.

MADIGAN, M. T. et al. *Microbiologia De Brock - 14ª Edição*. [S.l.]: Artmed Editora, 2016. 1032 p. ISBN 9788582712986. Citado 4 vezes nas páginas 17, 18, 19 e 20.

MANICHANH, C. et al. The gut microbiota in IBD. *Nature Reviews. Gastroenterology & Hepatology*, v. 9, n. 10, p. 599–608, oct 2012. Disponível em: <http://dx.doi.org/10.1038/nrgastro.2012.152>. Citado na página 15.

MATTHEWS, N. et al. Agglutinins to bacteria in crohn's disease. *Gut*, v. 21, n. 5, p. 376–380, may 1980. ISSN 0017-5749. Disponível em: <http://gut.bmj.com/cgi/doi/10.1136/gut.21.5.376>. Citado na página 64.

MEDZHITOV, R. Origin and physiological roles of inflammation. *Nature*, v. 454, n. 7203, p. 428–435, jul 2008. Disponível em: <http://dx.doi.org/10.1038/nature07201>. Citado 2 vezes nas páginas 20 e 22.

MESSER, J. S. The cellular autophagy/apoptosis checkpoint during inflammation. *Cellular and Molecular Life Sciences*, v. 74, n. 7, p. 1281–1296, 2017. Disponível em: <http://dx.doi.org/10.1007/s00018-016-2403-y>. Citado 2 vezes nas páginas 23 e 60.

MICHAIL, S.; BULTRON, G.; DEPAOLO, R. W. Genetic variants associated with crohn's disease. *The application of clinical genetics*, v. 6, p. 25–32, jul 2013. Disponível em: <http://dx.doi.org/10.2147/{TACG}.S339>. Citado 2 vezes nas páginas 14 e 27.

MITCHELL, A. L. et al. InterPro in 2019: improving coverage, classification and access to protein sequence annotations. *Nucleic Acids Research*, v. 47, n. D1, p. D351–D360, jan 2019. Disponível em: <http://dx.doi.org/10.1093/nar/gky1100>. Citado na página 43.

MIZUSHIMA, N. Autophagy: process and function. *Genes & Development*, v. 21, n. 22, p. 2861–2873, nov 2007. Disponível em: <http://dx.doi.org/10.1101/gad.1599207>. Citado 2 vezes nas páginas 25 e 27.

MONDAL, K.; KUGATHASAN, S. IBD: Genetic differences in crohn's disease susceptibility and outcome. *Nature Reviews. Gastroenterology & Hepatology*, v. 14, n. 5, p. 266–268, mar 2017. Disponível em: <http://dx.doi.org/10.1038/nrgastro.2017.24>. Citado na página 14.

MONDOT, S. et al. Structural robustness of the gut mucosal microbiota is associated with crohn's disease remission after surgery. *Gut*, v. 65, n. 6, p. 954–962, 2016. Disponível em: <http://dx.doi.org/10.1136/gutjnl-2015-309184>. Citado na página 63.

MONTERO-MELéNDEZ, T. et al. Identification of novel predictor classifiers for inflammatory bowel disease by gene expression profiling. *Plos One*, v. 8, n. 10, p. e76235, oct 2013. Disponível em: <http://dx.doi.org/10.1371/journal.pone.0076235>. Citado na página 50.

MORGAN, X. C. et al. Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment. *Genome Biology*, v. 13, n. 9, p. R79, apr 2012. Disponível em: <http://dx.doi.org/10.1186/gb-2012-13-9-r79>. Citado 2 vezes nas páginas 15 e 16.

MOTTAWEA, W. et al. Altered intestinal microbiota-host mitochondria crosstalk in new onset crohn's disease. *Nature Communications*, v. 7, p. 13419, nov 2016. Disponível em: <http://dx.doi.org/10.1038/ncomms13419>. Citado 2 vezes nas páginas 27 e 28.

MUKHOPADHYAY, S. et al. Autophagy and apoptosis: where do they meet? *Apoptosis: An International Journal on Programmed Cell Death*, v. 19, n. 4, p. 555–566, apr 2014. Disponível em: <http://dx.doi.org/10.1007/s10495-014-0967-2>. Citado 2 vezes nas páginas 28 e 29.

NAGATA, S. Apoptosis and clearance of apoptotic cells. *Annual Review of Immunology*, v. 36, p. 489–517, apr 2018. Disponível em: <http://dx.doi.org/10.1146/annurev-immunol-042617-053010>. Citado na página 29.

NG, S. C. et al. Worldwide incidence and prevalence of inflammatory bowel disease in the 21st century: a systematic review of population-based studies. *The Lancet*, v. 390, n. 10114, p. 2769–2778, dec 2018. Disponível em: <http://dx.doi.org/10.1016/S0140-6736(17)32448-0>. Citado na página 37.

NORRGARD, K. Genetic variation and disease: GWAS. *Nature Education*, v. 1, n. 1, p. 87, 2008. Disponível em: <https://www.nature.com/scitable/topicpage/genetic-variation-and-disease-gwas-682>. Citado na página 31.

NOURANI, E.; KHUNJUSH, F.; DURMUş, S. Computational approaches for prediction of pathogen-host protein-protein interactions. *Frontiers in microbiology*, v. 6, p. 94, feb 2015. Disponível em: <http://dx.doi.org/10.3389/fmicb.2015.00094>. Citado na página 33.

OBATA, Y. et al. Epithelial cell-intrinsic notch signaling plays an essential role in the maintenance of gut immune homeostasis. *Journal of Immunology*, v. 188, n. 5, p. 2427–2436, mar 2012. Disponível em: <http://dx.doi.org/10.4049/jimmunol.1101128>. Citado na página 24.

OGAWA, M. et al. Escape of intracellular shigella from autophagy. *Science*, v. 307, n. 5710, p. 727–731, feb 2005. ISSN 1095-9203. Disponível em: <http://dx.doi.org/10.1126/science.1106036>. Citado na página 64.

OLIVEIRA, A. L. Biotechnology, big data and artificial intelligence. *Biotechnology Journal*, p. e1800613, mar 2019. Disponível em: <http://dx.doi.org/10.1002/biot.201800613>. Citado na página 31.

ONALI, S.; FAVALE, A.; FANTINI, M. C. The resolution of intestinal inflammation: The peace-keeper's perspective. *Cells*, v. 8, n. 4, apr 2019. Disponível em: <http://dx.doi.org/10.3390/cells8040344>. Citado 2 vezes nas páginas 22 e 23.

PASCAL, V. et al. A microbial signature for crohn's disease. *Gut*, v. 66, n. 5, p. 813–822, feb 2017. Disponível em: <http://dx.doi.org/10.1136/gutjnl-2016-313235>. Citado 2 vezes nas páginas 15 e 64.

PETERSON, L. W.; ARTIS, D. Intestinal epithelial cells: regulators of barrier function and immune homeostasis. *Nature Reviews. Immunology*, v. 14, n. 3, p. 141–153, mar 2014. Disponível em: <http://dx.doi.org/10.1038/nri3608>. Citado 2 vezes nas páginas 7 e 26.

POWELL, D. W. et al. Mesenchymal cells of the intestinal lamina propria. *Annual Review of Physiology*, v. 73, p. 213–237, 2011. Disponível em: <http://dx.doi.org/10.1146/annurev.physiol.70.113006.100646>. Citado na página 24.

PRESLEY, L. L. et al. Host-microbe relationships in inflammatory bowel disease detected by bacterial and metaproteomic analysis of the mucosal-luminal interface. *Inflammatory Bowel Diseases*, v. 18, n. 3, p. 409–417, mar 2012. Disponível em: <http://dx.doi.org/10.1002/ibd.21793>. Citado na página 31.

RATH, E.; MOSCHETTA, A.; HALLER, D. Mitochondrial function - gatekeeper of intestinal epithelial cell homeostasis. *Nature Reviews. Gastroenterology & Hepatology*, v. 15, n. 8, p. 497–516, aug 2018. Disponível em: <http://dx.doi.org/10.1038/s41575-018-0021-x>. Citado na página 62.

RAUEN, T. et al. YB-1 acts as a ligand for notch-3 receptors and modulates receptor activation. *The Journal of Biological Chemistry*, v. 284, n. 39, p. 26928–26940, sep 2009. Disponível em: <http://dx.doi.org/10.1074/jbc.M109.046599>. Citado na página 62.

REDESTIG, H. et al. Data integration, metabolic networks and systems biology. In: ROBERTS, J. A. et al. (Ed.). *Annual Plant Reviews*. Chichester, UK: John Wiley & Sons, Ltd, 2018. p. 261–316. ISBN 9781119312994. Disponível em: <http://doi.wiley.com/10.1002/9781119312994.apr0469>. Citado na página 31.

SABATINO, A. D. et al. Increased enterocyte apoptosis in inflamed areas of crohns disease. *Diseases of the Colon & Rectum*, v. 46, n. 11, p. 1498–1507, nov 2003. ISSN 0012-3706. Disponível em: <https://insights.ovid.com/crossref?an=00003453-200346110-00008>. Citado 2 vezes nas páginas 29 e 61.

SAITOH, T. et al. Loss of the autophagy protein Atg16L1 enhances endotoxin-induced IL-$1\beta production. Nature, v.\ 456, n.\ 7219, p. 264--268, nov 2008. ISSN 0028-0836. Disponível em$: <>. Citado na página 30.

SCHULTZ, M. et al. Lactobacillus GG in inducing and maintaining remission of crohn's disease. *BMC Gastroenterology*, v. 4, p. 5, mar 2004. Disponível em: <http://dx.doi.org/10.1186/1471-{230X}-4>. Citado na página 65.

SEKIROV, I. et al. Gut microbiota in health and disease. *Physiological Reviews*, v. 90, n. 3, p. 859–904, jul 2010. Disponível em: <http://dx.doi.org/10.1152/physrev.00045.2009>. Citado na página 15.

SERAFINO, A. et al. WNT-pathway components as predictive markers useful for diagnosis, prevention and therapy in inflammatory bowel disease and sporadic colorectal cancer. *Oncotarget*, v. 5, n. 4, p. 978–992, feb 2014. Disponível em: <http://dx.doi.org/10.18632/oncotarget.1571>. Citado na página 61.

SHEEHAN, D.; SHANAHAN, F. The gut microbiota in inflammatory bowel disease. *Gastroenterology Clinics of North America*, v. 46, n. 1, p. 143–154, jan 2017. Disponível em: <http://dx.doi.org/10.1016/j.gtc.2016.09.011>. Citado 3 vezes nas páginas 14, 15 e 16.

SHI, J. Defensins and paneth cells in inflammatory bowel disease. *Inflammatory Bowel Diseases*, v. 13, n. 10, p. 1284–1292, oct 2007. Disponível em: <http://dx.doi.org/10.1002/ibd.20197>. Citado 2 vezes nas páginas 24 e 29.

SHIM, J. O. Recent advance in very early onset inflammatory bowel disease. *Pediatric gastroenterology, hepatology & nutrition*, v. 22, n. 1, p. 41–49, jan 2019. Disponível em: <http://dx.doi.org/10.5223/pghn.2019.22.1.41>. Citado na página 62.

SHOEMAKER, B. A.; PANCHENKO, A. R. Deciphering protein-protein interactions. part II. computational methods to predict protein and domain interaction partners. *PLoS Computational Biology*, v. 3, n. 4, p. e43, apr 2007. Disponível em: <http://dx.doi.org/10.1371/journal.pcbi.0030043>. Citado 2 vezes nas páginas 32 e 33.

SNOEKS, L. et al. Tumor suppressor FOXO3 participates in the regulation of intestinal inflammation. *Laboratory Investigation*, v. 89, n. 9, p. 1053–1062, sep 2009. Disponível em: <http://dx.doi.org/10.1038/labinvest.2009.66>. Citado na página 62.

SONG, D. et al. FOXO3 promoted mitophagy via nuclear retention induced by manganese chloride in SH-SY5Y cells. *Metallomics: Integrated Biometal Science*, v. 9, n. 9, p. 1251–1259, sep 2017. Disponível em: <http://dx.doi.org/10.1039/c7mt00085e>. Citado na página 62.

STANGE, E. F.; WEHKAMP, J. Recent advances in understanding and managing crohn's disease. *F1000Research*, v. 5, p. 2896, dec 2016. Disponível em: <http://dx.doi.org/10.12688/f1000research.9890.1>. Citado 3 vezes nas páginas 15, 24 e 29.

STEPHEN, B.; HAJJAR, J. Overview of basic immunology for clinical investigators. *Advances in Experimental Medicine and Biology*, v. 995, p. 1–31, 2017. Disponível em: <http://dx.doi.org/10.1007/978-3-319-53156-4_1>. Citado 2 vezes nas páginas 17 e 18.

SUDHAKAR, P. et al. Targeted interplay between bacterial pathogens and host autophagy. *Autophagy*, p. 1–14, mar 2019. ISSN 1554-8627. Disponível em: <https://www.tandfonline.com/doi/full/10.1080/15548627.2019.1590519>. Citado 2 vezes nas páginas 27 e 64.

TRIPATHI, L. P.; SOWDHAMINI, R. Genome-wide survey of prokaryotic serine proteases: analysis of distribution and domain architectures of five serine protease families in prokaryotes. *BMC Genomics*, v. 9, p. 549, nov 2008. Disponível em: <http://dx.doi.org/10.1186/1471-2164-9-549>. Citado na página 63.

TüREI, D.; KORCSMáROS, T.; SAEZ-RODRIGUEZ, J. OmniPath: guidelines and gateway for literature-curated signaling pathway resources. *Nature Methods*, v. 13, n. 12, p. 966–967, nov 2016. Disponível em: <http://dx.doi.org/10.1038/nmeth.4077>. Citado na página 44.

UHLéN, M. et al. Proteomics. tissue-based map of the human proteome. *Science*, v. 347, n. 6220, p. 1260419, jan 2015. Disponível em: <http://dx.doi.org/10.1126/science.1260419>. Citado na página 42.

VANCAMELBEKE, M. et al. Genetic and transcriptomic bases of intestinal epithelial barrier dysfunction in inflammatory bowel disease. *Inflammatory Bowel Diseases*, v. 23, n. 10, p. 1718–1729, 2017. Disponível em: <http://dx.doi.org/10.1097/{MIB}.00000000000012>. Citado na página 50.

VERES, D. V. et al. ComPPI: a cellular compartment-specific database for protein-protein interaction network analysis. *Nucleic Acids Research*, v. 43, n. Database issue, p. D485–93, jan 2015. Disponível em: <http://dx.doi.org/10.1093/nar/gku1007>. Citado 2 vezes nas páginas 42 e 60.

WANG, L. et al. Autophagy and ubiquitination in salmonella infection and the related inflammatory responses. *Frontiers in cellular and infection microbiology*, v. 8, p. 78, mar 2018.

ISSN 2235-2988. Disponível em: <http://journal.frontiersin.org/article/10.3389/fcimb.2018.00078/full>. Citado 3 vezes nas páginas 9, 29 e 30.

WEIMER, B. C. et al. Whole cell cross-linking to discover host-microbe protein cognate receptor/ligand pairs. *Frontiers in microbiology*, v. 9, p. 1585, jul 2018. ISSN 1664-302X. Disponível em: <https://www.frontiersin.org/article/10.3389/fmicb.2018.01585/full>. Citado na página 33.

WILLING, B. P. et al. A pyrosequencing study in twins shows that gastrointestinal microbial profiles vary with inflammatory bowel disease phenotypes. *Gastroenterology*, v. 139, n. 6, p. 1844–1854.e1, dec 2010. Disponível em: <http://dx.doi.org/10.1053/j.gastro.2010.08.049>. Citado 5 vezes nas páginas 16, 46, 55, 63 e 65.

WRIGHT, E. K. et al. Recent advances in characterizing the gastrointestinal microbiome in crohn's disease: a systematic review. *Inflammatory Bowel Diseases*, v. 21, n. 6, p. 1219–1228, jun 2015. Disponível em: <http://dx.doi.org/10.1097/{MIB}.00000000000003>. Citado na página 64.

YANO, T.; KURATA, S. An unexpected twist for autophagy in crohn's disease. *Nature Immunology*, v. 10, n. 2, p. 134–136, feb 2009. ISSN 1529-2908. Disponível em: <http://www.nature.com/doifinder/10.1038/ni0209-134>. Citado 3 vezes nas páginas 27, 29 e 30.

YATIM, K. M.; LAKKIS, F. G. A brief journey through the immune system. *Clinical Journal of the American Society of Nephrology*, v. 10, n. 7, p. 1274–1281, jul 2015. Disponível em: <http://dx.doi.org/10.2215/{CJN}.100310>. Citado na página 17.

YELLABOINA, S. et al. DOMINE: a comprehensive collection of known and predicted domain-domain interactions. *Nucleic Acids Research*, v. 39, n. Database issue, p. D730–5, jan 2011. Disponível em: <http://dx.doi.org/10.1093/nar/gkq1229>. Citado 2 vezes nas páginas 33 e 42.

YILMAZ, B. et al. Microbial network disturbances in relapsing refractory crohn's disease. *Nature Medicine*, v. 25, n. 2, p. 323–336, feb 2019. ISSN 1078-8956. Disponível em: <http://www.nature.com/articles/s41591-018-0308-z>. Citado na página 63.

ZUO, T.; NG, S. C. The gut microbiota in the pathogenesis and therapeutics of inflammatory bowel disease. *Frontiers in microbiology*, v. 9, p. 2247, sep 2018. Disponível em: <http://dx.doi.org/10.3389/fmicb.2018.02247>. Citado na página 63.

APÊNDICE A – Produção científica

A.1 Artigo em preparação para ser submetido

Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Manuscript Category

OXFORD

Subject Section

A robust integrated computational pipeline for inferring microbe-host interactions using protein-protein interaction predictions and heterogeneous biological datasets

Tahila Andrighetti [4], Padhmanand Sudhakar [1,2,3], Leila Gul [1], Ney Lemke [4], and Tamas Korcsmaros [1,2]*

[1]Earlham Institute, Norwich Research Park, Norwich, NR4 7UZ, UK
[2]Quadram Institute, Norwich Research Park, Norwich, NR4 7UA, UK
[3]Department of Chronic Diseases, Metabolism and Ageing, KU Leuven, Leuven, Belgium
[4]Sao Paulo University (UNESP), Institute of Biosciences, Botucatu, Brazil.

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Host-microbiome interactions are important to all living organisms. Understanding it has the potential to boost technology advance in diverse fields as medical, agriculture, ecology and others. Although, experimental techniques presents limitations and disavantages which hamper a deeper analysis. Regarding this, the development of computational approaches to predict host-microbiome interactions are indispensable. Here, we present a pipeline to predict host-microbiome interactions and construct a model to infere the human molecular mechanisms which can be modulated by the microrganisms.
Contact: Tamas.Korcsmaros@earlham.ac.uk
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Host-microbiome interactions happens in all plants and animals, shaping their metabolism and evolution [Barrett and Wu, 2017, Weimer *et al.*, 2018, Simon *et al.*, 2019]. In many ecosystems, microbial communities play important roles in the dynamics by decomposing organic matter and releasing nutrients as sulfur, carbon, nitrogen and oxygen, to be acquired by other organisms [Wooley *et al.*, 2010].

When associated with host organisms, microbial communities influence host physiology and health. For example, microbial communities present in and around plant roots potentially induce beneficial biochemical activity and hormonal responses to water stress [Fitzpatrick *et al.*, 2018]. In the intestine of cattle, microbial communities produce enzymes for cellulose degradation [Morgavi *et al.*, 2013].

In humans, the microbial count amounts to more than 100 trillion of bacterial cells: this is 10 times more than the number of human cells [Di Bella *et al.*, 2013]. This community of microorganisms are indispensable to the human life since they modulate and influence immunity and nutrient acquisition. For example, the skin microbiome helps in protecting and stimulating the immune system [Abdallah *et al.*, 2017]. Moreover, the gastrointestinal microbiome plays a crucial role in nutrient assimilation and energy yield by actively participating in metabolic pathways. Accordingly, dysbiosis of microbial communities can induce diseases such as type 2 diabetes, obesity and inflammatory bowel diseases as Crohn‘s disease [Devaraj *et al.*, 2013, Loh and Blaut, 2012].

However, microorganisms need not always be beneficial the host. Pathogenic microbes are known to cause harm to the host as well the microbial community by excluding beneficial species . Phytopathogens for example can be very deleterious to plants in their natural environments as well in agricultural settings, thus impairing agricultural productivity [Khater *et al.*, 2017]. Other bacteria, such as Shigella and Salmonella, are not only responsible for human diseases, but also for productivity loss in the poultry industry [Foley *et al.*, 2013].

Therefore, the study of host-microbiome interactions are important for the advancement of many fields as agriculture, livestock and human health. These cross-talks are mediated by molecular interactions between and among compounds and proteins expressed by the host and the microbiome. Bacterial molecules, as metabolites [Heinken *et al.*, 2013], proteins [Nourani *et al.*, 2015] and small RNAs [Katiyar-Agarwal and Jin, 2010], can interact with the host molecules launching a intracellular cascade able to modulate biological processes by affecting key genes and proteins.

Protein-protein interactions (PPIs) are one of the most relevant and studied type of molecular interactions between host and microbes. However, experimental techniques to probe inter-species PPIs are time consuming and financially expensive [Nourani *et al.*, 2015, Dyer *et al.*, 2008, 2010, Gould *et al.*, 2018, Weimer *et al.*, 2018, Foley *et al.*, 2013], and also limit available validated inter-species PPIs present in open databases. From the inferred microbe-host PPIs, it is possible to build interaction networks to better understand the interactions among the microbial and host components and how these interactions interfere on host metabolism and physiology [Guven-Maiorov *et al.*, 2017].

Here we describe a pipeline to analyse host-microbiome interactions at a molecular level using network approaches. A central point of our pipeline is the prediction of interactions between host and bacterial proteins using a domain-domain and domain-motif approach. From the filtered predicted interactions between the host and bacterial proteins, we compile a host multilayer network using experimentally verified downstream interactions which reach a selected set of user-provided gene/protein list. This list can be compiled either from apriori knowledge derived from phenotypic observations or contextual data obtained from -omic experiments. Various types of datasets, such as as gene expression, can then be used to validate the biological relevance of the network.

2 Pipeline description

As a first step of the workflow, bacterial and host protein lists for the interaction prediction are compiled. The bacterial proteins can be obtained from annotated bacterial proteomes or experimentally derived metaproteomes in the case of communities. As for the host proteins, the list is compiled based on their localization depending on whether or not the interactions between the bacterial and host proteins are spatially possible. Localization based filtering can be applied also to bacterial proteins depending on the context. For example, in the case of microbes which are extracellular to the host, the host proteins are confined to those present at the extracellular matrix or cellular membrane, because these are the locations where they are more prone to interact with the bacterial proteins. In this case, bacterial proteins are narrowed down to extracellular or secreted or membrane bound subsets. Gold standard databases which are currently used in the pipeline to compile the human proteins are ComPPI [Veres *et al.*, 2015], MatrixDB [Launay *et al.*, 2015] and Human Protein Atlas (HPA) [Uhlen *et al.*, 2015]. The user can also provide their own pre-compiled lists.

The next step is the interaction prediction. In this step, we compare the annotated domains and motifs from the compiled proteins with the databases DOMINE [Yellaboina *et al.*, 2011] and ELM [Gouw *et al.*, 2018] which contain gold-standard lists of domain-domain and domain-motif interactions [Bader *et al.*, 2008, Korcsmaros *et al.*, 2013] which are already known. This procedure helps determine which host proteins via their respective domains interact with bacterial proteins via matching domains and/or motifs. Some studies such as [Sudhakar *et al.*, 2019, Evans *et al.*, 2009, Wojcik and SchÃ¤chter, 2001] have already successfully used this approach to predict PPIs.

Subsequently, the interactions are filtered to keep only sterically possible interactions which is performed by excluding interactions with

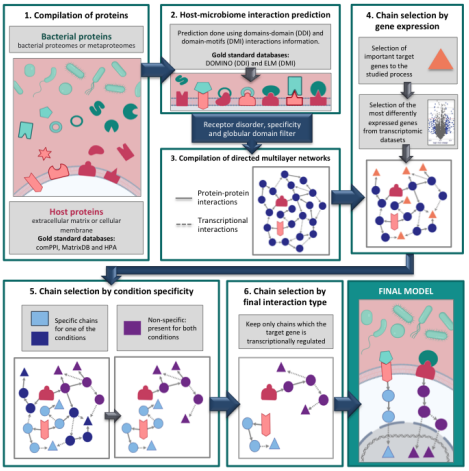

Fig. 1. Pipeline workflow

motifs outside disordered interactions (using IUPRED [DosztÃ¡nyi *et al.*, 2005]) and/or within globular domains (from PFAM [El-Gebali *et al.*, 2019], InterPro [Mitchell *et al.*, 2019]).

Subsequently, directed multilayered networks containing protein-protein interactions (PPIs) and transcriptional regulatory interactions (TRIs) connecting the host receptors potentially modulated by the bacterial proteins are compiled from molecular interaction databases. With this step, we get a list of potentially active molecular chains/pathways from the selected bacterial proteins to the host proteins of interest.

A final step involves prioritizing the molecular chains to those involving host proteins of interest such as those differentially expressed at the gene or protein level. Accordingly, the remaining network is constituted by directed chains which start with host receptor proteins and reaching selected target genes. Additionally, if the selected host target genes at the end of the chains are from a transcriptomic dataset, then the immediate upstream regulator of the selected host target genes is limited to a transcription factor.

Finally, the obtained model can be visualized by a multilayer network which starts with the bacterial-host receptor interactions, going through PPIs and TRIs and ultimately reaching the selected host target genes.

3 Pipeline use case

Crohn's disease (CD) is an Inflammatory Bowel Disease (IBD) characterized by chronic inflammation of colon and ileum which it's known to be caused by a combination of genetic and environmental factors [Baumgart and Sandborn, 2012]. Furthermore, there is a relevant change in the gut microbiome composition of CD patients in comparison with the healthy individuals [Mottawea *et al.*, 2016]. Together, these factors result in increased inflammation and altered cellular processes, and one of the most notorious impaired mechanisms in CD is autophagy [Henderson and Stevens, 2012].

Autophagy is a homeostatic process which is responsible for degrading non functional cellular components, as organelles and proteins, as well as invasive bacteria. This process has a basal activity in healthy cells and is modulated in stressful conditions such as nutrient starvation, oxidative

stress and pathogen infection [MÅ±zes *et al.*, 2013]. Moreover, autophagy can be modulated by the microorganisms from the microbiota or pathogens [Krokowski and Mostowy, 2016, Sudhakar *et al.*, 2019, Blacher *et al.*, 2017]. These conditions are observed in CD cells triggered by the increased inflammatory state.

Based on the fact that many meta-omic studies demonstrate significant differences in bacterial protein composition between healthy and CD individuals [Sheehan and Shanahan, 2017], we hypothesized that bacterial proteins are able to modulate autophagy in a positive manner in healthy cells, but impairing it in CD cells. With our pipeline, it is possible to study which are the differences in terms of the molecular mechanisms and pathways by which bacterial proteins can influence autophagy.

To address this, the user can first predict which are the human proteins which directly interact with bacterial proteins. For this step, they can use a differential gut metaproteome obtained from CD and healthy patients in order to consider the bacterial proteins uniquely present in either of the groups. The domains of these proteins are then compared with motifs of human proteins [Luck *et al.*, 2011] in order to predict if they are prone to interactions, according to domain-motif interaction information retrieved from ELM [Gouw *et al.*, 2018]. The human proteins selected to perform this prediction must be located in the cellular membrane and extracellular matrix because these are the locations where the interactions are more likely to happen. This information can be obtained from the databases ComPPI [Veres *et al.*, 2015], MatrixDB [Launay *et al.*, 2015] and Human Proteome Atlas [Uhlen *et al.*, 2015]. Subsequently, these interactions go through a disorder and structural filtering in order to eliminate physically unprovable interactions.

From the predicted microbial-host protein interactions, the user can compile multilayer networks containing protein-protein and transcriptional regulatory interactions. To this end, they can use various publicly available databases such as OmniPath [TÃ¼rei *et al.*, 2016], HTRI [Bovolenta *et al.*, 2012] and TRRUST [Han *et al.*, 2018] to compile the networks. Subsequently, bacterial proteins can then be added to the compiled network in order to obtain pathways which starts with bacterial proteins interacting with extracellular human proteins, followed by intracellular proteins and terminating with autophagy genes. These networks can then be filtered by retaining only specific pathways from healthy or CD conditions according to the bacterial protein specificity.

Thereafter, the user can obtain differential gene expression data from (healthy vs CD condition) publicly available CD expression datasets to select the most differentially expressed autophagy genes.

This compiled network allows the user to evaluate which modulated pathways were unique in CD and healthy conditions and how they regulate the expression of autophagy genes. From the analysis of such networks in our study, we higlight the role of mitophagy (autophagy of damaged mitochondrias), which in fact is impaired in CD and is a factor of increased oxidative stress and, consequently, enhanced inflammation. Besides autophagy, we could observe the influence of other biological pathways which are altered in CD cells as cellular proliferation and apoptosis, both of which are associated with increased inflammation when dysfunctional.

4 Conclusion

Host-microbiome interactions have considerable impacts on host phenotypes and their understanding is crucial making advances in various subjects as medicine, agriculture, environmental conservation to name a few. However, there are technological and methodological boundaries which challenge the science to investigate such interactions in a cheap and practical way. Therefore, the development of computation-enabled bioinformatic pipelines are in the right direction to understand the potential host-microbe interactions and the processes they modulate . Due to the infinitesimal possibilities in terms of all the interactions and pathways possible between the studied microbe and host, such computational pipelines will help prioritize on a selected number of pathways based on specific criteria such as contextual expression data etc.

Here, we presented a pipeline for the prediction of host-microbiome interactions using a domain-motif interaction theory and, subsequently, a systems biology approach to include the host-microbiome interactions in context of the host networks which capture the systemic context. The obtained networks allow us to yield insights about how the microbiome can interfere with host biological processes.

Funding

This work has been supported by the .....

References

Abdallah, F., Mijouin, L., and Pichon, C. (2017). Skin immune landscape: inside and outside the organism. *Mediators of Inflammation*, **2017**, 5095293.

Bader, S., KÃ¼hner, S., and Gavin, A.-C. (2008). Interaction networks for systems biology. *FEBS Letters*, **582**(8), 1220–1224.

Barrett, K. E. and Wu, G. D. (2017). Influence of the microbiota on host physiology - moving beyond the gut. *The Journal of Physiology*, **595**(2), 433–435.

Baumgart, D. C. and Sandborn, W. J. (2012). Crohn's disease. *The Lancet*, **380**(9853), 1590–1605.

Blacher, E., Levy, M., Tatirovsky, E., and Elinav, E. (2017). Microbiome-modulated metabolites at the interface of host immunity. *Journal of Immunology*, **198**(2), 572–580.

Bovolenta, L. A., Acencio, M. L., and Lemke, N. (2012). HTRIdb: an open-access database for experimentally verified human transcriptional regulation interactions. *BMC Genomics*, **13**, 405.

Devaraj, S., Hemarajata, P., and Versalovic, J. (2013). The human gut microbiome and body metabolism: implications for obesity and diabetes. *Clinical Chemistry*, **59**(4), 617–628.

Di Bella, J. M., Bao, Y., Gloor, G. B., Burton, J. P., and Reid, G. (2013). High throughput sequencing methods and analysis for microbiome research. *Journal of Microbiological Methods*, **95**(3), 401–414.

DosztÃ¡nyi, Z., Csizmok, V., Tompa, P., and Simon, I. (2005). IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics*, **21**(16), 3433–3434.

Dyer, M. D., Murali, T. M., and Sobral, B. W. (2008). The landscape of human proteins interacting with viruses and other pathogens. *PLoS Pathogens*, **4**(2), e32.

Dyer, M. D., Neff, C., Dufford, M., Rivera, C. G., Shattuck, D., Bassaganya-Riera, J., Murali, T. M., and Sobral, B. W. (2010). The human-bacterial pathogen protein interaction networks of bacillus anthracis, francisella tularensis, and yersinia pestis. *Plos One*, **5**(8), e12089.

El-Gebali, S., Mistry, J., Bateman, A., Eddy, S. R., Luciani, A., Potter, S. C., Qureshi, M., Richardson, L. J., Salazar, G. A., Smart, A., Sonnhammer, E. L. L., Hirsh, L., Paladin, L., Piovesan, D., Tosatto, S. C. E., and Finn, R. D. (2019). The pfam protein families database in 2019. *Nucleic Acids Research*, **47**(D1), D427–D432.

Evans, P., Dampier, W., Ungar, L., and Tozeren, A. (2009). Prediction of HIV-1 virus-host protein interactions using virus and host sequence motifs. *BMC Medical Genomics*, **2**, 27.

Fitzpatrick, C. R., Copeland, J., Wang, P. W., Guttman, D. S., Kotanen, P. M., and Johnson, M. T. J. (2018). Assembly and ecological function of the root microbiome across angiosperm plant species. *Proceedings of the National Academy of Sciences of the United States of America*, **115**(6), E1157–E1165.

Foley, S., Johnson, T., Ricke, S., Nayak, R., and Danzeisen, J. (2013). Salmonella pathogenicity and host adaptation in chicken-associated serovars. *Microbiology and Molecular Biology Reviews*, **77**(4), 582–607.

Gould, A. L., Zhang, V., Lamberti, L., Jones, E. W., Obadia, B., Korasidis, N., Gavryushkin, A., Carlson, J. M., Beerenwinkel, N., and Ludington, W. B. (2018). Microbiome interactions shape host fitness. *Proceedings of the National Academy of Sciences of the United States of America*, **115**(51), E11951–E11960.

Gouw, M., Michael, S., SÃ¡mano-SÃ¡nchez, H., Kumar, M., Zeke, A., Lang, B., Bely, B., Chemes, L. B., Davey, N. E., Deng, Z., Diella, F., GÃ¼rth, C.-M., Huber, A.-K., Kleinsorg, S., Schlegel, L. S., Palopoli, N., Roey, K. V., Altenberg, B., RemÃ©nyi, A., Dinkel, H., and Gibson, T. J. (2018). The eukaryotic linear motif resource - 2018 update. *Nucleic Acids Research*, **46**(D1), D428–D434.

Guven-Maiorov, E., Tsai, C.-J., and Nussinov, R. (2017). Structural host-microbiota interaction networks. *PLoS Computational Biology*, **13**(10), e1005579.

Han, H., Cho, J.-W., Lee, S., Yun, A., Kim, H., Bae, D., Yang, S., Kim, C. Y., Lee, M., Kim, E., Lee, S., Kang, B., Jeong, D., Kim, Y., Jeon, H.-N., Jung, H., Nam, S., Chung, M., Kim, J.-H., and Lee, I. (2018). TRRUST v2: an expanded reference database of human and mouse transcriptional regulatory interactions. *Nucleic Acids Research*, **46**(D1), D380–D386.

Heinken, A., Sahoo, S., Fleming, R. M. T., and Thiele, I. (2013). Systems-level characterization of a host-microbe metabolic symbiosis in the mammalian gut. *Gut microbes*, **4**(1), 28–40.

Henderson, P. and Stevens, C. (2012). The role of autophagy in crohn's disease. *Cells*, **1**(3), 492–519.

Katiyar-Agarwal, S. and Jin, H. (2010). Role of small RNAs in host-microbe interactions. *Annual Review of Phytopathology*, **48**, 225–246.

Khater, M., de la Escosura-MuÃ±iz, A., and MerkoÃ§i, A. (2017). Biosensors for plant pathogen detection. *Biosensors & Bioelectronics*, **93**, 72–86.

Korcsmaros, T., Dunai, Z. A., Vellai, T., and Csermely, P. (2013). Teaching the bioinformatics of signaling networks: an integrated approach to facilitate multi-disciplinary learning. *Briefings in Bioinformatics*, **14**(5), 618–632.

Krokowski, S. and Mostowy, S. (2016). Interactions between shigella flexneri and the autophagy machinery. *Frontiers in cellular and infection microbiology*, **6**, 17.

Launay, G., Salza, R., Multedo, D., Thierry-Mieg, N., and Ricard-Blum, S. (2015). MatrixDB, the extracellular matrix interaction database: updated content, a new navigator and expanded functionalities. *Nucleic Acids Research*, **43**(Database issue), D321–7.

Loh, G. and Blaut, M. (2012). Role of commensal gut bacteria in inflammatory bowel diseases. *Gut microbes*, **3**(6), 544–555.

Luck, K., Fournane, S., Kieffer, B., Masson, M., NominÃ©, Y., and TravÃ©, G. (2011). Putting into practice domain-linear motif interaction predictions for exploration of protein networks. *Plos One*, **6**(11), e25376.

Mitchell, A. L., Attwood, T. K., Babbitt, P. C., Blum, M., Bork, P., Bridge, A., Brown, S. D., Chang, H.-Y., El-Gebali, S., Fraser, M. I., Gough, J., Haft, D. R., Huang, H., Letunic, I., Lopez, R., Luciani, A., Madeira, F., Marchler-Bauer, A., Mi, H., Natale, D. A., Necci, M., Nuka, G., Orengo, C., Pandurangan, A. P., Paysan-Lafosse, T., Pesseat, S., Potter, S. C., Qureshi, M. A., Rawlings, N. D., Redaschi, N., Richardson, L. J., Rivoire, C., Salazar, G. A., Sangrador-Vegas, A., Sigrist, C. J. A., Sillitoe, I., Sutton, G. G., Thanki, N., Thomas, P. D., Tosatto, S. C. E., Yong, S.-Y., and Finn, R. D. (2019). InterPro in 2019: improving coverage, classification and access to protein sequence annotations. *Nucleic Acids Research*, **47**(D1), D351–D360.

Morgavi, D. P., Kelly, W. J., Janssen, P. H., and Attwood, G. T. (2013). Rumen microbial (meta)genomics and its application to ruminant production. *Animal : an international journal of animal bioscience*, **7 Suppl 1**, 184–201.

Mottawea, W., Chiang, C.-K., MÃ¼hlbauer, M., Starr, A. E., Butcher, J., Abujamel, T., Deeke, S. A., Brandel, A., Zhou, H., Shokralla, S., Hajibabaei, M., Singleton, R., Benchimol, E. I., Jobin, C., Mack, D. R., Figeys, D., and Stintzi, A. (2016). Altered intestinal microbiota-host mitochondria crosstalk in new onset crohn's disease. *Nature Communications*, **7**, 13419.

MÃ±zes, G., Tulassay, Z., and Sipos, F. (2013). Interplay of autophagy and innate immunity in crohn's disease: a key immunobiologic feature. *World Journal of Gastroenterology*, **19**(28), 4447–4454.

Nourani, E., Khunjush, F., and DurmuÅŸ, S. (2015). Computational approaches for prediction of pathogen-host protein-protein interactions. *Frontiers in microbiology*, **6**, 94.

Sheehan, D. and Shanahan, F. (2017). The gut microbiota in inflammatory bowel disease. *Gastroenterology Clinics of North America*, **46**(1), 143–154.

Simon, J.-C., Marchesi, J. R., Mougel, C., and Selosse, M.-A. (2019). Host-microbiota interactions: from holobiont theory to analysis. *Microbiome*, **7**(1), 5.

Sudhakar, P., Jacomin, A.-C., Hautefort, I., Samavedam, S., Fatemian, K., Ari, E., Gul, L., Demeter, A., Jones, E., Korcsmaros, T., and Nezis, I. P. (2019). Targeted interplay between bacterial pathogens and host autophagy. *Autophagy*, pages 1–14.

TÃ¼rei, D., KorcsmÃ¡ros, T., and Saez-Rodriguez, J. (2016). OmniPath: guidelines and gateway for literature-curated signaling pathway resources. *Nature Methods*, **13**(12), 966–967.

Uhlen, M., Fagerberg, L., HallstrÃ¶m, B. M., Lindskog, C., Oksvold, P., Mardinoglu, A., Sivertsson, ., Kampf, C., SjÃ¶stedt, E., Asplund, A., Olsson, I., Edlund, K., Lundberg, E., Navani, S., Szigyarto, C. A.-K., Odeberg, J., Djureinovic, D., Takanen, J. O., Hober, S., Alm, T., Edqvist, P.-H., Berling, H., Tegel, H., Mulder, J., Rockberg, J., Nilsson, P., Schwenk, J. M., Hamsten, M., von Feilitzen, K., Forsberg, M., Persson, L., Johansson, F., Zwahlen, M., von Heijne, G., Nielsen, J., and PontÃ©n, F. (2015). Proteomics. tissue-based map of the human proteome. *Science*, **347**(6220), 1260419.

Veres, D. V., GyurkÃ³, D. M., Thaler, B., Szalay, K. Z., Fazekas, D., KorcsmÃ¡ros, T., and Csermely, P. (2015). ComPPI: a cellular compartment-specific database for protein-protein interaction network analysis. *Nucleic Acids Research*, **43**(Database issue), D485–93.

Weimer, B. C., Chen, P., Desai, P. T., Chen, D., and Shah, J. (2018). Whole cell cross-linking to discover host-microbe protein cognate receptor/ligand pairs. *Frontiers in microbiology*, **9**, 1585.

Wojcik, J. and SchÃ¤chter, V. (2001). Protein-protein interaction map inference using interacting domain profile pairs. *Bioinformatics*, **17 Suppl 1**, S296–305.

Wooley, J. C., Godzik, A., and Friedberg, I. (2010). A primer on metagenomics. *PLoS Computational Biology*, **6**(2), e1000667.

Yellaboina, S., Tasneem, A., Zaykin, D. V., Raghavachari, B., and Jothi, R. (2011). DOMINE: a comprehensive collection of known and predicted domain-domain interactions. *Nucleic Acids Research*, **39**(Database issue), D730–5.

A.2 Capítulo de livro

O seguinte capítulo será inserido no livro "Bioinformática: Contexto Computacional e Aplicações", cuja publicação está prevista para o segundo semestre do ano de 2019. A referida obra está sendo organizada pelos professores Scheila de Avila e Silva e Daniel Notari, juntamente com o acadêmico Gabriel Dall'Alba, da Universidade de Caxias do Sul (Caxias do Sul, Rio Grande do Sul).

Ferramentas de análise e processamento de metagenomas

Os micro-organismos são os seres-vivos mais abundantes da Terra. Bactérias, arqueias, vírus e microeucariotos (fungos e protozoários) fazem parte de todos dos ecossistemas terrestres que têm condições de suportar vida, desde os mais amenos - como solo, tecidos animais e vegetais e oceanos - até ambientes extremos, como fumarolas, minas ácidas e geleiras, onde muitas vezes são os únicos habitantes. Comunidades microbianas cumprem papéis cruciais na dinâmica dos ecossistemas, decompondo matéria morta e disponibilizando novamente nutrientes como enxofre, carbono, nitrogênio e oxigênio, para serem adquiridos por outros organismos (COUNCIL, 2007; WOOLEY; GODZIK; FRIEDBERG, 2010).

A habilidade de reciclagem de nutrientes torna os micro-organismos indispensáveis para a vida na Terra e atrai o interesse humano para aplicações que podem ser úteis em diversas áreas. Comunidades microbianas associadas a outros organismos influenciam na fisiologia do hospedeiro e contribuem para sua saúde e crescimento. A microbiota no intestino de bovinos produz enzimas para a digestão de celulose; o entendimento sobre a relação entre a digestão e as enzimas produzidas pelos microrganismos, fornecem informações que podem servir de embasamento para a melhoria da produção de leite e carne e também para a diminuição do impacto ambiental causado pela criação de gado (MORGAVI et al., 2013). No corpo de seres humanos, há mais de 100 trilhões de células de bactérias, dez vezes mais do que a quantidade das células do próprio corpo (BELLA et al., 2013). Esses micro-organismos são indispensáveis para a vida do ser humano, habitam muitos de seus tecidos garantindo imunidade e aquisição de nutrientes a seu corpo. A microbiota da pele auxilia na imunidade e proteção dos humanos. A composição microbiológica do trato gastrointestinal influencia na aquisição de nutrientes, no rendimento de energia e em diversas vias metabólicas e seu desequilíbrio pode facilitar a indução de doenças como diabetes tipo 2 e obesidade; a partir de estudos dessas microbiotas, podem ser desenvolvidos meios alternativos de tratamento e de prevenção de diversas doenças (DEVARAJ; HEMARAJATA; VERSALOVIC, 2013).

Em solos, a associação da composição microbiana com plantas é indispensável, desempenhando papéis na qualidade do solo e produtividade e saúde das plantas hospedeiras através de mecanismos diretos ou indiretos, como na mineralização da matéria orgânica do solo, ativação dos mecanismos de defesa de plantas e produção de antibióticos contra patógenos; o melhoramento de microbiomas do solo pode auxiliar para maior rendimento agrícola e controle de pestes, bem como aprimoramento de alimentos como vinhos e queijos (ZARRAONAINDIA et al., 2015). Em oceanos, podem ser observadas diferenças significativas nas comunidades microbianas em diferentes profundidades, influenciadas

por características ambientais como oxigenação, salinidade e temperatura; em ambientes marinhos poluídos observou-se a presença de genes de resistência a arsênico e a metais pesados e de redução de sulfato, refletindo a alta capacidade de adaptação dos micro-organismos (BIK, 2014).

Entretanto, apenas 1% dos micro-organismos podem ser cultivados em laboratório (HANDELSMAN, 2004), limitando consideravelmente a extensão a que estudos de microbiomas podem ser conduzidos a partir de meios de cultura. Essa dificuldade foi superada com o advento das tecnologias de sequenciamento de DNA que possibilitaram o estabelecimento de um novo campo de estudo inserido na genômica: a metagenômica. O termo, cunhado por Handelsman em 1998 (HANDELSMAN et al., 1998), define o estudo dos genomas de comunidades microbianas presentes em um determinado habitat a partir do DNA extraído desse ambiente, sem a necessidade de cultivo dos micro-organismos. Deste modo, permitiu a revelação da diversidade microbiana e genética de diversos sistemas biológicos, relações genômicas entre função e filogenia de organismos não cultiváveis e perfis evolucionários de comunidades, além de outras interações biomoleculares (MARCO, 2011; THOMAS; GILBERT; MEYER, 2012).

Como exemplos de grandes iniciativas baseadas nessa tecnologia, temos o Projeto Microbioma Humano (Human Microbiome Project), financiado pelo NIH (National Institutes of Health), e o consórcio europeu MicroWine. O primeiro tem como objetivo sequenciar o metagenoma de partes do corpo humano, como cavidade gastro-intestinal, olhos, pele, vias aéreas, trato urogenital e sangue, para esclarecer o papel do microbioma na saúde e desenvolver novas ferramentas que possam ser utilizados posteriormente em prol de outras pesquisas (PETTERSSON; LUNDEBERG; AHMADIAN, 2009). Já o MicroWine, explora comunidades de micro-organismos que desempenham papéis importantes em todos os estágios da vinicultura - auxiliando o acesso das plantas a nutrientes do solo e na sua imunidade contra patógenos - até os processos de vinificação, que influenciam nos sabores e aromas característicos de cada vinho (MICROWINE, A MARIE CURIE INITIAL TRAINING NETWORK, 2016).

A primeira etapa de qualquer estudo metagenômico envolve a retirada das amostras do ambiente de estudo e posterior isolamento, fragmentação e sequenciamento do material genético dos micro-organismos relacionados àquele meio. Há três gerações de métodos de sequenciamento. Os métodos de primeira e segunda geração fragmentam o DNA em segmentos (*reads*) cujos comprimentos variam entre 35 e 700 pares de base. Devido à sua natureza, a análise dos dados metagenômicos resultantes dessas técnicas através de ferramentas computacionais torna-se bastante complexa. O método de primeira geração, também chamado de sequenciamento de Sanger, ainda é utilizado devido à sua baixa taxa de erros e *reads* relativamente longos, com mais de 700 pb, facilitando a análise pós-sequenciamento. Entretanto, seu custo é mais elevado do que

das plataformas de nova geração – U$ 400 mil por gigabase – e limita-se a até 96 Kb de informação por sequenciamento. Em contrapartida, as plataformas de segunda geração podem chegar a custar U$ 50,00 por gigabase e retornam mais de 1 Gb por sequenciamento. Consequentemente, essas tecnologias vêm substituindo o sequenciamento de Sanger, através de plataformas como Illumina/Solexa, 454/Roche e Applied Biosystems SOLiD. Por sua vez, o sequenciamento de terceira geração, também conhecido como sequenciamento de molécula única, propõe o rendimento de mais dados a menores custos e *reads* de tamanho maior do que 10 mil pb; as duas tecnologias de sequenciamento de molécula única mais utilizados são Pacific Biosciences e Oxford Nanopore. Apesar de suas vantagens, sequenciamentos de terceira geração ainda são pouco utilizados na metagenômica devido a sua alta taxa de erros (Tabela) (MOROZOVA; MARRA, 2008; THOMAS; GILBERT; MEYER, 2012; LAND et al., 2015; OULAS et al., 2015; LEE et al., 2016).

Tabela 1 – Lista de plataformas de sequenciamento, o tamanho de seus *reads* e seu custo por GB (MOROZOVA; MARRA, 2008; THOMAS; GILBERT; MEYER, 2012; LEE et al., 2016).

Geração	Tecnologia de sequenciamento	Tamanho dos *reads*	Custo por GB (aprox.)	Rendimento por corrida
1ª	Sanger	>700 pb	U$ 400 000	96 kb
2ª	454 / Roche	400 - 700 pb	U$ 20 000	80 - 120 Mb
2ª	Illumina	100 - 150 pb	U$ 50	1 Gb
2ª	Life Technologies / SOLiD	35 - 75 pb	U$ 130	1 - 3 Gb
3ª	Pacific Biosciences	10 - 15 kpb	U$ 500	5 Gb
3ª	Oxford Nanopore Technologies	5 - 10 kpb	U$ 1000	>40 Gb

A diminuição de preço das tecnologias de sequenciamento de segunda e terceira geração (NGS, do inglês *new generation sequencing*, ou sequenciamento de nova geração) permitiu a popularização da metagenômica entre os pesquisadores, e, consequentemente, o aumento na quantidade de dados disponibilizada em bancos de dados. Entretanto, o poder computacional e o desenvolvimento de algoritmos de análise de metagenomas não acompanha o crescimento na quantidade de dados produzidos. O primeiro problema está relacionado com a disponibilidade dos metagenomas: os sistemas de armazenamento de sequências não suportam quantidades de dados tão massivas e o formato dos dados não é padronizado. Outro obstáculo está relacionado às características dos dados produzidos: *reads* muito curtos e grande quantidade de erros gerados pelas plataformas de nova geração fazem com que a análise de metagenomas demande algoritmos mais complexos e custosos comp utacionalmente. Deste modo, é evidente a necessidade de novas ferramentas de análise de metagenomas para a maioria das etapas do processamento de dados pós-sequenciamento (KIM et al., 2013; KUMAR et al., 2015).

Existem duas categorias para o sequenciamento de metagenomas: na primeira, um gene marcador, mais frequentemente o 16S rRNA, é isolado através de PCR e sequenciado;

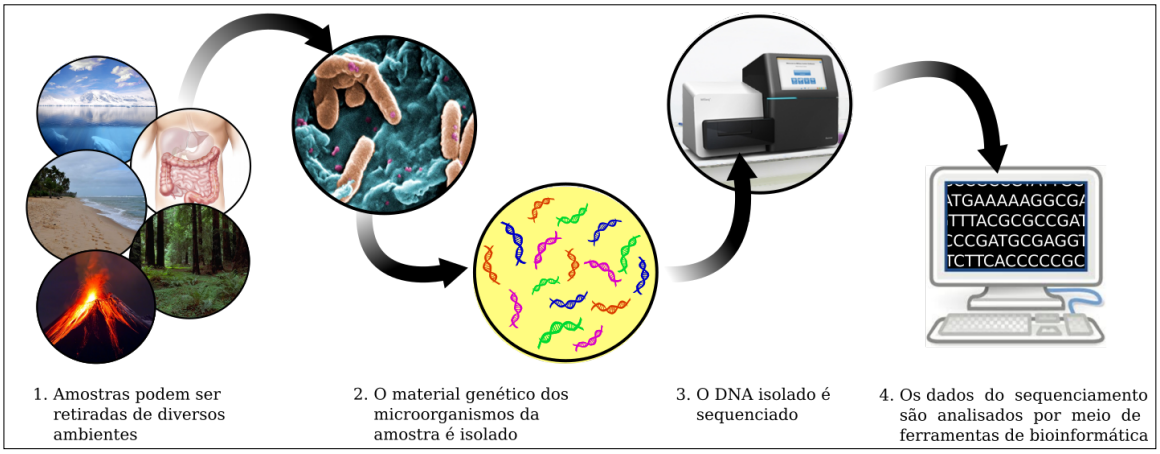

Figura 1 – Etapas da análise metagenômica de um ambiente.

já na metagenômica por WGS, do inglês, *whole genome shotgun*, todo o DNA dos microorganismos presentes na amostra é sequenciado (OULAS et al., 2015). A escolha da técnica mais adequada depende do objetivo de análise dos dados.

Há métodos de análise e ferramentas para trabalhar especificamente com os dados de cada abordagem. Portanto, entraremos em detalhes sobre as ferramentas utilizadas dentro das seções abaixo que discorrem separadamente sobre metagenômica a partir de 16S rRNA e de WGS.

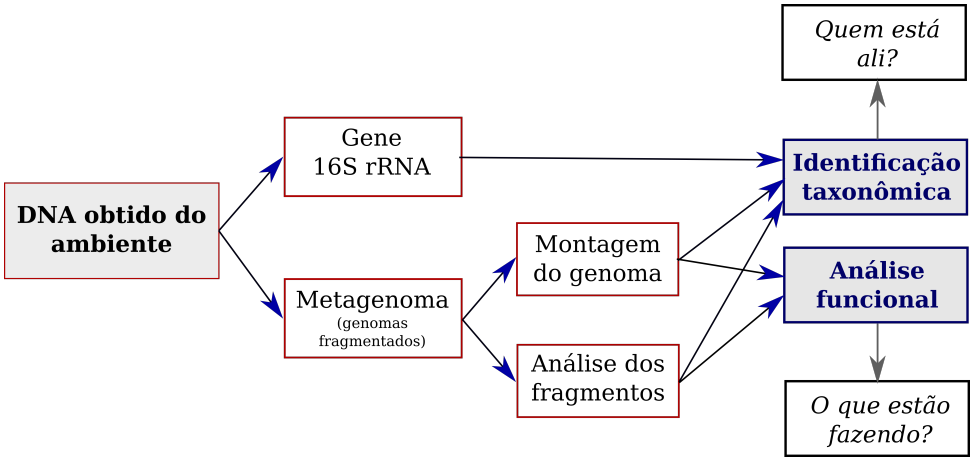

Figura 2 – O DNA obtido do ambiente pode ser analisado de duas formas: i) a partir da informação de somente um gene, sendo o mais utilizado o 16S rRNA, ou ii) a partir da informação de todo o DNA sequenciado, ou metagenoma.

1 Metagenômica a partir do gene 16S rRNA

Embora os estudos a partir de WGS estejam sendo desenvolvidos com frequência cada vez mais alta, a análise dos 16S rRNA ainda é amplamente aceita e é uma ferramenta poderosa para o estudo das comunidades microbianas em alta resolução (SUN et al., 2011). A utilização do gene 16S rRNA como marcador taxonômico possibilitou o desenvolvimento de um método de identificação de micro-organismos de uma amostra sem a necessidade de cultivo dos micro-organismos. A primeira execução bem sucedida ocorreu em 1991, quando foram registrados novas espécies a partir da análise dos genes 16S rRNA de amostras de oceano (SCHMIDT; DELONG; PACE, 1991; RIESENFELD; SCHLOSS; HANDELSMAN, 2004).

A inclusão da análise taxonômica de microbiomas a partir do gene 16S rRNA no conceito de "metagenômica"ainda é uma controvérsia entre os pesquisadores. Muitos deles sugerem que esse tipo de análise seja denominada "metagenética", por utilizar apenas um gene e não todo o genoma (ESPOSITO; KIRSCHBERG, 2014). Entretanto, para fins didáticos, nesse capítulo consideraremos que a análise taxonômica a partir do gene 16S rRNA também faz parte do campo da metagenômica.

O gene 16S rRNA codifica a subunidade pequena do RNA ribossômico de Archaeas e Bactérias e mostrou-se adequado por apresentar regiões hiperconservadas intercaladas com regiões variáveis ao longo de sua sequência. As regiões conservadas são quase idênticas dentre os micro-organismos, portanto são utilizadas para desenvolver *primers* universais para o isolamento dos genes da amostra. As outras regiões variam proporcionalmente à proximidade filogenética entre os táxons, sendo portanto utilizadas como parâmetros de comparação para a identificação dos micro-organismos (Figura 3) (KIM et al., 2013; NIKOLAKI; TSIAMIS, 2013).

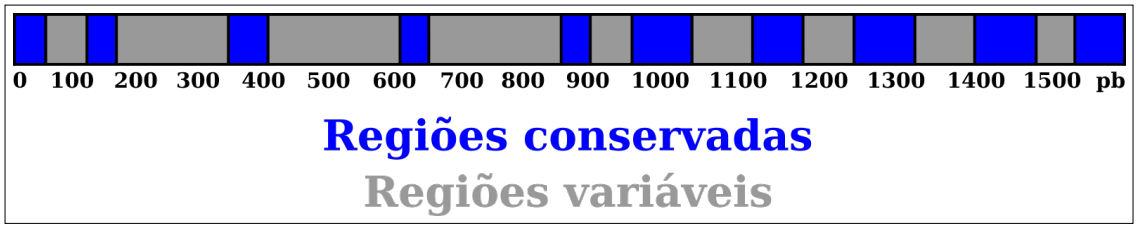

Figura 3 – 16S rRNA

Entretanto, as regiões variáveis do gene 16S rRNA apresentam baixa resolução entre as espécies, permitindo a classificação eficiente dos micro-organismos de um metagenoma somente até o nível de gênero. Outro obstáculo da técnica 16S rRNA na metagenômica é a alta susceptibilidade a vieses, pois os *primers* podem apresentar mais afinidade por sequências de determinadas espécies. Essa tendencia pode favorecer a seleção dos genes alguns organismos em detrimento de outros em uma amostra e impedir que novos

táxons sejam representadas se não forem compatíveis aos *primers* utilizados (NIKOLAKI; TSIAMIS, 2013; PORETSKY et al., 2014).

Mesmo com as limitações apresentadas, a metagenômica por 16S rRNA ainda tem sido mais utilizada para análises taxonômicas e filogenéticas do que a análise por WGS. Isso porque o advento das tecnologias de sequenciamento de nova geração não facilitou somente o sequenciamento de genomas inteiros, mas desenvolveram novas tecnologias específicas para as análises com o 16S rRNA, tornando esse método mais rápido, fácil e barato. Consequentemente, a quantidade dos dados de referência se ampliou significativamente, facilitando suas análises (KUNIN et al., 2008; TRINGE; HUGENHOLTZ, 2008).

O primeiro passo para a execução da metagenômica a partir do gene 16S rRNA é a obtenção do DNA dos micro-organismos de um meio ambiente. Depois de isolado, esse DNA passa pelo processo de PCR, onde ocorre a amplificação dos genes a partir de *primers* que identificam a localização do gene 16S rRNA e o replica diversas vezes. O produto dessa amplificação é submetido à técnica de eletroforese, onde o gene pode ser identificado em um gel por meio de bandas. As bandas que correspondem ao gene 16S rRNA são selecionadas e o DNA contido é purificado e submetido ao sequenciamento (SANSCHAGRIN; YERGEAU, 2014; ZHOU et al., 2015).

Depois do sequenciamento, os dados resultantes (*output*) são armazenados em um computador e devem ser processados e analisados a partir de ferramentas de bioinformática.

1.1 Pré-processamento dos dados brutos

O *output* do sequenciamento de DNA é um conjunto de dados brutos com erros de replicação e sequências de baixa qualidade que prejudicam a análise dos genes sequenciados. Para a análise mais precisa dos metagenomas, é necessário realizar uma etapa denominada, em inglês, *denoising*, que consiste no pré-processamento dos dados brutos. No *denoising*, realiza-se uma filtragem das sequências para que reste somente as que representam a comunidade microbiana com qualidade (KIM et al., 2013; OULAS et al., 2015).

Abaixo estão citados algumas ferramentas de *denoising* mais utilizadas:

- **PyroNoise:** Ferramenta de *denoising* dos dados obtidos pelo pirosequenciamento 454, da Roche, uma das mais utilizadas para a metagenômica a partir de 16S rRNA. Os dados brutos gerados por essa plataforma são fluxogramas que representam os *reads* sequenciados. O PyroNoise clusteriza esse fluxograma e utiliza uma medida de distância que modela o ruído do sequenciamento. Esse método permite a identi-

ficação das sequências verdadeiras em meio aos fragmentos sequenciados (QUINCE et al., 2009; GASPAR; THOMAS, 2013).

- **Amplicon-Noise:** Essa ferramenta utiliza primeiramente o algoritmo do PyroNoise para remover os erros a partir da clusterização de fluxogramas, mas sem necessitar de alinhamento, como é o caso do *software* original do PyroNoise (QUINCE et al., 2011; GASPAR; THOMAS, 2013).
- **QIIME:** É uma versão mais rápida do PyroNoise original. O algoritmo alinha e clusteriza os fluxogramas em um único passo, assim levando em conta tanto os erros do pirosequenciamento quando os do PCR. Aceita dados em formato fastq, portanto pode ser utilizado para outras plataformas além da 454 Roche (CAPORASO et al., 2010; GASPAR; THOMAS, 2013).
- **DADA, Divisive Amplicon Denoising Algorithm:** Realiza o *denoising* executando um modelo estatístico paramétrico de substituição de erros juntamente com um algoritmo de agrupamento hierárquico divisivo (ROSEN et al., 2012).

As sequências de baixa qualidade a serem retiradas dos dados incluem quimeras. Quimeras são recombinantes artificiais que se formam entre duas ou mais sequências durante a amplificação do DNA por PCR. Elas normalmente formam-se quando fragmentos de DNA que terminam prematuramente a amplificação anelam-se a outros. Essas moléculas artificiais dificultam a diferenciação das sequências originais das recombinantes, resultando na superestimação do nível de diversidade microbiana presente na amostra (KIM et al., 2013).

Entretanto, a detecção das quimeras na amostra não é um processo trivial, uma vez que a união das moléculas ocorre em posições aleatórias e as plataformas de NGS geram Bases de dados de 16S rRNA e *reads* curtos, dificultando a localização das sequências originais que possuam informação taxonômica suficiente (KIM et al., 2013).

Há algumas ferramentas especificas para a remoção de quimeras dos dados obtidos por NGS. O *software* ChimeraSlayer realiza o alinhamento das sequências obtidas com as de bancos de dados para identificar possíveis quimeras em meio aos dados e retirá-las (HAAS et al., 2011). A maioria das outras ferramentas, como Perseus (QUINCE et al., 2011), Decipher (WRIGHT; YILMAZ; NOGUERA, 2012) e UCHIME (EDGAR, 2010), utilizam informações de frequências de sequências para detectar as quimeras, assumindo que sequências quiméricas são menos frequentemente representadas do que as normalmente amplificadas.

1.2 Caracterização taxonômica do microbioma

Depois que a filtragem dos dados foi executada, realiza-se a classificação taxonômica dos genes 16S rRNA. Os pesquisadores que optam por realizar seus estudos a partir do 16S rRNA normalmente visam a obtenção de um perfil taxonômico ou filogenético da comunidade microbiana em questão, para responder a pergunta *"quem está no meio?"*. A partir dessas informações é possível realizar estudos relacionados à evolução da comunidade microbiana, associação entre micro-organismos e comparação, composição microbiótica de diferentes meios, entre outros.

A classificação dos genes podem ocorrer basicamente de duas maneiras: de forma taxonomicamente dependente ou independente. Nas análises taxonomicamente dependentes, as sequências desconhecidas são comparadas com outras já classificadas presentes em bancos de dados e então atribuídas à táxons cujas sequências apresentaram maior similaridade. Em análises taxonomicamente independentes, as sequências são agrupadas de acordo com determinados índices de similaridade apenas comparando umas com as outras, sem utilizar bases de dados como referência (SUN et al., 2011). Embora muitas ferramentas apliquem os métodos separadamente, essas abordagens podem ser utilizadas concomitantemente para uma análise mais prática e aprofundada.

Os métodos de análise taxonômica de 16S rRNA e suas ferramentas serão descritas mais detalhadamente a seguir.

1.2.1 Análises taxonomicamente dependentes

Na execução das análises taxonomicamente dependentes, as sequências de 16S rRNA desconhecidas são comparadas com sequências conhecidas disponíveis em bancos de dados e atribuídas aos táxons das que apresentam maior similaridade (SUN et al., 2011).

Uma das ferramentas mais utilizada para as análises taxonomicamente dependentes é o BLAST (ALTSCHUL et al., 1990). Esse *software* realiza o alinhamento das sequências desconhecidas com sequências de bancos de dados que podem ser fornecidos pelo usuário. Existem bancos de dados específicos com informações de RNA ribossomal que podem ser importados em análises que utilizam sequências de referencia, como SILVA (QUAST et al., 2013), EzTaxon-e (KIM et al., 2012), RDP (COLE et al., 2009) e Greengenes (DESANTIS et al., 2006).

As sequências alinhadas em algoritmos como BLAST precisam ser submetidas à análises do resultado em outras ferramentas que processam os *outputs* e apresentam as atribuições taxonômicas ao pesquisador. Exemplos ferramentas utilizadas para esse fim são MEGAN (HUSON et al., 2007), que utiliza algoritmo do menor ancestral em comum (LCA) e TANGO, que utiliza o método de (CLEMENTE; JANSSON; VALIENTE, 2011;

KIM et al., 2013).

assign ambiguous short reads to a node in the reference taxonomy minimizing a penalty score that generalizes our previous mapping based on the F-measure.

Outra alternativa para a abordagem taxonomicamente dependente é a comparação da similaridade das sequências por sua composição. RDP, um algoritmo que utiliza algoritmo de redes Bayesianas, utiliza a frequência de oligonucleotídeos de 8 bases das sequências para treinar as redes e atribuir táxons às sequências desconhecidas (WANG et al., 2007; KIM et al., 2013).

Existe ainda outro método de análise taxonomicamente dependente, que classifica as sequencias de acordo com sua alocação em uma árvore filogenética guia baseada em modelos evolucionários. E uma alternativa útil para casos em que não há sequencias de micro-organismos de táxons próximos as sequencias desconhecidas nos bancos de dados. Algoritmos que utilizam essa estrategia incluem SEPP (MIRARAB; NGUYEN; WARNOW, 2012), EPA (BERGER; KROMPASS; STAMATAKIS, 2011), pplacer (MATSEN; KODNER; ARMBRUST, 2010), QIIME (CAPORASO et al., 2010) e AMPHORA2 (WU; SCOTT, 2012; KIM et al., 2013).

1.2.2 Análises taxonomicamente independentes

A realização de análises taxonomicamente independentes normalmente baseiam-se na clusterização de OTUs. OTU, sigla em inglês *operational taxonomic unit*, significa “unidades taxonômicas operacionais"em português. OTUs são os agrupamentos dos genes 16S rRNA de acordo com sua similaridade. Níveis de similaridade maiores do que 97% para bactérias e archaeas correspondem à mesma espécie (PATIN et al., 2013; OULAS et al., 2015).

Algoritmos para clusterização de sequências de 16S rRNA utilizam basicamente duas estratégias: a partir de alinhamento de sequências, na qual as sequências desconhecidas de 16S rRNA são alinhadas entre elas, podendo haver ou não a utilização de sequências de referência; e estratégias independentes de alinhamento. Abaixo, estão citadas algumas ferramentas para clusterização de OTUs (KIM et al., 2013):

- **NAST:** compara cada sequência de 16S rRNA submetida alinhando-as com 10 mil sequências de referência alinhadas e não quiméricas de bactérias e archaeas conhecidas (DESANTIS et al., 2006).
- **SINA aligner:** realiza o alinhamento das sequências com um algoritmo baseado em ordem parcial (PRUESSE; PEPLIES; GLÖCKNER, 2012).
- **Infernal:** baseia-se nos alinhamentos de perfis de estruturas secundárias de RNA para realizar o agrupamento de OTUs (NAWROCKI; KOLBE; EDDY, 2009).

- **UCLUST:** organiza as sequências para diminuir o número de oligonucleotídeos em comum, explorando o fato que sequências similares tendem a ter pequenos oligonucleotídeos em comum (EDGAR, 2010).
- **CD-HIT:** primeiramente, utiliza a maior sequência de entrada como primeira representante do cluster. Depois, compara sequências restantes em ordem decrescente de tamanho para classificá-las como redundantes ou representativas em comparação com as representativas existentes. As similaridades são calculadas pela contagem dos oligonucleotídeos em comum (FU et al., 2012).
- **ESPIRIT-Tree:** emprega uma técnica de particionamento de espaço para organizar os objetos hierarquicamente em células. Deste modo, pode encontrar o vizinho mais próximo de cada objeto pode em células adjacentes utilizando uma estratégia de "dividir e conquistar" (CAI; SUN, 2011).

2 *Whole Genome Shotgun*

Embora a metagenômica a partir do gene 16S rRNA seja um modo eficaz para estudar a diversidade microbiana das comunidades e seu impacto nos ambientes em que estão presentes, é possível obter somente informações sobre a composição dos micro-organismos dos meios analisados. A partir de metagenômica por WGS, é possível adquirir as informações de biodiversidade relacionadas à composição funcional do meio, possibilitando a responder as questões *"quem está presente no ambiente?"*, *"o que esses micro-organismos estão fazendo?"* e *"como eles interagem?"* (FORDE; O'TOOLE, 2013; OULAS et al., 2015).

O procedimento para metagenômica por WGS também inclui o isolamento do material genético da amostra, assim como na metagenômica por 16S rRNA, embora no caso da WGS todo o DNA é submetido ao sequenciamento. Deste modo, o resultado do sequenciamento consiste em diversos segmentos de DNA dos genomas de micro-organismos presentes no meio. Como o genoma é aleatoriamente fragmentado, os *reads* do metagenoma pertencem a qualquer parte do genoma: desde genes taxonomicamente informativos, como o gene 16S rRNA, até sequências codificadoras que fornecem informações sobre funções que os micro-organismos podem realizar no ambiente (SHARPTON, 2014).

Por apresentar maior complexidade do que os metagenomas de 16S rRNA, o desenvolvimento de ferramentas para análise de metagenomas por WGS também é mais desafiador. Primeiramente, porque é mais difícil determinar a origem taxonômica dos *reads*. Depois, na maioria das vezes não é possível obter a representação de todos os micro-organismos do ambiente por causa de sua diversidade. Além disso, a etapa de filtragem é mais complicada do que a do 16S rRNA: a identificação e remoção de sequências

contaminantes é problemática devido à dificuldade de diferenciar as sequências dos micro-organismos das sequências de organismos indesejados. Também pode haver a necessidade de montar os genomas, etapa desafiadora para os bioinformatas devido ao alto custo computacional e demanda por grandes quantidades de informações (SHARPTON, 2014).

Essas limitações tem esmaecido com o avanço das tecnologias de informática, que vem apresentando computadores mais potentes ao longo do tempo, e com as tecnologias de sequenciamento, que prezam por maiores tamanhos de *reads*. Abaixo, estão detalhadas as etapas de análise de bioinformática para metagenômica por WGS e ferramentas utilizadas para cada finalidade.

2.1 Controle de qualidade das sequências

O primeiro passo a ser realizado depois que o metagenoma é sequenciado, é o controle de qualidade, ou filtragem dos dados retornados para retirar sequências de baixa qualidade do metagenoma. Essa é uma importante etapa a ser realizada, pois erros e sequências de organismos não desejados dificultam a montagem dos reads e sua análise, especialmente quando o contaminante é altamente abundante ou tem um grande genoma (BRAGG; TYSON, 2014; SHARPTON, 2014). Alguns dos programas utilizados para o controle de qualidade são plataforma-específicos. Os mais utilizados citados abaixo:

- FASTX toolkit: conjunto de linhas de comando para pré-processamento de reads curtos de dados FASTA/FASTQ (FASTX-TOOLKIT. . . , 2016);
- FASTQC: ferramenta de controle de qualidade para dados retornados de sequenciamento pela plataforma Illumina. FASTQC fornece uma interface gráfica para visualização e simplificação da filtragem desses dados (BABRAHAM BIOINFORMATICS, 2016);
- ngs backbone: aplicável a dados de sequenciamento de Sanger, 454, Illumina e SOLiD. Além da limpeza de dados, executa funções como: montagem e anotação de transcriptomas, leitura de mapeamento e busca e seleção por polimorfismos de nucleotídeo único (SNP, do inglês, single nucleotide polymorphism) (BLANCA et al., 2011);
- Pyrobayes: software de controle de qualidade de dados retornados por pirosequenciamento 454. Pyrobayes permite busca por SNPs em aplicações de resequenciamento, produzindo buscas mais confiáveis do que o programa nativo da plataforma (QUINLAN et al., 2008);
- Shore: realiza busca por polimorfismos em dados retornados por sequenciamento na plataforma Illumina (OSSOWSKI et al., 2008).

2.2 Montagem de fragmentos

Depois de pré-processados, os *reads* dos metagenomas podem ser submetidos à de montagem. Nessa etapa, os fragmentos são unidos com outros originados do mesmo genoma para formar sequências maiores e, assim, facilitar a análise. É praticamente impossível realizar a montagem de genomas inteiros a partir de dados de metagenômica, pois o genoma da maioria dos micro-organismos representados na amostra não é completamente sequenciada e é difícil atribuir exatamente a qual espécie cada *read* pertence. Entretanto, em alguns casos, é possível montar grande parte dos genomas para realizar estudos que requerem a estrutura do genoma, como por exemplo em análises funcionais que busquem por regiões codificantes. (WOOLEY; GODZIK; FRIEDBERG, 2010; KIM et al., 2013; SHARPTON, 2014).

A maioria dos *softwares* de montagem de genomas são desenvolvidos para junção de fragmentos obtidos por sequenciamento de genomas inteiros, do qual os *reads* são provindos de somente um organismo, sendo pouco eficazes para metagenomas onde as sequências apresentam diferentes origens. Além disso, a falta de genomas de referência de micro-organismos não culturalizáveis e o pequeno tamanho dos fragmentos gerados pelo sequenciamento tornam a tarefa ainda mais desafiadora (KUMAR et al., 2015).

Há dois tipos de montagem de genomas que pode ser executados: a montagem baseada em genomas referências e a montagem *de novo*.

2.2.1 Montagem baseada em genomas de referência

Utiliza-se um ou mais genomas de referência como “mapas"onde os *reads* podem ser posicionados. Quando dois ou mais *reads* dispõem-se um ao lado do outro, e possível realizar sua união e formar fragmentos maiores (Figura **??**).

Ferramentas utilizadas para a montagem a partir de genomas de referência incluem MetaAMOS (TREANGEN et al., 2013), Newbler (montagem de *reads* da 454-Roche) e MIRA4 (CHEVREUX et al., 2004). Essas ferramentas não são computacionalmente custosas, mas são adequadas para a aplicação em metagenomas de ambientes bem explorados, onde a composição microbiótica já é conhecida. Nesses casos, é mais provável que genomas de micro-organismos próximos aos presentes nesses ambientes já estejam disponíveis nos bancos de dados, podendo assim serem usados como referência.

2.2.2 Montagem de novo:

montagem de fragmentos sem a utilização de genomas de referência. Esses *softwares* apresentam algoritmos mais complexos do que os que utilizam genomas de referência, normalmente, como por exemplo grafos de-Bruijin (COMPEAU; PEVZNER; TESLER, 2011). Programas de montagem *de novo* também são computacionalmente mais caros,

demandando computadores com grande quantidades de memória e tomando longos tempos de execução. As montagens *de novo* de metagenomas são comumente executadas por ferramentas como Abyss (SIMPSON et al., 2009), Velvet (ZERBINO; BIRNEY, 2008), SOAP (LI et al., 2008) e EULER (PEVZNER; TANG; WATERMAN, 2001).

Próxima geração de ferramentas de montagem: a maioria dos algoritmos citados acima foram desenvolvidos para a montagem de *reads* sequenciados a partir de genomas únicos. Sua utilização para metagenomas, cujas sequências são originadas de diversos organismos, esbarram em alguns obstáculos: primeiramente, subespécies similares podem apresentar variações relevantes, assim como sequências de espécies diferentes podem ser muito similares; além disso, a diferença na quantidade de DNA de cada espécie na amostra também interfere na montagem do genoma. Outros algoritmos vem sido desenvolvidos para superar essas dificuldades.

Dois exemplos de *softwares* de nova geração são Meta-Velvet-SL (NAMIKI et al., 2012; AFIAHAYATI; MULYANA, 2015) e Meta-IDBA (PENG et al., 2011), que combinam ferramentas de *binning* (mais detalhes sobre *binning* são apresentados abaixo) e de montagem de *reads* para unir os fragmentos de metagenômica com mais acurácia. Esses programas utilizam valores de frequências de oligonucleotídeos (*k-mers*) para detectar torções nos grafos de-Bruijin e limiares de *k-mers* para decompor os grafos em sub-grafos. Deste modo, os fragmentos são conectados baseados nos sub-grafos decompostos e executando um agrupamento mais eficiente das sequências, possibilitando a separação das que pertencem a diferentes espécies.

2.3 Análise taxonômica e *binning*

Para analisar a diversidade taxonômica de uma amostra de metagenômica, pode ser realizado um processo denominado *binning* (MANDE; MOHAMMED; GHOSH, 2012). A partir desse procedimento os *reads*, que até então não têm sua origem taxonômica determinada, são agrupados em táxons de acordo com diferentes características da sequência. Além de quantificar os micro-organismos de diferentes táxons que estão presentes no meio, a partir do *binning* é possível reduzir a complexidade do conjunto de dados para facilitar posteriores fases do estudo, como a montagem de fragmentos ou análise funcional (SHARPTON, 2014).

A partir de análises taxonômicas é possível estudar o papel da composição das comunidades microbianas nos ecossistemas em que fazem parte. Zarraonaindia mostrou que a diversidade de espécies associadas aos órgão das parreiras (folhas, flores, frutos e raízes) e ao solo onde estão plantadas são importantes para definir o sabor final dos vinhos produzidos (ZARRAONAINDIA et al., 2015). Outro estudo demonstrou que a microbiota intestinal pode influenciar no desenvolvimento da obesidade (RIDAURA et al.,

2013). Resultados como os citados evidenciam a importância da composição microbiana nos ambientes em diversas áreas.

Apesar da importância das análises taxonômicas, esse procedimento representa grandes desafios para os pesquisadores e desenvolvedores de *software* principalmente por causa do pequeno tamanho dos *reads* obtidos pelas NGS. Por serem muito pequenos, muitas vezes não apresentam informações suficientes para que seja possível classificá-los. Consequentemente, muitos fragmentos acabam sendo excluídos do agrupamento. Para minimizar esses problemas, pode ser realizada a pré-montagem dos fragmentos e deve ser utilizada a ferramenta de análise mais adequada para o tipo de dados que se está trabalhando (KIM et al., 2013).

Existem basicamente duas categorias de ferramentas que utilizam diferentes abordagens para a classificação taxonômica dos *reads* de metagenômica: similaridade de sequências e de composição de sequências.

2.3.1 Ferramentas de similaridade de sequência:

classificam a sequência desconhecida de acordo com sua similaridade com as armazenadas em bancos de dados. Primeiramente, os dados são alinhados com ferramentas como BLAST. Posteriormente, os *softwares* de análise de metagenômica utilizam as informações para fazer inferências taxonômicas e filogenéticas (KIM et al., 2013). O método de similaridade de sequências para a classificação taxonômica dos metagenomas fornece maior resolução e acurácia da análise do que o binning a partir da composição de sequências (descrito na seção 3.3.2.). Entretanto, seu custo computacional é maior e aumenta exponencialmente com a diminuição do comprimento dos reads (LIU et al., 2013; SHARPTON, 2014). Abaixo, estão citadas ferramentas de análise taxonômica por similaridade popularmente utilizadas (SHARPTON, 2014): Abaixo, estão citadas ferramentas de análise taxonômica por similaridade popularmente utilizadas (SHARPTON, 2014):**MEGAN:** utiliza BLAST para comparar os *reads* de metagenômica com banco de dados de sequências que são anotadas com a taxonomia do NCBI. Depois, o *software* infere a taxonomia da sequência colocando o fragmento em um nó da árvore taxonômica do NCBI correspondente ao último ancestral comum (LCA, do inglês *last commom ancestor*) de todos os táxons que contém homologia com o *read* (HUSON et al., 2007). **MG-RAST:** utiliza reconstrução filogenética de sequências de banco de dados que sejam similares a cada *read* para classificá-lo taxonomicamente (MEYER et al., 2008). **CARMA:** utiliza os melhores alinhamentos recíprocos entre as sequências disponíveis em banco de dados e os *reads* de metagenômica e modelos de índices de evolução específicos gene-família para

inferir o rank taxonômica apropriado a cada fragmento de DNA (GERLACH; STOYE, 2011).

2.3.2 Ferramentas de composição da sequência

Utiliza as características intrínsecas da composição das sequências (e.g. conteúdo GC, frequência de oligonucleotídeos, utilização de códons, assinaturas periódicas) para classificá-las taxonomicamente. Essas características, também chamadas de *assinaturas genômicas*, são moldadas em cada grupo taxonômico ao longo da evolução de acordo com as pressões evolutivas às quais os micro-organismos estão sujeitos. Levando em conta que organismos filogeneticamente mais próximos apresentarão mais similaridade na composição de suas sequências, é possível agrupar ou classificar os *reads* de metagenômica de acordo com essas características (CAMPBELL; RUAN; WEI, 1999; THOMAS; GILBERT; MEYER, 2012).

As ferramentas que utilizam assinaturas genômicas são mais rápidas e custam menos computacionalmente do que as ferramentas de similaridade. Entretanto, apresentam dificuldade na identificação de fragmentos muito pequenos (i.e., 150 pb), pois eles não contém informação suficiente para uma classificação eficiente (MANDE; MOHAMMED; GHOSH, 2012).

Para agrupar e classificar os *reads*, os programas de composição de sequências frequentemente utilizam aprendizagem de máquina supervisionados ou não supervisionados. Os supervisionados utilizam genomas conhecidos para que o algoritmo reconheça os padrões de cada grupo taxonômico; os *reads* são atribuídos a um táxon de acordo com as características reconhecidas pela máquina. Algoritmos não supervisionados não utilizam sequências de referência, eles comparam um fragmento com outro reconhecendo padrões comuns entre eles e os agrupando em conjuntos de acordo com suas características compartilhadas. *Binning* a partir de algoritmos não supervisionados reúnem os *reads* em grupos taxonomicamente distintos, mas é necessária a utilização de ferramentas adicionais para atribuir táxons a cada agrupamento (BRAGG; TYSON, 2014).

Os *Softwares* mais popularmente utilizados para o *binning* por composição de sequências são (KIM et al., 2013; SHARPTON, 2014):

- **PhyloPithia e PhyloPithiaS:** utilizam o algoritmo de aprendizagem supervisionada de máquina *support vector machines*, que analisa sequências de treinamento já identificadas taxonomicamente para construir modelos de frequências de oligonucleotídeos que determinam se o *read* é um membro do grupo (MCHARDY et al., 2007; PATIL et al., 2011).
- **Phymm:** ferramenta supervisionada que utiliza modelos Markovianos interpolados que combinam probabilidades de predição derivadas de sequências

de treinamento com diversos tamanhos. Opcionalmente, aplica alinhamento com BLAST para a classificação dos *reads*. Adequado para sequências pequenas originadas de NGS.Adequado para sequências pequenas originadas de NGS. (BRADY; SALZBERG, 2009; BRADY; SALZBERG, 2011).

- **NBC:** utiliza classificador supervisionado de Naive Bayes baseando-se nos perfis de frequência de *k-mers* de cada grupo taxonômico. Adequado para sequências pequenas originadas de NGS (ROSEN; REICHENBERGER; ROSENFELD, 2011).
- **TACOA:** ferramenta não supervisionada que utiliza regra do vizinho mais próximo (*k-nearest neighbor*) para agrupar os *reads* (DIAZ et al., 2009).

2.3.3 Ferramentas híbridas

Para compensar as vantagens e desvantagens das categorias de algoritmos citados acima, há ferramentas que utilizam tanto a abordagem de composição de sequências quanto o alinhamento para a classificação taxonômica dos *reads* de metagenômica. Alguns exemplos estão descritos abaixo:

- **PhymmBL:** combina a ferramenta Phymm, descrita anteriormente, com alinhamentos em BLAST pata aumentar a precisão da classificação (BRADY; SALZBERG, 2009).
- **RITA:** combina BLAST com a ferramenta NBC (descrita anteriormente), mas considera os resultados do BLAST com mais peso (MACDONALD; PARKS; BEIKO, 2012).
- **SPHINX:** na primeira fase, SPHINX compara a composição de tetranucleotídeos do *read* com a dos genomas de referência para realizar uma pré-filtragem dos grupos taxonômicos a que ele possa pertencer. Depois, utiliza algoritmos de alinhamento de sequência para classificar o fragmento mais restritamente (MOHAMMED et al., 2011).

2.4 Análise funcional

A análise funcional dos metagenomas fornece informações sobre funções codificadas nos genomas dos micro-organismos da comunidade, respondendo à pergunta “o que os micro-organismos estão fazendo no ambiente estudado?". A partir da caracterização dos genes do metagenoma é possível traçar um perfil funcional da comunidade microbiana que pode ser utilizado para comparar metagenomas de diferentes ambientes, revelar a presença de novos genes ou fornecer informações do mesmo ambiente em diferentes condições (SHARPTON, 2014).

Para realizar a análise funcional do metagenoma, primeiramente identifica-se os fragmentos que contém sequências codificadoras e depois as compara com sequências de banco de dados de genes, proteínas, famílias de proteínas ou vias metabólicas com funções conhecidas para identificar qual a função dos genes desconhecidos (SHARPTON, 2014). As etapas estão detalhadas abaixo.

2.4.1 Identificação de genes

A busca por genes em meio ao metagenoma pode ser realizada com ou sem a montagem dos *reads*. Se os *reads* estiverem montados e as sequências codificadoras estiverem completas, a predição de genes pode ser muitas vezes realizada com os mesmos programas de busca de genes em genomas completos que não requiram parâmetros espécie-específicos, pois os *reads* de metagenomas são originados de diversas linhagens. A análise funcional de metagenomas não montados é mais desafiadora, pois envolve a predição de sequências codificadoras incompletas (SHARPTON, 2014). Independente da opção de montagem ou não dos fragmentos, a maioria dos programas de predição de genes utiliza informações de códons (i.e., *start* codon - AUG) para identificar quadros de leitura abertos (ORFs, do inglês, *open reading frames*) e assim classificar as sequências como codificadoras ou não codificadoras (OULAS et al., 2015). Os *softwares* de identificação de genes em metagenomas utilizam diferentes modelos de predição de genes, como aprendizagem de máquina (HAYES; BORODOVSKY, 1998), modelos ocultos de Markov (HMM, do inglês, *hidden Markov models*) (YADA et al., 1999) e tendência de utilização de dicódon (NGUYEN et al., 2009). Abaixo, estão citados alguns algoritmos de busca de genes em metagenomas (KIM et al., 2013).**MetaGeneMark:** utiliza tendência de uso de códons incorporados a HMMs (ZHU; LOMSADZE; BORODOVSKY, 2010). **Prodigal:** utiliza algoritmos de aprendizagem de máquina (HYATT et al., 2010). **MetaGene:** utiliza tendência de utilização de di-códons (TANENBAUM et al., 2010). **FragGeneScan:** utiliza modelos de erro de sequenciamento e HMM com tendência de utilização de códons incorporado (RHO; TANG; YE, 2010). **MetaGeneAnnotator:** utiliza algoritmos de aprendizagem de máquina com informações de tendência de utilização de di-códons (NOGUCHI; TANIGUCHI; ITOH, 2008). **Glimmer-MG:** utiliza HMM (SALZBERG et al., 1998). **Orphelia:** utiliza algoritmos de aprendizagem de máquina com informações de tendência de utilização de di-códons (HOFF et al., 2009).

2.4.2 Anotação funcional dos genes

Depois que os genes são identificados no metagenoma, a próxima etapa é encontrar a função desempenhada por eles no microbioma. Essa etapa é a mais computacionalmente custosa por causa do grande tamanho dos metagenomas e, muitas vezes, comprimento muito pequeno dos *reads*. Os genes encontrados no metagenoma são comparados com genes caracterizados disponíveis nos bancos de dados a partir de ferramentas de alinhamento de sequência, como BLAST, para inferir a função que desempenham no microbioma (OULAS et al., 2015).

Os bancos de dados fornecem informações sobre domínios e classificação de proteínas por suas funções. A comparação dos genes desconhecidos com as referências desses bancos de dados possibilitam a determinação de quais funções e vias estão presentes no metagenoma e sua quantidade. Os bancos de mais utilizados para a busca de informações de genes conhecidos incluem (BELLA et al., 2013): KEGG (*Kyoto Encyclopedia of Genes and Genomes*) (KANEHISA; GOTO, 2000), COG (*Clusters of Orthologous Groups system*) (TATUSOV et al., 2003), Pfam (BATEMAN et al., 2004), CDD (*Conserved Domains Database*) (MARCHLER-BAUER et al., 2005), SEED (OVERBEEK et al., 2005), TIGRFAM (SELENGUT et al., 2007) e eggNOG (MULLER et al., 2010).

2.4.3 Ferramentas generalizadas

Há ferramentas que foram desenvolvidas para executar a análise funcional dos metagenomas de forma mais simplificada, promovendo a interação de algoritmos de identificação de genes, alinhamento de sequencias e bancos de dados. Muitas delas também processam outras etapas como filtragem dos dados e comparação de metagenomas e fornecem visualização acessível dos resultados. Esses programas estão descritos abaixo (THOMAS; GILBERT; MEYER, 2012; BELLA et al., 2013; OULAS et al., 2015):

– **MG-RAST:** ferramenta que possui banco de dados próprio e executa controle de qualidade, predição de genes, anotação funcional e ambiente de comparação de metagenomas, retornando dados ao usuário em forma de perfil de abundância e informações taxonômicas.

  O *software* MG-RAST primeiramente prediz os genes dos metagenomas, executa o alinhamento com a ferramenta BLAT (BLAST-*like alignament tool*) (KENT, 2002) e então identifica os genes dos bancos de dados com melhores homologia (acima de 70% de identidade) em comparação aos *reads*. A partir desse ponto da análise, são utilizados os genes homólogos identificados pelo alinhamento, e não mais os genes encontrados nos metagenomas.

Embora a utilização dos genes homólogos e não dos originais ocasione uma série de limitações, fornece maior rapidez ao método, pois os dados dos homólogos já foram pré-processados. Desse modo, a única etapa computacionalmente intensa o alinhamento dos genes do metagenoma com os de bancos de dados, mas passada essa fase, o resto das comparações já foram pré-processadas (MEYER et al., 2008; GLASS et al., 2010).

– **IMG/MER 4:** *software* que executa controle de qualidade, predição de genes e anotação funcional. Utiliza perfis HMM para associar genes com o banco de dados PFAM, e então as funções são identificadas com COGs. Os bancos de dados com matrizes de corte posição-especificas (PSSMs, do inglês, *position-specific scoring matrix*) para COGs são obtidos do NCBI e são usados para a anotação de sequências de proteínas. Além disso, os genes são rotulados usando KEGG, números EC e sua filogenia é atribuída utilizando buscas de similaridade. IMG/MER pode utilizar seus próprios dados, uma vez que apresenta um grande repositório público de genomas.

  Inicia o processamento dos metagenomas com a predição de todos genes do metagenoma. Depois, utiliza os genes originais do metagenoma são submetidos à identificação de proteínas correspondentes no banco de dados PFAM.

  O banco de dados PFAM não é aceito pelo MG-RAST. PFAM fornece informações muito mais detalhadas do que o COGs, único banco de dados de proteínas utilizado pelo MG-RAST. Além disso, PFAM fornece uma análise com maior cobertura do que COGs, pois o número de metagenomas inseridos nos *clusters* do PFAM é maior do que o do COGs. Entretanto, a maior limitação do IMG/MER é o crescimento exponencial do número de genes, característica não vinculada ao MG-RAST, pois ele não mantém os metagenomas para análise (MARKOWITZ et al., 2013).

– **EBI Metagenomics service:** utiliza estrutura de metadata e formatos que obedecer os padrões do GSC (Consórcio de Padrões Genômicos, sigla em inglês, *Genomic Standards Consortium*). Além disso, está adotando um novo esquema de dados que atualmente sendo hospedado pelo EBI-EMBL: o ENA (Arquivo Europeu de Nucleotídeos, sigla em inglês, *European Nucleotide Archive*). O ENA tem o objetivo de integrar dados derivados das tecnologias de sequenciamento em um padrão mutualmente aceito.

  O EBI Metagenomics oferece um serviço de análise de dados genômicos obtidos por *shotgun* e também por genes marcadores. Isso permite a extração dos dados de rRNA dos metagenomas utilizando ferramentas como rRNASelector (LEE; YI; CHUN, 2011) para análise de metagenômica por genes marcadores. Também é compatível com ferramentas de análise 16S rRNA, como Qiime (citado

no item sobre “Metagenômica de 16S rRNA") para atribuição taxonômoica correta dessas sequências.

Para busca de sequências codificadoras nos metagenomas, EBI Metagenomics utiliza FragGeneScan para identificar as sequências codificadoras de proteínas. Para a anotação funcional, utiliza bancos de dados como Interpro, que é um sistema cumulativo e composto de múltiplos bancos de dados de famílias de proteínas e permite predição de domínios de proteínas e atribuição funcional aos genes.

EBI Metagenomics fornece arquivamento de dados via ENA e números de acesso únicos para cada conjunto de dados submetido. As políticas de arquivamento requerem que os dados sejam públicos, entretanto, há um período de 2 anos, a partir da submissão, durante o qual os dados são mantidos em modo privativo até que o usuário publique os resultados analisados (HUNTER et al., 2014).

– **CAMERA:** serviço de nuvem online que fornece ferramentas de *software* hospedadas e infraestrutura computacional de alta performance para a análise de dados de metagenômica. Permite a publicação do conjunto de dados e comparação entre os metagenomas.

  É uma ferramenta flexível, que permite que o usuário interfira no processo de análise. Entretanto, essa característica exige experiência e conhecimento do usuário para que a análise possa ser executada de maneira correta e os resultados possam ser corretamente interpretados (SESHADRI et al., 2007).

– **MEGAN:** ferramenta para visualização de resultados de análises taxonômicas ou funcionais derivados do BLAST. Apresenta diversas opções de visualização, como dendrogramas, gráficos de barras e outros tipos de gráficos que permitem que dados hierárquicos sejam explorados e torna a análise mais visualmente acessível (HUSON et al., 2007).

3 Conclusão e Perspectivas Futuras

Nas últimas duas décadas foram promovidos avanços importantes nas sub-áreas metagenômica. A disponibilidade de métodos de extração de DNA de quase todo tipo de amostra ambiental, diminuição brusca nos preços do sequenciamento, evolução das tecnologias NGS e progressos no campo da computação como poder de processamento e armazenamento e desenvolvimento de algoritmos mais complexos possibilitaram a obtenção de análises de comunidades microbióticas cada vez mais complexas e completas (KUMAR et al., 2015).

Os progressos no campo da metagenômica ainda apresentam potencial de irem cada vez mais longe: as tecnologias de sequenciamento em desenvolvimento, como PacBio ou sequenciamento por nanoporo prometem maiores facilidades para os protocolos de análise desde a montagem até o processo de anotação funcional. Além das plataformas, as próprias ferramentas computacionais vêm se aprimorando no decorrer do tempo: a crescente quantidade de genomas de referência de micro-organismos cultiváveis e não cultiváveis fornece um aumento progressivo na quantidade de informações disponíveis, consequentemente aumentando a precisão dos algoritmos de análises taxonômica e funcional (OULAS et al., 2015).

A falta de padronização dos dados é outra dificuldade que está sendo superada: o GSC vem desenvolvendo protocolos como o MARMS (em inglês, Mininum Analysis Requirements of Metagenome Sequences, ou "requerimentos mínimos para a análise de sequências metagenômicas"em português). MARMS será composta de metologogias padronizadas e consensos na escolha de *softwares*, etapas de análise, valores de limite e parâmetros. Esse projeto tem o objetivo de minimizar os viéses que podem ser gerados pela análise de múltiplas metodologias. GSC também pretende promover a integração de dados analisados atribuídos a formatos e estruturas mutualmente aceitáveis que facilitarão a troca de *insights* e informação valiosos no campo da microbiologia ambiental (OULAS et al., 2015).

Entretanto, o desenvolvimento de ferramentas e bancos de dados para estudos metagenômicos ainda está em seu princípio, e ainda há muitas limitações para ser contornadas (KIM et al., 2013). Primeiramente, a precisão das ferramentas necessita ser melhorada. Segundo, é necessária uma melhora na infraestrutura computacional para gerenciamento, disponibilidade e armazenamento de dados, pois o rápido aumento no tamanho e quantidade de dados está superando a capacidade computacional. Terceiro, é necessário o aperfeiçoamento de métodos estatísticos, especialmente para metagenomas originados e comunidades complexas onde os dados de táxons e genes podem ser esparços. Por último, é necessário o aprimoramento de sistemas experimentais para a manipulação de comunidades microbianas; modificando a composição dos meios de cultura (e.g., administração de antibióticos, suplementação probiótica, transplante de comunidades, mudanças físicas como de pH, temperatura, pressão), é possível estabelecer a relação dos micro-organismos com a composição do ambiente (SHARPTON, 2014). Com a superação dessas dificuldades, as comunidades microbianas poderão ser estudadas com cada vez mais rapidez e precisão, proporcionando o aumento do conhecimento sobre a dinâmica das comunidades microbianas e impulsionando o avanço da ciência em diversas áreas.

Referências

AFIAHAYATI, A.; MULYANA, S. Multiple sequence alignment menggunakan hidden markov model. In: *Seminar Nasional Informatika (SEMNASIF)*. [S.l.: s.n.], 2015. v. 1, n. 1. 13

ALTSCHUL, S. F. et al. Basic local alignment search tool. *Journal of molecular biology*, Elsevier, v. 215, n. 3, p. 403–410, 1990. 8

BABRAHAM BIOINFORMATICS. *FASTQC, A quality control tool for high throughput sequence data.* 2016. Data de acesso: 10 jun. 2016. Disponível em: <http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc/>. 11

BATEMAN, A. et al. The pfam protein families database. *Nucleic acids research*, Oxford Univ Press, v. 32, n. suppl 1, p. D138–D141, 2004. 18

BELLA, J. M. D. et al. High throughput sequencing methods and analysis for microbiome research. *Journal of microbiological methods*, Elsevier, v. 95, n. 3, p. 401–414, 2013. 1, 18

BERGER, S. A.; KROMPASS, D.; STAMATAKIS, A. Performance, accuracy, and web server for evolutionary placement of short sequence reads under maximum likelihood. *Systematic biology*, Oxford University Press, p. syr010, 2011. 9

BIK, H. M. Deciphering diversity and ecological function from marine metagenomes. *The Biological Bulletin*, MBL, v. 227, n. 2, p. 107–116, 2014. 2

BLANCA, J. M. et al. ngs_backbone: a pipeline for read cleaning, mapping and snp calling using next generation sequence. *BMC genomics*, BioMed Central, v. 12, n. 1, p. 1, 2011. 11

BRADY, A.; SALZBERG, S. Phymmbl expanded: confidence scores, custom databases, parallelization and more. *Nature methods*, Nature Publishing Group, v. 8, n. 5, p. 367–367, 2011. 16

BRADY, A.; SALZBERG, S. L. Phymm and phymmbl: metagenomic phylogenetic classification with interpolated markov models. *Nature methods*, Nature Publishing Group, v. 6, n. 9, p. 673–676, 2009. 16

BRAGG, L.; TYSON, G. W. Metagenomics using next-generation sequencing. *Environmental Microbiology: Methods and Protocols*, Springer, p. 183–201, 2014. 11, 15

CAI, Y.; SUN, Y. Esprit-tree: hierarchical clustering analysis of millions of 16s rrna pyrosequences in quasilinear computational time. *Nucleic acids research*, Oxford Univ Press, p. gkr349, 2011. 10

CAMPBELL, S. A.; RUAN, S.; WEI, J. Qualitative analysis of a neural network model with multiple time delays. *International Journal of Bifurcation and Chaos*, World Scientific, v. 9, n. 08, p. 1585–1595, 1999. 15

CAPORASO, J. G. et al. Qiime allows analysis of high-throughput community sequencing data. *Nature methods*, Nature Publishing Group, v. 7, n. 5, p. 335–336, 2010. 7, 9

CHEVREUX, B. et al. Using the miraest assembler for reliable and automated mrna transcript assembly and snp detection in sequenced ests. *Genome research*, Cold Spring Harbor Lab, v. 14, n. 6, p. 1147–1159, 2004. 12

CLEMENTE, J. C.; JANSSON, J.; VALIENTE, G. Flexible taxonomic assignment of ambiguous sequencing reads. *BMC bioinformatics*, BioMed Central, v. 12, n. 1, p. 1, 2011. 8, 9

COLE, J. R. et al. The ribosomal database project: improved alignments and new tools for rrna analysis. *Nucleic acids research*, Oxford Univ Press, v. 37, n. suppl 1, p. D141–D145, 2009. 8

COMPEAU, P. E.; PEVZNER, P. A.; TESLER, G. How to apply de bruijn graphs to genome assembly. *Nature biotechnology*, Nature Publishing Group, v. 29, n. 11, p. 987–991, 2011. 12

COUNCIL, N. R. *The New Science of Metagenomics: Revealing the Secrets of Our Microbial Planet*. The National Academies Press, 2007. ISBN 978-0-309-10676-4. Disponível em: <http://www.nap.edu/catalog/11902/the-new-science-of-metagenomics-revealing-the-secrets-of-our>. 1

DESANTIS, T. Z. et al. Greengenes, a chimera-checked 16s rrna gene database and workbench compatible with arb. *Applied and environmental microbiology*, Am Soc Microbiol, v. 72, n. 7, p. 5069–5072, 2006. 8, 9

DEVARAJ, S.; HEMARAJATA, P.; VERSALOVIC, J. The human gut microbiome and body metabolism: implications for obesity and diabetes. *Clinical chemistry*, Am Assoc Clin Chem, v. 59, n. 4, p. 617–628, 2013. 1

DIAZ, N. N. et al. Tacoa–taxonomic classification of environmental genomic fragments using a kernelized nearest neighbor approach. *BMC bioinformatics*, BioMed Central Ltd, v. 10, n. 1, p. 56, 2009. 16

EDGAR, R. C. Search and clustering orders of magnitude faster than blast. *Bioinformatics*, Oxford Univ Press, v. 26, n. 19, p. 2460–2461, 2010. 7, 10

ESPOSITO, A.; KIRSCHBERG, M. How many 16s-based studies should be included in a metagenomic conference? it may be a matter of etymology. *FEMS Microbiology Letters*, v. 351, p. 145–146, 2014. 5

FASTX-TOOLKIT, A short-reads pre-processing tools. 2016. Data de acesso: 10 jun. 2016. Disponível em: <http://hannonlab.cshl.edu/fastx_toolkit/index.html>. 11

FORDE, B. M.; O'TOOLE, P. W. Next-generation sequencing technologies and their impact on microbial genomics. *Briefings in functional genomics*, Oxford University Press, v. 12, n. 5, p. 440–453, 2013. 10

FU, L. et al. Cd-hit: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, Oxford Univ Press, v. 28, n. 23, p. 3150–3152, 2012. 10

GASPAR, J. M.; THOMAS, W. K. Assessing the consequences of denoising marker-based metagenomic data. *PLoS One*, Public Library of Science, v. 8, n. 3, p. e60458, 2013. 7

GERLACH, W.; STOYE, J. Taxonomic classification of metagenomic shotgun sequences with carma3. *Nucleic acids research*, Oxford Univ Press, p. gkr225, 2011. 15

GLASS, E. M. et al. Using the metagenomics rast server (mg-rast) for analyzing shotgun metagenomes. *Cold Spring Harbor Protocols*, Cold Spring Harbor Laboratory Press, v. 2010, n. 1, p. pdb–prot5368, 2010. 19

HAAS, B. J. et al. Chimeric 16s rrna sequence formation and detection in sanger and 454-pyrosequenced pcr amplicons. *Genome research*, Cold Spring Harbor Lab, v. 21, n. 3, p. 494–504, 2011. 7

HANDELSMAN, J. Metagenomics: application of genomics to uncultured microorganisms. *Microbiology and molecular biology reviews*, Am Soc Microbiol, v. 68, n. 4, p. 669–685, 2004. 2

HANDELSMAN, J. et al. Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chemistry & biology*, Elsevier, v. 5, n. 10, p. R245–R249, 1998. 2

HAYES, W. S.; BORODOVSKY, M. How to interpret an anonymous bacterial genome: machine learning approach to gene identification. *Genome research*, Cold Spring Harbor Lab, v. 8, n. 11, p. 1154–1171, 1998. 17

HOFF, K. J. et al. Orphelia: predicting genes in metagenomic sequencing reads. *Nucleic acids research*, Oxford Univ Press, v. 37, n. suppl 2, p. W101–W105, 2009. 17

HUNTER, S. et al. Ebi metagenomics—a new resource for the analysis and archiving of metagenomic data. *Nucleic acids research*, Oxford Univ Press, v. 42, n. D1, p. D600–D606, 2014. 20

HUSON, D. H. et al. Megan analysis of metagenomic data. *Genome research*, Cold Spring Harbor Lab, v. 17, n. 3, p. 377–386, 2007. 8, 14, 20

HYATT, D. et al. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC bioinformatics*, BioMed Central, v. 11, n. 1, p. 1, 2010. 17

KANEHISA, M.; GOTO, S. Kegg: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, Oxford Univ Press, v. 28, n. 1, p. 27–30, 2000. 18

KENT, W. J. Blat—the blast-like alignment tool. *Genome research*, Cold Spring Harbor Lab, v. 12, n. 4, p. 656–664, 2002. 18

KIM, M. et al. Analytical tools and databases for metagenomics in the next-generation sequencing era. *Genomics & informatics*, v. 11, n. 3, p. 102–113, 2013. 3, 5, 6, 7, 8, 9, 12, 14, 15, 17, 21

KIM, O.-S. et al. Introducing eztaxon-e: a prokaryotic 16s rrna gene sequence database with phylotypes that represent uncultured species. *International journal of systematic and evolutionary microbiology*, Microbiology Society, v. 62, n. 3, p. 716–721, 2012. 8

KUMAR, S. et al. Metagenomics: Retrospect and prospects in high throughput age. *Biotechnology research international*, Hindawi Publishing Corporation, v. 2015, 2015. 3, 12, 20

KUNIN, V. et al. A bioinformatician's guide to metagenomics. *Microbiology and molecular biology reviews*, Am Soc Microbiol, v. 72, n. 4, p. 557–578, 2008. 6

LAND, M. et al. Insights from 20 years of bacterial genome sequencing. *Functional & integrative genomics*, Springer, v. 15, n. 2, p. 141–161, 2015. 3

LEE, H. et al. Third-generation sequencing and the future of genomics. *bioRxiv*, Cold Spring Harbor Labs Journals, p. 048603, 2016. 3

LEE, J.-H.; YI, H.; CHUN, J. rrnaselector: a computer program for selecting ribosomal rna encoding sequences from metagenomic and metatranscriptomic shotgun libraries. *The Journal of Microbiology*, Springer, v. 49, n. 4, p. 689–691, 2011. 19

LI, R. et al. Soap: short oligonucleotide alignment program. *Bioinformatics*, Oxford Univ Press, v. 24, n. 5, p. 713–714, 2008. 13

LIU, Y. et al. Gene prediction in metagenomic fragments based on the svm algorithm. *BMC bioinformatics*, BioMed Central Ltd, v. 14, n. Suppl 5, p. S12, 2013. 14

MACDONALD, N. J.; PARKS, D. H.; BEIKO, R. G. Rapid identification of high-confidence taxonomic assignments for metagenomic data. *Nucleic acids research*, Oxford Univ Press, p. gks335, 2012. 16

MANDE, S. S.; MOHAMMED, M. H.; GHOSH, T. S. Classification of metagenomic sequences: methods and challenges. *Briefings in bioinformatics*, Oxford Univ Press, p. bbs054, 2012. 13, 15

MARCHLER-BAUER, A. et al. Cdd: a conserved domain database for protein classification. *Nucleic acids research*, Oxford Univ Press, v. 33, n. suppl 1, p. D192–D196, 2005. 18

MARCO, D. *Metagenomics: Current innovations and future trends*. [S.l.]: Horizon Scientific Press, 2011. 2

MARKOWITZ, V. M. et al. Img 4 version of the integrated microbial genomes comparative analysis system. *Nucleic acids research*, Oxford Univ Press, p. gkt963, 2013. 19

MATSEN, F. A.; KODNER, R. B.; ARMBRUST, E. V. pplacer: linear time maximum-likelihood and bayesian phylogenetic placement of sequences onto a fixed reference tree. *BMC bioinformatics*, BioMed Central, v. 11, n. 1, p. 1, 2010. 9

MCHARDY, A. C. et al. Accurate phylogenetic classification of variable-length dna fragments. *Nature methods*, Nature Publishing Group, v. 4, n. 1, p. 63–72, 2007. 15

MEYER, F. et al. The metagenomics rast server–a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC bioinformatics*, BioMed Central Ltd, v. 9, n. 1, p. 386, 2008. 14, 19

MICROWINE, A MARIE CURIE INITIAL TRAINING NETWORK. *MicroWine, A Marie Curie Initial Training Network*. 2016. Data de acesso: 10 jun. 2016. Disponível em: <http://www.microwine.eu/>. 2

MIRARAB, S.; NGUYEN, N.; WARNOW, T. Sepp: Saté-enabled phylogenetic placement. In: CITESEER. *Pac Symp Biocomput*. [S.l.], 2012. v. 17, p. 247–258. 9

MOHAMMED, M. H. et al. Sphinx—an algorithm for taxonomic binning of metagenomic sequences. *Bioinformatics*, Oxford Univ Press, v. 27, n. 1, p. 22–30, 2011. 16

MORGAVI, D. P. et al. Rumen microbial (meta) genomics and its application to ruminant production. *Animal*, Cambridge Univ Press, v. 7, n. s1, p. 184–201, 2013. 1

MOROZOVA, O.; MARRA, M. A. Applications of next-generation sequencing technologies in functional genomics. *Genomics*, Elsevier, v. 92, n. 5, p. 255–264, 2008. 3

MULLER, J. et al. eggnog v2. 0: extending the evolutionary genealogy of genes with enhanced non-supervised orthologous groups, species and functional annotations. *Nucleic acids research*, Oxford Univ Press, v. 38, n. suppl 1, p. D190–D195, 2010. 18

NAMIKI, T. et al. Metavelvet: an extension of velvet assembler to de novo metagenome assembly from short sequence reads. *Nucleic acids research*, Oxford Univ Press, v. 40, n. 20, p. e155–e155, 2012. 13

NAWROCKI, E. P.; KOLBE, D. L.; EDDY, S. R. Infernal 1.0: inference of rna alignments. *Bioinformatics*, Oxford Univ Press, v. 25, n. 10, p. 1335–1337, 2009. 9

NGUYEN, M. N. et al. Di-codon usage for gene classification. In: *Pattern Recognition in Bioinformatics*. [S.l.]: Springer, 2009. p. 211–221. 17

NIKOLAKI, S.; TSIAMIS, G. Microbial diversity in the era of omic technologies. *BioMed research international*, Hindawi Publishing Corporation, v. 2013, 2013. 5, 6

NOGUCHI, H.; TANIGUCHI, T.; ITOH, T. Metageneannotator: detecting species-specific patterns of ribosomal binding site for precise gene prediction in anonymous prokaryotic and phage genomes. *DNA research*, Kazusa DNA Resh Ins, v. 15, n. 6, p. 387–396, 2008. 17

OSSOWSKI, S. et al. Sequencing of natural strains of arabidopsis thaliana with short reads. *Genome research*, Cold Spring Harbor Lab, v. 18, n. 12, p. 2024–2033, 2008. 11

OULAS, A. et al. Metagenomics: tools and insights for analyzing next-generation sequencing data derived from biodiversity studies. *Bioinformatics and biology insights*, Libertas Academica, v. 9, p. 75, 2015. 3, 4, 6, 9, 10, 17, 18, 21

OVERBEEK, R. et al. The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic acids research*, Oxford Univ Press, v. 33, n. 17, p. 5691–5702, 2005. 18

PATIL, K. R. et al. Taxonomic metagenome sequence assignment with structured output models. *Nature methods*, Nature Publishing Group, v. 8, n. 3, p. 191–192, 2011. 15

PATIN, N. V. et al. Effects of otu clustering and pcr artifacts on microbial diversity estimates. *Microbial ecology*, Springer, v. 65, n. 3, p. 709–719, 2013. 9

PENG, Y. et al. Meta-idba: a de novo assembler for metagenomic data. *Bioinformatics*, Oxford Univ Press, v. 27, n. 13, p. i94–i101, 2011. 13

PETTERSSON, E.; LUNDEBERG, J.; AHMADIAN, A. Generations of sequencing technologies. *Genomics*, v. 93, n. 2, p. 105–111, Feb 2009. 2

PEVZNER, P. A.; TANG, H.; WATERMAN, M. S. An eulerian path approach to dna fragment assembly. *Proceedings of the National Academy of Sciences*, National Acad Sciences, v. 98, n. 17, p. 9748–9753, 2001. 13

PORETSKY, R. et al. Strengths and limitations of 16s rrna gene amplicon sequencing in revealing temporal microbial community dynamics. *PloS one*, Public Library of Science, v. 9, n. 4, p. e93827, 2014. 6

PRUESSE, E.; PEPLIES, J.; GLÖCKNER, F. O. Sina: accurate high-throughput multiple sequence alignment of ribosomal rna genes. *Bioinformatics*, Oxford Univ Press, v. 28, n. 14, p. 1823–1829, 2012. 9

QUAST, C. et al. The silva ribosomal rna gene database project: improved data processing and web-based tools. *Nucleic acids research*, Oxford Univ Press, v. 41, n. D1, p. D590–D596, 2013. 8

QUINCE, C. et al. Accurate determination of microbial diversity from 454 pyrosequencing data. *Nature methods*, v. 6, n. 9, p. 639, 2009. 7

QUINCE, C. et al. Removing noise from pyrosequenced amplicons. *BMC bioinformatics*, BioMed Central, v. 12, n. 1, p. 1, 2011. 7

QUINLAN, A. R. et al. Pyrobayes: an improved base caller for snp discovery in pyrosequences. *Nature methods*, Nature Publishing Group, v. 5, n. 2, p. 179–181, 2008. 11

RHO, M.; TANG, H.; YE, Y. Fraggenescan: predicting genes in short and error-prone reads. *Nucleic acids research*, Oxford Univ Press, v. 38, n. 20, p. e191–e191, 2010. 17

RIDAURA, V. K. et al. Gut microbiota from twins discordant for obesity modulate metabolism in mice. *Science*, American Association for the Advancement of Science, v. 341, n. 6150, p. 1241214, 2013. 14

RIESENFELD, C. S.; SCHLOSS, P. D.; HANDELSMAN, J. Metagenomics: genomic analysis of microbial communities. *Annu. Rev. Genet.*, Annual Reviews, v. 38, p. 525–552, 2004. 5

ROSEN, G. L.; REICHENBERGER, E. R.; ROSENFELD, A. M. Nbc: the naive bayes classification tool webserver for taxonomic classification of metagenomic reads. *Bioinformatics*, Oxford Univ Press, v. 27, n. 1, p. 127–129, 2011. 16

ROSEN, M. J. et al. Denoising pcr-amplified metagenome data. *BMC bioinformatics*, BioMed Central Ltd, v. 13, n. 1, p. 283, 2012. 7

SALZBERG, S. L. et al. Microbial gene identification using interpolated markov models. *Nucleic acids research*, Oxford Univ Press, v. 26, n. 2, p. 544–548, 1998. 17

SANSCHAGRIN, S.; YERGEAU, E. Next-generation sequencing of 16s ribosomal rna gene amplicons. *JoVE (Journal of Visualized Experiments)*, n. 90, p. e51709–e51709, 2014. 6

SCHMIDT, T. M.; DELONG, E.; PACE, N. Analysis of a marine picoplankton community by 16s rrna gene cloning and sequencing. *Journal of bacteriology*, Am Soc Microbiol, v. 173, n. 14, p. 4371–4378, 1991. 5

SELENGUT, J. D. et al. Tigrfams and genome properties: tools for the assignment of molecular function and biological process in prokaryotic genomes. *Nucleic acids research*, Oxford Univ Press, v. 35, n. suppl 1, p. D260–D264, 2007. 18

SESHADRI, R. et al. Camera: a community resource for metagenomics. *PLoS biology*, v. 5, n. 3, 2007. 20

SHARPTON, T. J. An introduction to the analysis of shotgun metagenomic data. Frontiers Research Foundation, 2014. 10, 11, 12, 13, 14, 15, 16, 17, 21

SIMPSON, J. T. et al. Abyss: a parallel assembler for short read sequence data. *Genome research*, Cold Spring Harbor Lab, v. 19, n. 6, p. 1117–1123, 2009. 13

SUN, Y. et al. A large-scale benchmark study of existing algorithms for taxonomy-independent microbial community analysis. *Briefings in bioinformatics*, Oxford Univ Press, p. bbr009, 2011. 5, 8

TANENBAUM, D. M. et al. The jcvi standard operating procedure for annotating prokaryotic metagenomic shotgun sequencing data. *Standards in genomic sciences*, Springer, v. 2, n. 2, p. 229–237, 2010. 17

TATUSOV, R. L. et al. The cog database: an updated version includes eukaryotes. *BMC bioinformatics*, BioMed Central, v. 4, n. 1, p. 1, 2003. 18

THOMAS, T.; GILBERT, J.; MEYER, F. Metagenomics-a guide from sampling to data analysis. *Microb Inform Exp*, v. 2, n. 3, 2012. 2, 3, 15, 18

TREANGEN, T. J. et al. Metamos: a modular and open source metagenomic assembly and analysis pipeline. *Genome Biol*, v. 14, n. 1, p. R2, 2013. 12

TRINGE, S. G.; HUGENHOLTZ, P. A renaissance for the pioneering 16s rrna gene. *Current opinion in microbiology*, Elsevier, v. 11, n. 5, p. 442–446, 2008. 6

WANG, Q. et al. Naive bayesian classifier for rapid assignment of rrna sequences into the new bacterial taxonomy. *Applied and environmental microbiology*, Am Soc Microbiol, v. 73, n. 16, p. 5261–5267, 2007. 9

WOOLEY, J. C.; GODZIK, A.; FRIEDBERG, I. A primer on metagenomics. *PLoS Comput Biol*, v. 6, n. 2, p. e1000667, Feb 2010. Disponível em: <http://dx.doi.org/10.1371/journal.pcbi.1000667>. 1, 12

WRIGHT, E. S.; YILMAZ, L. S.; NOGUERA, D. R. Decipher, a search-based approach to chimera identification for 16s rrna sequences. *Applied and environmental microbiology*, Am Soc Microbiol, v. 78, n. 3, p. 717–725, 2012. 7

WU, M.; SCOTT, A. J. Phylogenomic analysis of bacterial and archaeal sequences with amphora2. *Bioinformatics*, Oxford Univ Press, v. 28, n. 7, p. 1033–1034, 2012. 9

YADA, T. et al. Modeling and predicting transcriptional units of escherichia coligenes using hidden markov models. *Bioinformatics*, Oxford Univ Press, v. 15, n. 12, p. 987–993, 1999. 17

ZARRAONAINDIA, I. et al. The soil microbiome influences grapevine-associated microbiota. *mBio*, American Society for Microbiology, v. 6, n. 2, p. e02527–14, mar 2015. 1, 13

ZERBINO, D. R.; BIRNEY, E. Velvet: algorithms for de novo short read assembly using de bruijn graphs. *Genome research*, Cold Spring Harbor Lab, v. 18, n. 5, p. 821–829, 2008. 13

ZHOU, J. et al. High-throughput metagenomic technologies for complex microbial community analysis: open and closed formats. *MBio*, Am Soc Microbiol, v. 6, n. 1, p. e02288–14, 2015. 6

ZHU, W.; LOMSADZE, A.; BORODOVSKY, M. Ab initio gene identification in metagenomic sequences. *Nucleic acids research*, Oxford Univ Press, v. 38, n. 12, p. e132–e132, 2010. 17