FUNDAÇÃO OSWALDO CRUZ
INSTITUTO GONÇALO MONIZ

Curso de Pós-Graduação em Biotecnologia em Saúde e Medicina Investigativa

DISSERTAÇÃO DE MESTRADO

ANÁLISE MULTI-ÔMICA DAS VIAS ASSOCIADAS À TUBERCULOSE INFANTIL

EDUARDO FUKUTANI ROCHA

Salvador – BA
2022

FUNDAÇÃO OSWALDO CRUZ
INSTITUTO GONÇALO MONIZ

Curso de Pós-Graduação em Biotecnologia em Saúde e Medicina Investigativa

ANÁLISE MULTI-ÔMICA DAS VIAS ASSOCIADAS À TUBERCULOSE INFANTIL

EDUARDO FUKUTANI ROCHA

Dissertação apresentada ao Curso de Pós-Graduação em Biotecnologia em Saúde e Medicina Investigativa para a obtenção do grau de Mestre.

Orientador: Prof. Dr. Artur Trancoso Lopo de Queiroz

Salvador – BA
2022

Ficha Catalográfiaca elaborada pela Biblioteca do
Instituto Gonçalo Moniz/ FIOCRUZ – Bahia - Salvador

R672a Rocha, Eduardo Fukutani

Análise multi-ômica das vias associadas à tuberculose infantil/ Eduardo Fukutani Rocha. _ Salvador, 2022.

44 f.: il.: 30 cm

Orientador: Prof. Dr. Artur Trancoso Lopo de Queiroz

Dissertação (Mestrado em Biotecnologia em Saúde e Medicina Investigativa) – Instituto Gonçalo Moniz, Fundação Oswaldo Cruz, Salvador, 2022.

1. Bioinformática. 2 Tuberculose infantil. 3. Análise de dados. 4. Dados transcriptômicos. 5. Biomarcador. I. Título.

CDU 616-002.5

“ANÁLISE MULTI-ÔMICA DAS VIAS ASSOCIADAS À TUBERCULOSE INFANTIL”

EDUARDO FUKUTANI ROCHA

FOLHA DE APROVAÇÃO

Salvador, 27 de junho de 2022.

COMISSÃO EXAMINADORA

Dra. Marta Giovanetti
Pesquisadora
IOC FIOCRUZ

Dra. Theolis Costa Barbosa Bessa
Pesquisadora
IGM/FIOCRUZ

Dr. Pablo Iv~~an Pereir~~a Ramos
Pesquisador
IGM/FIOCRUZ

FONTES DE FINANCIAMENTO

"O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001."

"À CAPES pelo fomento, apoio financeiro e consolidação do programa de pós-graduação em Biotecnologia em Saúde e Medicina Investigativa."

ROCHA, Eduardo Fukutani. **Análise multi-ômica das vias associadas à Tuberculose infantil**. 44 f. Dissertação (Mestrado em Biotecnologia em Saúde e Medicina Investigativa) – Instituto Gonçalo Moniz, Fundação Oswaldo Cruz, Salvador, 2022.

RESUMO

INTRODUÇÃO: A tuberculose é uma das principais causas de mortalidade infantil relacionada a doenças infecciosas no mundo. Se trata de uma doença infecciosa que afeta principalmente os pulmões, causada pelo bacilo *Mycobacterium Tuberculosis*. Os métodos atuais para diagnosticar a TB infantil apresentam baixa performance. Assinaturas transcricionais e metabólicas para o diagnóstico da TB infantil são interessantes e promissoras, porém ainda precisam de mais validações para que o seu como testes de rotina seja possível. Assinaturas identificadas com integração de dados transcriptômicos e metabolômicos são uma alternativa promissora para o diagnóstico da TB infantil. **OBJETIVO**: O objetivo deste trabalho é identificar uma assinatura em dados transcriptômicos e metabolômicos integrados para classificar crianças portadoras de TB. **MATERIAIS e MÉTODOS**: Amostras de crianças de até 15 anos foram coletadas da cidade de Pune, do estado de Maharashtra, Índia. As amostras tiveram os seus metabólitos medidos por CL/EMAD e seus transcritos mensurados por NGS. No total, 40 amostras com dados transcriptômicos e metabolômicos foram obtidas, destas 16 são amostras do grupo caso (TB) e 24 são amostras controle. Os genes diferencialmente expressos entre os grupos caso e controle foram identificados nos dados transcriptômicos. O algoritmo de floresta aleatória foi utilizado para identificar os melhores genes classificadores entre os *DEGs*. A performance dos classificadores identificados foi avaliada por curvas *ROC*. Uma validação *in silico* da performance dos genes selecionados foi feita em outros conjuntos de dados. A integração dos dados transcriptômicos e metabolômicos foi realizada, para isso foi feita uma análise de correlação entre os genes selecionados e os valores de abundância dos metabólitos. Por fim, uma análise de enriquecimento de vias foi aplicada aos metabólitos correlacionados com os genes. As vias metabólicas enriquecidas foram estudadas e relacionadas à patologia da TB infantil. **RESULTADOS**: Os genes diferencialmente expressos foram analisados utilizando o pacote *DESeq2*, identificando 174 *DEGs*. O algoritmo de floresta aleatória foi aplicado nos dados de expressão dos *DEGs*, indicando 5 genes: *BPI*, *AZU1*, *C1QC*, *AC092580.4* e *MPO*. Curvas ROC foram utilizadas para mensurar a performance dos DEGs e dos 5 genes, apresentando AUCs de 0,86 e 0,91 respectivamente. Um total de 27 metabólitos foram correlacionados aos 5 genes no grupo caso, enquanto 33 metabólitos foram correlacionados no grupo controle. As vias metabólicas enriquecidas condizem com o quadro observado na TB infantil. A performance dos genes selecionados foi validada em outros conjuntos de dados: *GSE39939*, *GSE39940* e *GSE41055*. Para isto, curvas ROC mensuraram a performance dos genes na classificação de amostras dos outros conjuntos de dados. As principais AUCs nos outros conjuntos foram de 0,80 no *GSE39939*, 0,85 no *GSE39940* e 0,70 no *GSE41055*. **CONCLUSÃO**: O conjunto de genes classificadores proposto é uma alternativa que demonstrou bastante consistência na sua performance, inclusive em outros conjuntos de dados. Este conjunto classificador é um passo à frente para a identificação de um conjunto de genes como potenciais biomarcadores para a TB infantil.

Palavras-chave: Bioinformática. Tuberculose infantil. Análise de dados. Dados transcriptômicos. Dados metabolômicos. Biomarcador.

ROCHA, Eduardo Fukutani. **Multi-omic analysis of pathways associated with childhood tuberculosis**. 44f. Dissertação (Mestrado em Biotecnologia em Saúde e Medicina Investigativa) – Instituto Gonçalo Moniz, Fundação Oswaldo Cruz, Salvador, 2022.

ABSTRACT

INTRODUCTION: Tuberculosis is one of the main causes of infectious disease related infant mortality worldwide. It is an infectious disease that mainly affects the lungs, caused by the *Mycobacterium Tuberculosis* bacillus. Current methods for diagnosing childhood TB exhibit poor performance. Transcriptional and metabolic signatures for the diagnosis of childhood TB are interesting and promising, but they still need further validation to be used as routine tests. Signatures identified with integration of transcriptomic and metabolomic data are a promising alternative for the diagnosis of childhood TB. **OBJECTIVE**: The objective of this work is to identify a signature in integrated transcriptomic and metabolomic data to classify children with TB. **MATERIALS AND METHODS**: Samples from children with age up to 15 years were collected from the city of Pune, state of Maharashtra, India. The samples had their metabolites measured by *CL/EMAD* and their transcripts measured by *NGS*, a total of 54 samples with transcriptomic and metabolomic data have been retrieved, from which 16 samples comprises the case group (TB) and 24 samples comprises the control group. The differentially expressed genes between the case and control groups were identified in the transcriptomic data. The random forest algorithm was used to identify the best classifier genes among the *DEGs*. The performance of the identified classifiers was evaluated by *ROC* curves. An *in silico* validation of the performance of selected genes was performed on other datasets. The integration of transcriptomic and metabolomic data was performed, for which a correlation analysis was performed between the selected genes and the metabolite abundance values. Finally, a pathway enrichment analysis was applied to the metabolites correlated with the genes. The enriched metabolic pathways were studied and related to the pathology of childhood TB. **RESULTS**: The differentially expressed genes were analyzed using the *DESeq2* package, identifying 174 DEGs. The random forest algorithm was applied to *DEGs* expression data, indicating 5 genes: BPI, AZU1, C1QC, AC092580.4 and MPO. *ROC* curves were used to measure the *DEGs* and the 5 genes' classifying performance, presenting AUCs of 0.86 and 0.91 respectively. A total of 27 metabolites were correlated to the 5 genes in the case group, while 33 metabolites were correlated in the control group. The enriched metabolic pathways are consistent with the childhood TB pathology. The performance of selected genes was validated in other datasets: *GSE39939*, *GSE39940* and *GSE41055*. For this, *ROC* curves measured the performance of genes in the classification of samples from the other data sets. The top *AUCs* in the other sets were 0.80 in *GSE39939*, 0.85 in *GSE39940* and 0.70 in *GSE41055*. **CONCLUSION**: The proposed set of classifier genes is an alternative that has shown a lot of consistency in its performance, including in other datasets. This classifier set is a step forward towards the identification of a set of genes as potential biomarkers for childhood TB.

Keywords: Bioinformatics. Childhood tuberculosis. Data analysis. Transcriptomic data. Metabolomic data. Biomarker.

LISTA DE FIGURAS

SUMÁRIO

1 INTRODUÇÃO E JUSTIFICATIVA

A tuberculose (TB) é uma das principais causas de mortalidade infantil relacionada à doenças infecciosas no mundo (MARTINEZ et al., 2018). A mesma é causada pelo bacilo *Mycobacterium tuberculosis* e afeta aproximadamente 1 milhão de pacientes pediátricos por ano, o que leva à morte pelo menos 210 mil infectados (MURRAY et al., 2014). Ao infectar o indivíduo adulto, a bactéria pode permanecer latente em grande parte dos infectados (HANIFA et al., 2009), mas quando infecta crianças e bebês, tende a evoluir para formas mais severas de tuberculose, como a disseminada ou meningítica (CARVALHO et al., 2018; MARAIS et al., 2006). Apesar da vacinação pelo Bacilo de Calmette e Guérin (BCG) proteger crianças contra as formas mais severas de TB, ela não é capaz de impedir uma infecção primária ou reativação de uma infecção latente (GUTIÉRREZ-GONZÁLEZ et al., 2021).

O diagnóstico da tuberculose infantil é o principal desafio no controle da doença, uma vez que os testes diagnósticos tradicionais possuem sensibilidade reduzida nesta população. A coleta de amostras para a realização dos testes microbiológicos em crianças é dificultada, essa população tende a engolir as secreções do sistema respiratório por reflexo. Além disso, a coleta de escarro induzido ou aspirado gástrico/nasofaríngeo também costuma ser difícil nessa população, outro problema é que as amostras coletadas assim apresentam rendimento bacteriológico baixo, mesmo com o uso de cultura líquida ou *Xpert Ultra* (NICOL; ZAR, 2011; ZAR et al., 2019).

Além das dificuldades na coleta de amostras, os testes para o diagnóstico da TB em crianças também são menos eficazes, devido à natureza paucibacilar da doença nessa população. Testes microbiológicos padrão ouro, como a cultura micobacteriana e o exame baciloscópico do escarro, possuem baixa sensibilidade quando aplicados em crianças (THOMAS et al., 2014). Testes imunológicos, como o teste intradérmico da tuberculina (TST) e o ensaio de liberação do interferon gama (IGRA), não são capazes de distinguir a forma ativa e latente da doença, além de possuírem sensibilidade e especificidade reduzidas nessa população (DUNN; STARKE; REVELL, 2016). Por fim, o teste molecular *GeneXpert MTB/RIF*, o teste mais sensível para o diagnóstico da TB, também tem a sua sensibilidade reduzida ao ser aplicado em crianças (THOMAS, 2017).

As dificuldades na coleta de amostras e baixa sensibilidade no diagnóstico da TB em crianças faz com que aproximadamente 15~50% dos casos de TB infantil não sejam confirmados, mas sim presumidos de forma indireta através da avaliação física, dos exames de imagem e por estudo da epidemiologia da doença (. A ausência de um teste diagnóstico com

alta sensibilidade e com amostragem acessível demonstra a necessidade de se explorar alternativas modernas.

O desenvolvimento de testes diagnósticos baseados em assinaturas transcricionais ou metabólicas no sangue tem sido uma área muito explorada na última década. Com o desenvolvimento de novas tecnologias para quantificar transcritos e metabólitos no sangue e novas metodologias para análise de dados, diversas assinaturas transcricionais e metabólicas têm sido propostas para diagnosticar a TB em adultos (GOLETTI et al., 2018). Por outro lado, uma assinatura transcricional definitiva ainda não foi definida para crianças (TOGUN et al., 2018) e apenas uma assinatura metabólica foi proposta (SUN et al., 2016). Embora assinaturas transcricionais e metabólicas tenham sido propostas, estudos subsequentes para validar as suas eficiências como testes diagnósticos ainda são necessários (HOANG et al., 2021; WEINER et al., 2018).

Uma estratégia a ser explorada é a identificação de uma assinatura através da integração de dados transcriptômicos e metabolômicos. Nesta metodologia, os dados metabolômicos fornecem mensurações mais próximas do fenótipo do indivíduo, complementando as informações contidas nos dados transcriptômicos (CAVILL et al., 2016). Almejando propor uma assinatura transcricional que também tenha base em dados metabolômicos, nós analisamos os dados transcriptômicos e metabolômicos de crianças com idade inferior a 15 anos provenientes da cidade de Pune, Índia.

2 **OBJETIVOS**

2.1 OBJETIVO GERAL

Identificar um conjunto de alvos para a classificação de pacientes pediátricos portadores de tuberculose através da análise de dados transcriptômicos e metabolômicos integrados.

2.2 OBJETIVOS ESPECÍFICOS

- Identificar uma assinatura transcriptômica capaz de identificar crianças com tuberculose ativa nos dados transcriptômicos;
- Identificar os principais metabólitos correlacionados à assinatura transcriptômica nos dados metabolômicos;
- Identificar as vias metabólicas associadas aos metabólitos correlacionados e aos genes que compõem a assinatura transcriptômica;
- Validar os melhores alvos identificados em conjuntos de dados independentes.

3 **REVISÃO DE LITERATURA**

3.1 TUBERCULOSE INFANTIL

A tuberculose (TB) é uma doença bacteriana infecciosa causada pela bactéria gram-negativa *Mycobacterium tuberculosis* (MTB), que afeta principalmente os pulmões (PERLMAN et al., 1997). Atualmente, a TB infantil é uma das principais causas de mortalidade infantil no mundo, afetando aproximadamente 1 milhão de crianças e levando 210 mil à morte anualmente (MURRAY et al., 2014). Embora a vacinação pela BCG possa conferir proteção contra as formas mais severas de TB, a mesma não é capaz de evitar uma infecção primária ou uma possível reativação da doença em casos de infecções latentes (GUTIÉRREZ-GONZÁLEZ et al., 2021). A transmissão da TB ocorre através do contato com perdigotos de pessoas infectadas pela tuberculose pulmonar ativa, forma clínica e aflige cerca de 10% das pessoas infectadas pela MTB (HOLMES et al., 2018). Estima-se que um terço da população mundial está infectada pela MTB de forma latente, o que torna essa população um grande reservatório de bacilos (BAÑULS et al., 2015). No estado de latência, a bactéria fica contida dentro de granulomas nos pulmões do hospedeiro (SHALER et al., 2013), mas um descontrole pode levar à ativação da doença, principalmente em casos de imunossupressão (MARCY et al., 2014). Devido à imaturidade dos seus sistemas imunológicos, as crianças não vacinadas por BCG tendem a desenvolver formas mais severas da doença, principalmente formas extrapulmonares como a infecção disseminada ou meningite (BATRA et al., 2012; CARVALHO et al., 2018).

A TB ativa em crianças apresenta sintomas variados que podem ser confundidos com outras doenças respiratórias. É comum em crianças com TB ativa a presença de tosse, febre, sudorese noturna e perda de peso (LAMB; STARKE, 2017). O tratamento para a TB infantil é feito com a utilização de 4 antibióticos (isoniazida, rifampicina, etambutol e pirazinamida) durante 2 meses, seguidos de 4 meses de utilização de rifampicina e isoniazida (SUÁREZ et al., 2019). Devido à rápida progressão da doença, as crianças muitas vezes começam o tratamento contra a TB antes mesmo da confirmação do diagnóstico micobacteriano (CHIANG; SWANSON; STARKE, 2015).

3.2 DIAGNÓSTICO

Um dos maiores desafios no controle da TB infantil é a obtenção de um diagnóstico preciso, uma vez que os testes utilizados atualmente foram desenvolvidos para adultos, estes

não consideram o quadro clínico distinto que a TB infantil apresenta (PEREZ-VELEZ; MARAIS, 2012). Isso faz com que a maioria dos casos infantis sejam diagnosticados através da observação de sinais e sintomas, exames de imagem e dados epidemiológicos da doença (CHIANG; SWANSON; STARKE, 2015).

Os testes utilizados para a confirmação da TB podem ser classificados em três grupos: microbiológicos, imunológicos e moleculares. Os principais testes microbiológicos são a cultura micobacteriana, utilizado como padrão-ouro, e o exame baciloscópico do escarro. Ambos os testes possuem sensibilidade reduzida quando aplicados em crianças, o primeiro tem 7 a 40% de sensibilidade (THOMAS et al., 2014), enquanto o segundo possui sensibilidade de até 15% (KUMAR; KUMAR; SINGH, 2015; KUNKEL et al., 2016).

Os testes imunológicos são o teste intradérmico da tuberculina (TST) e o ensaio de liberação do interferon gama (IGRA). O TST não é capaz de distinguir a forma ativa e latente da doença, também possui sensibilidade e especificidade reduzidas em crianças, além de gerar resultados falsos-positivos quando o paciente foi vacinado previamente pela vacina BCG (ASHLEY; SIEBENMANN, 1967). Mesmo quando o resultado do TST é negativo, não se pode excluir totalmente a infecção por MTB: aproximadamente 20% das crianças portadoras de TB ativa imunocompetentes não apresentam reação a este teste, gerando resultados falsos-negativos (DUNN; STARKE; REVELL, 2016). O IGRA tem maior especificidade e não gera confusão em pacientes vacinados pelo BCG, mas a sua sensibilidade ainda é reduzida em crianças e não é possível distinguir a forma latente e ativa da doença com este teste (LAMB; STARKE, 2017). Por fim, o teste molecular mais utilizado *GeneXpert MTB/RIF* também possui sensibilidade reduzida quando utilizado em crianças, com sensibilidade de aproximadamente 66% (THOMAS, 2017).

3.3 ASSINATURAS TRANSCRICIONAIS E METABÓLICAS

Atualmente, estudos têm buscado o diagnóstico confirmatório da tuberculose infantil em assinaturas transcricionais no sangue de pacientes portadores de TB. No ano de 2010, Berry *et al.* identificaram uma assinatura composta por 393 genes capazes de distinguir a TB ativa (BERRY et al., 2010). (KAFOROU et al., 2013) realizaram um estudo em adultos para classificar pacientes com TB (latente e ativa) e HIV. Neste trabalho foram identificados 27 genes capazes de distinguir as formas ativa e latente da TB. Também em 2013, Bloom *et al*. realizaram um estudo para distinguir pacientes com TB, sarcoidosis e doenças granulomatosas. Este estudo identificou uma assinatura composta por 144 genes com sensibilidade de 80% e

especificidade de 90% para classificar adultos com TB (BLOOM et al., 2013).

Anderson et al (2014) realizam um estudo para distinguir casos de TB infantil e de coinfecção por MTB e HIV. O estudo identificou uma assinatura composta por 51 genes com sensibilidade de 82,9% e especificidade de 83,6%. ZAK et al (2016) realizaram um estudo com adolescentes de 12 a 18 anos para predizer uma progressão da TB latente para ativa. O estudo foi realizado em forma de coorte, com acompanhamento dos participantes de 6 em 6 meses. Por fim, uma assinatura composta por 16 genes foi identificada, tendo sensibilidade de 66,1% e especificidade de 80,6%. Também (SWEENEY et al, 2016) realizaram uma meta análise de dados transcriptômicos para identificar uma assinatura para a TB ativa. Apenas 3 genes foram propostos como assinatura, tendo área abaixo da curva de 0,9 para TB ativa e 0,88 para TB latente nas curvas ROC.

Embora menos estudadas em comparação com assinaturas transcricionais, as assinaturas metabólicas são uma importante fonte de informação. Biomarcadores metabólicos já foram propostos para doenças como Alzheimer (HAN et al., 2002), Parkinson (BOGDANOV et al., 2008), diabetes (WANG et al., 2005) e pneumonia (LAIAKIS et al., 2010). A primeira assinatura metabólica para a TB foi proposta por (WEINER et al., 2012) para distinguir TB ativa e latente. Ao analisar cerca de 400 metabólitos por CL/EMAD de participantes adultos da África do Sul, foi proposta uma assinatura composta por 20 metabólitos para o diagnóstico da TB. Porém, ainda é necessária uma validação destes resultados em outros conjuntos de dados e populações. Em 2018, o mesmo grupo fez outra proposta de assinatura metabólica, desta vez para predizer a possibilidade de progressão da TB latente para ativa. Pacientes adultos da África do Sul, Etiópia, Gâmbia e Uganda foram recrutados para este estudo prospectivo e tiveram os seus metabólitos analisados por CL/EMAD. O estudo propôs uma assinatura metabólica composta por 10 metabólitos capazes de predizer a progressão da doença com 69% de sensibilidade e 75% de especificidade (WEINER et al., 2018). A única assinatura metabólica para a TB infantil foi proposta por (SUN et al., 2016). Composta por 3 metabólitos, esta assinatura teve sensibilidade e especificidade acima de 80% para classificar crianças com TB.

Um estudo publicado por (DUTTA et al., 2020) biomarcadores metabólicos para identificar crianças com TB, durante o tratamento da TB e após o tratamento da TB. Apenas um único metabólito, N-acetilneuraminato, foi capaz de diferenciar crianças de TB e do grupo controle com uma área abaixo da curva de 0,66. O metabólito quinolinato foi identificado como metabólito classificador, separando crianças do grupo controle e durante o tratamento da TB com uma área abaixo da curva de 0,77. Por fim, o metabólito pyroxidato conseguiu separar

crianças pós-tratamento da TB e crianças controle com área abaixo da curva de 0,87. Este estudo integrou o estudo do metaboloma e do transcriptoma ao correlacionar ambos os conjuntos de dados, utilizando também para isso a análise multi-ômica de fatores (*MOFA*) (ARGELAGUET et al., 2018). Este estudo demonstrou uma metodologia recente, identificando uma assinatura em um tipo de dado biológico e integrando os seus achados com outro tipo.

4 **MATERIAIS E MÉTODOS**

4.1 DESENHO DO ESTUDO

A Coorte para Pesquisa de Tuberculose pela Parceria Médica Indo-Americana (CTRIUMPH) é um estudo prospectivo observacional de adultos e crianças com tuberculose, este visa estudar fatores associados com a transmissão e progressão da tuberculose (GUPTE et al., 2016). A CTRIUMPh recrutou crianças com idade inferior a 15 anos em uma parceria acadêmica com a Universidade de Johns Hopkins para o estudo. O local do recrutamento foi a Faculdade de medicina do Estado Byramjee Jeejeebhoy, um hospital universitário terciário em Pune, cidade do estado de Maharashtra, Índia.

4.2 CRITÉRIOS DA AMOSTRAGEM

Todas as crianças participantes do estudo tinham idade inferior a 15 anos e tiveram o sangue total periférico coletado para o estudo do transcriptoma e plasma para o estudo do metaboloma. As crianças portadoras de TB (grupo "caso") do estudo tiveram o diagnóstico confirmado pelo teste *GeneXpert MTB/RIF*, cultura micobacteriana, ou presença de granulomas caseosos na histopatologia de amostras extrapulmonares. No momento do recrutamento e coleta de amostras (anos 2015 e 2016) as crianças portadoras de TB iniciaram o tratamento da doença, seguindo o protocolo padrão com isoniazida, rifampicina, etambutol e pirazinamida. Ao fim do tratamento, as crianças haviam sido curadas e não apresentaram relapso clínico ou microbiológico por 1 ano de acompanhamento observado. Amostras do plasma e sangue periférico total do grupo "caso" foram coletadas no momento do recrutamento dos participantes.

As amostras de controle foram coletadas de crianças com perfis de idade e sexo similares às crianças que fizeram parte do grupo caso e que eram contatos familiares de pacientes portadores de TB pulmonar. As crianças que compõem o grupo controle não apresentaram sintomas para triagem de TB e sinais de TB não foram observados na radiografia do tórax, também tiveram resultado negativo no teste TST (<5 mm) e IGRA no momento do recrutamento. As amostras dos participantes controle foram coletadas no momento do recrutamento e em visitas subsequentes após 4 e 12 meses. Além disso, os participantes controle foram testados novamente para TB após 6 e 12 meses para verificar uma possível conversão de controle para TB.

4.3 CONSIDERAÇÕES ÉTICAS

Esse estudo foi aprovado pelos comitês de revisão institucional da Faculdade de Medicina do Estado de Byramjee Jeejeebhoy e da escola de medicina da niversidade Johns Hopkins. Todos os experimentos foram realizados seguindo as diretrizes e regulamentos institucionais estabelecidos. Todos os participantes com idade inferior a 18 anos assinaram o termo de consentimento livre e esclarecido, fornecido pelos seus responsáveis legais.

4.4 DESCRIÇÃO DA POPULAÇÃO DO ESTUDO

A população recrutada para este estudo é composta por crianças da cidade de Pune, uma cidade do estado de Maharashtra, Índia. Todas as crianças participantes do estudo tinham idade inferior a 15 anos e nenhuma tinha co-infecção por HIV. No total, 54 amostras das crianças foram coletadas durante o estudo, as amostras tiveram os transcritos do sangue periférico total e metabólicos do plasma sanguíneo mensurados, gerando os dados transcriptômicos e metabolômicos. Destas amostras, 16 foram coletadas do grupo caso e 24 foram coletadas do grupo controle no momento do recrutamento, 8 amostras do grupo controle foram coletadas após 4 meses e 6 amostras do grupo controle após 12 meses (Tabela 1). A média das idades das crianças presentes nos dados é de 9 anos (intervalo interquartil de 6 a 12). Em relação ao gênero dos participantes, há 22 participantes do sexo feminino e 29 do sexo masculino, 3 não tiveram o gênero descrito. Devido ao baixo número de amostras presentes nos grupos controle após 4 meses e controle após 12 meses, apenas as amostras controle e caso do momento do recrutamento foram incluídas no estudo. No total, o conjunto de dados analisado é composto por 40 amostras, sendo destas 16 amostras do grupo caso e 24 amostras do grupo controle.

Tabela 1 - Distribuição das amostras do conjunto de dados transcriptômico e metabolômico, sumarizando a quantidade de amostras presentes em cada grupo e em cada período de coleta realizado. As coletas do grupo ocorreram no momento do recrutamento (*baseline*), as coletas do grupo controle foram realizadas no momento do recrutamento (*baseline*), seguidas de novas coletas 4 e 12 meses depois.

	Número de amostras
Caso baseline	16
Controle baseline	24
Controle 4m	8
Controle 12m	6
Total	54

Fonte: Elaborado pelo o autor

4.5 OBTENÇÃO DOS DADOS TRANSCRIPTÔMICOS

O sangue periférico total e o plasma das crianças foram coletados pela CTRIUMPh em parceria com a universidade de John Hopkins, na cidade de Pune, Índia. A mensuração dos transcritos das amostras foi realizada em um estudo anterior participante desse coorte (TORNHEIM et al., 2020). As amostras foram coletadas em tubos *PAXgene* e tiveram os seus transcritos extraídos utilizando kits comerciais de extração (*PAXgene Blood RNA kit*). Os transcritos isolados foram sequenciados por *RNA-seq* utilizando a plataforma *Hiseq 2500* da *Illumina*, seguindo as instruções do fabricante. Um total de 54 amostras foram sequenciadas e os dados estão disponíveis para análise.

4.6 OBTENÇÃO DOS DADOS METABOLÔMICOS

As amostras de plasma foram processadas pela empresa *Metabolon*, na cidade de Durham, Carolina do Norte, EUA. Os procedimentos foram realizados utilizando o sistema automatizado *MicroLab STAR* da empresa *Hamilton*, análises de controle de qualidade foram utilizadas seguindo protocolos anteriormente escritos (DUTTA et al., 2019; EVANS et al., 2009). Primeiro, diversos padrões de recuperação foram aplicados antes da extração para ter um melhor controle de qualidade do processo. As amostras foram colocadas em um evaporador *Turbovap* (*Zymark*) para a remoção do solvente orgânico. O extrato das amostras foi armazenado durante a noite anterior em nitrogênio antes da preparação para análise por cromatografia líquida de ultra-alta performance acoplada a espectrometria de massa em tandem

(*UPLC-MS/MS*). Todos os métodos foram realizados utilizando a cromatografia líquida de ultra performance *ACQUITY*, da *Waters* e o espectrômetro de massa de alta resolução/acurácia *Q-Exactive*, da *ThermoFisher Scientific* acoplado a uma fonte de ionização por eletrospray aquecida (*HESI-II*) e analisador de massas *Orbitrap*, sendo operado a uma resolução de 35.000 massas. O extrato das amostras foi secado e reconstituído em solventes compatíveis com cada um dos 4 métodos de análise diferentes: Otimizar a análise para compostos hidrofílicos, para condições de íons positivos ácidos, íons negativos básicos e para conteúdo orgânico geral mais alto. Os solventes de reconstituição foram utilizados em concentrações padrões fixas para garantir uma injeção e cromatografia consistentes. A análise de espectrometria de massas foi feita utilizando o protocolo da empresa *Metabolon*, realizando varreduras que cobriam massas de 70 a 1000 m/z. O dado bruto foi extraído, os picos foram identificados e os processos de controle de qualidade foram aplicados também utilizando protocolos da *Metabolon* (DEHAVEN et al., 2010) Por fim, os compostos foram identificados ao serem comparados com a biblioteca de compostos da empresa.

4.7 CONTROLE DE QUALIDADE E MAPEAMENTO DOS DADOS

O controle de qualidade do sequenciamento e a trimagem dos dados de *NGS* foram feitos utilizando o *Trimmomatic* (BOLGER; LOHSE; USADEL, 2014) no modo *paired-end* de trimagem, este processo é feito para retirar os adaptadores utilizados durante o *NGS*, leituras de baixa qualidade e leituras feitas com menos de 36 bases nitrogenadas. O mapeamento dos *reads* foi concluído com o algoritmo de pseudo-mapeamento *Salmon* (PATRO et al., 2017), utilizando o genoma de referência *GRCh38.p13* (GenBank ID: 8687898), apenas as amostras com mais de 60% de taxa de mapeamento foram mantidas. O *Salmon* realiza um pseudo-mapeamento das sequências fornecidas com o genoma de referência, quantificando as expressões dos genes. Este algoritmo foi utilizado para criar a tabela de contagem de expressão devido ao seu custo computacional reduzido e velocidade de processamento acelerada. O controle de qualidade dos dados de metaboloma foi realizado utilizando o *hardware* e *software* da empresa *metabolon*, seguindo os procedimentos descritos em anteriormente por Dehaven *et al.* (DEHAVEN et al., 2010).

4.8 ANÁLISE DOS DADOS TRANSCRIPTÔMICOS

Os genes diferencialmente expressos entre os grupos caso e controle foram calculados estatisticamente pelo pacote *DESeq2*. Para isso, os valores de expressão entre os grupos são comparados usando modelos lineares generalizados, a partir dos valores de média e dispersão de cada gene presente no dado, os modelos são feitos e ajustados considerando que os dados possuem uma distribuição binomial negativa (também chamada de distribuição *gamma-Poisson*). Dentro do próprio algoritmo, os dados são normalizados, tendo a média da expressão gênica definida a partir da concentração de fragmentos de cDNA dos genes multiplicada pelo fator de normalização do algoritmo. Uma matriz de *design* do estudo contendo o metadado das amostras é utilizada em conjunto com os dados de expressão, a partir disso o modelo linear ajustado retorna coeficientes de expressão geral do gene e o *log2* de *fold change* na comparação entre os grupos caso e controle. Os testes estatísticos para significância dos achados consideram a hipótese nula de que o *log2* de *fold change* é igual a zero, não havendo diferenças de expressão entre os grupos (LOVE; HUBER; ANDERS, 2014).

Na plataforma R (https://cran.r-project.org), a função *DESeq()* foi utilizada com os parâmetros padrões e os resultados foram extraídos pela função *results()*, comparando os grupos Caso vs Controle. O valor de p dos resultados foi ajustado por *False discovery rate (FDR)* de 0,05 e os genes diferencialmente expressos (*DEGs*) foram filtrados por *log* de *fold change* maior do que 1 e menor do que -1. Após isso, para normalizar os dados, a transformação estabilizadora de variância (*Variance Stabilizing transformation, VST*) do mesmo pacote, foi aplicada. Esta transformação faz com que a variância seja aproximadamente constante em relação às médias dos dados observados. Os grupos das amostras (caso e controle) são respeitados durante essa normalização, evitando assim a perda de informações.

Após isso, aplicamos um algoritmo de aprendizagem de máquina de floresta aleatória (*random forest*) (LIAW; WIENER, 2002) nos dados de expressão dos *DEGs* para identificar os melhores genes capazes de classificar as amostras. Para isso, a função *randomForest()* foi utilizada, com os parâmetros *ntree* = 1000 e *mtry* = 100. O primeiro parâmetro (*ntree*) define que serão geradas 1000 árvores de decisões na floresta aleatória, cada árvore testando um conjunto diferente de genes sorteados, o segundo parâmetro define quantos genes serão sorteados por árvore (*mtry*). As variáveis mais importantes para a classificação das amostras foram selecionadas de acordo com os valores de decréscimo médio de precisão e índice Gini. O decréscimo médio de precisão quantifica o quanto a retirada de uma variável afeta na precisão do modelo de classificação, e o índice Gini quantifica o quanto a presença daquela

variável contribui para uma classificação mais homogênea nos ramos das árvores. Curvas de característica de operação do receptor (ROC) (ROBIN et al., 2011) foram utilizadas para examinar a performance dos genes selecionados, exibindo no gráfico a sensibilidade e 1 - especificidade. A área sob a curva é utilizada para avaliar a performance do classificador em uma curva ROC, é a probabilidade da classificação correta de amostras, os valores de uma *AUC* variam entre 0,5 e 1, sendo o mínimo equivalente à aleatoriedade e o máximo à uma classificação perfeita.

4.9 METABÓLITOS CORRELACIONADOS E ENRIQUECIMENTO DE VIAS

Os dados gerados pela empresa *Metabolon* compilaram valores de abundância de 538 metabólitos extraídos do plasma de 121 participantes. Os metabólitos que possuíam valores de abundância e identificadores KEGG não disponíveis foram retirados do conjunto de dados. Após isso, os valores de abundância dos metabólitos foram transformados em log2, resultando assim em um conjunto de dados compilando valores de 122 metabólitos. Com o intuito de avaliar o impacto da expressão diferencial dos genes classificadores no metaboloma das amostras, uma análise de correlação entre os valores de expressão dos genes e os valores de abundância dos metabólitos foi realizada. Para maior consistência na análise de correlação, as amostras em comum dos dados transcriptômicos e metabolômicos foram selecionadas. Além disso, por se tratar de genes que possuem expressão diferencial entre os grupos, as amostras dos grupos caso e controle foram separadas e duas análises de correlação foram feitas, uma para o grupo caso e uma para o grupo controle.

As análises de correlação foram feitas na plataforma R, utilizando a função *rcorr()* do pacote *Hmisc (“The Hmisc and rms Packages”, 2016)* . Esta função gera uma matriz de correlações entre todas as variáveis disponíveis no dado, os valores de correlação têm base no coeficiente de correlação de classificações (*rank*) *rho* de *Spearman* (LEHMAN, 2005). Este coeficiente é ideal como medida não-paramétrica de correlação utilizando *ranking* de duas ou mais variáveis, servindo para identificar correlações que tenham caráter monotônico, ou seja, que a correlação seja constantemente positiva ou negativa entre as variáveis. Os valores dos coeficientes de correlação *rho* podem variar de -1 até 1 (SCHOBER; BOER; SCHWARTE, 2018), foram considerados metabólitos correlacionados aqueles que tiveram valor absoluto de $rho \geq 0{,}4$ e valor de $p < 0{,}05$.

Por fim, identificamos as vias metabólicas relacionadas a estes metabólitos que possuíam correlações com os transcritos classificadores. A análise de enriquecimento de vias

foi feita utilizando o pacote *MetaboAnalystR* (PANG et al., 2020) e o banco de dados *KEGG* (KANEHISA; GOTO, 2000). Os identificadores *KEGG* dos metabólitos correlacionados foram utilizados para fazer a análise de sobre-representação, permitindo assim identificar as funções biológicas mais relevantes no conjunto de dados. Estas funções são definidas ao avaliar a quantidade de metabólitos e as suas funções descritas no banco de dados. O valor de *foldchange* das vias enriquecidas são gerados ao dividir a quantidade de acertos (*Hits*) observados da via pela quantidade esperada (*Expected)* normalmente em um conjunto de dados.

4.10 VALIDAÇÃO DO CONJUNTO DE BIOMARCADORES EM OUTROS CONJUNTOS DE DADOS

Um total de 3 conjuntos de dados publicamente disponíveis no geoncbi (https://www.ncbi.nlm.nih.gov/geo/) foram selecionados para avaliar a performance dos genes selecionados na classificação de amostras infantis: GSE33939, GSE33940 e GSE41055. Os dois primeiros conjuntos, GSE39939 e GSE39940, fazem parte da super série GSE39941 e foram publicados em estudos realizados pelo grupo de Anderson et al. (2014). Ambos os conjuntos de dados são focados em crianças da África do Sul, com idade inferior a 15 anos, portadoras de TB latente (LTBI), TB ativa com e sem coinfecção por HIV e outras doenças. O terceiro conjunto de dados, GSE41055, foi publicado em um estudo realizado por Verhagen et al. (2013). Este estudo possui dados de crianças indígenas da etnia Warao da Venezuela, também com idade inferior a 15 anos, agrupadas em três grupos: crianças saudáveis (controle), portadoras de LTBI e TB ativa. A distribuição das amostras de cada conjunto de dados está disponível na tabela suplementar 3.

Os três conjuntos de dados não processados e metadados foram baixados utilizando o pacote *GEOquery* (DAVIS; MELTZER, 2007), as amostras *outliers* foram retiradas utilizando a função *outlier(s)*, do pacote *arrayQualityMetrics* (KAUFFMANN; GENTLEMAN; HUBER, 2009). Os dados das sondas foram colapsados por símbolo de gene utilizando a função *collapseRows()*, do pacote *WGCNA* (LANGFELDER; HORVATH, 2008). Após baixar os conjuntos de dados e processá-los, apenas amostras pertencentes aos grupos controle, LTBI, TB ativa sem coinfecção por HIV e TB ativa com coinfecção por HIV foram mantidas.

Para avaliar o desempenho classificatório do nosso conjunto de genes, os valores de expressão destes genes presentes nos conjuntos de dados foram utilizados para classificar as suas amostras. Por fim, curvas ROC foram utilizadas para sumarizar a performance classificatória dos genes em cada conjunto de dados.

Tabela suplementar 2 - Distribuição das amostras disponíveis nos três conjuntos de dados utilizados para validação. Apenas as amostras que não são *outliers* e que fazem parte dos grupos controle, LTBI, TB ativa e TB ativa HIV+ foram sumarizadas.

	Controle	LTBI	TB ativa	TB ativa HIV+	Total
GSE39939	0	14	50	27	91
GSE39940	0	52	68	39	159
GSE41055	9	9	9	0	27

Fonte: Elaborado pelo o autor

5 **RESULTADOS**

5.1 ANÁLISE DOS DADOS TRANSCRIPTÔMICOS: GENES DIFERENCIALMENTE EXPRESSOS

O conjunto de dados analisado tem valores de expressão de 18.543 transcritos de 40 amostras, sendo 16 amostras de TB e 24 controles. Os *DEGs* entre os grupos caso e controle foram calculados, para isso um conjunto de dados *DESeq* foi criado utilizando a função *DESeqDataSetFromMatrix()*, gerando um objeto *deseq* que continha os dados de expressão e o metadado com o agrupamento e identificador das amostras, o *design* da análise foi de acordo com o agrupamento das mesmas. O conjunto de dados *DESeq* foi analisado pela função *DESeq()* com os parâmetros padrões e os resultados extraídos utilizando a função *results()*, usando *fdr* de 0,05 como método de ajuste de valor de p. A tabela resultante contém o valor base das médias de expressão, o *log2* de *fold change*, o valor de erro padrão, estatísticas de teste, valor de p e valor de p ajustado.

A tabela resultante contém 174 *DEGs* com *log2* de *fold change* acima de 1 e abaixo de -1, com valores de p e de *fdr* abaixo de 0,05. Destes, 119 *DEGs* são super-expressos (*log2* de *fold change* > 1) e 55 *DEGs* são sub-expressos (*log2* de *fold change* < 1) no grupo caso. Os resultados foram sumarizados no gráfico de vulcão, com os valores de *log2* de *fold change* no eixo x e -*log10* do valor de p no eixo y. As linhas no eixo x representam *log2* de *fold chage* = ±1 e as linhas no eixo y representam o valor de p = 0,05. (Figura 1).

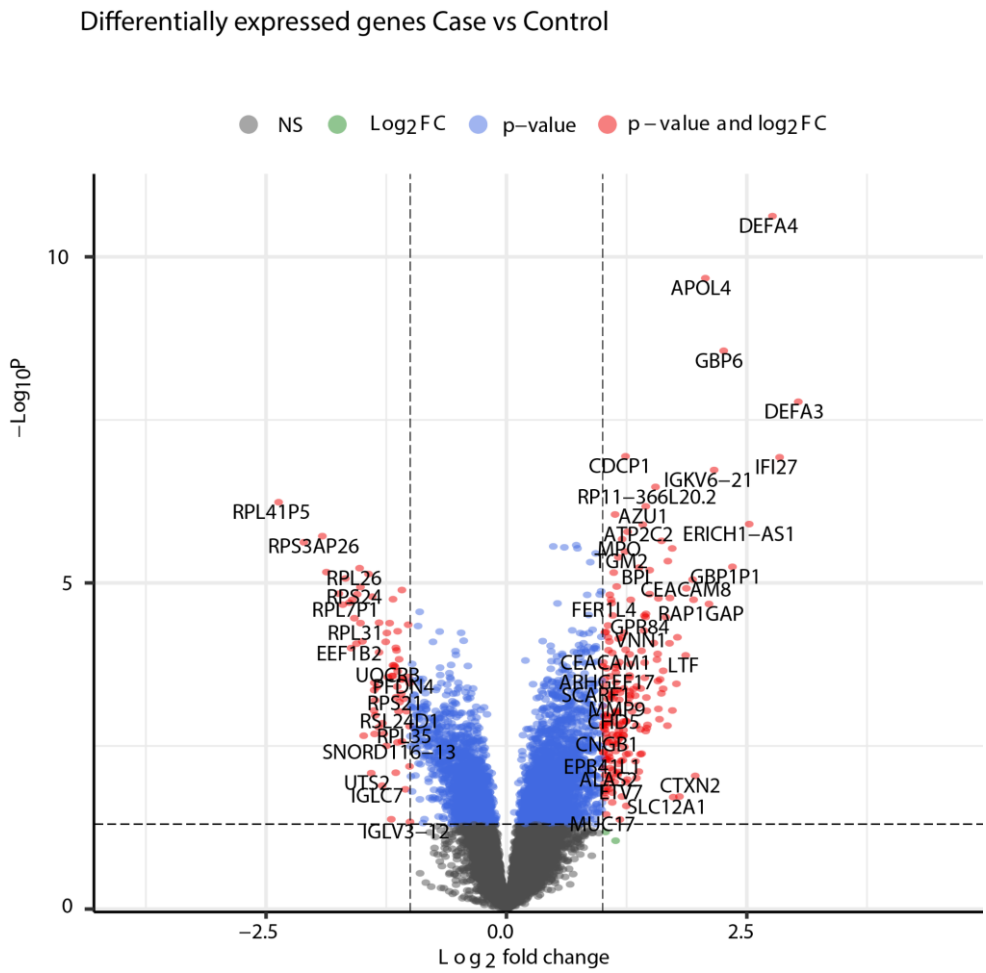

Figura 1 - *Volcano plot* sumarizando a distribuição dos genes diferencialmente expressos no baseline: no eixo X está quantificado o *log2* de *fold change* e no eixo Y o *-log10* do valor de P.
Fonte: Elaborado pelo o autor

5.2 APRENDIZAGEM DE MÁQUINA E CURVAS ROC

Para identificar os genes com maior poder de classificação das amostras, o algoritmo de aprendizagem de máquina de floresta aleatória foi aplicado. Os valores de expressão dos genes diferencialmente expressos foram utilizados para gerar 1000 árvores de decisões, sorteando aleatoriamente 100 genes para serem testados por árvore. Com isso, o algoritmo pôde avaliar a performance e identificar os melhores genes para classificar as amostras, os genes foram selecionados pelos valores de decréscimo médio de acurácia e índice Gini. Um total de 5 genes foram selecionados: *BPI*, *AZU1*, *C1QC*, *AC092580.4* e *MPO* (Figura 2).

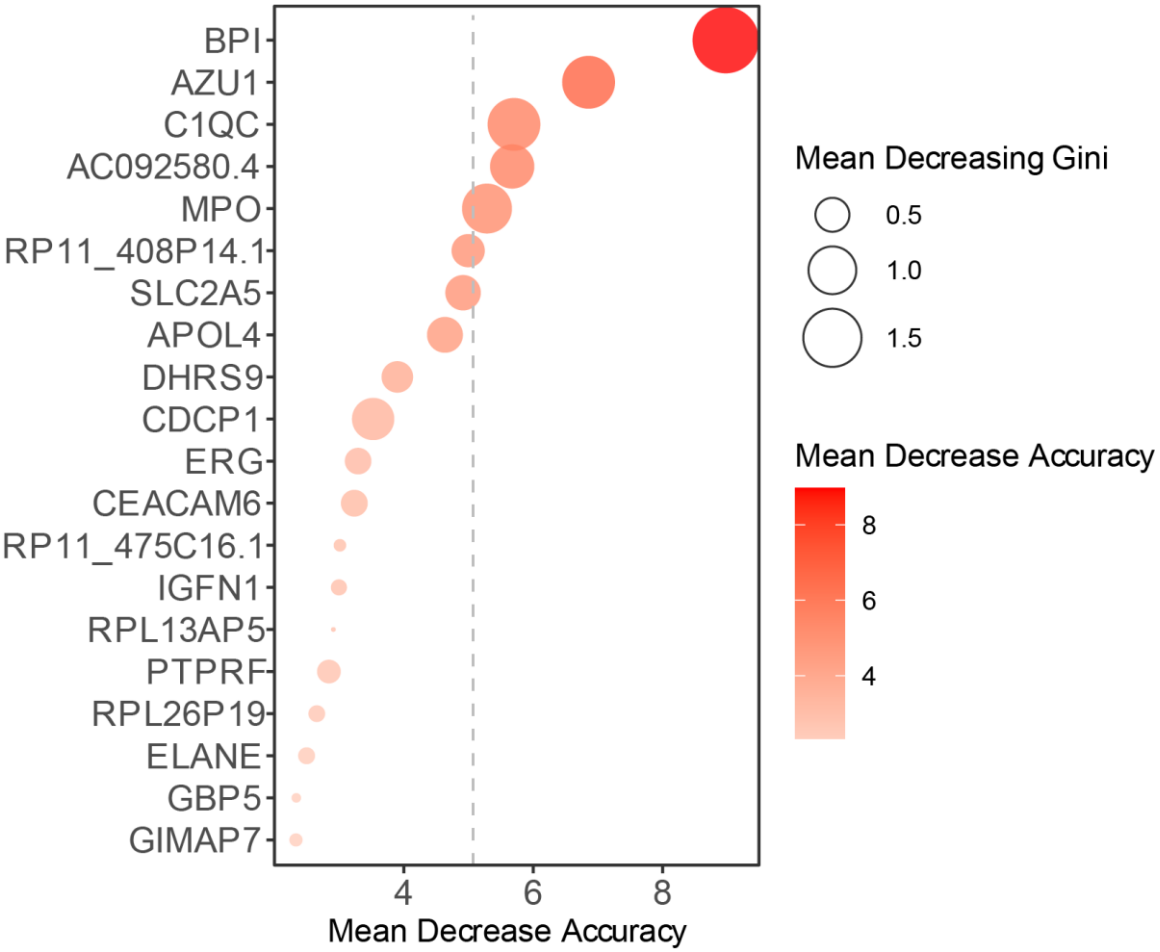

Figura 2 - Gráfico de decréscimo médio de acurácia da floresta aleatória. O tamanho das esferas indica o quanto a variável afeta o índice Gini, a coloração e posicionamento no eixo X indica o decréscimo médio de acurácia, a linha pontilhada é o limite do terceiro quartil de decréscimo médio de acurácia.
Fonte: Elaborado pelo o autor

As performances dos *DEGs* e dos genes selecionados pela floresta aleatória foram examinadas utilizando curvas ROC. Os valores de expressão desses genes tiveram as suas dimensões reduzidas por análise de componentes principais (JOLLIFFE, 1986) através da função *prcomp()*. Estes componentes foram utilizados em conjunto com o metadado das amostras para montar um modelo linear generalizado binomial, a função *fitted()* foi utilizada para extrair do modelo os valores de predição das amostras, valores acima de 0,5 indicam que o modelo classificou a amostra como caso e abaixo de 0,5 como controle. Comparando as predições do modelo com o agrupamento real das amostras, as curvas ROC foram geradas, ilustrando a sensibilidade e 1 - especificidade do modelo como classificador. As *AUCs* das curvas ROC dos *DEGs* e dos genes selecionados pela floresta aleatória foram de 0,86 (intervalo de confiança de 0,75 a 0,97) e de 0,91 (intervalo de confiança de 0,83 a 0,99) (Figura 3).

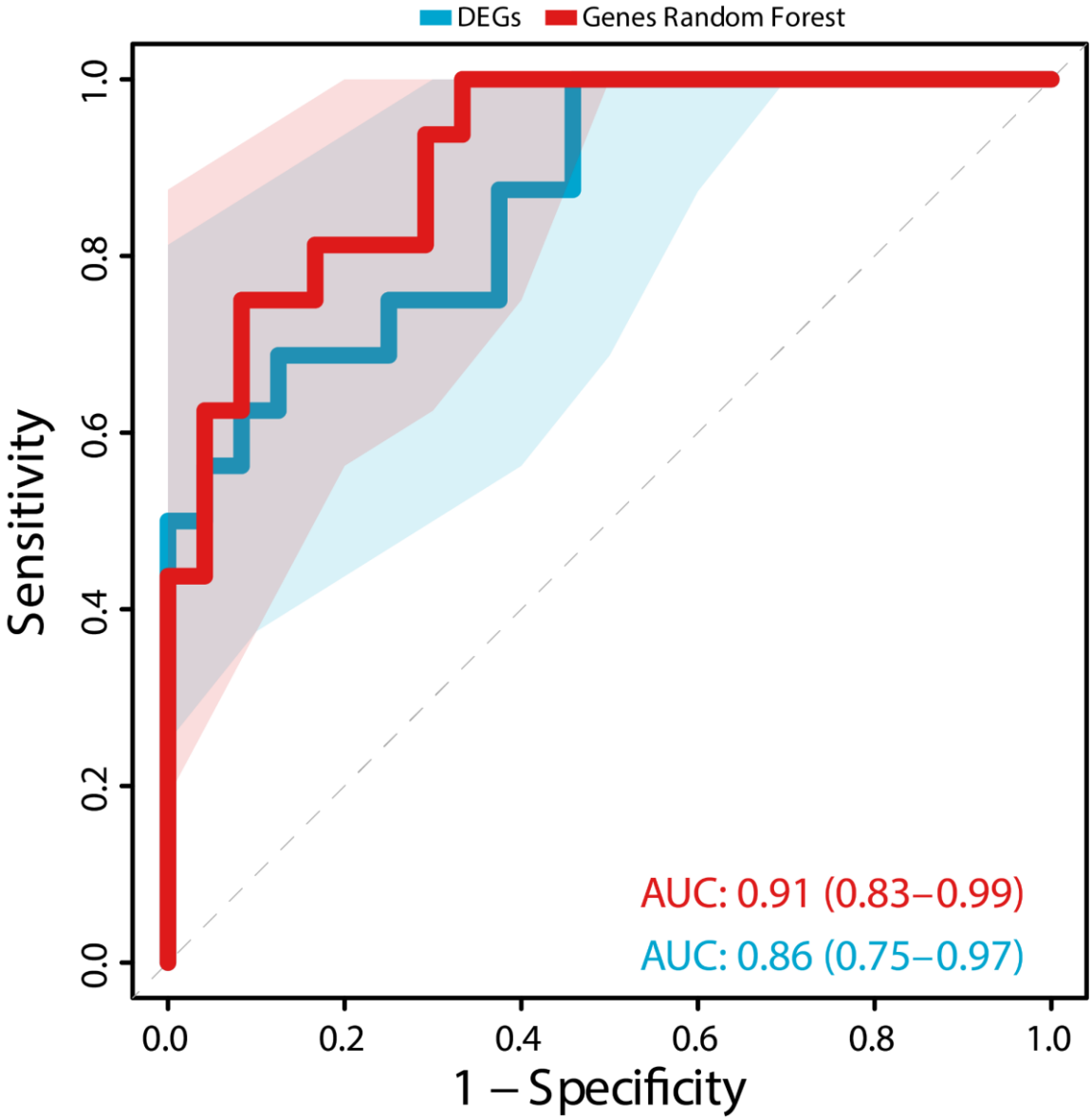

Figura 3 - Curva ROC demonstrando o desempenho dos 4 genes selecionados a partir da floresta aleatória e dos DEGs anteriormente identificados como classificadores. No eixo X está 1 - especificidade e no eixo Y a sensibilidade do classificador.
Fonte: Elaborado pelo o autor

5.3 CORRELAÇÃO COM O METABOLOMA

Os metabólitos que possuem correlações moderadas, fortes e muito fortes com os transcritos classificadores foram identificados, para isso foi gerada uma matriz de coeficientes de correlação de *ranking rho* de *Spearman*. Os coeficientes de correlação foram gerados pela função *rcorr*, correlacionando os valores de expressão dos genes classificadores com todos os metabólitos do conjunto de dados do metaboloma. Apenas as correlações com valor absoluto de $r \geq 0,4$ e valor de $p < 0,05$ foram consideradas significantes. Duas análises de correlação foram feitas: uma para as amostras do grupo caso e outra para as amostras do grupo controle.

No grupo caso foram observados 27 metabólitos correlacionados, sendo destes 25 com correlações positivas e apenas 2 com correlações negativas (Figura 4A). No grupo controle também foram observados 33 metabólitos, mas destes apenas 3 tiveram correlações positivas e 30 tiveram correlações negativas (Figura 4B). A lista dos metabólitos correlacionados está disponível como tabela suplementar.

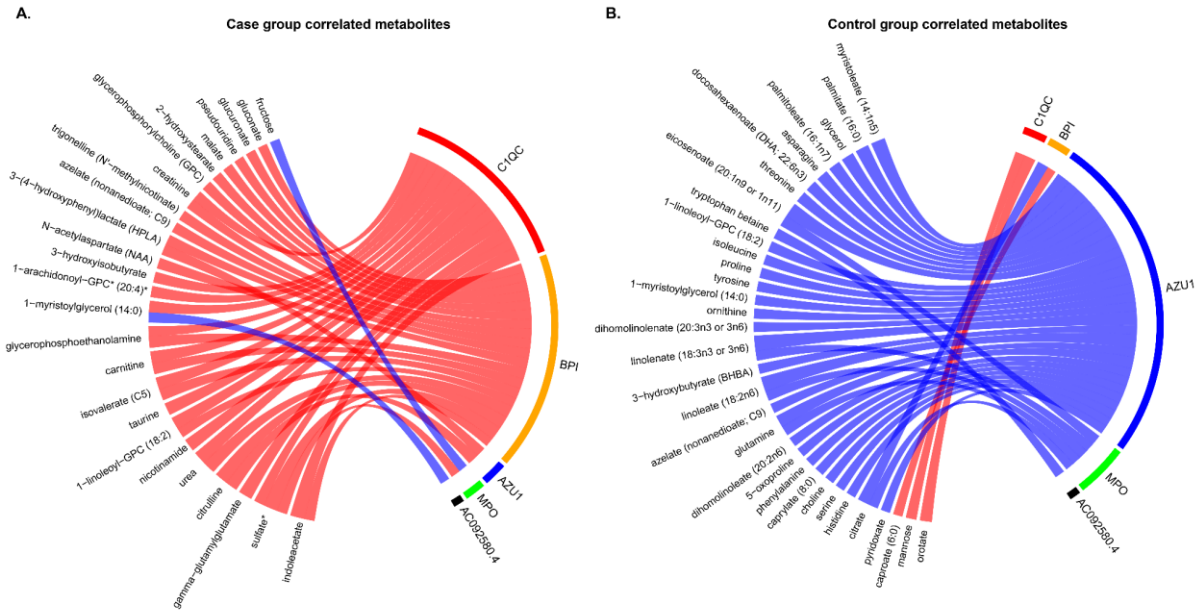

Figura 4 - A. Diagrama de acordes sumarizando os metabólitos correlacionados no grupo caso, as linhas de correlação em vermelho e azul indicam as correlações positivas e negativas respectivamente. B. Diagrama de acordes sumarizando os metabólitos correlacionados no grupo controle.
Fonte: Elaborado pelo o autor

5.4 ENRIQUECIMENTO DE VIAS

A análise de enriquecimento de vias do pacote *MetaboAnalystR* foi empregada para enriquecer as vias metabólicas mais representadas pelo conjunto de metabólitos que foram correlacionados aos transcritos. O enriquecimento compara os identificadores *KEGG* dos metabólitos com as suas funções biológicas anotadas no banco de dados, gerando assim uma tabela de vias que foram mais representadas. Esta análise foi feita separando os metabólitos de acordo com o sentido da correlação com os transcritos (positivo e negativo) e com o grupo em que a correlação foi observada (caso e controle). Os resultados do enriquecimento de vias foram sumarizados em um único gráfico para melhor comparação entre os grupos (Figura 5). As vias representadas por metabólitos que tiveram correlações negativas com os transcritos tiveram os seus *foldchanges* multiplicados por -1 para manter o sentido da correlação.

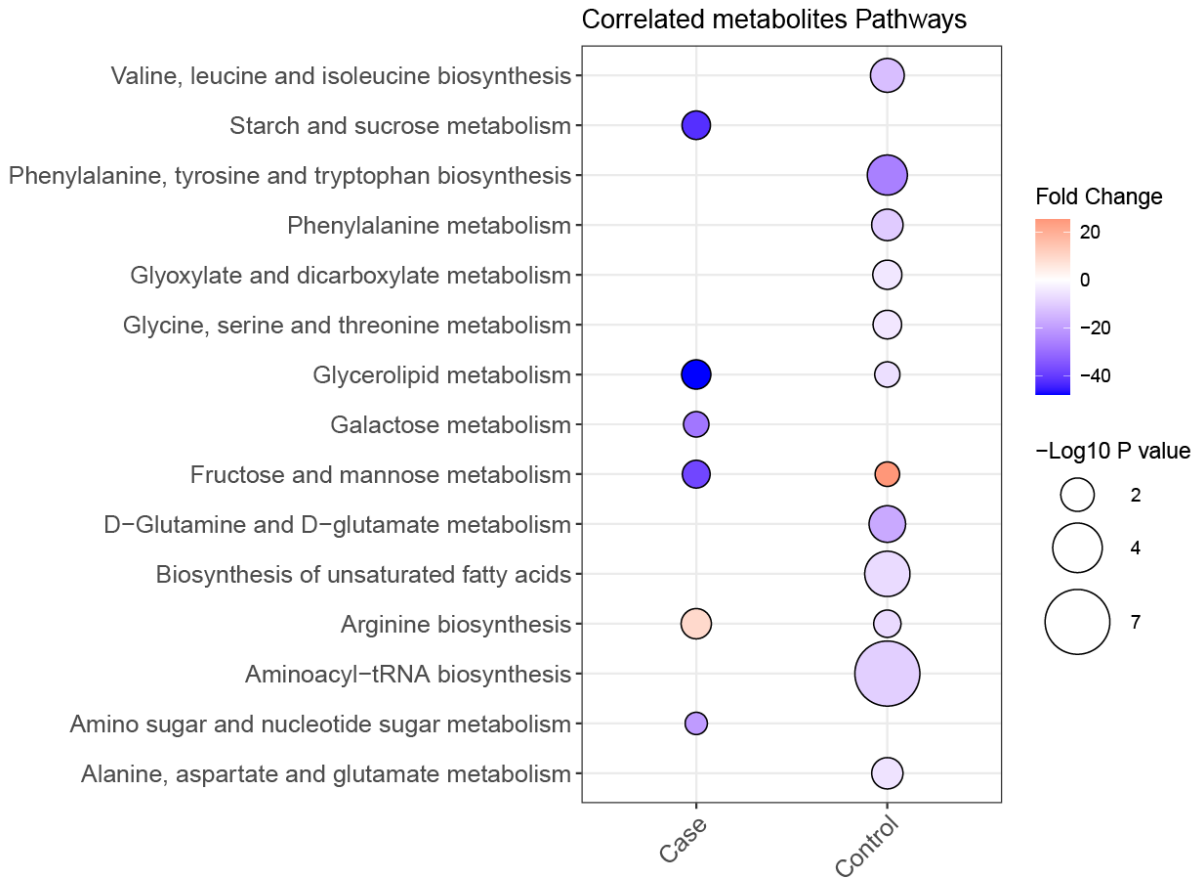

Figura 5 - Gráfico de pontos sumarizando as vias metabólicas enriquecidas. A cor do *Fold Change* de cada via varia de azul a vermelho de acordo com o seu valor, sendo azul para valores negativos e vermelho para valores positivos. O tamanho dos pontos é o valor de P transformado em -log10, sendo os menores valores de P representados por pontos maiores. Os grupos caso e controle são separados no eixo X do gráfico e as vias estão à esquerda do eixo Y.
Fonte: Elaborado pelo o autor

5.5 VALIDAÇÃO EM OUTROS CONJUNTOS DE DADOS

Os conjuntos de dados *GSE39939*, *GSE39940* e *GSE41055* foram baixados e processados individualmente para validação da performance classificatória do conjunto de genes. Para isso, os dados de expressão dos genes classificadores presentes nos conjuntos de dados foram isolados e utilizados para classificar as amostras. No conjunto de dados *GSE39939*, apenas 3 genes classificadores estavam presentes: *AZU1*, *BPI* e *MPO*. Este conjunto de dados possui 91 amostras, sendo 14 amostras de LTBI, 50 amostras de TB ativa e 27 amostras de TB ativa HIV+. Os dados de expressão destes genes foram utilizados para classificar amostras de LTBI vs TB ativa e LTBI vs TB ativa HIV+ com AUCs de 0,80 (I.C. 0.69~0,91) e 0,93 (I.C. 0,86~1,00) respectivamente (Figura 6A). No segundo conjunto de dados, *GSE39940*, também haviam apenas 3 genes classificadores: *AZU1*, *BPI* e *MPO*. Neste conjunto de dados há 159

amostras, sendo 52 amostras de LTBI, 68 amostras de TB ativa e 39 amostras de TB ativa HIV+. Estes genes conseguiram classificar amostras com AUCs de 0,85 (I.C. 0,78~0,92) e 0,93 (I.C. 0,87~0,98) ao separar LTBI vs TB ativa e LTBI vs TB ativa HIV+ respectivamente (Figura 6B). O conjunto de dados *GSE41055* possui 4 genes classificadores: *AZU1*, *BPI*, *C1QC* e *MPO*. No total, o conjunto de dados tem 27 amostras, sendo 9 amostras de LTBI, 9 amostras de TB ativa e 9 amostras controle. Os valores de expressão dos 4 genes classificaram as amostras dos grupos controle vs TB ativa e controle vs LTBI com AUCs de 0,70 (I.C. 0,44~0,96) e 0,77 (I.C. 0,54~1) respectivamente (Figura 6C).

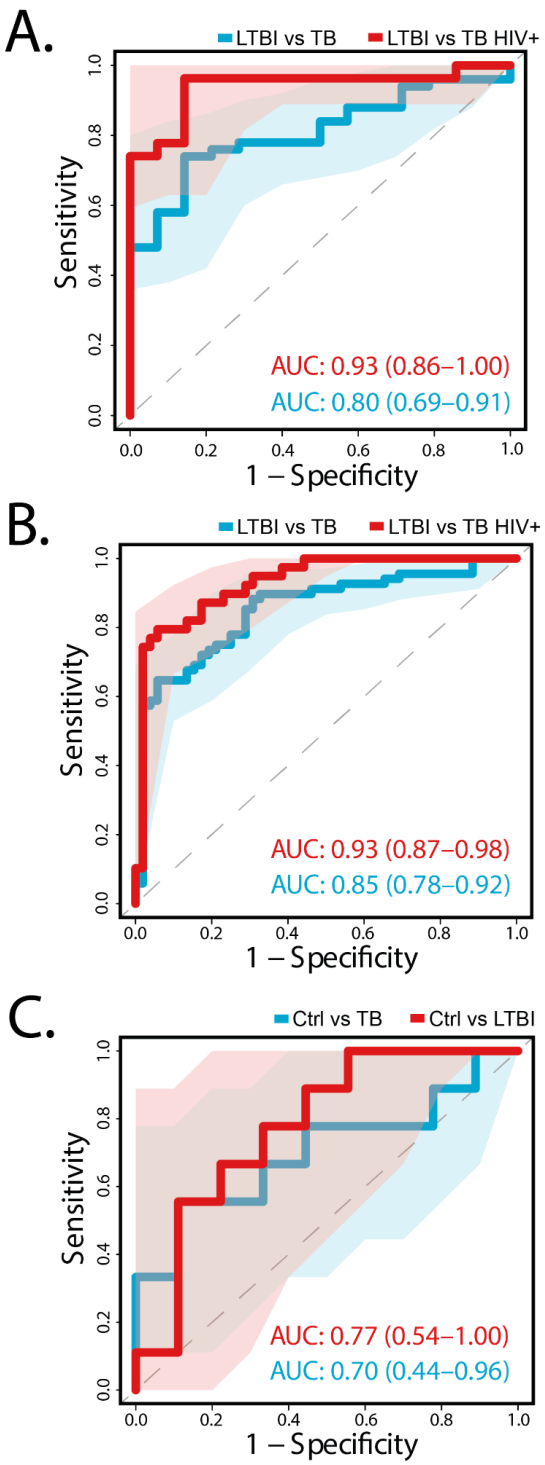

Figura 6 - A- Curva ROC para avaliar a performance classificatória dos genes *AZU1*, *BPI* e *MPO* ao classificar as amostras do conjunto de dados *GSE39939*. As curvas ROC em azul e vermelho sumarizam as performances dos genes na classificação de amostras LTBI vs TB e LTBI vs TB HIV+ respectivamente. **B-** Curva ROC sumarizando a performance dos genes *AZU1*, *BPI* e *MPO* ao classificar as amostras do conjunto de dados *GSE39940*. Em azul e vermelho estão as curvas ROC sumarizando a performance dos genes na classificação dos grupos LTBI vs TB e LTBI vs TB HIV+. **C. -** Curva ROC sumarizando a performance dos genes *AZU1, BPI, MPO* e *C1QC* ao classificarem amostras do conjunto de dados *GSE41055*. Em azul e vermelho estão as linhas com as performances na classificação de amostras dos grupos controle vs TB e controle vs LTBI.
Fonte: Elaborado pelo o autor

6 **DISCUSSÃO**

6.1 ESTUDO PRECEDENTE

Os dados utilizados para esse estudo foram utilizados anteriormente por *Dutta et al.* (DUTTA et al., 2020), em uma análise focada principalmente nos dados metabolômicos. Um metabólito classificador, o N-acetilneuraminato, foi identificado para separar pacientes portadores de TB de pacientes controle com AUC de 0,66; Outro metabólito, quinolinato, foi capaz de classificar pacientes durante o tratamento da TB e pacientes controle com uma *AUC* de 0,77; Por fim, piridoxato discriminou amostras pós-tratamento da TB de amostras controle com *AUC* de 0,87. Após identificar os metabólitos classificadores, o estudo fez a correlação destes metabólitos com o transcriptoma dos pacientes que tinham dados disponíveis. No total, 116 transcritos tiveram correlações com os 3 metabólitos classificadores do estudo. A análise de enriquecimento de vias foi feita utilizando estes transcritos, que tiveram correlações fortes com os metabólitos. O atual estudo foi feito com uma perspectiva de complementar o biomarcador metabolômico feito por *Dutta et al.*, realizando uma análise focada nos dados transcriptômicos para identificar alvos capazes de classificar as amostras. Os 5 genes classificadores identificados neste estudo não tiveram correlações com os metabólitos encontrados por *Dutta et al.*, sendo um novo conjunto de alvos para classificação de pacientes portadores de TB, uma alternativa que utiliza a mensuração dos transcritos em vez dos metabólitos. Ao avaliar a performance dos 3 metabólitos identificados por *Dutta et al.* em conjunto com os 5 genes classificadores identificados neste estudo, a AUC foi de 0,90 (I.C. 0,82~1) (Figura suplementar 1).

6.2 TRANSCRITOS CLASSIFICADORES

Os dados gerados por *NGS* de 40 amostras contendo os valores de expressão de 18543 transcritos foram analisados nesta parte do estudo. Os genes diferencialmente expressos entre os grupos caso e controle foram identificados utilizando o pacote *DESeq2*. No total, a análise de genes diferencialmente expressos identificou 174 *DEGs*, sendo 119 super-expressos e 55 sub-expressos no grupo caso em relação ao grupo controle. Após isso, o algoritmo de aprendizagem de máquina de floresta aleatória foi aplicado nos valores de expressão dos *DEGs*, com o intuito de identificar os genes com a maior acurácia ao classificar as amostras. O algoritmo identificou 5 genes capazes de classificar as amostras com TB e controle: *BPI, AZU1,*

C1QC, *AC092580.4* e *MPO*. Estes 5 genes classificadores tiveram uma *AUC* de 0,917 (Intervalo de confiança de 0,835~0,999) ao terem a sua performance avaliada por curva *ROC*.

O gene *BPI* é um gene que codificante de proteína de ligação a lipopolissacarídeo, sendo associado a grânulos de neutrófilos e tendo atividade antimicrobiana (https://www.genecards.org/cgi-bin/carddisp.pl?gene=BPI). Este gene já foi associado à ação antimicrobiana no controle da TB, tendo o aumento da sua expressão induzida pelos genes *RvD1* e *Mar1 (RUIZ et al., 2019)*. Os resultados observados pela análise dos genes diferencialmente expressos também apontou a expressão elevada do gene *BPI* no grupo portador de TB em relação ao grupo controle com *log* de *fold change* de 1,37.

O segundo gene selecionado, *AZU1*, codifica para uma proteína antibacteriana ligadora de heparina, derivada de grânulos de neutrófilos, específica de monócitos e fibroblastos (https://www.genecards.org/cgi-bin/carddisp.pl?gene=AZU1). A sua ação antibacteriana é limitada a maioria das bactérias gram negativas, uma vez que tende a se ligar ao envelope de lipopolissacarídeos que essas bactérias possuem. Além disso, também é um mediador de inflamação, promovendo o recrutamento de monócitos (MORGAN et al., 1991). Até o momento, este gene não foi associado como um gene importante para a TB, mas as suas funções são compatíveis com a fisiopatologia da doença, sendo relacionado ao processo inflamatório, mesmo que reduzido em crianças, que está infecção causa. Este gene foi superexpresso no grupo de TB em comparação com o grupo controle, tendo um *fold change* de 1,44.

O terceiro gene com maior desempenho classificatório, *C1QC*, codifica uma proteína que participa das cadeias *C1q*, do sistema de complemento do sistema imunológico humano (https://www.genecards.org/cgi-bin/carddisp.pl?gene=C1QC). A cadeia de componentes que este gene faz parte já foi proposta anteriormente como um biomarcador sérico para a TB ativa em adultos (LUBBERS et al., 2018), os resultados observados indicam uma boa performance de identificar TB ativa em crianças. A expressão deste gene foi muito elevada no grupo de TB em relação ao grupo controle, tendo log de *fold change* de 1,77. Por fim, o único gene sub-expresso no grupo TB em relação ao controle, *AC092580.4*, teve o log de *fold change* de -1,13. Este gene é um gene de RNA longo não-codificante que foi anteriormente associado à leucemia mielóide aguda (JIA et al., 2018), alterações na sua expressão podem afetar o sistema imunológico e os processos de diferenciação celular no sistema hematopoiético (FENG et al., 2018) para se conhecer totalmente as funções biológicas desse gene de RNA longo não-codificante ainda são necessários mais estudos. O último gene indicado, *MPO*, codifica para a enzima bactericida Myeloperoxidase, o principal componente dos grânulos azurófilos dos neutrófilos, produzido durante a diferenciação celular mielóide

(https://www.genecards.org/cgi-bin/carddisp.pl?gene=MPO). Este gene teve *log* de *fold change* de 1,20, sendo mais expresso no grupo caso do que no controle.

A maioria dos principais achados da análise dos dados transcriptômicos são genes codificantes para proteínas e enzimas antibacterianas presentes nos grânulos dos neutrófilos, o que corresponde com a patologia da TB pulmonar (PHILIPS; ERNST, 2012). Os neutrófilos são as primeiras células do sistema imune a entrar nos pulmões durante a infecção por MTB, são o tipo predominante de células nos pulmões de pessoas com TB pulmonar (EUM et al., 2010) e também carregam a maior carga bacteriana (EUM et al., 2010). Estas células já foram associadas e utilizadas como assinaturas para a TB pulmonar (BERRY et al., 2010) na doença, podendo ter atividade antimicobacteriana ou imunopatológica (LIU; LIU; GE, 2017).

Com estes achados, os resultados encontrados por *Dutta et al.* foram refinados, uma alternativa para classificar crianças portadoras de TB com uma assinatura transcriptômica composta por apenas 5 genes foi identificada. Além disso, essa assinatura também demonstrou boa performance ao classificar amostras de outros conjuntos de dados, incluindo as populações da África do sul (*GSE39939* e *GSE39940*) e da tribo indígena Warao, da Venezuela (*GSE41055*).

6.3 CORRELAÇÃO COM O METABOLOMA

Como uma análise exploratória a fim de avaliar a influência dos genes classificadores no metaboloma, os valores de expressão destes genes foram correlacionados com os valores de abundância de todos os metabólitos das amostras correspondentes. No total, 38 metabólitos foram correlacionados no grupo caso, sendo 36 positivamente correlacionados e apenas 2 negativamente. Os genes que mais tiveram correlações com os metabólitos foram o *BPI* e o *C1QC*, possuindo 19 e 14 correlações positivas respectivamente. No grupo controle houveram também 38 correlações significantes, mas apenas 3 destas foram positivas, enquanto 35 foram negativas. Os genes que mais tiveram correlações neste grupo foram o *AZU1* e o *MPO*, o primeiro gene teve 28 correlações negativas observadas, enquanto o segundo teve 5 correlações do mesmo tipo.

Os metabólitos que tiveram correlações significativas foram utilizados na análise de enriquecimento de vias pelo pacote *MetaboAnalystR*. Uma das principais vias identificadas em ambos os grupos foi a via de biossíntese de Arginina, esta foi enriquecida por metabólitos com correlações positivas no grupo caso e por metabólitos com correlações negativas no grupo controle. Na homeostase, a arginina é um aminoácido não-essencial, uma vez que a sua síntese

no eixo intestinal-renal consegue suprir as necessidades do organismo (BROSNAN; BROSNAN, 2004). Porém, a arginina passa a ser classificada como um aminoácido semi-essencial durante infecções crônicas, inflamações, ou após a ocorrência de lesões no intestino ou rins, uma vez que a sua síntese não é o suficiente para suprir as necessidades do organismo (MORRIS, 2016). Este aminoácido é de extrema importância na contenção da infecção por *Mycobacterium tuberculosis*, servindo como o substrato para a síntese de óxido nítrico, um dos principais fatores antimicobacterianos (KRAKAUER, 2019). Além disso, a arginina tem um papel importante na contenção do patógeno, sendo requerida por diversos tipos celulares do sistema imune, como linfócitos T, macrófagos e células *NK* para progressão do ciclo celular, proliferação e resposta imune (CROWTHER; QUALLS, 2020). O *fold change* positivo de 9,43 observado na via de biossíntese de arginina no grupo caso é condizente com o quadro de infecção por TB, no qual o organismo necessita sintetizar maiores quantidades de arginina para ter uma resposta imune contra o patógeno.

O metabolismo de frutose e manose também foi uma via metabólica enriquecida em ambos os grupos, mas com *fold change* negativo no grupo caso e positivo no grupo controle. Esta via não havia sido associada com reações do hospedeiro na patologia da TB anteriormente, apenas a frutose foi associada com o metabolismo da bactéria na forma de frutose-6-fosfato, sendo o substrato da fosfofrutoquinase, a principal enzima glicolítica do patógeno (SNÁŠEL et al., 2021). Mais estudos são necessários para elucidar a importância desta via na patologia da TB.

7 **CONCLUSÃO**

Os dados transcriptômicos e metabolômicos de crianças da cidade de Pune foram analisados para identificar alvos capazes de classificar crianças com TB ativa. Um novo conjunto de alvos composto por apenas 5 genes foi identificado, sendo estes os genes *BPI, AZU1, C1QC, AC092580.4* e *MPO*. A identificação deste conjunto de alvos é um passo à frente para o desenvolvimento de um teste diagnóstico para a TB infantil, uma alternativa ao conjunto anteriormente proposto por *Dutta et al.*, com dados transcriptômicos que possuem performance consistente.

Este conjunto demonstrou muita precisão ao classificar as amostras tanto no próprio conjunto de dados no qual foi definido, quanto em outros conjuntos de dados utilizados na validação. Ao ter a sua performance avaliada por curvas ROC, exibiu áreas sobre a curva de 0,917 (I.C. de 0,835~0,999) no conjunto *discovery*, 0,80 (I.C. 0.69~0,91) no conjunto *GSE39939*, 0,85 (I.C. 0,78~0,92) no conjunto *GSE39940* e 0,70 (I.C. 0,44~0,96) no conjunto *GSE41055*. Além disso, o fato deste conjunto de genes ter classificado também amostras de LTBI e TB ativa com coinfecção por HIV nos conjuntos de validação corrobora ainda mais para a sua consistência.

Observando as funções metabólicas dos genes, estas também condizem com a resposta do hospedeiro à bactéria *Mycobacterium tuberculosis*, sendo a maioria dos mesmos genes relacionados aos neutrófilos e resposta imune inflamatória. Também foi feita a correlação dos genes com os metabólitos e o enriquecimento de vias dos metabólitos correlacionados. As principais vias enriquecidas presentes em ambos os grupos também condizem com a patologia da TB, corroborando para a consistência dos nossos achados.

REFERÊNCIAS

ANDERSON, S. T. et al. Diagnosis of childhood tuberculosis and host RNA expression in Africa. **The New England Journal of Medicine**, v. 370, n. 18, p. 1712–1723, 1 maio 2014.

ARGELAGUET, R. et al. Multi-omics factor analysis-a framework for unsupervised integration of multi-omics data sets. **Molecular Systems Biology**, v. 14, n. 6, p. e8124, 20 jun. 2018.

ASHLEY, M. J.; SIEBENMANN, C. O. Tuberculin skin sensitivity following BCG vaccination with vaccines of high and low viable counts. **Canadian Medical Association Journal**, v. 97, n. 22, p. 1335–1339, 25 nov. 1967.

BAÑULS, A.-L. et al. Mycobacterium Tuberculosis: Ecology and Evolution of a Human Bacterium. **Journal of Medical Microbiology**, v. 64, n. 11, p. 1261–1269, nov. 2015.

BATRA, S. et al. Childhood tuberculosis in household contacts of newly diagnosed tb patients. **PloS One**, v. 7, n. 7, p. e40880, 31 jul. 2012.

BERRY, M. P. R. et al. An interferon-inducible neutrophil-driven blood transcriptional signature in human tuberculosis. **Nature**, v. 466, n. 7309, p. 973–977, 19 ago. 2010.

BLOOM, C. I. et al. Transcriptional blood signatures distinguish pulmonary tuberculosis, pulmonary sarcoidosis, pneumonias and lung cancers. **PloS One**, v. 8, n. 8, p. e70630, 5 ago. 2013.

BOGDANOV, M. et al. Metabolomic Profiling to Develop Blood Biomarkers for Parkinson's disease. **Brain: A Journal of Neurology**, v. 131, n. Pt 2, p. 389–396, fev. 2008.

BOLGER, A. M.; LOHSE, M.; USADEL, B. Trimmomatic: a flexible trimmer for illumina sequence data. **Bioinformatics**, v. 30, n. 15, p. 2114–2120, 1 ago. 2014.

BROSNAN, M. E.; BROSNAN, J. T. Renal arginine metabolism. **The Journal of Nutrition**, v. 134, n. 10 Suppl, p. 2791S–2795S; discussion 2796S–2797S, out. 2004.

CARVALHO, I. et al. Managing latent tuberculosis infection and tuberculosis in children. **Pulmonology**, v. 24, n. 2, p. 106–114, mar. 2018.

CAVILL, R. et al. Transcriptomic and metabolomic data integration. **Briefings in Bioinformatics**, v. 17, n. 5, p. 891–901, set. 2016.

CHIANG, S. S.; SWANSON, D. S.; STARKE, J. R. New Diagnostics for childhood tuberculosis. **Infectious Disease Clinics of North America**, v. 29, n. 3, p. 477–502, set. 2015.

CROWTHER, R. R.; QUALLS, J. E. metabolic regulation of immune responses to: a spotlight on l-arginine and l-tryptophan metabolism. **Frontiers in Immunology**, v. 11, p. 628432, 2020.

DAVIS, S.; MELTZER, P. S. Geoquery: a bridge between the gene expression omnibus (geo) and bioconductor. **Bioinformatics,** v. 23, n. 14, p. 1846–1847, 15 jul. 2007.

DEHAVEN, C. D. et al. Organization of GC/MS AND LC/MS metabolomics data into

chemical libraries. **Journal of Cheminformatics**, v. 2, n. 1, p. 9, 18 out. 2010.

DUNN, J. J.; STARKE, J. R.; REVELL, P. A. Laboratory diagnosis of mycobacterium tuberculosis infection and disease in children. **Journal of Clinical Microbiology**, v. 54, n. 6, p. 1434–1441, jun. 2016.

DUTTA, N. K. et al. Inhibiting the stringent response blocks entry into quiescence and reduces persistence. **Science Advances**, v. 5, n. 3, p. eaav2104, mar. 2019.

DUTTA, N. K. et al. Integration of metabolomics and transcriptomics reveals novel biomarkers in the blood for tuberculosis diagnosis in children. **Scientific Reports**, v. 10, n. 1, p. 19527, 11 nov. 2020.

EUM, S.-Y. et al. Neutrophils are the predominant infected phagocytic cells in the airways of patients with active pulmonary TB. **Chest**, v. 137, n. 1, p. 122–128, jan. 2010.

EVANS, A. M. et al. Integrated, Nontargeted Ultrahigh Performance Liquid Chromatography/electrospray Ionization Tandem Mass Spectrometry Platform for the Identification and Relative Quantification of the Small-Molecule Complement of Biological Systems. **Analytical Chemistry**, v. 81, n. 16, p. 6656–6667, 15 ago. 2009.

FENG, Y. et al. Expression profile analysis of long non-coding rna in acute myeloid leukemia by microarray and bioinformatics. **Cancer Science**, v. 109, n. 2, p. 340–353, fev. 2018.

GOLETTI, D. et al. Update on tuberculosis biomarkers: from correlates of risk, to correlates of active disease and of cure from disease. **Respirology** , v. 23, n. 5, p. 455–466, maio 2018.

GRAHAM, S. M. et al. Clinical case definitions for classification of intrathoracic tuberculosis in children: an update. **Clinical Infectious Diseases: An Official Publication of the Infectious Diseases Society of America**, v. 61Suppl 3, p. S179–87, 15 out. 2015.

GUPTE, A. et al. Cohort for tuberculosis research by the indo-us medical partnership (ctriumph): protocol for a multicentric prospective observational study. **BMJ Open**, v. 6, n. 2, p. e010542, 25 fev. 2016.

GUTIÉRREZ-GONZÁLEZ, L. H. et al. Immunological aspects of diagnosis and management of childhood tuberculosis. **Infection and Drug Resistance**, v. 14, p. 929–946, 8 mar. 2021.

HANIFA, Y. et al. Prevalence of Latent Tuberculosis Infection among Gold Miners in South Africa. **The International Journal of Tuberculosis and Lung Disease: The Official Journal of the International Union against Tuberculosis and Lung Disease**, v. 13, n. 1, p. 39–46, jan. 2009.

HAN, X. et al. Substantial sulfatide deficiency and ceramide elevation in very early alzheimer's disease: potential role in disease pathogenesis. **Journal of Neurochemistry**, v. 82, n. 4, p. 809–818, ago. 2002.

HOANG, L. T. et al. Transcriptomic signatures for diagnosing tuberculosis in clinical practice: a prospective, multicentre cohort study. **The Lancet Infectious Diseases**, v. 21, n. 3, p. 366–375, mar. 2021.

HOLMES, K. K. et al. (Org.). **Major Infectious Diseases**. Washington (DC): the international bank for reconstruction and development / the World Bank, 2018.

JIA, Y. et al. **Deregulated Expression of Long Non-Coding RNA AC092580.4 in Acute Myeloid Leukemia**. **Blood**. [S.l: s.n.]. Disponível em: <http://dx.doi.org/10.1182/blood-2018-99-114917>. 2018

JOLLIFFE, I. T. **Principal Component Analysis**. **Springer Series in Statistics**. [S.l: s.n.]. Disponível em: <http://dx.doi.org/10.1007/978-1-4757-1904-8>. , 1986.

KAFOROU, M. et al. Detection of tuberculosis in hiv-infected and -uninfected african adults using whole blood rna expression signatures: a case-control study. **PLoS Medicine**, v. 10, n. 10, p. e1001538, out. 2013.

KANEHISA, M.; GOTO, S. KEGG: Kyoto encyclopedia of genes and genomes. **Nucleic Acids Research**, v. 28, n. 1, p. 27–30, 1 jan. 2000.

KAUFFMANN, A.; GENTLEMAN, R.; HUBER, W. Arrayqualitymetrics--a bioconductor package for quality assessment of microarray data. **Bioinformatics** , v. 25, n. 3, p. 415–416, 1 fev. 2009.

KRAKAUER, T. Inflammasomes, autophagy, and cell death: the trinity of innate host defense against intracellular bacteria. **Mediators of Inflammation**, v. 2019, p. 2471215, 8 jan. 2019.

KUMAR, M. K.; KUMAR, P.; SINGH, A. Recent Advances in the Diagnosis and Treatment of Childhood Tuberculosis. **Journal of Natural Science, Biology, and Medicine**, v. 6, n. 2, p. 314–320, jul. 2015.

KUNKEL, A. et al. Smear positivity in paediatric and adult tuberculosis: systematic review and meta-analysis. **BMC Infectious Diseases**, v. 16, p. 282, 13 jun. 2016.

LAIAKIS, E. C. et al. Metabolomic analysis in severe childhood pneumonia in the Gambia, West Africa: findings from a pilot study. **PloS One**, v. 5, n. 9, 9 set. 2010. Disponível em: <http://dx.doi.org/10.1371/journal.pone.0012655>.

LAMB, G. S.; STARKE, J. R. Tuberculosis in Infants and Children. **Microbiology Spectrum**, v. 5, n. 2, abr. 2017. Disponível em: <http://dx.doi.org/10.1128/microbiolspec.TNMI7-0037-2016>.

LANGFELDER, P.; HORVATH, S. WGCNA: An R Package for Weighted Correlation Network Analysis. **BMC Bioinformatics**, v. 9, p. 559, 29 dez. 2008.

LEHMAN, A. **JMP for basic univariate and multivariate statistics**: a systematic guide. [S.l.]: SAS Press, 2005.

LIAW, A.; WIENER, M. **Classification and regression by randomforest**. [S.l: s.n.]. Disponível em: <https://CRAN.R-project.org/doc/Rnews/>. 2002.

LIU, C. H.; LIU, H.; GE, B. Innate Immunity in Tuberculosis: Host Defense vs Pathogen

evasion. **Cellular & Molecular Immunology**, v. 14, n. 12, p. 963–975, dez. 2017.

LOVE, M. I.; HUBER, W.; ANDERS, S. Moderated estimation of fold change and dispersion for RNA-Seq data with DESeq2. **Genome Biology**, v. 15, n. 12, p. 550, 2014.

LUBBERS, R. et al. Complement component C1q as serum biomarker to detect active tuberculosis. **Frontiers in Immunology**, v. 9, p. 2427, 23 out. 2018.

MARAIS, B. J. et al. The Spectrum of disease in children treated for tuberculosis in a highly ndemic area. **The International Journal of Tuberculosis and Lung Disease: The Official Journal of the International Union against Tuberculosis and Lung Disease**, v. 10, n. 7, p. 732–738, jul. 2006.

MARCY, O. et al. Causes and determinants of mortality in hiv-infected adults with tuberculosis: an analysis from the CAMELIA ANRS 1295-CIPRA KH001 randomized trial. **Clinical Infectious Diseases: An Official Publication of the Infectious Diseases Society of America**, v. 59, n. 3, p. 435–445, 1 ago. 2014.

MARTINEZ, L. et al. Tuberculin skin test conversion and primary progressive tuberculosis disease in the first 5 years of life: a birth cohort study from cape town, south africa. **The Lancet. Child & Adolescent Health**, v. 2, n. 1, p. 46–55, jan. 2018.

MORGAN, J. G. et al. Cloning of the cdna for the serine protease homolog cap37/azurocidin, a microbicidal and chemotactic protein from human granulocytes. **Journal of Immunology,** v. 147, n. 9, p. 3210–3214, 1 nov. 1991.

MORRIS, S. M., Jr. Arginine metabolism revisited. **The Journal of Nutrition**, v. 146, n. 12, p. 2579S–2586S, dez. 2016.

MURRAY, C. J. L. et al. Global, regional, and national incidence and mortality for hiv, tuberculosis, and malaria during 1990-2013: a systematic analysis for the global burden of disease study 2013. **The Lancet**, v. 384, n. 9947, p. 1005–1070, 13 set. 2014.

NICOL, M. P.; ZAR, H. J. New specimens and laboratory diagnostics for childhood pulmonary tb: progress and prospects. **Paediatric Respiratory Reviews**, v. 12, n. 1, p. 16–21, mar. 2011.

PANG, Z. et al. MetaboAnalystR 3.0: toward an optimized workflow for global metabolomics. **Metabolites**, v. 10, n. 5, 7 maio 2020. Disponível em: <http://dx.doi.org/10.3390/metabo10050186>.

PATRO, R. et al. Salmon Provides Fast and Bias-Aware Quantification of Transcript Expression. **Nature Methods**, v. 14, n. 4, p. 417–419, abr. 2017.

PEREZ-VELEZ, C. M.; MARAIS, B. J. Tuberculosis in children. **The New England Journal of Medicine**, v. 367, n. 4, p. 348–361, 26 jul. 2012.

PERLMAN, D. C. et al. Variation of Chest Radiographic Patterns in Pulmonary Tuberculosis by Degree of Human Immunodeficiency Virus-Related Immunosuppression. The Terry Beirn Community Programs for Clinical Research on AIDS (CPCRA). The AIDS Clinical Trials Group (ACTG). **Clinical Infectious Diseases: An Official Publication of the Infectious**

Diseases Society of America v. 25, n. 2, p. 242–246, ago. 1997.

PHILIPS, J. A.; ERNST, J. D. Tuberculosis pathogenesis and immunity. **Annual Review of Pathology**, v. 7, p. 353–384, 2012.

ROBIN, X. et al. pROC: an open-source package for r and s+ to analyze and compare ROC Curves. **BMC Bioinformatics**, v. 12, p. 77, 17 mar. 2011.

ROYA-PABON, C. L.; PEREZ-VELEZ, C. M. Tuberculosis exposure, infection and disease in children: a systematic diagnostic approach. **Pneumonia (Nathan Qld.)**, v. 8, p. 23, 24 nov. 2016.

RUIZ, A. et al. Resolvin D1 (RvD1) and Maresin 1 (Mar1) contribute to human macrophage control of m. tuberculosis infection while resolving inflammation. **International Immunopharmacology**, v. 74, p. 105694, set. 2019.

SCHOBER, P.; BOER, C.; SCHWARTE, L. A. Correlation coefficients: appropriate use and interpretation. **Anesthesia and Analgesia**, v. 126, n. 5, p. 1763–1768, maio 2018.

SHALER, C. R. et al. Within the enemy's camp: contribution of the granuloma to the dissemination, persistence and transmission of mycobacterium tuberculosis. **Frontiers in Immunology**, v. 4, p. 30, 14 fev. 2013.

SNÁŠEL, J. et al. Phosphofructokinases a and b from display different catalytic properties and allosteric regulation. **International Journal of Molecular Sciences**, v. 22, n. 3, 2 fev. 2021. Disponível em: <http://dx.doi.org/10.3390/ijms22031483>.

SUÁREZ, I. et al. The Diagnosis and Treatment of Tuberculosis. **Deutsches Arzteblatt International**, v. 116, n. 43, p. 729–735, 25 out. 2019.

SUN, L. et al. Utility of Novel Plasma Metabolic Markers in the Diagnosis of Pediatric tuberculosis: a classification and regression tree analysis approach. **Journal of Proteome Research**, v. 15, n. 9, p. 3118–3125, 2 set. 2016.

SWEENEY, T. E. et al. Genome-wide expression for diagnosis of pulmonary tuberculosis: A Multicohort Analysis. **The Lancet. Respiratory Medicine**, v. 4, n. 3, p. 213–224, mar. 2016.

The Hmisc and rms Packages. **Biostatistics and Computer-based Analysis of Health Data using R**. [S.l: s.n.]. Disponível em: <http://dx.doi.org/10.1016/b978-1-78548-088-1.50013-0>. 2016

THOMAS, T. A. et al. Intensified specimen collection to improve tuberculosis diagnosis in children from rural south africa, an observational study. **BMC Infectious Diseases**, v. 14, p. 11, 9 jan. 2014.

THOMAS, T. A. Tuberculosis in children. **Pediatric Clinics of North America**, v. 64, n. 4, p. 893–909, ago. 2017.

TOGUN, T. O. et al. Biomarkers for diagnosis of childhood tuberculosis: a systematic review. **PloS One**, v. 13, n. 9, p. e0204029, 13 set. 2018.

TORNHEIM, J. A. et al. Transcriptomic profiles of confirmed pediatric tuberculosis patients and household contacts identifies active tuberculosis, infection, and treatment response among indian children. **The Journal of Infectious Diseases**, v. 221, n. 10, p. 1647–1658, 27 abr. 2020.

VERHAGEN, L. M. et al. A Predictive Signature Gene Set for Discriminating Active from Latent Tuberculosis in Warao Amerindian Children. **BMC Genomics**, v. 14, p. 74, 1 fev. 2013.

WANG, C. et al. Plasma Phospholipid Metabolic Profiling and Biomarkers of Type 2 Diabetes Mellitus Based on High-Performance Liquid Chromatography/electrospray Mass Spectrometry and Multivariate Statistical Analysis. **Analytical Chemistry**, v. 77, n. 13, p. 4108–4116, 1 jul. 2005.

WEINER, J., 3rd et al. Biomarkers of inflammation, immunosuppression and stress with active disease are revealed by metabolomic profiling of tuberculosis patients. **PloS One**, v. 7, n. 7, p. e40221, 23 jul. 2012.

WEINER, J., 3rd et al. Metabolite Changes in Blood Predict the Onset of Tuberculosis. **Nature Communications**, v. 9, n. 1, p. 5208, 6 dez. 2018.

ZAK, D. E. et al. A Blood RNA Signature for Tuberculosis Disease Risk: A Prospective Cohort Study. **The Lancet**, v. 387, n. 10035, p. 2312–2322, 4 jun. 2016.

ZAR, H. J. et al. Tuberculosis Diagnosis in Children Using Xpert Ultra on Different Respiratory Specimens. **American Journal of Respiratory and Critical Care Medicine**, v. 200, n. 12, p. 1531–1538, 15 dez. 2019.

MATERIAL SUPLEMENTAR

Figura suplementar

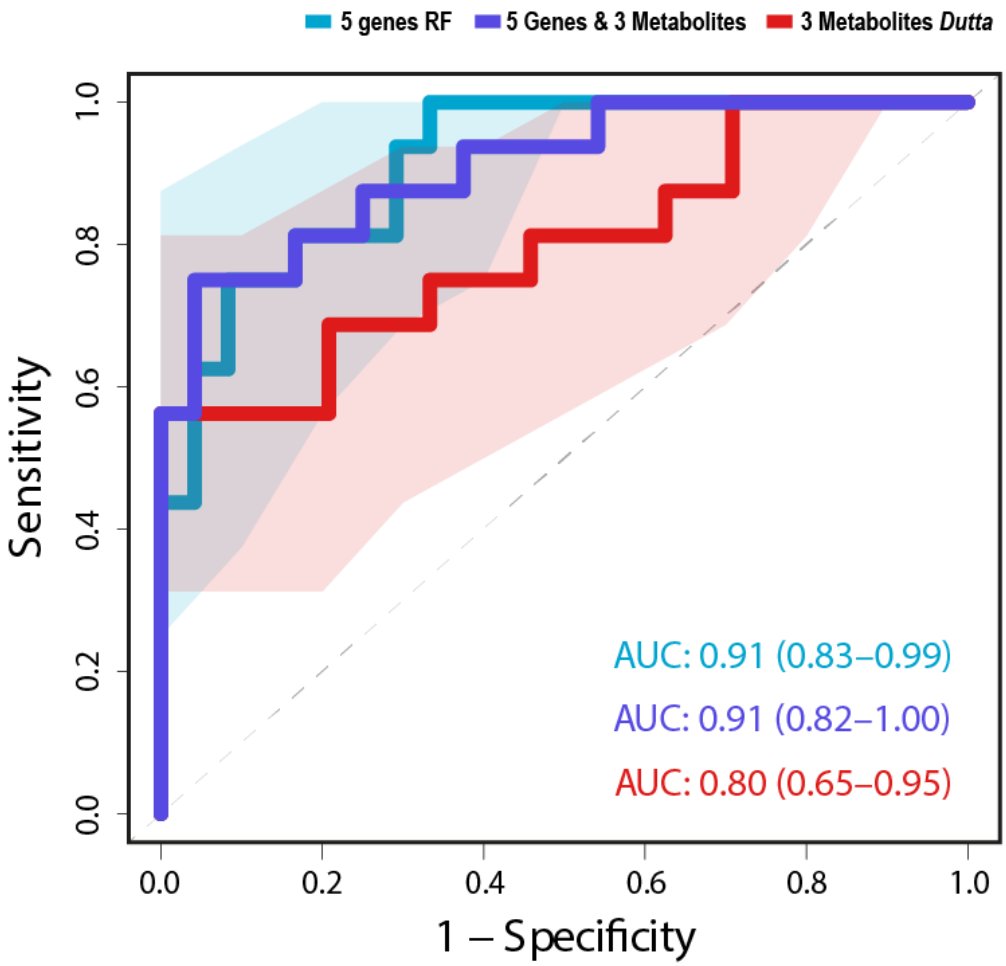

Figura suplementar 1 - Curva ROC sumarizando a performance classificatória dos conjuntos de biomarcadores ao classificar amostras dos grupos caso e controle no momento do recrutamento. Em azul claro está a curva dos 5 genes classificadores (*BPI*, *AZU1*, *C1QC*, *AC092580.4* e *MPO*). Em azul escuro está a curva dos 5 genes classificadores em combinação com os 3 metabólitos propostos por *Dutta el al.* (N-acetilneuraminato, piridoxato e quinolinato). Em vermelho está a curva dos 3 metabólitos propostos por *Dutta et al* isoladamente.