UNIVERSIDADE FEDERAL DE UBERLÂNDIA
INSTITUTO DE BIOTECNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM GENÉTICA E BIOQUÍMICA

ANÁLISE DE DADOS DE SEQUENCIAMENTO DE NOVA GERAÇÃO PARA MELHORIA DA SAÚDE HUMANA

Aluno: Tamires Caixeta Alves

Orientador: Dr. Matheus de Souza Gomes
Co-Orientadora: Dra. Leticia da Conceição Braga

UBERLÂNDIA - MG
2024

UNIVERSIDADE FEDERAL DE UBERLÂNDIA
INSTITUTO DE BIOTECNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM GENÉTICA E BIOQUÍMICA

ANÁLISE DE DADOS DE SEQUENCIAMENTO DE NOVA GERAÇÃO PARA MELHORIA DA SAÚDE HUMANA

Aluno: Tamires Caixeta Alves

Orientador: Dr. Matheus de Souza Gomes
Co-Orientadora: Dra. Leticia da Conceição Braga

Tese apresentada à Universidade Federal de Uberlândia como parte dos requisitos para obtenção do Título de Doutor em Genética e Bioquímica (Área Bioquímica).

UBERLÂNDIA - MG
2024

Dados Internacionais de Catalogação na Publicação (CIP)
Sistema de Bibliotecas da UFU, MG, Brasil.

A474a
2024

Alves, Tamires Caixeta, 1994-
Análise de dados de sequenciamento de nova geração para melhoria da saúde humana [recurso eletrônico] / Tamires Caixeta Alves. - 2024.

Orientador: Matheus de Souza Gomes.
Coorientadora: Leticia da Conceição Braga.
Tese (Doutorado) - Universidade Federal de Uberlândia, Programa de Pós-Graduação em Genética e Bioquímica.
Modo de acesso: Internet.
Disponível em: http://doi.org/10.14393/ufu.te.2023.8084
Inclui bibliografia.
Inclui ilustrações.

1. Bioquímica. I. Gomes, Matheus de Souza, 1981-, (Orient.). II. Braga, Leticia da Conceição, 1970-, (Coorient.). III. Universidade Federal de Uberlândia. Programa de Pós-Graduação em Genética e Bioquímica. IV. Título.

CDU: 577.1

André Carlos Francisco
Bibliotecário - CRB-6/3408

UNIVERSIDADE FEDERAL DE UBERLÂNDIA
Coordenação do Programa de Pós-Graduação em Genética e Bioquímica
Av. Pará 1720, Bloco 2E, Sala 244 - Bairro Umuarama, Uberlândia-MG, CEP 38400-902
Telefone: +55 (34) 3225-8438 - www.ppggb.ibtec.ufu.br - ppggb@ufu.br

ATA DE DEFESA - PÓS-GRADUAÇÃO

Programa de Pós-Graduação em:	Genética e Bioquímica				
Defesa de:	Doutorado Acadêmico - PPGGB				
Data:	Vinte e três de janeiro de dois mil e vinte e quatro	Hora de início:	13:30h	Hora de encerramento:	17:39
Matrícula do Discente:	11923GBI015				
Nome do Discente:	Tamires Caixeta Alves				
Título do Trabalho:	Análise de dados de Sequenciamento de Nova Geração para melhoria da Saúde Humana.				
Área de concentração:	Bioquímica				
Linha de pesquisa:	Biologia Molecular				
Projeto de Pesquisa de vinculação:	Aplicação de métodos moleculares e computacionais para genômica, proteômica e transcriptoma.				

Aos vinte e três dias do mês de janeiro de dois mil e vinte e quatro, às 13:30 horas, reuniu-se via web conferência pela Plataforma *Microsoft Teams*, em conformidade com a Portaria nº 36, de 19 de março de2020 da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – CAPES, Resolução de nº06/2020 e Resolução nº 19/2022 do Conselho de Pesquisa e Pós-graduação pela Universidade Federal de Uberlândia, a Banca Examinadora, designada pelo Colegiado do Programa de Pós-graduação em Genética e Bioquímica, assim composta: Drª. Carolina Pereira de Souza Melo, Drª. Renata Guerra de Sá Cota, Dr. Aulus Estevão Anjos de Deus Barbosa, Dr. Laurence Rodrigues do Amaral e Dr. Matheus de Souza Gomes, orientador (a) do (a) candidato (a) e demais convidados presentes conforme lista de presença. Iniciando os trabalhos o (a) presidente da mesa, Dr. Matheus de Souza Gomes apresentou a Comissão Examinadora e o (a) candidato (a), agradeceu a presença do público, e concedeu o (à) Discente a palavra para a exposição do seu trabalho. A duração da apresentação do (a) Discente e o tempo de arguição e resposta foram conforme as normas do Programa de Pós-graduação em Genética e Bioquímica. A seguir o (a) senhor (a) presidente concedeu a palavra, pela ordem sucessivamente, aos examinadores, que passaram a arguir o (a) candidato (a). Ultimada a arguição, que se desenvolveu dentro dos termos regimentais, a Banca, em sessão secreta, atribuiu os conceitos finais. Em face do resultado obtido, a Banca Examinadora considerou o candidato (a):

APROVADA

Esta defesa de Tese de Doutorado é parte dos requisitos necessários à obtenção do título de Doutor. O competente diploma será expedido após cumprimento dos demais requisitos, conforme as normas do Programa, a legislação pertinente e a regulamentação interna da UFU. Nada mais havendo a tratar foram encerrados os trabalhos. Foi lavrada a presente ata que após lida e achada conforme foi assinada pela Banca Examinadora.

Documento assinado eletronicamente por **Matheus de Souza Gomes**, **Professor(a) do Magistério Superior**, em 23/01/2024, às 17:40, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Laurence Rodrigues do Amaral**, **Professor(a) do Magistério Superior**, em 23/01/2024, às 17:40, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Renata Guerra de Sá Cota**, **Usuário Externo**, em 23/01/2024, às 17:40, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Aulus Estevão Anjos de Deus Barbosa**, **Professor(a) do Magistério Superior**, em 23/01/2024, às 17:41, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Carolina Pereira de Souza Melo**, **Usuário Externo**, em 23/01/2024, às 17:41, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

A autenticidade deste documento pode ser conferida no site https://www.sei.ufu.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **5061045** e o código CRC **5ED1AAEB**.

Referência: Processo nº 23117.089932/2023-86 SEI nº 5061045

UNIVERSIDADE FEDERAL DE UBERLÂNDIA
INSTITUTO DE BIOTECNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM GENÉTICA E BIOQUÍMICA

ANÁLISE DE DADOS DE SEQUENCIAMENTO DE NOVA GERAÇÃO PARA MELHORIA DA SAÚDE HUMANA

ALUNO: Tamires Caixeta Alves

COMISSÃO EXAMINADORA

Presidente: Dr. Matheus de Souza Gomes

Examinadores: Dr. Aulus Estevão Anjos de Deus
Dra. Carolina Pereira de Souza Melo
Dr. Laurence Rodrigues do Amaral
Dra. Renata Guerra de Sá Cota

Data da Defesa: 23 / 01 / 2024

As sugestões da Comissão Examinadora e as Normas do PGGB para o formato da Dissertação/Tese foram contempladas

MATHEUS DE SOUZA GOMES

DEDICO,

Aos meus estimados professores,

Àqueles que, ao longo da minha jornada acadêmica, me guiaram com sabedoria, paciência e dedicação, esta tese é dedicada a vocês. Das primeiras lições na educação infantil até o doutorado, cada um de vocês contribuiu de maneira significativa para minha formação e crescimento como acadêmica.

AGRADEÇO,

Ao meu orientador, Dr. Matheus de Souza Gomes, que sempre me desafiou a alcançar o meu potencial máximo. Seu apoio e orientação moldaram minha jornada acadêmica e profissional. Agradeço também pela confiança e perseverança em mim como pesquisadora.

À Dra. Letícia da Conceição Braga e ao Dr. Gabriel da Rocha Fernandes, que abriram as portas de seus laboratórios e disponibilizaram de seu tempo (e paciência) para me ensinar e orientar. A maior parte do que sei sobre câncer e análise de dados de sequenciamento, respectivamente, se deve a vocês.

À Ana Paula e ao Fábio por terem me ajudado na parte experimental dos dois artigos. Vocês foram fundamentais para que os sequenciamentos ocorressem da melhor forma possível!

À minha família e aos meus amigos, por estarem por perto quando precisei e compreenderem minha ausência quando foi necessário. Obrigada pelo suporte e admiração.

Aos demais colegas dos laboratórios LBAM (UFU), CEBio (FioCruz-Minas) e IEPI (IMP), com os quais compartilhei as dificuldades acadêmicas e pessoais, agradeço por sempre me ouvirem e proporcionarem um ambiente acolhedor. Isso desempenhou um papel fundamental na construção desta tese.

À Universidade Federal de Uberlândia, aos órgãos de fomento (Capes, FAPEMIG e CNPq) e ao Estado brasileiro pelos recursos e oportunidades em todos os níveis da minha formação, que foi inteiramente pública desde o ensino fundamental.

“Existem muitas hipóteses em ciência que estão erradas.
Isso é perfeitamente aceitável, eles são a abertura
para achar as que estão certas.”
Carl Sagan

SUMÁRIO

APRESENTAÇÃO.....11

CAPÍTULO I: AVANÇOS E DESAFIOS DO SEQUENCIAMENTO DE NOVA GERAÇÃO: aplicações e implicações na saúde humana.....13

CAPÍTULO II: Identification and characterization of microRNAs in *Biomphalaria tenagophila* and comparative analysis of their expression in *Schistosoma mansoni*-resistant and -susceptible snail populations.....37

CAPÍTULO III: Mitochondrial DNA Variants and OXPHOS Pathway in High-Grade Ovarian Cancer: Challenging the Metabolic Paradigm.....77

APRESENTAÇÃO

O Sequenciamento de Nova Geração (NGS) tem revolucionado a pesquisa científica, fornecendo insights mais profundos sobre a genética e pavimentando o caminho para uma melhor compreensão dos perfis genéticos individuais, diagnósticos mais acurados e o desenvolvimento de terapias mais efetivas e personalizadas. Nos últimos anos, testemunhamos avanços notáveis no campo das ciências ômicas, impulsionados pela tecnologia NGS. No entanto, a integração desta tecnologia na prática clínica rotineira e na saúde pública ainda enfrenta desafios significativos, apesar de suas numerosas implicações positivas.

Neste contexto, a presente tese é estruturada em capítulos que exploram este tema em profundidade. No primeiro capítulo, um artigo de revisão, submetido ao Mário Penna Journal – *Molecular and Clinical Research* (MPJ), aborda os métodos de sequenciamento, as metodologias de análise de dados gerados por essa tecnologia e suas aplicações no campo biológico e médico. As diversas plataformas de sequenciamento são descritas detalhadamente, abarcando desde o método de Sanger, passando por tecnologias como Roche/454, Illumina, Ion Torrent, ABI/SOLiD, PacBio até Oxford Nanopore. Após elucidar as metodologias empregadas por essas plataformas, são apresentadas as ferramentas utilizadas na análise de dados. Detalha-se um procedimento passo-a-passo que inclui a verificação da qualidade dos dados, remoção de adaptadores, alinhamento, montagem, chamada de variantes e análise da expressão gênica diferencial. Além disso, são discutidas abordagens como o sequenciamento de genoma completo (WGS), de exoma (WES), RNA-seq, entre outras, e suas respectivas aplicações.

No Capítulo II, é apresentado o artigo publicado "*Identification and Characterization of MicroRNAs in Biomphalaria tenagophila and Comparative Analysis of Their Expression in Schistosoma mansoni-Resistant and -Susceptible Snail Populations*". Este estudo foi conduzido com o objetivo de investigar diferenças na expressão de miRNAs entre populações de caramujos suscetíveis e resistentes, visando encontrar soluções para combater a esquistossomose.

Por fim, no Capítulo III, discute-se o artigo "*Mitochondrial DNA Variants and OXPHOS Pathway in High-Grade Ovarian Cancer: Challenging the Metabolic*

Paradigm", submetido à revista *Cancer Genetics*. Este trabalho envolve o uso de sequenciamento de DNA mitocondrial (mtDNA) para uma análise abrangente de mutações presentes no mtDNA, com o intuito de avaliar a relação da via de fosforilação oxidativa com o câncer de ovário seroso de alto grau.

CAPÍTULO I

(Artigo de revisão submetido ao *Mário Penna Journal - Molecular and Clinical Cancer Research*)

AVANÇOS E DESAFIOS DO SEQUENCIAMENTO DE NOVA GERAÇÃO: aplicações e implicações na saúde humana

Mário Penna Journal - molecular and clinical cancer research, Belo Horizonte, v. x, p. x-x, jan./jun. 20xx. DOI:

Mini Review

AVANÇOS E DESAFIOS DO SEQUENCIAMENTO DE NOVA GERAÇÃO: aplicações e implicações na saúde humana

ADVANCEMENTS AND CHALLENGES IN NEXT-GENERATION SEQUENCING: applications and implications in human health

Tamires Caixeta Alves
Programa de Pós-Graduação em Genética e Bioquímica, Instituto de Biotecnologia, Universidade Federal de Uberlândia - UFU, *campus* Patos de Minas, Patos de Minas, MG, Brasil
ORCID: 0000-0002-1507-2811
alvesctamires@gmail.com

Leticia da Conceição Braga
Instituto de Inovação, Ensino e Pesquisa do Instituto Mário Penna
ORCID: 0000-0002-6181-9410
leticia.braga@mariopenna.org.br

Matheus de Souza Gomes
Instituto de Biotecnologia, Universidade Federal de Uberlândia - UFU, *campus* Patos de Minas, Patos de Minas, MG, Brasil
ORCID: 0000-0001-7352-3089
matheusgomes@ufu.br

RESUMO

Este estudo revisa o Sequenciamento de Nova Geração (NGS) e explora suas aplicações na saúde humana. O objetivo é analisar os avanços e desafios dessa tecnologia. Utilizando uma abordagem de revisão, foram examinados estudos sobre os avanços tecnológicos no campo do NGS. Os resultados destacam o impacto significativo do NGS na identificação de variantes genéticas, diagnóstico de doenças genéticas e no desenvolvimento de terapias personalizadas. Além disso, aborda desafios como o gerenciamento e análise de grandes volumes de dados gerados pelo NGS. Conclui-se que o NGS desempenha um papel crucial na medicina personalizada, mas enfrenta desafios complexos que exigem soluções inovadoras para maximizar seu potencial na melhoria da saúde humana.

Palavras-chave: Sequenciamento de Nova Geração, Genética Médica, Medicina Personalizada

ABSTRACT

This study reviews Next-Generation Sequencing (NGS) and its applications in human health. The objective is to analyze the advancements and challenges of this technology. Using a review approach, studies, and technological developments in the field of NGS were examined. The results highlight the significant impact of NGS on the identification of genetic variants, diagnosis of genetic diseases, and the development of personalized therapies. Additionally, it addresses challenges such as managing and analyzing large volumes of data generated by NGS. It is concluded that NGS plays a crucial role in personalized medicine but

faces complex challenges that require innovative solutions to maximize its potential in improving human health.

***Keyword**: Next-Generation Sequencing, Medical Genetics, Personalized Medicine*

Data de submissão:

Data de aprovação:

1. INTRODUÇÃO

O dogma central da Biologia proposto por Crick[1] explica os mecanismos de expressão e transmissão da informação genética, afirmando que o DNA codifica a produção do RNA por transcrição e o RNA, por sua vez, codifica a produção de proteínas pelo processo de tradução. Assim, a informação armazenada no DNA pode ser acessada pelo conhecimento do conteúdo de nucleotídeos que compõe o mesmo. Neste contexto, o sequenciamento genético é o processo de inferência dos nucleotídeos que compõem os ácidos nucleicos (DNA ou RNA) em sua exata ordem. Com essa tecnologia, é possível determinar qual a sequência de nucleotídeos de determinado indivíduo estudado. Ao longo dos anos, diferentes tecnologias de sequenciamento foram surgindo até chegarmos ao *Next-Generation Sequencing* (NGS), ou Sequenciamento de Nova Geração. Esta tecnologia, relativamente recente, usada para sequenciar DNA e RNA em larga escala a fim de compreender a informação genética de um organismo desejado.

2. HISTÓRIA DO SEQUENCIAMENTO

O primeiro método para determinação da sequência de ácidos nucleicos foi o sequenciamento de Sanger[2], publicado em 1977. Esse sequenciamento foi amplamente utilizado, desde sua descoberta, e ainda é usado em algumas pesquisas até nos dias de hoje. Esse método consiste no uso de um nucleotídeo de terminação da cadeia nomeado didesoxinucleotídeo[2]. Para a realização do experimento são necessários: DNA molde, DNA polimerase, nucleotídeos, *primers* e os didesoxinucleotídeos. A fita de DNA molde terá o seu conteúdo sequenciado durante a reação de síntese da cadeia complementar pela DNA polimerase. Os

nucleotídeos e *primers* são elementos essenciais para a esse processo, enquanto os didesoxinucleotídeos são a chave do método, pois a DNA polimerase adiciona os nucleotídeos em cadeia baseado na estrutura química (com uma hidroxila na posição 3' da ribose dos nucleotídeos convencionais) e sem essa hidroxila a cadeia é interrompida. Sendo assim, fragmentos de tamanhos diferentes são gerados, todos contendo um didesoxinucleotídeo marcado em sua extremidade. A partir de uma eletroforese em gel, é possível obter a sequência formada baseada na informação destes fragmentos diversos[2]. Na mesma época, em 1977, Maxam e Gilbert também publicaram um método de sequenciamento por um processo químico que quebra parcialmente uma molécula de DNA marcada radioativamente na parte terminal em cada repetição de uma base[3]. Os comprimentos das sequências marcadas identificam as posições das bases e a leitura dos fragmentos é realizada através da eletroforese em gel de poliacrilamida[3]. Tanto Fredrick Sanger quanto Walter Gilbert foram laureados com o Prêmio Nobel pela descoberta do método de sequenciamento de DNA. No entanto, o método de Sanger foi muito mais utilizado, pela facilidade do seu protocolo e uso de reagentes menos nocivos à saúde humana. Por isso, o sequenciamento de Sanger dominou esse mercado por mais de 30 anos e foi amplamente utilizado para o sequenciamento do genoma de diversos organismos.

O primeiro genoma sequenciado foi o da bactéria *Haemophilus influenzae*, em 1995[4], utilizando o método de Sanger automatizado. Em 1996, o primeiro eucarioto, *Saccharomyces cerevisiae*, também utilizando sequenciamento de Sanger automatizado[5]. O projeto Genoma Humano foi sequenciado e publicado em 2001, usando duas abordagens e duas equipes independentes. As estratégias utilizadas foram *shotgun*, usado pela Celera Genomics[6], e BAC *contigs*, pelo National Human Genome Research Institute (NHGRI, NIH). Ambas as abordagens usaram o método de Sanger automatizado para determinar os nucleotídeos sequenciados.

No entanto, o método de Sanger tem suas limitações. Além do custo por base sequenciada ser muito alto, a processividade é baixa. Logo, o número de sequências de DNA a serem sequenciadas é muito menor do que com as tecnologias atuais. Com a modernização da Medicina, e cada vez mais emergente

a necessidade de uma Medicina Personalizada, novas tecnologias foram desenvolvidas e alcançaram maior protagonismo nesse cenário. O desenvolvimento de tecnologia de NGS de alto rendimento provou que pode gerar dados enormes (milhões de sequências) de forma rápida e econômica.

3. *NEXT-GENERATION SEQUENCING* (NGS)

Desde 2005, ocorreram mudanças significativas no cenário global de sequenciamento. As mudanças de geração são marcadas pelas mudanças na química e plataformas. O método de sequenciamento de DNA de Sanger e o método de Maxam-Gilbert são considerados sequenciamentos de primeira geração. Por esses métodos, fragmentos de DNA de até 1 quilobase podem ser sequenciados de uma vez só, com uma precisão altíssima.

A tecnologia que sucedeu a primeira geração, conhecida como tecnologia de sequenciamento de nova geração (NGS), ou tecnologias de sequenciamento de alto rendimento, é marcada pela capacidade de produzir uma enorme quantidade de dados com um custo muito baixo e uma velocidade rápida. Tecnologias NGS podem sequenciar em paralelo de milhões a bilhões de leituras em uma única execução e o tempo necessário para gerar as leituras do tamanho gigabase é de apenas alguns dias ou horas[7], tornando-o melhor do que o sequenciamento de primeira geração, como o sequenciamento Sanger. Para fins de comparação, exemplifico o próprio genoma humano. Ele possui 3 bilhões de pares de base (pbs) e é formado por moléculas de DNA com comprimentos que variam de 33 a ~247 milhões de pbs, distribuídas nos 23 cromossomos localizados em cada núcleo de células humanas. O sequenciamento do genoma humano utilizando o método de Sanger levou quase 15 anos, exigiu a cooperação de mais de um laboratório e teve custo aproximado de US$ 3 bilhões[7], enquanto o sequenciamento por sequenciadores NGS usando o 454 *Genome Sequencer FLX*, em 2008, levou dois meses e por aproximadamente um centésimo do custo[8].

As tecnologias de sequenciamento de segunda geração se referem às mais novas tecnologias de sequenciamento desenvolvidas após a primeira geração[9].

Roche 454, *Ion Torrent*, Illumina-Solexa e ABI-SOLiD fazem parte dessa segunda geração de sequenciadores. Essa segunda geração é caracterizada pela necessidade de preparar bancos de sequenciamento amplificados antes de iniciar o sequenciamento de clones de DNA amplificados. Há a necessidade da fragmentação do DNA, construção das bibliotecas e, só depois, o sequenciamento em si[9]. São gerados milhões de leituras curtas em paralelo. E, por isso, isso exige altos recursos computacionais, para posterior montagem ou mapeamento. As características básicas são: a geração de muitos milhões de leituras curtas em paralelo, velocidade de sequenciamento do processo em comparação com a primeira geração, o baixo custo de sequenciamento e a saída do sequenciamento é detectada diretamente sem a necessidade de eletroforese.

Outra divisão das tecnologias NGS são as tecnologias de sequenciamento de terceira geração que surgiram recentemente e, ao contrário da segunda geração, essas tecnologias são classificadas como Tecnologia de Sequenciamento de Molécula Única (SMRT), por sua capacidade de sequenciar uma única molécula sem a necessidade da criação de bibliotecas de amplificação. Elas são capazes de gerar leituras mais longas com custos muito menores e em menos tempo[10].

3.1 Princípios e Tecnologias do NGS

Dentro das tecnologias de segunda geração, com sequenciamento de leituras curtas, existe uma divisão de duas abordagens: sequenciamento por ligação e sequenciamento por síntese. E são classificadas principalmente por quatro plataformas: Roche/454 lançada em 2005[11], Illumina/Solexa em 2006[12], em 2007 a ABI/SOLiD (http://www.appliedbiosystems.com/)[13] e, por último, em 2010, Ion Torrent[14].

3.1.1 O sequenciamento Roche/454

O sequenciamento Roche/454 utiliza a técnica de pirosequenciamento para adição de nucleotídeos, uma abordagem de sequenciamento por síntese[11]. Esse método se baseia na detecção de pirofosfato liberado após cada incorporação de nucleotídeo na nova fita sintética de DNA. Amostras de DNA são fragmentadas e cada fragmento (400-600 pbs) se liga a adaptadores e esses adaptadores são

complementares a *primers* ligados a esferas magnéticas. Assim, cada esfera está associada a um único fragmento de DNA[15]. As esferas são filtradas para que permaneçam apenas esferas com fragmentos de DNA ligados e são emulsificadas e amplificadas por PCR, gerando cerca de um milhão de cópias de cada fragmento na superfície das esferas. Estas esferas são transferidas para uma placa de picotitulação e há ativação de uma série de reações produzindo luz a cada incorporação de nucleotídeo[16]. A luz emitida é detectada após cada incorporação de nucleotídeo e a sequência do fragmento de DNA é inferida.

No sequenciamento Roche/454 milhares de reações ocorram em paralelo, isso aumenta consideravelmente o rendimento do sequenciamento. Alguns equipamentos dessa plataforma geram leituras com comprimentos de até 1.000 pb e pode produzir aproximadamente 1 milhão de leituras por execução. Essas leituras são relativamente longas, logo, mais fáceis de serem mapeadas a um genoma de referência. O principal problema dessa plataforma são inserções e deleções devido à presença de regiões homopolímerica[17]. Como a técnica é dependente da intensidade da luz emitida por pirosequenciamento, os sinais com intensidade muito alta ou muito baixa levam à sub ou superestimação do número de nucleotídeos, o que causa erros na identificação de nucleotídeos. E isso diminui a confiabilidade dos resultados.

3.1.2 Sequenciamento *Ion Torrent*

A mais recente das tecnologias de sequenciamento de segunda geração, o Ion Torrent se baseia na detecção de íons liberados durante a síntese do DNA[14]. A amplificação clonal também é baseada em esferas em um micro poço por meio de PCR em emulsão. As sequências adaptadoras são ligadas aos fragmentos de DNA e capturadas juntamente com uma esfera coberta com adaptadores complementares, desoxinucleotídeos (dNTPs), *primers* e DNA polimerase. A detecção se difere do Roche/454. O sequenciamento Ion Torrent usa um chip que contém um conjunto de micro poços e cada um possui uma esfera com vários fragmentos idênticos[18]. Ao incorporar cada nucleotídeo com um fragmento, é liberado um íon hidrogênio que altera o pH da solução. Essa alteração é detectada por um sensor acoplado ao fundo do micro poço e convertida em um sinal de

voltagem proporcional ao número de nucleotídeos incorporados[10]. A tecnologia do Ion Torrent foi a primeira a realizar sequenciamento de semicondutores sem sensoriamento óptico.

A capacidade de produção desta tecnologia é de comprimentos de leitura de 200-600 pbs, maiores do que outros sequenciadores de segunda geração, e um tempo de execução extremamente rápido, entre 2 e 8 horas[10]. A principal desvantagem é a dificuldade de interpretação das sequências homopoliméricas (mais de 6 pb)[20]. Devido à própria química da técnica, na incorporação de múltiplas bases idênticas, pode haver uma perda de linearidade de resposta devido a medições imprecisas de magnitudes de pulso de tensão. Assim podem aparecer como erros de inserção/exclusão em uma única leitura, o que causa erros de inserção e deleção (indel) com uma taxa de cerca de ~1%. Por outro lado, a tecnologia apresenta uma baixa taxa de erro de substituição (por taxa básica de <0,1%)[21].

3.1.3 Sequenciamento Illumina/Solexa

A tecnologia Illumina também faz parte dos sequenciadores de segunda geração, tem abordagem de sequenciamento por síntese, com uma tecnologia de terminação reversível com etiqueta fluorescente [12]. É a tecnologia mais utilizada no mercado NGS.

O processo se inicia com a fragmentação aleatória do DNA. Os adaptadores são ligados a ambas as extremidades de cada fragmento. Em seguida, esses adaptadores são fixados aos respectivos adaptadores complementares, estes últimos já são diversas variantes de adaptadores (complementares) fixados em uma célula de fluxo[22]. Em seguida, cada fragmento ligado à célula de fluxo é amplificado por “amplificação de ponte PCR” que cria várias cópias idênticas de cada sequência; onde o conjunto de sequências feitas a partir de um mesmo fragmento original é chamado de cluster. Cada cluster contém aproximadamente um milhão de cópias do mesmo fragmento original[23]. Por fim, ocorre a determinação de cada nucleotídeo das sequências. A Illumina usa a abordagem de sequenciamento por síntese que emprega terminadores reversíveis, nos quais os quatro nucleotídeos modificados, *primers* e DNA polimerases são adicionados como um mix, e os *primers* são

hibridizados nas sequências. Então, polimerases são usadas para estender os *primers* usando os nucleotídeos modificados. Cada tipo de nucleotídeo é marcado com uma fluorescência diferente, para que cada tipo seja único. Os nucleotídeos possuem um grupo 3'-hidroxila inativo, para garantir que apenas um nucleotídeo seja incorporado por vez. Os clusters são excitados por laser para emitir um sinal de luz específico para cada nucleotídeo, que será detectado por uma câmera de dispositivo de carga acoplada e os programas de computador traduzirão esses sinais em uma sequência de nucleotídeos. O processo continua com a eliminação do terminador com o rótulo fluorescente e o início de um novo ciclo com uma nova incorporação.

As plataformas NGS da Illumina são capazes de sequenciar extremidades pareadas. Nesse tipo de sequenciamento ocorre em ambas as extremidades de um fragmento de DNA, o que gera dados de sequência de alta qualidade com cobertura profunda e alto número de leituras. Por possuir a tecnologia de terminação reversível com etiqueta fluorescente e a tecnologia das extremidades pareadas, a Illumina se torna a plataforma de sequenciamento base por base mais precisa do mercado, com uma taxa de erro de 0,1% (principalmente erros de substituição, muito raramente inserções/exclusões)[19].

No entanto, também existem desvantagens, como toda tecnologia. O tempo de execução relativamente longo, quando comparado às outras tecnologias; é necessário uma alta exigência de controle de carregamento de amostras porque a sobrecarga pode resultar em clusters sobrepostos e baixa qualidade de sequenciamento; e as substituições de nucleotídeos são o tipo de erro mais comum nesta tecnologia, devido à má identificação do nucleotídeo incorporado. A taxa geral de erro desta tecnologia de sequenciamento é de cerca de 1%[24].

3.1.4 Sequenciamento ABI/SOLiD

Diferentemente das outras tecnologias de segunda geração até então descritas, ABI/SOLiD utiliza da abordagem de sequenciamento por ligação, aplica DNA ligase em vez de DNA polimerase para o sequenciamento[13]. São realizadas múltiplas rodadas de sequenciamento. A princípio, os adaptadores são anexados aos fragmentos de DNA, ambos são fixados em esferas e clonados por emulsão de

PCR. Apenas um fragmento de DNA é mantido na superfície da esfera magnética individual. O DNA amplificado capturado por esferas é ancorado em um vidro e inundado com oligonucleotídeos marcados com fluorescência[25]. Se o oligonucleotídeo for complementar ao molde, ele será ligado e então duas bases serão detectadas ao mesmo tempo. O oligonucleotídeo é então clivado.

Após a deposição de esferas em uma lâmina, os fragmentos são sequenciados por ligação, na qual uma sonda codificada com duas bases marcada com fluoróforo, que é composta de nucleotídeos conhecidos na primeira e segunda posições, seguida por degenerados ou bases universais, são adicionadas à biblioteca de DNA. A sonda de duas bases é ligada a uma âncora que é complementar a um adaptador, e a lâmina é fotografada para identificar as duas primeiras bases em cada fragmento. As fitas não estendidas são cobertas por sondas não marcadas ou fosfatase para manter a sincronização do ciclo. Finalmente, as bases degeneradas terminais e o fluoróforo são clivados da sonda, deixando um fragmento estendido de 5 pb. O processo é repetido dez vezes até que duas em cada cinco bases sejam identificadas. Neste ponto, toda a cadeia é reiniciada removendo todas as sondas ligadas e o processo de ligação da sonda, ligação, imagem e clivagem é repetido quatro vezes, cada uma com um n + 1, n + 2, n + 3 ou n + 4 da âncora deslocada[18].

Então, o formato de saída é o espaço de cores, que é a forma codificada do nucleotídeo onde quatro cores fluorescentes são usadas para representar 16 combinações possíveis de duas bases. O sequenciador repete este ciclo de ligação e a cada ciclo a fita complementar é removida e um novo ciclo de sequenciamento começa na posição n-1 do molde. O ciclo é repetido até que cada base seja sequenciada duas vezes. Os dados recuperados do espaço de cores podem ser traduzidos em letras de bases de DNA e a sequência do fragmento de DNA é deduzida[23].

O comprimento das leituras pode chegar a 100 pb. Há alta precisão porque cada base é lida duas vezes, enquanto a desvantagem são as leituras relativamente curtas e os tempos de execução longos. Os erros de sequenciamento nesta tecnologia são devidos ao ruído durante o ciclo de ligação que causa erros na

identificação das bases[26]. Assim como Illumina, o principal tipo de erro é a substituição.

3.1.5 Sequenciamento *Pacific Biosciences* (PacBio)

O sequenciamento *Pacific Biosciences* (PacBio) ou SMRT (*Single Molecule, Real-Time*) é um método de sequenciamento de terceira geração mais amplamente utilizado[27]. A técnica se baseia na mesma marcação fluorescente das demais tecnologias, mas em vez de executar ciclos de amplificação de nucleotídeos, detecta os sinais em tempo real, pois são emitidos quando ocorrem as incorporações. Nanoestruturas chamadas guias de onda de modo zero (ZMWs - *zero-mode waveguides*) são acopladas às células SMRT para a realização do sequenciamento[28]. Os ZMWs exploram as propriedades da luz que passa através de aberturas com diâmetro menor que seu comprimento de onda, de modo que a luz não pode ser propagada. Devido ao seu pequeno diâmetro, a intensidade da luz diminui ao longo dos poços e o fundo dos poços iluminado. Cada ZMW contém uma DNA polimerase ligada à sua parte inferior e o fragmento de DNA alvo para sequenciamento. Durante a reação de sequenciamento, o fragmento de DNA é incorporado pela DNA polimerase com nucleotídeos marcados com fluorescência (com cores diferentes). Sempre que um nucleotídeo é incorporado, ele libera um sinal luminoso que é registrado por sensores[9]. A detecção dos nucleotídeos marcados permite determinar a sequência do DNA.

Comparado aos sequenciadores de segunda geração, PacBio possui algumas vantagens. A rápida preparação das amostras, de 4 a 6 horas no lugar de dias e longos comprimentos de leitura, em média ~ 10 kpbs (leituras individuais muito longas podem chegar a 60 kpb)[29]. No entanto, as plataformas PacBio possuem alta taxa de erro, cerca de 13%, dominada por erros de inserções e exclusões, que se distribuem aleatoriamente pela longa leitura[18].

3.1.6 Sequenciamento por nanoporos - *Oxford Nanopore Technology* (ONT)

ONT é um tipo de sequenciamento que se baseia em nanoporos. Essa tecnologia depende de porinas, proteínas transmembrana que criam um canal

poroso através de uma membrana[30]. ONT consiste em nanoporos biológicos ou nanoporos de estado sólido que são colocados em uma membrana polimérica de alta resistência elétrica. Uma voltagem é mantida através da membrana pela passagem de uma corrente iônica[31]. Quando um analito passa através do poro, a corrente é interrompida e a interrupção é utilizada para identificar a molécula específica. A amostra de DNA a ser analisada é mantida intacta e misturada com cópias de uma enzima processante. O complexo enzima-DNA se aproxima do nanoporo, o DNA de fita simples é puxado através da abertura e a enzima prende a fita de DNA através do nanoporo, uma base de cada vez[32]. A passagem de cada ácido nucleico através do nanoporo cria interrupções características no fluxo da corrente elétrica. O sinal é usado para determinar a sequência[23].

A grande vantagem da técnica é a eliminação da necessidade de nucleotídeos e ligases no preparo das amostras, que se torna muito mais simples e barata. O MinION, primeiro dispositivo lançado com essa técnica, é um sequenciador ONT móvel de molécula única que mede dez centímetros de comprimento e é conectado por uma porta USB 3.0 de um laptop[33]. Os dados são carregados na tela antes mesmo que o sequenciamento seja concluído. Então a praticidade é uma vantagem significativa também. O maior problema é a alta taxa de erro de ~ 12%. Cerca de 3% de incompatibilidades, ~4% de inserções e ~5% de exclusões[34].

ONT tem se dedicado a desenvolver novos sequenciadores baseado nessa técnica com mais células de fluxo e mais nanoporos para processar mais amostras e tentar solucionar o problema da velocidade de translocação do DNA pelo nanoporo.

4. BIOINFORMÁTICA E ANÁLISE DE DADOS

Com o avanço das tecnologias NGS, a Bioinformática se tornou ainda mais imprescindível. O alto volume de dados gerados pelos sequenciadores torna os *pipelines* de análise de dados fundamentais para utilização destas tecnologias. As ferramentas computacionais são essenciais para um bom gerenciamento, análise e interpretação do grande volume de dados de NGS gerados a todo instante.

Cada plataforma de NGS produz dados diferentes, cada um com suas especificidades. E há grande diferença entre as plataformas de NGS de segunda e terceira geração, devido ao tamanho das leituras. No entanto, o fluxo de trabalho é dividido em três etapas principais: análise primária, secundária e terciária[35]. Neste parágrafo, descrevo procedimentos realizados, de forma geral, quanto aos dados gerados por DNA-seq. A análise primária consiste na identificação das bases para cada fragmento de DNA (geralmente ocorre no software do próprio sequenciador) e avaliação da qualidade, filtragem e retirada dos adaptadores, também conhecido como "trimagem". A análise secundária inclui o alinhamento/mapeamento das leituras a um genoma de referência, ou montagem de novo, e a chamada de variantes. E, por fim, a análise terciária envolve a identificação, anotação e interpretação das variantes. É nessa terceira etapa que há associação dos dados gerados com o fenótipo das amostras.

A ideia do fluxo de trabalho dividido em três etapas também ocorre em análise de dados em RNA-seq[10]. Embora a análise possa ser mais desafiadora, devido às próprias características do transcritoma, como *splicing* alternativo e a limitação dos dados à expressão gênica do momento celular em que o sequenciamento ocorre. A análise primária é basicamente igual ao tratamento dos dados de DNA-seq: controle de qualidade e trimagem. Na análise secundária é realizado o mapeamento das leituras e a reconstrução do transcritoma. E a análise terciária, onde se busca o sentido biológico do dado, é realizada a quantificação da expressão e análise de expressão diferencial (quando se objetiva comparar mais de um grupo).

A maior diferença, em termos de software utilizado, entre as plataformas está na primeira parte da análise primária. Cada plataforma de NGS necessita de ferramentas específicas para identificação das bases que foram sequenciadas. A detecção e análise de dados brutos (análise de sinal), geração de leituras de sequenciamento legíveis (chamada de base) e pontuação da qualidade da base irá depender do software acoplado ao sequenciador que irá variar de acordo com a química utilizada por cada plataforma[36]. Ion Torrent, por exemplo, utiliza o Ion Torrent Suite, um software específico que irá processar o sinal de incorporação de base detectado pelo sensor na parte inferior do chip de sequenciamento. Já a plataforma da Illumina, que possui o sinal emitido por fluorescência, realiza a

chamada de bases a partir de medições de intensidade de sinal fluorescente resultantes dos nucleotídeos incorporados durante cada ciclo. O chamador de base das plataformas da Illumina mais utilizado é o Bustard[37]. O software operacional do ONT é o MinKNOW, que faz aquisição de dados, análise em tempo real e chamada de base local para produzir dados de sinal brutos que podem ser usados para chamada de base[10].

Após a chamada de bases, ainda na análise primária, temos a verificação da qualidade (dados que saíram dos softwares acoplados de cada plataforma) e a retirada dos adaptadores. Os erros na chamada de base são expressos em pontuações de qualidade da chamada de base com base na pontuação *Phred*, uma probabilidade de erro logarítmica[38]. A pontuação de *Phred* é relacionada a probabilidade de uma base estar incorreta. Um *Phred* de 10 (Q10) tem uma precisão de 90,0% (1 chance em 10 de estar incorreta), e quanto a Q30 significa probabilidade de 1 em 1000 de uma base incorreta ou precisão de 99,9%[38]. Os arquivos FASTQ possuem informações da qualidade inerentes a cada base e é a partir deste formato que se inicia o controle de qualidade. O software mais utilizado para avaliação da qualidade é o FastQC[36]. Para trimagem das sequências, tanto as bases com qualidade ruins quanto os adaptadores, o programa mais utilizado para dados advindos da Illumina é o Trimmomatic[39]. Para dados vindos do Ion Torrent, o software mais utilizado para trimagem é o Torrent Suite Software[36]. Tanto para DNA-seq como para RNA-seq, esses programas são utilizados com parâmetros ajustados para cada tipo de experimento.

Na análise secundária, é realizado o alinhamento ou mapeamento das leituras num genoma de referência. Mas pode ocorrer uma montagem de novo, que é uma montagem do zero, a depender do organismo que está sendo estudado. Para dados gerados a partir de sequencias humanas o mapeamento ao genoma de referência é a abordagem mais comum[40]. No entanto, em organismos que ainda não possuem genomas próximos filogeneticamente para serem usados como referência, ou organismos que possuem genomas/transcritomas mal montados, a montagem *de novo* é visada para criação ou correção das imprecisões no genoma de referência.

Uma das partes mais desafiadoras da análise de dados de NGS se encontra nesta etapa. Especialmente de dados de sequenciadores de segunda geração, que

possuem leituras curtas. Quanto menor o fragmento, maior a dificuldade de montar uma sequência consenso que se aproxime do que seriam os cromossomos de um indivíduo. O formato de entrada mais comum para as ferramentas de montagem e alinhamento é o FASTQ. E o formato de saída é o SAM (*Sequence Alignment Map*) ou BAM (*Binary Alignment Map*), a representação binária compactada de SAM. Esses arquivos são riquíssimos em informação, como dados de qualidade, regiões onde as sequencias foram mapeadas, profundidade e amplitude da cobertura, etc[41]. Os métodos mais utilizados para alinhamento para leituras curtas de dados de DNA-seq é o *Burrows-Wheeler Aligners* (BWA)[42] e o Bowtie[43], para dados da Illumina, e Torrent Mapping Alignment Program (TMAP) para *Ion Torrent*[36]. Uma ferramenta útil para a visualização dos dados mapeados a um genoma de referência é o *Interactive Genome Viewer* (IGV)[44]. Para dados de RNA-seq as ferramentas mais comuns para mapeamento são TopHat[45] e STAR[45]. Para fazer a reconstrução dos transcritos *de novo* o Trinity[46] é utilizado.

Ainda na análise secundária de dados de DNA-seq, há a necessidade da chamada de variantes genéticas. TMAP, alinhador do *Ion Torrent*, já possui essa função embutida. No entanto, para dados da Illumina, há a necessidade de alguns passos de processamento pós-alinhamento, como a remoção de leituras duplicadas, realinhamento local intensivo e recalibração do índice de qualidade base[47]. As ferramentas utilizadas para esse processamento são as mesmas da chamada de variantes para dados da Illumina: SAMtools[48] e *Genome Analysis Toolkit* (GATK)[49]. Uma vez processados, os arquivos BAM podem ser submetidos a chamada de variantes, que consiste na identificação de variantes. Os arquivos de entrada são arquivos SAM/BAM e os arquivos de saída são VCF (*Variant Caller Format*), um formato tabular. VCF é um arquivo de texto que contém um cabeçalho, seguido por linhas de dados, cada uma contendo informações sobre a posição cromossômica, a base de referência, a base ou bases alternativas identificadas. SAMtools, GATK e Freebayes são os chamadores de variantes mais comuns para dados gerados pela Illumina[50]. *Ion Torrent* tem o seu próprio chamador de variantes *Torrent Variant Caller* (TVC). As variantes podem ser SNVs (*Single Nucleotide Variant*), e INDELS (*insertions/deletions*) ou SV (*Structure Variants*), sendo que SVs

incluem grandes inserções/duplicações ou exclusões chamadas de CNV (*Copy Number Variation*)[51].

Por fim, na análise terciária dos dados de NGS, é dado o sentido biológico aos nucleotídeos sequenciados. Para dados de DNA-seq, essa análise começa com a anotação das variantes e, logo após, filtragem de variantes, priorização e visualização de dados. Diversas ferramentas estão disponíveis e cada uma dela utiliza de metodologias e bancos de dados diferentes. Essas ferramentas se baseiam em: grau de conservação de resíduos de aminoácidos, homologia de sequência, conservação evolutiva, estrutura proteica ou previsão estatística baseada em mutações conhecidas. Alguns algoritmos comuns são o SIFT[52], PolyPhen-2[53] e CADD[54]. Também existem preditores de efeito de variantes, que utilizam de vários anotadores para predizer o impacto delas, como o *Variant Effect Predictor* (VEP)[55]. Em sequenciamentos de genoma total ou exoma, inúmeras variantes são identificadas. Por isso, se faz necessário a filtragem destas variantes. Alguns filtros para retirada de falsos-positivos podem ser aplicados como a exclusão de variantes com baixa profundidade nas leituras e parâmetros de qualidade. Outro filtro comumente usado para dados de NGS é o MAF (*Minor Allele Frequency*)[56], usado para filtragem baseado na frequência alélica populacional. Por fim, algumas diretrizes padronizadas para a interpretação das variantes genômicas são utilizadas, como, por exemplo, as diretrizes do Colégio Americano de Genética Médica e Genômica (*American College of Medical Genetics* - ACMG)[57].

Para dados de RNA-seq, algumas ferramentas estão disponíveis para quantificação de transcritos e análise de expressão diferencial de genes. Duas ferramentas amplamente utilizadas são os pacotes do R, edgeR[58] e DESeq2[59]. Através destes pacotes, *heatmaps* podem ser gerados para a visualização de genes diferencialmente expressos. Também há a possibilidade de análises mais avançadas de interpretações funcionais como ontologia, através do *Gene Ontology* ou análise de enriquecimento de vias[60].

É válido ressaltar que é necessário um estudo detalhado dos programas utilizados em análises de Bioinformática a fim de escolher as melhores ferramentas e ajustar os parâmetros para responder pergunta biológica em questão da pesquisa.

Há necessidade do conhecimento da técnica utilizada no laboratório para escolha mais coerente dos softwares alinhados com o objetivo do desenho experimental.

5. APLICAÇÕES EM PESQUISA BIOLÓGICA E MÉDICA

O acesso à informação genética de um indivíduo pode revelar muito sobre suas condições fisiológicas. Visto que toda a expressão gênica tem sua informação armazenada em seu DNA. Os dados de NGS podem ajudar na compreensão de diversos mecanismos, desde a da biologia de patógenos, a virulência e o fenômeno da resistência bacteriana, até aplicação clínica de em doenças hereditárias, oncologia e doenças infecciosas em humanos. Alternativamente, dados de NGS podem auxiliar pesquisas voltadas para estudo do solo e genética de plantas para melhoramento.

5.1 Sequenciamento do genoma completo

São muitas as abordagens para a busca de respostas de problemas biológicos. A começar pelo sequenciamento do genoma completo (*Whole-Genome Sequencing* - WGS). Com a tecnologia de NGS é possível sequenciar genomas inteiros, inclusive os mais de 3 bilhões de pares de bases do genoma humano. Nesse tipo de sequenciamento há muita informação do organismo como um todo, onde se pode capturar variantes pequenas e grandes. Genomas inteiros são sequenciados mais comumente pela plataforma Illumina, principalmente quando já há uma referência para mapear os genes. E se utiliza a tecnologia SMRT para montagem de novo, quando há necessidade de grandes coberturas para tornar o desafio computacional de montagem menor[23].

5.2 Exoma Completo

Um tipo de sequenciamento muito comum é o do exoma completo (*Whole-Exome Sequencing* - WES). A parte de maior interesse clínico e comercial de um genoma é a parte codificadora. E apenas 1% do genoma codifica proteínas[6]. Grande parte das mutações patogênicas ou significativas ocorrem nas proteínas, onde elas possuem função alterada e desencadeiam processos fisiológicos

negativos. Sendo assim, o sequenciamento do exoma é mais funcional para responder esse tipo de pergunta e, adicionalmente, ainda é mais barato por sequenciar uma parte menor do genoma[61].

5.3 Sequenciamento de Alvo

Outro tipo comum de sequenciamento é o direcionado a um alvo (*Targeted Sequencing*). O sequenciamento direcionado pode produzir uma cobertura muito maior de regiões genômicas de interesse, ao mesmo tempo que reduz o custo e o tempo do sequenciamento[62]. Essa estratégia é muito visada quando os genes a serem explorados já são conhecidos. Assim, o sequenciamento é direcionado de genes ou regiões genômicas específicas[61]. Existem painéis disponíveis para fins de diagnóstico clínico, o que permite um diagnóstico rápido de muitas doenças. Como, por exemplo, painéis de mutações *hotspot* em câncer, onde se pode avaliar a carga mutacional ou mesmo painéis farmacogenômicos, usados para detectar a eficácia de medicamentos com segurança.

5.4 DNA mitocondrial

Temos também o sequenciamento de DNA mitocondrial. Muito utilizado na prática forense[63], em pesquisas de ancestralidade[64] e evolução[65]. Essa técnica se assemelha com o sequenciamento direcionado a um alvo e ao WES pois há uma redução no tamanho (e custo) do sequenciamento, quando comparado ao genoma completo. Outra razão para estudar o DNA mitocondrial (mtDNA) é a exploração da via de fosforilação oxidativa em diversas doenças, como o câncer.

5.5 Transcritoma

Outro sequenciamento amplamente utilizado é o RNA-seq (sequenciamento de transcritos). Inicialmente, o RNA é convertido em cDNA (DNA complementar) para que seja possível realizar o sequenciamento. O RNA-seq aplica tecnologias NGS para traçar o perfil qualitativo e quantitativo do conjunto completo de transcritos, incluindo mRNAs, pequenos RNAs e outros RNAs não codificantes[62]. Esse tipo de sequenciamento é usado para descobrir a sequência codificadora de genes expressos e estudar o nível de expressão gênica. O perfil do transcritoma

fornece uma informação instantânea dos padrões de expressão gênica e elementos reguladores em uma célula, tecido ou organismo sob diferentes estados fisiológicos[62]. Como se fosse uma fotografia do momento celular que aquele material sequenciado se encontrava. E isso é extremamente relevante para a compreensão dos processos biológicos no desenvolvimento e na doença. Essa abordagem é bastante utilizada para comparação de indivíduos saudáveis e doentes. No entanto, a técnica extrapola a prática clínica. Outras condições podem ser avaliadas. A expressão de genes na condição de exposição ou não a um parasita por duas populações de organismos hospedeiros, por exemplo, pode ser comparada com esse tipo de estratégia experimental. Esse tipo de abordagem é essencial para avaliar as características que diferem um grupo de outro, baseado na expressão gênica daquele momento celular. Também se pode comparar um mesmo indivíduo em diferentes estágios de desenvolvimento a fim de se conhecer os genes envolvidos em cada etapa do crescimento.

5.6 Metagenômica

A metagenômica é um campo da genômica que se concentra no estudo direto do material genético coletado de ambientes complexos, como amostras de solo, água, trato gastrointestinal humano ou ecossistemas marinhos. Ao contrário da genômica tradicional, que se concentra em organismos individuais, a metagenômica analisa comunidades microbianas inteiras presentes em um determinado ambiente. É o estudo da diversidade, taxonomia e do potencial funcional de uma comunidade microbiana coexistindo em um ambiente[67]. Além de identificar organismos, a metagenômica pode fornecer informações sobre as funções genômicas presentes na comunidade, incluindo genes envolvidos em processos metabólicos, produção de metabólitos e resistência a antibióticos. Essa abordagem é aplicada em diversos campos, como microbiologia ambiental, estudo de microbiomas humanos, monitoramento de ecossistemas, descoberta de novos genes e enzimas, bem como em aplicações industriais e médicas.

6. DISCUSSÃO E CONSIDERAÇÕES FINAIS

Todas estas estratégias vêm sendo utilizadas para melhoria da saúde humana, direta ou indiretamente. Seja na compreensão do próprio genoma humano, na busca de respostas genéticas, ou no sequenciamento de microrganismos que acometem nossa saúde, animais que possam ser hospedeiros destes microrganismos, plantas que nos alimentamos, dentre diversos outros objetivos. A Tabela 1 contém informações resumidas sobre prós e contras de cada plataforma, bem como suas aplicações.

Tabela 1 – Resumo das plataformas

Plataforma	Prós	Contras	Tamanho das Leituras
Illumina	- Alta precisão e confiabilidade - Custo relativamente baixo por base - Maior profundidade de cobertura - Variedade de aplicações	- Sequências curtas - Erros em regiões ricas em GC - Dificuldades em sequenciar regiões heterozigotas	Curto: 150-300 bp
454/Roche	- Sequências longas - Bons para regiões repetitivas - Sem viés de GC	- Custo mais elevado por base - Produção de dados mais limitada - Menor precisão em regiões homopoliméricas - Descontinuada	Médio: 700-800 bp
Ion Torrent	- Sem necessidade de fluoróforos - Rápido e baixo custo por base - Boa resolução para variantes	- Erros em homopolímeros - Menor profundidade de cobertura - Limitado em regiões repetitivas	Curto: 200-400 bp
ABI SOLiD	- Alta precisão de base - Boa resolução em regiões repetitivas	- Custo mais elevado por base - Leituras curtas	Curto: até 75 bp
PacBio	- Sequências longas - Baixo custo para longos fragmentos	- Maior custo por base - Menor precisão de leitura - Erros sistemáticos	Longo: >10 kb
ONT	-Sequenciamento em tempo real - Sequências longas - Baixo custo para longos fragmentos	- Menor precisão em comparação com outras plataformas	Longo: >10 kb

Para a seleção da plataforma de sequenciamento mais adequada a um experimento, é crucial considerar diversas questões. O custo e a infraestrutura, muitas vezes, desempenham um papel decisivo nessa escolha. No entanto, é fundamental destacar que a pergunta biológica em foco deve ser determinante na escolha da plataforma, dependendo do tipo específico de aplicação.

REFERÊNCIAS

1. CRICK F. Central Dogma of Molecular Biology. Nature. 1970 Aug;227(5258):561–3.
2. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. Proceedings of the National Academy of Sciences. 1977 Dec;74(12):5463–7.
3. Maxam AM, Gilbert W. A new method for sequencing DNA. Proc Natl Acad Sci U S A. 1977 Feb;74(2):560–4.
4. Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, et al. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. Science. 1995 Jul 28;269(5223):496–512.
5. Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, et al. Life with 6000 genes. Science. 1996 Oct 25;274(5287):546, 563–7.
6. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, et al. The sequence of the human genome. Science. 2001 Feb 16;291(5507):1304–51.
7. Gullapalli RR, Desai K V, Santana-Santos L, Kant JA, Becich MJ. Next generation sequencing in clinical medicine: Challenges and lessons for pathology and biomedical informatics. J Pathol Inform. 2012;3:40.
8. Wheeler DA, Srinivasan M, Egholm M, Shen Y, Chen L, McGuire A, et al. The complete genome of an individual by massively parallel DNA sequencing. Nature. 2008 Apr 17;452(7189):872–6.
9. Gupta N, Verma VK. Next-Generation Sequencing and Its Application: Empowering in Public Health Beyond Reality. In 2019. p. 313–41.
10. Hu T, Chitnis N, Monos D, Dinh A. Next-generation sequencing technologies: An overview. Hum Immunol. 2021 Nov;82(11):801–11.
11. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, et al. Genome sequencing in microfabricated high-density picolitre reactors. Nature. 2005 Sep 15;437(7057):376–80.
12. Bentley DR. Whole-genome re-sequencing. Curr Opin Genet Dev. 2006 Dec;16(6):545–52.
13. Shendure J, Porreca GJ, Reppas NB, Lin X, McCutcheon JP, Rosenbaum AM, et al. Accurate Multiplex Polony Sequencing of an Evolved Bacterial Genome. Science (1979). 2005 Sep 9;309(5741):1728–32.
14. Rothberg JM, Hinz W, Rearick TM, Schultz J, Mileski W, Davey M, et al. An integrated semiconductor device enabling non-optical genome sequencing. Nature. 2011 Jul 20;475(7356):348–52.
15. Mardis ER. Next-Generation DNA Sequencing Methods. Annu Rev Genomics Hum Genet. 2008 Sep 1;9(1):387–402.
16. Liu L, Li Y, Li S, Hu N, He Y, Pong R, et al. Comparison of Next-Generation Sequencing Systems. J Biomed Biotechnol. 2012;2012:1–11.

17. Huse SM, Huber JA, Morrison HG, Sogin ML, Welch D. Accuracy and quality of massively parallel DNA pyrosequencing. Genome Biol. 2007;8(7):R143.

18. Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. Nat Rev Genet. 2016 Jun 17;17(6):333–51.

19. Hu T, Chitnis N, Monos D, Dinh A. Next-generation sequencing technologies: An overview. Hum Immunol. 2021 Nov;82(11):801–11.

20. Loman NJ, Misra R V, Dallman TJ, Constantinidou C, Gharbia SE, Wain J, et al. Performance comparison of benchtop high-throughput sequencing platforms. Nat Biotechnol. 2012 May 22;30(5):434–9.

21. Merriman B, R&D Team IT, Rothberg JM. Progress in Ion Torrent semiconductor chip based sequencing. Electrophoresis. 2012 Dec 4;33(23):3397–417.

22. Mardis ER. Next-Generation DNA Sequencing Methods. Annu Rev Genomics Hum Genet. 2008 Sep 1;9(1):387–402.

23. Shetty PJ, Amirtharaj F, Shaik NA. Introduction to Nucleic Acid Sequencing. In: Essentials of Bioinformatics, Volume I. Cham: Springer International Publishing; 2019. p. 97–126.

24. Dohm JC, Lottaz C, Borodina T, Himmelbauer H. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. Nucleic Acids Res. 2008 Sep 1;36(16).

25. Voelkerding K V, Dames SA, Durtschi JD. Next-Generation Sequencing: From Basic Research to Diagnostics. Clin Chem. 2009 Apr 1;55(4):641–58.

26. Alic AS, Ruzafa D, Dopazo J, Blanquer I. Objective review of *de novo* stand-alone error correction methods for <scp>NGS</scp> data. WIREs Computational Molecular Science. 2016 Mar 11;6(2):111–46.

27. Schadt EE, Turner S, Kasarskis A. A window into third-generation sequencing. Hum Mol Genet. 2010 Oct 15;19(R2):R227–40.

28. Rhoads A, Au KF. PacBio Sequencing and Its Applications. Genomics Proteomics Bioinformatics. 2015 Oct;13(5):278–89.

29. Ari Ş, Arikan M. Next-Generation Sequencing: Advantages, Disadvantages, and Future. In: Plant Omics: Trends and Applications. Cham: Springer International Publishing; 2016. p. 109–35.

30. Clarke J, Wu HC, Jayasinghe L, Patel A, Reid S, Bayley H. Continuous base identification for single-molecule nanopore DNA sequencing. Nat Nanotechnol. 2009 Apr 22;4(4):265–70.

31. Jain M, Koren S, Miga KH, Quick J, Rand AC, Sasani TA, et al. Nanopore sequencing and assembly of a human genome with ultra-long reads. Nat Biotechnol. 2018 Apr 29;36(4):338–45.

32. Feng Y, Zhang Y, Ying C, Wang D, Du C. Nanopore-based Fourth-generation DNA Sequencing Technology. Genomics Proteomics Bioinformatics. 2015 Feb;13(1):4–16.

33. Mikheyev AS, Tin MMY. A first look at the Oxford Nanopore MinION sequencer. Mol Ecol Resour. 2014 Nov 24;14(6):1097–102.

34. Branton D, Deamer DW, Marziali A, Bayley H, Benner SA, Butler T, et al. The potential and challenges of nanopore sequencing. Nat Biotechnol. 2008 Oct 9;26(10):1146–53.

35. Gargis AS, Kalman L, Bick DP, da Silva C, Dimmock DP, Funke BH, et al. Good laboratory practice for clinical next-generation sequencing informatics pipelines. Nat Biotechnol. 2015 Jul 8;33(7):689–93.

36. Pereira R, Oliveira J, Sousa M. Bioinformatics and Computational Tools for Next-Generation Sequencing Analysis in Clinical Genetics. J Clin Med. 2020 Jan 3;9(1):132.

37. Kao WC, Stevens K, Song YS. BayesCall: A model-based base-calling algorithm for high-throughput short-read sequencing. Genome Res. 2009 Oct;19(10):1884–95.

38. Ewing B, Hillier L, Wendl MC, Green P. Base-Calling of Automated Sequencer Traces Using *Phred.* I. Accuracy Assessment. Genome Res. 1998 Mar 1;8(3):175–85.

39. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014 Aug 1;30(15):2114–20.

40. Flicek P, Birney E. Sense from sequence reads: methods for alignment and assembly. Nat Methods. 2009 Nov 15;6(S11):S6–12.

41. Sims D, Sudbery I, Ilott NE, Heger A, Ponting CP. Sequencing depth and coverage: key considerations in genomic analyses. Nat Rev Genet. 2014 Feb 17;15(2):121–32.

42. Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics. 2009 Jul 15;25(14):1754–60.

43. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 2009;10(3):R25.

44. Thorvaldsdottir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinform. 2013 Mar 1;14(2):178–92.

45. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013 Jan 1;29(1):15–21.

46. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol. 2011 Jul 15;29(7):644–52.

47. Robinson PN, Piro RM, Jäger M. Computational Exome and Genome Analysis. Chapman and Hall/CRC; 2017.

48. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009 Aug 15;25(16):2078–9.

49. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010 Sep;20(9):1297–303.

50. Hwang S, Kim E, Lee I, Marcotte EM. Systematic comparison of variant calling pipelines using gold standard personal exome variants. Sci Rep. 2015 Dec 7;5(1):17875.

51. Haraksingh RR, Snyder MP. Impacts of Variation in the Human Genome on Gene Regulation. J Mol Biol. 2013 Nov;425(21):3970–7.

52. Ng PC. SIFT: predicting amino acid changes that affect protein function. Nucleic Acids Res. 2003 Jul 1;31(13):3812–4.

53. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, et al. A method and server for predicting damaging missense mutations. Nat Methods. 2010 Apr;7(4):248–9.

54. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet. 2014 Mar 2;46(3):310–5.

55. McLaren W, Pritchard B, Rios D, Chen Y, Flicek P, Cunningham F. Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. Bioinformatics. 2010 Aug 15;26(16):2069–70.

56. The International HapMap 3 Consortium. Integrating common and rare genetic variation in diverse human populations. Nature. 2010 Sep;467(7311):52–8.

57. Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genetics in Medicine. 2015 May;17(5):405–24.

58. Robinson MD, McCarthy DJ, Smyth GK. <tt>edgeR</tt> : a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010 Jan 1;26(1):139–40.

59. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014 Dec 5;15(12):550.

60. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene Ontology: tool for the unification of biology. Nat Genet. 2000 May;25(1):25–9.

61. Grada A, Weinbrecht K. Next-Generation Sequencing: Methodology and Application. Journal of Investigative Dermatology. 2013 Aug;133(8):1–4.

62. Xuan J, Yu Y, Qing T, Guo L, Shi L. Next-generation sequencing in the clinic: Promises and challenges. Cancer Lett. 2013 Nov;340(2):284–95.

63. Kanthaswamy S. Review: Wildlife forensic genetics—Biological evidence, <scp>DNA</scp> markers, analytical approaches, and challenges. Anim Genet. 2023 Dec 20;

64. Pan Z, Xu S. Population genomics of East Asian ethnic groups. Hereditas. 2020 Dec 8;157(1):49.

65. Hammar F, Miller DL. Genetic Diversity in the mtDNA of Physarum polycephalum. Genes (Basel). 2023 Mar 2;14(3):628.

66. Pasquinelli AE. MicroRNAs and their targets: recognition, regulation and an emerging reciprocal relationship. Nat Rev Genet. 2012 Apr 13;13(4):271–82.

67. Fadiji AE, Babalola OO. Metagenomics methods for the study of plant-associated microbial communities: A review. J Microbiol Methods. 2020 Mar;170:105860.

CAPÍTULO II

(Artigo publicado na revista *Gene*)

IDENTIFICATION AND CHARACTERIZATION OF MICRORNAS IN *BIOMPHALARIA TENAGOPHILA* AND COMPARATIVE ANALYSIS OF THEIR EXPRESSION IN *SCHISTOSOMA MANSONI*-RESISTANT AND -SUSCEPTIBLE SNAIL POPULATIONS

Gene 884 (2023) 147742

Contents lists available at ScienceDirect

Gene

journal homepage: www.elsevier.com/locate/gene

Identification and characterization of microRNAs in *Biomphalaria tenagophila* and comparative analysis of their expression in *Schistosoma mansoni*-resistant and -susceptible snail populations

Tamires Caixeta Alves [a,1], Fábio Ribeiro Queiroz [b,1], Angelo Borges de Melo Neto [a], Gabriel da Rocha Fernandes [b], Fabiano Sviatopolk-Mirsky Pais [b], Wander de Jesus Jeremias [c], Elio Hideo Babá [b], Marina de Moraes Mourão [b], Enyara Rezende Morais [a], Fernanda Janku Cabral [d], Laurence Rodrigues do Amaral [a], Roberta Lima Caldeira [b], Paulo Marcos Zech Coelho [b], Matheus de Souza Gomes [a,*]

[a] *Bioinformatics and Molecular Analysis Laboratory, Federal University of Uberlândia, Patos de Minas, MG, Brazil*
[b] *René Rachou Institute, Oswaldo Cruz Foundation, Belo Horizonte, Minas Gerais, Brazil*
[c] *Department of Pharmacy, Federal University of Ouro Preto, Ouro Preto, MG, Brazil*
[d] *State University of Campinas, Campinas, SP, Brazil*

ARTICLE INFO

Edited by Isabel Mauricio

Keywords:
Schistosomiasis
miRNAs
Differential expression
Biomphalaria

ABSTRACT

Background: Schistosomiasis is a neglected tropical disease caused by *Schistosoma* and affects over 240 million people worldwide. One of the most prominent causative agents is *Schistosoma mansoni*, which develops inside the intermediate host. *Biomphalaria tenagophila* is the second most important vector of schistosomiasis in Brazil and the Taim population is completely resistant to infection by *S. mansoni*.
Objective: This study aims to identify and characterize *B. tenagophila* microRNAs (miRNAs) and evaluate their differential expression in *S. mansoni*-susceptible and -resistant populations of *B. tenagophila*.
Methods: Two populations of *B. tenagophila* snails, susceptible and resistant to *S. mansoni* infection, were used to investigate the small RNA response of these snails after being infected with the parasite. Small RNA sequencing and quantitative real-time PCR were employed to identify and validate differentially expressed miRNAs. Bioinformatics analysis were performed to identify miRNA precursors and mature and evaluate their differential expression.
Findings: The study predicted 173 mature miRNAs and 123 precursors. Among them were six Lophotrochozoa-specific miRNAs, three mollusk-specific miRNAs, and six pre-miRNAs in a cluster. The small RNA sequencing and RT-PCR of *B. tenagophila* samples allowed assessing the expression patterns of miRNAs.
Main conclusions: The results obtained may support future studies in *Biomphalaria* spp., generating a global impact on disease control.

1. Introduction

Schistosomiasis is a neglected disease that affects 240 million people, and more than 700 million people live in endemic areas, with most of these individuals living in low and middle-income countries in the world, mostly countries with poor sanitary conditions (WHO, 2022).

Abbreviations: A, adenine content; AMFE, Adjusted Minimum Free Energy; AU, adenine and uracil content; AU ratio, ratio of adenine and uracil; C, cytosine content; G, guanine content; GC, guanine and cytosine content; GC ratio, ratio of guanine and cytosine; HM, Herivelto Martins; ITS region, internal transcribed spacer region; MFEE, Minimum Free Energy of the Thermodynamic Set; MFEI, Minimum Free Energy; miRNA, microRNA; mRNA, Messenger RNA; PCR, Polymerase Chain Reaction; QLF test, quasi-likelihood F-test; RNA-seq, RNA sequencing; RT-qPCR, Real-time quantitative Polymerase Chain Reaction; U, uracil content.

* Corresponding author at: Institute of Biotechnology (IBTEC), Universidade Federal de Uberlândia (UFU), Av. Getúlio Vargas, 230, 37800-128 Patos de Minas, MG, Brazil.
E-mail address: matheusgomes@ufu.br (M. de Souza Gomes).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.gene.2023.147742
Received 10 May 2023; Received in revised form 22 August 2023; Accepted 24 August 2023
Available online 25 August 2023
0378-1119/© 2023 Elsevier B.V. All rights reserved.

Schistosoma infection is one of the most important causes of non-cirrhotic portal hypertension in Latin America, Africa, and Asia (Bodh and Chawla, 2014). The World Health Organization proposed the achievement of schistosomiasis elimination as a public health problem in endemic countries by 2030 (Spangenberg, 2021). Due to the complex scenario, that target will only be accomplished if different strategies are adopted to eradicate this public health problem, including sanitation, measures for intermediate host control, effective treatment, and sensitive diagnosis.

Currently, the most employed disease control relies on the chemotherapeutic use of praziquantel, which, disregarding the high cure rates, is not efficient against immature forms (schistosomula) and, therefore, outbreaks and reinfections persist (Spangenberg, 2021). Thus, the search for new strategies to control transmission is necessary.

Schistosomiasis is caused by parasites of the genus *Schistosoma*, organisms that rely on a complex life cycle, involving the development within an intermediate host for the transformation of miracidia into cercariae, the human infectant larvae. The intermediate hosts of *Schistosoma mansoni* (Plathyhelminthes:Trematoda) are freshwater mollusks from the genus *Biomphalaria* (Mollusca:Planorbidae). Thus, the control of the intermediate host would be one of the combined efforts required to interrupt the parasite life cycle, however, to date, the only measure available for intermediate control is the use of an environmental hazardous molluscicide, Niclosamine (Coelho and Caldeira, 2016).

B. tenagophila is the second most important species in the transmission of schistosomiasis in Brazil and, unlike several populations *B. glabrata*, which is highly susceptible to infection by *S. mansoni*, some populations of *B. tenagophila* are resistant to the parasite infection (Coelho et al., 2004). In 1963 were reported a wide variation in the level of susceptibility to *S. mansoni* in different geographic populations of *Biomphalaria* spp. from several locations in Brazil (Paraense and Corrêa, 1978). Among these population, *Biomphalaria* spp., a variation in the susceptibility profile of the mollusk was also noticed (Coelho et al., 2004), some populations were highly susceptible to the parasite and another totally resistant to *S. mansoni* infection, *B. tenagophila* from the Taim population (dos Santos et al., 1979). This population of *B. tenagophila* was identified in Rio Grande do Sul – Brazil and is of great interest given the characteristic of complete natural resistance to *S. mansoni*. Even with the exposure of this population to high loads of parasites, *S. mansoni* cannot develop (de Jesus Pereira et al., 2008) and do not mature into the cercariae form (Bezerra et al., 1997). Also, this characteristic has been shown to be a dominant genetic trait, highlighting the importance of this specimen to the understanding of intermediate host resistance (Rosa et al., 2005).

The investigation of small RNAs is also relevant, since small RNAs are critical for gene expression control and there is a gap of knowledge in *Biomphalaria tenagophila*. Micro-RNAs (miRNAs) are the main representatives of this class, and they regulate gene expression at the post-transcriptional level through the degradation or translation repression of their targets, the messenger RNAs (mRNAs) (Bartel, 2004).

There is a high degree of specificity in the mollusk-trematode interaction, which will result in foci maintenance (Basch, 1976). Therefore, a better understanding of the mechanisms involved in the *Biomphalaria-S. mansoni* interaction will enable to create new control strategies that block or interfere in the development of *S. mansoni* and, thus, reduce the transmission of the disease. Here, we describe 173 mature miRNAs and 123 precursors, mollusk-specific miRNAs.

2. Materials and methods

2.1. Mollusks

Two populations of *B. tenagophila* were used in this study: 1) Population resistant to infection by *S. mansoni*: named Taim, originally from Rio Grande do Sul, but maintained by our group on the farm "Sítio Santa Inês", in the county of Bananal/SP (Rosa et al., 2013). The mollusks were sent directly from the farm for use in the laboratory, being kept in the Moluscario Lobato Paraense of the Rene Rachou Institute under ideal conditions of temperature, water quality, and food (Pereira De Souza and Jannotti Passos, 2001); 2) Population susceptible to infection by *S. mansoni*: named Herivelto Martins (HM), collected in the Herivelto Martins stream, county of Bananal/SP and maintained in the laboratory of Parasitology, Microbiology, and Immunology of the Institute of Biological Sciences – Federal University of Juiz de Fora. Part of this population was transferred to the Moluscario Lobato Paraense at Rene Rachou Institute and maintained under the previously described conditions (Pereira De Souza and Jannotti Passos, 2001).

To confirm whether the HM population snails tested were *B. tenagophila*, genomic DNA was extracted from six individuals and the ITS region of the ribosomal DNA was amplified by PCR, followed by amplicons cleavage with the enzyme DdeI and profile evaluation (Tuan and Santos, 2007).

2.2. Infection of mollusks

For the infection assays, *B. tenagophila* mollusks from the Taim and HM populations, sexually mature and with an approximate shell diameter of 7 to 12 mm for the Taim and 4 to 7 mm for the HM, were selected (Rosa et al., 2013). All mollusks in this study come from spawning that hatched on the same date. Snails were separated into two groups for each population, experimental and control. Individual snails of experimental groups were infected with 50 miracidia of *S. mansoni* from SJ strain, provided by Moluscario Lobato Paraense, and exposed for 3 h to artificial light. Four hours after infection, the snails were recovered, washed in dechlorinated water, dried, and instantly frozen in liquid nitrogen, then stored in a freezer at −80 °C. Snails with the same characteristics were selected to be held without infection as the control group and frozen simultaneously in liquid nitrogen. Three biological replicates were created for each group, each with ten mollusks. Groups with 50 exposed mollusks of each population were also kept for 38 days after infection as a control for infection efficiency.

2.3. RNA extraction and purification

Total RNA from snails were isolated and subsequently used for sequencing and for qPCR assays. Whole snails were frozen directly in liquid nitrogen, macerated, and homogenized with Tri Reagent® (Sigma-Aldrich) according to the manufacturer's instructions. The total RNA was treated with DNase using TURBO RNAse-free kit (Ambion®), as recommended by the manufacturer. The quantification of DNA and RNA was performed by using NanoDrop nano-spectrometer (Thermo Fisher Scientific). The quality profile of all samples used for sequencing and gene expression assays were evaluated by Bioanalyzer® (Agilent), according to the manufacturer's instructions. All samples showed the same profile when their electropherogram was analyzed (S1 Figure).

2.4. Small RNA sequencing

For the small RNA sequencing, the total RNA from snails, exposed or not to *S. mansoni*, was used for library construction using NEBNext® Small RNA Library Prep Set for Illumina® (New England Biolabs). The small RNA fraction was purified at 150 nucleotides size according to manufacturer's instructions and sequenced using Illumina HiSeq™ 2500 system (Illumina®). The sequences, from Taim in BioSample SAMN25964963 and from HM in BioSample SAMN25964964, are available in project PRJNA807167 in the SRA database. To detect possible contaminants in the samples, Trace mode of the miRTrace program was used (Kang et al., 2018).

2.5. Real-time quantitative PCR (RT-qPCR)

To validate the miRNAs and transcripts, quantitative Real-Time PCR

(RT-qPCR) TaqMan® system (Life Technologies™) was performed using ViiA 7™ Real-Time PCR System (Applied Biosystems™). The TaqMan® system oligonucleotides were manufactured by Life Technologies™ (S1 Table). The miRNAs expression level of snail samples was measured in three biological replicates from each different population, gene expression levels were normalized using the U6 transcript as an internal control. The relative quantification of the expressed gene was determined by the $2^{-\Delta\Delta Ct}$ method (Livak and Schmittgen, 2001) and statistical analyses among the different groups were performed using the T-test on GraphPad Prism® 5.0 (GraphPad Software, Inc., San Diego, USA). Values of $p \leq 0.05$ were considered statistically significant and are denoted with asterisks in the figures.

2.6. Bioinformatics analysis

2.6.1. Identification of precursors and mature miRNAs

The putative miRNAs were identified from the assembled genome of *B. tenagophila* generated by our group. The raw genome data have been deposited in BioSample SAMN25964961 from the PRJNA807167 project, in the SRA database. To identify the miRNA precursors, the algorithm developed for the analysis of small RNAs in *B. glabrata* based on structural, thermodynamic and homology characteristics was used (Adema et al., 2017). The parameters used to describe the structural and thermodynamic characteristics of the precursors were: Adjusted Minimum Free Energy (AMFE), Minimum Free Energy Index (MFEI), Minimum Free Energy of the Thermodynamic Set (MFEE), size, A (adenine) content, U (uracil) content, C (cytosine) content, G (guanine) content, GC content, AU content, GC ratio, and AU ratio. From the precursors of miRNAs, the mature miRNAs were identified based on data from miRbase (miRBase: the microRNA database https://mirbase.org) by homology. Sequences identified as precursors of miRNA molecules were structurally and thermodynamically characterized using RNAfold software (ViennaRNA Services https://rna.tbi.univie.ac.at/cgi-bin/RNAWebSuite/RNAfold.cgi). This tool predicts the secondary structure of these sequences and indicates the thermodynamic characteristics of each molecule, such as Minimum Free Energy (MFE), diversity, and frequency of sequences (Hofacker, 2003). The alignments between the miRNA precursors and their homologs were performed by RNAalifold (ViennaRNA Services https://rna.tbi.univie.ac.at/cgi-bin/RNAWebSuite/RNAalifold.cgi). The phylogenetic trees were generated using the MEGA program (Kumar et al., 2018). Phylogeny was performed using the Neighbor-Joining Tree test with adjusted parameters of 5000 bootstrap and Kimura 2-parameter for Substitution Model. The precursors and mature identified and characterized were used to feed the miRDeep program (An et al., 2013) and evaluate the presence of mature miRNAs in next-generation sequencing samples from small RNAs.

2.6.2. Differential expression of miRNAs

Small RNAs from the HM and Taim populations were sequenced, under two conditions, exposed or not to *S. mansoni*. For counting the number of times each mature miRNA occurred in each sample was used the miRDeep program (An et al., 2013). Differential expression of miRNAs from the susceptible (HM) and resistant (Taim) *B. tenagophila* dataset were performed using the R program and EdgeR package. Two variables were analyzed: population and parasite exposition. Pair-by-pair tests were performed using exact test (Fisher) and for the complete model with the two variables, it was done with QLF test. The exact test is based on between negative binomial dispersions, which aims at modelling the mean-dispersion relationship across the entire dataset. Thus, it works well for comparing two conditions. The QLF test is based on quasi-likelihood dispersions, which aims at capturing the variability and uncertainty of the dispersion estimate for each gene, thus more suitable for gerating complete models, comparing two conditions and two variables simultaneously. For both tests, the normalization factors were calculated to scale the raw sizes of the library using the calcNormFactors function.

3. Results

3.1. Snail species identification and results of exposures to S. mansoni

According to Tuan and Santos (2007), there is a differentiation in the ITS region of the ribosomal region for the classification of species within the genus *Biomphalaria*. Thus, the six individuals tested showed a consistent pattern (S2 Figure) for mollusks of the species *B. tenagophila*, in the amplification of the ITS region of the ribosomal DNA. This result confirmed that the mollusks of the HM population belong to the species *B. tenagophila*.

The snails were exposed to the SJ strain of *S. mansoni,* which has been described as compatible with *B. tenagophila* in Brazil. After contact with light, all snails of Taim population, as expected, showed no elimination of cercariae, 100 % specimens were refractory to infection. On the other hand, the snails from HM population showed an infection rate of 82 %, thus confirming their status as susceptible to *S. mansoni*.

3.2. Bioinformatics analysis

3.2.1. Identification of precursors and mature miRNAs

Using the previously pipeline developed by our research group (Adema et al., 2017), 123 precursors of miRNAs and 173 mature miRNAs were identified (S2 Table), among which, some miRNA were used for bioinformatics analyzes and validation by Real-Time qPCR.

A total of 76 families of miRNAs were identified. The main homologue was *B. glabrata* (Adema et al., 2017) and the sequences found in common had more than 90 % identity. Many of the miRNAs identified in this study were also identified in the *B. pfeifferi* (Bu et al., 2023) genome. In Fig. 1 we compared the families conserved in mollusks between the species *B. glabrata* and *B. pfeifferi* (data obtained from the respective published papers) and two other gastropods available in miRBase, bivalves (Rosani et al., 2021) compared by Rosani, and the mollusks of the class Cephalopoda available in MirGeneDB.

Some of the identified miRNA families are known in several of the animal kingdom such as MIR-124 (Fig. 2) and MIR-190 (S3 Fig). The MIR-8 family was also characterized and is highly conserved among invertebrates, Mir-8, and vertebrates, Mir-200, Mir-141, and Mir-429, as shown in S4 Fig.

The MIR-71 (S5 fig) is a miRNA from protostome. This miRNA is clustered with MIR-2, as shown in Fig. 3. The miRNAs Mir-745a and Mir-745b were also found in a miRNA cluster (S6 fig).

The MIR-1986 (Fig. 4), MIR-1990 (S7 fig), and MIR-1994 (S8 fig) miRNAs families found in Lophotrochozoa clade.

Finally, three miRNA families were found in Mollusk clade, using the three mollusks (*Haliotis rufescens*, *Lottia gigantea* and *Melibe leonina*) that are available in the miRNAs database miRBase (https://mirbase.org/), five mollusks (*Crassostrea gigas*, *Nautilus pompilius*, *Euprymna scolopes*, *Octopus vulgaris* and *Octopus bimaculoides*) available in database MirGeneDB (https://mirgenedb.org/) and *B. glabrata* (group data), thus being mollusk-specific: MIR-1984 (Fig. 5), MIR-1985 (S9 Fig), and MIR-2722 (S10 Fig).

The secondary structures of the analyzed miRNAs from *B. tenagophila* matched with their *B. glabrata*. The distribution of the clades in the analyzed phylogenetic trees corresponded the distribution of the species in the tree of life. Regarding the alignments, the sequences displayed more than 80 % identity in the most important part of the mature sequences, the seed region (2–8nt), which is the part of the sequence that binds with 100 % complementarity to mRNA targets in animals.

3.2.2. Small-Seq analysis and differential expression of miRNAs

The initial evaluation of the sequencing dataset of small RNAs of *B. tenagophila* showed high quality reads, in which almost all presented a quality scale superior to Phred score 30 (Ewing and Green, 1998). The quality profile of the *Biomphalaria* mollusk samples can be seen in S1Figure.

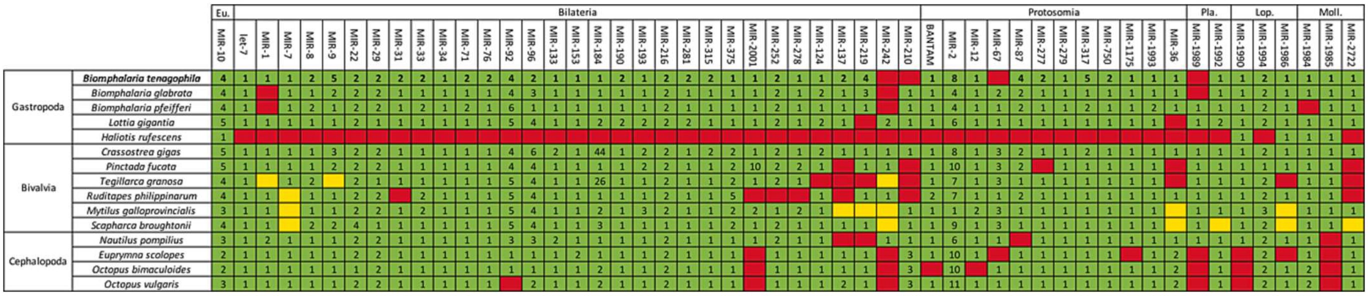


Fig. 1. Conserved miRNA families in mollusks. The comparison was performed in three classes Bivalvia, Gastropoda and Cephalopoda. Presence of miRNA families conserved in Eumetazoa (Eu.), Bilateria, Protosomia, Platytrochozoa (Pla.), Lophotrochozoa (Lop.), and Mollusca (Moll.) is indicated with green color and numbers are the miRNAs present in each family according to the availability of this information in databases and previously published papers. In yellow are the families that have been identified, in other studies, but have not yet been classified. The families of miRNAs not identified in the species were indicated by red color. The highlighted species is *Biomphalaria tenagophila*, the object of study of this work. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

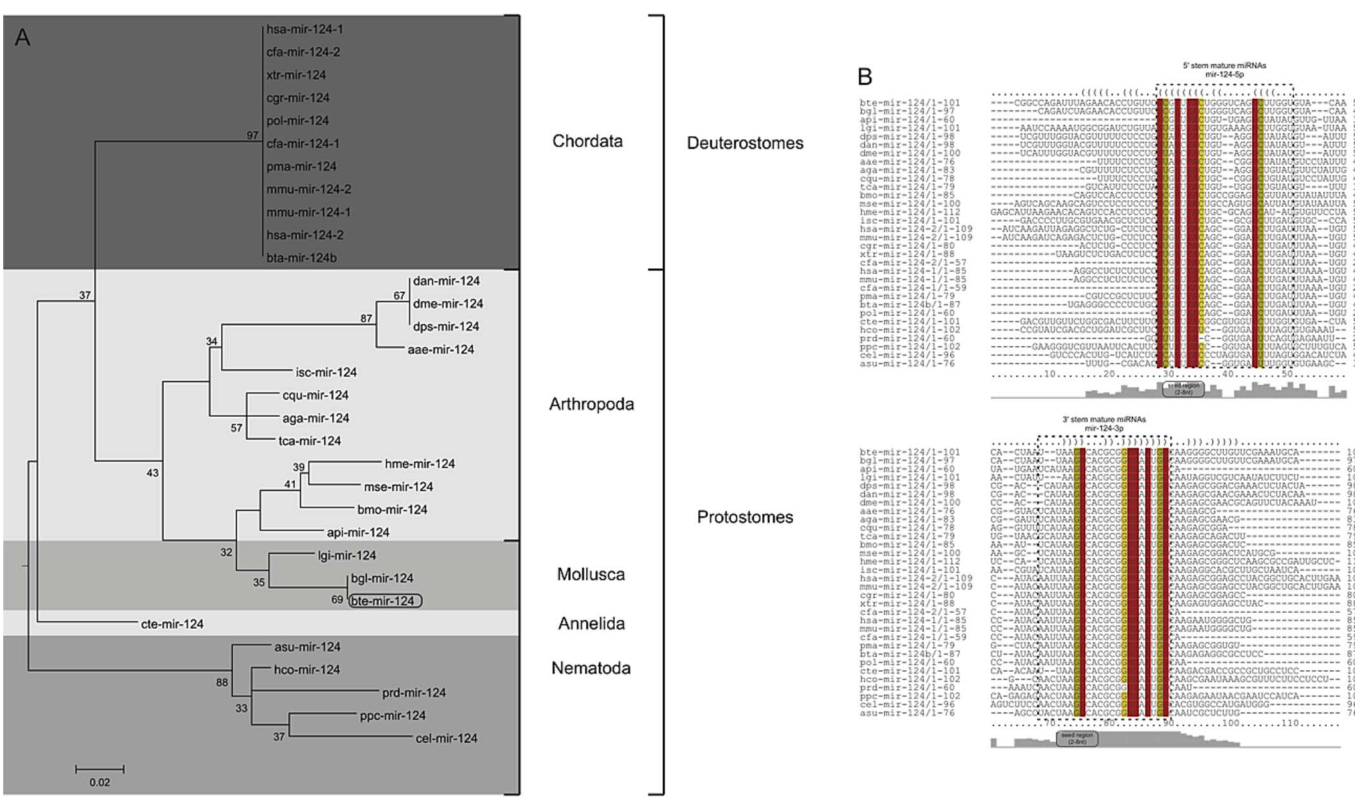


Fig 2. MIR-124 family. (A) Phylogenetic tree. (B) Global alignment among Bte-Mir-124 and its homologs.

After filtering our samples by quality and length, all libraries containing more than 10 million reads were maintained, except for replicate 1 of *B. tenagophila* HM in the unexposed condition with just over 7 million reads (S3 Table). All samples, of both populations and conditions, have enough reads to perform all the characterization and gene expression analysis of small *B. tenagophila* RNAs based on small RNA-seq. We also checked the origin of the sequencing data organisms to confirm whether the contained reads were mostly snail (S11 Fig).

To analyze the differential expression of miRNAs in these samples miRDeep was used to identify the number of times each mature miRNA occurred in each sample compared to the other samples. Thus, tables with the counts of each mature miRNAs for differential expression analysis were generated (S4 Table).

The results showed statistically significant at $p\text{-value} < 0.05$ and log2 fold-change (Table 1).

Differences in expression of five miRNAs were found between *B. tenagophila* HM population exposed to *S. mansoni* compared to exposed Taim population, as shown in Table 1. According to the differential expression analysis performed with the small RNA sequencing data, using the EdgeR package from the R software, Mir-124-3p and Mir-153-3p were less expressed in Taim than in the HM and other miRNAs were more expressed in Taim than in HM (Fig. 6).

Also, there were differences in Mir-124-3p and Bte-Mir-1986-5p expression between *B. tenagophila* Taim unexposed when compared to HM unexposed, with greater expression in *B. tenagophila* HM population than in Taim, as shown in the S4 table. In the last condition, as shown in the last line of Table 1, it is evident that there was no relationship between the population and the infection among the analyzed miRNAs. In other words, the miRNAs evaluated do not confer the characteristics of resistance or not to *S. mansoni*.

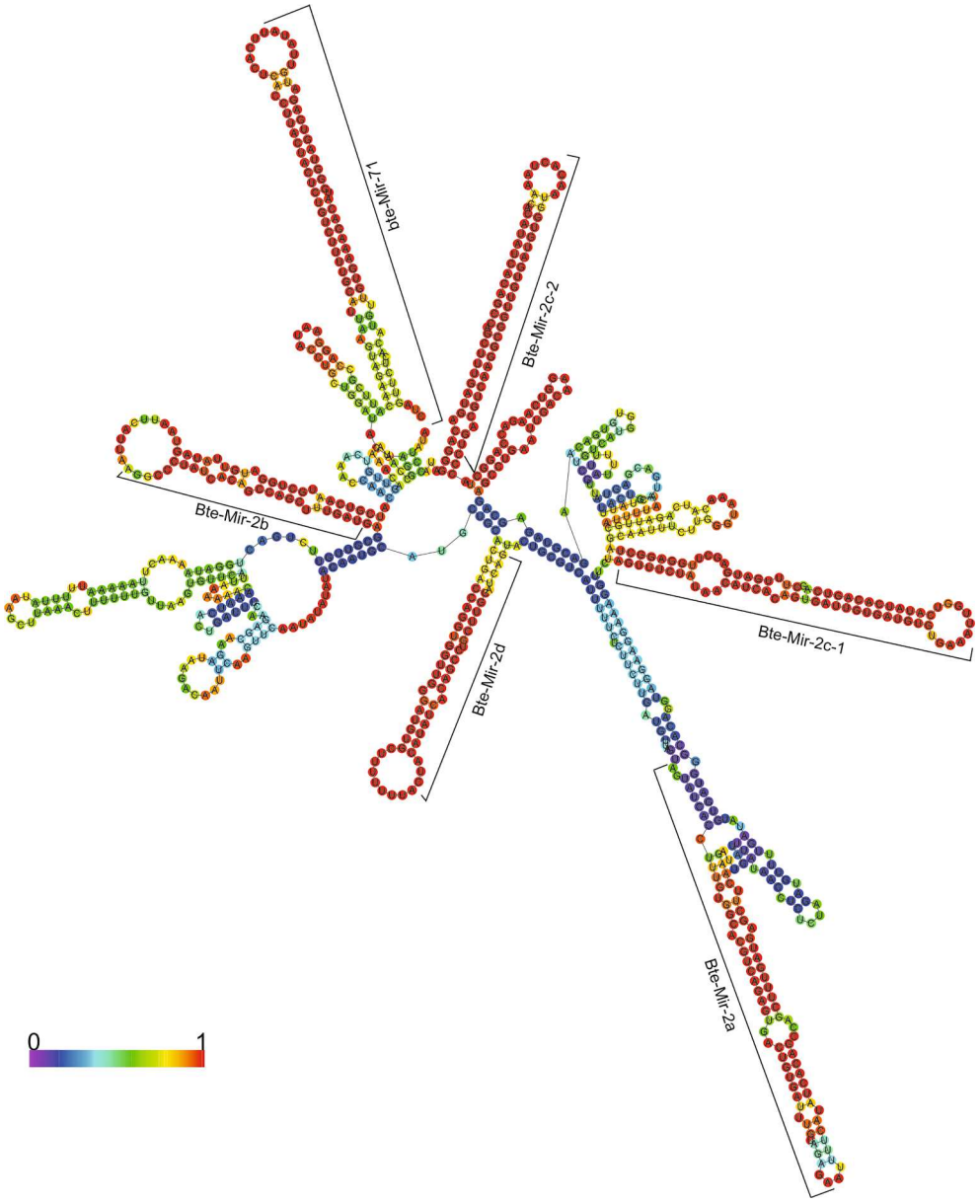


Fig 3. Cluster of MIR-71 and MIR-2.

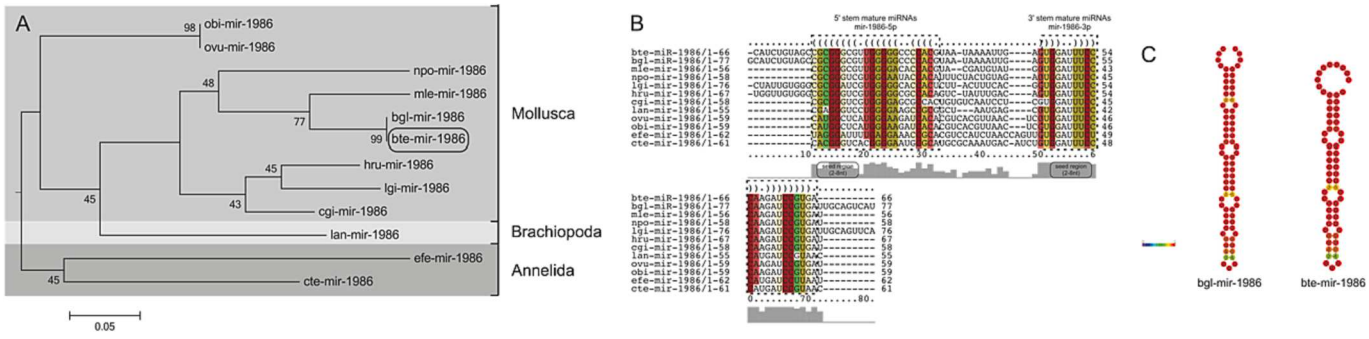


Fig 4. MIR-1986 family. (A) Phylogenetic tree. (B) Global alignment among Bte-Mir-1986 and its homologs. (C) Secondary structures of miRNA precursors.

3.3. Validation by RT-qPCR

Seven miRNAs (Mir-8, Mir-92, Mir-184, Mir-315, Mir-375, Mir-750, and Mir-1984) were chosen for RT-qPCR validation (Fig. 7) based on their high expression levels in the sequencing data and their biological relevance to snails, rather than being differentially expressed. These miRNAs were identified due to their conservation across various clades of the tree of life, highlighting their potential importance in snail

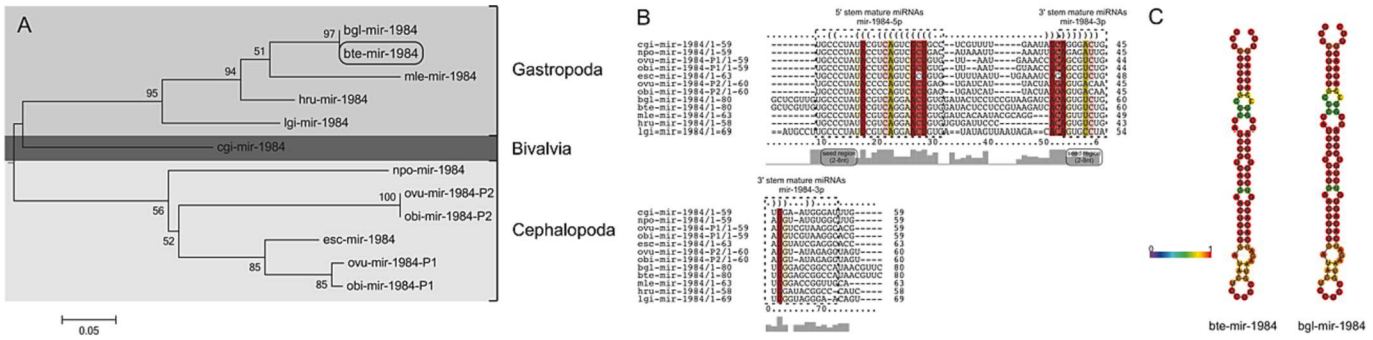


Fig 5. MIR-1984 family. (A) Phylogenetic tree. (B) Global alignment among Bte-Mir-1984 and its homologs. (C) Secondary structures of miRNA precursors.

Table 1
Differentially expressed miRNAs in small RNA samples under different conditions, with p-value of 0.05 and log2 fold-change.

Conditions	DE miRNAs
HM exposed × Taim exposed	Bte-Mir-184-3p, Bte-Mir-375-3p, Bte-Mir-124-3p, Bte-Mir-750–1-3p, Bte-Mir-153-3p
HM exposed × HM unexposed	0
Taim exposed × Taim unexposed	Bte-Mir-278-3p
Taim unexposed × HM unexposed	Bte-Mir-124-3p and Bte-Mir-1986-5p
HM exposed × Taim exposed × Taim unexposed × HM unexposed	0

biology.

The miRNAs Mir-8, Mir-92, Mir-184, Mir-375, Mir-750, and Mir-1984, were downregulated in the Taim and HM populations snails (Fig. 7 A, B, C, E, F and G) relative to the RT-qPCR control (myoglobin). The Mir-315 was upregulated in snails of the Taim population and downregulated in the HM population, relative to the control, as shown in Fig. 7D. Among the miRNAs analyzed, there was a statistically significant difference only in the expression of the miRNAs Mir-92, Mir-184, Mir-315, and Mir-1984.

4. Discussion

The identification of small RNAs is important for a better understanding of the biology of individuals, mainly organisms that participate as intermediate hosts in the infectious disease cycle for humans. In our study, we identified miRNAs and compared their presence from datasets generated from two *Biomphalaria* populations, one susceptible and one resistant, before and after exposure.

4.1. Identification of precursors and mature miRNAs from the *B. tenagophila* genome

To assess the influence of miRNAs in the gene expression regulation in different *Biomphalaria*, we sought to investigate the differential expression of miRNAs in extremely different populations regarding susceptibility phenotypes to *Schistosoma* parasite. The miRNAs are small non-coding RNAs (~22 nt) that play an important role in gene regulation at the post-transcriptional level. The mature form of miRNAs binds to the target mRNA and prevents its translation, thus silencing the gene that would be translated (Ambros, 2004). The miRNAs have already been identified in several species (Navarro and Baulcombe, 2019; Liu and Zhang, 2012; Marco et al., 2013; Wheeler et al., 2009). *B. tenagophila* precursor and mature miRNAs (123 miRNA precursors and 173 matures) were broadly comparable with the number of miRNAs found in other Mollusk species such as and *Melibe leonina* (90 precursors and 176 mature) (Fromm et al., 2015) and *B. glabrata* nnn (Adema et al., 2017).

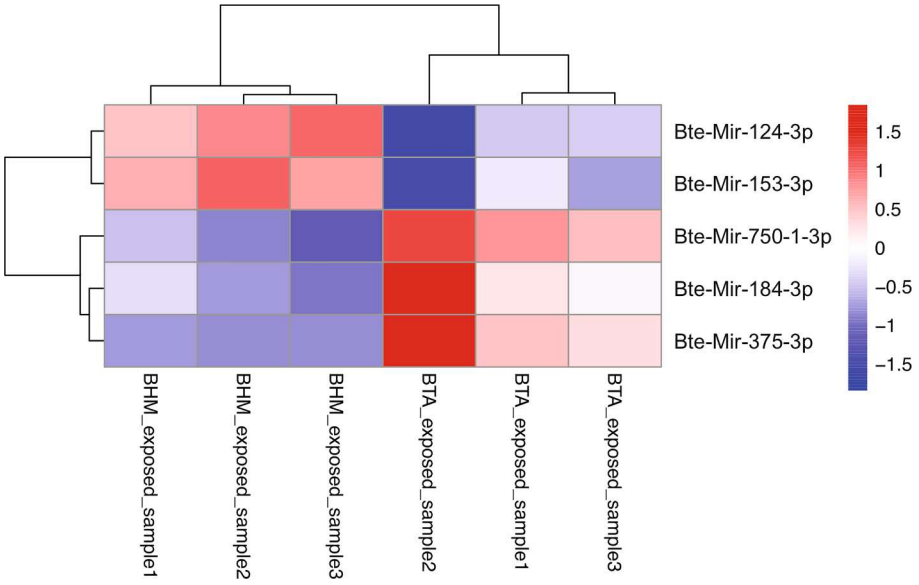


Fig 6. Heatmap generated from the result of EdgeR for the *B. tenagophila* HM population exposed × Taim population exposed condition. The heatmap was calculated using the pheatmap library. In the color gradient, red indicated a higher miRNA expression level and blue showed a lower miRNA expression level. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

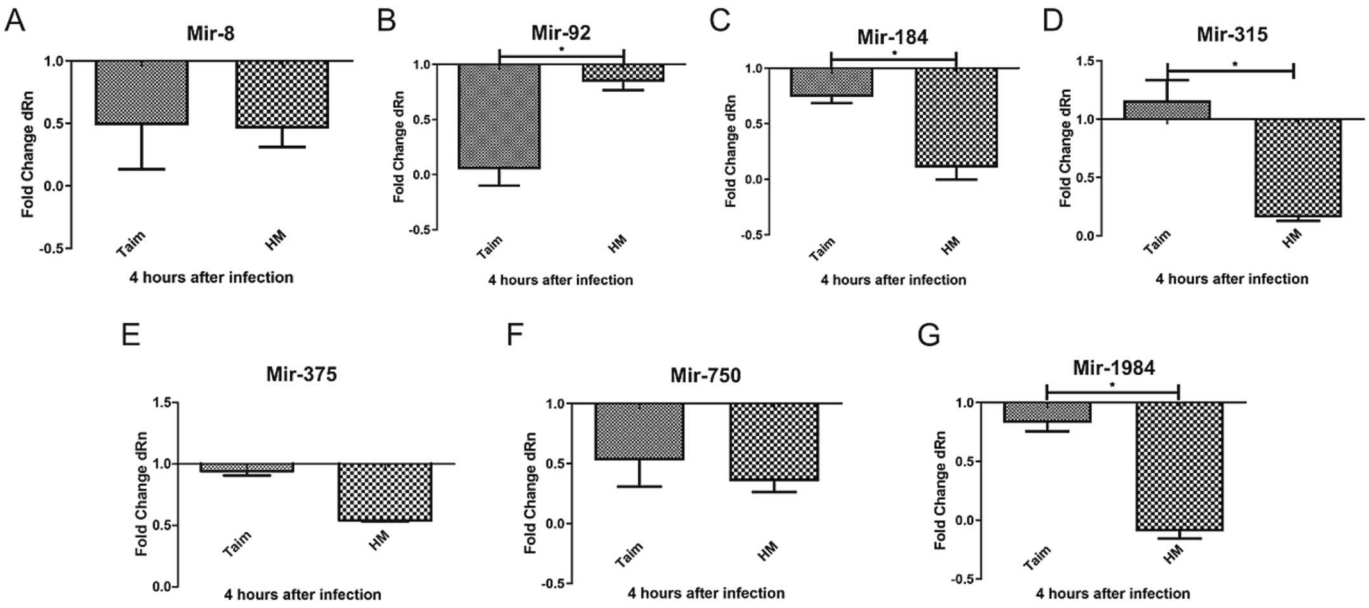


Fig 7. Relative expression of miRNAs four hours after exposure to *S. mansoni* from *B. tenagophila*. Statistically significant p values ≤ 0.05 are marked with asterisks.

4.1.1. General miRNAs

The Mir-124 was found in the genome and transcriptome of *B. tenagophila* with high similarity, especially in the seed region, as can be seen in the alignment of Fig. 2B. This family, found in several eukaryote species and some studies have shown that it is abundantly expressed in the brain and participates in several processes such as neurogenesis, synapse morphology, neurotransmission, neuronal inflammation, autophagy, and mitochondrial function in mammals (Huang et al., 2018). In mollusks, there are also studies showing its expression within cell bodies and neurites of regenerating motorneurons in *Lymnaea stagnalis* (Walker et al., 2018) and identified with neuronal function in squid, *Loligo pealei* (Bitel et al., 2012).

In contrast, the function of Mir-190, another miRNA found in *B. tenagophila* genome and transcriptome, has not yet been clarified in mollusks. Although it can be seen in Fig. S3B fig that there is similarity in the seed region with the protostome and deuterostome homologs, as well. Some studies have shown that it has several targets in humans leading to different types of cancers, such as the DPC4 genes in colorectal cancer (Xie et al., 2018), ZEB1 and SMAD2 in breast cancer (Yu et al., 2018), FOXP2 in gastric cancer (Jia et al., 2016), VEGF which is a gene involved in angiogenesis (Hao et al., 2014) and other human cancers such as cervical, prostate, and liver (Yu and Cao, 2019). As can be seen, Mir-190 regulates several genes and has different targets in humans. In *Drosophila*, Mir-190 has been identified as a regulator of the response to hypoxia (De Lella Ezcurra et al., 2016). Thus, it is expected that the Bte-Mir-190 to be involved in various processes as well as in other organisms.

Also, the Mir-8 found in *B. tenagophila* genome and transcriptome is part of a conserved family of miRNAs found in invertebrate and vertebrate organisms that encompasses not only Mir-8, but also Mir-200, Mir-141, and Mir-429. Mir-8 has been identified in invertebrates and the other miRNAs Mir-200, Mir-141, and Mir-429 play a role only in vertebrate organisms. This family appears to play a role as a regulator of neurogenesis and gliogenesis, and neural homeostasis in the central nervous system of several organisms (Trümbach and Prakash, 2015). In *Tribolium castaneum*, Mir-8 has been associated with a regulatory role in the larva-adult transition, in the development of wings, eyes, and legs (Wu et al., 2019). In *Crassostrea hongkongensis*, Mir-8 has been associated with the development of gonads (Wei et al., 2019).

Mir-71 was found conserved in *B. tenagophila.* Homologue Mir-71 is active in nematode on the nervous system and the transcriptional activity of DAF-16 in the intestine of *Caenorhabditis elegans* (Boulias and Horvitz, 2012). Some miRNAs can be processed from polycistronic transcripts, such as the Mir-71/2 cluster (De Souza et al., 2013). This cluster was found in our study, and it has been shown highly conserved in Protostome species (De Souza et al., 2013). In *Locusta migratory*, the Mir-2/13/71 cluster showed a modulatory action of insect reproduction (Song et al., 2019). Another miRNA cluster found in *B. tenagophila* was Mir-745a/b known to play critical roles in host-pathogen interaction, as identified in *Plutella xylostella* (Xu et al., 2017).

4.1.2. Lophotrochozoa specific miRNAs

MIR-1986, MIR-1990, and MIR-1994, Lophotrochozoa-specific miRNAs families, were also found in *B. tenagophila* genome and transcriptome. MiR-1990 and Mir-1986 were highly expressed in mature testes of *Crassostrea hongkongensis* and Mir-1994 was highly expressed in gonads (Wei et al., 2019). In our work, all these Lophotrochozoa specific miRNAs were expressed in the small RNA datasets. Mir-1986-5p, along with Mir-124-3p, was differentially expressed in the comparison condition of unexposed of the Taim population and unexposed HM. The conservation of these miRNAs in the clade may justify the important role in the reproductive system of Lophotrochozoa species.

4.1.3. Mollusk specific miRNAs

Finally, the mollusk-specific miRNAs families, MIR-1984, MIR-1985, and MIR-2722 were found in *B. tenagophila* genome and transcriptome. In the RT-qPCR validation, after four hours of exposure to *S. mansoni* of the two populations, Mir-1984 was under expressed in both *B. tenagophila* populations, Taim and HM. However, it was more under expressed in exposed Taim populations than in exposed HM population. In *Haliotis discus hannai*, also a marine mollusk, Mir-1984 targeted the bone morphogenetic protein 7 (BMP7) (Huang et al., 2018). Mainly the mature Mir-1984-5p and Mir-1985-5p, showed high expression in the sequenced samples from both populations, suggesting a role in the regulation of immunological mechanisms by this miRNA in the *B. tenagophila*.

4.2. Small-Seq analysis and differential expression of miRNAs

There was a difference in the expression profile in five miRNAs,

according to Table 1, in the challenged snail condition of the two populations considering host reaction to the parasite exposure in a different manner. The Bte-Mir-124-3p and Bte-Mir-153-3p were statistically more expressed in the susceptible HM population, when compared to exposed to *S. mansoni*. The Bte-Mir-750-1-3p, Bte-Mir-375-3p, and Bte-Mir-184-3p had lower expression in the *S. mansoni* exposure condition of the HM population.

To assess the correlation among populations concerning their responses to exposure, employing the complete model, our findings indicate the absence of any miRNA displaying a statistically significant variance in expression. One plausible hypothesis put forth to elucidate this phenomenon posits that the genes implicated in Taim resistance may be constitutive in nature. Consequently, their expression remains unaltered, irrespective of exposure to or absence of *S. mansoni*, as constitutive genes are incessantly expressed at consistent levels across all cells and tissues of an organism, independent of environmental circumstances or specific metabolic requirements. Another conceivable avenue entails the sequencing of small RNAs exclusively from the hemolymph, which plays a pivotal role in pathogen defense, thereby facilitating an assessment of miRNA expression discrepancies.

4.3. Validation by RT-qPCR

Our group identified that snails from the resistant Taim population can attack and destroy *S. mansoni* sporocysts a few hours after the exposure (de Jesus Pereira et al., 2008; Nacif-Pimenta et al., 2012). Therefore, it was decided to evaluate the role of small RNAs in the *B. tenagophila/S. mansoni* interaction. The period of four hours after the host exposure was chosen as the most suitable time to test, since, in this period, there can be an intense response of the mollusk against the pathogen and probably with the involvement of small RNAs. The miRNAs chosen for validation by RT-qPCR were highly expressed in the small RNAs datasets from both, exposed and non-exposed conditions of the two populations (Taim and HM) analyzed.

The differentially expressed miRNAs in RT-PCR validation were Mir-92, Mir-184, Mir-315 and Mir-1984. The Mir-92 was less expressed in the Taim resistant population compared to the susceptible HM (Fig. 7B). This miRNA was associated with the tumor suppressor gene, PTEN, in humans (Zhao et al., 2013). The Mir-184, Mir-315 and Mir-1984 (Fig. 7C, D and G) were higher expressed in Taim, *S. mansoni*-resistant population, compared to HM, *S. mansoni*-susceptible. The Mir-184 was found in *Drosophila* with a dynamic expression pattern during embryo development, especially in the central nervous system (Li et al., 2011). Thus, being associated in important functions in fly development. The Mir-315, which had higher expression in Taim, compared to HM, (Fig. 7D) was also found in *Drosophila* involved in a neural development pathway (Yuan et al., 2020). Finally, the Mir-1984, a mollusk-specific miRNA, was higher expressed in Taim compared to HM (Fig. 7G). It has been reported that this miRNA is in the regulation of the innate immune system response in *C. gigas* (Zhou et al., 2014).

4.4. Main conclusions

Considering the great importance of schistosomiasis in tropical and subtropical countries, being the second most impacting parasitic disease in the world, the sequencing of small RNAs to compare two populations of vector snails, *S. mansoni*-resistant and *S. mansoni*-susceptible, can improve control programs. The identified of miRNAs can contribute to the study of gene expression regulation not only in *Biomphalaria*, but also in other mollusks. Further investigations of this resistant population, such as the analysis of differential gene expression and miRNA depletion, may help to elucidate their roles in the resistance phenotype of the intermediate host *S. mansoni*.

CRediT authorship contribution statement

Tamires Caixeta Alves: Software, Formal analysis, Data curation, Writing – original draft, Visualization. **Fábio Ribeiro Queiroz:** Conceptualization, Methodology, Validation, Investigation, Writing – review & editing, Supervision. **Angelo Borges de Melo Neto:** Software, Formal analysis, Data curation. **Gabriel da Rocha Fernandes:** Formal analysis, Data curation, Supervision. **Fabiano Sviatopolk-Mirsky Pais:** Formal analysis, Data curation. **Wander de Jesus Jeremias:** Conceptualization, Methodology, Writing – review & editing, Formal analysis. **Elio Hideo Babá:** Conceptualization, Methodology, Writing – review & editing. **Marina de Moraes Mourão:** Conceptualization, Methodology, Writing – review & editing. **Enyara Rezende Morais:** Conceptualization, Methodology, Writing – review & editing. **Fernanda Janku Cabral:** Conceptualization, Methodology, Writing – review & editing. **Laurence Rodrigues do Amaral:** Software, Data curation. **Roberta Lima Caldeira:** Conceptualization, Methodology, Writing – review & editing, Resources. **Paulo Marcos Zech Coelho:** Conceptualization, Methodology, Resources, Writing – review & editing. **Matheus de Souza Gomes:** Resources, Software, Writing – review & editing, Supervision, Project administration, Funding acquisition.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

The authors thank FAPEMIG, CAPES, and CNPq for fellowships and financial support - grant numbers: FAPEMIG PPM-00700-18 [M. de Souza Gomes], CNPq 306053/2017-9 [M. de Souza Gomes], CNPq 427855/2018-7 [M. de Souza Gomes] CNPq 313310/2020-3 [M. de Souza Gomes]; the Laboratory of Bioinformatics and Molecular Analysis (LBAM) – UFU and Bioinformatics facility, Fiocruz-Minas, René Rachou Research Center, Belo Horizonte, Brazil for providing the computing facilities for the analyses and for data mining and organization. We also thank the Laboratory of Helminthology and Medical Malacology, Fiocruz-Minas, René Rachou Institute, Belo Horizonte, Brazil for technical support, supply of snail samples, and laboratory facilities.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.gene.2023.147742.

References

Adema, C.M., Hillier, L.D.W., Jones, C.S., Loker, E.S., Knight, M., Minx, P., et al., 2017. Whole genome analysis of a schistosomiasis-transmitting freshwater snail. Nat. Commun. 8 (May).

Ambros, V., 2004. The functions of animal microRNAs. Nature 431 (7006), 350–355.

An, J., Lai, J., Lehman, M.L., Nelson, C.C., 2013. MiRDeep*: An integrated application tool for miRNA identification from RNA sequencing data. Nucleic Acids Res. 41 (2), 727–737.

Bartel, D.P., 2004 Jan 23. MicroRNAs: genomics, biogenesis, mechanism, and function. Cell 116 (2), 281–297. https://doi.org/10.1016/s0092-8674(04)00045-5. PMID: 14744438.

Basch, P.F., 1976. PARASITOLOGICAL Intermediate Host Specificity in *Schistosoma mansoni*. Exp. Parasitol. 169 (39), 150–169.

Bezerra, F.S.D.M., Nogueira-Machado, J.A., Chaves, M.M., Martins, R.L., Coelho, P.M.Z., 1997. Quantification of the population and phagocytary activity of hemocytes of resistant and susceptible strains of *Biomphalaria glabrata* and *Biomphalaria tenagophila* infected with *Schistosoma mansoni*. Rev. Inst. Med. Trop. Sao Paulo 39, 197–201.

Bitel, C.L., Singh, V., Frederikse, P.H., 2012. MiR-124, miR-125b, let-7 and vesicle transport proteins in squid lenses in *L. pealei*. Curr. Eye Res. 37 (5), 388–394.
Bodh, V., Chawla, Y., 2014. Noncirrhotic intrahepatic portal hypertension. Clin. Liver Dis. 3 (6), 129–132.
Boulias, K., Horvitz, H.R., 2012. The *C. elegans* MicroRNA mir-71 acts in neurons to promote germline-mediated longevity through regulation of DAF-16/FOXO. Cell Metab. 15 (4), 439–450.
Bu, L., Lu, L., Laidemitt, M., Zhang, S.M., Mutuku, M., et al., 2023. A genome sequence for *Biomphalaria pfeifferi*, the major vector snail for the human-infecting parasite *Schistosoma mansoni*. PLoS Negl. Trop. Dis. 17 (3), e0011208.
Coelho, P., Caldeira, R., 2016. Critical analysis of molluscicide application in schistosomiasis control programs in Brazil. Infect. Dis. Poverty 5, 57. https://doi.org/10.1186/s40249-016-0153-6.
Coelho, P.M.Z., Carvalho, O.S., Andrade, Z.A., Martins-Sousa, R.L., Rosa, F.M., Barbosa, L., et al., 2004. *Biomphalaria tenagophila/Schistosoma mansoni* interaction: Premises for a new approach to biological control of schistosomiasis. Mem. Inst. Oswaldo Cruz 99 (5), 109–111.
de Jesus Pereira, C.A., Martins-Souza, R.L., Corrêa Jr, A., Coelho, P.M.Z., Negrão-Corrêa, D., 2008. The participation of cell-free hemolymph of Biomphalaria tenagophila in the defense mechanism against to Schistosoma mansoni sporocysts. Parasite Immunol. https://doi.org/10.1111/j.1365-3024.2008.01062.x.
De Lella Ezcurra, A.L., Bertolin, A.P., Kim, K., Katz, M.J., Gándara, L., Misra, T., et al., 2016. miR-190 Enhances HIF-Dependent Responses to Hypoxia in *Drosophila* by Inhibiting the Prolyl-4-hydroxylase Fatiga. PLoS Genet. 12 (5), 1–24.
De Souza, G.M., Donoghue, M.T.A., Muniyappa, M., Pereira, R.V., Guerra-Sá, R., Spillane, C., 2013. Computational identification and evolutionary relationships of the MicroRNA gene cluster miR-71/2 in protostomes. J. Mol. Evol. 76 (6), 353–358.
dos Santos, M.B.L., de Freitas, J.R., dos Reis Correia, M.C., Coelho, P.M.Z., 1979. Susceptibility of Biomphalaria tenagophila hybrids to Schistosoma mansoni: crossing between strains from Taim (RS), Cabo-Frio (RJ), and Belo Horizonte (MG), Brasil. Revista do Instituto de Medicina Tropical de Sao Paulo 21 (6), 281–286.
Ewing, B., Green, P., 1998. Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res. 8 (3), 186–194.
Fromm, B., Billipp, T., Peck, L.E., Johansen, M., Tarver, J.E., King, B.L., et al., 2015. A Uniform System for the Annotation of Vertebrate microRNA Genes and the Evolution of the Human microRNAome. Annu. Rev. Genet. 49 (v), 213–242.
Hao, Y., Yang, J., Yin, S., Zhang, H., Fan, Y., Sun, C., et al., 2014. The synergistic regulation of VEGF-mediated angiogenesis through miR-190 and target genes. RNA 20 (8), 1328–1336.
Hofacker, I.L., 2003. Vienna RNA secondary structure server. Nucleic Acids Res. 31 (13), 3429–3431.
Huang, S., Ge, X., Yu, J., Han, Z., Yin, Z., Li, Y., Chen, F., Wang, H., Zhang, J., Lei, P., 2018. Increased miR-124-3p in microglial exosomes following traumatic brain injury inhibits neuronal inflammation and contributes to neurite outgrowth via their transfer into neurons. FASEB J. 32 (1), 512–528. https://doi.org/10.1096/fj.201700673R.
Huang, J., Luo, X., Huang, M., et al., 2018. Identification and characteristics of muscle growth-related microRNA in the Pacific abalone. Haliotis discus hannai. BMC Genomics 19, 915. https://doi.org/10.1186/s12864-018-5347-952.
Jia, W.Z., Yu, T., An, Q., Yang, H., Zhang, Z., Liu, X., et al., 2016. MicroRNA-190 regulates FOXP2 genes in human gastric cancer. Onco Targets Ther. 9, 3643–3651.
Kang, W., Eldfjell, Y., Fromm, B., et al., 2018. miRTrace reveals the organismal origins of microRNA sequencing data. Genome Biol. 19, 213. https://doi.org/10.1186/s13059-018-1588-9.
Kumar, S., Stecher, G., Li, M., Knyaz, C., Tamura, K., 2018 Jun 1. MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. Mol. Biol. Evol. 35 (6), 1547–1549. https://doi.org/10.1093/molbev/msy096. PMID: 29722887; PMCID: PMC5967553.
Li, P., Peng, J., Hu, J., et al., 2011. Localized expression pattern of miR-184 in Drosophila. Mol. Biol. Rep. 38, 355–358. https://doi.org/10.1007/s11033-010-0115-1.
Liu, Q., Zhang, H., 2012. Molecular Identification and Analysis of Arsenite Stress-Responsive miRNAs in Rice. J. Agric. Food Chem. 60 (26), 6524–6536.
Livak, K.J., Schmittgen, T.D., 2001. Analysis of relative gene expression data using real-time quantitative PCR and the 2-ΔΔCT method. Methods 25 (4), 402–408.
Marco, A., Kozomara, A., Hui, J.H.L., Emery, A.M., Rollinson, D., Griffiths-Jones, S., et al., 2013. Sex-Biased Expression of MicroRNAs in *Schistosoma mansoni*. PLoS Negl. Trop. Dis. 7 (9), 1–8.
Nacif-Pimenta, R., Mattos, A.C.A.D., Orfanó, A.D.S., Barbosa, L., Pimenta, P.F.P., Coelho, P.M.Z., 2012. *Schistosoma mansoni* in susceptible and resistant snail strains *Biomphalaria tenagophila*: in vivo tissue response and in vitro hemocyte interactions. PLoS One. 7 (9), 1–12.
Navarro, F.J., Baulcombe, D.C., 2019. miRNA-Mediated Regulation of Synthetic Gene Circuits in the Green Alga *Chlamydomonas reinhardtii*. ACS Synth. Biol. 8 (2), 358–370.
Paraense, W.L., Corrêa, L.R., 1978. Differential Susceptibility of Biomphalaria tenagophila Populations to Infection with a Strain of Schistosoma mansoni. J. Parasitol. 64 (5), 822–826. https://doi.org/10.2307/3279509.
Pereira De Souza, C., Jannotti Passos, L.K., 2001. Resistance of *Biomphalaria* occidentalis from Varzea das Flores Dam, Minas Gerais, to *Schistosoma mansoni* Infection Detected by Low Stringency Polymerase Chain Reaction. Mem. Inst. Oswaldo Cruz 96 (3), 381–383.
Rosa, F.M., Godard, A.L., Azevedo, V., Coelho, P.M., 2005 Feb. Biomphalaria tenagophila: dominant character of the resistance to Schistosoma mansoni in descendants of crossbreedings between resistant (Taim, RS) and susceptible (Joinville, SC) strains. Mem. Inst. Oswaldo Cruz 100 (1), 19–23. https://doi.org/10.1590/s0074-02762005000100004. Epub 2005 Apr 12 PMID: 15867958.
Rosa, F.M., Marques, D.P.A., Maciel, E., Couto, J.M., Negrão-Corrêa, D.A., Teles, H.M.S., et al., 2013. BREEDING OF *Biomphalaria tenagophila* IN MASS SCALE. Rev. Inst. Med. Trop. Sao Paulo 55 (1), 39–44.
Rosani, U., Bortoletto, E., Bai, C.M., Novoa, B., Figueras, A., Venier, P., Fromm, B., 2021. Digging into bivalve miRNAomes: between conservation and innovation. Philos Trans R Soc Lond B Biol Sci. 376 (1825), 20200165 https://doi.org/10.1098/rstb.2020.0165.
Song, J., Li, W., Zhao, H., Zhou, S., 2019. Clustered miR-2, miR-13a, miR-13b and miR-71 coordinately target Notch gene to regulate oogenesis of the migratory locust *Locusta migratoria*. Insect Biochem Mol Biol. 106, 39–46. https://doi.org/10.1016/j.ibmb.2018.11.004.
Spangenberg, T., 2021. Alternatives to Praziquantel for the Prevention and Control of Schistosomiasis. ACS Infect. Dis. 7 (5), 939–942.
Trümbach, D., Prakash, N., 2015. The conserved miR-8/miR-200 microRNA family and their role in invertebrate and vertebrate neurogenesis. Cell Tissue Res. 359 (1), 161–177.
Tuan, R., Santos, P., 2007. ITS2 variability of Biomphalaria (Mollusca, Planorbidae) species from the Paranapanema Valley (São Paulo State, Brazil): diversity patterns, population structure and phylogenetic relationships. Genet. Mol. Biol. 30, 139–144. https://doi.org/10.1590/S1415-47572007000100024.
Walker, S.E., Spencer, G.E., Necakov, A., Carlone, R.L., 2018. Identification and characterization of microRNAs during retinoic acid-induced regeneration of a molluscan central nervous system. Int. J. Mol. Sci. 19 (9).
Wei, P., He, P., Zhang, X., Li, W., Zhang, L., Guan, J., et al., 2019. Identification and characterization of microRNAs in the gonads of *Crassostrea hongkongensis* using high-throughput sequencing. Comp Biochem Physiol - Part D Genomics Proteomics [Internet]. 31 (July), 100606 https://doi.org/10.1016/j.cbd.2019.100606.
Wheeler, B.M., Heimberg, A.M., Moy, V.N., Sperling, E.A., Holstein, T.W., Heber, S., et al., 2009. The deep evolution of metazoan microRNAs. Evol. Dev. 11 (1), 50–68.
World Health Organization. Schistosomiasis (Bilharzia). 2022. https://www.who.int/health-topics/schistosomiasis#tab=tab_1. Accessed 22 Aug 2022.
Wu, W., Zhai, M., Li, C., Yu, X., Song, X., Gao, S., et al., 2019. Multiple functions of miR-8-3p in the development and metamorphosis of the red flour beetle, *Tribolium castaneum*. Insect Mol Biol. 28 (2), 208–221.
Xie, B., Deng, Z., Pan, Y., Fu, C., Fan, S., Tao, Y., et al., 2018. Post-transcriptional regulation DPC4 gene by miR-190 in colorectal cancer cells. J. Cancer Res. Ther. 14 (4), 838–842.
Xu, J., Xu, X., Li, S., Wang, S., Xu, X., Zhou, X., Yu, J., Yu, X., Shakeel, M., Jin, F., 2017 Dec. Genome-Wide Profiling of Plutella xylostella Immunity-Related miRNAs after Isaria fumosorosea Infection. Front. Physiol. 14 (8), 1054. https://doi.org/10.3389/fphys.2017.01054. PMID: 29311981; PMCID: PMC5735356.
Yu, Y., Cao, X.C., 2019. MiR-190-5p in human diseases. Cancer Cell. Int. [Internet]. 19 (1), 1–9. https://doi.org/10.1186/s12935-019-0984-x.
Yu, Y., Luo, W., Yang, Z.J., Chi, J.R., Li, Y.R., Ding, Y., et al., 2018. miR-190 suppresses breast cancer metastasis by regulation of TGF-β-induced epithelial-mesenchymal transition. Mol. Cancer 17 (1), 1–12.
Yuan, L., Ren, X., Zheng, Y., Qian, J., Xu, L., Sun, M., 2020. MiR-315 is required for neural development and represses the expression of dFMR1 in *Drosophila melanogaster*. Biochem. Biophys. Res. Commun. [Internet]. 525 (2), 469–476. https://doi.org/10.1016/j.bbrc.2020.02.127.
Zhao, B., Zhu, Y., Cui, K., Gao, J., Yu, F., Chen, L., et al., 2013. Expression and significance of PTEN and miR-92 in hepatocellular carcinoma. Mol. Med. Rep. 7 (5), 1413–1416.
Zhou, Z., Wang, L., Song, L., Liu, R., Zhang, H., Huang, M., Chen, H., 2014 Feb 6. The identification and characteristics of immune-related microRNAs in haemocytes of oyster Crassostrea gigas. PLoS One 9 (2), e88397. https://doi.org/10.1371/journal.pone.0088397.

Supplementary Materials

SUPPLEMENTAL FILE

S1 Table - TaqMan® assays.

miRNA	Assay ID	Assay name
miR-8	269	dme-miR-8
miR-92	241870_mat	dpu-miR-92
miR-184	290	dme-miR-184
miR-315	325	dme-miR-315
miR-375	007201_mat	aca-miR-375
miR-750	008337_mat	cte-miR-750
miR-1984	006673_mat	hru-miR-1984
U6 snRNA	1973	U6 snRNA

S2 Table - Conserved precursors and mature miRNAs *B. tenagophila* and homologs.

Precursor name	Pre-miRNA Sequence	miRNA length (nt)	Mature miRNA ID (-5p and -3p)	Mature miRNA sequence	Mature miRNA start position	Mature miRNA end position	Homologue miRNA
bte-miR-29b	UGCUGUGGAUGUUACAGUUGGAAUUCUGGUCUUGAAUGGUGGAUAGAUGCUGUCUAUUUAUCUCUAGCACCAUUUGAAAUCAGUGCUUCCGAUGUGUAAG	100	bte-miR-29b-3p	UAGCACCAUUUGAAAUCAGUGCU	65	87	bbe-miR-29b-3p
bte-miR-8-2	GAGCAACUGCAUUUCAUCUUACCAAACAGCAUUAGUUCGAUUUUCUUGAAUUGCUAAUACUGUCAGGUAAAGAUGGCUUAAGGACGC	87	bte-miR-8-2-3p	UAAUACUGUCAGGUAAAGAUG	55	75	mle-miR-8-3p
			bte-miR-8-2-5p	CAUCUUACCAAACAGCAUUAG	15	35	pte-miR-8-5p
bte-miR-33a	AAUUUGAGAAAGGAUACUGUUAGUGGUGCAUUGUAGUUGCAUUGCGUGUCAAGCUUGUGCAAUGUCUCUGCAGUGCAAUCACCAAUGUAUUCUCAUGCUUCA	102	bte-miR-33a-5p	GUGCAUUGUAGUUGCAUUGC	26	45	ocu-miR-33a-5p
			bte-miR-33a-3p	CAAUGUCUCUGCAGUGCA	60	77	oni-miR-33a-3p
bte-miR-31	ACCAGAAUAGAGUUCAGACCAGGCAAGAUGUUGGCAUAGCUGAGUCAAACAACUGUCAGCUGUGUCAUAUGUUGCCAAUUUUGGACUCGUUGCGUCAU	98	bte-miR-31-5p	AGGCAAGAUGUUGGCAUAGCUGA	21	43	mle-miR-31-5p
			bte-miR-31-3p	AGCUGUGUCAUAUGUUGCCAAU	58	79	mle-miR-31-3p
bte-miR-71-1	GAAUUGAAAGAAAGUAUUCUUUGAAAGACAUGGGUAGUGAGAUGUUUAUUUUUAAAAAUUCAUCUUACUAGCCUGUCUUUCGAAAGAAAACUCUUCCUAAU	101	bte-miR-71-1-5p	UGAAAGACAUGGGUAGUGAGAUG	22	44	mle-miR-71-5p
			bte-miR-71-1-3p	UCUUACUAGCCUGUCUUU	63	80	pte-miR-71-3p
bte-miR-190-1	GAUAGAGGCCAGCGCCCUGAGAGAUAUGUUUGAUAUAUUUGGUGGUGUGUACAGAUUCACCAGAUAUCAGACAUGUCACAACAGAGGUUGGCUAGUCA	98	bte-miR-190-1-5p	AGAUAUGUUUGAUAUAUUUGGUGG	22	45	cte-miR-190
			bte-miR-190-1-3p	ACCAGAUAUCAGACAUGU	59	76	ssa-miR-190b-3p

bte-miR-2a-2	CAAUAUAUAUACAAUGCAUGCUGCACUGACCAGGUGGUUGGGAUGUGCUUUUUUUACUACAUAUCACAGCCUGCUUGGAUCAGUACUGCGUCAUUUUUCUC	101	bte-miR-2a-2-3p	UAUCACAGCCUGCUUGGAUCAGUA	62	85	mle-miR-2a-3p
			bte-miR-2a-2-5p	CUGACCAGGUGGUUGGGAUGUG	26	47	mle-miR-2a-5p
bte-miR-96b	CGCUCAUUUGGCACUUGUGGAAUAAUCGUUACAAAGGAUUGGAUUAUACAUCGGUGCCAAAUAAGA	66	bte-miR-96b-5p	AUUUGGCACUUGUGGAAUAAUCG	6	28	lgi-miR-96b
			bte-miR-96b-3p	AUUAUACAUCGGUGCCAAAUAA	43	64	mle-miR-96b-3p
bte-miR-153	CCAACUAACUGCUGCUUUUCCUGGCAGCUUUUGUGAUUUAGCAAUUGUACGGUGGGCAAAUUGCAUAGUCACAAAAGUGAUCGGGACUGCAUGCUUCCAC	100	bte-miR-153-3p	UUGCAUAGUCACAAAAGUGAUC	61	82	nle-miR-153
			bte-miR-153-5p	GCUUUUGUGAUUUAGCAAUU	27	46	mle-miR-153-5p
bte-miR-184	AAUGACAAACUUAAAGCCUACGUUUCCUCAUCACUUUUCCAUCCCGUCUGUAACAGAGAACUGGACGGAGAACUGAUAAGGGCUUGUGGGUGACCUUUCACC	102	bte-miR-184-3p	ACUGGACGGAGAACUGAUAAGGGC	60	83	bmo-miR-184-3p
			bte-miR-184-5p	UCCUCAUCACUUUUCCAUCCC	25	45	xla-miR-184-5p
bte-miR-281-2	AAGAGAAUUUCUGACCCAAACGAUGAAGGGAGCAUCCGUCGACAGUCGGAAAGAAGGCACUGUCAUGGAGUUGCUCUCUUUACCGCUAAGUCAAAAUCGA	100	bte-miR-281-2-3p	UGUCAUGGAGUUGCUCUCUUUA	61	82	mle-miR-281-3p
			bte-miR-281-2-5p	AAGGGAGCAUCCGUCGACAGU	26	46	mle-miR-281-5p
bte-miR-375	AAGCUUCAAUCAAAACGCCAGGUGACACGAGCAGUCCGCAACAAGACCUCAACAGGUGUUUUGUUCGUUCGGCUCGCGUUAUUAAGGCGGUCAUUCCACC	100	bte-miR-375-3p	UUUGUUCGUUCGGCUCGCGUUAU	60	82	bbe-miR-375-3p
			bte-miR-375-5p	ACACGAGCAGUCCGCAACAAG	25	45	mle-miR-375-5p

bte-miR-1986	CAUCUGUAGCCGCGGGCGUUGGGGGCCCCACGUAAUAAAAUUGAGUGGAUUUCCCAAGAUCCGUGA	66	bte-miR-1986-3p	UGGAUUUCCCAAGAUCCGUGA	46	66	mle-miR-1986-3p
			bte-miR-1986-5p	CGCGGGCGUUGGGGGCCCCACG	11	32	mle-miR-1986-5p
bte-miR-2722	GUUUACCAAACUGGAGAGGUGUUUCAUAGGCGCUACAUUGUGAGGACCGUGGCGCCGUGGAAACAUCUACCCAGUGAGUAGAC	83	bte-miR-2722-3p	UGGCGCCGUGGAAACAUCUACC	50	71	lgi-miR-2722
			bte-miR-2722-5p	GAGGUGUUUCAUAGGCGCUACA	16	37	mle-miR-2722-5p
bte-miR-29a	UCCAUCUGAUCAGCAUCUAUGCUGAACUACCUCAAGGAAACCUGGUCUCUUCUGGUGCUUAGAUUUUAAAAAGGGUCUAGCACCAUUUGAAAUCAGUUUUCUACUGUGGUACAAUCGACUGAGAUGUAAUG	131	bte-miR-29a-3p	UAGCACCAUUUGAAAUCAGUUU	78	99	mle-miR-29a-3p
			bte-miR-29a-5p	CCUGGUCUCUUCUGGUGCUUAGA	41	63	mle-miR-29a-5p
bte-let-7l	ACGAUGAGAAUAUCCCGAAGUGAGGUAGUAGGUUGUAUUGUUAAGAACUACACCAUUACUGAAGAACAGUACAAUCUGCUAGCUUUCCCCGUGGUAUGAACUCGCUGC	108	bte-let-7l-5p	UGAGGUAGUAGGUUGUAUUGUU	21	42	gga-let-7l-5p
bte-miR-279	AAUUAAAAAGUCCACAUUUUGUGGAUGGCUGUUGGUCUGGUGCAUGUCCCUUAUAAUCAUGACUAGAUCCACACUCAUCCAUAAUGGUGGGGACUAGCAUCC	102	bte-miR-279-3p	UGACUAGAUCCACACUCAUCCA	60	81	mle-miR-279-3p
			bte-miR-279-5p	GAUGGCUGUUGGUCUGGUGCAUG	24	46	mle-miR-279-5p
bte-miR-7	CUUUCUGCACUAACUCUAACCCUGACUGUUGUGGUUUGGAAGACUAGUGAUUUAGUUGUUUUGUAAUGAGACCAACAAUAAAUCACAAUCUUCUAGGACUACGGCAGUCACCUAACCUCUG	121	bte-miR-7-5p	UGGAAGACUAGUGAUUUAGUUGU	37	59	mle-miR-7-5p
			bte-miR-7-3p	CAAUAAAUCACAAUCUUCUAGGA	76	98	mle-miR-7-3p

bte-miR-137-2	UGUUUGCAUUGCCGUGUCUCUUACGGGUAUUCUUGGGUAAAUAAUACAACAUCAGGCUGUUAUUGCUUGAGAAUACACGUAACGGCCUCGGUUAAAUCACAGCCA	105	bte-miR-137-2-3p	UAUUGCUUGAGAAUACACGUAA	61	82	mle-miR-137-3p
			bte-miR-137-2-5p	ACGGGUAUUCUUGGGUAAAUAAUA	23	46	mdo-miR-137b-5p
bte-miR-9-1	UACGGCUUCCAGUCUUGCCUUUGGUAACCUAGCUUUAUGAAUUUGUUCUACAAUCAUACAGCUAGAUAACCAAAGACAAGAACGGCGCCCUCCCA	95	bte-miR-9-1-3p	UCAUACAGCUAGAUAACCAAAGA	54	76	oha-miR-9-3p
			bte-miR-9-1-5p	CUUUGGUAACCUAGCUUUAUGA	19	40	bmo-miR-9b-5p
bte-miR-278	GGCCUCUAUCGGAUGAAAGCCUGACAGGAUCACUUCAAUGAUCGGUGGGACUUUCGUUCGUUUGAGUUC	69	bte-miR-278-3p	UCGGUGGGACUUUCGUUCGUUU	42	63	api-miR-278
bte-miR-4974-2	ACGUUAUCUAGUUAUGGCUAAAAUCUCUCUAUAUAUAAAAUAUAUAUAUUCCUAUUUAACUCUGUAGAGAGAUUGCUUUGGAUAUCUAAUAAAAA	95	bte-miR-4974-2-5p	UAUAAAAUAUAUAUAUUCCU	34	53	dme-miR-4974-3p
bte-miR-277	CUGCUGCCAGAGUUUCCUUCCGUACCAGCAAUGCAUUCUACAUGUACUUGCAUGCUGUGUAAAUGCAUUAUCUGGUAUCUGAAGGAGGCCUGGCUAGUUC	100	bte-miR-277-3p	UAAAUGCAUUAUCUGGUAUCUGA	60	82	mle-miR-277-3p
			bte-miR-277-5p	GUACCAGCAAUGCAUUCUACA	22	42	mle-miR-277-5p
bte-miR-750-2	GUUCUAGUUAGAAGUUGGAGGUUAGAUCUAAAAGUAGUAUUCUAGAGAUAUAGAUAUAAUCUUUAUAUAUUGACCUUGAUUCUAAUCUAGUCU	93	bte-miR-750-2-5p	AGUUGGAGGUUAGAUCUA	13	30	asu-miR-750-5p
bte-miR-5605	UGUUUUAAGAGAUCAUAAUUAGGAUGGGGGAAGAUGUGCACAACAUACAUAUUCUAGGGUAUACCUAUUCUAGAUCUAUUAAAGGG	86	bte-miR-5605-5p	UAGGAUGGGGGAAGAUGUG	20	38	cin-miR-5605-3p
bte-miR-8482-2	GAAAUUAUGUAAUAACCCAUCUCUAUCUAGCGUACUUAUGACUAUUAAUUAAAUAGUUCUACUCAAGAAAGUAAUGGGUUAAUUAUAAUAUG	92	bte-miR-8482-2-3p	AAUGGGUUAAUUAUAAUAUG	73	92	sma-miR-8482-5p
bte-miR-4829c	UUUAGUAUUGUACUUUUCCCAAUGCGUACAAAUAAAUGUGGCAUAUAUGGAUAAAAGUAACAAAUGCAUCGACAUUGGAGUAAAACCAUACUCAU	95	bte-miR-4829c-5p	UGUACUUUUCCCAAUGCGU	9	27	sko-miR-4829c

bte-miR-8412	UCUAUGCUCUCUUAAAGUGAAUGAUAGUCUUGGUUCUAACUGCUUAGCUUCCGUGUGCAGUGAUAUUGUCUAUCACUAAACUUAACACGAUCAUGGC	97	bte-miR-8412-3p	CAGUGAUAUUGUCUAUCACU	58	77	sma-miR-8412-5p
bte-miR-76	GACGAUACAAAUGAAACAAGUUUCGUUGUUGAUGAAGUUUCCAAUGUUAUUAAAGAAUCGAUCGAAGCAACGAUCGGUUCUCAAUCGUAUCAAC	94	bte-miR-76-5p	UUCGUUGUUGAUGAAGUUU	22	40	cbn-miR-76
bte-miR-2b-3	UAUAGGUAUCCCUGUCAUCAAAGCUGGCUGUGAUAUGUGUUUAGUGUUACCACAUCACAACCGCCUUGACGUCAGGGUAGCCUGGU	86	bte-miR-2b-3-5p	UCAUCAAAGCUGGCUGUGAUAUG	15	37	ame-miR-2b-5p
bte-miR-2b-2	UAAAUCAAUCUUGAAGCUCAUCAAAGCUGGCUGUGAUAUGAAAAUUCUCUACAAAUCACAGUCACUCUGACGUGCCACAAAGGUGAUACU	90	bte-miR-2b-2-5p	UCAUCAAAGCUGGCUGUGAUAUGA	18	41	ame-miR-2b-5p
bte-miR-1004	AGAUAUAUGAAUGUUUAAAAAGAAAGUGGAGUGCUCCCCUUAGAAAGUUGGGGGACAUUGAUCUACUUGAAACAUCUCAAAUAGGC	86	bte-miR-1004-3p	GUUGGGGGACAUUGAUCU	47	64	dme-miR-1004-5p
bte-miR-8482-1	GUGAAAUAAUGUAUUAACCCAUCUCUAUCUAGCGUACUUAUGACUAUUAAUUAAAUAGUUCUACUCAAGAAAGUAAUGGGUUAAUUAUAAUAAUUAUU	98	bte-miR-8482-1-3p	AAUGGGUUAAUUAUAAUAA	75	93	sma-miR-8482-5p
bte-miR-12	CACGAAGCUUAAGGUCUGAGUAUUACUUCAGGUACUGAGUGAAUGGGCUGCGGUCAGUACCUAAUGUGAUAUUCUUACCUUGCGCUCUCACA	92	bte-miR-12-5p	UGAGUAUUACUUCAGGUACUGA	17	38	mle-miR-12-5p
			bte-miR-12-3p	AGUACCUAAUGUGAUAUUCUUA	56	77	mle-miR-12-3p
bte-miR-8-1	UUGCGCGAGGUGAAGGUCGUCUUACCUAGCAGCAUUGGACGUUGAUCUGACGAGCGAGUCUCUAAUACUGUCAGGUAAAGAUGUCUUUCGCCCGCGCCC	99	bte-miR-8-1-3p	UAAUACUGUCAGGUAAAGAUGUC	63	85	mle-miR-8-3p
			bte-miR-8-1-5p	GUCUUACCUAGCAGCAUUGGA	19	39	mle-miR-8-5p
bte-miR-745a	GAAUGAGACAAGUCCAUGUAGGCUCUUCACCUGGUAGACUUGUUGGCUUGGUCAAGCUGCCUGAUGAAGAGCUGUCCAUGGGUUGGCUACACAC	94	bte-miR-745a-3p	AGCUGCCUGAUGAAGAGCUGUCC	55	77	mle-miR-745a-3p
			bte-miR-745a-5p	GGCUCUUCACCUGGUAGACUUG	21	42	mle-miR-745a-5p

bte-miR-745b	CCAUGACAUCUGGGGUAGCAGCUCCUCCUUCGGUCAGCUAUCCUAUUCCCAUGUUGGGAGCUGCCAAAUGAAGGGCUGUUGCCAAAUGUCAGAA	94	bte-miR-745b-3p	GAGCUGCCAAAUGAAGGGCUGU	58	79	mle-miR-745b-3p
			bte-miR-745b-5p	GCUCCUCCUUCGGUCAGCUAUCCU	21	44	mle-miR-745b-5p
bte-miR-1175	UGUCAGGCGUGUUAUGGAGUGGAGAGAGUUUUAUCUCAUCGUGUUUGGCGAAAUCUGAUGAGAUUCAACUCCUCCAACUGCAGAGCACGCUUGGAA	96	bte-miR-1175-3p	UGAGAUUCAACUCCUCCAACUGC	59	81	mle-miR-1175-3p
			bte-miR-1175-5p	AGUGGAGAGAGUUUUAUCUCAUC	18	40	mle-miR-1175-5p
bte-miR-7d	AAGUCUUGUGACAACAUUUGGACACAAUUGUAGCUUAAAAUAGGUCAACAUUAUUUCUGUUUGUGUUGCCAAUGUAUCUAAUUUUGUUAGAAGAAUA	97	bte-miR-7d-3p	CAACAUUAUUUCUGUUUGUG	46	65	sme-miR-7d-3p
bte-miR-4974-1	UAAUUAAAAUAUAAAAUAUAUAUAUUAGGCCUAGAUCUACUUUAUUUAGUAGUCUAGACUCUGUGCAGUUGUAAAUGUAUAGUAUACUAUUGUAACUA	98	bte-miR-4974-1-5p	UAUAAAAUAUAUAUUAGG	10	29	dme-miR-4974-3p
bte-miR-3849	CGUUGGCUAUCAACCAGUUGCAAUGCCAGGUCAGCAUUGUGUUGGAAAUCAACAGGUACAGCAGUGUAGUGGUUAUCAGCCAGGA	85	bte-miR-3849-5p	AGCAUUGUGUUGGAAAUCAA	33	52	tca-miR-3849-3p
bte-miR-9539	AUUUAGAAUCCACAUUGAGACUCUGAGUAGACUCUAGAUUCUAGAAGCUAUCACUUACAGGCUUAGUAGUGCCGUUCUAGUU	82	bte-miR-9539-5p	ACAUUGAGACUCUGAGUAG	12	30	dvi-miR-9539-5p
bte-miR-2491	UUCCAACAGAUGAUGGGUGCUGUUAAUCUAACUGGGGUGCAGCAACAACAGCAGCAGCAACUGCAGUUGCUUCAACUGCAACAGCUAUUGUUGCAA	96	bte-miR-2491-3p	CAACAACAGCAGCAGCAA	43	60	dme-miR-2491-3p
bte-miR-1990	AGAUGAGCUUGUUAUGAGAAGUAAGUUGAUGGGGUCCCAGGCGGAUGAAAUGUUCCGGGACUACGUCAACUACUUGCUUGUACCAAGCUACACCC	95	bte-miR-1990-5p	AGUAAGUUGAUGGGGUCCCAGG	20	41	mle-miR-1990-5p
			bte-miR-1990-3p	CGGGACUACGUCAACUACUUG	56	76	mle-miR-1990-3p
bte-miR-219d	AAGGUCAAGUGUCAAGCACUGAUGUCCACACACAGUUCUUGGUGUGAUCAGGAACAAGAAUUGCGUUUGGACAAUCAUAGCCAGAUUACUUGGACUGC	98	bte-miR-219d-3p	AGAAUUGCGUUUGGACAAUCA	57	77	oni-miR-219d

bte-miR-252b	GUCCAUAGGUCAGUCUCCUUCUAUAAGUAGUGGUGCCGCAGGUACGUCCUGGUCACUCUACCUGCACACCGGUGCUUACAUUGGGAGACGGACAUUGCUG	100	bte-miR-252b-5p	AUAAGUAGUGGUGCCGCAGGUA	23	44	mle-miR-252b-5p
			bte-miR-252b-3p	CCUGCACACCGGUGCUUAC	61	79	mle-miR-252b-3p
bte-miR-252a	GAAUUUCUUACCCAUUUCUAAGUACUGGUGCCGCGGGACUGUAUAACACUGGCCUCCAGUGGUCCCGGUACUUAUCAGAGGGUGGAAACCA	91	bte-miR-252a-5p	CUAAGUACUGGUGCCGCGGGA	18	38	mle-miR-252a-5p
			bte-miR-252a-3p	CCAGUGGUCCCGGUACUUA	56	74	mle-miR-252a-3p
bte-miR-3904	AAGUAAAUAAUGUGGACUACACUUAUUUAAAAAUAAUGUGGACUACACUUAUUUAAAAAUAAUGUGGACUACACUUAUUUUAAAA	84	bte-miR-3904-5p	CUUAUUUAAAAAUAAUGU	22	39	tca-miR-3904-3p
			bte-miR-3904-3p	CUUAUUUAAAAAUAAUGU	48	65	tca-miR-3904-3p
bte-miR-4109	UAUGAAAUUCGGUUAAUAGAGCCAGUCGGUGGUAACUACCAAACAAAAAUUUCGCUUGGAAGUGGUCCGAUUAAGCGGAAUUGUCCGU	88	bte-miR-4109-3p	GUGGUCCGAUUAAGCGGA	62	81	cin-miR-4109-3p
bte-miR-34	GCAGUCCCUGAUUUUUUGGCAGUGUGGUUAGCUGGUUGUUUCGCCUGCAACAACCACUCUCCACAUUACCGCCAAAUCAGUCGACCAU	88	bte-miR-34-5p	UGGCAGUGUGGUUAGCUGGUUGU	17	39	mle-miR-34-5p
bte-miR-125b-1	UUGACUAGCUCUAACAAGUCAGGUUCUUGGUAGCCUUGAGGUAAAGAAUCAAGAGUAAGAUCUUUGACUUUCAAGUAGUUGG	82	bte-miR-125b-1-5p	ACAAGUCAGGUUCUUGGUAGC	14	34	mmu-miR-125b-2-3p
bte-miR-216b	CACCUGUCUAAUAUCAGCUGGUAAUCCUGAGUUAAAGAAAAUCUCAGGAAACUGGUUGGUAUUGAACAGAGG	72	bte-miR-216b-5p	UAAUAUCAGCUGGUAAUCCUGAG	9	31	mle-miR-216b-5p
			bte-miR-216b-3p	CAGGAAACUGGUUGGUAUU	45	63	mle-miR-216b-3p
bte-miR-87-1	CAUGUGCAAGGUGGGUGUGCCUUAAGCCUUGACUCAGGCCUGGUUUUGUGUAAAGGUGAGCAAAGUUUCAGGUGUAUUCACUUUGCACUAA	91	bte-miR-87-1-3p	GUGAGCAAAGUUUCAGGUGUAU	56	77	lgi-miR-87
			bte-miR-87-1-5p	GUGCCUUAAGCCUUGACUCAGGCCU	17	41	mle-miR-87-5p

bte-miR-8433	UACAUUGUUUAGUAUCGAAGACCGUGUAAUUAAAGUUAGUUUUGGAAAUAAAAUACCAUACUUUUGAUUCACUUUUCGGUACAAAAAAUAAU	92	bte-miR-8433-5p	UAGUUUUGGAAAUAAAAUA	37	55	sma-miR-8433-5p
bte-miR-9-2	CCUCUCAAUGUAGUUCAUAUGAACAUUACAAUUCUUUGGUUUUCUUUAGAUUGAGUGUGAAGGGAAGAUUUUAAUGAUUUGAAUUUCACGUUGAGCCC	98	bte-miR-9-2-5p	UCUUUGGUUUUCUUUAGAU	33	51	gsa-miR-9-5p
bte-miR-8423	UAAAUAUCAAAUGAAAAAAGGUUGUACCAGGUGGGAGGGGCGAUACAUGCAAUCGCGUUCUGCCCAUCGGACAAACCAAUACUUUUUCUGUUUGUAUUUU	100	bte-miR-8423-3p	UUUUUCUGUUUGUAUUUU	83	100	sma-miR-8423-5p
bte-miR-190-2	CUUUUGUCAUGAUAUGUUUGAUAUUUGGUUGUUGGAAUCUCUCAUUCAACCGAAUUUCAUUCAUAUCUGCACAAUUU	77	bte-miR-190-2-5p	UGAUAUGUUUGAUAUUUGGUUGU	10	32	bbe-miR-190-5p
bte-miR-2001	UUUAAAAAUCAACAUGUGCUCCCGUUAUAAUGGGCAUUGCAGGUCUAUCUCAACAUUGUGACCGUUAUAAUGGGCAUUGCAGGUCUUGCUUGUUGCC	97	bte-miR-2001-3p	UUGUGACCGUUAUAAUGGGCAUU	56	78	lgi-miR-2001
bte-miR-87-2	UGAGCACAAGGUGGGUGUGCCUGAAAUUUGGCUCAGGCCUAGCUAAGACGAAGGUGAGCAAAGUUUCAGGUGUAUUCACUUUGUGCAUU	89	bte-miR-87-2-3p	GUGAGCAAAGUUUCAGGUGUAU	54	75	lgi-miR-87
bte-miR-315-1	CUCUACGUUUUGAUUGUUGCUCAGAGAGUCUUCGCUGAAAAAUGACUCUUUGACAAUGGUCAAAAUCUUAAAC	73	bte-miR-315-1-5p	UUUUGAUUGUUGCUCAGAGAGUCU	8	31	pca-miR-315-5p
bte-miR-317-3	GCUAAUAGCUGGUGGUGAACACAGCUGAUGGUGAACACAGCUGGUGGUGAGCACAGCCUCUUUUG	65	bte-miR-317-3-3p	GUGAACACAGCUGGUGGU	31	48	bmo-miR-317-3p
			bte-miR-317-3-5p	GUGAACACAGCUGAUGGU	15	32	bmo-miR-317-3p
bte-miR-9419	CCCUUUACAGACAUGAAGGAAAUCAAUGACGCCACACAUGGGGUACAUUUUCUUCUUGAAGAUGCUUGACCUCCGCCUGUUGUAAACUA	89	bte-miR-9419-3p	UUCUUGAAGAUGCUUGACCU	53	72	nve-miR-9419
bte-miR-4160	AUGGAGCCUAAAGUCCUGAGUCCUGCUGUAAGUGUCCGUCAGGAAGAUUGAAACAAACAAUGAGGCCAGGACUUUCAAUUUUAGGCUCAAA	91	bte-miR-4160-3p	GAAGAUUGAAACAAACAAUG	43	62	cin-miR-4160-3p
bte-miR-1984	GCUCGUUGUGCCCUAUCCGUCAGGAACUGUGGAUACUCCUCCGUAAGAUCACAGUGUCUGUCGGAGCGGCCAUAACGUUC	80	bte-miR-1984-5p	UGCCCUAUCCGUCAGGAACUGUG	9	31	lgi-miR-1984
			bte-miR-1984-3p	AGUGUCUGUCGGAGCGGCCA	53	72	mle-miR-1984-3p

bte-miR-1985	GUCAGAAGGACAGUUAUGCCAUUUUUAUCAGUCACUGUGUCAUGACUAAAUCGCAGGGCUGGUGACAAUGGCUUGAUAACCUUCUAUU	88	bte-miR-1985-5p	UGCCAUUUUUAUCAGUCACUGUG	17	39	lgi-miR-1985
bte-miR-1994a	UCAGUGUUUGAUUCAAGGGAAGCCAUUCUGUCAGCAUGUUACAGAUUUAUCAUGAGACAGUGUGUCCUCCCUUGCUUCACCACCAA	86	bte-miR-1994a-3p	UGAGACAGUGUGUCCUCCCUUG	53	74	mle-miR-1994a-3p
bte-miR-4164	GACCUCCAGCUGUCUCGUUGUGAGGCCGCAUACAACCAGCUGCUCUUUGUUAUGUCAGCUAAGGUAGUUGGGGCCUCAGAGAAACAGUUGAAGAGC	96	bte-miR-4164-3p	UCUUUGUUAUGUCAGCUAAG	44	63	cin-miR-4164-3p
bte-miR-137-1	UAUUGCUUAAGAAUACGGAAAGAGCUUCAAUUUUGGGGCUCUUUUUUUAUUAAAUCAAAUCCUACCAAUAUUUUACAGAUUUGGCCGUUUCAUAACAGUG	100	bte-miR-137-1-5p	UAUUGCUUAAGAAUACGGAA	1	20	ocu-miR-137-3p
bte-miR-7a	AUUUUCGGCAUUUCAGUACAAGAAGUUGCAUCUUGAGUUUCUCCAGUGUCUUUAAGCGCUUAGGAAACAUCAACUGAUUCUAGUCAAAUGUCGAAUUU	98	bte-miR-7a-3p	UCAACUGAUUCUAGUCAAAU	70	89	sme-miR-7a-3p
bte-miR-317-2	AUGGAACUGCCCCUGUCUAGGGUACUACGUGUGUUGCACAGUGAGUAAUGUUGUGAACACAGCUGGUGGUAUCUUAUCUAGAGGUGAGUUCAAG	94	bte-miR-317-2-3p	UGAACACAGCUGGUGGUAUCUU	54	75	lgi-miR-317
bte-miR-317-1	UGAGUUGGCACUGACUAAGGUAUCAUCCUGUGUUCACAGUAAGUACUGACCAUGUGAACACAGCUGGUGGUAUCUUACUUUGUGGUCAACAGU	93	bte-miR-317-1-3p	UGAACACAGCUGGUGGUAUCUU	55	76	lgi-miR-317
			bte-miR-317-1-5p	AAGGUAUCAUCCUGUGUUCACA	17	38	mle-miR-317-5p
bte-miR-281-1	UAAAGGAAGAACAAUUCGAUGAUCAGCUAGUUGAAUGAAAAAGAACGCUGUCAUGGAAUUGCUCUCCUCGC	71	bte-miR-281-1-3p	UGUCAUGGAAUUGCUCUC	49	66	pte-miR-281-3p
bte-miR-8335-1	UGUUGGUGUUUUCUUGUUGUUGUUGUUGUUGUUGUUUUGAUGAAGGAUGUCACCAGCACAGGAAAAUACCAUUA	74	bte-miR-8335-1-5p	GUUGUUGUUGUUGUUUUG	22	39	ppc-miR-8335-3p
bte-miR-1992	CGGACAUACGGUCAGUCAGUGGUCGAUUGCUUGGAUAGUGGCUAUAAUGCUAUCAGCAGUUGUACCACUGAUUUGAUGGUGUGUUUU	87	bte-miR-1992-3p	UCAGCAGUUGUACCACUGAUUUG	53	75	mle-miR-1992-3p
			bte-miR-1992-5p	AGUCAGUGGUCGAUUGCU	14	31	mle-miR-1992-5p
bte-miR-219	GCAGUCCAAGUAAUCUGGCUAUGAUUGUCCAAACGCAAUUCUUGUUCCUGAUCACACCAAGAACUGUGUGUGGACAUCAGUGCUUGACACUUGACCUU	98	bte-miR-219-5p	UGAUUGUCCAAACGCAAUUCUUG	22	44	mle-miR-219-5p

			bte-miR-219-3p	AGAACUGUGUGUGGACAUCAGU	60	81	mle-miR-219-3p
bte-miR-9-3	CGUUCUUGUCUUUGGUUAUCUAGCUGUAUGAUUGUAGAACAAAUUCAUAAAGCUAGGUUACCAAAGGCAAGACUG	75	bte-miR-9-3-3p	AUAAAGCUAGGUUACCAAAGGCA	47	69	mle-miR-9-3p
			bte-miR-9-3-5p	UCUUUGGUUAUCUAGCUGUAUGA	9	31	mle-miR-9-5p
bte-miR-4968	AACAGCAGCAACAGCAGCAGCAGCAGCUGCAACACAUGUCCAUGAAAAGCCAACAGGACAGCAACAUUGUGGAGCAGCUCAAGGCCAUUGCUGCUCAG	98	bte-miR-4968-5p	CAGCAACAGCAGCAGCAGCAG	6	26	dme-miR-4968-3p
bte-miR-71-2	CUAGUUCUCACAUGUUGUGAAAGACAUGGGUAGUGAGAUGUUAUAUUCACUCACCUUACUACUCUGUCUUUUGCAUUAAGUAGAACAUU	89	bte-miR-71-2-5p	UGAAAGACAUGGGUAGUGAGAUG	18	40	mle-miR-71-5p
			bte-miR-71-2-3p	CCUUACUACUCUGUCUUUUGCA	54	75	mle-miR-71-3p
bte-miR-2b-1	GUUGUCAAACCAACAUCGUCAAUGCUGGAUGUUAUAGUAAUUCAUUAAGGCCCUAUCACAGCCAGCUUUGAUGAGCGUUGUUCUGACUGG	90	bte-miR-2b-1-3p	UAUCACAGCCAGCUUUGAUGAGCG	54	77	pte-miR-2b-3p
			bte-miR-2b-1-5p	UCAAUGCUGGAUGUUAUAGU	19	38	mle-miR-2b-5p
bte-miR-2a-1	AGUAUCACCUUUGUGGCACGUCAGAGUGACUGUGAUUUGUAGAGAAUUUUCAUAUCACAGCCAGCUUUGAUGAGCUUCAAGAUUGAUUUA	90	bte-miR-2a-1-3p	UAUCACAGCCAGCUUUGAUGAGCU	53	76	bdo-miR-2a
			bte-miR-2a-1-5p	CGUCAGAGUGACUGUGAUUUG	19	39	asu-miR-2a-5p
bte-miR-2c-2	ACCAGGCUACCCUGACGUCAAGGCGGUUGUGAUGUGGUAACACUAAACACAUAUCACAGCCAGCUUUGAUGACAGGGAUACCUAUA	86	bte-miR-2c-2-3p	UAUCACAGCCAGCUUUGAUGACA	52	74	mle-miR-2c-3p
			bte-miR-2c-2-5p	ACGUCAAGGCGGUUGUGAUGUG	15	36	mle-miR-2c-5p

bte-miR-277a	CCACCAGCUAGUUUACAGUAGACAAAUUGCUCAAAGUACCAAUUAGUUGAGCCAUUUCUUUGUGUUAUUACUGGCUA	77	bte-miR-277a-3p	GAGCCAUUUCUUUGUGUU	49	66	gsa-miR-277a-5p
bte-miR-981	UAGGCGUUUACUCGGGUUUCGCGAUGUGCGAACAGUCCUCUGCACUGUGUUCGUUGUCGUCGAAACCUGCCUUAAACGCUGU	82	bte-miR-981-3p	UUCGUUGUCGUCGAAACCUGCCU	50	72	mle-miR-981-3p
			bte-miR-981-5p	CGGGUUUCGCGAUGUGCGAACA	13	34	mle-miR-981-5p
bte-miR-750-1	UCUCCUUGCGUUGGAGGAUUAGAUCUUAGCAGUAUUGACUUUGGUGCCAGAUCUAACUCUUCCAGCUCAUGGCGU	75	bte-miR-750-1-3p	CCAGAUCUAACUCUUCCAGCUCA	47	69	mle-miR-750-3p
			bte-miR-750-1-5p	CGUUGGAGGAUUAGAUCUUAGCA	9	31	mle-miR-750-5p
bte-miR-2c-1	CUUUUCCGCCAAGUGGCUGUGAGUAUGUGUAUUUUAAUUGUCAAAUCAUAUCACAGCCAGCUUUGGUGGAAUGG	74	bte-miR-2c-1-3p	UAUCACAGCCAGCUUUGGUGG	49	69	dya-miR-2c
bte-miR-124	CGGCCAGAUUUAGAACACCUGUUCGCGUUCACUGGGUCAGCCUUGGUGUACAACACUAAUUAAGGCACGCGGUGAAUGCCAAGGGGCUUGUUCGAAAUGCA	101	bte-miR-124-3p	UUAAGGCACGCGGUGAAUGCCAA	60	82	ocu-miR-124-3p
			bte-miR-124-5p	GUUCACUGGGUCAGCCUUGGU	27	47	mle-miR-124-5p
bte-miR-315-2	UGACAGAAAGGUUGGCUCAUUUUUUGAUUGUUGCUCAGAAAGCCGGUGAUAAAAGUUCGGCGCUCGGGCUUCAAUCAAACACUGACGCUGGCUAAUAUGCU	101	bte-miR-315-2-5p	UUUUGAUUGUUGCUCAGAAAGCC	22	44	mle-miR-315-5p
			bte-miR-315-2-3p	CGCUCGGGCUUCAAUCAAACA	61	81	mle-miR-315-3p
bte-miR-1	ACGACAGACUUGGAGAGAGACCGUUAGCGCAUGUAACAUACUUCUUUGCUAUCCCAUAUUACAAUCAAGCUAUGGAAUGUAAAGAAGUAUGUAUCAUACGUUAGCAUUCUACAACCAGAUCAAACAUGU	129	bte-miR-1-3p	UGGAAUGUAAAGAAGUAUGUAU	73	94	mle-miR-1-3p
			bte-miR-1-5p	ACAUACUUCUUUGCUAUCCCAUA	36	58	mle-miR-1-5p

bte-miR-317-4	ACUACGUGUGUUGCACGGUGAGUAAUGUUGUGAACACAGCUGGUGGUAUCUUACUUUGUGGUCAACAGUCCAGCUGCAACGACCGUUUC	89	bte-miR-317-4-5p	UGAACACAGCUGGUGGUAUCUU	31	52	lgi-miR-317
bte-miR-317-5	AUGGAACUGCCCCUGUCUAGGGUACUACGUGUGUUGCACGGUGAGUAAUGUUGUGAACACAGCUGGUGGUAUCUUAUCUAGAGGUGAGUUCAAG	94	bte-miR-317-5-3p	UGAACACAGCUGGUGGUAUCUU	54	75	lgi-miR-317
bte-miR-2580	UAUCAUAAAUUAAUCUGUGUAGCAUGAGCGUGUCUGUGUGUAUGUACGAGGCCACAUUCUAUAUUGCUGUCUCAUCUACUCUAGAUUGUUAUGGCU	96	bte-miR-2580-5p	CGUGUCUGUGUGUAUGUACG	29	48	dsi-miR-2580
bte-miR-6302-1	GUGAAGUGAUUAAUAAACAAUGCCAGCCCCUUCAGUCUAGACUGACGGGGGGCUUAAGAGACUUUAUUUUUAAUCGCUUUGA	82	bte-miR-6302-1-3p	AGACUGACGGGGGGCUUAAG	39	58	hme-miR-6302-1-5p
bte-miR-8335-2	UUUUAUUUCCCAAUAGAUUUAAGUGUUUUUUAUGCUCCAUACACAGUUGUUGUUGUUGUUUUGGAGCAUAAAAAACACUUAAAUAUUUGAAAAAUAAAA	99	bte-miR-8335-2-3p	GUUGUUGUUGUUGUUUUG	46	63	ppc-miR-8335-3p
bte-miR-3350	AGGAGGAACUAGAAAGAGUAUACUUAGAGUUUGAAUACUUUGAUCUAUCAUCAUCGUCAUGUAUUCUAAUUUUAGGUUCACUUGUUGCUUUUUCCUUCU	99	bte-miR-3350-3p	UCAUGUAUUCUAAUUUUAGG	57	76	bmo-miR-3350
bte-miR-7952	GACCAGUAGCUAGAGUGCAUUUUUGUUAAAUUAAUAAAAAGUGGAUGAUUCUACAGAGCACAAAUAAAAAUGUGCUUUUGCAUAAUGAGA	90	bte-miR-7952-5p	UUAAUAAAAAGUGGAUGA	31	48	prd-miR-7952-3p
bte-bantam	ACAACAAAACUGGUUUUCCCAUUGGUCUGGCAGAUAUUUCGUAUUCUGAGAUCAUUGUGAAAACUGAUUUUGUAC	75	bte-bantam-3p	UGAGAUCAUUGUGAAAACUGAUU	47	69	bgl-bantam-3p
bte-miR-92	GAAUGUCCAUGUCAUGUAGGUUGUGAUGUUUGCAAUCAUUGGGUAAACAUGUUUUGACCAAUUGCACUCGUCCCGGCCCUGCUUGAUAUGAUGAUUAAAG	99	bte-miR-92-3p	AAUUGCACUCGUCCCGGCCUGC	60	81	bgl-miR-92-3p
bte-miR-92a-2	GGCACAGAAGAAAGGUCAUGAAAUGUGCAAUGUGUUCCUUGUACAAAUAAACUCAUAUUGCACUUUUCCAGGCCUUUUUUUACUGUCA	88	bte-miR-92a-2-3p	UAUUGCACUUUUCCAGGCCUUU	56	77	bgl-miR-92a-3p
bte-miR-92a-1	AGUUUCUACCUUUAUGUCAUCAGGUUGUGAUAAGUGCAGUCUGUCUAACAAGUUCACAUUGCACUUUUCCCGGCCUGUUGCUGUAAAGUCAUAACA	96	bte-miR-92a-1-3p	CAUUGCACUUUUCCCGGCCUGU	57	78	bgl-miR-92a-1-3p
bte-miR-96a	UGUUCCUGUCCUUGGCACUGGCGGAAUAGUCACGGGCAUCAGAAAGUGAUUGUUUCCUGGUGUCAAAAACAGUGG	75	bte-miR-96a-5p	CUUGGCACUGGCGGAAUAGUCA	11	32	bgl-miR-96a-5p

bte-miR-125	AAGAGAUUGCCCCUACCUUUUCCCUGAGACCAUAAUUUGUGCCUCCGAGCCCAGGUUGUGGCCUUAGGAACUAGGAAGUGUGGCAUCAAA	90	bte-miR-125-5p	UCCCUGAGACCAUAAUUUGUGCC	21	43	bgl-miR-125-5p
bte-miR-125b-2	GUGUUGGUCAUCAGCUCUGGGUCUGGAAUGAGUUGUCCACAAGUCACGCUCUUGGGACUAAUCAUGCCAGAGAGGGUCAGGUGUUGACCAGUG	93	bte-miR-125b-2-3p	ACAAGUCACGCUCUUGGG	39	56	bgl-miR-125b-3p
bte-miR-1991	GUAGCCGUGUCGUGAGCUUACCCUGUUAUACUGAGAAGUUGAGAACCUCUCUGACAGAUAUUUCUCGUUUCACAGGGUAUUUUCACUGCACGGUGA	96	bte-miR-1991-5p	CUUACCCUGUUAUACUGAGAAGU	17	39	bgl-miR-1991-5p
bte-miR-1993	GAAACUUUCGGGAAUAAUGGCAUGAUGCAGUGCAAUGACUUGUGUAUUAUGCUGCUAUUCACGAGAUCAA	70	bte-miR-1993-3p	UAUUAUGCUGCUAUUCACGAGA	45	66	bgl-miR-1993-3p
bte-miR-1994b	GGAGUGAUUGCUCUCAGGGAUGGCACUCUGUCUACUUGUGUUUUUGACCAUGAGACAGUGCGUCCUCCCUCAGAUUCAAAUCACCA	86	bte-miR-1994b-3p	UGAGACAGUGCGUCCUCCCUCA	51	72	bgl-miR-1994b-3p
bte-miR-1a	CAGACUUGGAGAGAGACCGUUAGCGCAUGUAACAUACUUCUUUGCUAUCCCAUAUUACAAUCAAGCUAUGGAAUGUAAAGAAGUAUGUAUCAUACGUUAGCAUUCUA	107	bte-miR-1a-3p	UGGAAUGUAAAGAAGUAUGUAU	69	90	bgl-miR-1a-3p
bte-miR-216a	GGCGUUUGCAUAAUCUCAGCUGGUAAUUCAGAGUGAUUUCAUAGAGUCUCAAGUUACCAGCCGAGGUUACACAAAUAC	78	bte-miR-216a-5p	UAAUCUCAGCUGGUAAUUCAGAG	11	33	bgl-miR-216a-5p
bte-miR-219b	GAAGGUCAAGUGUCAAGCACUGAUGUCCACACACAGUUCUUGGUGUGAUCAGGAACAAGAAUUGCGUUUGGACAAUCAUAGCCAGAUUACUUGGACUGC	99	bte-miR-219b-3p	AGAAUUGCGUUUGGACAAUCAUA	58	80	bgl-miR-219b-3p
bte-miR-36a	GGCAGUUGUGCAUGGUGGGUGUUUACCAAGGUGGUGCAGUUUAUCUCCCUUGCCAUCACCGGGUAUACAUUCAUCCGCGUACGGCUUGCAA	91	bte-miR-36a-3p	UCACCGGGUAUACAUUCAUCCG	56	77	bgl-miR-36a-3p
bte-miR-7620a-2	UGUCAUACUGUAGGACUUUUUUGAGAACGUCCUAUUGUAGGACUUUUUUUUUAGAUCGUCAUAUUGUAGGACUUAUUUUUAGAUCGUCCUAUUUUAGGGC	100	bte-miR-7620a-2-3p	UAUUGUAGGACUUAUUUUUA	62	81	bgl-miR-7620a-2-3p
			bte-miR-7620a-2-5p	UAUUGUAGGACUUUUUUUU	33	51	bgl-miR-7620a-2-3p
bte-miR-7620a-1	CUUUUUUGAGAACGUCCUAUUGUAGGACUUUUUUUUUAGAUCGUCAUAUUGUAGGACUUAUUUUUAGAUCGUCCUAUUUUAGGGCUGAUUUUCAGAAUG	99	bte-miR-7620a-1-3p	UAUUGUAGGACUUAUUUUUA	47	66	bgl-miR-7620a-2-3p

			bte-miR-7620a-1-5p	UAUUGUAGGACUUUUUUUU	18	36	bgl-miR-7620a-2-3p
bte-miR-9b	CCAGUCUUGCCUUUGGUAACCUAGCUUUAUGAAUUUGUUCUACAAUCAUACAGCUAGAUAACCAAAGACAAGAACG	76	bte-miR-9b-5p	CCUUUGGUAACCUAGCUUUAUGA	10	32	bgl-miR-9b-5p
bte-miR-33	GAAAGGAUACUGUUAGUGGUGCAUUGUAGUUGCAUUGCGUGUCAAGCUUGUGCAAUGUCUCUGCAGUGCAAUCACCAAUGUAUUCUCAU	89	bte-miR-33-5p	GUGCAUUGUAGUUGCAUUGCGUG	19	41	bgl-miR-33-5p
bte-miR-133	AGCCAUGCUAGUCUGGAGCUGGUUGAAUCUGGGCCAAAUUGGAUUGUUGAAUUGCAUUUGGUCCCCUUCAAUCAGUUGUAGCUAGCUGACA	91	bte-miR-133-5p	AGCUGGUUGAAUCUGGGCCAAAU	17	39	bgl-miR-133-5p
			bte-miR-133-3p	UUUGGUCCCCUUCAAUCAGUUG	57	78	bgl-miR-133-3p
bte-miR-87b-2	CAUGUGCAAGGUGGGUGUGCCUUAAGCCUUGACUCAGGCCUGGUUUUGUGUAAAGGUGAGCAAAGUUUCAGGUGUAUUCACUUUGCACUAA	91	bte-miR-87b-2-3p	GUGAGCAAAGUUUCAGGUGUAU	56	77	bgl-miR-87b-2-3p
bte-miR-219a	GCAGUCCAAGUAAUCUGGCUAUGAUUGUCCAAACGCAAUUCUUGUUCCUGAUCACACCAAGAACUGUGUGUGGACAUCAGUGCUUGACACUUGACCUU	98	bte-miR-219a-5p	UGAUUGUCCAAACGCAAUUCUUG	22	44	bgl-miR-219a-5p
			bte-miR-219a-3p	AGAACUGUGUGUGGACAUCA	60	79	bgl-miR-219a-3p
bte-miR-72	AUAGAGUUCAGACCAGGCAAGAUGUUGGCAUAGCUGAGUCAAACAACUGUCAGCUGUGUCAUAUGUUGCCAAUUUUGGACUCGUU	85	bte-miR-72-5p	AGGCAAGAUGUUGGCAUAGCUGA	15	37	bgl-miR-72-5p
bte-miR-87b-1	UGAGCACAAGGUGGGUGUGCCUGAAAUUUGGCUCAGGCCUAGCUAAGACGAAGGUGAGCAAAGUUUCAGGUGUAUUCACUUUGUGCAUU	89	bte-miR-87b-1-3p	GUGAGCAAAGUUUCAGGUGUAU	54	75	bgl-miR-87b-2-3p
bte-miR-193	GUGCCUAACUCACUGGGGAUUUUGGUUGCCAACUAUGUCCUGUAAAAUGCAUACUGGCCUUCAAAAUCCCAAACGAGGAGGUCU	84	bte-miR-193-3p	UACUGGCCUUCAAAAUCCCAAA	52	73	bgl-miR-193-3p
bte-miR-92b	UAUUGUCACUGCCAUGAGGUUGUGAUUAGAGCAAUCUGUUGUUAUACAAACCAAAUUGCACUAAUCCCGGCCUACAUCCACGUGACAUGA	90	bte-miR-92b-3p	AAUUGCACUAAUCCCGGCCUAC	54	75	bgl-miR-92b-3p

bte-miR-2d	UACAAUGCAUGCUGCACUGACCAGGUGGUUGGGAUGUGCUUUUUUUACUACAUAUCACAGCCUGCUUGGAUCAGUACUGCGUCAUUUUU	89	bte-miR-2d-3p	UAUCACAGCCUGCUUGGAUCAGU	53	75	bgl-miR-2d-3p
bte-miR-9a	CGUUCUUGUCUUUGGUUAUCUAGCUGUAUGAUUGUAGAACAAAUUCAUAAAGCUAGGUUACCAAAGGCAAGACUG	75	bte-miR-9a-5p	UCUUUGGUUAUCUAGCUGUAUGA	9	31	bgl-miR-9a-5p
			bte-miR-9a-3p	AUAAAGCUAGGUUACCAAAGGC	47	68	bgl-miR-9a-3p
bte-mir-2c-1	GAAGGAAAGGUUCUAGUUUCUAUAACAUCACAGUGAUUGUGAUGUGUGAAAUUGGUCAUAUCACAGUCAGCUUUGAUGAGCUUGGAGGCUAGCAAUUUC	101	bte-mir-2c-1-3p	UAUCACAGUCAGCUUUGAUGAGC	59	80	lgi-mir-2c

S3 Table: Quantity and quality of sequenced reads.

Sample	ID	Total reads	Phred 30	Reads 18-30	Nt percentage after Phred 30	Nt percentage between 18-30
B. tenagophila Taim exposed replicate 1	BTA_exposed_sample1	45407622	45407621	16765705	1	36,92%
B. tenagophila Taim exposed replicate 2	BTA_exposed_sample2	26712682	26712678	12147204	1	45,47%
B. tenagophila Taim exposed replicate 3	BTA_exposed_sample3	32043600	32043593	13396035	1	41,81%
B. tenagophila Taim unexposed replicate 1	BTA_unexposed_sample1	44861950	44861944	18951582	1	42,24%
B. tenagophila Taim unexposed replicate 2	BTA_unexposed_sample2	33768148	33768144	14275658	1	42,28%
B. tenagophila Taim unexposed replicate 3	BTA_unexposed_sample3	27420733	27420729	12281722	1	44,79%
B. tenagophila HM exposed replicate 1	BHM_exposed_sample1	25454139	25454137	10459644	1	41,09%
B. tenagophila HM exposed replicate 2	BHM_exposed_sample2	27891097	27891092	14554099	1	52,18%
B. tenagophila HM exposed replicate 3	BHM_exposed_sample3	32728489	32728481	16856946	1	51,51%

B. tenagophila HM unexposed replicate 1	BHM_unexposed_sample1	15587360	15587355	7362539	1	47,23%
B. tenagophila HM unexposed replicate 2	BHM_unexposed_sample2	22575223	22575222	13003278	1	57,60%
B. tenagophila HM unexposed replicate 3	BHM_unexposed_sample3	22617505	22617504	13534890	1	59,84%

S4 Table – Representation of mature miRNA counts detected in small-Seq samples.

miRNA	BHM_exposed_sample1	BHM_exposed_sample2	BHM_exposed_sample3	BHM_unexposed_sample1	BHM_unexposed_sample2	BHM_unexposed_sample3	BTA_exposed_sample1	BTA_exposed_sample2	BTA_exposed_sample3	BTA_unexposed_sample1	BTA_unexposed_sample2	BTA_unexposed_sample3
bte-miR-1-3p_bte-miR-1-a-3p	305669	653051	848591	333093	721458	740535	372063	156093	367501	644642	529620	472380
bte-miR-184-3p	262409	361910	344161	204251	361331	331038	818171	2145545	535155	518915	573022	544738
bte-miR-315-2-5p	176983	301900	320279	125817	290841	257992	292846	199830	250538	322360	316320	311824
bte-miR-750-1-3p	132170	262736	252615	111618	239080	220620	479771	221289	402713	404811	377631	311495
bte-miR-71-1-5p_bte-miR-71-2-5p	102553	210765	215975	97854	204138	197246	219175	76898	192052	197444	193290	173813
bte-miR-1985-5p	99114	232038	294919	78621	211428	215888	175022	44421	146232	178236	169684	145771
bte-miR-1984-5p	93044	185329	233052	60897	185045	180380	140390	74064	130385	155127	131845	114667
bte-miR-279-3p	89330	256547	304488	76786	147427	177507	115008	65946	113993	147104	95428	82089
bte-let-7-l-5p	62475	140655	164484	56730	116022	122626	155650	75734	187867	191587	156587	125968
bte-miR-125-5p	49758	148248	161369	46990	103044	107687	112306	20301	114634	133388	106545	92630

bte-bantam-3p	49488	74158	92665	23241	73853	68405	73615	32812	56097	78288	77077	73291
bte-miR-92-3p	48635	83128	96315	47487	79755	78570	70958	75620	83548	86558	82222	72610
bte-miR-375-3p	43057	99453	117788	36216	84816	81888	160775	124745	138628	133232	112788	78477
bte-miR-8-1-3p_bte-miR-8-2-3p	34419	89118	100258	28484	60397	59992	47022	9062	44074	66039	53934	51082
bte-miR-216-b-5p	31673	85795	102343	30997	76546	68721	60638	16019	62717	92096	76160	67714
bte-miR-981-3p	25806	65738	77447	29611	54991	55044	37994	11829	39846	51748	36391	30660
bte-miR-92-a-1-3p	20533	41797	50338	23349	38216	39298	33761	13384	36892	40565	33927	28676
bte-miR-34-5p	19459	55059	53910	22020	45465	39567	46686	9262	54547	64419	42878	32140
bte-miR-278-3p	18803	31862	34564	14569	28714	29155	39739	42479	32634	27380	22732	14881
bte-miR-133-3p	18348	49821	66863	18589	41758	41595	20971	7370	27471	36065	24441	20502
bte-miR-92-b-3p	16986	38643	55479	20480	33823	33780	23448	4263	27327	34378	28766	26269
bte-miR-9-3-5p_bte-miR-9-a-5p	14543	31890	38527	13227	28969	27801	22304	10894	20986	29120	30342	28228
bte-miR-96-a-5p	13548	24784	28017	10543	23305	22938	23323	35808	21800	21714	22318	21147
bte-miR-277-3p	12744	60471	47394	20560	40468	33433	33576	2548	32519	54436	40208	36189
bte-miR-315-1-5p	11964	23353	23540	7365	17943	16445	21059	17285	17391	21481	19614	19721

bte-miR-216-a-5p	10584	29239	26057	11053	23045	19488	16146	4478	15093	23649	19803	18826
bte-miR-190-1-5p	10243	20325	18708	6460	16553	12491	14868	6987	11018	19015	18264	19456
bte-miR-1175-3p	9324	31627	37118	11360	24391	21959	18488	1825	21827	32359	18210	15764
bte-miR-2-a-1-3p	8034	23861	29139	6811	17569	17825	10676	1683	10156	16343	12605	11279
bte-miR-281-2-3p	7216	19747	23915	6355	15243	14584	11901	2401	11539	17599	11516	10788
bte-miR-252-a-5p	6960	21150	24912	6862	14286	15986	15581	2985	15003	14228	11709	9451
bte-miR-1994-b-3p	5092	12044	14248	3683	9639	9561	9237	1152	10330	12836	10723	10237
bte-miR-1994-a-3p	4622	14393	21041	4100	9773	10913	8521	975	9101	11861	7987	7062
bte-miR-2-a-2-3p_bte-miR-2-d-3p	3899	11236	14099	3427	8412	8018	6335	813	6059	8397	5821	5181
bte-miR-317-1-3p_bte-miR-317-2-3p_bte-miR-317-4-5p_bte-miR-317-5-3p_bte-miR-317-3-3p	3603	12147	15733	3321	7705	7573	4918	886	5320	8142	5195	4487
bte-miR-31-5p_bte-miR-72-5p	3559	9638	11478	5029	8399	8415	3066	881	3364	5683	4348	3932
bte-miR-2722-3p	3298	5125	5499	2018	5092	4735	4981	8978	4044	4259	5307	5317

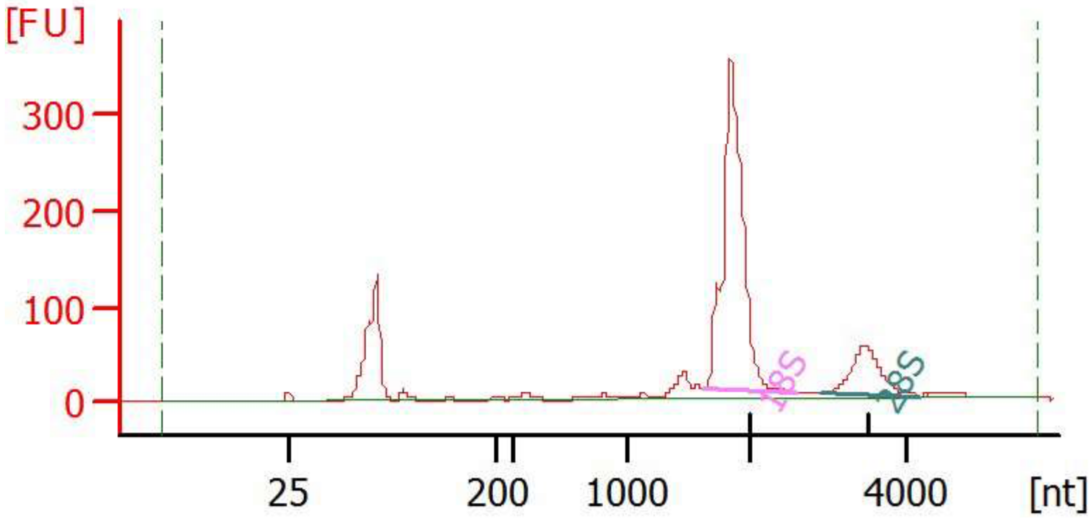
bte-miR-1991-5p	2853	7200	7827	1956	5216	5234	3731	1126	3622	4923	3729	3499
bte-miR-2-b-1-3p	2753	9157	12120	2947	7208	6851	4026	514	3858	6600	4818	4348
bte-miR-1990-3p	2697	5482	6660	3128	5961	5530	4711	3607	4439	5697	5095	4589
bte-miR-29-a-3p	2648	9315	12399	2597	6799	7251	3388	349	4534	6433	3818	2905
bte-miR-252-b-5p	2647	6158	8694	2591	5272	5490	4391	1527	4289	4786	3983	3513
bte-miR-87-1-3p_bte-miR-87-2-3p_bte-miR-87-b-2-1-3p_bte-miR-87-b-2-2-3p	2507	7828	9570	3089	5561	5717	2478	282	2627	4676	2686	2280
bte-miR-96-b-5p	2122	4515	5590	2364	4250	4502	3848	4705	4075	4655	4514	4529
bte-miR-281-2-5p	1963	5065	3957	1400	3020	2577	2731	1778	1991	2850	2089	2283
bte-miR-2-c-2-3p	1478	4981	6759	1285	3613	3320	2232	335	2014	3689	2341	2008
bte-miR-745-a-3p	1254	2935	3571	854	2312	1971	2397	860	2345	3237	2233	2012
bte-miR-745-a-5p	817	1539	1540	556	1751	1432	1122	1169	1282	2099	2115	1714
bte-miR-193-3p	640	1504	1852	435	1139	1106	1038	299	707	1140	872	854
bte-miR-745-b-3p	631	1166	1473	392	877	866	1472	466	1146	1648	1300	1363
bte-miR-1175-5p	595	1351	1419	438	1111	947	1160	172	1058	2036	1592	1586

bte-miR-137-2-3p	536	1341	1341	348	978	814	559	97	454	720	557	532
bte-miR-92-a-2-3p	457	1309	1870	434	1128	997	927	105	751	1254	717	649
bte-miR-2001-3p	396	1262	1416	189	1229	1048	766	388	624	943	609	532
bte-miR-29-b-3p	378	1332	1731	406	874	1025	572	73	689	987	550	480
bte-miR-124-3p	370	1199	1648	314	850	816	279	41	282	488	266	194
bte-miR-12-5p	344	1173	1751	414	848	767	625	100	607	1005	754	616
bte-miR-279-5p	341	725	752	280	590	546	370	740	469	678	601	544
bte-miR-1993-3p	301	666	839	282	814	617	463	206	501	615	562	528
bte-miR-1984-3p	273	702	768	421	824	656	333	316	289	455	530	539
bte-miR-71-2-3p	231	577	654	221	652	550	342	385	404	502	500	418
bte-miR-9-1-3p	203	1057	2209	291	1202	1993	833	139	663	1028	775	940
bte-miR-7-5p	154	453	617	150	317	397	143	51	125	243	207	188
bte-miR-133-5p	139	264	451	185	382	326	171	122	164	269	196	243
bte-miR-153-3p	124	403	379	91	246	285	113	17	80	201	207	192
bte-miR-219-3p_bte-miR-219-a-3p	111	241	296	99	209	199	252	152	171	241	211	216
bte-miR-1-5p	102	232	383	121	299	275	169	41	210	251	205	208

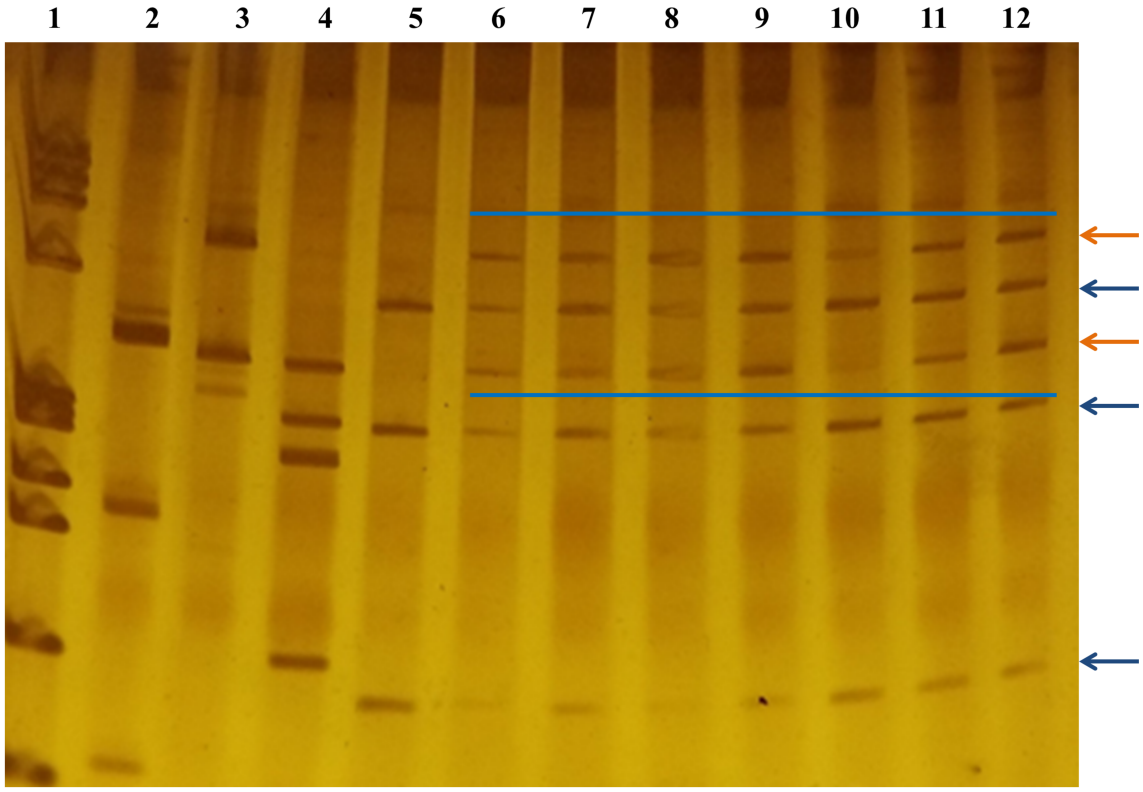
bte-miR-1990-5p	99	271	347	130	240	275	192	40	287	396	276	241
bte-miR-9-3-3p_bte-miR-9-a-3p	93	317	397	127	228	227	137	19	133	301	131	132
bte-miR-8-1-5p	90	144	133	45	130	128	81	106	94	234	187	187
bte-miR-33-a-3p	77	290	421	94	178	222	133	25	126	212	120	109
bte-miR-750-1-5p	64	185	165	75	146	157	89	104	137	254	183	137
bte-miR-184-5p	58	132	142	47	107	101	92	31	96	127	93	100
bte-miR-252-a-3p	57	183	204	46	129	145	132	52	142	163	121	112
bte-miR-2-c-2-5p	56	118	136	39	87	78	90	41	59	119	96	83
bte-miR-315-2-3p	56	110	135	55	82	104	55	87	75	113	81	78
bte-miR-1992-3p	49	92	98	28	78	67	61	17	36	65	49	68
bte-miR-33-5p_bte-miR-33-a-5p	47	92	124	48	64	65	42	6	29	71	68	66
bte-miR-216-b-3p	43	194	199	43	243	153	30	17	60	170	77	90
bte-miR-29-a-5p	43	149	153	50	120	116	90	41	98	129	93	84
bte-miR-281-1-3p	42	63	61	22	40	60	50	16	34	58	32	30
bte-miR-96-b-3p	41	128	198	50	88	90	153	50	107	149	108	118

bte-miR-1986-3p	28	35	65	18	38	20	56	3	42	35	52	22
bte-miR-2-a-2-5p	27	57	76	29	62	61	4	13	24	52	37	35
bte-miR-2-a-1-5p	26	49	73	30	63	63	38	20	27	65	82	72
bte-miR-9-1-5p_bte-miR-9-b-5p	26	70	154	30	77	122	109	36	56	105	67	74
bte-miR-375-5p	21	70	58	9	63	36	22	12	51	67	44	44
bte-miR-277-5p	15	47	50	4	22	19	9	6	15	31	11	22
bte-miR-31-3p	15	31	37	8	16	17	10	1	13	19	13	17
bte-miR-190-1-3p	13	74	66	21	67	58	23	1	24	37	35	16
bte-miR-219-5p_bte-miR-219-a-5p	12	28	57	6	33	30	14	3	9	22	21	14
bte-miR-1992-5p	11	34	26	20	39	33	10	7	15	29	31	39
bte-miR-87-1-5p	11	22	34	3	18	15	16	2	18	27	22	15
bte-miR-2-c-1-3p	7	18	22	6	20	12	12	1	7	21	17	9
bte-miR-12-3p	6	46	66	9	38	30	25	3	6	28	27	24
bte-miR-7-3p	6	30	40	9	26	29	9	0	17	14	21	23
bte-miR-317-1-5p	4	17	17	5	10	2	7	1	7	12	10	4

bte-miR-137-2-5p	3	11	16	3	6	9	1	2	10	7	9	2
bte-miR-1986-5p	3	8	11	0	6	11	8	8	10	21	16	17
bte-miR-190-2-5p	1	3	10	1	6	1	2	0	0	0	4	3
bte-miR-124-5p	0	5	5	1	1	4	1	1	2	1	6	8
bte-miR-153-5p	0	0	1	0	0	0	4	0	0	1	2	0
bte-miR-2-b-1-5p	0	0	0	0	0	0	0	0	0	1	1	1
bte-miR-2722-5p	0	4	5	5	8	1	3	1	2	5	4	0
bte-miR-36-a-3p	0	1	1	3	1	1	0	0	3	0	0	1
bte-miR-7-d-3p	0	0	2	0	0	0	0	0	0	0	0	0
bte-miR-745-b-5p	0	0	1	0	1	1	1	0	1	3	0	0
bte-miR-8335-1-5p_bte-miR-8335-2-3p	0	0	0	0	0	0	0	0	0	0	1	1
bte-miR-981-5p	0	4	4	0	4	1	1	0	1	0	3	1

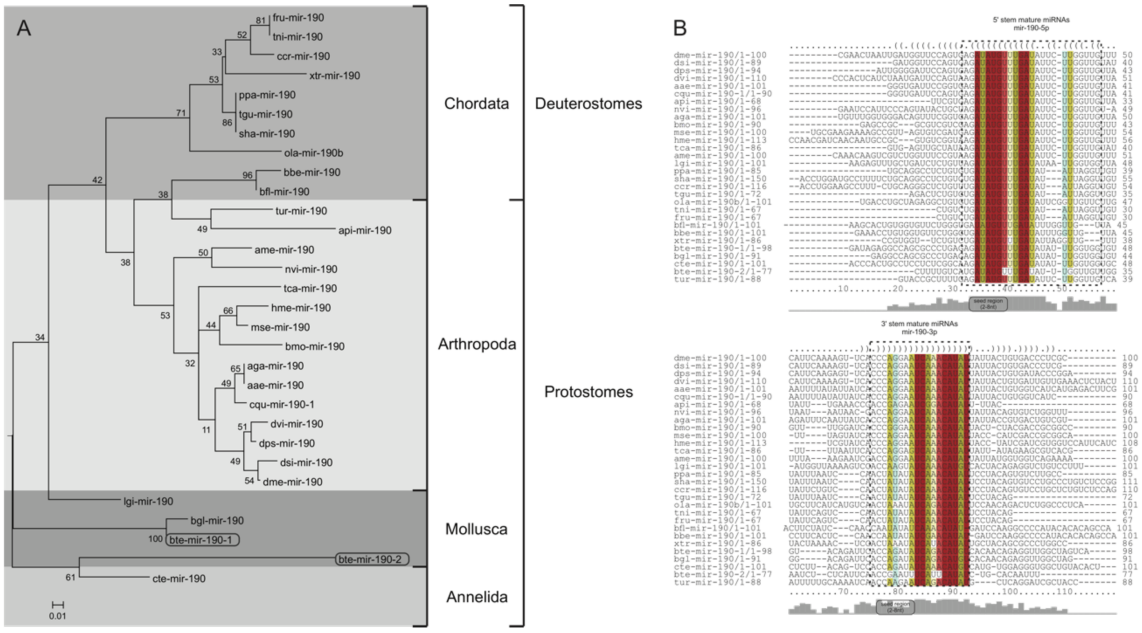


S1 Figure: Electropherogram of total RNA from *Biomphalaria* sp to the Bioanalyzer. There is a peak corresponding to 18S ribosomal RNA and a smaller peak showing traces of 28S RNA. Average RIN of 8.

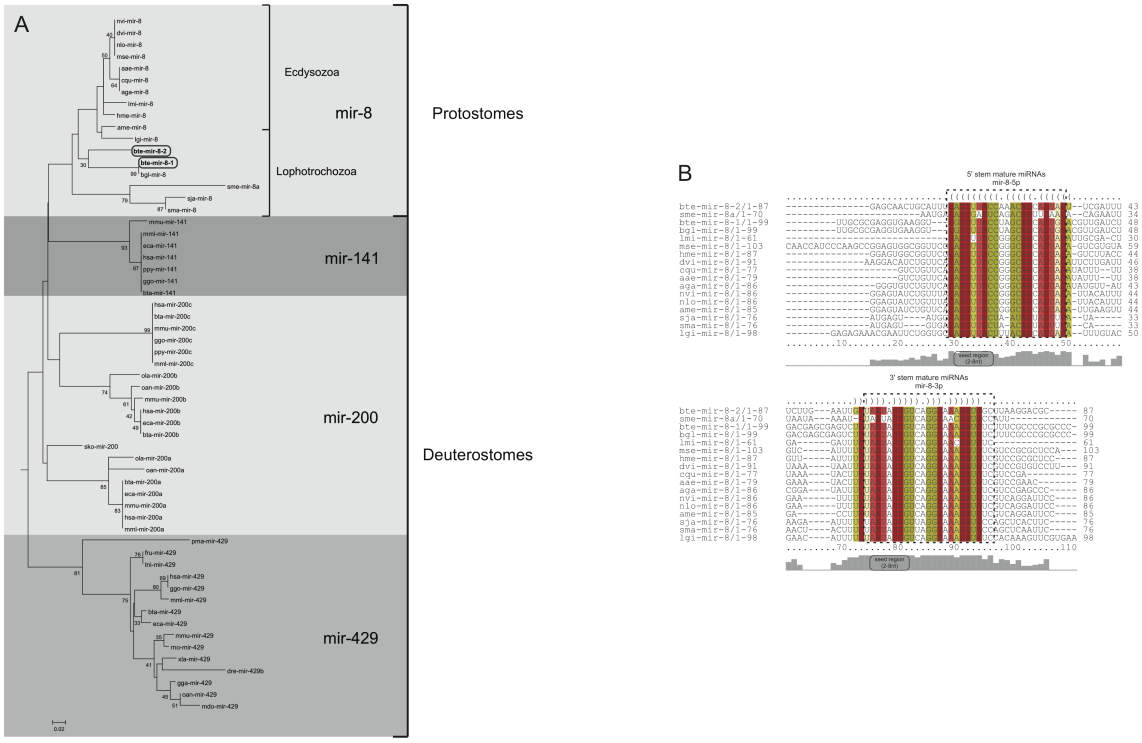


S2 Figure: PCR-RFLP of the ITS region of the DNA cleaved with DdeI Samples of *Biomphalaria tenagophila* from Bananal/SP. Channel 1 represents the molecular weight standard, channel 2 the *Biomphalaria glabrata*,

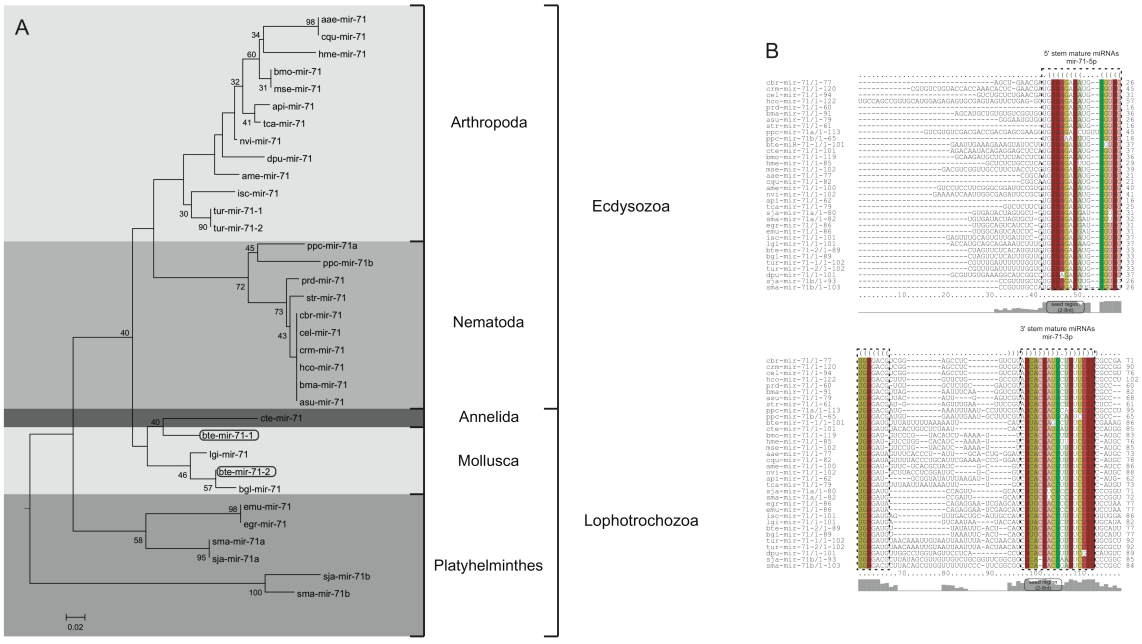
channel 3 the *Biomphalaria tenagophila* (orange arrow), channel 4 the *Biomphalaria straminea*, channel 5 the *Schistosoma mansoni* (blue arrow), channels 6 to 12 represent *B. tenagophila* + *S. mansoni* (samples from Bananal/SP) (blue line).



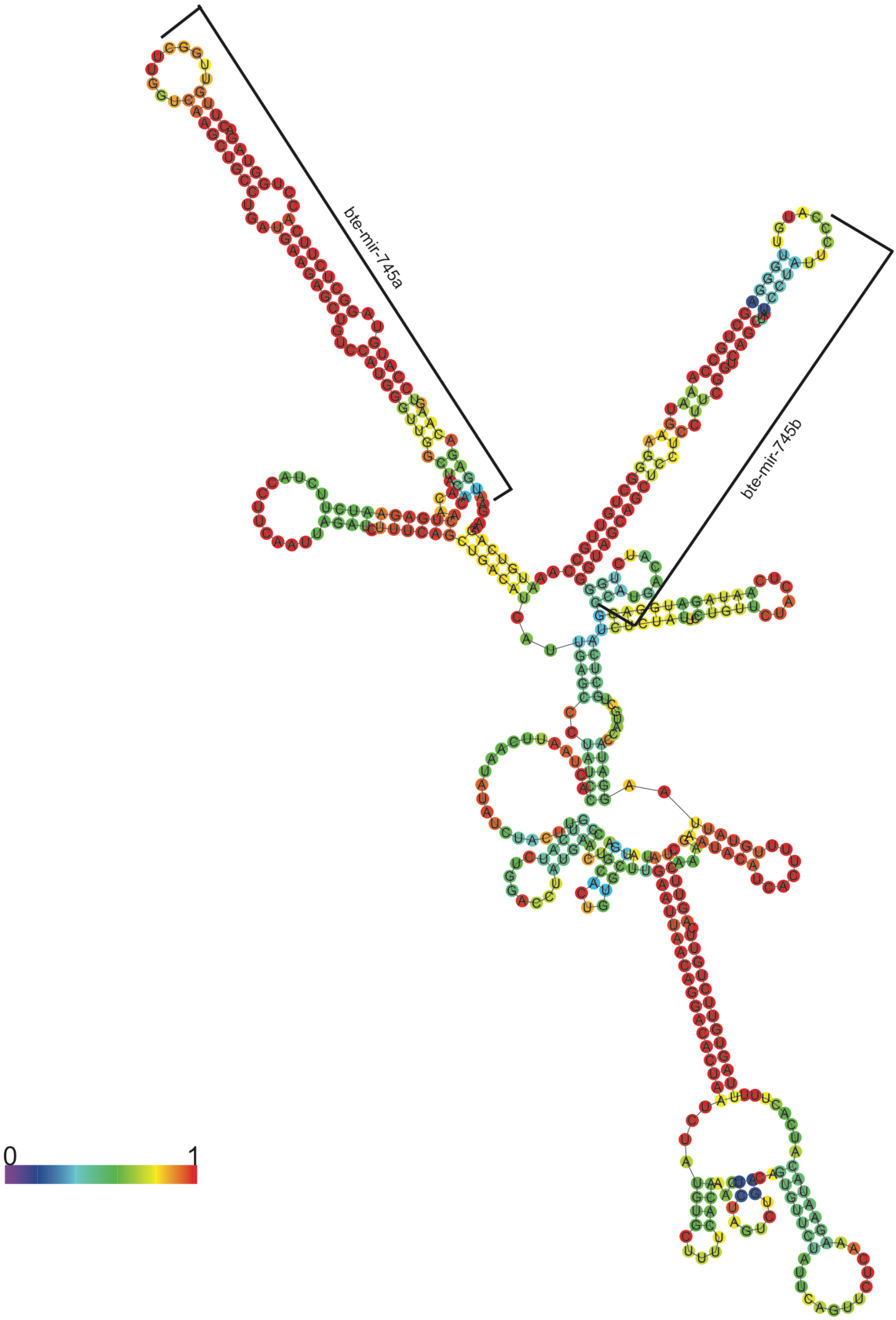
S3 Figure: miR-190 family. (A) Phylogenetic tree. (B) Global alignment among bte-miR-190 and its homologs.



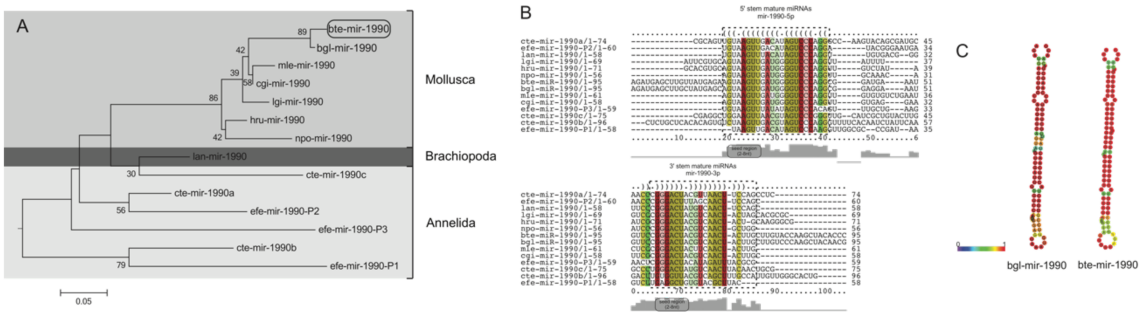
S4 Figure: Family of miR-8, miR-141, miR-200 and miR-429. (A) Phylogenetic tree. (B) Global alignment among bte-miR-8 and its homologs.



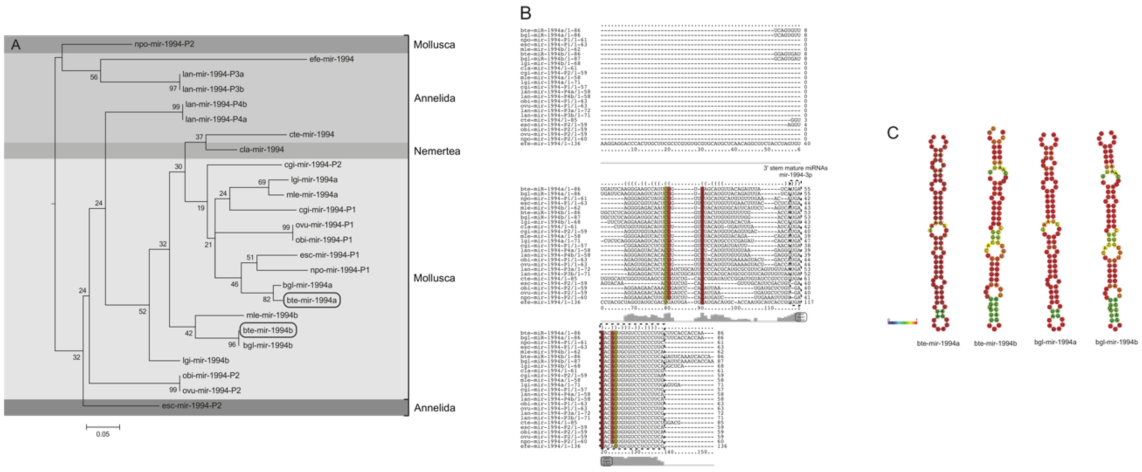
S5 Figure: miR-71 family. (A) Phylogenetic tree. (B) Global alignment among bte-miR-71 and its homologs.



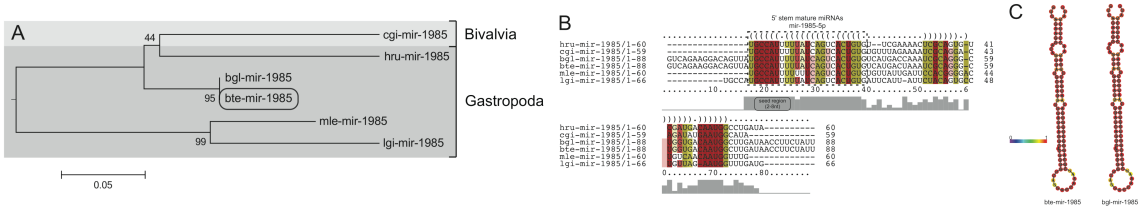
S6 Figure: Cluster of miR-745a and miR-745b.



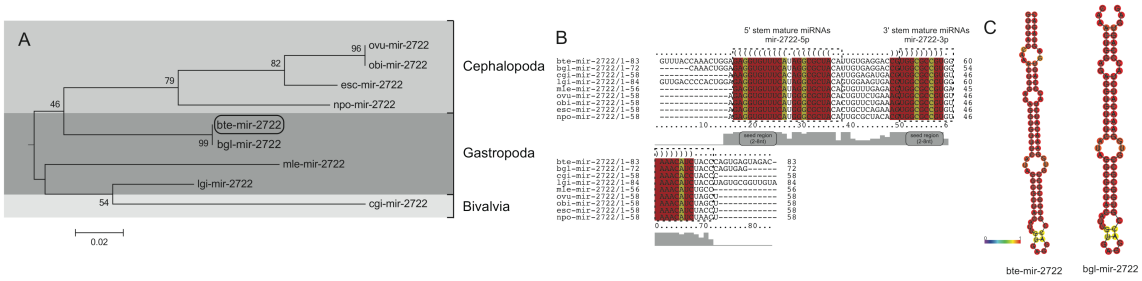
S7 Figure: miR-1990 family. (A) Phylogenetic tree. (B) Global alignment among bte-miR-1990 and its homologs.



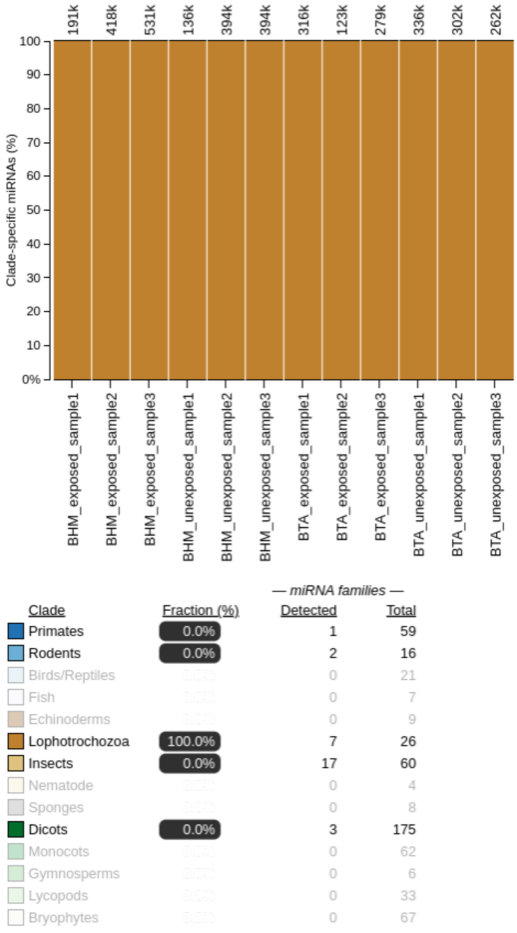
S8 Figure: miR-1994 family. (A) Phylogenetic tree. (B) Global alignment among bte-miR-1994 and its homologs.



S9 Figure: miR-1985 family. (A) Phylogenetic tree. (B) Global alignment among bte-miR-1985 and its homologs.



S10 Figure: miR-2722 family. (A) Phylogenetic tree. (B) Global alignment among bte-miR-2722 and its homologs.



S11 Figure: Result of the Trace mode of the miRTrace tool. 100% of the libraries are Lophotrocozoa group of the species analyzed in this study, *Biomphalaria tenagophila*.

CAPÍTULO III

(Artigo submetido a revista *Cancer Genetics*)

Mitochondrial DNA Variants and OXPHOS Pathway in High-Grade Ovarian Cancer: Challenging the Metabolic Paradigm

Mitochondrial DNA Variants and OXPHOS Pathway in High-Grade Ovarian Cancer: Challenging the Metabolic Paradigm

Tamires Caixeta Alves[1], Ana Paula Alvares da Silva Ramos[2], Antônio Facciolo Filho[1], Fábio Ribeiro Queiroz[3], Higgor Gonçalves Dornelas[4], Agnaldo Lopes da Silva Filho[5], Letícia da Conceição Braga[2,3*], Matheus de Souza Gomes[1*]

[1] Laboratório de Bioinformática e Análises Moleculares, Universidade Federal de Uberlândia, Patos de Minas, MG, Brazil.

[2] Programa de Pós-graduação em Ciências Morfofuncionais, Universidade Federal de São João D'El Rey, São João D'El Rey, MG, Brazil

[3] Laboratório de Pesquisa Básica e Translacional em Oncologia, Instituto Mario Penna, Belo Horizonte, MG, Brazil

[4] Seção Técnica de Biologia e Bacteriologia Legal, Instituto de Criminalística, Polícia Civil de Minas Gerais

[5] Departamento de Ginecologia e Obstetrícia, Faculdade de Medicina, Universidade Federal de Minas Gerais, Belo Horizonte, MG, Brazil

Correspondence:

*Shared Corresponding authors:

Matheus de Souza Gomes, PhD

Universidade Federal de Uberlândia - Campus Patos de Minas

R. Padre Pavoni, 290 - Rosário, Patos de Minas – Minas Gerais - Brazil

ZIP Code: 38701-002

Phone: +55 34 3823-3714

E-mail: matheusgomes@ufu.br

and

Letícia da Conceição Braga, PhD

Laboratório de Pesquisa Básica e Translacional em Oncologia - Instituto Mário Penna

Rua Gentios, 1420 - Belo Horizonte - Minas Gerais - Brazil

ZIP code: 30380-472

Phone: +55 31 3299-9526

E-mail: leticia.braga@mariopenna.org.br

HIGHLIGHTS

- Few significant mutations in mitochondrial DNA were found in samples from ovarian cancer patients
- Ovarian cancer growth may benefit from the oxidative phosphorylation pathway
- Targeting the oxidative phosphorylation pathway could be a promising strategy for ovarian cancer treatment

ABSTRACT

This study explores the role of mitochondrial DNA (mtDNA) and oxidative phosphorylation (OXPHOS) in high-grade ovarian cancer (HGOC). Recognizing dysregulated energy metabolism as a crucial hallmark in cancer development, the study challenges the widespread belief in universal OXPHOS impairment across all cancer types. The analysis of mtDNA mutations in HGOC patients revealed a lower frequency of genetic variants in the samples, with 14.4% categorized as missense mutations, predominantly characterized by pyrimidine transitions. Notably, an elevated mutation frequency in the D-loop region of mtDNA was observed. Haplogroup analysis indicated genetic heterogeneity within the patient cohort. Contrary to the traditional Warburg effect, findings suggest a predominant utilization of OXPHOS for energy production in ovarian tumors, particularly primary tumors. This novel understanding opens avenues for OXPHOS as a viable therapeutic target in ovarian cancer treatment. While preliminary, further research is essential to fully comprehend the implications of mtDNA mutations and OXPHOS in ovarian cancer for effective clinical interventions.

Keywords: Ovarian cancer, Mitochondrial DNA, Oxidative phosphorylation, Missense mutations, Personalized medicine, Therapeutic strategies

1 INTRODUCTION

Mitochondria are central organelles at the intersection of several energy metabolisms. According to the Hanahan and Weinberg's hallmarks of cancer, the reprogramming of energy metabolism is an emerging hallmark for the development and progression of cancer since the metabolism of various macronutrients can be altered to promote cell proliferation [1]. In humans, mitochondrial DNA (mtDNA) is organized as a circular double-stranded DNA molecule and is approximately 16,569 bp in length. It is distinct from nuclear DNA and encodes essential components of the mitochondrial oxidative phosphorylation (OXPHOS), which, given its high energy yield, is crucial to produce cellular energy [2].

The OXPHOS machinery consists of five multiprotein complexes, in which electron transfer occurs and an electrochemical proton gradient is created to drive the synthesis of adenosine triphosphate (ATP). The mtDNA contains 13 of the [3] 92 structural genes of the electron transport chain complexes: the MT-ND1, MT-ND2, MT-ND3, MT-ND4, MT-ND4L, MT-ND5 and MT-ND6 genes of complex I; the MT-CYB of complex III; the MT-CO1, MT-CO2 and MT-CO3 genes of complex IV; and the MT-ATP6 and MT-ATP8 genes of complex V. Complex I is responsible for catalyzing the transfer of electrons to ubiquinone from NADH. Complex II catalyzes the transfer of electrons to ubiquinone from succinate and is composed of four subunits, all from nuclear DNA. Complex III carries electrons from reduced ubiquinone to cytochrome c. Complex IV transfers electrons from cytochrome c to oxigen. These four complexes are responsible for electron transfer and the formation of a proton gradient in the respiratory chain of the OXPHOS system. The last complex, complex V, is formed by an enzyme that synthesizes ATP using the proton motive force generated by complexes I, III and IV [4].

In addition to the protein-coding genes, the mtDNA also has genes for 22 tRNAs and 2 rRNAs which are useful for synthesizing mitochondrial proteins and a control region which regulates the transcription and replication of the mtDNA, called the displacement loop (D-loop), which is approximately 1000 nucleotides in length [5]. Given the importance of mtDNA-encoded genes in the electron

transport chain and the importance of OXPHOS for producing energy in cells, functional variants in this region may affect the copy number and regulation of expression of mtDNA genes and influence cancer progression. Alterations in mtDNA may be associated with mitochondrial dysfunction, impaired OXPHOS, and increased production of reactive oxygen species (ROS), consequently affecting cell growth, signaling, inflammation, and susceptibility to infections; Ca^{2+} and pH levels, regulators of homeostasis; apoptotic initiation, and alteration of tricarboxylic cycle intermediates [6–9].

Aerobic glycolysis, or the so-called Warburg effect [10], is a phenomenon that occurs in some cancer cells that favors glycolysis, instead of OXPHOS, even under aerobic conditions, for energy production. It was therefore assumed that OXPHOS would be negatively regulated in cancer cells. Since this effect was discovered, various therapies have been aimed at pharmacological inhibition of glycolysis. However, recent studies show that cancers are metabolically heterogeneous and that some types of tumors depend on OXPHOS for their bioenergetics and consequent tumor growth [11–14]. To highlight the importance of the genes encoded by mtDNA, a study was carried out by completely removing mtDNA from tumors and it was observed that tumor cells without mtDNA showed delayed tumor growth and formed tumors only after acquiring mtDNA from the host organism [15]. Regarding mtDNA copy number, a study by Wang et al. shows that mtDNA content in tumor cells was significantly higher than in the normal ovary [16]. Emphasizing the importance of mtDNA for OXPHOS and cell energy generation in ovarian cancer. Therefore, OXPHOS could be an emerging target in cancer therapy and has already been studied as a possibility in several cancers [17,18], including ovarian cancer [19,20].

Ovarian cancer is a complex and heterogeneous disease characterized by the uncontrolled growth of abnormal cells in the ovaries. It is a significant health concern worldwide, ranking as one of the leading causes of cancer-related mortality among women [21]. Ovarian cancer presents unique challenges due to its asymptomatic nature in the early stages, resulting in late diagnosis and limited treatment options [22]. Understanding the molecular mechanisms underlying ovarian cancer development and progression is crucial for improving early

detection, personalized therapies, and ultimately, patient outcomes. Extensive research efforts are focused on identifying biomarkers, deciphering genetic alterations, and exploring therapeutic targets to advance our knowledge and develop more effective strategies for the prevention, diagnosis, and treatment of this devastating disease. The mtDNA has emerged as a topic of interest in ovarian cancer research due to its characteristics and potential involvement in disease development and progression [23].

Some studies have been carried out on ovarian cancer focusing on mitochondria [24–26]. Interestingly, mtDNA-deficient ovarian cancer cells highly expressed several genes related to anti-apoptosis, proliferation regulation, and cell stemness [27]. In short, mtDNA in ovarian cancer appears to be essential for tumor cell growth, suggesting that OXPHOS should be highly utilized to provide energy support for this type of tumor. Understanding the intricate relationship between mtDNA and ovarian cancer can provide valuable information on disease mechanisms, biomarkers, and personalized treatment strategies.

Therefore, our aim was to investigate the mutations that occur in mtDNA to better understand the impact of the OXPHOS pathway on ovarian cancer.

2 METHODOLOGY

2.1 Patients

A total of 24 patients were sequenced. Of these, there were seven normal ovaries, 17 primary high-grade ovarian cancer (HGOC) and six metastatic HGOC (from patients who also had primary tumors). The cohort was selected from the Gynecologic Oncology Service of the School of Medicine of Minas Gerais Federal University, between February 2008 and July 2010. The ethics committee of Institution approved the study (protocol CAAE: 85375317.0.0000.5149), and all patients signed an informed consent form.

Before the study, all participants underwent clinical and gynecological assessment and transvaginal ultrasound. In patients diagnosed with ovarian serous cystadenocarcinoma, surgical staging was carried out following the guidelines of the International Federation of Gynecology and Obstetrics (FIGO). Non-cancerous ovarian tissue samples were obtained from individuals

undergoing oophorectomy as part of total abdominal hysterectomy for the treatment of benign gynecological conditions. A pathologist conducted the histological examination of the samples. None of the patients had previously received chemotherapy or had an acute infectious peritoneal condition.

Thus, 24 libraries were prepared for sequencing: 17 primary tumor libraries, six metastatic tumor libraries and a pool with the normal ovary samples.

2.2 DNA extraction, Libraries Construction and Sequencing

Mitochondrial DNA extraction was performed from fresh frozen tissue. Phase separation was performed using TRIzol-based method and total DNA was recovered with using Back extraction buffer. Columns from the Wizard® Plus SV Minipreps DNA Purification System kit were used for mtDNA filtration. In the mean, 250 ng of mtDNA mass was rescued to construct the sequencing libraries. 35 ng of each of the seven samples, used for the library of the pool of normal ovaries. The Illumina DNA Prep was used to prepare the sequencing libraries according to manufacture protocol. Quality parameters as size (580 bp average) and concentration (2311 pg/µl average) were measured using High Sensitivity D1000 on 2200 TapeStation System (Agilent, CA) and Qubit™ dsDNA HS Assay Kit on a Qubit 3 Fluorometer (Thermo Fisher Scientific, MA), respectively. Libraries that showed the typical size profile, with an average of 600 bp analyzed in a range of 150-1500 bp, as suggested by the manufacturer, were normalized to 4 nM. The 24 libraries were pooled and sequenced using Illumina's MiSeq™ FGx, using the 300-cycle MiSeq Reagent Kit v2 cartridge from the Forensic Biology and Bacteriology Technical Section of the Minas Gerais Civil Police.

2.3 Mitochondrial DNA variants analysis

First, we evaluate the quality of the reads using FastQC Version 0.12.0 [28]. Then, the 3' adapters were removed by Trimmomatic version 0.40 [29]. To select variants with confidence, filtering was performed requiring a minimum Phred quality score of 30. The mapping of the sequences was performed with the BWA version 0.7.17 [30] using the chrM, referring to the mtDNA, of the reference human genome GRCh38. Variant calling was performed using Mutect2 of GATK

version 4.2.3.0 [31], with mitochondria mode flag, and variants with a minimum depth of 30 and without the week evidence flag, parameters adjusted to exclude false positives. The variants were annotated using Ensembl Variant Effect Predictor (VEP) version 110.1 [32] and the MITOMAP - A human mitochondrial genome database [33]. The Haplogrep version 3.2.1 program [34] was used to classify the haplogroups. And for generation of MAF files, and subsequent visualization of the graphs, Vcf2maf, of VEP, was used. The VEP, Variant Effect Predictor, uses some tools to evaluate the impact of mutations on proteins, such as SIFT, PolyPhen-2 and CLIN_SIG (which presents the ClinVar clinical significance of the dbSNP variant and is in accordance with the American College of Medical Genetics - ACMG guideline). The package of R program Maftools was used for data visualization and processing. All steps are summarized in the Supplementary File 1.

2.4 Validation data in publicly available mtDNA data

To validate our data, we matched our results with mtDNA variants available in public databases - the International Cancer Genome Consortium (ICGC) (https://dcc.icgc.org/ and RRID: SCR_021722) and The Cancer Mitochondrial Atlas (TCMA) (https://bioinformatics.mdanderson.org/public-software/tcma/). Information was obtained from 100 ovarian cancer samples from the TCMA and 90 samples from the ICGC - some samples overlap. The available results were the VCFs with the mutations already defined. We didn't do the variant calling.

3 RESULTS

Of the 24 libraries sequenced, only the normal ovary pool, one metastatic HGOC sample, and five primary HGOC samples obtained more than 600,000 sequenced reads and were considered in bioinformatic analysis.

3.1 Casuistic

The clinical and pathological data of patients included in the analyses for complete mitochondrial DNA sequencing are presented in Table 1.

Table 1. Clinicopathological characteristics of the patients analyzed.

	Normal Ovary	HGOC primary	HGOC metastasis
	n = 7	n = 5	n = 1
Age mean (years)	45	55	56
Parity (births)		1,8	3
CA125			
CA>330 (U/mL)		1	1
CA<330 (U/mL)		4	
Cytoreduction			
Optimal (<1 cm)		3	1
Suboptimal (>1 cm)		2	
Stage (FIGO)			
Stages I-II		3	1
Stages III-IV		1	
Ascites			
Yes		4	1
No		1	
Menarche (mean age)		14	11
Menopause			
Yes		4	1
No		1	

3.2 Identification of mtDNA mutations in primary and metastatic HGOC and normal ovarian samples of patients

After constructing and evaluating the libraries, sequencing was carried out at the Minas Gerais Civil Police using the MiSeq™ FGx by Illumina. The raw sequencing data can be found in accession PRJNA1082308 in the SRA database.

The data was filtered only for the mitochondrial "chromosome" of the most recent human genome, to eliminate any remnants of non-mitochondrial DNA in the samples. After this filtering, quality control was performed using FastQC. A total of 118 different variants (there is overlap of variants within samples), were found (Supplementary File 2). The average depth of the mutations was 115 per patient independent of the histological sample type. As can be seen in the Supplementary File 3 and Supplementary File 4, respectively the Mitomap and MAF tools annotations, despite we had detected all types of variants, the most common were the substitutions of single nucleotide (SNPs), and the transitions between pyrimidines were the most prevalent, following the transversion presented in 8,5% of all mutations. The transition T>C was the most frequent (Figure 1).

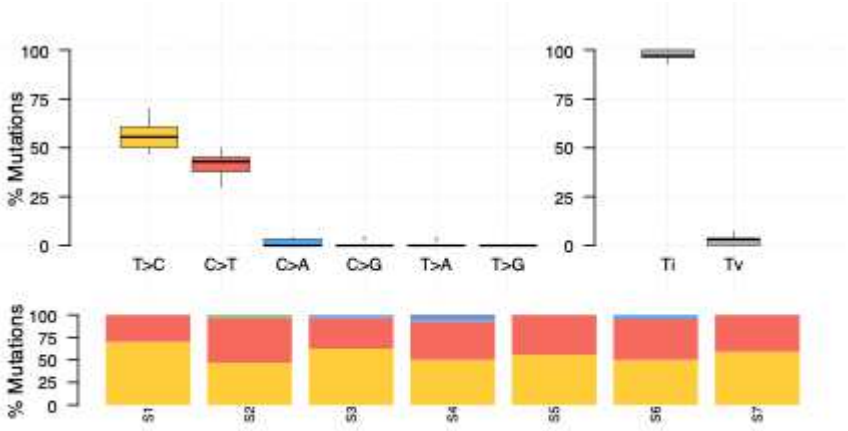


Figure 1. Types of substitutions of mutations. The top of the figure shows the percentage of substitutions in all the samples, represented as TMB. The bottom part shows the substitutions in each sample. Most substitutions were transitions, and the most common transition was from thymine to cytosine. Ti refers to transitions, most of the substitutions in this cohort, and Tv are transversions.

In Figure 2, the individual variant counts for each sample are illustrated. The normal ovary pool exhibited 41 variants. Primary tumors displayed variant counts ranging from 12 to 34, averaging 26.4. The metastatic HGOC sample (S1) revealed 32 variants.

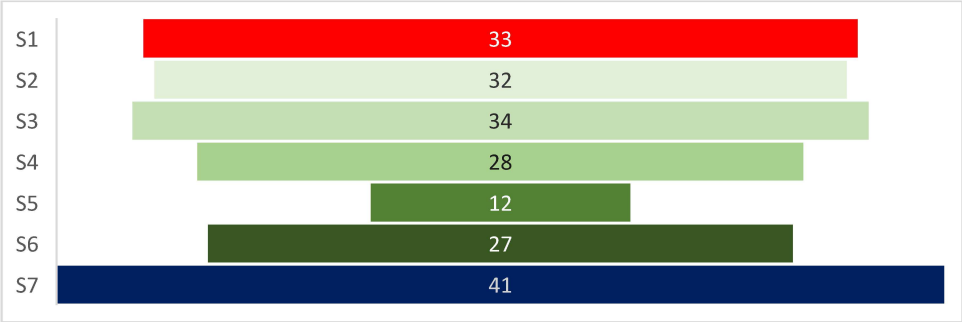


Figure 2. Number of mutations per sample. On the y-axis we have the samples and on the x-axis the number of mutations in each sample. S7, a pool of normal ovaries (S7), had the highest number of mutations compared to the primary HGOC samples (S2-S6), followed by S1, metastatic HGOC.

The major number of mutations were found in the regulatory region of the mtDNA, the D-loop, following MT-CYB, MT-ND5, MT-ND4 and MT-RNR1. (Figure 3). The presence of mutations in the D-loop and ribosome genes (MT-RNR1 and MT-RNR2) was presented in all three groups. Mutations in transporter RNA genes occurred only in tumor samples. Three occurred in primary (two in S3 and one in S4), while three in metastatic HGOC. However, mutations in MT-ATP6 and MT-CO1 occurred only in tumor samples too, but in both, primary and metastatic, at the same time (Figure 3).

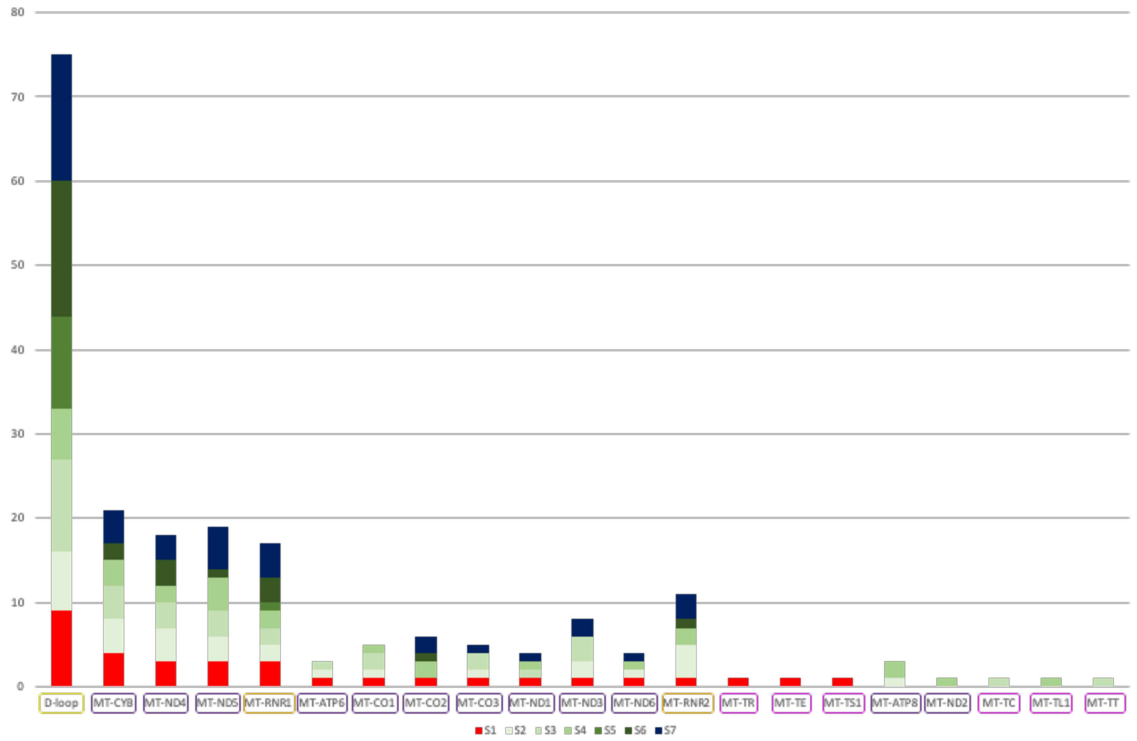


Figure 3. Number of mutations presented in each gene. On the y-axis we have the cumulative number of mutations, the colors of the bars indicate the samples and, on the x-axis, the mutated genes. D-loop is a hypervariable region, highlighted in yellow. Protein-coding genes are highlighted in purple. In orange, the rRNA genes and, in pink, the tRNAs.

Figure 4 shows the missense and Multi_hits mutations that occured in protein-coding genes only. Multi_Hits in the oncoplot mean those genes that have mutated more than once missense mutation in the same sample. The missense mutations in protein-coding genes occurred in 86% (6/7) of samples, including the metastatic HGOC, S1, primary HGOC, S2, S3, S4, S6, and the normal ovary pool, S7. Sample S5 did not have any missense mutation identified. Other variant classifications, such as silent and 3' and 5' flank variants, were identified in all samples and can be seen in the MAF tools annotations table (supplementary file 4). No frameshift or nonsense mutations were identified in the samples. The TMB, shown at the top of Figure 4, estimates tumor mutation burden. The samples S1, S2, and S3 had higher TMB levels. In all samples, except S5, more than one missense mutation occurred in each sample. MT-CYB was the gene with the highest number of mutations. MT-ATP6 and MT-ND3 were the two second most mutated (Figure 4). In MT-ATP6 there was only one mutation per tumor sample, metastatic, S1, and primary, S2, S3, and S4 (Figure 4). There were mutations in all three groups (normal ovary, primary, and metastatic HGOC) in the MT-ND3 gene, two mutations were observed in S3. In the MT-ND5 gene, the third with the most missense mutations, there was more than one missense mutation in the secondary tumor sample (Figure 4), and a single missense mutation in two primary HGOC samples (S2 and S3). In MT-CO2, MT-ATP8, and MT-ND4 genes, the mutations were observed only in primary tumor samples (Figure 4). And in the MT-ND1 gene, there was a missense mutation in the secondary tumor sample.

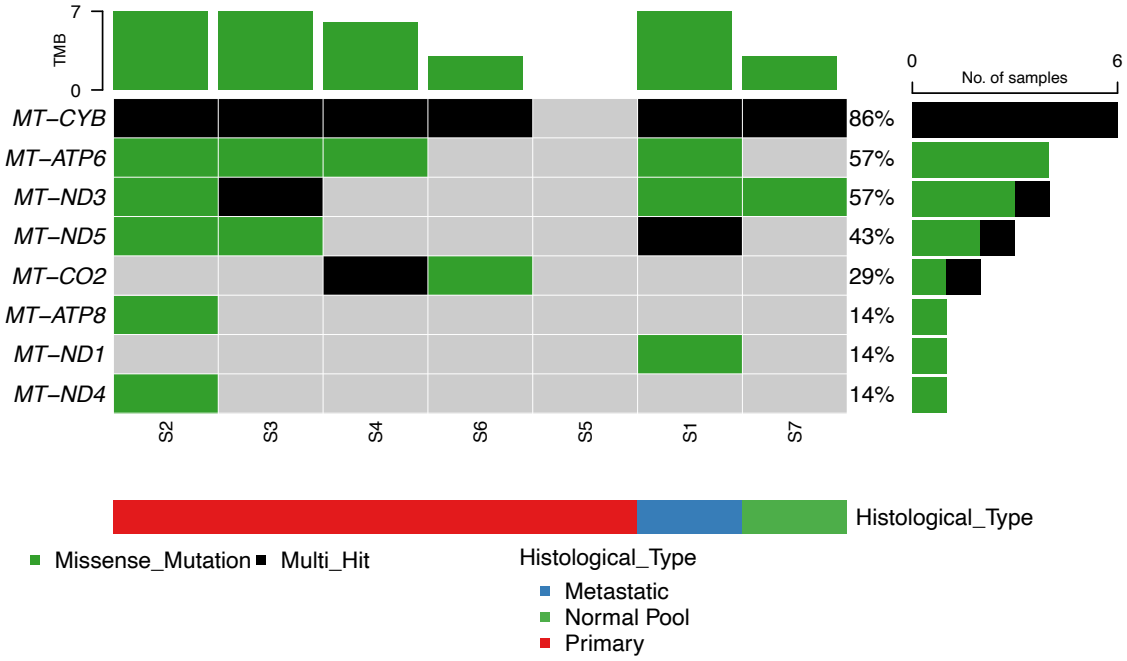


Figure 4. Oncoplot of genes with a missense mutation in coding genes. mtDNA genes that had missense mutations in each sample, separated by their histologic type, were evaluated in this study. The predominant green color indicates that all mutations that occurred in the sample are missense. The number, in this case from 0 to 7, indicates how many total mutations occurred in each sample named as TMB.

In table 2, the information on the missense mutations available in each gene according to each sample is clearer to observe. All missense mutations were classified according to the pathogenicity degree and probability of damage to protein function, this results were summarized in Table 2 (information taken from supplementary file 4). Two variants reported were not reported in dbSNP before (named as 'Novel'). One in MT-CO2 and the other in MT-ND5. Both were present in patients with primary HGOC (S4 and S2, respectively). The same variant was present in more the one patient, such as MT-ATP6:c.1754A>G was presented in three patients metastatic and primary HGOC (S1, S2, and S3), MT-CYB:c.20C>T and MT-CYB:c.580A>G occurred in six patients of three groups (S1, metastatic HGOC, S2, S3, S4, S6, primary HGOC and S7, pool of normal ovary), MT-ND3:c.340A>G was presented in four patients of three groups too (S1, metastatic HGOC, S2 and S3, primary HGOC and S7, pool of normal ovary) and MT-ND5:c.769A>G occured in a metastatic HGOC (S1) and a primary HGOC (S3). In SIFT, 75,75% of variants had their effect considered tolerant and

24,25% deleterious while in PolyPhen, 78,78% were benign, 15,15% were probably damaging and 6,07% were possibly damaging. Two variants are classified as pathogenic: MT-CYB:c.239G>A and MT-ND1:c.490A>G. 29 of 32 variants were classified as benign in at least one ClinVar study, 1 likely benign, 12 as likely pathogenic, and one variant of uncertain significance (VOUS) were detected.

Table 2. Classification of Pathogenicity of the missense mutations found in this study.

Gene	Samples	Position	dbSNP IDs	HGVSc	HGVSp	SIFT	PolyPhen	CLIN_SIG
MT-ATP6	S1	8701	rs2000975	c.175A>G	p.Thr59Ala	tolerated_low_confidence(0.64)	benign(0.005)	benign
	S2	8701	rs2000975	c.175A>G	p.Thr59Ala	tolerated_low_confidence(0.64)	benign(0.005)	benign
	S3	8701	rs2000975	c.175A>G	p.Thr59Ala	tolerated_low_confidence(0.64)	benign(0.005)	benign
	S4	8557	rs386829040	c.31G>C	p.Ala11Pro	deleterious_low_confidence(0.01)	possibly_damaging(0.594)	likely_benign
MT-ATP8	S2	8414	rs28358884	c.49C>T	p.Leu17Phe	tolerated(0.23)	probably_damaging(0.992)	benign
MT-CO2	S4	7805	rs879119797	c.220G>A	p.Val74Ile	tolerated_low_confidence(1)	benign(0.005)	benign
		8061	novel	c.476T>C	p.Val159Ala	deleterious_low_confidence(0)	probably_damaging(1)	
	S6	8027	rs1116904	c.442G>A	p.Ala148Thr	tolerated_low_confidence(0.12)	probably_damaging(0.998)	benign
MT-CYB	S1	14766	rs193302980	c.20C>T	p.Thr7Ile	tolerated_low_confidence(0.05)	benign(0.003)	benign,likely_pathogenic
		15326	rs2853508	c.580A>G	p.Thr194Ala	tolerated_low_confidence(0.21)	benign(0.009)	benign,likely_pathogenic
	S2	14766	rs193302980	c.20C>T	p.Thr7Ile	tolerated_low_confidence(0.05)	benign(0.003)	benign,likely_pathogenic
		15326	rs2853508	c.580A>G	p.Thr194Ala	tolerated_low_confidence(0.21)	benign(0.009)	benign,likely_pathogenic
	S3	14766	rs193302980	c.20C>T	p.Thr7Ile	tolerated_low_confidence(0.05)	benign(0.003)	benign,likely_pathogenic
		15326	rs2853508	c.580A>G	p.Thr194Ala	tolerated_low_confidence(0.21)	benign(0.009)	benign,likely_pathogenic
		15824	rs28357376	c.1078A>G	p.Thr360Ala	tolerated_low_confidence(0.5)	benign(0)	benign
	S4	14766	rs193302980	c.20C>T	p.Thr7Ile	tolerated_low_confidence(0.05)	benign(0.003)	benign,likely_pathogenic
		14985	rs207459995	c.239G>A	p.Arg80His	deleterious_low_confidence(0.01)	probably_damaging(0.997)	uncertain_significance,pathogenic
		15326	rs2853508	c.580A>G	p.Thr194Ala	tolerated_low_confidence(0.21)	benign(0.009)	benign,likely_pathogenic
	S6	14766	rs193302980	c.20C>T	p.Thr7Ile	tolerated_low_confidence(0.05)	benign(0.003)	benign,likely_pathogenic
		15326	rs2853508	c.580A>G	p.Thr194Ala	tolerated_low_confidence(0.21)	benign(0.009)	benign,likely_pathogenic
	S7	14766	rs193302980	c.20C>T	p.Thr7Ile	tolerated_low_confidence(0.05)	benign(0.003)	benign,likely_pathogenic
		15326	rs2853508	c.580A>G	p.Thr194Ala	tolerated_low_confidence(0.21)	benign(0.009)	benign,likely_pathogenic
MT-ND1	S1	3796	rs28357970	c.490A>G	p.Thr164Ala	deleterious_low_confidence(0.05)	benign(0.013)	benign,pathogenic
MT-ND3	S1	10398	rs2853826	c.340A>G	p.Thr114Ala	tolerated_low_confidence(0.3)	benign(0)	protective,benign
	S2	10398	rs2853826	c.340A>G	p.Thr114Ala	tolerated_low_confidence(0.3)	benign(0)	protective,benign
	S3	10086	rs28358274	c.28A>G	p.Asn10Asp	deleterious_low_confidence(0.03)	possibly_damaging(0.903)	benign
		10398	rs2853826	c.340A>G	p.Thr114Ala	tolerated_low_confidence(0.3)	benign(0)	protective,benign
	S7	10398	rs2853826	c.340A>G	p.Thr114Ala	tolerated_low_confidence(0.3)	benign(0)	protective,benign
MT-ND4	S2	11150	rs386829118	c.391G>A	p.Ala131Thr	deleterious_low_confidence(0)	benign(0.03)	benign,not_provided
MT-ND5	S1	13105	rs2853501	c.769A>G	p.Ile257Val	tolerated_low_confidence(0.77)	benign(0.007)	benign
		13886	rs28359182	c.1550T>C	p.Leu517Pro	deleterious_low_confidence(0.04)	benign(0.003)	benign
	S2	14005	novel	c.1669T>C	p.Trp557Arg	deleterious_low_confidence(0)	probably_damaging(0.998)	
	S3	13105	rs2853501	c.769A>G	p.Ile257Val	tolerated_low_confidence(0.77)	benign(0.007)	benign

3.3 Position of missense mutations in the genetic structure of the two genes with pathogenic variants

The positions of the missense mutations in the MT-CYB and MT-ND1 gene are shown in Figure 5. The variants occur in four positions of the MT-CYB protein, with amino acid positions 7 and 194 recurring in the six samples (Figure 5A). And both were classified as tolerance effect (Table 2). There was a mutation in the MT-ND1 gene only in sample S1, metastatic HGOC. The missense mutation occurred at position 164 of the protein, within the conserved domain region (Figure 5B). Besides benign effect reported in PolyPhen, it was classified as deleterious effect in SIFT.

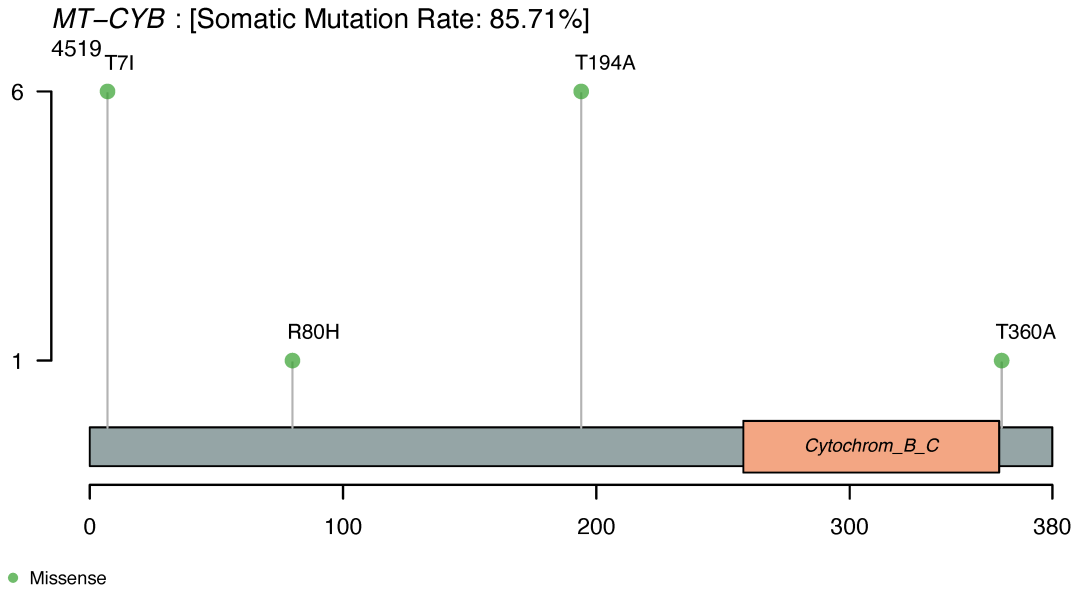


Figure 5A. Mutation spectrum across functional domains of MT-CYB.

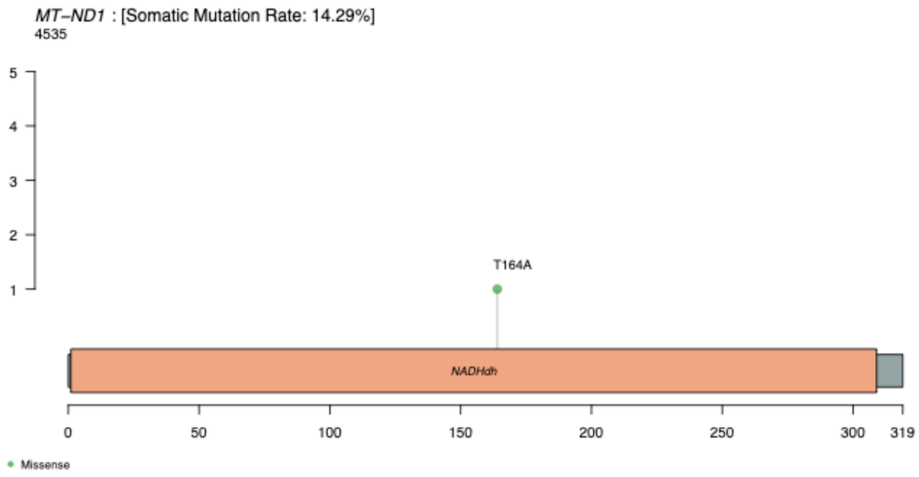


Figure 5B. Mutation spectrum across functional domains of MT-ND1.

3.4 Haplogroup of primary and metastatic HGOC and normal ovarian samples of patients in our cohort

The haplogroup classification of samples 1 to 6 was also performed (sample 7 is a pool of samples). As can be seen in Table 3, four haplogroups were observed with two of them (L3 and A2). with more than one sample

Table 3. Distribution of the haplogroup of samples S1-6

Sample	Predicted Haplogroup
S1	L3d (L3d1a2)
S2	D1 (D1)
S3	L3b (L3b1a)
S4	U6a (U6a6b)
S5	A2 (A2+(64))
S6	A2a (A2aa)

3.6 Mutations found in the TCMA and ICGC public databases

A total of 371 mutations were found in 100 samples in the TCMA and 166 mutations in 90 samples in ICGC. The most mutated gene in these samples was MT-RNR2, as shown in Figure 7, following the complex I genes: MT-ND4 and MT-ND5.

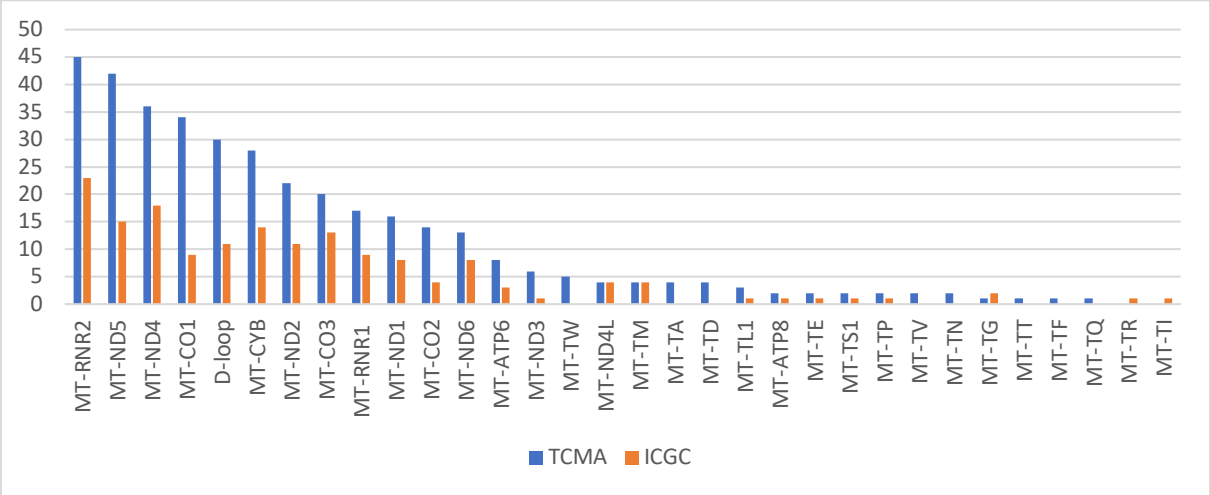


Figure 7. Number of mutations in DBs. The most mutated gene in both databases was an rRNA gene: MT-RNR2.

4 DISCUSSION

In some types of cancer, energy consumption is extremely high, considering the metabolic demands for the cancer's development. It has been observed that several cancers require the OXPHOS system for energy production [13,35], including ovarian cancer [36]. In breast cancer, an experiment was carried out by removing mtDNA from metastatic cells and replacing it with mtDNA from normal cells, simultaneously restoring mitochondrial respiratory function and suppressing highly metastatic potential [37]. These observations suggest that the mitochondrial respiratory defects observed in the cells are caused by mutations in the mtDNA, but not in the nuclear DNA, and are responsible for the expression of the high metastatic potential without using ROS-mediated pathways. This highlights the importance of mtDNA in cancer progression.

A low rate of mtDNA mutations in tumor samples would make sense, and our study agrees hardly with a hypothesis (Figure 2 – Supplementary File 2). Some filters are applied to remove false positives. We are aware that we are losing information with this decision, mainly due to heteroplasmic mutations having low representation concerning coverage. However, we preferred the certainty that the mutations presented were indeed positive and not sequencing artifacts.

For the development of tumors, mutations would play the role of gain or loss of function for the benefit or harm of the tumor. Therefore, they suffer greater selective pressure for variations than normal ovaries, which could only accumulate meaningless mutations. Interestingly, Grandhi and colleagues observed that metastatic tumors usually have more mutations in mtDNA than primary tumors [38]. Perhaps for metastatic tumors, there is greater selective pressure for gain of function due to random mutations. The exact mechanism is unknown, but one hypothesis could be that secondary tumors have different environments compared to primary tumors. Thinking about the oxygen-dependent pathway, OXPHOS, it may be that there is greater oxygen availability for primary tumors, and therefore, there is greater selective pressure for fewer (and less significant) mutations in OXPHOS genes, to ensure efficient energy

production. Or even pressure for advantageous mutations that lead to greater energy production under these conditions. In Figure 2, we can see that the average number of mutations in primary tumor samples is lower than in secondary tumor samples, corroborating this assumption.

In many studies of the molecular characterization of mitochondrial genomes in various types of cancer, D-loop region was the most mutated, followed by the MT-ND4 gene [39]. In other studies, it has been observed that mutations located in this region are associated with a reduction in the number of mtDNA copies [40] and an increase in D-loop mutations and tumor progression [41]. However, in ovarian cancer, no significant relationship was found between mitochondrial alterations and the clinical outcome of patients [42]. Similarly, as shown in Figure 3, the regulatory region, D-loop, was where the highest number of mutations occurred in all the samples in this study. Mutations in tRNAs can compromise the structure of these transporters and affect the coding of mtDNA proteins. Our study observed that these mutations occurred only in tumor samples (Figure 3). Three mutations in MT-TR, MT-TR, and MT-TS1 happened in the metastatic tumor patient. While in primary tumors, we detected two mutations in MT-TC and MT-TT and one in MT-TL1. A higher tRNA mutant allele frequency occurred in a study evaluating 24 types of cancer, comparing tumors and normal ovaries [38].

Interestingly, the MT-RNR2 gene was the most mutated in the data we collected from the TCMA and ICGC databases. In our results, MT-RNR1 was mutated more than MT-RNR2 in most samples (except for S2, which had four mutations in MT-RNR2 and two in MT-RNR1) (Figure 3). Once MT-RNR1 and MT-RNR2 genes are rRNA coding, these mutations could also affect protein-coding, reducing energy production by OXPHOS. In hepatocellular carcinoma, the MT-RNR1 G709A polymorphism was a risk factor and resulting in an unfavorable prognosis [43]. A study that sought a clinical association between ovarian cancer and mtDNA mutations showed that MT-RNR1 expression predicts platinum sensitivity [42]. In our cohort, apparently this association could not be detected once it was presented in all histological types. More important than the number of mutations in mtDNA genes is the disruptive impact of these mutations.

Regarding mutations in protein-coding genes, we will focus on the most significant ones, missense mutations. This mutation can alter the resulting protein's function, as it alters the DNA that causes different amino acids to be encoded in a specific position. The mutations must not be deleterious and prevent protein function for ATP synthesis to occur via OXPHOS to correspond to the energy demands of the tumor (especially the primary one). And in our data, we can see this (Table 2).

Most of the missense mutations observed were in tumor samples (Figure 4) which is why the TMB of the normal ovary pool shown in the figure is lower. The mutational load assessed refers only to the missense mutations identified in the study. As we can see in Table 2, not all the mutations in S1, metastatic HGOC, were in conserved domain regions of the proteins, especially since not all the proteins had mutations in this region. However, all proteins carrying mutations in conserved domains were identified in this sample. This suggests that the impact of the mutation would be higher in metastatic HGOC.

The importance lies not only in the number of variants, but in the individual impact of each mutation. Few variants have been detected and none of the mutations identified were classified as deleterious or damaging by the SIFT and PolyPhen classifiers, respectively. Two mutations were classified as pathogenic, by the ClinVar classification, one is located outside the conserved domain of MT-CYB (Figure 5), and the other was localized within a domain of MT-ND1 in metastatic HGOC (Figure 6).

The MT-CYB gene encodes a cytochrome b protein, OXPHOS complex III [44]. Complex III is responsible for catalyzing the transfer of electrons from coenzyme Q to cytochrome c in the mitochondrial respiratory chain, translocating protons concomitantly across the mitochondrial inner membrane [45]. The transfer of electrons then contributes to the generation of a proton gradient across the mitochondrial membrane which is then used for the synthesis of ATP. Therefore, a deficiency in cytochrome b could influence the proton gradient between the membranes and reduce ATP production capacity. Missense mutations in MT-CYB have been found in oral cavity cancers [46] and brain

tumors [47]. This suggests that there is an advantage over normal cells for cancer progression.

Complex I is a vital enzyme complex necessary for OXPHOS. The mitochondrially encoded NADH: ubiquinone oxidoreductase central subunit 1 (MT-ND1) is the largest subunit encoded by complex I mitochondria [48]. Mutations in MT-ND1 can cause an alteration in the components of electron transport and damage the normal flow of electrons, leading to an increase in the generation of superoxidase radicals and increasing oxidative stress [49]. In colorectal cancer, some variants in MT-ND1 have been found and associated with this type of cancer [50]. In addition to having a higher content and mutational load of circulating MT-ND1, patients with colorectal cancer had a poorer prognosis. Mutations in MT-ND1 have also been associated with breast cancer [51] and kidney cancer [52], showing that there may be a link between this gene and cancer.

As the OXPHOS system appears to be active in ovarian cancer, mitochondria could be targeted as therapeutic targets. This paper [53] shows some mitochondrial functions that can be targeted in cancer therapy, such as the electron transport chain and mitochondrial RNA polymerase (MT-POLRMT). Two mechanisms entirely linked to OXPHOS. Metformin, tamoxifen, α-tocopheryl succinate and 3-bromopyruvate are drugs that target the electron respiratory chain complexes, disrupting the function of the complexes and inducing the production of high levels of reactive oxygen species [54]. Therefore, other drugs and chemotherapies can be designed to target OXPHOS complexes or even mtDNA.

Most of the substitutions were pyrimidine transitions, T>C and C>T. This type of substitution has less of an impact than transversions since the type of nitrogenous base is categorized by the structure of the molecule. Pyrimidines have only one aromatic ring, while purines have two. Chemically, the steric effect of substitution is less when it occurs within the same category. In the study of the molecular characterization of mtDNA in human cancers, this pattern was also observed, both in ovarian cancer and in other types of cancer [39]. In ulcerative colitis the pyrimidine transition substitution was also the most common [41].

Haplogroups are a single group of mtDNA polymorphisms strongly linked to geographically specific genotypes [55]. In other words, the groups of polymorphisms emerged with evolution and became a kind of "signature" of each region of the planet as Homo sapiens moved to the different continents. Some haplogroups confer specific characteristics, such as a population's adaptation to a low-oxygen environment [56] or protection from myocardial infarction in a specific population [57]. Therefore, some studies have been conducted to relate haplogroups to an increased or decreased risk of cancer. Haplogroup M was associated with an increased risk of breast cancer and haplogroup D4a with an increased risk of thyroid cancer [58]. Haplogroup M33a2'3 was associated with susceptibility to pulmonary edema at high altitudes in Indian men [59]. In ovarian cancer some studies [26,42] have been performed, however no association was made in relation to risk. Our results (table 3) were also inconclusive and there was a diversity of haplogroups represented in the samples.

5 CONCLUSIONS

We encountered some difficulties with databases with little information on mtDNA. Although the importance of studying this mitochondrial genome has been highlighted, it was difficult to find data to correlate with our findings. There is a need for studies with more patients represented for more precise conclusions. However, we observed a potential marker that should be explored.

A few mutations, as well as a few pathogenic mutations, suggest the underestimated role of mitochondria in ovarian cancer. Our results indicate that the OXPHOS system is widely used, especially in primary tumors, which require much energy for proliferation. This raises the possibility of a therapeutic target for inhibiting tumorigenesis. And because mtDNA is connected to the mitochondrial internal membrane, it is extremely sensitive to fat-soluble substances, unlike nDNA. Therefore, it can be targeted as a binding and attack site for lipophilic carcinogens. One limitation of our work was the number of patients in each sequencing group. Even though work with other cancers supports this idea,

further work needs to be produced to be convinced that OXPHOS plays an important role in the progression of ovarian cancer.

ACKNOWLEDGMENTS

I would like to thank the Mario Penna Institute for its structure and support and Carolina who took part in meetings where the experimental design and discussions of the results took place. I would also like to thank the Civil Police of Minas Gerais for donating the reagents and the sequencer to make the experiment possible. To FUNED and UFMG for the samples. To Capes for funding the scholarship and to UFU for the structure and training.

REFERENCES

[1] Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. Cell 2011;144:646–74. https://doi.org/10.1016/j.cell.2011.02.013.

[2] Chinnery PF, Hudson G. Mitochondrial genetics. Br Med Bull 2013;106:135–59. https://doi.org/10.1093/bmb/ldt017.

[3] Xu Y, Xue D, Bankhead A, Neamati N. Why All the Fuss about Oxidative Phosphorylation (OXPHOS)? J Med Chem 2020;63:14276–307. https://doi.org/10.1021/acs.jmedchem.0c01013.

[4] Signes A, Fernandez-Vizarra E. Assembly of mammalian oxidative phosphorylation complexes I-V and supercomplexes. Essays Biochem 2018;62:255–70. https://doi.org/10.1042/EBC20170098.

[5] Wallace DC. A mitochondrial bioenergetic etiology of disease. J Clin Invest 2013;123:1405–12. https://doi.org/10.1172/JCI61398.

[6] Picard M, Zhang J, Hancock S, Derbeneva O, Golhar R, Golik P, et al. Progressive increase in mtDNA 3243A>G heteroplasmy causes abrupt transcriptional reprogramming. Proc Natl Acad Sci U S A 2014;111:E4033-42. https://doi.org/10.1073/pnas.1414028111.

[7] Wallace DC. A mitochondrial paradigm of metabolic and degenerative diseases, aging, and cancer: a dawn for evolutionary medicine. Annu Rev Genet 2005;39:359–407. https://doi.org/10.1146/annurev.genet.39.110304.095751.

[8] Wallace DC, Fan W. Energetics, epigenetics, mitochondrial genetics. Mitochondrion 2010;10:12–31. https://doi.org/10.1016/j.mito.2009.09.006.

[9] Wallace DC. Mitochondrial DNA variation in human radiation and disease. Cell 2015;163:33–8. https://doi.org/10.1016/j.cell.2015.08.067.

[10] Warburg O, Wind F, Negelein E. THE METABOLISM OF TUMORS IN THE BODY. J Gen Physiol 1927;8:519–30. https://doi.org/10.1085/jgp.8.6.519.

[11] Hu J, Locasale JW, Bielas JH, O'Sullivan J, Sheahan K, Cantley LC, et al. Heterogeneity of tumor-induced gene expression changes in the human metabolic network. Nat Biotechnol 2013;31:522–9. https://doi.org/10.1038/nbt.2530.

[12] Roesch A, Vultur A, Bogeski I, Wang H, Zimmermann KM, Speicher D, et al. Overcoming intrinsic multidrug resistance in melanoma by blocking the mitochondrial respiratory chain of slow-cycling JARID1B(high) cells. Cancer Cell 2013;23:811–25. https://doi.org/10.1016/j.ccr.2013.05.003.

[13] Sriskanthadevan S, Jeyaraju D V, Chung TE, Prabha S, Xu W, Skrtic M, et al. AML cells have low spare reserve capacity in their respiratory chain that renders them susceptible to oxidative metabolic stress. Blood 2015;125:2120–30. https://doi.org/10.1182/blood-2014-08-594408.

[14] Birsoy K, Wang T, Chen WW, Freinkman E, Abu-Remaileh M, Sabatini DM. An Essential Role of the Mitochondrial Electron Transport Chain in Cell Proliferation Is to Enable Aspartate Synthesis. Cell 2015;162:540–51. https://doi.org/10.1016/j.cell.2015.07.016.

[15] Tan AS, Baty JW, Dong L-F, Bezawork-Geleta A, Endaya B, Goodwin J, et al. Mitochondrial genome acquisition restores respiratory function and tumorigenic potential of cancer cells without mitochondrial DNA. Cell Metab 2015;21:81–94. https://doi.org/10.1016/j.cmet.2014.12.003.

[16] Wang Y, Liu VWS, Xue WC, Cheung ANY, Ngan HYS. Association of decreased mitochondrial DNA content with ovarian cancer progression. Br J Cancer 2006;95:1087–91. https://doi.org/10.1038/sj.bjc.6603377.
[17] Sica V, Bravo-San Pedro JM, Stoll G, Kroemer G. Oxidative phosphorylation as a potential therapeutic target for cancer therapy. Int J Cancer 2020;146:10–7. https://doi.org/10.1002/ijc.32616.
[18] Ashton TM, McKenna WG, Kunz-Schughart LA, Higgins GS. Oxidative Phosphorylation as an Emerging Target in Cancer Therapy. Clin Cancer Res 2018;24:2482–90. https://doi.org/10.1158/1078-0432.CCR-17-3070.
[19] Wu Y, Zhang X, Wang Z, Zheng W, Cao H, Shen W. Targeting oxidative phosphorylation as an approach for the treatment of ovarian cancer. Front Oncol 2022;12:971479. https://doi.org/10.3389/fonc.2022.971479.
[20] Nayak AP, Kapur A, Barroilhet L, Patankar MS. Oxidative Phosphorylation: A Target for Novel Therapeutic Strategies Against Ovarian Cancer. Cancers (Basel) 2018;10. https://doi.org/10.3390/cancers10090337.
[21] Siegel RL, Miller KD, Wagle NS, Jemal A. Cancer statistics, 2023. CA Cancer J Clin 2023;73:17–48. https://doi.org/10.3322/caac.21763.
[22] Javadi S, Ganeshan DM, Qayyum A, Iyer RB, Bhosale P. Ovarian Cancer, the Revised FIGO Staging System, and the Role of Imaging. AJR Am J Roentgenol 2016;206:1351–60. https://doi.org/10.2214/AJR.15.15199.
[23] Shen L, Zhan X. Mitochondrial Dysfunction Pathway Alterations Offer Potential Biomarkers and Therapeutic Targets for Ovarian Cancer. Oxid Med Cell Longev 2022;2022:5634724. https://doi.org/10.1155/2022/5634724.
[24] Li N, Li H, Wang Y, Cao L, Zhan X. Quantitative proteomics revealed energy metabolism pathway alterations in human epithelial ovarian carcinoma and their regulation by the antiparasite drug ivermectin: data interpretation in the context of 3P medicine. EPMA J 2020;11:661–94. https://doi.org/10.1007/s13167-020-00224-z.
[25] Li N, Zhan X. Signaling pathway network alterations in human ovarian cancers identified with quantitative mitochondrial proteomics. EPMA J 2019;10:153–72. https://doi.org/10.1007/s13167-019-00170-5.
[26] Earp MA, Brooks-Wilson A, Cook L, Le N. Inherited common variants in mitochondrial DNA and invasive serous epithelial ovarian cancer risk. BMC Res Notes 2013;6:425. https://doi.org/10.1186/1756-0500-6-425.
[27] Huang R, Wang J, Zhong Y, Liu Y, Stokke T, Trope CG, et al. Mitochondrial DNA Deficiency in Ovarian Cancer Cells and Cancer Stem Cell-like Properties. Anticancer Res 2015;35:3743–53.
[28] Andrews S. FastQC: a quality control tool for high throughput sequence data 2010.
[29] Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 2014;30:2114–20. https://doi.org/10.1093/bioinformatics/btu170.
[30] Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 2009;25:1754–60. https://doi.org/10.1093/bioinformatics/btp324.
[31] Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, del Angel G, Levy-Moonshine A, et al. From FastQ Data to High-Confidence Variant Calls: The

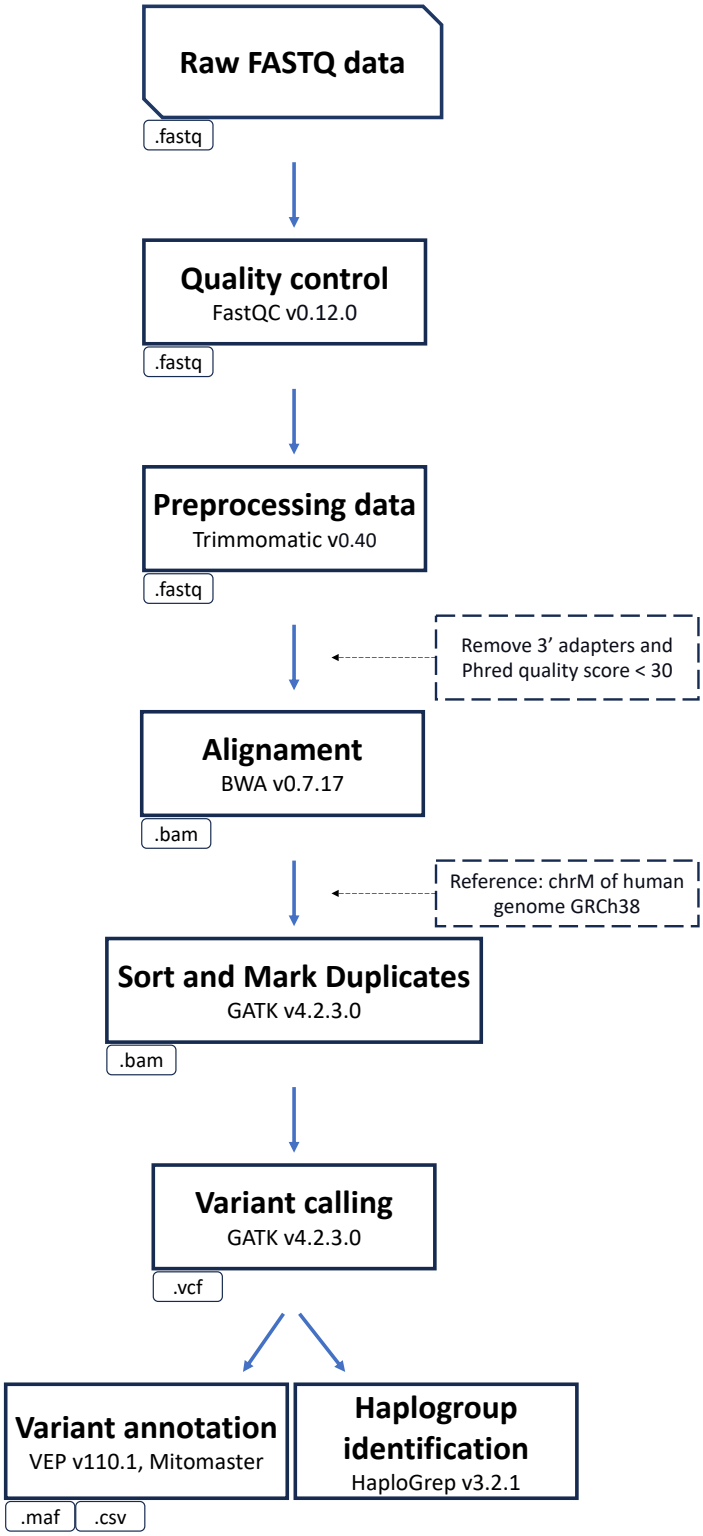
Genome Analysis Toolkit Best Practices Pipeline. Curr Protoc Bioinformatics 2013;43. https://doi.org/10.1002/0471250953.bi1110s43.
[32] McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, et al. The Ensembl Variant Effect Predictor. Genome Biol 2016;17:122. https://doi.org/10.1186/s13059-016-0974-4.
[33] Lott MT, Leipzig JN, Derbeneva O, Xie HM, Chalkia D, Sarmady M, et al. mtDNA Variation and Analysis Using Mitomap and Mitomaster. Curr Protoc Bioinformatics 2013;44:1.23.1-26. https://doi.org/10.1002/0471250953.bi0123s44.
[34] Schönherr S, Weissensteiner H, Kronenberg F, Forer L. Haplogrep 3 - an interactive haplogroup classification and analysis platform. Nucleic Acids Res 2023;51:W263–8. https://doi.org/10.1093/nar/gkad284.
[35] Birsoy K, Wang T, Chen WW, Freinkman E, Abu-Remaileh M, Sabatini DM. An Essential Role of the Mitochondrial Electron Transport Chain in Cell Proliferation Is to Enable Aspartate Synthesis. Cell 2015;162:540–51. https://doi.org/10.1016/j.cell.2015.07.016.
[36] Lim HY, Ho QS, Low J, Choolani M, Wong KP. Respiratory competent mitochondria in human ovarian and peritoneal cancer. Mitochondrion 2011;11:437–43. https://doi.org/10.1016/j.mito.2010.12.015.
[37] Imanishi H, Hattori K, Wada R, Ishikawa K, Fukuda S, Takenaga K, et al. Mitochondrial DNA mutations regulate metastasis of human breast cancer cells. PLoS One 2011;6:e23401. https://doi.org/10.1371/journal.pone.0023401.
[38] Grandhi S, Bosworth C, Maddox W, Sensiba C, Akhavanfard S, Ni Y, et al. Heteroplasmic shifts in tumor mitochondrial genomes reveal tissue-specific signals of relaxed and positive selection. Hum Mol Genet 2017;26:2912–22. https://doi.org/10.1093/hmg/ddx172.
[39] Yuan Y, Ju YS, Kim Y, Li J, Wang Y, Yoon CJ, et al. Comprehensive molecular characterization of mitochondrial genomes in human cancers. Nat Genet 2020;52:342–52. https://doi.org/10.1038/s41588-019-0557-x.
[40] Coskun PE, Beal MF, Wallace DC. Alzheimer's brains harbor somatic mtDNA control-region mutations that suppress mitochondrial transcription and replication. Proc Natl Acad Sci U S A 2004;101:10726–31. https://doi.org/10.1073/pnas.0403649101.
[41] Baker KT, Nachmanson D, Kumar S, Emond MJ, Ussakli C, Brentnall TA, et al. Mitochondrial DNA Mutations are Associated with Ulcerative Colitis Preneoplasia but Tend to be Negatively Selected in Cancer. Mol Cancer Res 2019;17:488–98. https://doi.org/10.1158/1541-7786.MCR-18-0520.
[42] Bragoszewski P, Kupryjanczyk J, Bartnik E, Rachinger A, Ostrowski J. Limited clinical relevance of mitochondrial DNA mutation and gene expression analyses in ovarian cancer. BMC Cancer 2008;8:292. https://doi.org/10.1186/1471-2407-8-292.
[43] Lin Y-H, Chu Y-D, Lim S-N, Chen C-W, Yeh C-T, Lin W-R. Impact of an MT-RNR1 Gene Polymorphism on Hepatocellular Carcinoma Progression and Clinical Characteristics. Int J Mol Sci 2021;22. https://doi.org/10.3390/ijms22031119.
[44] Baum H, Rieske JS, Silman HI, Lipton SH. ON THE MECHANISM OF ELECTRON TRANSFER IN COMPLEX III OF THE ELECTRON TRANSFER

CHAIN. Proceedings of the National Academy of Sciences 1967;57:798–805. https://doi.org/10.1073/pnas.57.3.798.
[45] Gil Borlado MC, Moreno Lastres D, Gonzalez Hoyuela M, Moran M, Blazquez A, Pello R, et al. Impact of the Mitochondrial Genetic Background in Complex III Deficiency. PLoS One 2010;5:e12801. https://doi.org/10.1371/journal.pone.0012801.
[46] Toure S, Mbaye F, Gueye MD, Fall M, Dem A, Sembene M. Somatic Mitochondrial Mutations in Oral Cavity Cancers among Senegalese Patients. Asian Pacific Journal of Cancer Prevention 2019;20:2203–8. https://doi.org/10.31557/APJCP.2019.20.7.2203.
[47] Kozakiewicz P, Grzybowska-Szatkowska L, Ciesielka M, Całka P, Osuchowski J, Szmygin P, et al. Mitochondrial DNA Changes in Genes of Respiratory Complexes III, IV and V Could Be Related to Brain Tumours in Humans. Int J Mol Sci 2022;23:12131. https://doi.org/10.3390/ijms232012131.
[48] Brandt U. Energy Converting NADH: Quinone Oxidoreductase (Complex I). Annu Rev Biochem 2006;75:69–92. https://doi.org/10.1146/annurev.biochem.75.103004.142539.
[49] Beckman KB, Ames BN. Oxidative Decay of DNA. Journal of Biological Chemistry 1997;272:19633–6. https://doi.org/10.1074/jbc.272.32.19633.
[50] Xu Y, Zhou J, Yuan Q, Su J, Li Q, Lu X, et al. Quantitative detection of circulating MT-ND1 as a potential biomarker for colorectal cancer. Bosn J Basic Med Sci 2021. https://doi.org/10.17305/bjbms.2021.5576.
[51] Grzybowska-Szatkowska L, Ślaska B. Mitochondrial NADH dehydrogenase polymorphisms are associated with breast cancer in Poland. J Appl Genet 2014;55:173–81. https://doi.org/10.1007/s13353-013-0190-9.
[52] Kim H, Komiyama T, Inomoto C, Kamiguchi H, Kajiwara H, Kobayashi H, et al. Mutations in the Mitochondrial ND1 Gene Are Associated with Postoperative Prognosis of Localized Renal Cell Carcinoma. Int J Mol Sci 2016;17:2049. https://doi.org/10.3390/ijms17122049.
[53] Moindjie H, Rodrigues-Ferreira S, Nahmias C. Mitochondrial Metabolism in Carcinogenesis and Cancer Therapy. Cancers (Basel) 2021;13. https://doi.org/10.3390/cancers13133311.
[54] Dong L, Neuzil J. Targeting mitochondria as an anticancer strategy. Cancer Commun (Lond) 2019;39:63. https://doi.org/10.1186/s40880-019-0412-6.
[55] Ayala FJ. The myth of Eve: molecular biology and human origins. Science 1995;270:1930–6. https://doi.org/10.1126/science.270.5244.1930.
[56] Luo Y, Gao W, Chen Y, Liu F, Gao Y. Rare mitochondrial DNA polymorphisms are associated with high altitude pulmonary edema (HAPE) susceptibility in Han Chinese. Wilderness Environ Med 2012;23:128–32. https://doi.org/10.1016/j.wem.2012.02.003.
[57] Nishigaki Y, Yamada Y, Fuku N, Matsuo H, Segawa T, Watanabe S, et al. Mitochondrial haplogroup N9b is protective against myocardial infarction in Japanese males. Hum Genet 2007;120:827–36. https://doi.org/10.1007/s00439-006-0269-z.
[58] Fang H, Shen L, Chen T, He J, Ding Z, Wei J, et al. Cancer type-specific modulation of mitochondrial haplogroups in breast, colorectal and thyroid cancer. BMC Cancer 2010;10:421. https://doi.org/10.1186/1471-2407-10-421.

[59] Sharma S, Singh S, Gupta RK, Ganju L, Singh SB, Kumar B, et al. Mitochondrial DNA sequencing reveals association of variants and haplogroup M33a2'3 with High altitude pulmonary edema susceptibility in Indian male lowlanders. Sci Rep 2019;9:10975. https://doi.org/10.1038/s41598-019-47500-1.

SUPPLEMENTARY MATERIAL

Supplementary File 1 – Fluxogram of Bioinformatics steps.



Supplementary File 2 – Variants identified in this study. The occurrence of variants found in each sample.

#CHROM	POS	REF	ALT	S1	S2	S3	S4	S5	S6	S7
NC_012920.1	57	TTT	AAC					x		
NC_012920.1	64	C	T					x	x	
NC_012920.1	73	A	G	x	x	x	x	x	x	x
NC_012920.1	143	G	A							x
NC_012920.1	146	T	C					x	x	x
NC_012920.1	152	TA	CA,CG	x						x
NC_012920.1	153	A	G					x	x	
NC_012920.1	195	T	C							x
NC_012920.1	210	A	G		x					
NC_012920.1	235	AT	GT,GC					x	x	
NC_012920.1	263	A	G	x	x	x	x		x	x
NC_012920.1	302	A	AC							x
NC_012920.1	310	T	C,TC	x	x	x	x		x	x
NC_012920.1	372	T	C			x				
NC_012920.1	513	GCA	G	x		x			x	
NC_012920.1	663	A	G						x	
NC_012920.1	750	A	G	x	x	x	x	x	x	x
NC_012920.1	769	G	A							x
NC_012920.1	921	T	C	x						
NC_012920.1	1018	G	A							x
NC_012920.1	1438	A	G	x	x	x	x		x	x
NC_012920.1	1736	A	G						x	
NC_012920.1	2092	C	T		x					
NC_012920.1	2416	T	C							x
NC_012920.1	2483	T	C							x
NC_012920.1	2652	G	A		x					
NC_012920.1	2706	A	G	x	x		x			

NC_012920.1	2789	C	T							x
NC_012920.1	3010	G	A		x					
NC_012920.1	3169	C	T				x			
NC_012920.1	3348	A	G				x			
NC_012920.1	3450	C	T			x				
NC_012920.1	3594	C	T							x
NC_012920.1	3796	A	G	x						
NC_012920.1	4883	C	T		x					
NC_012920.1	5471	G	A				x			
NC_012920.1	5773	G	A			x				
NC_012920.1	6221	T	C			x				
NC_012920.1	6680	T	C	x						
NC_012920.1	7028	C	T		x	x	x			
NC_012920.1	7648	C	T	x						
NC_012920.1	7771	A	G							x
NC_012920.1	7805	G	A				x			
NC_012920.1	8027	G	A						x	
NC_012920.1	8061	T	C				x			
NC_012920.1	8206	G	A							x
NC_012920.1	8407	C	A				x			
NC_012920.1	8414	C	T		x					
NC_012920.1	8557	G	C				x			
NC_012920.1	8701	A	G	x	x	x				
NC_012920.1	9449	C	T			x				
NC_012920.1	9540	T	C	x	x	x				x
NC_012920.1	10086	A	G			x				
NC_012920.1	10115	T	C							x
NC_012920.1	10373	G	A			x				
NC_012920.1	10398	A	G	x	x	x				x
NC_012920.1	10400	C	T		x					
NC_012920.1	10463	T	C	x						

NC_012920.1	10873	T	C	x	x	x				x
NC_012920.1	11002	A	G			x				
NC_012920.1	11150	G	A		x					
NC_012920.1	11233	T	C		x					
NC_012920.1	11467	A	G				x			
NC_012920.1	11719	G	A	x	x	x	x		x	
NC_012920.1	11887	G	A	x						
NC_012920.1	11914	G	A							x
NC_012920.1	11944	T	C						x	x
NC_012920.1	12007	G	A						x	
NC_012920.1	12236	GC	G	x						
NC_012920.1	12308	A	G				x			
NC_012920.1	12372	G	A				x			
NC_012920.1	12501	G	A				x			
NC_012920.1	12618	G	A				x			
NC_012920.1	12693	A	G							x
NC_012920.1	12705	C	T	x	x	x			x	x
NC_012920.1	12819	A	T		x					
NC_012920.1	13105	A	G	x		x				
NC_012920.1	13440	C	T				x			
NC_012920.1	13590	G	A							x
NC_012920.1	13650	C	T							x
NC_012920.1	13803	A	G							x
NC_012920.1	13886	T	C	x						
NC_012920.1	13914	C	A			x				
NC_012920.1	14005	T	C		x					
NC_012920.1	14179	A	G				x			
NC_012920.1	14284	C	T	x						
NC_012920.1	14566	A	G							x
NC_012920.1	14668	C	T		x					
NC_012920.1	14693	A	G	x						

NC_012920.1	14766	C	T	x	x	x	x		x	x
NC_012920.1	14985	G	A				x			
NC_012920.1	15061	A	G	x						
NC_012920.1	15106	G	A		x					
NC_012920.1	15301	G	A	x	x	x				x
NC_012920.1	15326	A	G	x	x	x	x		x	x
NC_012920.1	15784	T	C							x
NC_012920.1	15824	A	G			x				
NC_012920.1	15939	CT	C			x				
NC_012920.1	16093	T	C			x				
NC_012920.1	16111	C	T					x	x	
NC_012920.1	16124	T	C	x		x				
NC_012920.1	16172	T	C				x			
NC_012920.1	16189	T	C							x
NC_012920.1	16209	T	C			x				
NC_012920.1	16218	C	T						x	
NC_012920.1	16219	A	G				x			
NC_012920.1	16223	C	T	x	x	x		x	x	x
NC_012920.1	16239	C	T,A		x				x	
NC_012920.1	16266	C	T						x	
NC_012920.1	16278	C	T			x	x			x
NC_012920.1	16290	CC	TC,TT					x	x	
NC_012920.1	16294	C	T							x
NC_012920.1	16300	A	G	x						
NC_012920.1	16309	A	G							x
NC_012920.1	16319	G	A	x				x	x	
NC_012920.1	16362	T	C		x			x	x	
NC_012920.1	16390	G	A							x
NC_012920.1	16519	T	C			x				x

Supplementary File 3 – Mitomap annotation. All variants identified in the study separated by samples.

Query	rCRS Position	rCRS NT	Query NT	Alt name †	Mut type	Locus	Other	GB SeqsCount (Freq)*‡	gnomAD 3.1Frequency‡‡(count/total)	Helix Frequency‡‡in 195983 seqs	Refs	Conservation	Patient Report◊
S1	73	A	G		transition	CR:HVS2ATTCR:Control RegionCR:7S-like	Non-coding	FL:45883 (75.011%)CR:42505 (52.937%)	40199/5616471.574%	11993 461%	153 refs	NA	
S1	152	TA	CA		transversion	CR:HVS2CR:OHATTCR:Control RegionCR:7S-like	non-coding	FL:0 (0.000%)CR:0 (0.000%)	NR	NR		NA	
S1	263	A	G		transition	CR:HVS2CR:OHATTCR:Control Region	non-coding	FL:58432 (95.527%)CR:45942 (57.217%)	55895/5642099.070%	19386 199%	151 refs	NA	

S1	310	T	C		transition	CR:HVS2CR:OHCR:CSB2ATTCR:Control Region	non-coding	FL:24266 (39.671%)CR:23172 (28.859%)	375/70805.297%	NR	34 refs	NA	Status: ReportedPossible protective factor for normal tension glaucoma
S1	513	GCA	G		deletion	CR:HVS3CR:Control Region	non-coding	FL:0 (0.000%)CR:0 (0.000%)	NR	NR		NA	
S1	750	A	G		transition	12S	rRNA	FL:60122 (98.290%)CR:0 (0.000%)	55419/5635498.341%	19298698%	114 refs	98%	
S1	921	T	C		transition	12S	rRNA	FL:435 (0.711%)CR:0 (0.000%)	986/564271.747%	5570%	15 refs	67%	Status: ReportedPossibly LVNC-associated
S1	1438	A	G		transition	12S	rRNA	FL:58231 (95.198%)CR:0 (0.000%)	53745/5621395.610%	18988997%	124 refs	87%	
S1	2706	A	G		transition	16S	rRNA	FL:47574 (77.776%)CR:0 (0.000%)	41427/5602173.949%	12413563%	124 refs	84%	

S1	3796	A	G		transition	ND1	ND1:T164A	FL:290 (0.474%)CR:0 (0.000%)	466/5642 40.826%	19801 %	19 refs	71 %	Status: ReportedAdult-Onset Dystonia
S1	6680	T	C		transition	COI	COI:T259T	FL:1223 (1.999%)CR:0 (0.000%)	849/5642 21.505%	11381 %	26 refs	84 %	
S1	7648	C	T		transition	COII	COII:I21I	FL:180 (0.294%)CR:0 (0.000%)	140/5643 20.248%	960%	2 refs	20 %	
S1	8701	A	G		transition	ATPase6	ATPase6:T59A	FL:19100 (31.225%)CR:0 (0.000%)	17065/56 28530.31 9%	17154 9%	105 refs	42 %	
S1	9540	T	C		transition	COIII	COIII:L112L	FL:19059 (31.158%)CR:0 (0.000%)	17141/56 32330.43 3%	17229 9%	81 refs	98 %	
S1	10398	A	G		transition	ND3	ND3:T114A	FL:25866 (42.287%)CR:0 (0.000%)	23506/56 17041.84 8%	49322 25%	224 refs	51 %	Status: Reported; lineage L & M marker, also hg IJKPD protective

													factor / longevity / altered cell pH / metabolic syndrome / breast cancer risk / Leigh Syndrome risk / ADHD / cognitive decline / SCA2 age of onset / Fuchs endothelial corneal dystrophy
S1	10463	T	C		transition	R	MitoTIP4.80%	FL:3121 (5.102%)CR:0 (0.000%)	3274/564 285.802 %	17572 9%	48 refs	80 %	

S1	10873	T	C		transition	ND4	ND4:P38P	FL:19056 (31.154%)CR:0 (0.000%)	17166/56 29130.49 5%	17312 9%	98 refs	20 %	
S1	11719	G	A		transition	ND4	ND4:G320G	FL:46499 (76.019%)CR:0 (0.000%)	39882/56 09071.10 4%	11553 459%	135 refs	98 %	
S1	11887	G	A		transition	ND4	ND4:L376L	FL:273 (0.446%)CR:0 (0.000%)	225/5641 40.399%	3130 %	10 refs	31 %	
S1	12236	GC	G		deletion	S(AGY)		FL:0 (0.000%)CR:0 (0.000%)	NR	NR		NA	
S1	12705	C	T		transition	ND5	ND5:I123I	FL:24187 (39.542%)CR:0 (0.000%)	22373/45 77248.87 9%	35323 18%	115 refs	29 %	Status: ReportedPossible protective factor for normal tension glaucoma
S1	13105	A	G		transition	ND5	ND5:I257V	FL:4134 (6.758%)CR:0 (0.000%)	6556/514 2112.750 %	48822 %	36 refs	44 %	

S1	13886	T	C		transition	ND5	ND5:L517P	FL:458 (0.749%)CR:0 (0.000%)	1024/564 231.815 %	6150 %	5 refs	7%	
S1	14284	C	T		transition	ND6	ND6:E130E	FL:450 (0.736%)CR:0 (0.000%)	997/5643 21.767%	5750 %	4 refs	22 %	
S1	14693	A	G		transition	E	MitoTIP39.50%	FL:343 (0.561%)CR:0 (0.000%)	109/5642 60.193%	3660 %	32 refs	91 %	Status: ReportedMELAS / LHON / DEAF / hypertension helper
S1	14766	C	T		transition	Cytb	Cytb:T7I	FL:46188 (75.510%)CR:0 (0.000%)	39641/56 08670.67 9%	11470 059%	117 refs	49 %	
S1	15061	A	G		transition	Cytb	Cytb:G105G	FL:242 (0.396%)CR:0 (0.000%)	162/5642 80.287%	2520 %	10 refs	98 %	
S1	15301	G	A		transition	Cytb	Cytb:L185L	FL:16549 (27.055%)CR:0 (0.000%)	13722/56 09624.46 2%	15883 8%	91 refs	96 %	Status: ReportedPossible factor in high altitude

													sickness, LHON
S1	15326	A	G		transition	Cytb	Cytb:T194A	FL:60315 (98.605%)CR:0 (0.000%)	56027/56 39899.34 2%	19397 099%	120 ref s	18 %	
S1	16124	T	C		transition	ATTCR:Control RegionCR:HVS 1/HV1CR:TAS2 CR:7S-like	non-coding	FL:879 (1.437%)CR:1207 (1.503%)	1993/562 853.541 %	15321 %	51 refs	NA	
S1	16223	C	T		transition	ATTCR:Control RegionCR:HVS 1/HV1CR:7S-like	non-coding	FL:23015 (37.626%)CR:4059 0 (50.552%)	22182/56 28639.40 9%	35639 18%	186 ref s	NA	
S1	16300	A	G		transition	ATTCR:Control RegionCR:HVS 1/HV1CR:7S-like	non-coding	FL:311 (0.508%)CR:525 (0.654%)	279/5642 90.494%	5680 %	42 refs	NA	Status: Repo rtedBD-associated
S1	16319	G	A		transition	ATTCR:Control RegionCR:HVS 1/HV1CR:7S-like	non-coding	FL:3549 (5.802%)CR:7073 (8.809%)	3513/564 106.228 %	91035 %	134 ref s	NA	

S2	73	A	G		transition	CR:HVS2ATTCR:Control RegionCR:7S-like	non-coding	FL:45883 (75.011%)CR:42505 (52.937%)	40199/5616471.574%	11993461%	153 refs	NA	
S2	210	A	G		transition	CR:HVS2CR:OHATTCR:Control Region	non-coding	FL:733 (1.198%)CR:729 (0.908%)	112/564280.198%	6320%	24 refs	NA	
S2	263	A	G		transition	CR:HVS2CR:OHATTCR:Control Region	non-coding	FL:58432 (95.527%)CR:45942 (57.217%)	55895/5642099.070%	19386199%	151 refs	NA	
S2	310	T	C		transition	CR:HVS2CR:OHCR:CSB2ATTCR:Control Region	non-coding	FL:24266 (39.671%)CR:23172 (28.859%)	375/70805.297%	NR	34 refs	NA	Status: ReportedPossible protective factor for normal tension glaucoma
S2	750	A	G		transition	12S	rRNA	FL:60122 (98.290%)CR:0 (0.000%)	55419/5635498.341%	19298698%	114 refs	98%	

S2	1438	A	G		transition	12S	rRNA	FL:58231 (95.198%)CR:0 (0.000%)	53745/56 21395.61 0%	18988 997%	124 refs	87 %	
S2	2092	C	T		transition	16S	rRNA	FL:501 (0.819%)CR:0 (0.000%)	216/5643 10.383%	8750 %	12 refs	18 %	
S2	2652	G	A		transition	16SHumanin	rRNA	FL:13 (0.021%)CR:0 (0.000%)	12/56422 0.021%	370%	1 ref	18 %	
S2	2706	A	G		transition	16S	rRNA	FL:47574 (77.776%)CR:0 (0.000%)	41427/56 02173.94 9%	12413 563%	124 refs	84 %	
S2	3010	G	A		transition	16S	rRNA	FL:9385 (15.343%)CR:0 (0.000%)	9208/563 5016.341 %	41014 21%	110 refs	20 %	Status: ReportedCyclic Vomiting Syndrome with Migraine / high altitude adaptation
S2	4883	C	T		transition	ND2	ND2:P138P	FL:2774 (4.535%)CR:0 (0.000%)	615/5642 71.090%	22951 %	56 refs	10 0%	Status: Conflicting reportsGlaucoma

S2	7028	C	T		transition	COI	COI:A375A	FL:48522 (79.326%)CR:0 (0.000%)	41590/55 49874.94 0%	12348 563%	165 refs	10 0%	
S2	8414	C	T		transition	ATPase8	ATPase8:L17F	FL:2307 (3.772%)CR:0 (0.000%)	507/5643 30.898%	19761 %	60 refs	31 %	Status: ReportedIncreased risk of T2DM and high altitude polycythemia (HAPC) in haplogroup D4 / Longevity
S2	8701	A	G		transition	ATPase6	ATPase6:T59A	FL:19100 (31.225%)CR:0 (0.000%)	17065/56 28530.31 9%	17154 9%	105 refs	42 %	
S2	9540	T	C		transition	COIII	COIII:L112L	FL:19059 (31.158%)CR:0 (0.000%)	17141/56 32330.43 3%	17229 9%	81 refs	98 %	
S2	10398	A	G		transition	ND3	ND3:T114A	FL:25866 (42.287%)CR:0 (0.000%)	23506/56 17041.84 8%	49322 25%	224 refs	51 %	Status: Reported; lineage L & M

													marker, also hg IJKPD protective factor / longevity / altered cell pH / metabolic syndrome / breast cancer risk / Leigh Syndrome risk / ADHD / cognitive decline / SCA2 age of onset / Fuchs endothelial corneal dystrophy

S2	10400	C	T		transition	ND3	ND3:T114T	FL:12357 (20.202%)CR:0 (0.000%)	2974/563795.275%	90365%	108 refs	51%	
S2	10873	T	C		transition	ND4	ND4:P38P	FL:19056 (31.154%)CR:0 (0.000%)	17166/5629130.495%	173129%	98 refs	20%	
S2	11150	G	A		transition	ND4	ND4:A131T	FL:152 (0.248%)CR:0 (0.000%)	190/564060.337%	3610%	6 refs	7%	Status: ReportedFound in a Multiple Sclerosis patient
S2	11233	T	C		transition	ND4	ND4:L158L	FL:33 (0.054%)CR:0 (0.000%)	38/5642700.067%	1480%	4 refs	93%	
S2	11719	G	A		transition	ND4	ND4:G320G	FL:46499 (76.019%)CR:0 (0.000%)	39882/5609071.104%	11553459%	135 refs	98%	
S2	12705	C	T		transition	ND5	ND5:I123I	FL:24187 (39.542%)CR:0 (0.000%)	22373/4577248.879%	3532318%	115 refs	29%	Status: ReportedPossible protective factor for normal

													tension glaucoma
S2	12819	A	T		transversion	ND5	ND5:R161R	FL:0 (0.000%)CR:0 (0.000%)	NR	NR		96 %	
S2	14005	T	C		transition	ND5	ND5:W557R	FL:0 (0.000%)CR:0 (0.000%)	NR	NR		10 0%	
S2	14668	C	T		transition	ND6	ND6:M2M	FL:2408 (3.937%)CR:0 (0.000%)	550/5643 20.975%	20901 %	49 refs	24 %	Status: Repo rtedDepressi ve Disorder associated
S2	14766	C	T		transition	Cytb	Cytb:T7I	FL:46188 (75.510%)CR:0 (0.000%)	39641/56 08670.67 9%	11470 059%	117 ref s	49 %	
S2	15106	G	A		transition	Cytb	Cytb:L120L	FL:162 (0.265%)CR:0 (0.000%)	59/56418 0.105%	2160 %	8 refs	10 0%	
S2	15301	G	A		transition	Cytb	Cytb:L185L	FL:16549 (27.055%)CR:0 (0.000%)	13722/56 09624.46 2%	15883 8%	91 refs	96 %	Status: Repo rtedPossible factor in high altitude sickness, LHON

S2	15326	A	G		transition	Cytb	Cytb:T194A	FL:60315 (98.605%)CR:0 (0.000%)	56027/56 39899.34 2%	19397 099%	120 ref s	18 %	
S2	16223	C	T		transition	ATTCR:Control RegionCR:HVS 1/HV1CR:7S-like	non-coding	FL:23015 (37.626%)CR:4059 0 (50.552%)	22182/56 28639.40 9%	35639 18%	186 ref s	NA	
S2	16239	C	T		transition	ATTCR:Control RegionCR:HVS 1/HV1CR:7S-like	non-coding	FL:568 (0.929%)CR:710 (0.884%)	744/5640 81.319%	22021 %	59 refs	NA	
S2	16362	T	C		transition	ATTCR:Control RegionCR:HVS 1/HV1CR:7S-like	non-coding	FL:10255 (16.765%)CR:2172 1 (27.052%)	8423/562 3414.979 %	23527 12%	168 ref s	NA	
S3	73	A	G		transition	CR:HVS2ATTC R:Control RegionCR:7S-like	non-coding	FL:45883 (75.011%)CR:4250 5 (52.937%)	40199/56 16471.57 4%	11993 461%	153 ref s	NA	
S3	263	A	G		transition	CR:HVS2CR:O HATTCR:Contro l Region	non-coding	FL:58432 (95.527%)CR:4594 2 (57.217%)	55895/56 42099.07 0%	19386 199%	151 ref s	NA	

S3	310	T	C		transition	CR:HVS2CR:OHCR:CSB2ATTCR:Control Region	non-coding	FL:24266 (39.671%)CR:23172 (28.859%)	375/70805.297%	NR	34 refs	NA	Status: ReportedPossible protective factor for normal tension glaucoma
S3	372	T	C		transition	CR:HVS2CR:OHCR:mt4HATTCR:Control Region	non-coding	FL:205 (0.335%)CR:35 (0.044%)	60/564330.106%	1280%	2 refs	NA	
S3	513	GCA	G		deletion	CR:HVS3CR:Control Region	non-coding	FL:0 (0.000%)CR:0 (0.000%)	NR	NR		NA	
S3	750	A	G		transition	12S	rRNA	FL:60122 (98.290%)CR:0 (0.000%)	55419/5635498.341%	19298698%	114 refs	98%	
S3	1438	A	G		transition	12S	rRNA	FL:58231 (95.198%)CR:0 (0.000%)	53745/5621395.610%	18988997%	124 refs	87%	
S3	3450	C	T		transition	ND1	ND1:P48P	FL:446 (0.729%)CR:0 (0.000%)	1123/564301.990%	5910%	3 refs	100%	

S3	5773	G	A		transition	OLC	MitoTIP3.50%	FL:801 (1.310%)CR:0 (0.000%)	1279/564 242.267 %	16091 %	34 refs	36 %	
S3	6221	T	C		transition	COI	COI:P106P	FL:1774 (2.900%)CR:0 (0.000%)	2731/563 714.845 %	42982 %	23 refs	98 %	
S3	7028	C	T		transition	COI	COI:A375A	FL:48522 (79.326%)CR:0 (0.000%)	41590/55 49874.94 0%	12348 563%	165 refs	10 0%	
S3	8701	A	G		transition	ATPase6	ATPase6:T59A	FL:19100 (31.225%)CR:0 (0.000%)	17065/56 28530.31 9%	17154 9%	105 refs	42 %	
S3	9449	C	T		transition	COIII	COIII:Y81Y	FL:665 (1.087%)CR:0 (0.000%)	1240/564 282.197 %	8410 %	6 refs	98 %	
S3	9540	T	C		transition	COIII	COIII:L112L	FL:19059 (31.158%)CR:0 (0.000%)	17141/56 32330.43 3%	17229 9%	81 refs	98 %	
S3	10086	A	G		transition	ND3	ND3:N10D	FL:470 (0.768%)CR:0 (0.000%)	1197/564 332.121 %	5730 %	14 refs	87 %	Status: ReportedHypertensive end-

													stage renal disease
S3	10373	G	A		transition	ND3	ND3:E105E	FL:643 (1.051%)CR:0 (0.000%)	925/564131.640%	7680 %	11 refs	100%	
S3	10398	A	G		transition	ND3	ND3:T114A	FL:25866 (42.287%)CR:0 (0.000%)	23506/5617041.848%	4932225%	224 refs	51 %	Status: Reported; lineage L & M marker, also hg IJKPD protective factor / longevity / altered cell pH / metabolic syndrome / breast cancer risk / Leigh Syndrome risk / ADHD / cognitive

													decline / SCA2 age of onset / Fuchs endothelial corneal dystrophy
S3	10873	T	C		transition	ND4	ND4:P38P	FL:19056 (31.154%)CR:0 (0.000%)	17166/56 29130.49 5%	17312 9%	98 refs	20 %	
S3	11002	A	G		transition	ND4	ND4:Q81Q	FL:714 (1.167%)CR:0 (0.000%)	981/5642 11.739%	10241 %	12 refs	98 %	
S3	11719	G	A		transition	ND4	ND4:G320G	FL:46499 (76.019%)CR:0 (0.000%)	39882/56 09071.10 4%	11553 459%	135 ref s	98 %	
S3	12705	C	T		transition	ND5	ND5:I123I	FL:24187 (39.542%)CR:0 (0.000%)	22373/45 77248.87 9%	35323 18%	115 ref s	29 %	Status: Repo rtedPossible protective factor for normal tension glaucoma

S3	13105	A	G		transition	ND5	ND5:I257V	FL:4134 (6.758%)CR:0 (0.000%)	6556/5142112.750 %	48822 %	36 refs	44 %	
S3	13914	C	A		transversion	ND5	ND5:L526L	FL:482 (0.788%)CR:0 (0.000%)	1126/564321.995 %	5800 %	4 refs	100%	
S3	14766	C	T		transition	Cytb	Cytb:T7I	FL:46188 (75.510%)CR:0 (0.000%)	39641/5608670.679%	11470059%	117 refs	49 %	
S3	15301	G	A		transition	Cytb	Cytb:L185L	FL:16549 (27.055%)CR:0 (0.000%)	13722/5609624.462%	158838%	91 refs	96 %	Status: ReportedPossible factor in high altitude sickness, LHON
S3	15326	A	G		transition	Cytb	Cytb:T194A	FL:60315 (98.605%)CR:0 (0.000%)	56027/5639899.342%	19397099%	120 refs	18 %	
S3	15824	A	G		transition	Cytb	Cytb:T360A	FL:444 (0.726%)CR:0 (0.000%)	1117/564331.979 %	5680 %	3 refs	33 %	

S3	15939	CT	C		deletion	ATTT		FL:0 (0.000%)CR:0 (0.000%)	NR	NR		NA	
S3	16093	T	C		transition	ATTCR:Control RegionCR:HVS 1/HV1CR:TAS2	non-coding	FL:3472 (5.676%)CR:5188 (6.461%)	2979/560 935.311 %	50753 %	167 refs	NA	Status: ReportedCyclic Vomiting Syndrome
S3	16124	T	C		transition	ATTCR:Control RegionCR:HVS 1/HV1CR:TAS2 CR:7S-like	non-coding	FL:879 (1.437%)CR:1207 (1.503%)	1993/562 853.541 %	15321 %	51 refs	NA	
S3	16209	T	C		transition	ATTCR:Control RegionCR:HVS 1/HV1CR:7S-like	non-coding	FL:1501 (2.454%)CR:1902 (2.369%)	1635/563 432.902 %	36962 %	90 refs	NA	
S3	16223	C	T		transition	ATTCR:Control RegionCR:HVS 1/HV1CR:7S-like	non-coding	FL:23015 (37.626%)CR:4059 0 (50.552%)	22182/56 28639.40 9%	35639 18%	186 refs	NA	
S3	16278	C	T		transition	ATTCR:Control RegionCR:HVS 1/HV1CR:7S-like	non-coding	FL:6016 (9.835%)CR:7427 (9.250%)	10603/53 98419.64 1%	11334 6%	129 refs	NA	

S3	16519	T	C		transition	ATTCR:Control RegionCR:7S-like	non-coding	FL:38030 (62.173%)CR:33864 (42.175%)	35648/5423865.725%	12599264%	118 refs	NA	
S4	73	A	G		transition	CR:HVS2ATTCR:Control RegionCR:7S-like	non-coding	FL:45883 (75.011%)CR:42505 (52.937%)	40199/5616471.574%	11993461%	153 refs	NA	
S4	263	A	G		transition	CR:HVS2CR:OHATTCR:Control Region	non-coding	FL:58432 (95.527%)CR:45942 (57.217%)	55895/5642099.070%	19386199%	151 refs	NA	
S4	310	T	C		transition	CR:HVS2CR:OHCR:CSB2ATTCR:Control Region	non-coding	FL:24266 (39.671%)CR:23172 (28.859%)	375/70805.297%	NR	34 refs	NA	Status: ReportedPossible protective factor for normal tension glaucoma
S4	750	A	G		transition	12S	rRNA	FL:60122 (98.290%)CR:0 (0.000%)	55419/5635498.341%	19298698%	114 refs	98 %	

S4	1438	A	G		transition	12S	rRNA	FL:58231 (95.198%)CR:0 (0.000%)	53745/5621395.610%	18988997%	124 refs	87%	
S4	2706	A	G		transition	16S	rRNA	FL:47574 (77.776%)CR:0 (0.000%)	41427/5602173.949%	12413563%	124 refs	84%	
S4	3169	C	T		transition	16S	rRNA	FL:7 (0.011%)CR:0 (0.000%)	6/564330.011%	270%	1 ref	31%	
S4	3348	A	G		transition	ND1	ND1:L14L	FL:583 (0.953%)CR:0 (0.000%)	262/564280.464%	5910%	12 refs	96%	
S4	5471	G	A		transition	ND2	ND2:T334T	FL:755 (1.234%)CR:0 (0.000%)	545/563420.967%	20381%	17 refs	71%	
S4	7028	C	T		transition	COI	COI:A375A	FL:48522 (79.326%)CR:0 (0.000%)	41590/5549874.940%	12348563%	165 refs	100%	
S4	7805	G	A		transition	COII	COII:V74I	FL:483 (0.790%)CR:0 (0.000%)	365/564240.647%	8120%	8 refs	18%	
S4	8061	T	C		transition	COII	COII:V159A	FL:0 (0.000%)CR:0 (0.000%)	0	NR		100%	

S4	8407	C	A		transversion	ATPase8	ATPase8:T14T	FL:11 (0.018%)CR:0 (0.000%)	4/564290.007%	100%	1 ref	18%	
S4	8557	G	C		transversion	ATPase8ATPase6	ATPase8:L64FATPase6:A11P	FL:25 (0.041%)CR:0 (0.000%)	7/564210.012%	250%	2 refs	71.11%17.78%	
S4	11467	A	G		transition	ND4	ND4:L236L	FL:7880 (12.883%)CR:0 (0.000%)	8766/5634315.558%	3923120%	63 refs	100%	Status: ReportedAltered brain pH / sCJD patients
S4	11719	G	A		transition	ND4	ND4:G320G	FL:46499 (76.019%)CR:0 (0.000%)	39882/5609071.104%	11553459%	135 refs	98%	
S4	12308	A	G		transition	L(CUN)	MitoTIP42.00%	FL:7860 (12.850%)CR:0 (0.000%)	8754/5633615.539%	3918320%	108 refs	96%	Status: Reported [B]CPEO / Stroke / CM / Breast & Renal & Prostate Cancer Risk /

													Altered brain pH /sCJD
S4	12372	G	A		transition	ND5	ND5:L12L	FL:8446 (13.808%)CR:0 (0.000%)	8901/5635315.795%	3960420%	103 refs	80%	Status: ReportedAltered brain pH / sCJD patients
S4	12501	G	A		transition	ND5	ND5:M55M	FL:1595 (2.608%)CR:0 (0.000%)	1540/564032.730%	62193%	26 refs	16%	
S4	12618	G	A		transition	ND5	ND5:L94L	FL:760 (1.242%)CR:0 (0.000%)	1012/564151.794%	15691%	17 refs	100%	
S4	13440	C	T		transition	ND5	ND5:L368L	FL:21 (0.034%)CR:0 (0.000%)	51/564120.090%	440%	1 ref	24%	
S4	14179	A	G		transition	ND6	ND6:Y165Y	FL:324 (0.530%)CR:0 (0.000%)	179/564210.317%	4570%	5 refs	24%	Status: ReportedRecurrent pregnancy loss
S4	14766	C	T		transition	Cytb	Cytb:T7I	FL:46188 (75.510%)CR:0 (0.000%)	39641/5608670.679%	11470059%	117 refs	49%	

S4	14985	G	A		transition	Cytb	Cytb:R80H	FL:0 (0.000%)CR:0 (0.000%)	0	0	1 ref	100%	
S4	15326	A	G		transition	Cytb	Cytb:T194A	FL:60315 (98.605%)CR:0 (0.000%)	56027/5639899.342%	19397099%	120 refs	18%	
S4	16172	T	C		transition	ATTCR:Control RegionCR:HVS1/HV1CR:7S-likeCR:TAS	non-coding	FL:4471 (7.309%)CR:6807 (8.478%)	3589/562046.386%	87054%	148 refs	NA	
S4	16219	A	G		transition	ATTCR:Control RegionCR:HVS1/HV1CR:7S-like	non-coding	FL:603 (0.986%)CR:314 (0.391%)	345/564250.611%	12341%	39 refs	NA	
S4	16278	C	T		transition	ATTCR:Control RegionCR:HVS1/HV1CR:7S-like	non-coding	FL:6016 (9.835%)CR:7427 (9.250%)	10603/5398419.641%	113346%	129 refs	NA	
S5	57	TTT	AAC		transversion	CR:HVS2CR:OH57ATTCR:Control RegionCR:7S-like	non-coding	FL:0 (0.000%)CR:0 (0.000%)	NR	NR		NA	

S5	64	C	T		transition	CR:HVS2ATTCR:Control RegionCR:7S-like	non-coding	FL:1810 (2.959%)CR:2050 (2.553%)	2317/562014.123%	55003%	46 refs	NA	
S5	73	A	G		transition	CR:HVS2ATTCR:Control RegionCR:7S-like	non-coding	FL:45883 (75.011%)CR:42505 (52.937%)	40199/5616471.574%	11993461%	153 refs	NA	
S5	146	T	C		transition	CR:HVS2CR:OHATTCR:Control RegionCR:7S-like	non-coding	FL:11525 (18.842%)CR:10126 (12.611%)	11272/5534020.369%	2552513%	114 refs	NA	Status: ReportedAbsence of Endometriosis
S5	153	A	G		transition	CR:HVS2CR:OHATTCR:Control RegionCR:7S-like	non-coding	FL:2207 (3.608%)CR:2739 (3.411%)	3123/564115.536%	84884%	60 refs	NA	
S5	235	AT	GT		transversion	CR:HVS2CR:OHCR:CSB1CR:TFXATTCR:Control Region	non-coding	FL:0 (0.000%)CR:0 (0.000%)	NR	NR		NA	

S5	750	A	G		transition	12S	rRNA	FL:60122 (98.290%)CR:0 (0.000%)	55419/56 35498.34 1%	19298 698%	114 refs	98 %	
S5	16111	C	T		transition	ATTCR:Control RegionCR:HVS 1/HV1CR:TAS2 CR:7S-like	non-coding	FL:1680 (2.747%)CR:3367 (4.193%)	3047/563 705.405 %	70404 %	95 refs	NA	
S5	16223	C	T		transition	ATTCR:Control RegionCR:HVS 1/HV1CR:7S-like	non-coding	FL:23015 (37.626%)CR:4059 0 (50.552%)	22182/56 28639.40 9%	35639 18%	186 refs	NA	
S5	16290	CC	TC		transversion	ATTCR:Control RegionCR:HVS 1/HV1CR:7S-like	non-coding	FL:0 (0.000%)CR:0 (0.000%)	NR	NR		NA	
S5	16319	G	A		transition	ATTCR:Control RegionCR:HVS 1/HV1CR:7S-like	non-coding	FL:3549 (5.802%)CR:7073 (8.809%)	3513/564 106.228 %	91035 %	134 refs	NA	
S5	16362	T	C		transition	ATTCR:Control RegionCR:HVS	non-coding	FL:10255 (16.765%)CR:2172 1 (27.052%)	8423/562 3414.979 %	23527 12%	168 refs	NA	

						1/HV1CR:7S-like							
S6	64	C	T		transition	CR:HVS2ATTCR:Control RegionCR:7S-like	non-coding	FL:1810 (2.959%)CR:2050 (2.553%)	2317/562014.123%	55003%	46 refs	NA	
S6	73	A	G		transition	CR:HVS2ATTCR:Control RegionCR:7S-like	non-coding	FL:45883 (75.011%)CR:42505 (52.937%)	40199/5616471.574%	11993461%	153 refs	NA	
S6	146	T	C		transition	CR:HVS2CR:OHATTCR:Control RegionCR:7S-like	non-coding	FL:11525 (18.842%)CR:10126 (12.611%)	11272/5534020.369%	2552513%	114 refs	NA	Status: ReportedAbsence of Endometriosis
S6	153	A	G		transition	CR:HVS2CR:OHATTCR:Control RegionCR:7S-like	non-coding	FL:2207 (3.608%)CR:2739 (3.411%)	3123/564115.536%	84884%	60 refs	NA	
S6	235	AT	GT		transversion	CR:HVS2CR:OHCR:CSB1CR:T	non-coding	FL:0 (0.000%)CR:0 (0.000%)	NR	NR		NA	

						FXATTCR:Control Region							
S6	263	A	G		transition	CR:HVS2CR:OHATTCR:Control Region	non-coding	FL:58432 (95.527%)CR:45942 (57.217%)	55895/5642099.070%	19386199%	151 refs	NA	
S6	310	T	C		transition	CR:HVS2CR:OHCR:CSB2ATTCR:Control Region	non-coding	FL:24266 (39.671%)CR:23172 (28.859%)	375/70805.297%	NR	34 refs	NA	Status: ReportedPossible protective factor for normal tension glaucoma
S6	513	GCA	G		deletion	CR:HVS3CR:Control Region	non-coding	FL:0 (0.000%)CR:0 (0.000%)	NR	NR		NA	
S6	663	A	G		transition	12S	rRNA	FL:1757 (2.872%)CR:0 (0.000%)	2681/564204.752%	62143%	47 refs	60%	Status: ReportedCoronary atherosclerosis risk
S6	750	A	G		transition	12S	rRNA	FL:60122 (98.290%)CR:0 (0.000%)	55419/5635498.341%	19298698%	114 refs	98%	

S6	1438	A	G		transition	12S	rRNA	FL:58231 (95.198%)CR:0 (0.000%)	53745/5621395.610%	18988997%	124 refs	87%	
S6	1736	A	G		transition	16S	rRNA	FL:1738 (2.841%)CR:0 (0.000%)	2677/564224.745%	61963%	30 refs	31%	
S6	8027	G	A		transition	COII	COII:A148T	FL:1987 (3.248%)CR:0 (0.000%)	4171/563897.397%	69694%	16 refs	22%	
S6	11719	G	A		transition	ND4	ND4:G320G	FL:46499 (76.019%)CR:0 (0.000%)	39882/5609071.104%	11553459%	135 refs	98%	
S6	11944	T	C		transition	ND4	ND4:L395L	FL:1795 (2.935%)CR:0 (0.000%)	4743/563838.412%	28551%	30 refs	93%	
S6	12007	G	A		transition	ND4	ND4:W416W	FL:3683 (6.021%)CR:0 (0.000%)	4017/564027.122%	90415%	41 refs	9%	
S6	12705	C	T		transition	ND5	ND5:I123I	FL:24187 (39.542%)CR:0 (0.000%)	22373/4577248.879%	3532318%	115 refs	29%	Status: ReportedPossible protective factor for

													normal tension glaucoma
S6	14766	C	T		transition	Cytb	Cytb:T7I	FL:46188 (75.510%)CR:0 (0.000%)	39641/5608670.679%	11470059%	117 refs	49 %	
S6	15326	A	G		transition	Cytb	Cytb:T194A	FL:60315 (98.605%)CR:0 (0.000%)	56027/5639899.342%	19397099%	120 refs	18 %	
S6	16111	C	T		transition	ATTCR:Control RegionCR:HVS1/HV1CR:TAS2CR:7S-like	non-coding	FL:1680 (2.747%)CR:3367 (4.193%)	3047/563705.405 %	70404 %	95 refs	NA	
S6	16218	C	T		transition	ATTCR:Control RegionCR:HVS1/HV1CR:7S-like	non-coding	FL:373 (0.610%)CR:655 (0.816%)	366/545620.671%	11621 %	50 refs	NA	
S6	16223	C	T		transition	ATTCR:Control RegionCR:HVS1/HV1CR:7S-like	non-coding	FL:23015 (37.626%)CR:40590 (50.552%)	22182/5628639.409%	3563918%	186 refs	NA	

S6	16239	C	T		transition	ATTCR:Control RegionCR:HVS 1/HV1CR:7S-like	non-coding	FL:568 (0.929%)CR:710 (0.884%)	744/5640 81.319%	22021 %	59 refs	NA	
S6	16266	C	T		transition	ATTCR:Control RegionCR:HVS 1/HV1CR:7S-like	non-coding	FL:933 (1.525%)CR:1866 (2.324%)	403/5639 10.715%	13331 %	72 refs	NA	
S6	16290	CC	TC		transversion	ATTCR:Control RegionCR:HVS 1/HV1CR:7S-like	non-coding	FL:0 (0.000%)CR:0 (0.000%)	NR	NR		NA	
S6	16319	G	A		transition	ATTCR:Control RegionCR:HVS 1/HV1CR:7S-like	non-coding	FL:3549 (5.802%)CR:7073 (8.809%)	3513/564 106.228 %	91035 %	134 refs	NA	
S6	16362	T	C		transition	ATTCR:Control RegionCR:HVS 1/HV1CR:7S-like	non-coding	FL:10255 (16.765%)CR:2172 1 (27.052%)	8423/562 3414.979 %	23527 12%	168 refs	NA	

S7	73	A	G		transition	CR:HVS2ATTCR:Control RegionCR:7S-like	non-coding	FL:45883 (75.011%)CR:42505 (52.937%)	40199/5616471.574%	11993461%	153 refs	NA	
S7	143	G	A		transition	CR:HVS2CR:OHATTCR:Control RegionCR:7S-like	non-coding	FL:1181 (1.931%)CR:1507 (1.877%)	1726/564253.059%	23981%	44 refs	NA	
S7	146	T	C		transition	CR:HVS2CR:OHATTCR:Control RegionCR:7S-like	non-coding	FL:11525 (18.842%)CR:10126 (12.611%)	11272/5534020.369%	2552513%	114 refs	NA	Status: ReportedAbsence of Endometriosis
S7	152	TA	CA		transversion	CR:HVS2CR:OHATTCR:Control RegionCR:7S-like	non-coding	FL:0 (0.000%)CR:0 (0.000%)	NR	NR		NA	
S7	195	T	C		transition	CR:HVS2CR:OHATTCR:Control Region	non-coding	FL:11562 (18.902%)CR:9347 (11.641%)	15480/5607427.606%	3449618%	124 refs	NA	Status: ReportedBD-associated / melanoma pts

S7	263	A	G		transition	CR:HVS2CR:OHATTCR:Control Region	non-coding	FL:58432 (95.527%)CR:45942 (57.217%)	55895/5642099.070%	19386199%	151 refs	NA	
S7	309	C	CC	A302AC	insertion	CR:HVS2CR:OHCR:CSB2ATTCR:Control Region	non-coding	FL:616 (1.007%)CR:981 (1.222%)	9/155510.058%	NR	100 refs	NA	Status: ReportedAD-weakly associated
S7	310	T	C		transition	CR:HVS2CR:OHCR:CSB2ATTCR:Control Region	non-coding	FL:24266 (39.671%)CR:23172 (28.859%)	375/70805.297%	NR	34 refs	NA	Status: ReportedPossible protective factor for normal tension glaucoma
S7	750	A	G		transition	12S	rRNA	FL:60122 (98.290%)CR:0 (0.000%)	55419/5635498.341%	19298698%	114 refs	98%	
S7	769	G	A		transition	12S	rRNA	FL:4394 (7.183%)CR:0 (0.000%)	8568/5611415.269%	48402%	27 refs	16%	

S7	1018	G	A		transition	12S	rRNA	FL:4435 (7.251%)CR:0 (0.000%)	8503/5626115.113%	48122%	22 refs	9%	
S7	1438	A	G		transition	12S	rRNA	FL:58231 (95.198%)CR:0 (0.000%)	53745/5621395.610%	18988997%	124 refs	87%	
S7	2416	T	C		transition	16S	rRNA	FL:1792 (2.930%)CR:0 (0.000%)	4784/563818.485%	26961%	11 refs	24%	
S7	2483	T	C		transition	16S	rRNA	FL:109 (0.178%)CR:0 (0.000%)	252/564300.447%	3150%	9 refs	16%	
S7	2789	C	T		transition	16S	rRNA	FL:1137 (1.859%)CR:0 (0.000%)	2956/564175.240%	17421%	10 refs	38%	
S7	3594	C	T		transition	ND1	ND1:V96V	FL:4227 (6.910%)CR:0 (0.000%)	8364/5629514.857%	46332%	31 refs	20%	
S7	7771	A	G		transition	COII	COII:E62E	FL:1168 (1.909%)CR:0 (0.000%)	2957/564205.241%	17591%	9 refs	100%	

S7	8206	G	A		transition	COII	COII:M207M	FL:1555 (2.542%)CR:0 (0.000%)	4744/563 968.412 %	26581 %	13 refs	10 0%	
S7	9540	T	C		transition	COIII	COIII:L112L	FL:19059 (31.158%)CR:0 (0.000%)	17141/56 32330.43 3%	17229 9%	81 refs	98 %	
S7	10115	T	C		transition	ND3	ND3:I19I	FL:1502 (2.456%)CR:0 (0.000%)	4720/563 978.369 %	25371 %	10 refs	24 %	
S7	10398	A	G		transition	ND3	ND3:T114A	FL:25866 (42.287%)CR:0 (0.000%)	23506/56 17041.84 8%	49322 25%	224 ref s	51 %	Status: Repo rted; lineage L & M marker, also hg IJKPD protective factor / longevity / altered cell pH / metabolic syndrome / breast cancer

													risk / Leigh Syndrome risk / ADHD / cognitive decline / SCA2 age of onset / Fuchs endothelial corneal dystrophy
S7	10873	T	C		transition	ND4	ND4:P38P	FL:19056 (31.154%)CR:0 (0.000%)	17166/56 29130.49 5%	17312 9%	98 refs	20 %	
S7	11914	G	A		transition	ND4	ND4:T385T	FL:6409 (10.478%)CR:0 (0.000%)	6108/562 5610.857 %	10444 5%	64 refs	20 %	Status: Repo rtedPossible susceptibility to bullous pemphigoid
S7	11944	T	C		transition	ND4	ND4:L395L	FL:1795 (2.935%)CR:0 (0.000%)	4743/563 838.412 %	28551 %	30 refs	93 %	

S7	12693	A	G		transition	ND5	ND5:K119K	FL:1093 (1.787%)CR:0 (0.000%)	2941/564085.214 %	16631 %	9 refs	93 %	
S7	12705	C	T		transition	ND5	ND5:I123I	FL:24187 (39.542%)CR:0 (0.000%)	22373/4577248.879%	3532318%	115 refs	29 %	Status: ReportedPossible protective factor for normal tension glaucoma
S7	13590	G	A		transition	ND5	ND5:L418L	FL:3338 (5.457%)CR:0 (0.000%)	6120/5640210.851 %	63843 %	28 refs	69 %	Status: ReportedPossible protective factor for high altitude sickness
S7	13650	C	T		transition	ND5	ND5:P438P	FL:4243 (6.937%)CR:0 (0.000%)	8342/5616614.852 %	46432 %	18 refs	51 %	
S7	13803	A	G		transition	ND5	ND5:T489T	FL:1131 (1.849%)CR:0 (0.000%)	2954/564255.235 %	17201 %	9 refs	60 %	

S7	14566	A	G		transition	ND6	ND6:G36G	FL:1122 (1.834%)CR:0 (0.000%)	2865/564265.077%	16611%	8 refs	89%	
S7	14766	C	T		transition	Cytb	Cytb:T7I	FL:46188 (75.510%)CR:0 (0.000%)	39641/5608670.679%	11470059%	117 refs	49%	
S7	15301	G	A		transition	Cytb	Cytb:L185L	FL:16549 (27.055%)CR:0 (0.000%)	13722/5609624.462%	158838%	91 refs	96%	Status: ReportedPossible factor in high altitude sickness, LHON
S7	15326	A	G		transition	Cytb	Cytb:T194A	FL:60315 (98.605%)CR:0 (0.000%)	56027/5639899.342%	19397099%	120 refs	18%	
S7	15784	T	C		transition	Cytb	Cytb:P346P	FL:1998 (3.266%)CR:0 (0.000%)	3527/563966.254%	31022%	36 refs	100%	Status: ReportedPOAG - potential for association
S7	16189	T	C		transition	ATTCR:Control RegionCR:HVS	non-coding	FL:14962 (24.461%)CR:20753 (25.846%)	13910/5596224.856%	NR	226 refs	NA	Status: ReportedDiabetes/

						1/HV1CR:7S-like							Cardiomyopathy /cancer risk / mtDNA copy nbr / Metabolic Syndrome / Melanoma patients
S7	16223	C	T		transition	ATTCR:Control RegionCR:HVS 1/HV1CR:7S-like	non-coding	FL:23015 (37.626%)CR:40590 (50.552%)	22182/5628639.409%	3563918%	186 refs	NA	
S7	16278	C	T		transition	ATTCR:Control RegionCR:HVS 1/HV1CR:7S-like	non-coding	FL:6016 (9.835%)CR:7427 (9.250%)	10603/5398419.641%	113346%	129 refs	NA	
S7	16294	C	T		transition	ATTCR:Control RegionCR:HVS 1/HV1CR:7S-like	non-coding	FL:5510 (9.008%)CR:5712 (7.114%)	7997/5635614.190%	2081111%	135 refs	NA	

S7	16309	A	G		transition	ATTCR:Control RegionCR:HVS 1/HV1CR:7S-like	non-coding	FL:1571 (2.568%)CR:2433 (3.030%)	2541/564 064.505 %	18851 %	64 refs	NA	
S7	16390	G	A		transition	ATTCR:Control RegionCR:7S-like	non-coding	FL:3308 (5.408%)CR:4666 (5.811%)	5466/544 7810.033 %	61263 %	88 refs	NA	Status: Repo rtedPOAG - potential for association
S7	16519	T	C		transition	ATTCR:Control RegionCR:7S-like	non-coding	FL:38030 (62.173%)CR:3386 4 (42.175%)	35648/54 23865.72 5%	12599 264%	118 ref s	NA	

Supplementary File 4 – MAF annotation. Part of the information resulting from the VEP and MAFtools annotation.

Hugo_Symbol	Entrez_Gene_Id	Variant_Classification	Variant_Type	dbSNP_RS	Tumor_Sample_Barcode	HGVSc	HGVSp	HGVSp_Short	Transcript_ID	t_depth	t_ref_count	t_alt_count	Allele	Gene	Consequence	cDNA_position	CDS_position	Protein_position	Amino_acids	Codons	Existing_variation	SIFT	PolyPhen	CLIN_SIG	IMPACT	VARIANT_CLASS	vcf_pos
MT-ND1	4535	5'Flank	SNP	rs869183622	S7				ENST00000361390	40	0	40	G	ENSG00000198888	upstream_gene_variant						rs869183622				MODIFIER	SNV	73
MT-ND1	4535	5'Flank	SNP	rs375589100	S7				ENST00000361390	55	31	24	A	ENSG00000198888	upstream_gene_variant						rs375589100				MODIFIER	SNV	143
MT-ND1	4535	5'Flank	SNP	rs370482130	S7				ENST00000361390	55	9	46	C	ENSG00000198888	upstream_gene_variant						rs370482130				MODIFIER	SNV	146
MT-ND1	4535	5'Flank	DNP	rs117135796	S7				ENST00000361390	55	7	40	CA	ENSG00000198888	upstream_gene_variant						rs117135796				MODIFIER	substitution	152
MT-ND1	4535	5'Flank	SNP	rs2857291	S7				ENST00000361390	54	2	52	C	ENSG00000198888	upstream_gene_variant						rs2857291				MODIFIER	SNV	195
MT-ND1	4535	5'Flank	SNP	rs2853515	S7				ENST00000361390	42	0	42	G	ENSG00000198888	upstream_gene_variant						rs2853515			not_provided	MODIFIER	SNV	263
MT-ND1	4535	5'Flank	INS	rs878871521	S7				ENST00000361390	30	6	24	C	ENSG00000198888	upstream_gene_variant						rs878871521				MODIFIER	insertion	302
MT-ND1	4535	5'Flank	SNP	rs1556422421	S7				ENST00000361390	28	0	12	C	ENSG00000198888	upstream_gene_variant						rs1556422421				MODIFIER	indel	310
MT-ND1	4535	5'Flank	SNP	rs2853518	S7				ENST00000361390	43	0	43	G	ENSG00000198888	upstream_gene_variant						rs2853518			not_provided	MODIFIER	SNV	750
MT-ND1	4535	5'Flank	SNP	rs2853519	S7				ENST00000361390	48	5	43	A	ENSG00000198888	upstream_gene_variant						rs2853519				MODIFIER	SNV	769
MT-ND1	4535	5'Flank	SNP	rs2856982	S7				ENST00000361390	32	6	26	A	ENSG00000198888	upstream_gene_variant						rs2856982			benign	MODIFIER	SNV	1018
MT-ND1	4535	5'Flank	SNP	rs2001030	S7				ENST00000361390	52	0	52	G	ENSG00000198888	upstream_gene_variant						rs2001030			not_provided,benign	MODIFIER	SNV	1438
MT-ND1	4535	5'Flank	SNP	rs28358580	S7				ENST00000361390	39	8	31	C	ENSG00000198888	upstream_gene_variant						rs28358580				MODIFIER	SNV	2416
MT-ND1	4535	5'Flank	SNP	rs28445203	S7				ENST00000361390	34	20	14	C	ENSG00000198888	upstream_gene_variant						rs28445203			uncertain_significance	MODIFIER	SNV	2483
MT-ND1	4535	5'Flank	SNP	rs28358581	S7				ENST00000361390	42	8	34	T	ENSG00000198888	upstream_gene_variant						rs28358581				MODIFIER	SNV	2789
MT-ND1	4535	Silent	SNP	rs193303025	S7	c.288C>T	p.Val96=	p.V96=	ENST00000361390	47	8	39	T	ENSG00000198888	synonymous_variant	288/956	288/956	96/318	V	gtC/gtT	rs193303025,COSV104421141				LOW	SNV	3594
MT-CO2	4513	Silent	SNP	rs368038563	S7	c.186A>G	p.Glu62=	p.E62=	ENST00000361739	33	8	25	G	ENSG00000198712	synonymous_variant	186/684	186/684	62/227	E	gaA/gaG	rs368038563			benign	LOW	SNV	7771
MT-CO2	4513	Silent	SNP	rs28358883	S7	c.621G>A	p.Met207=	p.M207=	ENST00000361739	42	9	33	A	ENSG00000198712	synonymous_variant	621/684	621/684	207/227	M	atG/atA	rs28358883			benign	LOW	SNV	8206
MT-CO3	4514	Silent	SNP	rs2248727	S7	c.3347>C	p.Leu112=	p.L112=	ENST00000362079	46	1	45	C	ENSG00000198938	synonymous_variant	334/784	334/784	112/261	L	Tta/Cta	rs2248727,COSV104419768				LOW	SNV	9540
MT-ND3	4537	Silent	SNP	rs3899188	S7	c.57T>C	p.Ile19=	p.I19=	ENST00000361227	46	9	37	C	ENSG00000198840	synonymous_variant	57/346	57/346	19/115	I	attT/atC	rs3899188				LOW	SNV	10115
MT-ND3	4537	Missense_Mutation	SNP	rs2853826	S7	c.340A>G	p.Thr114Ala	p.T114A	ENST00000361227	33	0	33	G	ENSG00000198840	missense_variant	340/346	340/346	114/115	T/A	Acc/Gcc	rs2853826	tolerated_low_confidence(0.3)	benign(0)	protective,benign	MODERATE	SNV	10398
MT-ND4	4538	Silent	SNP	rs2857284	S7	c.1147>C	p.Pro38=	p.P38=	ENST00000361381	46	2	44	C	ENSG00000198886	synonymous_variant	114/1378	114/1378	38/459	P	ccT/ccC	rs2857284,COSV104419772				LOW	SNV	10873
MT-ND4	4538	Silent	SNP	rs2853496	S7	c.1155G>A	p.Thr385=	p.T385=	ENST00000361381	33	0	33	A	ENSG00000198886	synonymous_variant	1155/1378	1155/1378	385/459	T	acG/acA	rs2853496,COSV62294301				LOW	SNV	11914
MT-ND4	4538	Silent	SNP	rs3087901	S7	c.1185T>C	p.Leu395=	p.L395=	ENST00000361381	35	2	33	C	ENSG00000198886	synonymous_variant	1185/1378	1185/1378	395/459	L	ctT/ctC	rs3087901				LOW	SNV	11944
MT-ND5	4540	Silent	SNP	rs28359173	S7	c.357A>G	p.Lys119=	p.K119=	ENST00000361567	56	2	54	G	ENSG00000198786	synonymous_variant	357/1812	357/1812	119/603	K	aaA/aaG	rs28359173				LOW	SNV	12693
MT-ND5	4540	Silent	SNP	rs193302956	S7	c.369C>T	p.Ile123=	p.I123=	ENST00000361567	55	0	55	T	ENSG00000198786	synonymous_variant	369/1812	369/1812	123/603	I	atC/atT	rs193302956,COSV62293963				LOW	SNV	12705
MT-ND5	4540	Silent	SNP	rs28359177	S7	c.1254G>A	p.Leu418=	p.L418=	ENST00000361567	50	7	43	A	ENSG00000198786	synonymous_variant	1254/1812	1254/1812	418/603	L	ctG/ctA	rs28359177				LOW	SNV	13590
MT-ND5	4540	Silent	SNP	rs2854123	S7	c.1314C>T	p.Pro438=	p.P438=	ENST00000361567	43	3	40	T	ENSG00000198786	synonymous_variant	1314/1812	1314/1812	438/603	P	ccC/ccT	rs2854123,COSV62378070				LOW	SNV	13650
MT-ND5	4540	Silent	SNP	rs28359180	S7	c.1467A>G	p.Thr489=	p.T489=	ENST00000361567	27	4	23	G	ENSG00000198786	synonymous_variant	1467/1812	1467/1812	489/603	T	acA/acG	rs28359180				LOW	SNV	13803
MT-ND6	4541	Silent	SNP	rs28357677	S7	c.108T>C	p.Gly36=	p.G36=	ENST00000361681	42	6	36	G	ENSG00000198695	synonymous_variant	108/525	108/525	36/174	G	ggT/ggC	rs28357677				LOW	SNV	14566
MT-CYB	4519	Missense_Mutation	SNP	rs193302980	S7	c.20C>T	p.Thr7Ile	p.T7I	ENST00000361789	53	4	49	T	ENSG00000198727	missense_variant	20/1141	20/1141	7/380	T/I	aCt/aTt	rs193302980	tolerated_low_confidence(0.05)	benign(0.003)	benign,likely_pathogenic	MODERATE	SNV	14766
MT-CYB	4519	Silent	SNP	rs193302991	S7	c.555G>A	p.Leu185=	p.L185=	ENST00000361789	42	2	40	A	ENSG00000198727	synonymous_variant	555/1141	555/1141	185/380	L	ttG/ttA	rs193302991,COSV104419804			likely_pathogenic	LOW	SNV	15301
MT-CYB	4519	Missense_Mutation	SNP	rs2853508	S7	c.580A>G	p.Thr194Ala	p.T194A	ENST00000361789	40	0	40	G	ENSG00000198727	missense_variant	580/1141	580/1141	194/380	T/A	Aca/Gca	rs2853508	tolerated_low_confidence(0.21)	benign(0.009)	benign,likely_pathogenic	MODERATE	SNV	15326
MT-CYB	4519	Silent	SNP	rs527236194	S7	c.1038T>C	p.Pro346=	p.P346=	ENST00000361789	46	8	38	C	ENSG00000198727	synonymous_variant	1038/1141	1038/1141	346/380	P	ccT/ccC	rs527236194			likely_pathogenic	LOW	SNV	15784
MT-ND4	4538	3'Flank	SNP	rs28693675	S7				ENST00000361381	60	12	48	C	ENSG00000198886	downstream_gene_variant						rs28693675				MODIFIER	SNV	16189
MT-ND4	4538	3'Flank	SNP	rs2853513	S7				ENST00000361381	63	1	62	T	ENSG00000198886	downstream_gene_variant						rs2853513				MODIFIER	SNV	16223
MT-ND4	4538	3'Flank	SNP	rs41458645	S7				ENST00000361381	62	11	51	T	ENSG00000198886	downstream_gene_variant						rs41458645				MODIFIER	SNV	16278
MT-ND4	4538	3'Flank	SNP	rs140662392	S7				ENST00000361381	60	10	50	T	ENSG00000198886	downstream_gene_variant						rs140662392				MODIFIER	SNV	16294
MT-ND4	4538	3'Flank	SNP	rs373517769	S7				ENST00000361381	59	15	44	G	ENSG00000198886	downstream_gene_variant						rs373517769				MODIFIER	SNV	16309
MT-ND4	4538	3'Flank	SNP	rs41378955	S7				ENST00000361381	62	12	50	A	ENSG00000198886	downstream_gene_variant						rs41378955				MODIFIER	SNV	16390
MT-ND4	4538	3'Flank	SNP	rs3937033	S7				ENST00000361381	59	21	38	C	ENSG00000198886	downstream_gene_variant						rs3937033				MODIFIER	SNV	16519
MT-ND1	4535	5'Flank	TNP	novel	S5				ENST00000361390	32	0	32	AAC	ENSG00000198888	upstream_gene_variant										MODIFIER	substitution	57
MT-ND1	4535	5'Flank	SNP	rs3883917	S5				ENST00000361390	38	0	38	T	ENSG00000198888	upstream_gene_variant						rs3883917				MODIFIER	SNV	64
MT-ND1	4535	5'Flank	SNP	rs869183622	S5				ENST00000361390	40	0	40	G	ENSG00000198888	upstream_gene_variant						rs869183622				MODIFIER	SNV	73
MT-ND1	4535	5'Flank	SNP	rs370482130	S5				ENST00000361390	47	0	47	C	ENSG00000198888	upstream_gene_variant						rs370482130				MODIFIER	SNV	146
MT-ND1	4535	5'Flank	SNP	rs370716192	S5				ENST00000361390	44	0	44	G	ENSG00000198888	upstream_gene_variant						rs370716192				MODIFIER	SNV	153
MT-ND1	4535	5'Flank	DNP	novel	S5				ENST00000361390	31	.	.	GC	ENSG00000198888	upstream_gene_variant										MODIFIER	substitution	235
MT-ND1	4535	5'Flank	SNP	rs2853518	S5				ENST00000361390	38	1	37	G	ENSG00000198888	upstream_gene_variant						rs2853518			not_provided	MODIFIER	SNV	750
MT-ND4	4538	3'Flank	SNP	rs35315169	S5				ENST00000361381	45	0	45	T	ENSG00000198886	downstream_gene_variant						rs35315169				MODIFIER	SNV	16111
MT-ND4	4538	3'Flank	SNP	rs2853513	S5				ENST00000361381	43	0	43	T	ENSG00000198886	downstream_gene_variant						rs2853513				MODIFIER	SNV	16223
MT-ND4	4538	3'Flank	DNP	novel	S5				ENST00000361381	48	.	.	TT	ENSG00000198886	downstream_gene_variant										MODIFIER	substitution	16290
MT-ND4	4538	3'Flank	SNP	rs35105996	S5				ENST00000361381	48	0	48	A	ENSG00000198886	downstream_gene_variant						rs35105996			uncertain_significance	MODIFIER	SNV	16319
MT-ND4	4538	3'Flank	SNP	rs62581341	S5				ENST00000361381	60	0	60	C	ENSG00000198886	downstream_gene_variant						rs62581341				MODIFIER	SNV	16362
MT-ND1	4535	5'Flank	SNP	rs3883917	S6				ENST00000361390	48	0	48	T	ENSG00000198888	upstream_gene_variant						rs3883917				MODIFIER	SNV	64
MT-ND1	4535	5'Flank	SNP	rs869183622	S6				ENST00000361390	51	0	51	G	ENSG00000198888	upstream_gene_variant						rs869183622				MODIFIER	SNV	73
MT-ND1	4535	5'Flank	SNP	rs370482130	S6				ENST00000361390	61	0	61	C	ENSG00000198888	upstream_gene_variant						rs370482130				MODIFIER	SNV	146
MT-ND1	4535	5'Flank	SNP	rs370716192	S6				ENST00000361390	57	1	56	G	ENSG00000198888	upstream_gene_variant						rs370716192				MODIFIER	SNV	153
MT-ND1	4535	5'Flank	DNP	rs3937037	S6				ENST00000361390	52	.	.	GT	ENSG00000198888	upstream_gene_variant						rs3937037				MODIFIER	substitution	235
MT-ND1	4535	5'Flank	SNP	rs2853515	S6				ENST00000361390	42	0	42	G	ENSG00000198888	upstream_gene_variant						rs2853515			not_provided	MODIFIER	SNV	263
MT-ND1	4535	5'Flank	SNP	rs1556422421	S6				ENST00000361390	31	0	13	C	ENSG00000198888	upstream_gene_variant						rs1556422421				MODIFIER	indel	310
MT-ND1	4535	5'Flank	DEL	rs78907894	S6				ENST00000361390	29	2	27	-	ENSG00000198888	upstream_gene_variant						rs78907894			uncertain_significance	MODIFIER	deletion	513
MT-ND1	4535	5'Flank	SNP	rs56489998	S6				ENST00000361390	51	0	51	G	ENSG00000198888	upstream_gene_variant						rs56489998			benign	MODIFIER	SNV	663
MT-ND1	4535	5'Flank	SNP	rs2853518	S6				ENST00000361390	48	0	48	G	ENSG00000198888	upstream_gene_variant						rs2853518			not_provided	MODIFIER	SNV	750
MT-ND1	4535	5'Flank	SNP	rs2001030	S6				ENST00000361390	42	0	42	G	ENSG00000198888	upstream_gene_variant						rs2001030			not_provided,benign	MODIFIER	SNV	1438
MT-ND1	4535	5'Flank	SNP	rs193303006	S6				ENST00000361390	33	0	33	G	ENSG00000198888	upstream_gene_variant						rs193303006			uncertain_significance	MODIFIER	SNV	1736
MT-CO2	4513	Missense_Mutation	SNP	rs1116904	S6	c.442G>A	p.Ala148Thr	p.A148T	ENST00000361739	54	0	54	A	ENSG00000198712	missense_variant	442/684	442/684	148/227	A/T	Gcc/Acc	rs1116904	tolerated_low_confidence(0.12)	probably_damaging(0.998)	benign	MODERATE	SNV	8027
MT-ND4	4538	Silent	SNP	rs2853495	S6	c.960G>A	p.Gly320=	p.G320=	ENST00000361381	38	0	38	A	ENSG00000198886	synonymous_variant	960/1378	960/1378	320/459	G	ggG/ggA	rs2853495,COSV104668829				LOW	SNV	11719
MT-ND4	4538	Silent	SNP	rs3087901	S6	c.1185T>C	p.Leu395=	p.L395=	ENST00000361381	40	0	40	C	ENSG00000198886	synonymous_variant	1185/1378	1185/1378	395/459	L	ctT/ctC	rs3087901				LOW	SNV	11944
MT-ND4	4538	Silent	SNP	rs2853497	S6	c.1248G>A	p.Trp416=	p.W416=	ENST00000361381	48	0	48	A	ENSG00000198886	synonymous_variant	1248/1378	1248/1378	416/459	W	tgG/tgA	rs2853497				LOW	SNV	12007
MT-ND5	4540	Silent	SNP	rs193302956	S6	c.369C>T	p.Ile123=	p.I123=	ENST00000361567	54	0	54	T	ENSG00000198786	synonymous_variant	369/1812	369/1812	123/603	I	atC/atT	rs193302956,COSV62293963				LOW	SNV	12705
MT-CYB	4519	Missense_Mutation	SNP	rs193302980	S6	c.20C>T	p.Thr7Ile	p.T7I	ENST00000361789	55	0	55	T	ENSG00000198727	missense_variant	20/1141	20/1141	7/380	T/I	aCt/aTt	rs193302980	tolerated_low_confidence(0.05)	benign(0.003)	benign,likely_pathogenic	MODERATE	SNV	14766
MT-CYB	4519	Missense_Mutation	SNP	rs2853508	S6	c.580A>G	p.Thr194Ala	p.T194A	ENST00000361789	54	0	54	G	ENSG00000198727	missense_variant	580/1141	580/1141	194/380	T/A	Aca/Gca	rs2853508	tolerated_low_confidence(0.21)	benign(0.009)	benign,likely_pathogenic	MODERATE	SNV	15326
MT-ND4	4538	3'Flank	SNP	rs35315169	S6				ENST00000361381	68	0	68	T	ENSG00000198886	downstream_gene_variant						rs35315169				MODIFIER	SNV	16111
MT-ND4	4538	3'Flank	SNP	rs386829280	S6				ENST00000361381	78	0	78	T	ENSG00000198886	downstream_gene_variant						rs386829280				MODIFIER	SNV	16218
MT-ND4	4538	3'Flank	SNP	rs2853513	S6				ENST00000361381	77	0	77	T	ENSG00000198886	downstream_gene_variant						rs2853513				MODIFIER	SNV	16223
MT-ND4	4538	3'Flank	SNP	novel	S6				ENST00000361381	77	.	.	A	ENSG00000198886	downstream_gene_variant										MODIFIER	SNV	16239
MT-ND4	4538	3'Flank	SNP	rs879098011	S6				ENST00000361381	80	0	80	T	ENSG00000198886	downstream_gene_variant						rs879098011				MODIFIER	SNV	16266
MT-ND4	4538	3'Flank	DNP	rs1556424849	S6				ENST00000361381	74	.	.	TC	ENSG00000198886	downstream_gene_variant						rs1556424849				MODIFIER	substitution	16290
MT-ND4	4538	3'Flank	SNP	rs35105996	S6				ENST00000361381	74	0	74	A	ENSG00000198886	downstream_gene_variant						rs35105996			uncertain_significance	MODIFIER	SNV	16319
MT-ND4	4538	3'Flank	SNP	rs62581341	S6				ENST00000361381	79	0	79	C	ENSG00000198886	downstream_gene_variant						rs62581341				MODIFIER	SNV	16362

MT-ND1	4535	5'Flank	SNP	rs869183622	S1				ENST00000361390	57	0	57	G	ENSG00000198888	upstream_gene_variant						rs869183622				MODIFIER	SNV	73
MT-ND1	4535	5'Flank	DNP	rs117135796	S1				ENST00000361390	82	.	.	CA	ENSG00000198888	upstream_gene_variant						rs117135796				MODIFIER	substitution	152
MT-ND1	4535	5'Flank	SNP	rs2853515	S1				ENST00000361390	66	0	66	G	ENSG00000198888	upstream_gene_variant						rs2853515			not_provided	MODIFIER	SNV	263
MT-ND1	4535	5'Flank	SNP	rs1556422421	S1				ENST00000361390	44	0	20	C	ENSG00000198888	upstream_gene_variant						rs1556422421				MODIFIER	indel	310
MT-ND1	4535	5'Flank	DEL	rs78907894	S1				ENST00000361390	35	4	31	-	ENSG00000198888	upstream_gene_variant						rs78907894			uncertain_significance	MODIFIER	deletion	513
MT-ND1	4535	5'Flank	SNP	rs2853518	S1				ENST00000361390	59	0	59	G	ENSG00000198888	upstream_gene_variant						rs2853518			not_provided	MODIFIER	SNV	750
MT-ND1	4535	5'Flank	SNP	rs28358570	S1				ENST00000361390	57	0	57	C	ENSG00000198888	upstream_gene_variant						rs28358570				MODIFIER	SNV	921
MT-ND1	4535	5'Flank	SNP	rs2001030	S1				ENST00000361390	63	0	63	G	ENSG00000198888	upstream_gene_variant						rs2001030			not_provided,benign	MODIFIER	SNV	1438
MT-ND1	4535	5'Flank	SNP	rs2854128	S1				ENST00000361390	51	0	51	G	ENSG00000198888	upstream_gene_variant						rs2854128				MODIFIER	SNV	2706
MT-ND1	4535	Missense_Mutation	SNP	rs28357970	S1	c.490A>G	p.Thr164Ala	p.T164A	ENST00000361390	57	0	57	G	ENSG00000198888	missense_variant	490/956	490/956	164/318	T/A	Acc/Gcc	rs28357970	deleterious_low_confidence(0.05)	benign(0.013)	benign,pathogenic	MODERATE	SNV	3796
MT-CO1	4512	Silent	SNP	rs41352249	S1	c.777T>C	p.Thr259=	p.T259=	ENST00000361624	49	2	47	C	ENSG00000198804	synonymous_variant	777/1542	777/1542	259/513	T	acT/acC	rs41352249				LOW	SNV	6680
MT-CO2	4513	Silent	SNP	rs1603221059	S1	c.63C>T	p.Ile21=	p.I21=	ENST00000361739	35	1	34	T	ENSG00000198712	synonymous_variant	63/684	63/684	21/227	I	atC/atT	rs1603221059				LOW	SNV	7648
MT-ATP6	4508	Missense_Mutation	SNP	rs2000975	S1	c.175A>G	p.Thr59Ala	p.T59A	ENST00000361899	30	1	29	G	ENSG00000198899	missense_variant	175/681	175/681	59/226	T/A	Acc/Gcc	rs2000975,COSV104419767	tolerated_low_confidence(0.64)	benign(0.005)	benign	MODERATE	SNV	8701
MT-CO3	4514	Silent	SNP	rs2248727	S1	c.334T>C	p.Leu112=	p.L112=	ENST00000362079	57	0	57	C	ENSG00000198938	synonymous_variant	334/784	334/784	112/261	L	Tta/Cta	rs2248727,COSV104419768				LOW	SNV	9540
MT-ND3	4537	Missense_Mutation	SNP	rs2853826	S1	c.340A>G	p.Thr114Ala	p.T114A	ENST00000361227	47	0	47	G	ENSG00000198840	missense_variant	340/346	340/346	114/115	T/A	Acc/Gcc	rs2853826	tolerated_low_confidence(0.3)	benign(0)	protective,benign	MODERATE	SNV	10398
MT-ND3	4537	3'Flank	SNP	rs28358279	S1				ENST00000361227	56	0	56	C	ENSG00000198840	downstream_gene_variant						rs28358279			benign	MODIFIER	SNV	10463
MT-ND4	4538	Silent	SNP	rs2857284	S1	c.114T>C	p.Pro38=	p.P38=	ENST00000361381	55	1	54	C	ENSG00000198886	synonymous_variant	114/1378	114/1378	38/459	P	ccT/ccC	rs2857284,COSV104419772				LOW	SNV	10873
MT-ND4	4538	Silent	SNP	rs2853495	S1	c.960G>A	p.Gly320=	p.G320=	ENST00000361381	47	0	47	A	ENSG00000198886	synonymous_variant	960/1378	960/1378	320/459	G	ggG/ggA	rs2853495,COSV104668829				LOW	SNV	11719
MT-ND4	4538	Silent	SNP	rs1556424007	S1	c.1128G>A	p.Leu376=	p.L376=	ENST00000361381	48	0	48	A	ENSG00000198886	synonymous_variant	1128/1378	1128/1378	376/459	L	ctG/ctA	rs1556424007				LOW	SNV	11887
MT-ND3	4537	3'Flank	DEL	rs1603223633	S1				ENST00000361227	51	1	50	-	ENSG00000198840	downstream_gene_variant						rs1603223633				MODIFIER	deletion	12236
MT-ND5	4540	Silent	SNP	rs193302956	S1	c.369C>T	p.Ile123=	p.I123=	ENST00000361567	73	0	73	T	ENSG00000198786	synonymous_variant	369/1812	369/1812	123/603	I	atC/atT	rs193302956,COSV62293963				LOW	SNV	12705
MT-ND5	4540	Missense_Mutation	SNP	rs2853501	S1	c.769A>G	p.Ile257Val	p.I257V	ENST00000361567	46	0	46	G	ENSG00000198786	missense_variant	769/1812	769/1812	257/603	I/V	Atc/Gtc	rs2853501,COSV62294516	tolerated_low_confidence(0.77)	benign(0.007)	benign	MODERATE	SNV	13105
MT-ND5	4540	Missense_Mutation	SNP	rs28359182	S1	c.1550T>C	p.Leu517Pro	p.L517P	ENST00000361567	57	0	57	C	ENSG00000198786	missense_variant	1550/1812	1550/1812	517/603	L/P	cTa/cCa	rs28359182	deleterious_low_confidence(0.04)	benign(0.003)	benign	MODERATE	SNV	13886
MT-ND6	4541	Silent	SNP	rs28357673	S1	c.390G>A	p.Glu130=	p.E130=	ENST00000361681	71	1	70	T	ENSG00000198695	synonymous_variant	390/525	390/525	130/174	E	gaG/gaA	rs28357673				LOW	SNV	14284
MT-ND3	4537	3'Flank	SNP	rs386829226	S1				ENST00000361227	65	0	65	G	ENSG00000198840	downstream_gene_variant						rs386829226			benign	MODIFIER	SNV	14693
MT-CYB	4519	Missense_Mutation	SNP	rs193302980	S1	c.20C>T	p.Thr7Ile	p.T7I	ENST00000361789	64	0	64	T	ENSG00000198727	missense_variant	20/1141	20/1141	7/380	T/I	aCt/aTt	rs193302980	tolerated_low_confidence(0.05)	benign(0.003)	benign,likely_pathogenic	MODERATE	SNV	14766
MT-CYB	4519	Silent	SNP	rs527236205	S1	c.315A>G	p.Gly105=	p.G105=	ENST00000361789	59	1	58	G	ENSG00000198727	synonymous_variant	315/1141	315/1141	105/380	G	ggA/ggG	rs527236205,COSV62378179			likely_pathogenic	LOW	SNV	15061
MT-CYB	4519	Silent	SNP	rs193302991	S1	c.555G>A	p.Leu185=	p.L185=	ENST00000361789	69	0	69	A	ENSG00000198727	synonymous_variant	555/1141	555/1141	185/380	L	ttG/ttA	rs193302991,COSV104419804			likely_pathogenic	LOW	SNV	15301
MT-CYB	4519	Missense_Mutation	SNP	rs2853508	S1	c.580A>G	p.Thr194Ala	p.T194A	ENST00000361789	74	0	74	G	ENSG00000198727	missense_variant	580/1141	580/1141	194/380	T/A	Aca/Gca	rs2853508	tolerated_low_confidence(0.21)	benign(0.009)	benign,likely_pathogenic	MODERATE	SNV	15326
MT-ND4	4538	3'Flank	SNP	rs386829272	S1				ENST00000361381	93	0	93	C	ENSG00000198886	downstream_gene_variant						rs386829272				MODIFIER	SNV	16124
MT-ND4	4538	3'Flank	SNP	rs2853513	S1				ENST00000361381	83	1	82	T	ENSG00000198886	downstream_gene_variant						rs2853513				MODIFIER	SNV	16223
MT-ND4	4538	3'Flank	SNP	rs879082592	S1				ENST00000361381	92	1	91	G	ENSG00000198886	downstream_gene_variant						rs879082592				MODIFIER	SNV	16300
MT-ND4	4538	3'Flank	SNP	rs35105996	S1				ENST00000361381	88	1	87	A	ENSG00000198886	downstream_gene_variant						rs35105996			uncertain_significance	MODIFIER	SNV	16319

MT-ND1	4535	5'Flank	SNP	rs869183622	S2				ENST00000361390	62	0	62	G	ENSG00000198888	upstream_gene_variant						rs869183622				MODIFIER	SNV	73
MT-ND1	4535	5'Flank	SNP	rs368534078	S2				ENST00000361390	79	0	79	G	ENSG00000198888	upstream_gene_variant						rs368534078				MODIFIER	SNV	210
MT-ND1	4535	5'Flank	SNP	rs2853515	S2				ENST00000361390	69	0	69	G	ENSG00000198888	upstream_gene_variant						rs2853515			not_provided	MODIFIER	SNV	263
MT-ND1	4535	5'Flank	SNP	rs1556422421	S2				ENST00000361390	46	1	18	C	ENSG00000198888	upstream_gene_variant						rs1556422421				MODIFIER	indel	310
MT-ND1	4535	5'Flank	SNP	rs2853518	S2				ENST00000361390	70	0	70	G	ENSG00000198888	upstream_gene_variant						rs2853518			not_provided	MODIFIER	SNV	750
MT-ND1	4535	5'Flank	SNP	rs2001030	S2				ENST00000361390	81	0	81	G	ENSG00000198888	upstream_gene_variant						rs2001030			not_provided,benign	MODIFIER	SNV	1438
MT-ND1	4535	5'Flank	SNP	rs28588170	S2				ENST00000361390	61	0	61	T	ENSG00000198888	upstream_gene_variant						rs28588170				MODIFIER	SNV	2092
MT-ND1	4535	5'Flank	SNP	novel	S2				ENST00000361390	82	11	71	A	ENSG00000198888	upstream_gene_variant										MODIFIER	SNV	2652
MT-ND1	4535	5'Flank	SNP	rs2854128	S2				ENST00000361390	68	0	68	G	ENSG00000198888	upstream_gene_variant						rs2854128				MODIFIER	SNV	2706
MT-ND1	4535	5'Flank	SNP	rs3928306	S2				ENST00000361390	77	0	77	A	ENSG00000198888	upstream_gene_variant						rs3928306			not_provided	MODIFIER	SNV	3010
MT-ND2	4536	Silent	SNP	rs200763872	S2	c.414C>T	p.Pro138=	p.P138=	ENST00000361453	62	0	62	T	ENSG00000198763	synonymous_variant	414/1042	414/1042	138/347	P	ccC/ccT	rs200763872				LOW	SNV	4883
MT-CO1	4512	Silent	SNP	rs2015062	S2	c.1125C>T	p.Ala375=	p.A375=	ENST00000361624	29	0	29	T	ENSG00000198804	synonymous_variant	1125/1542	1125/1542	375/513	A	gcC/gcT	rs2015062			benign	LOW	SNV	7028
MT-ATP8	4509	Missense_Mutation	SNP	rs28358884	S2	c.49C>T	p.Leu17Phe	p.L17F	ENST00000361851	49	0	49	T	ENSG00000228253	missense_variant	49/207	49/207	17/68	L/F	Ctc/Ttc	rs28358884	tolerated(0.23)	probably_damaging(0.992)	benign	MODERATE	SNV	8414
MT-ATP6	4508	Missense_Mutation	SNP	rs2000975	S2	c.175A>G	p.Thr59Ala	p.T59A	ENST00000361899	37	0	37	G	ENSG00000198899	missense_variant	175/681	175/681	59/226	T/A	Acc/Gcc	rs2000975,COSV104419767	tolerated_low_confidence(0.64)	benign(0.005)	benign	MODERATE	SNV	8701
MT-CO3	4514	Silent	SNP	rs2248727	S2	c.334T>C	p.Leu112=	p.L112=	ENST00000362079	76	0	76	C	ENSG00000198938	synonymous_variant	334/784	334/784	112/261	L	Tta/Cta	rs2248727,COSV104419768				LOW	SNV	9540
MT-ND3	4537	Missense_Mutation	SNP	rs2853826	S2	c.340A>G	p.Thr114Ala	p.T114A	ENST00000361227	67	0	67	G	ENSG00000198840	missense_variant	340/346	340/346	114/115	T/A	Acc/Gcc	rs2853826	tolerated_low_confidence(0.3)	benign(0)	protective,benign	MODERATE	SNV	10398
MT-ND3	4537	Silent	SNP	rs28358278	S2	c.342C>T	p.Thr114=	p.T114=	ENST00000361227	67	0	67	T	ENSG00000198840	synonymous_variant	342/346	342/346	114/115	T	acC/acT	rs28358278				LOW	SNV	10400
MT-ND4	4538	Silent	SNP	rs2857284	S2	c.114T>C	p.Pro38=	p.P38=	ENST00000361381	81	0	81	C	ENSG00000198886	synonymous_variant	114/1378	114/1378	38/459	P	ccT/ccC	rs2857284,COSV104419772				LOW	SNV	10873
MT-ND4	4538	Missense_Mutation	SNP	rs386829118	S2	c.391G>A	p.Ala131Thr	p.A131T	ENST00000361381	81	0	81	A	ENSG00000198886	missense_variant	391/1378	391/1378	131/459	A/T	Gct/Act	rs386829118,COSV104668810	deleterious_low_confidence(0)	benign(0.03)	benign,not_provided	MODERATE	SNV	11150
MT-ND4	4538	Silent	SNP	rs1556423912	S2	c.474T>C	p.Leu158=	p.L158=	ENST00000361381	84	0	84	C	ENSG00000198886	synonymous_variant	474/1378	474/1378	158/459	L	ctT/ctC	rs1556423912				LOW	SNV	11233
MT-ND4	4538	Silent	SNP	rs2853495	S2	c.960G>A	p.Gly320=	p.G320=	ENST00000361381	55	0	55	A	ENSG00000198886	synonymous_variant	960/1378	960/1378	320/459	G	ggG/ggA	rs2853495,COSV104668829				LOW	SNV	11719
MT-ND5	4540	Silent	SNP	rs193302956	S2	c.369C>T	p.Ile123=	p.I123=	ENST00000361567	80	0	80	T	ENSG00000198786	synonymous_variant	369/1812	369/1812	123/603	I	atC/atT	rs193302956,COSV62293963				LOW	SNV	12705
MT-ND5	4540	Silent	SNP	novel	S2	c.483A>T	p.Arg161=	p.R161=	ENST00000361567	78	0	78	T	ENSG00000198786	synonymous_variant	483/1812	483/1812	161/603	R	cgA/cgT					LOW	SNV	12819
MT-ND5	4540	Missense_Mutation	SNP	novel	S2	c.1669T>C	p.Trp557Arg	p.W557R	ENST00000361567	79	20	59	C	ENSG00000198786	missense_variant	1669/1812	1669/1812	557/603	W/R	Tga/Cga		deleterious_low_confidence(0)	probably_damaging(0.998)		MODERATE	SNV	14005
MT-ND6	4541	Silent	SNP	rs28357678	S2	c.6G>A	p.Met2=	p.M2=	ENST00000361681	81	0	81	T	ENSG00000198695	synonymous_variant	6/525	6/525	2/174	M	atG/atA	rs28357678			uncertain_significance	LOW	SNV	14668
MT-CYB	4519	Missense_Mutation	SNP	rs193302980	S2	c.20C>T	p.Thr7Ile	p.T7I	ENST00000361789	81	0	81	T	ENSG00000198727	missense_variant	20/1141	20/1141	7/380	T/I	aCt/aTt	rs193302980	tolerated_low_confidence(0.05)	benign(0.003)	benign,likely_pathogenic	MODERATE	SNV	14766
MT-CYB	4519	Silent	SNP	rs1603225085	S2	c.360G>A	p.Leu120=	p.L120=	ENST00000361789	91	0	91	A	ENSG00000198727	synonymous_variant	360/1141	360/1141	120/380	L	ctG/ctA	rs1603225085				LOW	SNV	15106
MT-CYB	4519	Silent	SNP	rs193302991	S2	c.555G>A	p.Leu185=	p.L185=	ENST00000361789	100	0	100	A	ENSG00000198727	synonymous_variant	555/1141	555/1141	185/380	L	ttG/ttA	rs193302991,COSV104419804			likely_pathogenic	LOW	SNV	15301
MT-CYB	4519	Missense_Mutation	SNP	rs2853508	S2	c.580A>G	p.Thr194Ala	p.T194A	ENST00000361789	99	0	99	G	ENSG00000198727	missense_variant	580/1141	580/1141	194/380	T/A	Aca/Gca	rs2853508	tolerated_low_confidence(0.21)	benign(0.009)	benign,likely_pathogenic	MODERATE	SNV	15326
MT-ND4	4538	3'Flank	SNP	rs2853513	S2				ENST00000361381	97	0	97	T	ENSG00000198886	downstream_gene_variant						rs2853513				MODIFIER	SNV	16223
MT-ND4	4538	3'Flank	SNP	rs386829287	S2				ENST00000361381	98	.	.	T	ENSG00000198886	downstream_gene_variant						rs386829287				MODIFIER	SNV	16239
MT-ND4	4538	3'Flank	SNP	rs62581341	S2				ENST00000361381	95	1	94	C	ENSG00000198886	downstream_gene_variant						rs62581341				MODIFIER	SNV	16362

MT-ND1		4535	5'Flank	SNP	rs869183622	S3				ENST00000361390	72	0	72	G	ENSG00000198888	upstream_gene_variant						rs869183622				MODIFIER	SNV	73
MT-ND1		4535	5'Flank	SNP	rs2853515	S3				ENST00000361390	108	0	108	G	ENSG00000198888	upstream_gene_variant						rs2853515			not_provided	MODIFIER	SNV	263
MT-ND1		4535	5'Flank	SNP	rs1556422421	S3				ENST00000361390	76	0	37	C	ENSG00000198888	upstream_gene_variant						rs1556422421				MODIFIER	indel	310
MT-ND1		4535	5'Flank	SNP	novel	S3				ENST00000361390	66	0	66	C	ENSG00000198888	upstream_gene_variant										MODIFIER	SNV	372
MT-ND1		4535	5'Flank	DEL	rs78907894	S3				ENST00000361390	63	3	60	-	ENSG00000198888	upstream_gene_variant						rs78907894			uncertain_significance	MODIFIER	deletion	513
MT-ND1		4535	5'Flank	SNP	rs2853518	S3				ENST00000361390	88	0	88	G	ENSG00000198888	upstream_gene_variant						rs2853518			not_provided	MODIFIER	SNV	750
MT-ND1		4535	5'Flank	SNP	rs2001030	S3				ENST00000361390	102	0	102	G	ENSG00000198888	upstream_gene_variant						rs2001030			not_provided,benign	MODIFIER	SNV	1438
MT-ND1		4535	Silent	SNP	rs28358583	S3	c.144C>T	p.Pro48=	p.P48=	ENST00000361390	62	0	62	T	ENSG00000198888	synonymous_variant	144/956	144/956	48/318	P	ccC/ccT	rs28358583				LOW	SNV	3450
MT-ND3		4537	5'Flank	SNP	rs9659239	S3				ENST00000361227	55	0	55	A	ENSG00000198840	upstream_gene_variant						rs9659239			benign	MODIFIER	SNV	5773
MT-CO1		4512	Silent	SNP	rs370472320	S3	c.318T>C	p.Pro106=	p.P106=	ENST00000361624	48	0	48	C	ENSG00000198804	synonymous_variant	318/1542	318/1542	106/513	P	ccT/ccC	rs370472320				LOW	SNV	6221
MT-CO1		4512	Silent	SNP	rs2015062	S3	c.1125C>T	p.Ala375=	p.A375=	ENST00000361624	37	0	37	T	ENSG00000198804	synonymous_variant	1125/1542	1125/1542	375/513	A	gcC/gcT	rs2015062			benign	LOW	SNV	7028
MT-ATP6		4508	Missense_Mutation	SNP	rs2000975	S3	c.175A>G	p.Thr59Ala	p.T59A	ENST00000361899	51	0	51	G	ENSG00000198899	missense_variant	175/681	175/681	59/226	T/A	Acc/Gcc	rs2000975,COSV104419767	tolerated_low_confidence(0.64)	benign(0.005)	benign	MODERATE	SNV	8701
MT-CO3		4514	Silent	SNP	rs28358272	S3	c.243C>T	p.Tyr81=	p.Y81=	ENST00000362079	65	0	65	T	ENSG00000198938	synonymous_variant	243/784	243/784	81/261	Y	taC/taT	rs28358272				LOW	SNV	9449
MT-CO3		4514	Silent	SNP	rs2248727	S3	c.3347>C	p.Leu112=	p.L112=	ENST00000362079	106	0	106	C	ENSG00000198938	synonymous_variant	334/784	334/784	112/261	L	Tta/Cta	rs2248727,COSV104419768				LOW	SNV	9540
MT-ND3		4537	Missense_Mutation	SNP	rs28358274	S3	c.28A>G	p.Asn10Asp	p.N10D	ENST00000361227	102	0	102	G	ENSG00000198840	missense_variant	28/346	28/346	10/115	N/D	Aac/Gac	rs28358274	deleterious_low_confidence(0.03)	possibly_damaging(0.903)	benign	MODERATE	SNV	10086
MT-ND3		4537	Silent	SNP	rs28358277	S3	c.315G>A	p.Glu105=	p.E105=	ENST00000361227	96	0	96	A	ENSG00000198840	synonymous_variant	315/346	315/346	105/115	E	gaG/gaA	rs28358277				LOW	SNV	10373
MT-ND3		4537	Missense_Mutation	SNP	rs2853826	S3	c.340A>G	p.Thr114Ala	p.T114A	ENST00000361227	99	0	99	G	ENSG00000198840	missense_variant	340/346	340/346	114/115	T/A	Acc/Gcc	rs2853826	tolerated_low_confidence(0.3)	benign(0)	protective,benign	MODERATE	SNV	10398
MT-ND4		4538	Silent	SNP	rs2857284	S3	c.1147>C	p.Pro38=	p.P38=	ENST00000361381	119	0	119	C	ENSG00000198886	synonymous_variant	114/1378	114/1378	38/459	P	ccT/ccC	rs2857284,COSV104419772				LOW	SNV	10873
MT-ND4		4538	Silent	SNP	rs386829114	S3	c.243A>G	p.Gln81=	p.Q81=	ENST00000361381	98	1	97	G	ENSG00000198886	synonymous_variant	243/1378	243/1378	81/459	Q	caA/caG	rs386829114				LOW	SNV	11002
MT-ND4		4538	Silent	SNP	rs2853495	S3	c.960G>A	p.Gly320=	p.G320=	ENST00000361381	84	0	84	A	ENSG00000198886	synonymous_variant	960/1378	960/1378	320/459	G	ggG/ggA	rs2853495,COSV104668829				LOW	SNV	11719
MT-ND5		4540	Silent	SNP	rs193302956	S3	c.369C>T	p.Ile123=	p.I123=	ENST00000361567	109	0	109	T	ENSG00000198786	synonymous_variant	369/1812	369/1812	123/603	I	atC/atT	rs193302956,COSV62293963				LOW	SNV	12705
MT-ND5		4540	Missense_Mutation	SNP	rs2853501	S3	c.769A>G	p.Ile257Val	p.I257V	ENST00000361567	90	0	90	G	ENSG00000198786	missense_variant	769/1812	769/1812	257/603	I/V	Atc/Gtc	rs2853501,COSV62294516	tolerated_low_confidence(0.77)	benign(0.007)	benign	MODERATE	SNV	13105
MT-ND5		4540	Silent	SNP	rs1556424345	S3	c.1578C>A	p.Leu526=	p.L526=	ENST00000361567	101	0	101	A	ENSG00000198786	synonymous_variant	1578/1812	1578/1812	526/603	L	ctC/ctA	rs1556424345				LOW	SNV	13914
MT-CYB		4519	Missense_Mutation	SNP	rs193302980	S3	c.20C>T	p.Thr7Ile	p.T7I	ENST00000361789	95	0	95	T	ENSG00000198727	missense_variant	20/1141	20/1141	7/380	T/I	aCt/aTt	rs193302980	tolerated_low_confidence(0.05)	benign(0.003)	benign,likely_pathogenic	MODERATE	SNV	14766
MT-CYB		4519	Silent	SNP	rs193302991	S3	c.555G>A	p.Leu185=	p.L185=	ENST00000361789	111	0	111	A	ENSG00000198727	synonymous_variant	555/1141	555/1141	185/380	L	ttG/ttA	rs193302991,COSV104419804			likely_pathogenic	LOW	SNV	15301
MT-CYB		4519	Missense_Mutation	SNP	rs2853508	S3	c.580A>G	p.Thr194Ala	p.T194A	ENST00000361789	108	0	108	G	ENSG00000198727	missense_variant	580/1141	580/1141	194/380	T/A	Aca/Gca	rs2853508	tolerated_low_confidence(0.21)	benign(0.009)	benign,likely_pathogenic	MODERATE	SNV	15326
MT-CYB		4519	Missense_Mutation	SNP	rs28357376	S3	c.1078A>G	p.Thr360Ala	p.T360A	ENST00000361789	104	0	104	G	ENSG00000198727	missense_variant	1078/1141	1078/1141	360/380	T/A	Aca/Gca	rs28357376	tolerated_low_confidence(0.5)	benign(0)	benign	MODERATE	SNV	15824
MT-ND4		4538	3'Flank	DEL	rs1603225607	S3				ENST00000361381	104	0	104	-	ENSG00000198886	downstream_gene_variant						rs1603225607				MODIFIER	deletion	15939
MT-ND4		4538	3'Flank	SNP	rs2853511	S3				ENST00000361381	119	17	102	C	ENSG00000198886	downstream_gene_variant						rs2853511				MODIFIER	SNV	16093
MT-ND4		4538	3'Flank	SNP	rs386829272	S3				ENST00000361381	122	1	121	C	ENSG00000198886	downstream_gene_variant						rs386829272				MODIFIER	SNV	16124
MT-ND4		4538	3'Flank	SNP	rs386829278	S3				ENST00000361381	126	0	126	C	ENSG00000198886	downstream_gene_variant						rs386829278				MODIFIER	SNV	16209
MT-ND4		4538	3'Flank	SNP	rs2853513	S3				ENST00000361381	128	0	128	T	ENSG00000198886	downstream_gene_variant						rs2853513				MODIFIER	SNV	16223
MT-ND4		4538	3'Flank	SNP	rs41458645	S3				ENST00000361381	125	0	125	T	ENSG00000198886	downstream_gene_variant						rs41458645				MODIFIER	SNV	16278
MT-ND4		4538	3'Flank	SNP	rs3937033	S3				ENST00000361381	90	0	90	C	ENSG00000198886	downstream_gene_variant						rs3937033				MODIFIER	SNV	16519

MT-ND1		4535	5'Flank	SNP	rs869183622	S4				ENST00000361390	68	0	68	G	ENSG00000198888	upstream_gene_variant						rs869183622				MODIFIER	SNV	73
MT-ND1		4535	5'Flank	SNP	rs2853515	S4				ENST00000361390	67	0	67	G	ENSG00000198888	upstream_gene_variant						rs2853515			not_provided	MODIFIER	SNV	263
MT-ND1		4535	5'Flank	SNP	rs1556422421	S4				ENST00000361390	51	0	19	C	ENSG00000198888	upstream_gene_variant						rs1556422421				MODIFIER	indel	310
MT-ND1		4535	5'Flank	SNP	rs2853518	S4				ENST00000361390	61	0	61	G	ENSG00000198888	upstream_gene_variant						rs2853518			not_provided	MODIFIER	SNV	750
MT-ND1		4535	5'Flank	SNP	rs2001030	S4				ENST00000361390	74	0	74	G	ENSG00000198888	upstream_gene_variant						rs2001030			not_provided,benign	MODIFIER	SNV	1438
MT-ND1		4535	5'Flank	SNP	rs2854128	S4				ENST00000361390	64	0	64	G	ENSG00000198888	upstream_gene_variant						rs2854128				MODIFIER	SNV	2706
MT-ND1		4535	5'Flank	SNP	rs1057520173	S4				ENST00000361390	72	0	72	T	ENSG00000198888	upstream_gene_variant						rs1057520173			uncertain_significance	MODIFIER	SNV	3169
MT-ND1		4535	Silent	SNP	rs41423746	S4	c.42A>G	p.Leu14=	p.L14=	ENST00000361390	71	0	71	G	ENSG00000198888	synonymous_variant	42/956	42/956	14/318	L	ctA/ctG	rs41423746				LOW	SNV	3348
MT-ND2		4536	Silent	SNP	rs879108598	S4	c.1002G>A	p.Thr334=	p.T334=	ENST00000361453	67	0	67	A	ENSG00000198763	synonymous_variant	1002/1042	1002/1042	334/347	T	acG/acA	rs879108598				LOW	SNV	5471
MT-CO1		4512	Silent	SNP	rs2015062	S4	c.1125C>T	p.Ala375=	p.A375=	ENST00000361624	29	0	29	T	ENSG00000198804	synonymous_variant	1125/1542	1125/1542	375/513	A	gcC/gcT	rs2015062			benign	LOW	SNV	7028
MT-CO2		4513	Missense_Mutation	SNP	rs879119797	S4	c.220G>A	p.Val74Ile	p.V74I	ENST00000361739	50	0	50	A	ENSG00000198712	missense_variant	220/684	220/684	74/227	V/I	Gtc/Atc	rs879119797,COSV62294099	tolerated_low_confidence(1)	benign(0.005)	benign	MODERATE	SNV	7805
MT-CO2		4513	Missense_Mutation	SNP	novel	S4	c.476T>C	p.Val159Ala	p.V159A	ENST00000361739	60	21	39	C	ENSG00000198712	missense_variant	476/684	476/684	159/227	V/A	gTc/gCc		deleterious_low_confidence(0)	probably_damaging(1)		MODERATE	SNV	8061
MT-ATP8		4509	Silent	SNP	rs879190788	S4	c.42C>A	p.Thr14=	p.T14=	ENST00000361851	44	0	44	A	ENSG00000228253	synonymous_variant	42/207	42/207	14/68	T	acC/acA	rs879190788				LOW	SNV	8407
MT-ATP6		4508	Missense_Mutation	SNP	rs386829040	S4	c.31G>C	p.Ala11Pro	p.A11P	ENST00000361899	35	2	33	C	ENSG00000198899	missense_variant	31/681	31/681	11/226	A/P	Gcc/Ccc	rs386829040	deleterious_low_confidence(0.01)	possibly_damaging(0.594)	likely_benign	MODERATE	SNV	8557
MT-ND4		4538	Silent	SNP	rs2853493	S4	c.708A>G	p.Leu236=	p.L236=	ENST00000361381	61	0	61	G	ENSG00000198886	synonymous_variant	708/1378	708/1378	236/459	L	ttA/ttG	rs2853493,COSV62294568			benign,likely_pathogenic	LOW	SNV	11467
MT-ND4		4538	Silent	SNP	rs2853495	S4	c.960G>A	p.Gly320=	p.G320=	ENST00000361381	48	0	48	A	ENSG00000198886	synonymous_variant	960/1378	960/1378	320/459	G	ggG/ggA	rs2853495,COSV104668829				LOW	SNV	11719
MT-ND3		4537	3'Flank	SNP	rs2853498	S4				ENST00000361227	64	0	64	G	ENSG00000198840	downstream_gene_variant						rs2853498			benign	MODIFIER	SNV	12308
MT-ND5		4540	Silent	SNP	rs2853499	S4	c.36G>A	p.Leu12=	p.L12=	ENST00000361567	62	0	62	A	ENSG00000198786	synonymous_variant	36/1812	36/1812	12/603	L	ctG/ctA	rs2853499			benign,likely_pathogenic	LOW	SNV	12372
MT-ND5		4540	Silent	SNP	rs28397767	S4	c.165G>A	p.Met55=	p.M55=	ENST00000361567	67	0	67	A	ENSG00000198786	synonymous_variant	165/1812	165/1812	55/603	M	atG/atA	rs28397767				LOW	SNV	12501
MT-ND5		4540	Silent	SNP	rs878955011	S4	c.282G>A	p.Leu94=	p.L94=	ENST00000361567	68	2	66	A	ENSG00000198786	synonymous_variant	282/1812	282/1812	94/603	L	ttG/ttA	rs878955011,COSV62294215				LOW	SNV	12618
MT-ND5		4540	Silent	SNP	rs879201736	S4	c.1104C>T	p.Leu368=	p.L368=	ENST00000361567	76	0	76	T	ENSG00000198786	synonymous_variant	1104/1812	1104/1812	368/603	L	ctC/ctT	rs879201736				LOW	SNV	13440
MT-ND6		4541	Silent	SNP	rs878941565	S4	c.495T>C	p.Tyr165=	p.Y165=	ENST00000361681	74	0	74	G	ENSG00000198695	synonymous_variant	495/525	495/525	165/174	Y	taT/taC	rs878941565				LOW	SNV	14179
MT-CYB		4519	Missense_Mutation	SNP	rs193302980	S4	c.20C>T	p.Thr7Ile	p.T7I	ENST00000361789	71	0	71	T	ENSG00000198727	missense_variant	20/1141	20/1141	7/380	T/I	aCt/aTt	rs193302980	tolerated_low_confidence(0.05)	benign(0.003)	benign,likely_pathogenic	MODERATE	SNV	14766
MT-CYB		4519	Missense_Mutation	SNP	rs207459995	S4	c.239G>A	p.Arg80His	p.R80H	ENST00000361789	61	24	37	A	ENSG00000198727	missense_variant	239/1141	239/1141	80/380	R/H	cGc/cAc	rs207459995	deleterious_low_confidence(0.01)	probably_damaging(0.997)	uncertain_significance,pathogenic	MODERATE	SNV	14985
MT-CYB		4519	Missense_Mutation	SNP	rs2853508	S4	c.580A>G	p.Thr194Ala	p.T194A	ENST00000361789	65	0	65	G	ENSG00000198727	missense_variant	580/1141	580/1141	194/380	T/A	Aca/Gca	rs2853508	tolerated_low_confidence(0.21)	benign(0.009)	benign,likely_pathogenic	MODERATE	SNV	15326
MT-ND4		4538	3'Flank	SNP	rs2853817	S4				ENST00000361381	91	1	90	C	ENSG00000198886	downstream_gene_variant						rs2853817				MODIFIER	SNV	16172
MT-ND4		4538	3'Flank	SNP	rs878960666	S4				ENST00000361381	83	1	82	G	ENSG00000198886	downstream_gene_variant						rs878960666				MODIFIER	SNV	16219
MT-ND4		4538	3'Flank	SNP	rs41458645	S4				ENST00000361381	78	1	77	T	ENSG00000198886	downstream_gene_variant						rs41458645				MODIFIER	SNV	16278