TIAGO ANTONIO DE SOUZA

Análise das alterações genéticas em exomas de camundongos

Tese apresentada ao Programa de Pós-Graduação Interunidades em Biotecnologia da Universidade de São Paulo, Instituto Butantan e Instituto de Pesquisas Tecnológicas para obtenção do Título de Doutor em Biotecnologia.

.

São Paulo

2018

TIAGO ANTONIO DE SOUZA

Análise das alterações genéticas em exomas de camundongos

Tese apresentada ao Programa de Pós-Graduação Interunidades em Biotecnologia da Universidade de São Paulo, Instituto Butantan e Instituto de Pesquisas Tecnológicas para obtenção do Título de Doutor em Biotecnologia.

Área de concentração: Biotecnologia

Orientador: Prof. Dr. Carlos F. M. Menck

Coorientador: Dra. Silvia M. G. Massironi

Versão original.

São Paulo

2018

CATALOGAÇÃO NA PUBLICAÇÃO (CIP)
Serviço de Biblioteca e informação Biomédica
do Instituto de Ciências Biomédicas da Universidade de São Paulo

Ficha Catalográfica elaborada pelo(a) autor(a)

Souza, Tiago Antonio de
Análise das alterações genéticas em exomas de camundongos / Tiago Antonio de Souza; orientador Carlos Frederico Martins Menck; coorientador Silvia Maria Gomes Massironi. -- São Paulo, 2018.
155 p.

Tese (Doutorado)) -- Universidade de São Paulo, Instituto de Ciências Biomédicas.

1. Sequenciamento genético. 2. Bioinformática. 3. Camundongos. 4. Mutações Induzidas. 5. Mutagênese. I. Menck, Carlos Frederico Martins , orientador. II. Massironi, Silvia Maria Gomes, coorientador. III. Título.

UNIVERSIDADE DE SÃO PAULO

Programa de Pós-Graduação Interunidades em Biotecnologia

Universidade de São Paulo, Instituto Butantan, Instituto de Pesquisas Tecnológicas

Candidato: Tiago Antonio de Souza

Título da Tese: Análise das alterações genéticas em exomas de camundongos

Orientador: Prof. Dr. Carlos Frederico Martins Menck

Coorientadora: Dra. Silvia Maria Gomes Massironi

A Comissão Julgadora dos trabalhos de Defesa da Tese de Doutorado, em sessão pública realizada a / / , considerou

( ) Aprovado (a) ( ) Reprovado (a)

Examinador (a): Assinatura: .....

Nome: .....

Instituição: .....

Examinador (a): Assinatura: .....

Nome: .....

Instituição: .....

Examinador (a): Assinatura: .....

Nome: .....

Instituição: .....

Examinador (a): Assinatura: .....

Nome: .....

Instituição: .....

Presidente: Assinatura: .....

Nome: .....

Instituição: .....

UNIVERSIDADE DE SÃO PAULO
INSTITUTO DE CIÊNCIAS BIOMÉDICAS
COMISSÃO DE ÉTICA NO USO DE ANIMAIS

Cidade Universitária "Armando de Salles Oliveira", Butantã, São Paulo, SP · Av. Professor Lineu Prestes, 2415 - ICB III - 05508 000
Comissão de Ética em Pesquisa - Telefone (11) 3091-7733 - e-mail: cep@icb.usp.br

CERTIFICADO

Certificamos que a solicitação de licença de uso de animais intitulada "*Análise das alterações genéticas em exomas de camundongos*", registrada sob nº 053, nas fls. 32, do livro 03, foi analisada e aprovada pela COMISSÃO DE ÉTICA NO USO DE ANIMAIS (CEUA-ICB/USP) em 05 de maio de 2015.

Por esta licença, estão autorizados a manipular animais dentro dos limites do projeto proposto e no âmbito da **Lei Federal nº 11.794**, o Dr. (Dra.) Carlos Frederico Martins Menck (Investigador Principal) e os membros da equipe Tiago Antonio de Souza, Silvia Maria Gomes Massironi. Esta licença de uso de animais expira em 05/05/2019.

Havendo interesse na renovação da proposta, a solicitação deverá ser protocolada pela secretaria da CEUA-ICB/USP **até o último dia de validade da atual proposta**. Após essa data, uma nova proposta deverá ser encaminhada.

CERTIFICATE

We hereby certify that permission for the use of animals was granted to the research proposal *Analysis of genetic alterations in exomes from mice*, registered as Number 053, in pages 32 of book 03, by the local ETHICS COMMITTEE ON THE USE OF ANIMALS (CEUA-ICB/USP) in 05/05/2015.

Under this license, Carlos Frederico Martins Menck (Principal Investigator) and team members Tiago Antonio de Souza, Silvia Maria Gomes Massironi are authorized to make use of animals within the limits of the research proposal presented to this committee and of the **Brazilian Federal Law nº 11.794**.

This license expires in 05/05/2019. In case the investigators wish to renew this license, this must be presented to CEUA-ICB/USP before the last day of validity of the present license. After such date, a new research proposal must be presented.

São Paulo, 05 de maio de 2015.

Prof. Dr. Wothan Tavares de Lima
Coordenador-CEUA- ICB/USP

Profa. Dra. Ana Paula Lepique
Secretária- CEUA - ICB/USP

Para meus pais, **Maria** e **Mauri** (*in memoriam*), minha irmã **Maíra**
e para **Marcela**.

“A vida é arte do encontro. Embora haja tanto desencontro pela vida.”

V. Moraes

AGRADECIMENTOS

Ao meu orientador, **Carlos Menck**. Tenho uma profunda admiração pelos seus ideais e projetos, pela sua enorme sabedoria e competência, pela gentileza que você naturalmente envolve todos a sua volta – sem nenhuma exceção – e por sua incansável e admirável paixão pela ciência e pela educação. Sou muito grato por você ter me dado a oportunidade e confiança para que eu realizasse esse doutorado, apesar de todas as circunstâncias. Guardo comigo várias conversas que tivemos durante todos esses anos e considero isso uma das coisas mais valiosas que tenho comigo. É um orgulho e uma alegria imensa tê-lo conhecido e convivido com você. Muito obrigado Menck!

À minha co-orientadora **Sílvia Massironi**, pelo enorme apoio, orientação, motivação, oportunidades e por ter permitido que eu pudesse contribuir um pouco com seu belo trabalho. Um precioso exemplo de dedicação, responsabilidade e competência científica junto a uma rara gentileza e serenidade, em todas as circunstâncias. Obrigado por partilhar comigo um pouco do seu inestimável conhecimento na ciência de animais de laboratório, de uma forma tão paciente e carinhosa.

À minha mãe, **Maria Aparecida da Trindade Souza**. Essa tese definitivamente não aconteceria se não fosse pelas suas grandes e pequenas decisões, feitas corajosamente ao longo de toda uma vida e sempre cheias de amor. Passamos por momentos difíceis, de perda e de incertezas, mas você escolheu lutar e continuar. E eu tive a sorte acompanhá-la. Agora somos mais fortes, ou pelo menos um pouco mais experientes. Tenho muito orgulho em ser filho de uma mulher tão inteligente e que se mantem cada vez mais jovem, otimista e alegre e sempre disposta a abraçar todos a sua volta.

À **Marcela Mineiro,** pelo companheirismo, amor e por sempre me lembrar da preciosidade dos pequenos momentos, e que podemos ser felizes e plenos se soubermos aproveitá-los. Compartilhamos um jeito parecido de encarar a vida, mas também os desafios e as dúvidas, e talvez esse seja uma das grandes razões de estarmos juntos. Obrigado pelo carinho, pela compreensão e pela paciência, e também por acreditar que essa tese poderia se tornar de fato uma realidade.

À **Cláudia Mori** e todo seu grupo pelas sugestões e pelas construtivas reuniões, onde sempre aprendia um pouco mais sobre análises de comportamento animal. Tenho certeza que trabalhos bonitos relacionados aos mutantes surgirão em breve. Ao **Márcio Caldas**, pela ajuda inestimável no início do projeto com os camundongos e a todos do **Biotério de Experimentação do Departamento de Imunologia**, autores de um trabalho de excelente qualidade e essencial para que trabalhos de qualidade em pesquisa biomédica possam florescer.

Ao **Niels Olsen**, um grande cientista, professor, *chief* e ser humano. Seu apoio, aconselhamento e compreensão, em todas as etapas dessa trajetória, mesmo que indiretamente, foram essenciais. Espero assim, levar um pouco dos valores e princípios que você naturalmente deixa transparecer em suas ações.

Ao grande **Fernando Almeida** (Fernandinho), pela inestimável amizade cuidadosamente construída ao longo de todos esses anos de USP. A **Susan Ienne**, obrigado pelo apoio, compreensão e equilíbrio durante todos esses anos de convivência diária no laboratório, regados ao famoso cappuccino de astronauta. Ao **Maurício Lopes**, pelas incríveis e agradáveis conversas sobre quase tudo - e pela apresentação ao mundo literário dos grandes escritores argentinos. Ao **Alexandre Defelicibus** (Difiícilimus), companheiro de idéias mirabolantes e dono de uma presença sempre muito agradável. Obrigado por ter me introduzir – sem volta – ao mundo do Git, Trello,

Sublime e SnakeMake. Ao **Jonas Gaiarsa**, verdadeiro, competente e brilhante bioinformata, pelas ajudas com os servidores, bash e nosso amado Perl. E a todos os demais membros e ex-membros do **CEFAP**!

À todos os ex-membros e membros do **Laboratório de Reparo de DNA**, um grupo único e unido, repleto de grandes cientistas e seres humanos, tão raro de se encontrar. E particularmente pela ótima receptividade diária nos últimos meses do doutorado!

Aos meus grandes amigos e amigas da **UNICAMP** e do **COLUNI-UFV,** que proporcionam o que só as verdadeiras amizades são capazes de dar – e que não mudam – ignorando microvariáveis como a distância ou o tempo.

Aos membros da banca de qualificação, professores **Emmanuel D. Neto**, **Pedro Galante** e **Maria Z. Dagli** pela disponibilidade e pelas excelentes discussões e sugestões.

À **Laura Reinholdt** (The Jackson Laboratory) pelo envio dos dados relativos ao sequenciamento dos camundongos C57BL/6J e BALB/cJ.

Ao excelente apoio e ajuda de todos da **Secretaria de Biotecnologia**, em especial à **Fabia** e **Eliane**, sempre muito competentes e dispostas – mesmo nos momentos mais complicados.

Ao **Centro de Facilidades de Apoio a Pesquisa – Universidade de São Paulo** (CEFAP-USP), em especial à *facility* **GENIAL**, pela disponibilização das plataformas de sequenciamento NGS e toda a sua infraestrutura. Um especial agradecimento ao apoio dos professores **Rui Curi** e **Benedito Correa**, e à toda valorosa **comunidade de funcionários do ICB**.

À **Universidade de São Paulo** e ao **Instituto de Ciências Biomédicas** pela oportunidade dada para a realização deste doutorado.

Ao **Conselho Nacional de Desenvolvimento Científico e Tecnológico** (CNPq) – processo 141411/2017-1 e à **Fundação de Amparo à Pesquisa do Estado de São Paulo** (FAPESP) – processos 2012/25387-2 e 2009/53994-8, pelo valioso apoio financeiro. Espero que o cuidado na manutenção das oportunidades dadas por essas agências se mantenha, principalmente nos momentos mais adversos – quando é mais preterido e por isso deve se mostrar mais resiliente.

Sobre o Rigor na Ciência

"Naquele império, a Arte da Cartografia alcançou tal Perfeição que o mapa de uma única Província ocupava uma cidade inteira, e o mapa do Império uma Província inteira. Com o tempo, estes Mapas Desmedidos não bastaram e os Colégios de Cartógrafos levantaram um Mapa do Império que tinha o Tamanho do Império e coincidia com ele ponto por ponto.
Menos Dedicadas ao Estudo da Cartografia, as gerações seguintes decidiram que esse dilatado Mapa era Inútil e não sem Impiedade entregaram-no às Inclemências do sol e dos Invernos. Nos Desertos do Oeste perduram despedaçadas Ruínas do Mapa habitadas por Animais e por Mendigos; em todo o País não há outra relíquia das Disciplinas Geográficas.

(Suaréz Miranda: Viajes de Varones Prudentes, livro quarto, cap. XLV, Lérida, 1658)"

***Jorge Luis Borges**: "Del rigor en la ciencia" na seção Museo de El Hacedor (1960)*

RESUMO

DE SOUZA TA. **Análise das alterações genéticas em exomas de camundongos**. 2018. 155 f. [Tese (Doutorado em Biotecnologia)]. São Paulo: Instituto de Ciências Biomédicas, Universidade de São Paulo; 2018.

Camundongos são modelos valiosos para o entendimento dos processos e mecanismos moleculares e fisiológicos em mamíferos. A maioria do nosso conhecimento sobre esses processos e mecanismos vem de experimentos realizados com camundongos de linhagens isogênicas. Essas linhagens, criadas normalmente por sucessivos cruzamentos irmão-irmã, surgiram no início do século XX visando reduzir a interferência da variabilidade genética, aumentando a reprodutibilidade dos experimentos. Caracterizar o *background* genético das linhagens isogênicas permite não só traçar possíveis relações de parentesco entre linhagens, mas também permite o controle genético oriundo de possíveis contaminações e mutações espontâneas que possam surgir na população. Além das linhagens isogênicas, os camundongos mutantes também são importantes como modelos para o estudo de doenças humanas. O uso desses modelos murinos permite a elucidação e associação de fatores genéticos a manifestações fenotípicas diversas, como síndromes hereditárias e predisposições a doenças. Esses mutantes podem ser gerados por uma abordagem de varredura de mutagênese pelo agente mutagênico ENU, que inclui a caracterização de fenótipos interessantes e a busca pelas mutações causativas induzidas. O presente trabalho teve como objetivo utilizar o sequenciamento completo de exomas para caracterizar o *background* genético das linhagens isogênicas C57BL/6ICBI e BALB/cICBI, mantidas há quase 20 anos no Brasil e distribuídas pelo ICB-USP a pesquisadores de todo país. O trabalho também usou o sequenciamento de nova geração (NGS) para a busca das mutações causadores de fenótipo em um grupo de sete mutantes induzidos por ENU oriundos de uma varredura prévia. Através da aplicação de uma estratégia de análise de dados e filtragem de mutações foi possível encontrar mutações candidatas com alto potencial de impacto para todos os mutantes avaliados, validadas por sequenciamento Sanger. Os genes afetados pelas mutações encontradas indicam que os mutantes possam se tornar interessantes modelos para o estudo de doenças neuromusculares e neurológicas. A avaliação do exoma das linhagens C57BL/6ICBI e BALB/cICBI descartou a possibilidade de contaminação das colônias com outras linhagens, e revelou similaridades relacionadas com o parentesco das sublinhagens brasileiras em relação a linhagens *gold-standard.* As informações obtidas serão uma fonte importante de informação no planejamento e análise dos resultados obtidos com o uso tanto dos mutantes quanto com as linhagens fornecidas pelo Biotério do Departamento de Imunologia ao ICB a instituições de todo o Brasil.

Palavras-chave: Sequenciamento genético. Bioinformática. Camundongos. Mutações Induzidas. Mutagênese.

ABSTRACT

DE SOUZA TA. **Analysis of genetic alterations in mice exomes**. 2018. 155 f. [Ph.D. Thesis (Biotechnology)]. São Paulo: Instituto de Ciências Biomédicas, Universidade de São Paulo; 2018.

Mice are valuable models for the comprehension of molecular processes and underlying physiological mechanisms in mammals. Most of the knowledge about those processes came from experiments with isogenic mice. Those strains, arose in the 1900's by successive inbreeding, are very important as they reduce genetic variability across the experiments increasing reproducibility. Isogenic lineages are kept as isolated colonies in animal facilities and supplied to researchers, as they needed. Thus, is possible to trace relationships among strains all over the world using the characterization of their genetic backgrounds. It is also possible to detect putative contaminations and spontaneous mutations which can arise in the populations. Mutant mice are also important tools as human disease models, allowing associations between genetic factors and phenotypes. Those mutants could be generated in forward genetics approaches by screenings using mutagens as ENU. The aims of this work were to characterize the genetic background of two mouse strains used at ICB-USP – C57BL/6ICBI and BALB/cICBI – and to find causative mutations of seven mutants generated by a previous ENU-mutagenesis screening. We used whole-exome sequencing followed by resequencing data-analysis approaches to detect SNVs for both isogenic strains and mutants. Exome evaluation of isogenic strains C57BL/6ICBI and BALB/cICBI did not reveal any evidence for cross-contamination and provided insightful details related to other strains and substrains. A specific filtering strategy was applied to select candidates for phenotype-causative mutations in the seven ENU-induced mutants. We are able to select candidates for all mutants at a high global Sanger validation rate when considering only the main candidates for each mutant. Considering affected genes and phenotypes all mutants have potential to become interesting mouse models for human diseases. Taken together, our results are a reliable and confident source of genetic information for experimental analysis for researchers who use isogenic strains provided by animal facility at ICB-USP and research groups interested in further characterization of mutant study neuromuscular, neuronal or development processes using mice as animal models.

Keywords: DNA Sequencing. Bioinformatics. Mice. Induced mutations. Mutagenesis.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

Ataxrec1: mutante atáxico-1

BAM: *Binary version of SAM (Sequence Alignment Map)*

Bapa: mutante bate palmas

BED: *Browser Extensible Data*

BSA: *Bovine Serum Albumine. S*oroalbumina bovina

Carc: mutante careca

CC: *Colaborative Cross*

CCDS: *Consensus Coding Sequence*

CNV: *Copy Number Variation.* Variação no número de cópias

CRISPR: *Clustered Regularly Interspaced Short Palindromic Repeats*

Crm: cromossomo

Crup: mutante cruza pernas

CTD: *C-terminal domain,* domínio C-terminal

DHT: dihidrotestosterona

DNA: *Deoxyribonucleic acid*. ADN ou Ácido desoxirribonucleico

DO: *Diversity Outbred*

Ds/Dn: Razão entre Mutações sinônimas (Ds) e Mutações não-sinônimas (Dn)

Dst: *Dystonin* (Distonina)

EAE: *Experimental Autoimmune Encephalomyelitis.*

EDTA: *Ethylenediamine tetraacetic acid*

EMS: *Ethyl methanesulfonate*

ENCODE: *Encyclopaedia of DNA Elements*

ENU: *N-ethyl-N-nitrosourea*

Eqlb: Mutante equilíbrio

Fraq: Mutante fraqueza

GATK: *Genome Analysis Toolkit*

GO: *Gene Ontology*

GWAS: *Genome Wide Association Studies*

H3K4me1, H3K4me2, H3K4me3: Mono,di,tri metilação em K4 da histona H3

Het: heterozigoto

HGNC: *Hugo Gene Nomenclature Committee*

Hom: homozigoto

HRP: *horseradish peroxidase*

HUGO: *Human Genome Organisation*

ICB: Instituto de Ciências Biomédicas

INDEL: *Insertion or deletion.* Inserção ou deleção

JAX: *The Jackson Laboratory,* EUA.

Kmt2d: *Lysine Methyltransferase 2D*

LUSC: *Lung squamous cell carcinoma*

MGI: *Mouse Genome Informatics*

MGP: *Mouse Genomes Project*

Mm10: *Mus musculus genome version 10*

Mm9: *Mus musculus genome version 9*

NADPH: *Nicotinamide Adenine Dinucleotide Phosphate*

NGS*: Next-Generation Sequencing,* Sequenciamento de Nova Geração

NIH: *National Institutes of Health*

NK: Célula *Natural Killer*

NMD: *Nonsense-mediated mRNA decay*

OMIM: *Online Mendelian Inheritance in Man*

PBS: *Phosphate-buffered saline*

PCR: *Polimerase chain reaction*

QV: *Quality value*

RFLP: *Restriction fragment length polymorphism*

RNA: *Ribonucleic acid*. ARN ou Ácido ribonucleico.

RT-PCR: *Reverse transcription polimerase chain reaction*

Sacc: mutante sacudidor.

SAET: *SOLiD Accuracy Enhancement Tool*

SKCM: *Skin cutaneous melanoma*

SNP: *Single nucleotide polymorphism.*

SNV: *Single nucleotide variant*

SOLiD: *Sequencing by Oligonucleotide Ligation and Detection*

T: testosterona

TCGA: *The Genome Cancer Atlas*

HNSC: *Head and neck squamous cell carcinoma*

Tween-20: *Polyoxyethylene (20) sorbitan monolaurate*

UCSC: *University of California – Santa Cruz,* EUA.

UTR: *Untranslated region*

VCF: *Variant call format*

WES: *Whole exome sequencing*

WFA: *Workflow analysis*

WGS: *Whole genome sequencing*

WHO: *World Health Organization*. OMS ou Organização Mundial da Saúde.

WT: *wild-type,* selvagem

XSQ: *eXtensible SeQuence format*

PREFÁCIO

Esta tese teve como objetivos o desenvolvimento e aplicação das tecnologias de sequenciamento de nova geração (NGS) para detecção e caracterização de variantes genéticas em camundongos. Foi um trabalho que abrangeu toda a parte experimental para a obtenção das sequências de NGS e todas as análises *in silico* utilizadas para a interpretação desses dados. A tese também inclui, em seu último capítulo, o desenvolvimento de uma ferramenta para a análise do padrão de mutações pontuais em dados de resequenciamento.

O primeiro capítulo trata de uma introdução geral ao uso dos camundongos como organismo modelo e apresenta os objetivos e metodologias gerais utilizados no trabalho, relacionados principalmente ao próprio sequenciamento do exoma dos camundongos.

O Capítulo 2 trata da caracterização de duas linhagens isogênicas mantidas no Biotério do Departamento de Imunologia e distribuídas a diversos pesquisadores do ICB e de outras unidades. É um esforço pioneiro de caracterização do *background* genético de camundongos isogênicos mantidos em território brasileiro.

O Capítulo 3 abrange a estratégia para encontrar as mutações causais em um grupo de camundongos mutantes com fenótipos comportamentais e fisiológicos diversos, originados a partir de um estudo anterior utilizando mutagênese induzida por ENU. Esse capítulo também apresenta, além da indicação e validação de importantes candidatos em todos os mutantes, análises do impacto das mutações na função gênica.

Finalmente, no Capítulo 4, é apresentado o desenvolvimento de uma ferramenta *in silico* para a análise de padrões de mutação utilizando dados de NGS. Apesar do desenvolvimento da ferramenta ter sido motivado dentro do contexto da exploração das mutações provocadas pelo agente ENU, seu uso tem se mostrado satisfatório e até mesmo promissor na avaliação exploratória de processos gerais de mutagênese utilizando dados de resequencimento.

SUMÁRIO

CAPÍTULO 1- SEQUENCIAMENTO COMPLETO DO EXOMA DE CAMUNDONGOS

CAPÍTULO 1- SEQUENCIAMENTO COMPLETO DO EXOMA DE CAMUNDONGOS

1.1 Introdução

A compreensão da biologia humana é um dos aspectos principais da própria biologia como ciência. Entendê-la significa melhorar a qualidade de vida em termos de prevenção e tratamento de fatores que afetam a saúde humana. Todo o funcionamento do corpo humano é dependente de maquinarias moleculares codificadas pelo nosso genoma, por suas interações com o microambiente interno fisiológico e também com a influência de todo um macroambiente externo. Dessa forma, o desenvolvimento do ser humano em um organismo envolve uma base genética determinística, que se torna incrivelmente complexa quando consideramos os fatores ambientais. Correlacionar essas bases genéticas – cada vez mais conhecidas devido ao crescente uso das tecnologias de sequenciamento de DNA – às características físicas de um indivíduo significa melhorar nossa compreensão sobre o próprio papel das funções moleculares e fisiológicas determinadas pelo nosso genoma na manutenção da vida. Uma melhor compreensão desses mecanismos possibilita, por exemplo, melhores métodos de intervenção para quaisquer tipos de doenças que possam prejudicar o indivíduo e a população humana em geral.

1.1.1 Panorama geral das relações genótipo-fenótipo em humanos

Associar uma base genética - ou genótipo - a uma determinada manifestação de uma característica “física” – ou fenótipo – ou seja, compreender a função de um ou mais genes, não é uma tarefa trivial (ANTONARAKIS; BECKMANN, 2006). Essa associação começou a ser compreendida pelos estudos pioneiros realizados por Gregor Mendel, em meados do século XIX, e só redescoberta no início do século XX por William Bateson. Bateson cunhou pela primeira vez o termo “genética” em 1905 e traduziu o trabalho de Mendel para o mundo no livro *Mendel’s principles of heredity* em 1909 (BATESON; MENDEL, 1909). Os conceitos de hereditariedade definidos por Mendel em plantas foram pela primeira vez utilizados no estudo da alcaptonúria em 1902, considerada a primeira doença genética a ser descrita (GARROD, 1902). Teve início então a era da genética humana, abrindo possibilidade para a realização de predições sobre o fenótipo dos indivíduos a partir de suposições de seus genótipos. Essas doenças hereditárias, cujo padrão hereditário estava associado à

segregação de alelos de genes específicos segundo as leis propostas por Mendel, ficaram conhecidas como síndromes mendelianas. Esses alelos mendelianos, como são conhecidos, são muito raros na população humana, refletindo a baixa incidência das síndromes mendelianas – elas afetam somente cerca de 10 em cada 1000 pessoas no mundo (WHO, 2017[1]). Atualmente são conhecidas mais de 5.000 síndromes mendelianas, clinicamente associadas a um padrão de herança por um único gene, por isso denominadas síndromes mendelianas monogênicas ou simplesmente síndromes monogênicas (OMIM, 2017[2]).

Um dos principais motivos para a euforia da divulgação do sequenciamento do genoma humano, no início do século XXI, era a descoberta das sequencias de todos os genes humanos (CROLLIUS et al., 2000). Supunha-se, na época, que o genoma humano poderia conter 100.000 ou mais genes, mas esse número mostrou-se quase 5 vezes menor que a maioria das previsões (HGP, 2001; VENTER et al., 2001). Genes que codificam para proteínas correspondem apenas a uma pequena parcela de todo o genoma, em termos de número de pares de bases. Não há dúvidas, portanto, que o sequenciamento do genoma humano trouxe grandes descobertas e informação para o estudo das funções dos genes em relação ao desenvolvimento dos estudos de doenças mendelianas e da própria genética médica em si (HEARD et al., 2010), que iriam surgir ao longo dos anos 2000 até os dias de hoje, em forma de bancos de dados e catálogos.

O catálogo OMIM (*Online Mendelian Inheritance In Man*), criado no começo dos anos 1960 por Victor A. McKusick, acumula dados funcionais de todas as síndromes mendelianas e também informações associadas de mais de 15.000 genes (AMBERGER et al., 2015). Apesar da grande quantidade de dados, apenas 3.825 genes com mutações foram associados como causais em relação à um determinado fenótipo clínico (OMIM, 2017[3]). A grande maioria dos genes - 3.445 - está relacionada com doenças e traços monogênicos, enquanto 499 estão associados com susceptibilidade a doenças complexas ou infecções. Apenas 115 genes estão associados a variações benignas em testes laboratoriais - deficiência na lactato desidrogenase B, por exemplo. Apenas 121 genes estão associados a

---

[1] World Health Organization, WHO, 27/10/2017: http://www.who.int/genomics/public/geneticdiseases
[2] Online Mendelian Inheritance in Man, OMIM®. McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University (Baltimore, MD), 27/10/2017: https://omim.org/
[3] OMIM® *Dissected Morbid Map Scorecard*, 16/10/2017: https://www.omim.org/statistics/geneMap

doenças genéticas em células somáticas - como a síndrome de McCune-Albright e o gliobastoma multiforme. O número de genes humanos segundo os parâmetros do comitê HUGO[4] é de 20.212 genes sendo 19.109 genes classificados como codificadores de proteínas segundo o HGNC[5]. Considerando esses dados, apenas 18% dos genes têm até o momento algum tipo de fenótipo clínico associado em humanos, ou seja, são genes cujo impacto fenotípico e papel em doenças hereditárias pode ser inferido com um grau considerável de precisão. Se considerarmos uma definição de gene um pouco mais abrangente, incluindo além de segmentos de DNA que codificam proteínas também outros segmentos que codificam para RNAs longos não codificantes, pseudogenes, pequenos RNAs não codificantes e miRNAs, o número de associações fenotípicas descobertas é ainda menor (ZHANG; LUPSKI, 2015).

O panorama fica um pouco mais complexo ao considerarmos todas as regiões do genoma que não são consideradas como genes que codificam proteínas. Essas regiões não-codificantes perfazem cerca de 98,5% de todo o genoma humano – e ficaram conhecidas pelo famoso termo “DNA lixo”, que se popularizou após o sequenciamento do genoma humano (HEARD et al., 2010). Esse termo se mostrou de fato muito equivocado, principalmente depois dos estudos do projeto ENCODE - *Encyclopaedia of DNA Elements*[6] - iniciado em 2003. Segundo o ENCODE quase 80% do genoma possui algum tipo de atividade regulatória, entre mais de 70.000 promotores e quase 400.000 *enhancers* (ENCODE, 2012). O projeto, ainda em andamento, empregou um vasto conjunto de técnicas de identificação de sequências regulatórias, em mais de 10.000 amostras de linhagens celulares, tecidos, células primárias e células tronco somente em humanos, revelando uma rede de regulação gênica descomunal (ENCODE, 2012; KELLIS et al., 2014). Vários desses elementos possuem uma correspondência estatística a variantes associadas a doenças humanas e constituem uma base essencial para guiar a interpretação dos mecanismos moleculares associados a doenças e processos biológicos (SCHAUB et al., 2012).

---

[4] HUGO - *Human Genome Organisation:* www.hugo-international.org
[5] HGNC - *Hugo Gene Nomenclature Committee, 16/10/2017:* https://www.genenames.org/cgi-bin/statistics
[6] ENCODE – Encyclopaedia of DNA Elements: https://www.encodeproject.org/

No caso de manifestações fenotípicas em que não é possível estabelecer um padrão claro de herança mendeliana, a situação é um pouco mais dramática. Na tentativa de estabelecer conexões entre variações genotípicas e fenotípicas são utilizadas abordagens chamadas de GWAS - *Genome Wide Association Studies* (BUSH; MOORE, 2012). Essas abordagens consistem no agrupamento sistemático de indivíduos com um determinado traço fenotípico e a correlação de variantes comuns entre os membros desse grupo com a incidência do fenótipo, seguida pela inferência independente da predisposição associada às variantes detectadas (VISSCHER et al., 2012). Esses estudos envolvem comumente a genotipagem por microarranjos de um grande número de indivíduos, podendo chegar de 10.000 a 100.000 indivíduos (BUSH; MOORE, 2012). A aplicação clínica dessas abordagens tem contribuído muito para a avaliação da heritabilidade e o cálculo dos fatores de risco associados a características complexas, como o câncer (SUD et al., 2017), obesidade (GHOSH; BOUCHARD, 2017), diabetes e esquizofrenia (KURE; ANTONIO, 2017), por exemplo.

No entanto, abordagens do tipo GWAS têm tido um sucesso limitado (PASANIUC et al., 2012) na elucidação das bases genética no contexto molecular, funcional e biológico. Ou seja, a falta de um padrão de heritabilidade interpretável biologicamente em manifestações fenotípicas é mais uma regra do que uma exceção. Mais de 80% dos indivíduos presentes nos estudos de GWAS até 2016 tinham ancestralidade europeia, revelando um viés preocupante desses estudos (POPEJOY; FULLERTON, 2016). Outros fatores podem também afetar associações com o fenótipo, como os efeitos de segregação conjunta em variantes próximas, epistasia e variações ambientais. Esses fatores também estão associados a taxas de sucesso reduzidas em abordagens GWAS, justificando a relevância do uso de modelos animais (ERMANN; GLIMCHER, 2012).

1.1.2 Variabilidade genética em populações humanas

Dado esse panorama de regiões que podem estar diretamente envolvidas em manifestações fenotípicas em humanos é necessário compreender a origem e identificar a quantidade de variações genéticas hereditárias presentes na população humana. Estima-se que dois indivíduos não relacionados compartilhem entre si cerca entre 99,5 e 99,9% de identidade em seus genomas (LEVY et al., 2007; VENTER et al., 2001). Dessa forma, apenas 0,5% - ou ~ 15 milhões de bases -

constituem variações em termos de sequência, tanto polimorfismos quanto variantes raras que podem estar potencialmente associadas a doenças, síndromes mendelianas e predisposições genéticas variadas.

O projeto *1000 Genomes* iniciou-se em 2008 com o objetivo de mapear os polimorfismos presentes na população humana através do sequenciamento de indivíduos saudáveis provenientes da América, Europa, África e Ásia. No total 2.504 indivíduos de 26 populações foram sequenciados, incluindo genomas completos, exomas e genotipagem por microarranjos. Mais de 88 milhões de variantes foram encontradas – cerca de 84,7 milhões consistiam apenas de SNPs - *Single Nucleotide Polymorphisms* - sendo que mais de 99% possuía frequência maior que 1% (1000 GENOMES PROJECT et al., 2015). Esse número de variantes é proveniente dos processos mutagênicos espontâneos que originam mutações pontuais em células germinativas a uma taxa estimada de 1×10^{-8} nucleotídeos por sítio por geração em humanos. A taxa de mutação somática espontânea em células epiteliais do intestino é cerca de 13 vezes maior do que em células germinativas e por extrapolação cerca de 5 vezes maior em fibroblastos e linfócitos (LYNCH, 2010). Embora as taxas de mutação sejam relativamente baixas, a fixação dessas mutações germinativas pode ser seriamente agravada em populações isoladas, onde a consanguinidade é comum e, portanto, a taxa de síndromes mendelianas é maior (HAMAMY et al., 2011).

De fato, a consanguinidade familiar ainda é um dos fatores que mais colaboram com a caracterização da base genética de doenças mendelianas e a determinação da função dos genes em humanos (CHONG et al., 2015). Em grande parte do mundo essa taxa é menor que 5%, mas atinge taxas elevadas em algumas regiões - entre 20% e 50% no Oriente Médio, norte da África e no sudoeste da Ásia por exemplo - devido principalmente a particularidades culturais e religiosas. A análise dos *pedigrees* familiares aliado ao acompanhamento clínico permite a inferência do padrão de herança de uma determinada síndrome, restringindo a quantidade de genes que podem estar afetados (HAMAMY, 2012; HAMAMY et al., 2011). Um exemplo marcante no Brasil é a comunidade de Araras, situada no interior do estado de Goiás, onde a incidência de uma doença grave chamada de xeroderma pigmentosum é cerca de 100 vezes maior em comparação com a taxa de incidência estimada no mundo. Essa alta incidência é devido à presença de dois alelos mutantes no gene codificante para a POLH – DNA polimerase eta – ou XPV,

fixados na população devido a elevada taxa de casamentos consanguíneos entre os moradores da comunidade (MUNFORD, CASTRO et al., 2017).

1.1.3 O uso de camundongos como modelo animal

A principal consequência do acúmulo de informações de variantes provenientes dos maiores projetos de genômica humana, como os citados anteriormente, é justamente a dificuldade em conectá-las de forma a compreender a base genética das doenças. Isso acontece porque a genética humana é difícil: estudos de gerações são complicados, os acasalamentos não são pré-definidos e o tempo de uma geração é muito longo. Além disso, o acesso a tecidos é limitado e a genética humana é basicamente observacional e não interventiva. O reflexo dessas questões é a dificuldade em prever que determinadas mutações influenciem e até causem determinado fenótipo. Felizmente, praticamente todas essas limitações são superadas com o uso de modelos animais (ERMANN; GLIMCHER, 2012).

Um dos modelos animais mais usados é o camundongo, cujo genoma, anatomia e fisiologia são muito similares ao homem. Homens e camundongos têm uma linguagem similar de regulação gênica e compartilham cerca de 80% das sequências codificadoras de proteínas (CHURCH et al., 2009). Camundongos também possuem período de gestação e idade fértil bem menores que o homem, além de serem dóceis e com custo de manutenção relativamente baixo. Além de todos esses motivos, a maioria das técnicas de genética clássica e de genética molecular estão muito bem estabelecidas em camundongos, facilitando a manipulação em nível molecular dos genes-alvo (WEYDEN et al., 2011).

O uso de camundongos como modelo animal (**Figura 1.1**) se relaciona principalmente a difusão do uso de linhagens isogênicas de camundongos (**Figura 1.1A e B**), que se destacam pela maior similaridade genética entre seus membros. As estratégias de cruzamento para a obtenção de linhagens isogênicas consistem basicamente de sequências de cruzamentos irmão-irmã. Uma linhagem é considerada isogênica quando ocorre o processo de acasalamento irmão-irmã por pelo menos 20 gerações sequenciais. A identidade alélica da geração 20 de um esquema de cruzamento desse tipo é estimada em cerca de 98%, sendo praticamente idêntica ao genoma compartilhado entre às matrizes da geração anterior (SILVER, 1995).

Figura 1.1 Camundongos isogênicos como modelo animal. Linhagens isogênicas, como camundongos BALB/c (**A**) e C57BL/6 (**B**) estão entre os mais utilizados. Camundongos geneticamente diversos, como os Diversity Outbred – DO (**C**) podem ser usados para mapeamento de alta resolução e validação de *loci* ligados a susceptibilidade a doenças, resistência a drogas e fenótipos de comportamento. Fonte: (**A**) e (**B**) – Charles River Laboratories. (**C**) The Jackson Laboratory.

A utilização de linhagens isogênicas permitiu muitos avanços em estudos fisiológicos, anatômicos, comportamentais e genéticos, devido principalmente a um aumento na reprodutibilidade e no efeito do *background* genético nos experimentos (JUSTICE; DHILLON, 2016; MANDILLO et al., 2008; RICHTER et al., 2011). Em alguns casos, esse *background* genético homogêneo não mimetiza a variabilidade genética naturalmente encontrada em populações humanas. Um exemplo é o estudo da resposta a medicamentos na população humana (BOGUE; CHURCHILL; CHESLER, 2015). Atualmente iniciativas de criação de linhagens geneticamente diversas - mas com background conhecido e que ainda mantêm reprodutibilidade alta - como o *Collaborative Cross* (CC), RIX e *Diversity Outcross* ou DO **(Figura 1.1C**) têm sido cada vez mais utilizadas para espelhar a diversidade encontrada naturalmente em populações humanas e são atualmente tendências nesse sentido (BOGUE; CHURCHILL; CHESLER, 2015; YANG et al., 2011). Essas linhagens, criadas a partir de cruzamentos aleatórios de 8 linhagens isogênicas seguida por pelo menos 20 gerações de cruzamentos irmão-irmã possibilitam a inclusão de diversidade genética controlada em experimentos, preservando a possibilidade de o uso de painéis de genotipagem para identificação das variações oriundas das linhagens isogênicas utilizadas (BOGUE; CHURCHILL; CHESLER, 2015).

Os camundongos, junto com as moscas *Drosophila melanogaster,* foram os primeiros organismos mais complexos utilizados para o desenvolvimento de técnicas de manipulação genética (GUÉNET et al., 2015). As tecnologias de manipulação e também de detecção de macromoléculas biológicas tiveram uma importância substancial para o entendimento da biologia humana e na compreensão de

processos biológicos, formando a base do que hoje entende-se por de genética molecular. As tecnologias de manipulação genética em camundongos podem ser consideradas as mais avançadas entre os modelos de mamíferos e várias dessas tecnologias foram desenvolvidas e aprimoradas em camundongos isogênicos (EPPIG et al., 2015). Estratégias clássicas de mutagênese (MORESCO; LI; BEUTLER, 2013) e transgenia (BOUABE; OKKENHAUG, 2015) foram utilizadas de maneira clássica em camundongos. Abordagens modernas como o silenciamento por RNA de interferência (PREMSRIRUT et al., 2011) e suas variações bem como as técnicas de edição genômica *in situ* utilizando sistemas de edição genômica CRISPR-Cas9 (STAAHL et al., 2017) também podem ser consideradas tecnologias estabelecidas e frequentemente utilizadas em modelos murinos.

1.1.4 Genômica de camundongos

Embora muito utilizado como modelo animal, o sequenciamento completo do genoma de camundongo foi publicado somente em 2002 (MOUSE GENOME CONSORTIUM, 2002). Esse foi um marco muito comemorado, figurando como o segundo genoma de um mamífero a ser completamente sequenciado. Porém, muito antes de 2002, já era conhecido muito a respeito do genoma desses animais devido ao acúmulo de informações de sequenciamentos de genes isolados, experimentos de mutagênese, ensaios de *linkage,* mapas citogenéticos, painéis de marcadores moleculares, ensaios funcionais diversos e também descrições fenotípicas detalhadas de mutantes (GUÉNET, 2005). Em 1991, todas essas informações começaram a ser compiladas de maneira informatizada por um esforço pioneiro de pesquisadores do *The Jackson Laboratory* - JAX, que criaram a primeira versão do que ia se tornar a maior e mais completa fonte de informação sobre genômica de camundongos, o *Mouse Genome Informatics*[7] - MGI (EPPIG et al., 2015).

Sem dúvidas o sequenciamento do genoma completo do camundongo revelou importantes similaridades com o homem e ainda hoje é utilizado como referência para diversos estudos de genômica funcional, por ser um dos genomas mais estudados e com uma grande quantidade de informações funcionais associadas (BROWN et al., 2009). O genoma de camundongo, em sua primeira

---

[7] MGI – Mouse Genome Informatics: http://www.informatics.jax.org/

versão, possuía tamanho de 2,5 Gb – cerca de 14% menor que o genoma humano – a uma taxa de erro de sequenciamento estimada de 1 x 10^{-5} (MGS, 2002) - valor que não se alterou muito com a liberação da última versão, atualmente com cerca de 2,6 Gb. Em média, 85% das regiões codificantes de camundongo podem ser alinhadas com o genoma humano, ao passo que regiões não codificantes são pouco similares - 50% das sequência ou menos. Cerca de quase 90% dos genomas de camundongo e humano podem ser agrupados em regiões de considerada sintenia, revelando a conservação da organização estrutural dos cromossomos humanos e murinos em relação a um ancestral comum (GUÉNET, 2005; MGS, 2002).

Curiosamente, o genoma sequenciado em 2002 não corresponde a nenhum camundongo selvagem e sim majoritariamente ao camundongo da linhagem isogênica C57BL/6J (MGS, 2002). Já na época era suposto que o genoma de outras linhagens seria consideravelmente diferente do recente genoma sequenciado e que seria considerado como genoma referência (GUÉNET, 2005). De toda forma, o objetivo do sequenciamento do genoma de camundongo não era *a priori* investigar a diversidade genética da ampla gama de linhagens isogênicas disponíveis; isso só seria se realizar com o *Mouse Genomes Project*, a partir de 2011 (KEANE et al., 2011; YALCIN et al., 2011). Inicialmente o projeto sequenciou e analisou 17 linhagens de camundongos. Mais de 120 milhões de SNPs totais foram identificados, sendo que a linhagem C57BL/6NJ, mais próxima ao genoma referência, somou cerca de 1400 SNPs privados e a linhagem PWK/PhJ, mais distante e próxima de camundongos selvagens, somou cerca de 4 milhões de SNPs privados (KEANE et al., 2011). O sequenciamento dessas linhagens em um período tão curto só foi possível devido ao desenvolvimento das tecnologias de sequenciamento NGS, que revolucionaram a genômica como um todo e serão discutidas a seguir.

1.1.5 Sequenciamento de exomas

O surgimento das novas tecnologias de sequenciamento, chamadas comumente de sequenciamento de nova geração (NGS, do inglês *Next Generation Sequencing*), possibilitaram um salto enorme em termos de custo, rapidez e eficiência em genômica. Até 2005, a única abordagem disponível era o sequenciamento tradicional pelo método Sanger – utilizado de maneira integral nos

primeiros projetos Genoma – efetivo para o sequenciamento de segmentos de DNA pequenos mas com custo muito alto para genomas inteiros (MARDIS, 2013).

A introdução de técnicas de enriquecimento de segmentos de DNA correspondentes ao conjunto de exons aumentou muito a eficiência e os custos de sequenciamento NGS. Um exoma, por definição, é o conjunto de exons dos genes anotados em um genoma referência (WARR et al., 2015). Tipicamente são utilizadas bases de anotação confiáveis, como o CCDS (*consensus coding sequence*). O alvo-genômico a ser sequenciado, no caso de humanos e camundongos, passa de cerca de 3 bilhões de bases para algo em torno de 50 milhões de bases, ou cerca de 1,5% do genoma total do homem ou do camundongo. Uma síndrome relacionada a um tipo de perda auditiva causada pelo gene *dfn82* (WALSH et al., 2010) e uma outra síndrome relacionada a malformações do cérebro (BILGÜVAR et al., 2010) foram as duas primeiras doenças mendelianas humanas descobertas por sequenciamento de exoma, em 2010.

Atualmente, a abordagem mais utilizada para enriquecer amostras de DNA genômico com sequências anotadas como exons é a chamada de captura híbrida em solução (WARR et al., 2015). A captura em solução consiste na hibridização de uma biblioteca de DNA genômico com uma grande variedade de sondas que correspondem à maioria dos exons. As sondas podem ser biotiniladas facilitando a seleção dos fragmentos através de purificação por *beads* magnéticas (**Figura 1.2**). Existem vários métodos disponíveis que utilizam a captura em solução em exons, possuindo características singulares se considerarmos a especificidade, cobertura uniforme, maior ou menor número de duplicatas, maior taxa de alinhamento e cobertura de regiões UTR (SEKHAR et al., 2014). Quase todas as abordagens de enriquecimento também estão disponíveis para camundongos, apresentando métricas parecidas em relação às abordagens desenvolvidas para humanos (GAO et al., 2013).

Uma das principais vantagens do sequenciamento de regiões exônicas é a redução dos custos e tempo de análise seguida pela simplificação na interpretação dos resultados em relação ao sequenciamento do genoma completo. A principal desvantagem é a impossibilidade de sequenciamento de todos os exons, devido a restrições de desenho de sondas em determinadas regiões e amplificação diferencial de determinadas regiões por PCR. Além disso, temos a óbvia falta de informação de regiões não exônicas, que podem conter variantes importantes para o

desenvolvimento de um determinado fenótipo (WARR et al., 2015). Mesmo assim o sequenciamento completo de exomas é considerado uma técnica revolucionária em genômica, principalmente quando aplicada no diagnóstico e descoberta de doenças humanas (KOBOLDT et al., 2013; WARR et al., 2015).

Em camundongos, a primeira utilização do sequenciamento de exomas foi publicada logo em 2011, visando o mapeamento de mutantes induzidos por ENU (FAIRFIELD et al., 2011), baseada nos mesmos princípios utilizados em exomas humanos (**Figura 1.2**). As abordagens disponíveis para enriquecimento de exons por captura em solução em camundongos são muito similares às metodologias disponíveis para humanos.

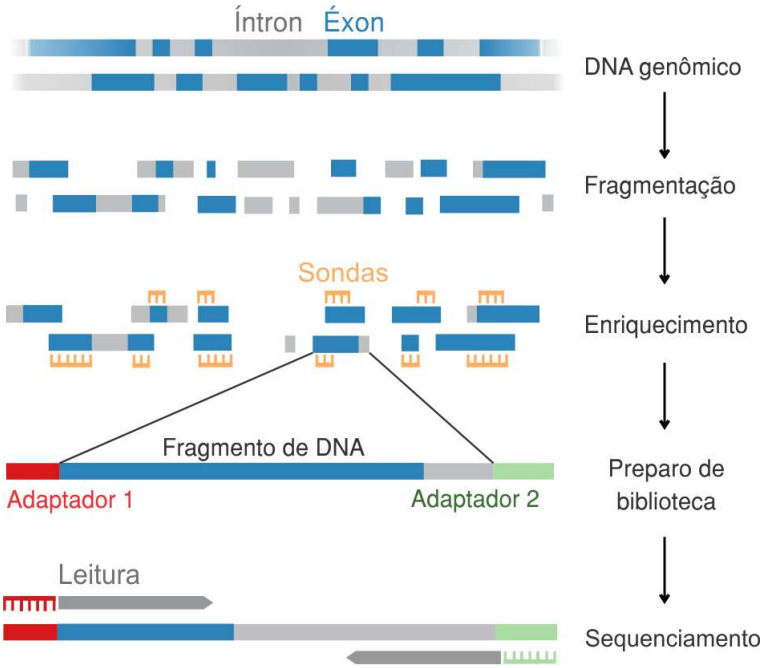

Figura 1.2 Etapas de enriquecimento de regiões exônicas por captura em solução. O DNA genômico é submetido a fragmentação, na maioria das vezes mecânica – por sonicação. Os fragmentos de DNA são então submetidos a hibridização com sondas específicas cujos alvos são regiões exônicas. A captura dos fragmentos hibridizados é feita por *beads* magnéticas. Os fragmentos de DNA enriquecidos são submetidos ao preparo das bibliotecas, que consiste basicamente na ligação de adaptadores e seleção de tamanho. As bibliotecas são então sequenciadas em plataformas NGS, em modo simples ou em modo *paired-end* onde são produzidas leituras para ambas as extremidades do fragmento. Adaptado de (DE SOUZA; IENNE, 2018).

O presente estudo, teve como principal objetivo global a aplicação e o desenvolvimento de uma metodologia capaz de identificar variantes em exomas de camundongos tanto em dois grupos distintos: um grupo composto por duas linhagens isogênicas C57BL/6 e BALB/c, mantidas pelo Biotério do ICB e um grupo composto por 7 camundongos mutantes oriundos de um estudo anterior desenvolvido por Massironi e colaboradores (MASSIRONI et al., 2006). Será

discutido nesse Capítulo o desenvolvimento da abordagem para a execução do sequenciamento do exoma dessas 9 amostras e obtenção das leituras para as análises posteriores de variantes, que serão detalhadas nos Capítulos 2 e 3.

1.2 Material e Métodos

1.2.1 Camundongos

Animais mutantes e das linhagens C57BL/6ICBI e BALB/cICBI foram obtidos do Biotério de Experimentação do Departamento de Imunologia do Instituto de Ciências Biomédicas da USP. Durante os experimentos os camundongos foram mantidos em microisoladores em ambiente com barreiras sanitárias. A temperatura foi mantida a 21±2°C com intervalos de luz e escuro de 12 h. Os animais foram alimentados com ração comercial Nuvilab (Quimtia, Curitiba, Paraná, Brasil) e água *ad libitum*. Dois animais machos de cada linhagem com idades entre 1 mês e 1 ano foram eutanasiados com excesso do anestésico Ketamina/Xylasina via intraperitoneal para remoção do baço e parte da cauda utilizados nos procedimentos de extração de DNA genômico, descrito em (**1.2.2**). O projeto, cujo título é “Análise das alterações genéticas em exomas de camundongos”, foi aprovado pela Comissão de Ética no uso de Animais (CEUA-ICB/USP) do ICB-USP em 5 de maio de 2015 com o número de registro 053/32 - livro 03.

1.2.2 Extração de DNA

Amostras da ponta da cauda e do baço de indivíduos machos de cada população mutante e das duas linhagens isogênicas C57BL/6ICBI e BALB/cICBI com aproximadamente 12 meses de vida foram submetidas à extração de DNA genômico com o *kit Genomic DNA from Tissue* (Macherel-Nagel, Düren, Alemanha) com tratamento com RNAase A e duas rodadas de eluição de 50 µL. Aproximadamente cerca de 25 mg do baço e pedaços da ponta da cauda de 0,6 cm (30 mg) foram utilizados para a extração com o kit, segundo orientações do fabricante. Foi utilizado cerca de 1 µL de DNA genômico das amostras para cada etapa do controle de qualidade: quantificação por fluorescência pelo Qubit 2.0 (Thermo Fisher Scientific, Waltham, Massachusetts, EUA) utilizando o *kit* DNA HS; avaliação por eletroforese em gel de agarose; análise espectrofotométrica de relações de absorbância 260:280 nm, 260:270 nm e 260:230 nm pelo equipamento Nanodrop (Thermo Fisher Scientific, Waltham, Massachusetts, EUA).

1.2.3 Preparo das bibliotecas, enriquecimento e PCR de emulsão

As amostras de DNA genômico da ponta da cauda sem indícios de degradação aparente em eletroforese em gel de agarose e com medidas razoáveis de relações de absorbância (razão 260:280 nm = 1,8 a 2,0; razão 260:270 nm = 1,2; razão 260:230 nm = 1,9) foram utilizadas para o enriquecimento dos exons e preparo das bibliotecas. O kit utilizado para o enriquecimento e preparo das bibliotecas foi o kit SureSelect Mouse All-Exon (Agilent Technologies, Santa Clara, California, EUA), com sondas derivadas de sequências de exons do banco RefSeq e Ensembl (conjunto de sondas S0276129) para o genoma NCBI37/mm9. As amostras de DNA genômico foram diluídas com tampão *Low EDTA TE* (Thermo Fisher) para 4 µg em um total de 130 µL e utilizadas para o preparo das bibliotecas, de acordo com instruções contidas no protocolo padrão do kit SureSelect^{XT} Target Enrichment Kits for AB SOLiD Multiplexed Sequencing (Agilent Technologies). A fragmentação mecânica por sonicação foi realizada no equipamento Covaris S2 (Covaris, Massachusetts, EUA) seguida pelo reparo de pontas, ligação dos adaptadores, modificação das extremidades do adaptador e captura híbrida em solução seguida por amplificação por PCR. As etapas de purificação de DNA foram realizadas utilizando o sistema de purificação de fase sólida de *beads* paramagnéticas Agencourt AMPure XP (Beckman Coulter, Pasadena, California, EUA). Um total de 12 *barcodes* (índices) foram utilizados para identificar 9 amostras diferentes, correspondentes a 7 mutantes e duas linhagens isogênicas. Visando aumentar a variabilidade e a acurácia na detecção e identificação dos *barcodes*, as amostras C57BL/6ICBI, BALB/cICBI e o mutante *fraqueza* (selecionado de maneira aleatória) foram duplicadas com *barcodes* diferentes, totalizando 12 bibliotecas e 12 *barcodes* diferentes.

As 12 bibliotecas foram quantificadas por fluorescência pelo Qubit 2.0 (Thermo Fisher) utilizando o kit de quantificação DNA HS e avaliadas quanto a distribuição dos fragmentos de DNA por eletroforese microfluídica pelo Bioanalyzer (Agilent Technologies) usando o chip DNA 1000. Após a quantificação e análise de distribuição de fragmentos, as bibliotecas foram então misturadas de forma eqüimolar em um *pool* totalizando 48,5 ng de DNA. As reações de PCR de emulsão foram realizadas utilizando kits E120 e kits E80 através das plataformas de automação EZBeads Emulsifier, Amplifier e Enricher (Thermo Fisher) de acordo com as instruções padrão para a produção de *beads* por PCR em emulsão e

enriquecimento de *beads* com *templates* amplificados. A deposição das *beads* nas *lanes* da plataforma SOLiD 5500XL (Thermo Fisher) foram quantificadas de acordo com as instruções do fabricante, descritas a seguir.

1.2.4 Sequenciamento NGS

Kits para preparo da PCR em emulsão e enriquecimento E120 e E80, que diferem quanto a quantidade de *beads* geradas, foram utilizados com as plataformas EZBeads para a preparação das *beads* para sequenciamento e quantificadas por comparação colorimétrica. As *beads* foram modificadas covalentemente para deposição nas *flowcell*s visando o limite de 250.000 *beads* por painel da *flowcell.* Primeiramente foi realizada uma corrida-teste chamada de *WFA* (*Workflow Analysis)* em que apenas 1 milhão de *beads* são depositadas e sequenciadas, em uma única lane, de modo a quantificar de forma mais precisa as *beads* P2, proporção de *beads* monoclonais e policlonais, dentre outros parâmetros.

As *beads* produzidas através de kits de PCR em emulsão foram avaliadas por WFA e depositadas com uma quantidade-alvo de 330 milhões por *lane.* Foram utilizadas três *flowcells* em três diferentes corridas na plataforma SOLiD 5500xl, (Thermo Fisher) em modo fragmento de 1x75 pb, cujos sequenciamentos foram realizados no Centro de Facilidades de Apoio à Pesquisa (CEFAP-USP), totalizando 11 *lanes* sequenciadas.

*1.2.5 Identificação e divisão (*demultiplexing*) das leituras*

Cada *lane* sequenciada resulta em um único arquivo XSQ, que armazena todos os tipos de leituras (*tags*) sequenciadas para cada fragmento detectado em uma *bead*. Cada *bead* válida gera uma leitura de 75 pb e uma leitura de 5 pb correspondente à *tag* do *barcode*. O formato também inclui um cabeçalho de informações relativas à corrida, como identificação das amostras e seus respectivos *barcodes*. Dessa forma, as leituras em cada uma das *lanes* sequenciadas foram divididas (ou separadas) de acordo com as leituras correspondentes aos *barcodes* das amostras pelo *script* convertFromXSQ.sh disponibilizado pela ferramenta XSQTools (Thermo Fisher).

A qualidade e quantidade de leituras produzidas foram avaliadas através dos relatórios de corridas gerados pelo ICS (Instrument Control Software for SOLiD Next-Generation Sequencing, Thermo Fisher).

1.3 Resultados

O sequenciamento das 11 *lanes* gerou um total de 947.951.581 leituras de 1x75 pb que correspondem a aproximadamente 71 Gb (71 x 10^6 pb) após a análise primária de *demultiplexing* (**Tabela 1.1**). Isso equivale a aproximadamente 105 milhões de leituras e 7,9 x 10^6 bases ± 4,4 por amostra. A capacidade de produção de leituras da plataforma SOLiD 5500xl, supondo níveis aceitáveis de deposição, qualidade de *beads* P2 e qualidade de detecção é de aproximadamente 100 milhões de leituras por *lane*. Sendo assim, o máximo ideal de quantidade de dados gerados em 11 *lanes* é de 1,1 bilhão de leituras no modo 1x75 bp.

Tabela 1.1 - Número de leituras de 1x75 pb e número de bases sequenciadas

.

Camundongo	Número de leituras	Bases (10^6)
ataxico-1	55471554	4.16
batepalmas	84590152	6.34
careca	55471554	4.16
cruzapernas	64304657	4.82
equilibrio	62364613	4.68
fraqueza	217138559	16.29
Sacudidor	94192080	7.06
C57BL/6ICBI	140456591	10.53
BALB/cICBI	173961821	13.05
Total	947951581	71.10

A quantidade de dados gerados depende de inúmeros fatores como acurácia na quantificação do DNA e das *beads*, densidade e eficiência de deposição, baixa proporção de *beads* policlonais e qualidade das bibliotecas, sendo uma questão multifatorial basicamente dependente de tentativas empíricas de otimização. Em média, o uso da plataforma atingiu cerca de 86,2% da capacidade total de sequenciamento, sendo que nos dois últimos sequenciamentos (total de 7 *lanes*) essa eficiência foi próxima de 100%, valores superiores aos alcançados por outros usuários da plataforma no CEFAP-USP no período de funcionamento da plataforma.

1.4 Discussão

O uso do sequenciamento de exomas em camundongos têm aumentado bastante nos últimos anos (CARUANA et al., 2013; FAIRFIELD et al., 2015; TANISAWA et al., 2013; WEI et al., 2017), visando principalmente a análise rápida de mutações candidatas em camundongos mutantes, de forma a diminuir os esforços de validação (SIMON et al., 2015).

A abordagem utilizada foi fortemente inspirada em um dos primeiros esforços para descoberta de mutações em camundongos mutantes utilizando o sequenciamento de exomas, descrito por (FAIRFIELD et al., 2011). Embora esse estudo tenha utilizado tecidos provenientes do baço (DERDAK et al., 2015; FAIRFIELD et al., 2011) ou fígado (MASUMURA et al., 2016; TANISAWA et al., 2013), demonstramos que a extração de tecido da ponta da cauda – muito menos trabalhoso e prático – também é suficiente para a obtenção de DNA genômico com qualidade para o preparo de bibliotecas de sequenciamento (**1.3.1**). De fato sequenciamentos recentes de exoma de camundongos têm utilizado pedaços da ponta de cauda, como o sequenciamento pioneiro de camundongos *knockin* CRISPR/Cas (NAKAJIMA et al., 2016).

Em relação ao preparo das bibliotecas, escolhemos o kit SureSelect para o preparo das bibliotecas em substituição ao kit Nimblegen, utilizado por (FAIRFIELD et al., 2011, 2015). Essa escolha foi feita principalmente pelos fatores de alta cobertura de regiões alvo, baixo número de duplicatas geradas pelo SureSelect e custo por amostra em grupos de 12 ou menos amostras (ASAN et al., 2011; CLARK et al., 2011). Apesar da maior dificuldade no preparo devido à manipulação e hibridização com sondas de RNA – exclusivas dos kits SureSelect – as bibliotecas atingiram todos os controles de qualidade (**6.2.2**). Para a análise de distribuição dos fragmentos das bibliotecas foi utilizado o kit de eletroforese microfluídica Bioanalyzer DNA 1000, cuja sensibilidade é menor, cerca de no mínimo 0,5 ng/µL, mas com um custo por amostra cerca de 2 a 3 vezes menor em relação ao kit sugerido DNA HS. A distribuição dos fragmentos ideal para a emulsão de PCR e sequenciamento é de, em média, ente 250 e 300 pb e todas as bibliotecas apresentaram tamanho médio dentro dessa variação (**6.2.2**).

Com relação ao sequenciamento em si optamos por uma estratégia de sequenciamento mais cuidadosa, controlada e econômica, que consistiu no uso de 11 *lanes* em um total de 3 *flowcells* diferentes em três corridas diferentes – uma

nova corrida era realizada somente após a análise do relatório e ajuste da quantidade de *beads* a serem depositadas. Essa estratégia se mostrou muito efetiva, visto que foi possível reajustar a quantidade de *beads* depositadas sem afetar o custo total de reagentes ou a qualidade das bases (**6.2.3**) a custo 25% menor – demonstrando as vantagens do compartilhamento de corridas entre projetos. Um outro ponto muito importante é a incompatibilidade do kit SureSelect Mouse All-Exon kit com os adaptadores reversos de sequenciamento da plataforma SOLiD, impossibilitando abordagens do tipo *paired-end.* Dessa forma, era importante maximizar a produção de leituras de forma a obter pelo menos uma cobertura efetiva de 30X do exoma – estimando 20% de perdas devido a leituras não mapeadas e filtragens de qualidade. Isso corresponderia a aproximadamente 36 milhões de leituras de 75 pb ou 2,7 Gb por amostra. Como demonstrado em (**6.2.3** e **Tabela 1.1**), todas as amostras foram sequenciadas acima desse limite – a menor quantidade de *reads*, praticamente o dobro do projetado, correspondeu a amostra Atáxico-1. Em média, a quantidade de dados gerados por amostra foi de 7,9 Gb, que corresponde a mais de 130X de cobertura estimada do exoma. Em termos comparativos, no estudo pioneiro de (FAIRFIELD et al., 2011) sobre o sequenciamento de camundongos mutantes, a maior quantidade de dados produzidos por amostra foi de 6,8 Gb enquanto no estudo de (TANISAWA et al., 2013) a amostra com mais quantidade de dados se limitou a cerca de 3 Gb, demonstrando o sucesso da abordagem de sequenciamento utilizada (**Tabela 1.1**).

O custo total em reagentes do sequenciamento das 9 amostras - incluindo extração de DNA, preparo das bibliotecas, PCR em emulsão, uso da plataforma SOLiD 5500XL – foi de aproximadamente US$ 1770,00 por amostra com no mínimo 2,7 Gb de dados produzidos. Esse custo reduzido, comparável inclusive com valores de serviços de sequenciamento atuais[8], foi possível devido ao compartilhamento de leituras e ao ajuste da quantidade de amostras a serem sequenciadas dentro do limite de uso dos kits.

Finalmente, demonstramos que foi possível obter com sucesso dados de sequenciamento com qualidade e custo razoável, segundo a proposta inicial. Nos próximos capítulos (**Capítulo 2** e **Capítulo 3**), serão discutidos o desenvolvimento

---

[8] Australian Phenomics Facility: http://www.apf.edu.au/products-pricing

das estratégias de análise dos dados obtidos – das estratégias de alinhamento e detecção de variantes – que foram aplicadas de maneira diferentes nas amostras de linhagens isogênicas (**Capítulo 2**) e para os camundongos mutantes (**Capítulo 3**).

1.4 Referências

1000GENOMES. A global reference for human genetic variation. **Nature**, v. 526, p. 68–74, 2015.

AMBERGER, J. S. et al. OMIM . org : Online Mendelian Inheritance in Man (OMIM R), an online catalog of human genes and genetic disorders. **Nucleic Acids Research**, v. 43, n. November 2014, p. 789–798, 2015.

ANTONARAKIS, S. E.; BECKMANN, J. S. Focus on Monogenic Disorders. **Nature reviews. Genetics**, v. 7, n. April, p. 277–282, 2006.

ASAN et al. Comprehensive comparison of three commercial human whole-exome capture platforms. **Genome biology**, v. 12, n. 9, p. R95, jan. 2011.

BARBA, M.; CZOSNEK, H.; HADIDI, A. Historical perspective, development and applications of next-generation sequencing in plant virology. **Viruses**, v. 6, n. 1, p. 106–136, 2013.

BATESON, W.; MENDEL, G. **Mendel's principles of heredity**. Cambridge: Cambridge: University Press, 1909, 1909.

BILGÜVAR, K. et al. Whole-exome sequencing identifies recessive WDR62 mutations in severe brain malformations. **Nature**, v. 467, n. 7312, p. 207–210, 2010.

BOGUE, M. A.; CHURCHILL, G. A.; CHESLER, E. J. Collaborative Cross and Diversity Outbred data resources in the Mouse Phenome Database. **Mammalian genome : official journal of the International Mammalian Genome Society**, v. 26, n. 9, p. 511–520, 2015.

BOUABE, H.; OKKENHAUG, K. Europe PMC Funders Group Gene Targeting in Mice : a Review. **Methods in Molecular Biology**, p. 315–336, 2015.

BROWN, S. D. M. et al. The functional annotation of mammalian genomes: the challenge of phenotyping. **Annual review of genetics**, v. 43, p. 305–33, 2009.

BUSH, W. S.; MOORE, J. H. Chapter 11 : Genome-Wide Association Studies. **PLoS Biology**, v. 8, n. 12, 2012.

CARUANA, G. et al. Genome-wide ENU mutagenesis in combination with high density SNP analysis and exome sequencing provides rapid identification of novel mouse models of developmental disease. **PloS one**, v. 8, n. 3, p. e55429, jan. 2013.

CHONG, J. X. et al. The Genetic Basis of Mendelian Phenotypes : Discoveries , Challenges , and Opportunities. **The American Journal of Human Genetics**, p. 199–215, 2015.

CHURCH, D. M. et al. Lineage-Specific Biology Revealed by a Finished Genome Assembly of the Mouse. **PLoS Biology**, v. 7, n. 5, p. e1000112, 2009.

CLARK, M. J. et al. Performance comparison of exome DNA sequencing technologies. **Nature biotechnology**, v. 29, n. 10, p. 908–14, out. 2011.

CROLLIUS, H. R. et al. Estimate of human gene number provided by genome- wide analysis using Tetraodon nigroviridis DNA sequence. v. 25, n. june, p. 235–238, 2000.

DE SOUZA, T. A.; IENNE, S. **Capítulo 11: Sequenciamento de DNA**. In: Biologia Molecular, Editor: Nancy Rebouças, submetido. Sao Paulo: Editora Atheneu, 2018.

DERDAK, S. et al. Genomic characterization of mutant laboratory mouse strains by exome sequencing and annotation lift-over. **BMC genomics**, v. 16, p. 351, 2015.

ENCODE. An integrated encyclopedia of DNA elements in the human genome. **Nature**, v. 489, p. 57–74, 2012.

EPPIG, J. T. et al. Mouse Genome Informatics (MGI): reflecting on 25 years. **Mammalian Genome**, v. 26, n. 7, p. 272–284, 2015.

ERMANN, J.; GLIMCHER, L. H. After GWAS: mice to the rescue? **Current opinion in immunology**, v. 24, n. 5, p. 564–70, out. 2012.

FAIRFIELD, H. et al. Mutation discovery in mice by whole exome sequencing. **Genome biology**, v. 12, n. 9, p. R86, 2011.

FAIRFIELD, H. et al. Exome sequencing reveals pathogenic mutations in 91 strains of mice with Mendelian disorders. **Genome Research**, v. 25, p. 948–957, 2015.

GAO, Q. et al. A systematic evaluation of hybridization-based mouse exome capture system. **BMC genomics**, v. 14, n. 1, p. 492, 21 jul. 2013.

GARROD, A. E. The Incidence of Alkaptonuria: A Study in Chemical Individuality. **Lancet**, v. ii, p. 1616–1620, 1902.

GHOSH, S.; BOUCHARD, C. Convergence between biological, behavioural and genetic determinants of obesity. **Nature Publishing Group**, 2017.

GUÉNET, J. et al. **Genetics of the Mouse**. Berlin Heidelberg: Springer-Verlag Berlin Heidelberg, 2015.

GUÉNET, J. L. The mouse genome. **Genome Research**, v. 15, p. 1729–1740, 2005.

HAMAMY, H. et al. Consanguineous marriages , pearls and perils : Geneva. **Genetics in Medicine**, v. 13, n. 9, 2011.

HAMAMY, H. Consanguineous marriages Preconception consultation in primary health care settings. **Journal of Community Genetics**, v. 3, p. 185–192, 2012.

HEARD, E. et al. Ten years of genetics and genomics: what have we achieved and where are we heading? **Nature Publishing Group**, v. 11, n. 10, p. 723–733, 2010.

JUSTICE, M. J.; DHILLON, P. Using the mouse to model human disease : increasing validity and reproducibility. **Disease Models & Mechanisms**, v. 9, p. 101–103, 2016.

KEANE, T. M. et al. Mouse genomic variation and its effect on phenotypes and gene regulation. **Nature**, v. 477, n. 7364, p. 289–294, 2011.

KELLIS, M. et al. Defining functional DNA elements in the human genome. **PNAS**, v. 111, n. 17, p. 6131–6138, 2014.

KOBOLDT, D. C. et al. Review The Next-Generation Sequencing Revolution and Its Impact on Genomics. **Cell**, v. 155, n. 1, p. 27–38, 2013.

KURE, E.; ANTONIO, F. A molecular pathway analysis stresses the role of inflammation and oxidative stress towards cognition in schizophrenia. **Journal of Neural Transmission**, v. 124, n. 7, p. 765–774, 2017.

LEVY, S. et al. The Diploid Genome Sequence of an Individual Human. **PLoS Biology**, v. 5, n. 10, p. 2113–2144, 2007.

LYNCH, M. Evolution of the mutation rate. **Trends in Genetics**, v. 26, n. 8, p. 345–352, 2010.

MANDILLO, S. et al. Reliability , robustness , and reproducibility in mouse behavioral phenotyping : a cross-laboratory study. **Physiological Genetics**, v. 34, p. 243–255, 2008.

MARDIS, E. R. Next-Generation DNA Sequencing Methods. **Annual Reviews Genomics and Human Genetics**, v. 9, p. 387–402, 2008.

MARDIS, E. R. Next-Generation Sequencing Platforms. **Annual Reviews Genomics and Human Genetics**, v. 6, p. 287–303, 2013.

MARDIS, E. R. DNA sequencing technologies: 2006–2016. **Nature Protocols**, v. 12, n. 2, p. 213–218, 2017.

MASSIRONI, S. M. G. et al. Inducing mutations in the mouse genome with the chemical mutagen ethylnitrosourea. **Brazilian journal of medical and biological research**, v. 39, n. 9, p. 1217–26, set. 2006.

MASUMURA, K. et al. Dose-dependent de novo germline mutations detected by whole-exome sequencing in progeny of ENU-treated male gpt delta mice. **Mutation Research - Genetic Toxicology and Environmental Mutagenesis**, v. 810, p. 30–39, 2016.

MGS. Initial sequencing and comparative analysis of the mouse genome. **Nature**, v. 420, n. December, p. 520–562, 2002.

MORESCO, E. M. Y.; LI, X.; BEUTLER, B. Going forward with genetics: recent technological advances and forward genetics in mice. **The American journal of pathology**, v. 182, n. 5, p. 1462–73, maio 2013.

MUNFORD, V. et al. A genetic cluster of patients with variant xeroderma pigmentosum with two different founder mutations *. **British Journal of Dermatology**, v. 176, p. 1270–1278, 2017.

NAKAJIMA, K. et al. Exome sequencing in the knockin mice generated using the CRISPR/Cas system. **Scientific Reports**, v. 6, n. 1, p. 34703, 2016.

PASANIUC, B. et al. Extremely low-coverage sequencing and imputation increases power for genome-wide association studies. **Nature genetics**, v. 44, n. 6, p. 631–5, jun. 2012.

PGH. Initial sequencing and analysis of the human genome. **Nature**, v. 409, n. February, p. 860–921, 2001.

POPEJOY, A. B.; FULLERTON, S. M. Genomics is falling on diversity. **Nature**, v. 538, n. 161, p. 164, 2016.

PREMSRIRUT, P. K. et al. Resource A Rapid and Scalable System for Studying Gene Function in Mice Using Conditional RNA Interference. **Cell**, v. 145, n. 1, p. 145–158, 2011.

RICHTER, S. H. et al. Effect of Population Heterogenization on the Reproducibility of Mouse Behavior : A Multi-Laboratory Study. **PLoS ONE**, v. 6, 2011.

RIEBER, N. et al. Coverage bias and sensitivity of variant calling for four whole-genome sequencing technologies. **PloS one**, v. 8, n. 6, p. e66621, jan. 2013.

SCHAUB, M. A. et al. Linking disease associations with regulatory information in the human genome. **Genome Research**, v. 22, p. 1748–1759, 2012.

SEKHAR, C. et al. Performance comparison of four exome capture systems for deep sequencing Performance comparison of four exome capture systems for deep sequencing. **BMC genomics**, v. 15, n. 449, 2014.

SIMON, M. M. et al. Current strategies for mutation detection in phenotype-driven screens utilising next generation sequencing. **Mammalian Genome**, v. 26, n. 9, p. 486–500, 2015.

STAAHL, B. T. et al. Efficient genome editing in the mouse brain by local delivery of engineered Cas9 ribonucleoprotein complexes. **Nature Publishing Group**, n. August 2016, 2017.

SUD, A. et al. Risk of Second Cancer in Hodgkin Lymphoma Survivors and Influence of Family History. **Journal of Clinical Oncology**, v. 35, n. 14, p. 1584–1591, 2017.

TANISAWA, K. et al. Exome sequencing of senescence-accelerated mice (SAM)

reveals deleterious mutations in degenerative disease-causing genes. **BMC genomics**, v. 14, p. 248, jan. 2013.

VENTER, J. C. et al. The Sequence of the Human Genome. **Science**, v. 291, n. February, 2001.

VISSCHER, P. M. et al. Five Years of GWAS Discovery. **The American Journal of Human Genetics**, v. 90, n. 1, p. 7–24, 2012.

WALSH, T. et al. Whole exome sequencing and homozygosity mapping identify mutation in the cell polarity protein GPSM2 as the cause of nonsyndromic hearing loss DFNB82. **American Journal of Human Genetics**, v. 87, n. 1, p. 90–94, 2010.

WARR, A. et al. Exome Sequencing : Current and Future Perspectives. **G3 (Bethesda, Md.)**, v. 5, n. August, p. 1543–1550, 2015.

WEI, L. et al. Exome sequencing analysis of murine medulloblastoma models identifies WDR11 as a potential tumor suppressor in Group 3 tumors. **Oncotarget**, v. 8, n. 39, p. 64685–64697, 2017.

WEYDEN, L. VAN DER et al. The mouse genetics toolkit : revealing function and mechanism. **Genome biology**, v. 12, 2011.

YALCIN, B. et al. Sequence-based characterization of structural variation in the mouse genome. **Nature**, v. 477, n. 7364, p. 326–329, 2011.

YANG, H. et al. Subspecific origin and haplotype diversity in the laboratory mouse. **Nature Genetics**, v. 43, n. 7, p. 648–655, 2011.

ZHANG, F.; LUPSKI, J. R. Non-coding genetic variants in human disease. **Human Molecular Genetics**, v. 24, n. July, p. 102–110, 2015.

CAPÍTULO 2 – CARACTERIZAÇÃO DO EXOMA DOS CAMUNDONGOS ISOGÊNICOS C57BL/6-ICBI e BALB/c-ICBI

CAPÍTULO 2 – CARACTERIZAÇÃO DO EXOMA DOS CAMUNDONGOS ISOGÊNICOS C57BL/6-ICBI e BALB/c-ICBI

2.1 Introdução

2.1.1 Origem das linhagens isogênicas de camundongos

Camundongos isogênicos são ferramentas básicas em múltiplas áreas da ciência, atuando como modelos tanto para a pesquisa básica quanto em pesquisa aplicada. Considera-se uma linhagem isogênica os indivíduos gerados após a 20ª geração de acasalamentos seguidos irmão-irmã. Estima-se que nessa geração, os camundongos irmãos possuam 98,6% de todos os seus *loci* em homozigose (BECK et al., 2000; GREEN, 1966). A grande maioria das linhagens isogênicas utilizadas hoje são muito relacionadas em termos de genealogia entre si e foram originadas no início do século XX por pesquisadores do nordeste dos EUA. De fato, evidências sobre a uniformidade do DNA mitocondrial de 16 linhagens isogênicas sugerem que grande parte das linhagens isogênicas foi originada de uma mesma fêmea ancestral há cerca de 150 ou 200 anos atrás (GOIOS et al., 2007).

As linhagens isogênicas de camundongos têm sua história ligada direta ou indiretamente aos trabalhos dos pesquisadores americanos Clarence C. Little, Halsey J. Bagg, Leonell C. Strong, Miss Lathrop e E. Carleton MacDowell no início do século XX (GREEN, 1966). Little tinha como objetivo reduzir o efeito da variabilidade genética no estudo de doenças neoplásicas em camundongos e usou acasalamentos irmão-irmã em sequência para criar uma população que apresentaria a menor variação possível, criando assim os princípios básicos da primeira linhagem isogênica. Os camundongos usados por Little tinham características peculiares de pelagem, chamadas de *dilution, brown* e *nonagouti*. Essa linhagem isogênica criada por Little em 1909 ficou conhecida por DBA a partir de 1950 e ainda é muito utilizada em estudos sobre o câncer (GREEN, 1966; GUÉNET et al., 2015).

Em 1913, um outro pesquisador de Nova York, chamado Halsey Bagg, obteve camundongos albinos de um criador no estado de Ohio e os manteve em uma pequena colônia isolada, também com o intuito de estudar o desenvolvimento de neoplasias em camundongos. Seu colega Leonell C. Strong utilizou um camundongo da criação de Bagg e um camundongo também albino que Little mantinha em sua criação para o estudo de envelhecimento e câncer. Desse cruzamento a linhagem A começou a ser estabelecida, dando origem às linhagens hoje conhecidas como A/J e A/He. Strong realizou uma série de cruzamentos dos camundongos de Bagg e a

linhagem DBA, em 1920, dando origem às linhagens C3H, CBA, C, CHI e CI2I (GREEN, 1966). George D. Snell, geneticista de Harvard, recebeu também alguns camundongos albinos de Bagg e utilizou a nomenclatura BALB/c para identificar os *Bagg alb* e a letra "c" para indicar a coloração branca dos mesmos (POTTER; MORRIS, 1985). Em 1980, Snell seria agraciado com o prêmio Nobel em medicina por suas descobertas do complexo H-2 em camundongos, sistema análogo ao complexo de histocompatibilidade MHC em humanos (SNELL, 1980).

Estima-se hoje, que as linhagens isogênicas denominadas de *standard,* como a C57BL/6J e DBA/2J derivem de contribuições de três subespécies de camundongos diferentes, chamadas de *Mus musculus musculus, Mus m. domesticus e Mus m. castaneus* (**Figura 2.1**). Linhagens isogênicas derivadas de uma subespécie única são conhecidas como *wild-derived* e constituem um *pool* genético distinto das linhagens *standard*. São exemplos de linhagens *wild-derived:* CAST/EiJ, originada a partir de *Mus m. castaneus;* PWK/PhJ, originada de *Mus m. musculus* e SPRET/EiJ, originada da subespécie *Mus m. spretus* (BECK et al., 2000).

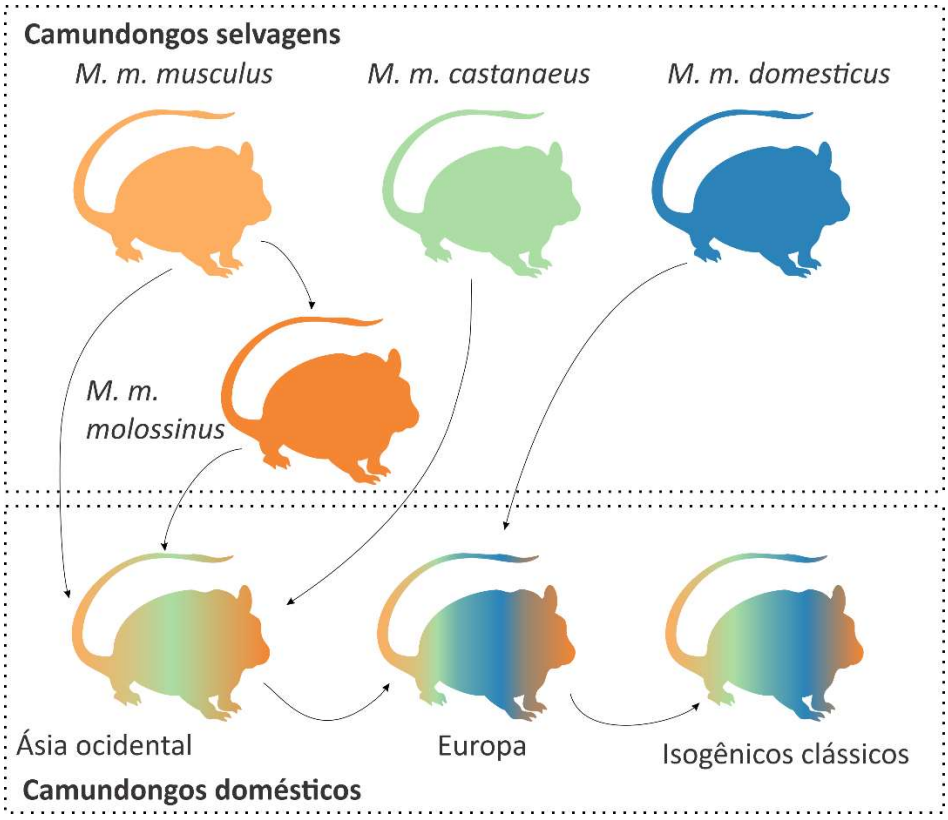

Figura 2.1 Origem dos camundongos isogênicos clássicos. A grande maioria dos camundongos isogênicos de laboratório são derivados de cruzamentos entre três subespécies principais de camundongos (*musculus, castanaeus* e *domesticus*) e também com cruzamentos com camundongos japoneses *molossinus*. Apesar de começarem a ser criados nos EUA, os camundongos foram trazidos da Europa após sua domesticação. Adaptado de (FRAZER et al., 2007).

Duas linhagens isogênicas *standard* estão entre as mais utilizadas no mundo: a linhagem C57BL/6, conhecida como *black* ou C57 e a linhagem BALB/c, conhecida como *albino* ou BALB. A seguir serão dados mais detalhes sobre a origem e características dessas duas linhagens e as sublinhagens derivadas, principalmente sobre os camundongos mantidos no Biotério do Departamento de Imunologia do Instituto de Ciências Biomédicas da USP, alvos desse estudo.

2.1.2 Origens e características da linhagem C57BL/6

Em 1921, nos EUA, Little utilizou camundongos da criação de Miss Lathrop em cruzamentos gerando uma linhagem de pelagem escura, que deu origem às atuais C57BR e C57BL (GREEN, 1966). As sublinhagens 6 e 10 foram separadas antes de 1937, dando origem às atuais C57BL/6 e C57BL/10. Atualmente sabe-se que 6 e 10 se diferenciam pelo menos nos loci *H9*, *Igh2* e *Lv* (MEKADA et al., 2009).

Camundongos C57BL/6 possuem características clássicas que os distinguem de outras sublinhagens C57BL. Eles se reproduzem de maneira rápida (HANSEN, 1973), porém são menos dóceis que outras linhagens (THOMPSON, 1953). Possuem mais atividade locomotora (DAVIS; KING; BABBINI, 1967) e odores (WYSOCKI; WHITNEY; TUCKER, 1977) e preferência aumentada por álcool (FULLER, 1964; LE et al., 1994) e opiáceos, como a morfina (BELKNAP et al., 1993) em relação a outras linhagens. Também possuem uma baixa incidência espontânea de tumores (HOAG, 1963) e são menos responsivos a diversos tipos de infecção por Toxoplasma (MACARIO; STAHL; MILLER, 1980), Leishmania (MONROY-OSTRIA et al., 1994) e Trypanosoma (ROWLAND; LOZYKOWSKI; MCCORMICK, 1992).

Pelo menos nove sublinhagens C57BL/6 foram estabelecidas até 1970 (FESTING, 1996). Duas dessas linhagens tornaram-se as mais famosas e mais utilizadas: a sublinhagem C57BL/6J proveniente do laboratório *The Jackson Laboratory* – JAX e sublinhagem C57BL/6N, mantida pelo *National Institutes of Health* – NIH (ALTMAN; KATZ, 1979; BAILEY, 1978). Em 1948, os camundongos C57BL/6 mantidos na JAX - oriundos da criação original de Little nos anos 1930 - foram nomeados como C57BL/6J (ALTMAN; KATZ, 1979).

Em 1951, na geração F32 a linhagem foi enviada para o NIH e mantida por décadas em uma colônia isolada, resultando em uma sublinhagem separada, nomeada como C57BL/6N. Animais vivos foram enviados do NIH para empresas como o Laboratório Charles River em 1974 (C5BL/6NCrl) e Harlan (C57BL/Hsd). Em

1984, o laboratório JAX[9] recebeu embriões congelados da geração F126 e os disponibilizou como C57BL/6NJ. Em 1991 animais da geração F151 também foram enviados do NIH para a Taconic[10] (C57BL/6NTac). Embora tenham origens diferentes e tenham sido separados por múltiplas gerações, não foram detectadas diferenças em painéis de genotipagem entre as sublinhagens C57BL/6N (ZURITA et al., 2011). A **Figura 2.2** detalha as relações entre as sublinhagens C57BL/6 mais utilizadas e fornecidas pelas maiores empresas. Todos os fornecedores, no entanto, aconselham que sejam adquiridos animais da mesma fonte, evitando que variantes genéticas não detectadas possam influenciar os experimentos.

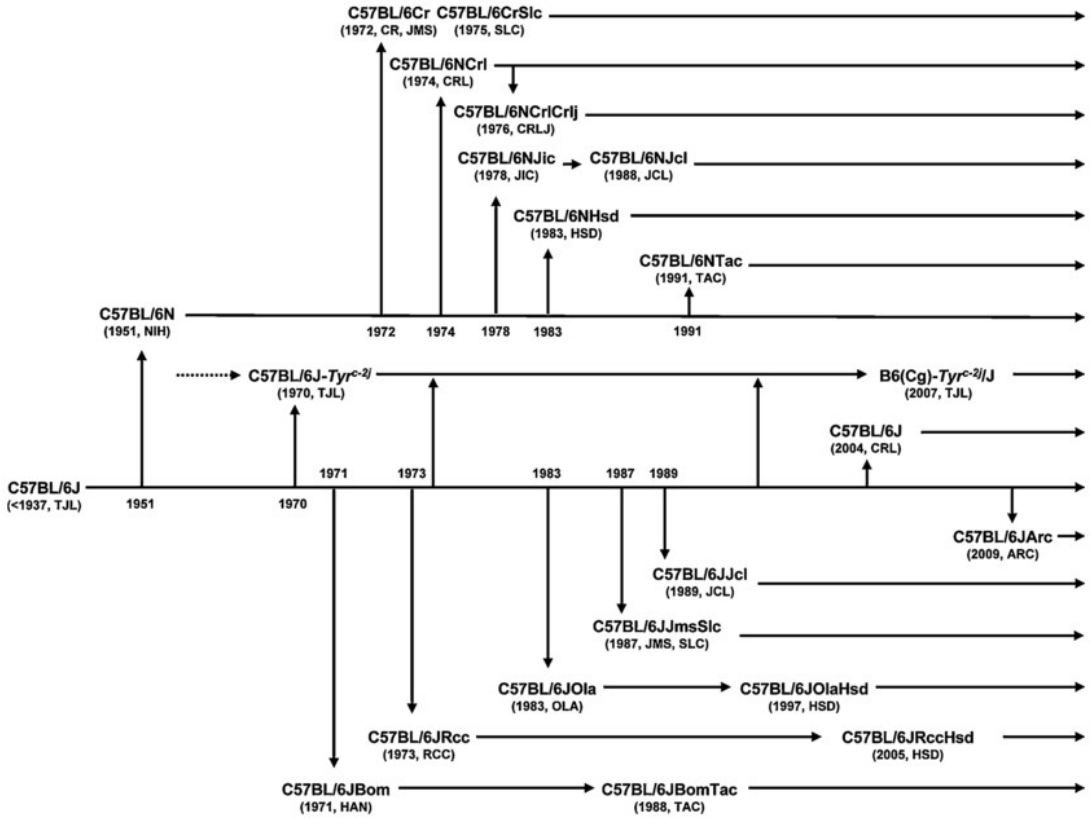

Figura 2.2 Relações entre as sublinhagens mais comuns do camundongo C57BL/6. A genealogia acima, desenhada fora da escala temporal, foi retirada de (ZURITA et al., 2011) com informações de (MEKADA et al., 2009). Fonte: (ZURITA et al., 2011).

Os camundongos *gold standard* C57BL/6J também podem ser adquiridos de diferentes fontes, mas ainda continuam disponíveis para aquisição direta pela JAX.

---

[9] JAX strain C57BL/6J, 27/10/2017: https://www.jax.org/strain/000664
[10] Taconic strain C57B/6NTac, 27/10/2017: https://www.taconic.com/mouse-model/black-6-b6ntac

O laboratório Charles River[11] fornece camundongos para a Ásia e Europa, mas reintroduz frequentemente camundongos da colônia da JAX para evitar efeitos de deriva genética (FONTAINE; DAVIS, 2016). Porém outras instituições mantém camundongos 6J em colônias separadas por muito tempo. Os laboratórios ENVIGO[12] oferecem duas sublinhagens 6J – C57BL/6OlaHsd e C57BL/6JRccHsd, adquiridas em 1973-1974, que não possuem a mutação *nnt* encontrada nas outras sublinhagens 6J. Camundongos enviados para o Japão em 1987 constituíram as linhagens C57BL/6JJcl e C57BL/6JJmsSlc não possuem a mutação no gene *nnt* (MEKADA et al., 2009; SIMON et al., 2013) surgindo que a mutação tenha ocorrido entre 1974 e 1987 (**Figura 2.2**). Camundongos comercializados na França como C57BL/6JRj desde 1993 possuem diferenças genéticas conhecidas embora possuam a mutação *nnt* (KERN et al., 2012).

2.1.3 Diferenças genotípicas e fenotípicas das sublinhagens C57BL/6J e C57BL/6N

Uma das diferenças genotípicas mais conhecidas foi reportada em 2005 na sublinhagem C57BL/6J, que possui uma deleção nos exons 7-11 do gene que codifica para a nicotinamida nucleotídeo transferase (*nnt*), intacto nas linhagens C57BL/6N. Essa mutação espontânea que implica na ausência da proteína NNT explica diferenças no controle da secreção de insulina e homeostase de glicose nos camundongos C57BL/6J (TOYE et al., 2005) e também anormalidades na biologia redox em mitocôndrias (RONCHI et al., 2013). A sublinhagem C57BL/6N também é a mais utilizada na produção de células-tronco embrionárias em camundongos, utilizados para nocautes e manipulação gênica (PETTITT et al., 2009). As sublinhagens 6N e 6J podem ser distinguidas através de vários SNPs em painéis de genotipagem disponíveis, consistindo em uma maneira confiável de controle genético de distinção entre sublinhagens *black* (MEKADA et al., 2009; ZURITA et al., 2011).

Em 2013 um extenso estudo de comparação entre as sublinhagens 6N e 6J validaram 34 SNPs em regiões codificantes, 2 *small indels* em regiões codificantes e 43 variantes estruturais, incluindo a mutação *nnt* (SIMON et al., 2013). Vários genes

---

[11] Charles River C57BL/6N, 27/10/2017: http://www.criver.com/files/pdfs/rms/c57bl6/rm_rm_d_c57bl6n_mouse.aspx
[12] ENVIGO, C57BL/6Hsd, 27/10/2017: http://www.envigo.com/products-services/research-models-services/models/research-models/mice/inbred/c57bl-6-inbred-mice/

contendo essas variantes foram analisados e se destacam os genes *vmn2r65* (vomeronasal 2 receptor 65), c*yp2a22* (citocromo P450 família 2, subfamília a, polipeptídeo 22), *rptor* (Raptor – proteína chave na regulação do complexo de sinalização mTORC1), *plk1* (importante no ciclo celular), *herpud2* (envolvida no processamento de proteínas do retículo endoplasmático), *crb1* (mutações provocam degeneração na retina), *cyfip2* (função neuronal e polimerização de actina), *Ide* (envolvida na degradação de insulina) e *fgfbp3* (fator de crescimento de fibroblastos) (SIMON et al., 2013).

As principais diferenças clássicas entre os camundongos 6N e 6J implicam em diferentes respostas para o álcool (KHISTI et al., 2006), susceptibilidade na formação de tumores (DIWAN; BLACKMAN, 1980), pancreatite crônica (ULMASOV et al., 2013), diferenças no desenvolvimento ocular (SIMON et al., 2013) – incluindo catarata e resposta a luz – que podem explicar diferenças na manutenção do ciclo circadiano (BANKS et al., 2015). Outras diferenças como pressão arterial sistólica e habilidade motora (MATSUO et al., 2014) também foram reportadas assim como diferenças na resposta imune ativada por células NK (SIMON et al., 2013). Nem todas as diferenças são explicadas pela falta da proteína Nnt em camundongos 6J apesar do impacto na função mitocondrial (RONCHI et al., 2013), que pode afetar inúmeros processos metabólicos pela regulação de vias redox dependentes de glutationa e tireodoxina/peroxiredoxina (FONTAINE; DAVIS, 2016).

2.1.4 Origens e características da linhagem BALB/c

Ao redor de 1930 foi originada a linhagem BALB/c, através do cruzamento de camundongos albinos de Bagg, por MacDowell (GREEN, 1966; GUÉNET et al., 2015). Convenientemente um outro pesquisador, George D. Snell, recebeu os camundongos albinos de Bagg e cunhou a nomenclatura BALB/c (POTTER; MORRIS, 1985), conforme mostrado na **Figura 2.3**. A linhagem isogênica BALB/c é uma das mais utilizadas no mundo, juntamente com os camundongos C57BL/6 (EPPIG et al., 2015). É particularmente conhecida por fornecer plasmocitomas, tipos de células tumorais que constituem a base para a produção de anticorpos monoclonais (GREENFIELD, 2014).

Três principais sublinhagens BALB/c remontam a 1940 e divergiram devido à heterozigosidade residual presente nas matrizes iniciais e não devido a contaminação com cruzamentos errôneos (POTTER; MORRIS, 1985). Essas três

sublinhagens são conhecidas como BALB/cHeAn, BALB/cJ e BALB/cRI, diferenciadas por mutações nos locus *raf1* – que controla a expressão da alfa-fetoproteína, *qa2* – que controla antígenos de superfície, *gdc1* – que codifica para a L-glicerol 3-fosfato desidrogenase no fígado e a sequência repetitiva *pr1* (POTTER; MORRIS, 1985).

A colônia da sublinhagem BALB/cJ, fornecida atualmente pela JAX[13] acumulou 243 gerações em agosto de 2014. Essa linhagem foi originada em 1947 a partir da geração 41 de uma colônia de J. Paul Scott – criada a partir das colônias originais providas por Snell em 1938/1939. A origem dos camundongos fornecidos pela Taconic[14] é uma colônia em F184 antes de 1988, recebidas pelo NIH – provenientes do estoque original de Snell e chamadas de linhagem Andervont. Em 2005 a colônia da Taconic estava em geração F229 muito similar às colônias F221 da JAX.

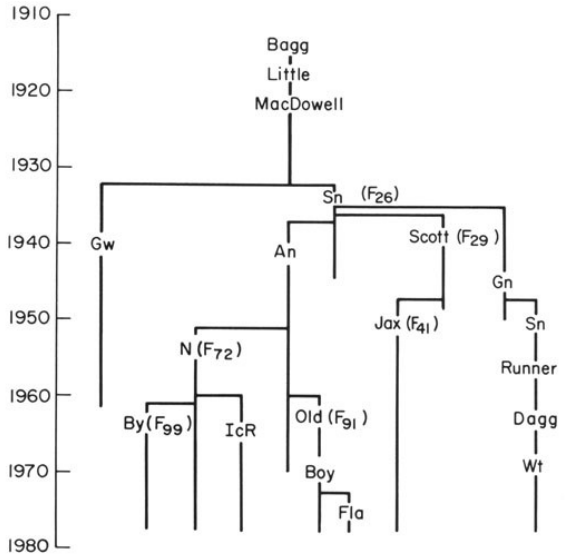

Figura 2.3 Origens da linhagem BALB/c e suas sublinhagens. A genealogia acima foi retirada de (BAILEY, 1978), Sn = G.D Snell, Na = H.B. Andervont, N = NIH, By = D.W. Bailey, Boy – E.A. Boyse, Fla = L. Flaherty, Gw = J.W. Gowen, Gn = E.L. e M.C. Green, JAX = Jackson Laboratory, IcR = Institute for Cancer Research, Wt = W.K. Whitten. Fonte: (BAILEY, 1978).

Camundongos BALB/c são relativamente dóceis, com boa performance reprodutiva e expectativa de vida alta. Também possuem baixa incidência de tumores de mama, mas desenvolvem vários tipos de câncer em idades avançadas –

[13] JAX Strain BALB/cJ, 27/10/2017: https://www.jax.org/strain/000651
[14] Taconic Strain BALBc, 27/10/2017: https://www.taconic.com/mouse-model/balbc

incluindo neoplasias reticulares e tumores nos pulmões e rins (HESTON; VLAHAKIS, 1971)- além de constituírem um modelo muito utilizado em pesquisa cardiovascular, por serem resistentes a arteriosclerose induzida por dieta (PAIGEN et al., 1990). Camundongos BALB/c também desenvolvem uma forma atípica de encefalomielite autoimune experimental (EAE) quando imunizados com $PLP_{180-199}$ (TEUSCHER et al., 1998).

2.1.5 Breve história das sublinhagens C57BL/6-ICBI e BALB/c-ICBI

O Instituto de Ciências Biomédicas, através do biotério de experimentação do Departamento de Imunologia, fornece diversas linhagens isogênicas e co-isogênicas – camundongos transgênicos ou mutantes. A linhagem C57BL/6 é atualmente a linhagem mais utilizada pelos pesquisadores seguida da linhagem BALB/c (MASSIRONI, comunicação pessoal).

A colônia C57BL/6 do ICB foi originada de matrizes do CEMIB/UNICAMP, denominada atualmente de C57BL/6JUnib[15] (RONCHI et al., 2013), oriundas originalmente da Universidade de Medicina Veterinária em Hannover-Alemanha (Zentralinstitut für Versuchstierzucht, Hannover), em meados de 1980 (MASSIRONI, comunicação pessoal), presumidamente de uma colônia C57BL/6N. A colônia BALB/c do ICB foi originada com matrizes originadas da sublinhagem BALB/cJ, provenientes do Instituto Pasteur (França), no início dos anos 90 (**Figura 2.4**), como doação do Prof. Jean Louis Guénet. Essa colônia também é mantida há pelo menos 25 anos e também pode ser considerada uma sublinhagem originada a partir da sublinhagem BALB/cJ (MASSIRONI, comunicação pessoal).

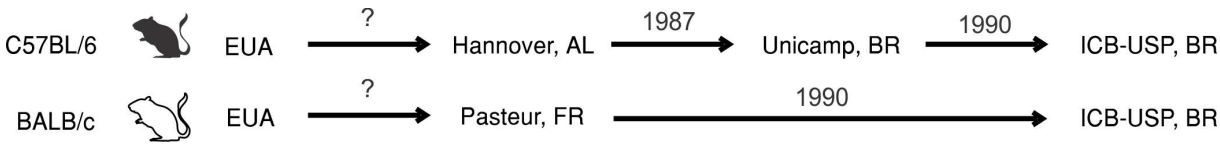

Figura 2.4 Origens das sublinhagem C57BL/6ICBI e BALB/cICBI. O camundongo C57BL/6ICBI foi trazido de Hannover, Alemanha, em 1987 para o CEMIB-UNICAMP, em Campinas. Alguns anos mais tarde foi introduzido no ICB, formando a colônia atual do Biotério. Não se sabe a origem da colônia de Hannover, mas presume-se que foi originada do NIH (C57BL/6N) nos EUA. Os camundongos BALB/cIBI foram trazidos no início da década de 90 diretamente do Instituto Pasteur para o ICB-USP.

---

[15] CEMIB, UNICAMP: www.cemib.unicamp.br

Desde então, as linhagens são mantidas através de sucessivos cruzamentos irmão-irmã na colônia sendo tanto a linhagem C57BL/6 quanto a linhagem BALB/c mantidas em populações isoladas. Considerando a duração de uma geração como 12 semanas, os camundongos do ICB estariam em sua 108ª geração desde 1990. Portanto, segundo os critérios clássicos, as colônias são consideradas sublinhagens e por esse motivo, as sublinhagens disponíveis no ICB-USP serão identificadas como C57BL/6ICBI e BALB/cICBI, segundo as recomendações de nomenclatura sugeridas pela JAX[16].

2.1.6 Sublinhagens e seu impacto na reprodutibilidade e interpretação de experimentos

Ao contrário do senso comum, camundongos isogênicos não são completamente idênticos geneticamente entre si e novas variações podem ser fixadas devido à inúmeros fatores de seleção e fixação de alelos (CASELLAS, 2011), originando sublinhagens que podem se comportar de maneira muito heterogênea em termos de fenótipo. Um estudo de 2008 (CASELLAS; MEDRANO, 2008) estimou em ~4,5% do total da variabilidade fenotípica em camundongos C57BL/6J foi devido a novas mutações que acumularam na colônia em poucas gerações, provavelmente por deriva genética. Quando essas variações são fixadas em uma determinada colônia - isolada geograficamente e fundada por poucos animais oriundos de uma linhagem parental – surge o que chamamos de sublinhagem, cujos membros podem diferir consideravelmente em termos de *background* genético com sua linhagem parental (COLETTI et al., 2013; NICHOLSON et al., 2010).

Sublinhagens podem surgir devido a vários fatores, como heterozigosidade residual (BAILEY, 1978); mutações espontâneas fixadas por deriva genética (CASELLAS, 2011; WOTJAK, 2003); contaminação acidental com outras linhagens (NAGGERT et al., 1995; WOTJAK, 2003); um novo status em uma colônia – como certificações livre de patógeno, por exemplo, e finalmente a própria separação de uma subcolônia da sua colônia parental por 20 ou mais gerações. Todas essas razões são suficientes para analisar cuidadosamente respostas de um modelo

---

[16] JAX Strain Nomenclature, 27/10/2017: http://www.informatics.jax.org/nomen/strains.shtml

murino de uma maneira reprodutível e interpretá-las da melhor forma possível (JUSTICE; DHILLON, 2016; MANDILLO et al., 2008).

Infelizmente, todo o grau de diversidade genética entre sublinhagens e até mesmo linhagens não é muito bem caracterizado. Apesar do conhecimento crescente sobre as variações fenotípicas entre linhagens e toda a complexidade da base genética no fenótipo (SHAO et al., 2008), das variantes mendelianas até a complexidade infinitesimal da maioria dos traços fenotípicos quantitativos (FLINT; MACKAY, 2009), não conhecemos a base molecular da maioria dos fenótipos que são influenciados diretamente pelo background genético. Os estudos mantidos pelo *Mouse Genomes Project* vêm colaborando muito com o conhecimento sobre a variação genética em camundongos e seu impacto no fenótipo (KEANE et al., 2011), potencializados pelo sequenciamento do genoma de várias linhagens isogênicas (DORAN et al., 2016; YALCIN et al., 2012). No entanto, muitos estudos apontam para diferenças fenotípicas significativas entre sublinhagens, como por exemplo diferenças comportamentais (STIEDL et al., 1999), rejeição de tecidos em sublinhagens 129 (SIMPSON et al., 1997) e susceptibilidade diferencial a tumores em C3H (GLANT et al., 2001). O exemplo mais marcante das diferenças entre sublinhagens permanece com os camundongos C57BL/6J e C57BL/6N, usados durante muito tempo como sendo de certa forma equivalentes. Esses camundongos *black* apresentam respostas completamente diferentes a diversos experimentos como: resposta ao medo (STIEDL et al., 1999), anestésicos e função cardíaca (ROTH et al., 2002) e limiares de eletro convulsão (YANG et al., 2003).

Considerando a era pós-genômica em que estamos inseridos, é muito importante entender o *background* genético dos modelos utilizados, o que pode significar um salto qualitativo enorme para a interpretação e reprodutibilidade dos dados (COLETTI et al., 2013; SITTIG et al., 2014). Caracterizar geneticamente sublinhagens de camundongos deve ser vista como uma realidade, visto os custos cada vez menores de sequenciamento (YALCIN et al., 2012).

A caracterização genética das sublinhagens isogênicas mantidas no ICB é, portanto, essencial para a avaliação comparativas dos animais no seu contexto genético e possibilita o rastreio de mutações espontâneas fixadas na população, servindo de referência para os estudos desenvolvidos com esses animais por pesquisadores do Instituto e de outras instituições pelo Brasil. Esse estudo pode ser considerado um esforço pioneiro no sentido de aprofundar o nível de informação

levada ao pesquisador em termos de controle genético e manutenção de colônias, de forma a facilitar a detecção de mutações espontâneas e contaminações, estimulando a enorme importância de estudos frequentes e sistematizados nesse sentido.

2.2 Material e Métodos

2.2.1 Sequenciamento do exoma completo: mapeamento das leituras

As leituras de 1x75 pb das amostras C57BL/6ICBI e BALB/cICBI foram produzidas pela plataforma de sequenciamento NGS SOLiD 5500XL, conforme descrito no Capítulo 1. As análises de mapeamento foram realizadas com o genoma mm10 (GRCm38/mm10), obtido do *UCSC Genome Browser.* As leituras foram mapeadas em *colorspace* com a ferramenta LifeScope 2.1 (Thermo Fisher Scientific, Waltham, Massachusetts, EUA) utilizando os módulos presentes na *pipeline* de análise de resequenciamento de regiões-alvo - *Targeted Resequencing Analyses*. Como parâmetros-base para essa análise foram considerados alinhamentos primários e o mapeamento no exoma foi considerado apenas se pelo menos uma base tiver sido alinhada nas regiões-alvo.

O primeiro estágio foi a utilização do módulo SAET - SOLiD Accuracy Enhancement Tool (Thermo Fisher) - que usa um algoritmo de correção espectral de alinhamento semelhante à abordagem de caminho euleriano (PEVZNER; TANG; WATERMAN, 2001) e ao algoritmo SHREC (SALMELA, 2010), que realiza correções de erro nas leituras em *colorspace* implementando correções espectrais em *k-mers*, cuja acurácia gira em torno de 99,99%. Os estágios de mapeamento envolveram etapas de distribuição das leituras nos nós computacionais; mapeamento; geração do arquivo BAM (uma etapa de junção final para cada conjunto de leituras) e tradução do alinhamento de *colorspace* para *basespace*. O mapeamento foi realizado utilizando os genomas referências de camundongo mm9 (NCBI37/mm9) e mm10 (GRCm38/mm10) com parâmetros-padrão da *pipeline,* utilizando mapeamento global no esquema 75.6.0 para os mapeamentos primários e o mapeamento local 20.1.0:30 para os mapeamentos secundários.

Todas as análises de mapeamento foram avaliadas através do módulo *bamstats* quanto às estatísticas globais de mapeamento, cobertura e qualidade do mapeamento. Para o cálculo e seleção das regiões alvo foi utilizado o conjunto de coordenadas cobertas (*covered*) das sondas de enriquecimento do kit SureSelect

Mouse All-Exon kit (S0276129) em formato BED, totalizando 49,6 milhões de bases de regiões-alvo. As coordenadas, originalmente desenhadas para o genoma NCBI37/mm9 foram convertidas para o genoma GRCm38/mm10 pela ferramenta LiftOver[17].

2.2.2 Sequenciamento do exoma completo: chamada de SNPs

O módulo utilizado para a chamada de SNPs na suíte Lifescope é o módulo *diBayes* - baseado no algoritmo diBayes - que executa a identificação das variantes utilizando uma abordagem bayesiana aliada a um algoritmo frequentista. Para análises de regiões-alvo, como os exomas, apenas leituras com qualidade mínima de mapeamento - *mapping QV* - maiores que 8 e com valor mínimo de qualidade por base igual a 26.

A chamada de SNPs considerou apenas: variantes que foram detectadas nas duas fitas; proporção de leituras válidas – *mapping QV* >8 – por total de leituras maior que 70%; número mínimo de posições únicas de início de leituras mapeadas maior que 3 para variantes em heterozigose e homozigose; qualidade da base mínima igual a 26 para o alelo não-referência; proporção mínima igual a 0,15 para o alelo alterado. Esse conjunto de parâmetros forma o tipo de chamada mais estringente, com o menor número de falsos-positivos, a um custo de aumento de falsos-negativos. Após a utilização dos algoritmos frequentista e bayesiano utilizando a codificação em *colorspace* e chamada dos SNPs baseada nos critérios anteriores foi construída a listagem bruta dos SNPs chamados, que foi filtrado como descrito em **2.2.3**.

2.2.3 Filtragem, comparações e anotação de SNPs

A aplicação de filtros de SNPs foi realizada pela ferramenta VCFTools (DANECEK et al., 2011), com implementações compiladas pela biblioteca HTSLib. As comparações entre conjuntos de SNPs, comuns e exclusivos, foi realizada pelo parâmetro *vcf-isec* de forma posicional e de troca sem considerar a zigosidade. O filtro adicional de cobertura mínima de 20X foi aplicado somente em condições de seleção rigorosa, para seleção de mutações importantes.

---

[17] Genome Browser LiftOver Tool: https://genome.ucsc.edu/cgi-bin/hgLiftOver

A lista de SNPs bruta obtida diretamente da chamada de SNPs ou filtrada foi reformatada e anotada com a ferramenta ANNOVAR (WANG; LI; HAKONARSON, 2010) utilizando as anotações RefGene/RefSeq para os genomas mm9 e mm10. A análise de concordância com o banco de dados de polimorfismos dbSNP142 foi também realizada com a ferramenta ANNOVAR.

2.2.4 Avaliação do impacto dos SNVs e análises de enriquecimento de vias metabólicas

As análises de impacto das variantes foram realizadas através das ferramentas PolyPhen2 (ADZHUBEI; JORDAN; SUNYAEV, 2013), SIFT (KUMAR; HENIKOFF; NG, 2009), PROVEAN (CHOI; CHAN, 2015) e SpliceMan (LIM; FAIRBROTHER, 2012). As ferramentas SIFT e PROVEAN foram utilizadas através de implementações *web* na plataforma Variant Effect Predictor (VEP-ENSEMBL)[18], como valores de *cutoff* de predição deletéria iguais a -2.5 e 0.05, respectivamente. Para a análise pela ferramenta PolyPhen-2 foi utilizada uma implementação local da ferramenta em um servidor dedicado com a configuração dos bancos de dados para sequências de camundongos[19], sendo as coordenadas anotadas em função dos indicadores UNIPROT.

Foram selecionados apenas os SNVs que implicavam em trocas não-sinônimas, ou seja, aqueles que potencialmente impactavam no produto gênico considerado. Os genes que continham esse tipo de SNV foram submetidos à uma análise de enriquecimento de classes funcionais pela ferramenta ENRICHR (CHEN et al., 2013), de acordo com três tipos de anotação: GO-Molecular Function, MGI-Mammalian Phenotype Level 3 e MGI-Mammalian Phenotype Level 4. Os mesmos tipos de SNPs das linhagens C57BL/6J e BALB/cJ foram utilizados como controles.

2.2.5 Bancos de dados e sequências dos camundongos gold standard

Os dados de SNPs para as amostras C57BL/6J (MMR-664) e BALB/c (MMR-651) foram gentilmente cedidos pelo grupo da Dra. Laura Reinholdt (Jackson Laboratory, Bar Harbor, Maine, EUA). Os SNPs foram obtidos através de uma *pipeline* própria de alinhamento e chamada de SNPs da JAX, utilizando leituras

---

[18] Variant Effect Predictor (VEP-ENSEMBL): http://www.ensembl.org/info/docs/tools/vep/index.html
[19] Polyphen-2 Standalone configuration protocol: http://genetics.bwh.harvard.edu/pph2/dokuwiki/docs

obtidas pela plataforma Illumina HiSeq (Illumina, San Diego, California, EUA) e o genoma mm10 - GRCm38/mm10 - como referência.

As listas exclusivas de SNPs das linhagens C57BL/6NJ, C57BL/10J, A/J e BALB/cJ e o banco de dados de SNPs de camundongos dbSNP142 foram obtidas do *Mouse Genomes Project*[20] do Wellcome Trust Sanger Institute, UK (KEANE et al., 2011) e foram chamadas em relação ao genoma referência mm10 (GRCm38/mm10). Foram considerados apenas os polimorfismos de um único nucleotídeo - SNPs, excluindo das análises os pequenos INDELs.

2.3 Resultados

2.3.1 Visão geral do sequenciamento de exomas das linhagens C57BL/6ICBI e BALB/cICBI

O mapeamento das leituras pode ser considerado a primeira etapa em análises de ressequenciamento. Para efeitos comparativos e também como controle externo aos experimentos foram utilizados dados relativos ao sequenciamento completo de exomas de duas linhagens isogênicas, cedidas gentilmente pela Dra. Laura Reinholdt (*The Jackson Laboratory - JAX*). As duas amostras, sequenciadas e alinhadas pela JAX, foram obtidas através de método de enriquecimento diferente – NimbleGen (Roche), sequenciado por outra plataforma - Illumina HiSeq, em modo *paired-end* - e analisado de forma totalmente independente – GATK[21] adaptado. A **Tabela 2.1** fornece as estatísticas de alinhamento das duas amostras sequenciadas no ICB e também duas amostras sequenciadas e analisadas independentemente pela JAX.

A quantidade total de leituras produzidas foi superior à obtida pelo sequenciamento da JAX, mesmo considerando que o último produziu sequencias *paired-end* de 100 pb. Um outro parâmetro importante, a taxa de alinhamento das leituras, foi ligeiramente superior na análise da JAX – maior que 96,9% - mas ainda superior a 92,6 % (**Tabela 2.1**). A eficiência do enriquecimento, medida pela proporção de leituras alinhadas nas regiões-alvo (exons), foi de 78% e 80% nas amostras do ICB, sendo que a quantidade de regiões-alvo não cobertas foi menor que 3,87%. A cobertura média das regiões-alvo foi ligeiramente superior nas

---

[20] Mouse Genomes Project Download Database FTP: ftp://ftp-mouse.sanger.ac.uk/
[21] Genome Analysis Toolkit, Broad Institute: https://software.broadinstitute.org/gatk/

amostras do ICB – maior que 100X - reflexo da maior produção de leituras no sequenciamento, sendo que mais que 84% das regiões alvo tiveram no mínimo 20X de cobertura, enquanto as amostras da JAX tiveram cerca de 79% de regiões-alvo cobertas pelo menos 20 vezes.

Apesar disso, o número total de SNPs chamados e a zigosidade dos mesmos é ainda assim muito semelhante entre as linhagens sequenciadas e analisadas em comparação com os dados da JAX (**Tabela 2.1**), mesmo utilizando plataformas de sequenciamento e *pipelines* distintas. As amostras da linhagem BALB/c apresentam cerca de 50 vezes mais SNPs chamados em relação às linhagens C57BL/6, evidenciando a interferência do genoma referência ser majoritariamente proveniente de linhagens C57BL/6.

Tabela 2.1 - Estatísticas de alinhamento das linhagens isogênicas sequenciadas do ICB e de dois sequenciamentos realizados na JAX.

Linhagem	**BALB/cICBI**	**BALB/cJ (MMR651)**	**C57BL/6ICBI**	**C57BL/6J (MMR664)**
Origem dos dados	ICB-USP	JAX	ICB-USP	JAX
Enriquecimento	SureSelect	NimbleGen	SureSelect	NimbleGen
Genoma referência	mm10	mm10	mm10	mm10
Plataforma	SOLiD	Illumina	SOLiD	Illumina
Num. leituras	173961821	71943600	140456591	75029224
Num. leituras mapeadas	161127172	69838892	131228128	72717949
% leituras mapeadas	92,6	97,1	93,4	96,9
Leituras mapeadas nos alvos	96882310	-	82263009	-
% leituras mapeadas nos alvos	78,72%	-	80,29%	-
Leituras mapeadas fora dos alvos	26196777	-	20190276	-
% leituras mapeadas fora dos alvos	21,28%	-	19,71%	-
Número de regiões-alvo sem cobertura	3102	-	3166	-
Porcentagem de bases-alvo sem cobertura	1912736 (3,87%)	-	1899983 (3,85%)	-
Porcentagem de bases-alvo cobertas >= 1X:	96,13%	-	96,15%	-
Porcentagem de bases-alvo cobertas >= 5X:	93,55%	-	93,43%	-
Porcentagem de bases-alvo cobertas >= 10X:	91,07%	90,96%	90,59%	91,13%
Porcentagem de bases-alvo cobertas >= 20X:	85,78%	79,40%	84,28%	78,79%
Média de cobertura nas regiões-alvo	121,23	44,31	101,78	44,95
Número de SNPs chamados	76763	78851	1619	1350
Número de SNPs em homozigose	71422	72117	200	167
Número de SNPs em heterozigose	5341	6734	1419	1183

O número e zigosidade dos SNPs brutos chamados nas amostras foram praticamente idênticos entre as linhagens isogênicas do ICB e da JAX. A variação no número de SNPs brutos entre as linhagens C57BL/6 foi de aproximadamente 15% e da linhagem BALB/c cerca de 3% (**Tabela 2.1**). A relação de SNPs homozigotos e heterozigotos foi praticamente a mesma entre as linhagens, porém as linhagens C57BL/6 apresentam a maioria de SNPs heterozigotos e as linhagens BALB/c apresentam a maioria dos SNPs detectados em homozigose (**Tabela 2.1**). Vale destacar que 1350 polimorfismos foram detectados na amostra C57BL6/J, linhagem que corresponde justamente à linhagem majoritária sequenciada do genoma referência. Os valores detectados na linhagem C57BL6/JICB foram apenas ligeiramente superiores, cerca de 1619 SNPs.

2.3.2 Concordância dos SNPs encontrados em relação à banco de dados de outras linhagens

A semelhança nas estatísticas de alinhamento e na chamada de SNPs entre as amostras sequenciadas no ICB e suas respectivas amostras de referência de linhagens sequenciadas na JAX permitiram algumas comparações interessantes. Utilizamos o banco de dados oriundo do *Mouse Genomes Project (Wellcome Trust Sanger Institute),* que sequenciou o genoma completo de 36 diferentes linhagens de camundongos com uma cobertura média de 40X, com leituras de 100 pb provenientes da plataforma Illumina HiSeq. Nas análises de ressequenciamento desse projeto foi utilizada a versão mm10 como referência, e foram disponibilizados os SNPs privativos relativos à cada linhagem e também um banco de polimorfismos total das linhagens sequenciadas. Foram utilizados em nossa análise o banco de SNPs global de todas as linhagens sequenciadas (dbSNP-Sanger) e das linhagens BALB/cJ (amostra BALB/cJ-Sanger), A/J (amostra AJ-Sanger), C57BL/10J (amostra C57BL10J-Sanger) e C57BL/6NJ (amostra C57BL6NJ-Sanger).

O perfil de concordância dos SNVs das linhagens BALB/cICBI e BALB/cJ (amostra MMR651) em relação aos dados da linhagem BALB/c do *Mouse Genomes Project* foi de 95,6% e 94,6% respectivamente, ou seja, praticamente o mesmo número de SNVs considerados como polimorfismos comuns entre linhagem BALB/c (**Figura 2.5A** e **Figura 2.5B**). A concordância dessas amostras com SNPs de linhagens C57BL/10J e C57BL/6NJ, distantes evolutivamente da linhagem BALB/c,

foi de 3,72% e 0,75% (BALB/cICBI) e 4,37% e 0,92% (BALBc/J-MMR651) respectivamente, enquanto a concordância com SNPs da linhagem A/J-Sanger, mais próxima em termos evolutivos, foi de 75,16% (BALB/cICBI) e 71,97% (BALB/cJ-MMR651) (**Figura 2.5A** e **Figura 2.5B**).

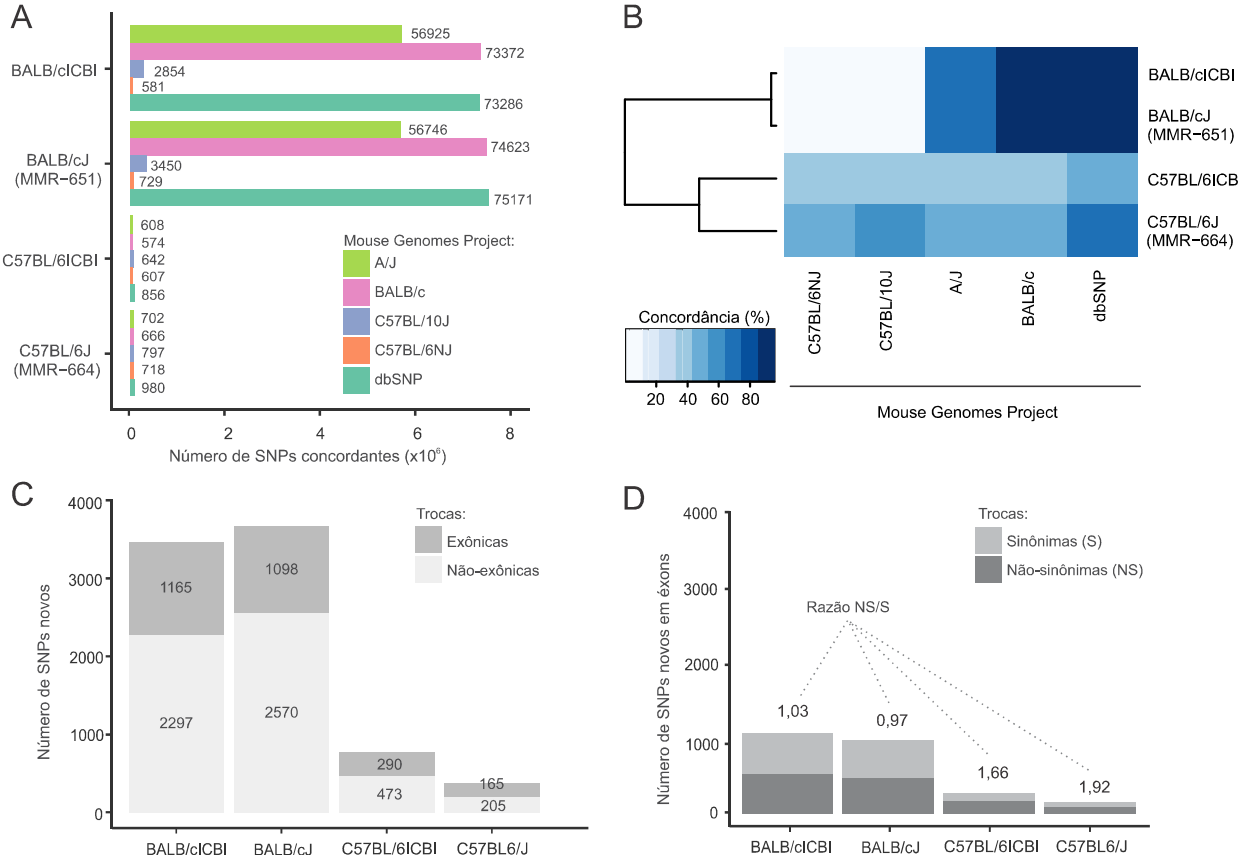

Figura 2.5 – Concordância dos SNPs encontrados nas amostras de camundongos do ICB (BALB/cICBI e C57BL/6ICBI) e da JAX (BALB/cJ e C57BL/6J). (**A**) Número absoluto de SNPs brutos chamados em concordância com o mesmo grupo de dados do Mouse Genomes Project (Wellcome Trust Sanger Institute). (**B**) *Heatmap* indicando a porcentagem de concordância das amostras em relação aos SNPs exclusivos de 4 linhagens próximas (C57BL/6NJ, C57BL/10J, A/J e BALB/cJ) e do banco de dados global de SNPs de camundongos (versão dbSNP142). (**C**) Número de SNPs novos e porcentagem de concordância de todos os SNPs em relação ao dbSNP142. (**D**) Classificação da troca dos SNPs novos em trocas não sinônimas, de ganho de stop ou em sítios de splicing (não-sinônima) e sinônimas. NS/S indica a a razão entre trocas sinônimas e não sinônimas dos SNPs novos.

A comparação do perfil dos SNVs das amostras C57BL/6ICBI e C57BL/6J (MMR664) com o banco de dados do *Mouse Genomes Project* manteve-se semelhante em termos de proporção da concordância específica com os SNPs de cada linhagem analisada (**Figura 2.5A** e **Figura 2.5B**). Porém, a linhagem C57BL/6ICBI tem cerca de 20% a mais de SNPs novos em relação à amostra C57BL/6J, ou seja, que não foram encontrados em nenhuma das 36 linhagens sequenciadas (**Figura 2.5C**). Em termos absolutos, C57BL/6ICBI tem mais que o

dobro de SNPs novos (n=763) do que a amostra C57BL/6J (n=370), sendo que o número de SNPs novos das amostras BALB/cICBI (n=3477) e BALB/cJ (n=3680) é similar (**Figura 2.5C**).

2.3.3 Principais diferenças genéticas encontradas nos exomas dos camundongos do ICB

As linhagens C57BL/6ICBI e C57BL/6J (MMR-664) apresentaram um menor número de SNPs novos em relação às linhagens BALB/cICBI e BALB/cJ (MMR-651), considerando o banco de dados de polimorfismos do *Mouse Genomes Project* (**Figura 2.5C**). Em relação à linhagem C57BL/6ICBI cerca de 38% dos SNPs novos estão situados em regiões exônicas ou de sítios de *splicing* e apenas 178 SNPs são trocas não sinônimas e estão provavelmente associadas à um potencial impacto na função de seus respectivos produtos gênicos (**Figura 2.5D**) e apenas 20 desses SNPs foram detectados em homozigose. Foram detectados 107 novos SNPs que implicaram em trocas sinônimas no camundongo C57BL/6ICBI e 178 SNPs não sinônimos. Em comparação com a amostra C57BL/6J, cerca de 45% dos SNVs estão associados a regiões exônicas ou de *splicing* e 102 SNVs são trocas não sinônimas (**Figura 2.5D**) e 53 SNPs implicam em trocas sinônimas. A relação de trocas sinônimas/não sinônimas (Ds/Dn) é de cerca de 1,66 na linhagem C57BL/6ICBI e de 1,92 na linhagem C57BL/6J (**Figura 2.5D**).

Na linhagem BALB/cICBI cerca de 34% dos SNPs novos estão em regiões exônicas ou associadas à *splicing,* 581 SNPs acarretam em trocas não sinônimas e 566 SNPs sinônimos (**Figura 2.5D**). A linhagem BALB/cJ (MMR651) apresenta resultados semelhantes: cerca de 30% do total de SNPs novos estão em regiões exônicas, 514 SNPs são trocas não sinônimas e 530 SNPs são trocas sinônimas (**Figura 2.5D**). A relação Ds/Dn das duas linhagens também é semelhante, de cerca de 1, que implica um número semelhante de trocas sinônimas e não sinônimas.

A classificação em termos GO (**Figura 2.6**) identificou um enriquecimento comum, nas sublinhagens BALB/cICBI e BALB/cJ, da classe *calcium ion binding* (GO:0005509). Já na sublinhagem C57BL/6ICBI foi identificado um enriquecimento do termo *actinin binding* (GO:0042805), não enriquecido na sublinhagem C57BL/6J, cujos temos mais enriquecidos foram *translation regulator activity, nucleic acid binding* (GO:0090079) e *histone deacetylase activity (H4-K16 specific)* (GO:0034739), ambos relacionados à regulação da transcrição gênica.

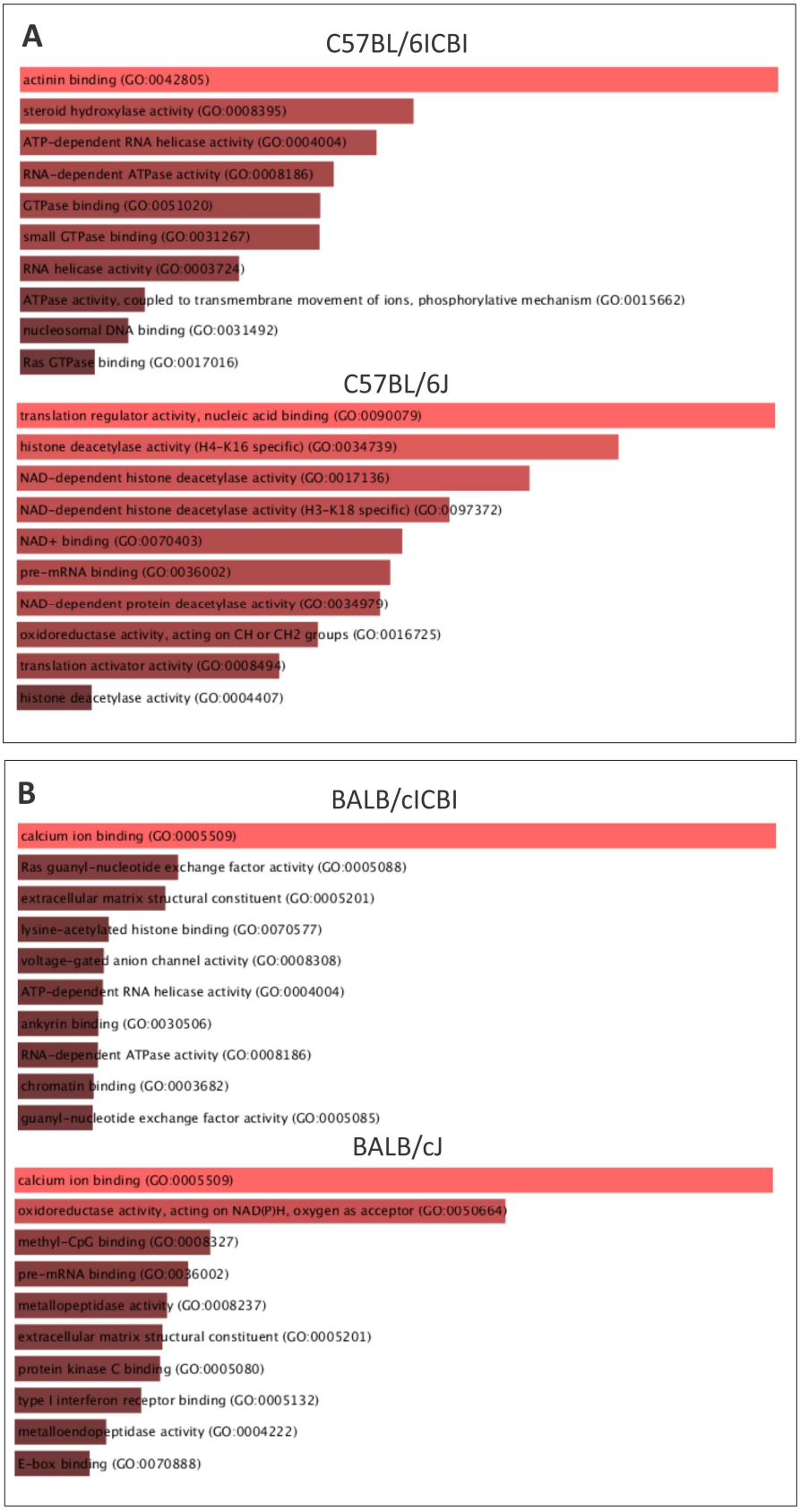

Figura 2.6 – Enriquecimento de termos GO associado aos SNPs exclusivos e não sinônimos nos camundongos do ICB (BALB/cICBI e C57BL/6ICBI) e nos camundongos da JAX (BALB/cJ e C57BL/6J). (A) Termos GO enriquecidos nos camundongos C57BL/6 e nos camundongos BALB/c (B), com os respectivos códigos GO. As análises de enriquecimento foram realizadas pela ferramenta Enrichr.

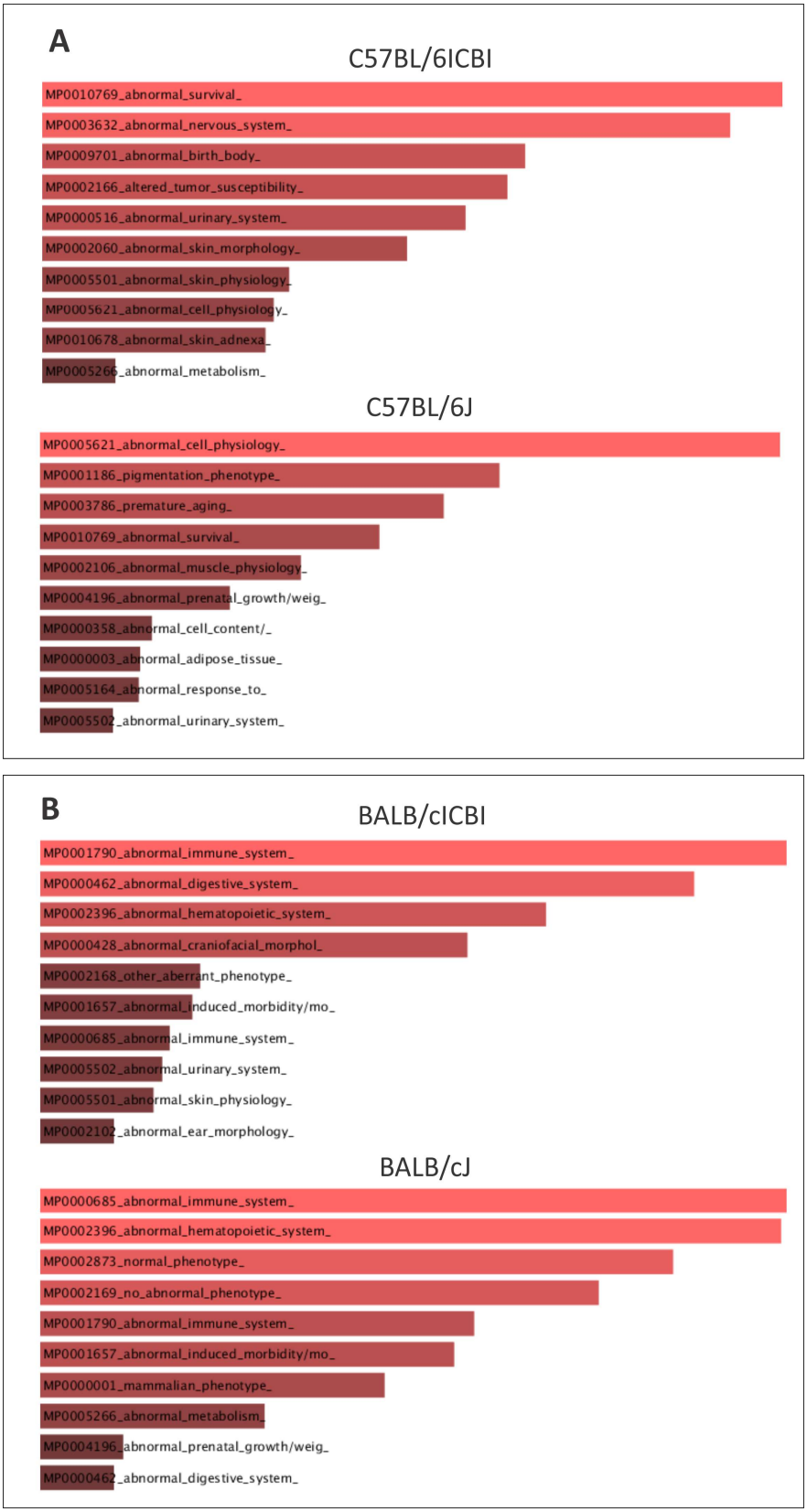

Figura 2.7 - Análise de enriquecimento de classes de genes afetados com SNPs novos que implicam em trocas não-sinônimas pela ferramenta Enrichr. Classes de genes enriquecidas segundo a classificação MGI Mammalian Phenotype Level 3 das amostras C57BL/6ICBI, C57BL/6J (A), BALB/cICBI e BALB/cJ (B) com destaque para as classes compartilhadas.

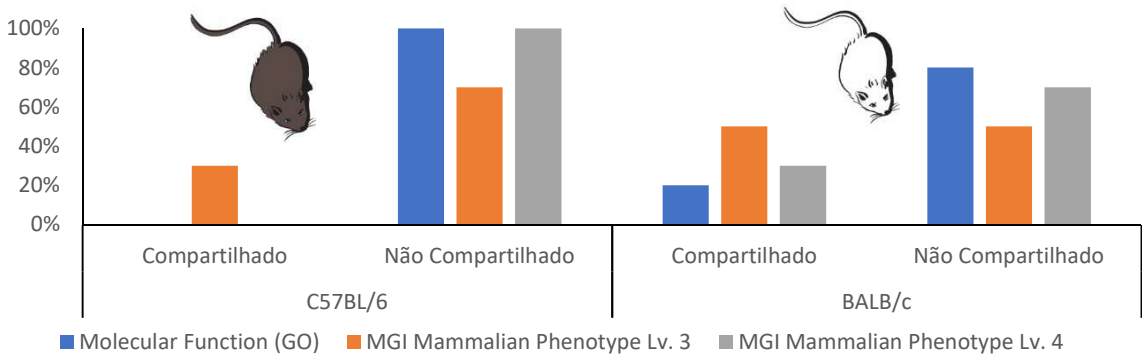

Figura 2.8 – Compartilhamento de classes enriquecidas em genes com variantes não-sinônimas. Proporção de classes enriquecidas compartilhadas e não compartilhadas segundo a classificação *Gene Ontology (GO) – Molecular function*, MGI Mammalian Phenotype Level 3 e Level 4 das amostras correspontes às linhagens C57BL/6 e BALB/cJ.

Os grupos de linhagens C57BL/6ICBI e BALB/cICBI apresentaram resultados diferentes quanto ao perfil de classes funcionais de genes com SNPs não-sinônimos (**Figura 2.6** e **Figura 2.7**). As linhagens C57BL/6ICBI e C57BL/6J compartilham apenas 30% das 10 classes funcionais mais enriquecidas do MGI Phenotype Level 3 (**Figura 2.7A e Figura 2.8**) enquanto as linhagens BALB/c compartilham 50% das classes (**Figura 2.7B e Figura 2.8**). Essa relação se mantém em relação às outras classes funcionais analisadas (**Figura 2.8**).

O perfil de troca de nucleotídeos é similar entre as linhagens, sendo que a frequência de trocas C>T é maior nos camundongos C57BL/6 (JAX) em relação ao ICB e menor em camundongos BALB/c (**Figura 2.9A**). Devido provavelmente à frequência de trocas C>T alterada, o número de transições também acompanha o mesmo padrão observado (**Figura 2.9B**). O espectro de trocas baseado em trinucleotídeos também não revelou nenhuma assinatura típica em relação às amostras, mas foi observado que o padrão de trocas entre as sublinhagens BALB/c é muito semelhante. Essa semelhança não é tão óbvia em relação às sublinhagens C57BL/6, sendo que o padrão de trocas em relação aos trinucleotídeos em todos os seis tipos de mutações (**Figura 2.9C**).

Figura 2.9 - Análise do padrão dos SNPs nas linhagens isogênicas. (A) Frequência de cada tipo de troca em relação ao total de mutações pontuais detectadas. (B) Proporção de tranversões e transições. (C) Espectro de mutações baseada em trinucleotídeos vizinhos nas 4 amostras analisadas, considerando apenas variantes novas em exons, realizada com o pacote SomaticSignatures (GEHRING et al., 2015; ALEXANDROV et al., 2013).

2.3.4 Análise comparativa entre as sublinhagens oriundas da Jackson e do ICB

As sublinhagens C57BL/6ICBI e C57BL/6J compartilham 332 SNPs, sendo que 232 deles estão localizados em exons e sítios de *splicing* (**Tabela 2.2**). Isso corresponde a uma concordância de 32,5% do total de SNPs detectados em regiões exônicas em C57BL/6ICBI e cerca de aproximadamente 36% na sublinhagem C57BL/6J. A linhagem C57BL/6ICBI possui 1287 SNVs únicos enquanto a sublinhagem C57BL/6J possui 1018 SNVs únicos em relação à sublinhagem do ICB.

Tabela 2.2 – Análise comparativa entre SNPs encontrados nas linhagens C57BL/6 e BALB/c provenientes do ICB (ICBI) e da JAX (J).

Sublinhagens	Únicos	Comuns (em exons)	% Concordância (em exons)
C57BL/6ICBI vs. C57BL/6J	1287	332 (232)	20,5 (32.5)
C57BL/6J vs. C57BL/6ICBI	1018	332 (232)	24,6 (35,8)
BALB/cICBI vs. BALB/cJ	34796	41967 (23045)	54,7 (84,7)
BALB/cJ vs. BALB/cICBI	36884	41967 (23045)	53,2 (79,7)

Em relação às sublinhagens BALB/c temos aproximadamente mais que 80% dos SNPs em exons ou sítios de *splicing* compartilhados entre as amostras BALB/cICBI e BALB/cJ, correspondentes a 23045 SNPs. Mais de 34,000 SNPs são únicos na sublinhagem BALB/cICBI e quase 37,000 SNPs estão presentes somente em BALB/cJ em relação à linhagem BALB/cICBI.

2.3.5 Mutações potencialmente impactantes encontradas na sublinhagem C57BL/6ICBI.

Para encontrar mutações que se fixaram, provavelmente por deriva genética, na colônia de camundongos do ICB utilizamos o seguinte filtro severo: mutações não sinônimas em homozigose, pelo menos 20X de cobertura local e exclusividade em relação ao banco de dados dbSNP142 (**Tabela 2.3**). Para a sublinhagem C57BL/6ICBI um total de 12 mutações foram encontradas, seguindo os critérios acima, e 6 mutações foram preditas como deletérias por pelo menos dois algoritmos de predição de impacto (Provean, SIFT ou Polyphen-2) nos genes *vmn2r15*, *ran, ths7a, taar2, ispd* e *kat6b*. Por outro lado, apenas três mutações foram preditas como benignas pelos três algoritmos, sugerindo que pelo menos 9 mutações fixadas nessa sublinhagem têm alta probabilidade impactar em algum tipo de manifestação fenotípica (**Tabela 2.3**).

Tabela 2.3 - Mutações potencialmente impactantes encontradas na sublinhagem C57BL6/ICBI

Crm.	Coordenadas	Transcrito	Provean	SIFT	Polyphen-2
1	11089730	Prex2:NM_029525:c.G533A:p.R178Q	Neutral	Tolerated	Possibly damaging
5	34633922	Mfsd10:NM_026660:c.G1348A:p.A450T	Neutral	Tolerated	Benign
5	**109286462**	**Vmn2r15:NM_001104626:c.G2375T:p.C792F**	**Deleterious**	**Tolerated**	**Probably damaging**
5	**129020946**	**Ran:NM_009391:c.G202A:p.G68S**	**Deleterious**	**Damaging**	**Benign**
6	**12471104**	**Thsd7a:NM_001164805:c.C1514T:p.S505L**	**Deleterious**	**Damaging**	**Probably damaging**
7	24612617	Phldb3:NM_001102613:c.G355A:p.V119M	Neutral	Tolerated	Probably damaging
10	**23941248**	**Taar2:NM_001007266: c.C685A:p.R229S**	**Deleterious**	**Damaging**	**Benign**
12	28563449	Allc:NM_053156:c.G493A:p.D165N	Neutral	Tolerated	Benign
12	**36381783**	**Ispd:NM_178629:c.G10T:p.G4W**	**Neutral**	**Damaging**	**Probably damaging**
13	91461196	Ssbp2:NM_024272:c.A10C:p.K4Q	Neutral	Damaging	Benign
14	**21664512**	**Kat6b:NM_001205241:c.G2513T:p.S838I**	**Deleterious**	**Damaging**	**Probably damaging**
15	74841588	Cyp11b1:NM_001033229:c.A25G:p.T9A	Neutral	Tolerated	Benign

Os genes da família *v2r* (Vmn2r15) estão envolvidos na detecção de feromônios em camundongos (YOUNG; HAMMOCK, 2007). Ran é uma proteína pequena G envolvida no transporte do núcleo celular e relacionada ao controle da síntese de DNA (SAZER; DASSO, 2000). *Thsd7a* codifica para uma trombospondina do tipo 1 que está associada a células endoteliais e também com baixa densidade óssea em osteoporose (LU; KIPNIS, 2010). TAAR2 é um proteína G receptora que está envolvida em diversos mecanismos de transdução de sinal (BABUSYTE et al., 2013). O gene *ispd* codifica para uma proteína que produz um tipo de glicano componente na cadeia de glicosilação da proteína alfa-distroglicana, responsável pela estabilização e proteção de fibras musculares (ROSCIOLI et al., 2012). Kat6b é uma histona acetil-transferase, responsável pela acetilação de histonas e regulação gênica de ativação de MAPK (KRAFT et al., 2011).

2.3.6 Mutações potencialmente impactantes encontradas na sublinhagem BALB/cICBI.

Os mesmos critérios utilizados em (**2.3.3**) foram utilizados para identificar os SNVs novos da sublinhagem BALB/cICBI, que poderiam impactar na função dos respectivos genes e com potencial de impacto em manifestações fenotípicas. Um total de 61 SNVs foram encontrados (**Tabela 2.4**), número cerca de 5 vezes mais que os SNVs encontrados na sublinhagem C57BL/6ICBI (**Tabela 2.3**).

Tabela 2.4 - Mutações potencialmente impactantes encontradas na sublinhagem BALB/c/ICBI

Crm.	Coordenadas	Transcrito	Provean	SIFT	PolyPhen-2
1	**167445992**	**Lrrc52:NM_001013382:exon2:c.G754A:p.G252S**	**Deleterious**	**Damaging**	**P. damaging**
1	173936625	Ifi203:NM_001045481:exon4:c.G280A:p.E94K	Neutral	Tolerated	ND
1	174332354	Mptx1:NM_025470:exon2:c.C225G:p.N75K	Neutral	Tolerated	ND
1	174431343	Olfr414:NM_146761:exon1:c.T914C:p.V305A	Neutral	Damaging	ND
2	**72234729**	**Rapgef4:NM_001204165:exon25:c.G2501T:p.C834F**	**Deleterious**	**Damaging**	ND
2	104992241	Ccdc73:NM_177600:exon16:c.A2534T:p.H845L	Neutral	Tolerated	Benign
2	154234522	Bpifb5:NM_144890:exon11:c.G1217A:p.R406K	Neutral	Tolerated	ND
3	15990741	Gm5150:NM_001081687:exon2:c.T319A:p.Y107N	Neutral	Tolerated	ND
3	98619274	Hsd3b5:NM_008295:exon4:c.C855G:p.S285R	Neutral	Tolerated	benign
4	**45058370**	**Fbxo10:NM_001024142:exon3:c.C1366T:p.R456W**	**Deleterious**	**Damaging**	**P. damaging**
4	88816661	Ifna7:NM_008334:exon1:c.C434T:p.T145I	Neutral	Damaging	ND
4	88850126	Ifna1:NM_010502:exon1:c.C40A:p.L14M	Neutral	Tolerated	Benign
4	113535709	Skint5:NM_001167876:exon51:c.C3683G:p.T1228S	Neutral	-	ND
4	113805154	Skint5:NM_001167876:exon21:c.A1793C:p.Y598S	Neutral	Tolerated	ND
4	113819198	Skint5:NM_001167876:exon19:c.C1648A:p.Q550K	Neutral	Damaging	ND
4	113827870	Skint5:NM_001167876:exon18:c.A1607G:p.N536S	Neutral	Tolerated	ND
4	113937660	Skint5:NM_001167876:exon5:c.G725T:p.C242F	Neutral	Damaging	ND
4	123935646	Rragc:NM_017475:exon7:c.T1154C:p.L385P	Neutral	Tolerated	Benign
5	30482167	1700001C02Rik:NM_029285:exon3:c.G437A:p.R146Q	Neutral	Tolerated	Benign
5	**45493517**	**Lap3:NM_024434:exon1:c.A58G:p.R20G**	**Neutral**	**Damaging**	**P. damaging**
5	123511373	B3gnt4:NM_198611:exon1:c.T800G:p.M267R	Neutral	Tolerated	Benign
6	85867854	Cml2:NM_053096:exon3:c.G525C:p.E175D	Neutral	Tolerated	Benign
6	120862938	Bcl2l13:NM_153516:exon3:c.A196T:p.T66S	Neutral	Tolerated	Benign
6	129775385	Gm156:NM_001014997:exon2:c.T166A:p.W56R	Neutral	Tolerated	ND
6	130059561	Klra15:NM_013793:exon4:c.A511G:p.S171G	Neutral	Tolerated	Benign
6	130062116	Klra33:NM_001039118:exon3:c.T312G:p.N104K	Deleterious	Tolerated	Benign
6	130062140	Klra33:NM_001039118:exon3:c.C288A:p.N96K	Neutral	Tolerated	Benign
6	130229980	Klra7:NM_014194:exon2:c.C157T:p.L53F	Deleterious	Tolerated	ND
7	44016829	Klk1b26:NM_010644:exon5:c.A695T:p.E232V	Neutral	Tolerated	benign
7	125850538	D430042O09Rik:NM_001081022:exon16:c.G2851A:p.G951S	Neutral	Tolerated	benign
8	39624276	Msr1:NM_001113326:exon4:c.A292G:p.T98A	Neutral	Tolerated	ND
8	39624276	Msr1:NM_001113326:exon4:c.A292G:p.T98A	Neutral	Tolerated	ND
9	**95019168**	**Slc9a9:NM_177909:exon10:c.T1153A:p.F385I**	**Deleterious**	**Damaging**	**P.damaging**
11	**29753186**	**Eml6:NM_146016:exon34:c.T4731G:p.D1577E**	**Deleterious**	**Tolerated**	**P. damaging**
11	**60749505**	**Top3a:NM_009410:exon12:c.G1366A:p.V456I**	**Neutral**	**Damaging**	**P. damaging**
11	71169761	Nlrp1b:NM_001162414:exon5:c.T2229G:p.N743K	Neutral	Tolerated	ND
11	71173024	Nlrp1b:NM_001162414:exon3:c.T1932A:p.S644R	Deleterious	Tolerated	ND
11	71181708	Nlrp1b:NM_001162414:exon2:c.G1308T:p.M436I	Neutral	Tolerated	ND
11	71182454	Nlrp1b:NM_001162414:exon2:c.A562G:p.T188A	Neutral	Tolerated	ND
11	71182709	Nlrp1b:NM_001162414:exon2:c.T307C:p.Y103H	Neutral	Tolerated	ND
11	73103887	Itgae:NM_008399:exon2:c.T64A:p.L22M	Neutral	Tolerated	Benign
11	88952543	Scpep1:NM_029023:exon2:c.G88C:p.D30H	Neutral	Tolerated	P. damaging
11	94938066	Col1a1:NM_007742:exon2:c.G233A:p.R78Q	Neutral	Damaging	Unknown
12	11241642	Gen1:NM_177331:exon14:c.G2340C:p.K780N	Neutral	Damaging	Benign
13	24990078	Gpld1:NM_008156:exon26:c.C2485T:p.R829W	Neutral	Damaging	ND
13	48902257	Fam120a:NM_001033268:exon11:c.C1954A:p.Q652K	Neutral	Tolerated	P. damaging
13	89691634	Vcan:NM_019389:exon7:c.C2910G:p.D970E	Neutral	Tolerated	ND
16	32752359	Muc4:NM_080457:exon2:c.T2236C:p.Y746H	Neutral	Tolerated	ND
16	32752369	Muc4:NM_080457:exon2:c.T2246G:p.I749S	Neutral	Tolerated	ND
16	32755171	Muc4:NM_080457:exon4:c.C5044A:p.P1682T	Neutral	Tolerated	ND
17	33999352	H2-K1:NM_001001892:exon3:c.C529T:p.L177F	Neutral	Tolerated	Benign
17	34284327	H2-Aa:NM_010378:exon2:c.G295A:p.V99I	Neutral	Tolerated	Benign
17	34284344	H2-Aa:NM_010378:exon2:c.T278A:p.V93E	Neutral	Tolerated	Benign
17	34284509	H2-Aa:NM_010378:exon2:c.G113C:p.S38T	Neutral	Tolerated	Benign
17	34508315	Btnl6:NM_030747:exon9:c.A1240G:p.N414D	Neutral	Tolerated	ND
17	**36813020**	**H2-M10.6:NM_201611:exon3:c.C375G:p.N125K**	**Deleterious**	**Damaging**	ND
19	**9878228**	**Incenp:NM_016692:exon12:c.G1649A:p.R550H**	**Deleterious**	**Damaging**	ND
19	**21832178**	**Tmem2:NM_001033759:exon17:c.G2921A:p.R974H**	**Deleterious**	**Damaging**	ND
X	**95026054**	**Spin4:NM_178753:exon1:c.C141G:p.H47Q**	**Deleterious**	**Damaging**	**P. damaging**
X	112557019	2010106E10Rik:NM_001168590:exon6:c.T629A:p.L210H	Neutral	Damaging	ND
X	**141723152**	**Irs4:NM_010572:exon1:c.C2047T:p.L683F**	**Neutral**	**Damaging**	**P. damaging**

Dentre os genes potencialmente afetados pelas variantes encontradas na sublinhagem BALB/cICBI (**Tabela 2.4**) destacamos os genes *Lrrc52*, *Rapgef4, Fbxo10, Slc9a9, H2-M10, Incenp, Tmem2* e *Spin4*, cujas variantes foram preditas

por pelo menos dois algoritmos como deletérias. O gene *Lrrc52* codifica para a enzima *leucine rich repeat containing 52,* cuja expressão é restrita nos testículos de animais adultos e controla a ativação de canais de íons KSPER que coordenam a motilidade dos espermatozoides (ZENG et al., 2015). O gene *Slc9a9* codifica uma proteína de membrana que atua como bomba de sódio e cujos camundongos *knock-out* apresentam traços relacionados ao autismo (YANG; FARAONE; ZHANG-JAMES, 2016). Finalmente, o gene *Spin4* codifica para uma proteína da família *spindlin*, cujos membros são proteínas efetoras que podem reconhecer sítios metilados no DNA e estão associados com regiões de repetições no DNA, ativando a transcrição de genes de RNA ribossomal (WANG, 2011).

2.4 Discussão

Conforme descrito no **Capítulo 1**, os dados gerados dos camundongos de linhagens isogênicas também foram utilizados para compor uma base de dados controle de polimorfismos para uma das etapas de filtragem em relação aos camundongos mutantes. Para essa base de dados todas as leituras foram alinhadas no genoma de referência mm9 (NCBI37/mm9), que foi a mesma versão utilizada para o mapeamento genético e para o desenho das sondas de enriquecimento de exons. Essas análises em relação aos camundongos mutantes, que usam as mesmas leituras produzidas descritas no **Capítulo 1** estão descritas no **Capítulo 3**.

Todas as análises descritas nesse capítulo se deram com o alinhamento das leituras com a versão mais recente do genoma de camundongo, chamada de mm10 (GRCm38/mm10), de 2011. Há uma clara tendência para a utilização do genoma mm10 em substituição ao genoma mm9, sendo que a maior parte dos dados disponíveis para comparação utilizam o genoma mm10 como referência. Sendo assim, as coordenadas das sondas de enriquecimento originalmente fornecidas no genoma mm9 foram convertidas para o genoma mm10, utilizando a ferramenta liftOver[22] (Genome Browser, UCSC). As estatísticas de alinhamento obtidas com o genoma mm10 foram muito semelhantes às obtidas com o genoma mm9, descritas no **Capítulo 3**.

---

[22] Genome Browser LiftOver Tool: https://genome.ucsc.edu/cgi-bin/hgLiftOver

Com relação à utilização do LifeScope e o módulo diBayes, eles têm se mostrado de fato superiores inclusive a *pipelines* GATK modernas - pelo menos em dados produzidos pela plataforma SOLiD 5500XL (PRANCKEVIČIENE et al., 2015). De fato, não houve diferenças grandes de performance quando comparamos as estatísticas de alinhamento entre a análise dos dados sequenciados pela JAX e pelo presente estudo (**Tabela 2.1**). A pequena diferença na taxa de alinhamento global das leituras pela *pipeline* do Lifescope pode estar relacionada diretamente mais ao tipo de sequenciamento realizado nas amostras da JAX do que a erros de sequenciamento ou performance de alinhamento. Normalmente o alinhamento de leituras *paired-end* e com leituras maiores, como no caso do sequenciamento da JAX – realizado 2x100 pb – apresenta melhor taxa de alinhamento em comparação a leituras pequenas únicas – como no caso do sequenciamento 1x75 pb (CHHANGAWALA et al., 2015; LANGMEAD, 2011).

A forma com que os SNPs são chamados pode ser determinada por um conjunto de parâmetros agrupados que definem a chamada quanto ao nível de estringência (LIFETECHNOLOGIES, 2011). Dessa forma, foram realizados diversos testes com variação dirigida de cobertura média e estringência na chamada de SNPs, de acordo com os parâmetros do módulo de chamada de SNPS *diBayes,* visando diminuir o número de falsos-positivos, mesmo que isso significasse a perda de eventuais falsos-negativos. Um dos parâmetros mais utilizados, pelo menos em humanos, para a avaliação da acurácia na chamada é a concordância com bancos de dados de polimorfismos (LIU et al., 2012). Quanto maior essa concordância maior é considerada a confiabilidade na chamada de SNPs, ou seja, menor é a taxa de falsos positivos. Utilizando a cobertura média de cerca de 100X e a estringência máxima na chamada de SNPs foi possível aumentar cerca de 35% a concordância com o banco de dados de polimorfismos da amostra C57BL/6ICBI e cerca de 5% da amostra BALB/cICBI (dados não mostrados).

Portanto, partimos do pressuposto que seria permitido uma taxa muito baixa de falsos-positivos para as comparações entre linhagens isogênicas, mesmo que isso implicasse na perda de variantes. Esse tipo de chamada, bem pouco agressiva, mas muito estringente, foi utilizada para a chamada das variantes nas amostras de linhagens isogênicas C57BL/6ICBI e BALB/cICBI, diminuindo ao máximo o número de falso-positivos.

Resumidamente, podemos concluir que o camundongo C57BL/6ICBI pode ser considerado uma sublinhagem com pelo menos 20% a mais de SNPs que a linhagem C57BL/6J. Uma mutação pontual no exon 1 e uma deleção nos exons 7-11 do gene *nnt* no cromossomo 13 é característica da linhagem C57BL/6J (TAKADA et al., 2013) e não está presente na linhagem C57BL/6NJ (NICHOLSON et al., 2010) e nem na sublinhagem C57BL/6/JUnib (CEMIB/UNICAMP) (RONCHI et al., 2013). Dada a origem comum das matrizes do ICB e do Centro Multidisciplinar para Pesquisa em Animais de Laboratório (CEMIB/UNICAMP), cuja colônia também foi estabelecida em 1987 por fundadores fornecidos pela Zentralinstitut für Versuchstierzucht (ZfV) (Hannover, Germany) (RONCHI et al., 2013), é esperado que os camundongos C57BL/6ICBI também não tenham a mutação no gene *nnt.* Foi possível confirmar, considerando a cobertura de leituras em todos os exons do gene *nnt,* que o gene também está intacto na sublinhagem C57BL/6ICBI, o que sugere uma proximidade com a linhagem C57BL/6NJ, proveniente do NIH. De qualquer forma, a caracterização do exoma do camundongo C57BL/6ICBI também pode ser considerada uma fonte de informação genética sobre os camundongos C57BL/6/JUnib.

É muito pouco provável, portanto, que a linhagem C57BL/6ICBI ser originada de cruzamentos indesejados com outras linhagens, devido às proporções de concordância praticamente idênticas com a amostra MMR664 da linhagem C57BL/6J (**Figura 2.5A** e **Figura 2.5B**). A hipótese mais provável é que ela realmente seja originada de matrizes anteriores à separação da atual linhagem C57BL/6J, somado a um posterior isolamento prolongado da colônia. Embora seja controverso, a manutenção de uma linhagem por mais de 20 gerações já a diverge em uma sublinhagem, principalmente pelo efeito de fixação de mutações por deriva genética. A maior proporção de mutações sinônimas também indica a ausência de algum tipo de seleção direcional (**Figura 2.5C**) sendo a deriva genética a principal causa da fixação de mutações nas sublinhagem C57BL/6ICBI.

A sublinhagem BALB/cICBI é de fato muito semelhante em termos de SNPs com a linhagem BALB/cJ, e pode ser considerada nesses termos como uma sublinhagem equivalente ao uso de animais BALB/cJ (**Figura 2.5** e **Tabela 2.2**). Mutações espontâneas impactantes podem ter surgido e se fixado na população, porém no contexto geral de variações as linhagens se comportam de maneira muito parecida (**Figura 2.5A**, **Figura 2.5B** e **Tabela 2.2**). Dessa forma, a lista de variações

encontradas pode ser usada como referência de causa em relação com a eventuais observações de fenótipos anômalos.

Embora a caracterização dos SNPs detectados em relação aos bancos de dados do *Mouse Genomes Project* e o número e tipo de SNVs novos detectados não apontem para nenhum indício óbvio de contaminação com outras linhagens ou seleção positiva nas linhagens do ICB, foram realizadas algumas análises para verificar o perfil mutagênico dos SNVs (**Figura 2.9**). Esse tipo de análise visa identificar uma possível alteração em algum mecanismo celular ou externo que possa aumentar a frequência de mutações germinativas na colônia. Apesar de contarmos apenas com o exoma de um animal de cada linhagem do ICB, foi possível realizar as comparações utilizando os dados de linhagens similares das amostras cedidas pela JAX. Essa caracterização do padrão de mutação foi realizada com a ferramenta *woland*, descrito no Capítulo 3 e com a ferramenta SomaticSignatures (GEHRING et al., 2015; ALEXANDROV et al., 2013) com o conjunto de SNVs novos de cada amostra em relação ao banco de dados de polimorfismos do *Mouse Genomes Project*, que assumimos corresponder a mutações que possam eventualmente ter surgido nas colônias.

Dessa forma, entende-se que a diferença entre as amostras não está associada a fatores como a utilização de diferentes sondas de enriquecimento de exons, plataformas de sequenciamento, alinhamento ou chamada de SNPs, visto que também é observado o mesmo padrão com as amostras MMR-664 e MMR-651. Observamos um elevado padrão de concordância dos SNPs das linhagens BALB/cJ e BALB/cICBI em relação aos bancos de dados (**Figura 2.5**) e também uma proporção elevada de SNPs comuns entre as duas sublinhagens (**Tabela 2.2**). Em relação à sublinhagem C57BL/6ICBI uma proporção menor de SNPs em comum em relação aos bancos de dados (**Figura 2.5**) e também em comparação à linhagem C57BL/6J (**Tabela 2.2**), o que sugere uma maior distância da sublinhagem fornecida pela JAX, reforçando a hipótese de origem da sublinhagem C57BL/6NJ do NIH, devido à presença indicativa do gene *nnt.*

A identificação de SNVs novos que podem ter surgido espontaneamente na linhagem e que foram indiretamente selecionados na manutenção da colônia são extremamente importantes para a caracterização de fenótipos que possam interferir em experimentos (**Figura 2.6** e **Figura 2.7**). Embora os dados das linhagens C57BL/6J e BALB/cJ fornecidas pela JAX sejam animais provenientes de renovação

por matrizes criopreservadas, os camundongos fornecidos pelo ICB são originados de matrizes que já passaram por várias gerações. Por exemplo, alguns pesquisadores relataram que camundongos da linhagem C57BL/6ICBI, por exemplo, se mostraram resistentes ao processo de desenvolvimento de encefalomielite experimental autoimune (EAE), cujo modelo canônico é o camundongo C57BL/6J (MASSIRONI, comunicação pessoal). Notavelmente uma classe de genes relacionados à ligação de alfa-actinina (*actinin binding* - GO:0042805) (**Figura 2.6**) está enriquecida com mutações não sinônimas. Interessantemente, a regulação de citoesqueleto de actina (DAGLEY et al., 2014) é uma das sinalizações afetadas pelo desenvolvimento de EAE em camundongos.

Um dos principais objetivos desse projeto é disponibilizar para a comunidade científica um banco de dados atualizável de polimorfismos associados às linhagens fornecidas pelo biotério de experimentação do ICB. O primeiro passo é fornecer os SNVs novos não sinônimos das linhagens C57BL/6ICBI (**Tabela 2.3** e **Tabela 2.4**) e BALB/cICBI e também a tabela completa dos SNPs de cada uma dessas linhagens, de forma que isso possibilite o acesso a informações sobre o *background* genético das linhagens pelos usuários do biotério.

2.5 Referências

ADZHUBEI, I.; JORDAN, D.; SUNYAEV, S. Predicting Functional Effect of Human Missense Mutations Using PolyPhen-2. **Current Protocols in Human Genetics**, v. 2, 2013.

ALEXANDROV, L.B.; NIK-ZAINAL, S.; WEDGE, D.C.; CAMPBELL, P.J., STRATTON, M.R. Deciphering signatures of mutational processes operative in human cancer. **Cell Reports**, v. 1, n.3, p. 246-259, 2013.

ALTMAN, P.; KATZ, D. **Inbred and Genetically Defined Strains of Laboratory Animals, Part 2, Hamster, Guinea Pig, Rabbit and Chicken**. Bethesda, MD: Federation of American Societies for Experimental Biology, 1979.

BABUSYTE, A. et al. Biogenic amines activate blood leukocytes via trace amine-associated receptors TAAR1 and TAAR2. **Journal of Leukocyte Biology**, v. 93, n. 3, p. 387–394, 2013.

BAILEY, D. Sources of subline divergence and their relative importance for sublines of six major inbred strains of mice. In: **Origins of Inbred Mice**. New York, NY: ed. Morse III HC, 1978.

BANKS, G. et al. Neurobiology of Aging Genetic background in fl uences age-related decline in visual and nonvisual retinal responses , circadian rhythms , and sleep q.

NBA, v. 36, n. 1, p. 380–393, 2015.

BECK, J. A et al. Genealogies of mouse inbred strains. **Nature genetics**, v. 24, n. 1, p. 23–5, jan. 2000.

BELKNAP, J. K. et al. Voluntary consumption of morphine in 15 inbred mouse strains. **Psychopharmacology**, v. 112, n. 2–3, p. 352–358, 1993.

CASELLAS, J. Inbred mouse strains and genetic stability: a review. **Animal**, v. 5, n. 1, p. 1–7, 2011.

CASELLAS, J.; MEDRANO, J. F. Within-Generation Mutation Variance for Litter Size in Inbred Mice. v. 2155, n. August, p. 2147–2155, 2008.

CHEN, E. Y. et al. Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. **BMC bioinformatics**, v. 14, p. 128, 2013.

CHHANGAWALA, S. et al. The impact of read length on quantification of differentially expressed genes and splice junction detection. **Genome Biology**, v. 16, n. 1, p. 131, 2015.

CHOI, Y.; CHAN, A. P. PROVEAN web server: A tool to predict the functional effect of amino acid substitutions and indels. **Bioinformatics**, v. 31, n. 16, p. 2745–2747, 2015.

COLETTI, D. et al. Substrains of inbred mice differ in their physical activity as a behavior. **The Scientific World Journal**, v. 2013, 2013.

DAGLEY, L. F. et al. Discovery of novel disease-specific and membrane-associated candidate markers in a mouse model of multiple sclerosis. **Molecular & cellular proteomics : MCP**, v. 13, n. 3, p. 679–700, 2014.

DANECEK, P. et al. The variant call format and VCFtools. **Bioinformatics (Oxford, England)**, v. 27, n. 15, p. 2156–8, 1 ago. 2011.

DAVIS, W. M.; KING, W. T.; BABBINI, M. Placebo effect of saline on locomotor activity in several strains of mice. **Journal of pharmaceutical sciences**, v. 56, n. 10, p. 1347–1349, out. 1967.

DIWAN, B. A.; BLACKMAN, K. E. Differential susceptibility of 3 sublines of C57BL/6 mice to the induction of colorectal tumors by 1,2-dimethyl-hydrazine. **Cancer Letters**, v. 9, p. 111–115, 1980.

DORAN, A. G. et al. Deep genome sequencing and variation analysis of 13 inbred mouse strains defines candidate phenotypic alleles , private variation and homozygous truncating mutations. **Genome Biology**, p. 1–16, 2016.

EPPIG, J. T. et al. Mouse Genome Informatics (MGI): reflecting on 25 years. **Mammalian Genome**, v. 26, n. 7, p. 272–284, 2015.

FESTING, M. **Origins and characteristics of inbred strains of mice**. 3rd. ed. Oxford: Oxford University Press, 1996.

FLINT, J.; MACKAY, T. F. C. Genetic architecture of quantitative traits in mice , flies , and humans. **Genome research**, v. 19, p. 723–733, 2009.

FONTAINE, D. A.; DAVIS, D. B. Attention to Background Strain Is Essential for Metabolic Research : C57BL / 6 and the International Knockout Mouse Consortium. **Diabetes**, v. 65, n. January, p. 25–33, 2016.

FRAZER, K. A. et al. A sequence-based variation map of 8.27 million SNPs in inbred mouse strains. **Nature**, v. 448, n. 7157, p. 1050–1053, 2007.

FULLER, J. L. Measurement of alcohol preference in genetic experiments. **Journal of comparative and physiological psychology**, v. 57, p. 85–88, fev. 1964.

GEHRING, J.S.; FISCHER, B.; LAWRENCE, M.; HUBER, W. SomaticSignatures: inferring mutational signatures from single-nucleotide variants. **Bioinformatics,** v. 22, n. 31, p. 2673-3675, 2015.

GLANT, T. et al. Variations in Susceptibility to Proteoglycan-Induced Arthritis and Spondylitis Among C3H Substrains of Mice. **Arthirits & rheumatism**, v. 44, n. 3, p. 682–692, 2001.

GOIOS, A. et al. mtDNA phylogeny and evolution of laboratory mouse strains. **Genome research**, v. 17, p. 293–298, 2007.

GREEN, E. L. **Biology of the Laboratory Mouse**. Second Edi ed. New York: Dover Publications, 1966.

GREENFIELD, E. A. **Generating Monoclonal Antibodies**. Second Edi ed. [s.l.] Cold Spring Harbor Laboratory Press, 2014.

GUÉNET, J. et al. **Genetics of the Mouse**. Berlin Heidelberg: Springer-Verlag Berlin Heidelberg, 2015.

HANSEN, C. T. **Catalog of NIH rodents.** Bethesda: National Institutes of Health, 1973.

HESTON, W. E.; VLAHAKIS, G. Alveolar nodules in various combinations of the mouse inbred strains and the different lines of the mammary tumor virus. **International Journal of Cancer**, v. 148, p. 141–148, 1971.

HOAG, W. G. SPONTANEOUS CANCER IN MICE. **Annals of the New York Academy of Sciences**, v. 108, p. 805–831, nov. 1963.

JUSTICE, M. J.; DHILLON, P. Using the mouse to model human disease : increasing validity and reproducibility. **Disease Models & Mechanisms**, v. 9, p. 101–103, 2016.

KEANE, T. M. et al. Mouse genomic variation and its effect on phenotypes and gene regulation. **Nature**, v. 477, n. 7364, p. 289–94, 15 set. 2011.

KERN, M. et al. Biochemical and Biophysical Research Communications C57BL / 6JRj mice are protected against diet induced obesity (DIO). **Biochemical and Biophysical Research Communications**, v. 417, n. 2, p. 717–720, 2012.

KHISTI, R. T. et al. Characterization of the ethanol-deprivation effect in substrains of C57BL / 6 mice. **Alcohol**, v. 40, p. 119–126, 2006.

KRAFT, M. et al. Disruption of the histone acetyltransferase MYST4 leads to a noonan syndrome-like phenotype and hyperactivated MAPK signaling in humans and mice. **Journal of Clinical Investigation**, v. 121, n. 9, p. 3479–3491, 2011.

KUMAR, P.; HENIKOFF, S.; NG, P. C. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. **Nature protocols**, v. 4, n. 8, p. 1073–1082, 2009.

LANGMEAD, B. Alignment with Bowtie. **Current Protocols in Bioinformatics**, p. 1–24, 2011.

LE, A. D. et al. Alcohol consumption by C57BL/6, BALB/c, and DBA/2 mice in a limited access paradigm. **Pharmacology, biochemistry, and behavior**, v. 47, n. 2, p. 375–378, fev. 1994.

LIFETECHNOLOGIES. LifeScope ™ Genomic Analysis Software 2 . 5 Command Shell. n. 4471875, 2011.

LIM, K. H.; FAIRBROTHER, W. G. Spliceman — a computational web server that predicts sequence variations in pre-mRNA splicing. **Bioinformatics**, v. 28, n. 7, p. 1031–1032, 2012.

LIU, Q. et al. Steps to ensure accuracy in genotype and SNP calling from Illumina sequencing data. **BMC genomics**, v. 13 Suppl 8, n. Suppl 8, p. S8, jan. 2012.

LU, Z.; KIPNIS, J. Thrombospondin 1--a key astrocyte-derived neurogenic factor. **The FASEB Journal**, v. 24, n. 6, p. 1925–1934, 2010.

MACARIO, A. J.; STAHL, W.; MILLER, R. Lymphocyte subpopulations and function in chronic murine toxoplasmosis. **Clinical and experimental immunology**, v. 41, n. 3, p. 415–422, set. 1980.

MANDILLO, S. et al. Reliability , robustness , and reproducibility in mouse behavioral phenotyping : a cross-laboratory study. **Physiological Genetics**, v. 34, p. 243–255, 2008.

MATSUO, H. et al. Effects of time of L -ornithine administration on the diurnal rhythms of plasma growth hormone , melatonin , and corticosterone levels in mice. **Chronobiology International**, p. 1–10, 2014.

MEKADA, K. et al. Genetic differences among C57BL/6 substrains. **Experimental animals / Japanese Association for Laboratory Animal Science**, v. 58, n. 2, p. 141–9, abr. 2009.

MONROY-OSTRIA, A. et al. Infection of BALB/c, C57B1/6 mice and F1 hybrid CB6F1 mice with strains of Leishmania mexicana isolated from Mexican patients with localized or diffuse cutaneous leishmaniasis. **Archives of medical research**, v. 25, n. 4, p. 401–406, 1994.

NAGGERT, J. K. et al. Genomic analysis of the C57BL/Ks mouse strain. **Mammalian Genome**, v. 133, p. 131–133, 1995.

NICHOLSON, A. et al. Diet-induced obesity in two C57BL/6 substrains with intact or mutant nicotinamide nucleotide transhydrogenase (Nnt) gene. **Obesity (Silver Spring, Md.)**, v. 18, n. 10, p. 1902–5, 2010.

PAIGEN, B. et al. Atherosclerosis Susceptibility Differences among Progenitors of Recombinant Inbred Strains of Mice. **Arteiosclerosis, thrombosis and vascular biology**, p. 316–323, 1990.

PETTITT, S. J. et al. Agouti C57BL / 6N embryonic stem cells for mouse genetic resources. **Nature Publishing Group**, v. 6, n. 7, p. 493–495, 2009.

PEVZNER, P. A.; TANG, H.; WATERMAN, M. S. An Eulerian path approach to DNA fragment assembly. **PNAS**, v. 98, n. 17, p. 9748–9753, 2001.

POTTER, S. W.; MORRIS, J. E. Development of mouse embryos in hanging drop culture. **The Anatomical record**, v. 211, n. 1, p. 48–56, jan. 1985.

PRANCKEVIČIENE, E. et al. Challenges in exome analysis by LifeScope and its alternative computational pipelines. **BMC Research Notes**, p. 1–15, 2015.

RONCHI, J. A et al. A spontaneous mutation in the nicotinamide nucleotide transhydrogenase gene of C57BL/6J mice results in mitochondrial redox abnormalities. **Free radical biology & medicine**, v. 63, p. 446–56, out. 2013.

ROSCIOLI, T. et al. Mutations in ISPD cause Walker-Warburg syndrome and defective glycosylation of α-dystroglycan. **Nature Genetics**, v. 44, n. 5, p. 581–585, 2012.

ROTH, D. M. et al. Impact of anesthesia on cardiac function during echocardiography in mice. **Heart and Circulatory Phisiology**, v. 92161, p. 2134–2140, 2002.

ROWLAND, E. C.; LOZYKOWSKI, M. G.; MCCORMICK, T. S. Differential cardiac histopathology in inbred mouse strains chronically infected with Trypanosoma cruzi. **The Journal of parasitology**, v. 78, n. 6, p. 1059–1066, dez. 1992.

SALMELA, L. Correction of sequencing errors in a mixed set of reads. **Bioinformatics**, v. 26, n. 10, p. 1284–1290, 2010.

SAZER, S.; DASSO, M. The ran decathlon: multiple roles of Ran. **Journal of cell science**, v. 113 (Pt 7, p. 1111–1118, 2000.

SHAO, H. et al. Genetic architecture of complex traits : Large phenotypic effects and pervasive epistasis. **PNAS**, v. 105, n. 50, p. 19910–19914, 2008.

SIMON, M. M. et al. A comparative phenotypic and genomic analysis of C57BL / 6J and C57BL / 6N mouse strains. **Genome biology**, p. 1–22, 2013.

SIMPSON, E. et al. Genetic variation among 129 substrains and its importance for targeted mutagenesis in mice. **Nature genetics**, v. 16, p. 19–27, 1997.

SITTIG, L. J. et al. Phenotypic instability between the near isogenic substrains BALB/cJ and BALB/cByJ. **Mammalian genome : official journal of the International Mammalian Genome Society**, v. 25, n. 11–12, p. 564–72, dez. 2014.

SNELL, G. D. Studies in Histocompatibility. **The Nobel Lectures**, n. December, 1980.

STIEDL, O. et al. Strain and substrain differences in context- and tone-dependent fear conditioning of inbred mice. **Behavioural Brain Research**, v. 104, p. 1–12, 1999.

TAKADA, T. et al. The ancestor of extant Japanese fancy mice contributed to the mosaic genomes of classical inbred strains. **Genome research**, v. 23, n. 8, p. 1329–38, ago. 2013.

TEUSCHER, C. et al. A Common Immunoregulatory Locus Controls Susceptibility to Actively Induced Experimental Allergic Encephalomyelitis and Experimental Allergic Orchitis in BALB/c Mice. **The Journal of Immunology**, p. 2751–2756, 1998.

THOMPSON, W. R. The inheritance of behaviour: behavioural differences in fifteen mouse strains. **Canadian Journal of Psychology**, v. 7, p. 145–155, 1953.

TOYE, A. A. et al. A genetic and physiological study of impaired glucose homeostasis control in C57BL/6J mice. **Diabetologia**, v. 48, n. 4, p. 675–686, 2005.

ULMASOV, B. et al. Differences in the Degree of Cerulein-Induced Chronic Pancreatitis in C57BL / 6 Mouse Substrains Lead to New Insights in Identi fi cation of Potential Risk Factors in the Development of Chronic Pancreatitis. **The American Journal of Pathology**, v. 183, n. 3, p. 692–708, 2013.

WANG, K.; LI, M.; HAKONARSON, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. **Nucleic acids research**, v. 38, n. 16, p. e164, set. 2010.

WOTJAK, C. T. C57BLack / BOX ? The importance of exact mouse strain nomenclature. **Trends in Genetics**, v. 19, n. 4, p. 183–184, 2003.

WYSOCKI, C. J.; WHITNEY, G.; TUCKER, D. Specific anosmia in the laboratory mouse. **Behavior genetics**, v. 7, n. 2, p. 171–188, mar. 1977.

YALCIN, B. et al. Next-generation sequencing of experimental mouse strains. **Mammalian genome : official journal of the International Mammalian Genome Society**, v. 23, n. 9–10, p. 490–8, 2012.

YANG, Y. et al. Spontaneous deletion of epilepsy gene orthologs in a mutant mouse with a low electroconvulsive threshold. **Human Molecular Genetics**, v. 12, n. 9, p. 975–984, 2003.

YANG, L.; FARAONE, S.V.; ZHANG-JAMES, Y. Autism spectrum disorder traits in Slc9a9 knock-out mice. Am J Med Genet B Neuropsychiatr Genet, v. 3, p. 363-376, 2016.

YOUNG, L. J.; HAMMOCK, E. A. D. On switches and knobs, microsatellites and monogamy. **Trends in Genetics**, v. 23, n. 5, p. 209–212, 2007.

ZURITA, E. et al. Genetic polymorphisms among C57BL/6 mouse inbred strains. **Transgenic research**, v. 20, n. 3, p. 481–9, jun. 2011.

ZENG, X.H.; YANG, C. XIA, X.M.; LIU, M.; LINGLE, C.J. SLO3 auxiliary subunit LRRC52 controls gating of sperm KSPER currents and is critical for normal fertility. **PNAS**, v.8, n.122, p. 2599-2604, 2015.

WANG, W.; CHEN, Z.; MAO, Z.; ZHANG, H.; DING, X.; CHEN, S.; ZHANG, X.; XU, R.; ZHU, B.; Nucleolar protein Spindlin1 recognizes H3K4 methylation and stimulates the expression of rRNA genes. **EMBO Rep,** v. 11, n. 12, p. 1160-1166, 2011.

CAPÍTULO 3 - SEQUENCIAMENTO DOS CAMUNDONGOS MUTANTES POR MUTAGÊNESE INDUZIDA POR ENU.

CAPÍTULO 3 - SEQUENCIAMENTO DOS CAMUNDONGOS MUTANTES POR MUTAGÊNESE INDUZIDA POR ENU.

3.1 Introdução

3.1.1 Genética direta ou forward genetics

Camundongos mutantes constituem uma importante ferramenta na identificação de fatores genéticos responsáveis diretamente por fenótipos complexos. A geração de camundongos nocautes ou *knockouts* (**Figura 3.1**) tem sido uma abordagem frequente para a inativação de genes seguida pela caracterização funcional dos mesmos (ANTONARAKIS; BECKMANN, 2006; GONDO, 2010). Porém, doenças humanas em sua maioria são causadas por mutações que alteram funcionalmente os produtos gênicos ou sua regulação ao invés de simplesmente desativá-los (COOPER et al., 2011). Portanto, o estudo de modelos animais obtidos por mutagênese aleatória ou dirigida (CRISPR-Cas9) (**Figura 3.1**) pode contribuir tanto com a elucidação da função dos genes quanto com a caracterização de doenças genéticas humanas (OLIVER; DAVIES, 2012).

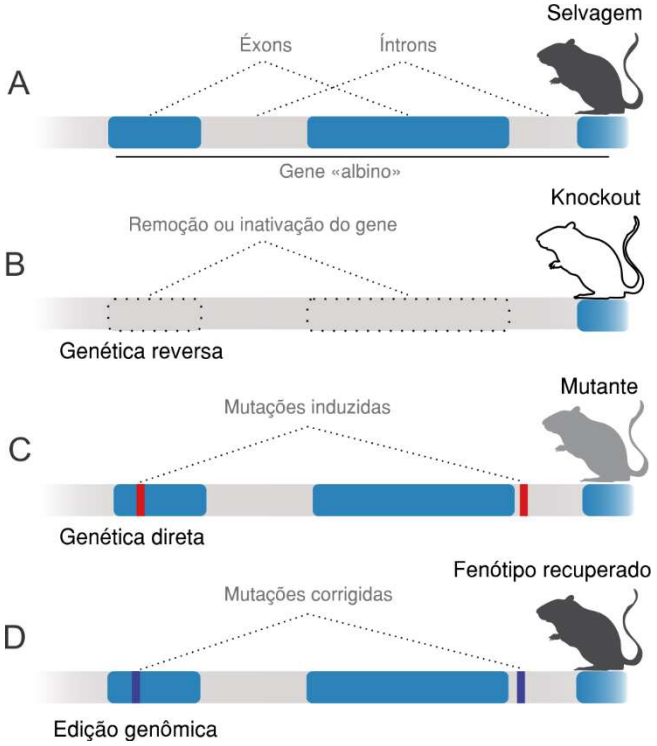

Figura 3.1 Abordagens para a compreensão da função dos genes usando modelos. A função de um hipotético gene "albino", envolvido na cor da pelagem em camundongos (**A**), pode ser investigada a partir da deleção ou inativação do gene (um *knockout) e* observação direta do fenótipo (**B**). Mutantes induzidos podem revelar detalhes mais finos sobre a função do gene (**C**) por genética direta, já que são observadas mutações pontuais que mimetizam alelos naturais. Recentemente, o desenvolvimento de ferramentas de edição genômica, como CRISPR-Cas, possibilitam a correção direcionada de mutações pontuais, restaurando o fenótipo selvagem.

A indução de mutações aleatórias através do uso de agentes mutagênicos é uma forma de acelerar e aumentar a diversidade de fenótipos em contrapartida à observação de fenótipos gerados a partir de mutações espontâneas (BULL et al., 2013). Vários agentes mutagênicos podem ser usados em varreduras de genética direta, desde agentes mutagênicos físicos como a radiação gama como agentes químicos como o etilmetanosulfonato (EMS) e N-etil-N-nitrosuréia (ENU) ou agentes mutagênicos endógenos como elementos de transposição (FARRELL et al., 2014).

A genética direta possui vantagens em relação à genética reversa, principalmente por ser guiada por variações fenotípicas e de certa forma por mimetizar variações pontuais que são mais prováveis de serem encontradas na natureza, como em síndromes mendelianas clássicas (BULL et al., 2013). É um tipo de abordagem que permite a seleção de fenótipos interessantes sem que haja nenhuma hipótese relacionada à base genética envolvida. Além disso, é um dos melhores métodos para a determinação da função de um gene, ao observar os fenótipos de organismos individuais que possuem uma mutação no próprio gene (MORESCO; LI; BEUTLER, 2013).

Uma das vantagens do uso de varreduras de genética direta é a liberdade em relação ao conhecimento prévio da sequência de DNA mutada. Essa vantagem se torna uma desvantagem devido à dificuldade dos procedimentos de localização ou mapeamento da mutação (BULL et al., 2013). Durante muito tempo essas abordagens não triviais consistiram em utilizar marcadores moleculares, como microssatélites e RFLPs, para mapear as mutações no genoma (JUSTICE et al., 1999). Através do uso de microssatélites, a mutação era localizada em um mapa de recombinação e somente a região-alvo que segregava com o fenótipo era identificada através de sequenciamento (HERRON et al., 2002). Esse tipo de estudo é chamado de triagem de mutagênese e pode ser usado para caracterizar genes específicos que afetam uma determinada função ou fenótipo complexo (CORDES, 2005; JUSTICE et al., 1999). O sequenciamento completo do genoma de camundongo facilitou muito a procura e seleção de candidatos, e ficou ainda mais rápido e menos custoso com o advento do uso de tecnologias de sequenciamento de nova geração (NGS) na procura pelos candidatos (ARNOLD et al., 2011; BOLES et al., 2009; BULL et al., 2013; FAIRFIELD et al., 2011).

3.1.2 Mutagênese por ENU e varredura de mutantes

O agente mutagênico N-etil-N-nitrosouréia ou ENU (**Figura 3.2**) tem sido usado em larga escala desde a década de 1990 para o estabelecimento de modelos de doenças humanas em camundongos (GUÉNET, 2005) e peixe-zebra (PATTON; ZON, 2001). O ENU pode introduzir mutações pontuais randômicas em virtualmente qualquer região do genoma de células tronco de espermatogônias a uma frequência de ~150×10^{-5} por locus (SHIBUYA; MORIMOTO, 1993). A mutagênese ocorre principalmente devido a trocas de uma única base pela alquilação direta dos ácidos nucleicos. O grupo etil do ENU pode ser transferido a um oxigênio - preferencialmente em O4-timina, O2-timina ou O2-citosina ou a um nitrogênio nas bases do DNA (KOELSCH; KINDLER-RÖHRBORN, 2009) (**Figura 3.2**). Durante a replicação do DNA, essas bases contendo aduto etil podem causar erro no pareamento das bases, introduzindo principalmente mutações pontuais (substituição de base), mesmo que eventualmente possam ocorrer também pequenas deleções e inserções (TAKAHASI; SAKURABA; GONDO, 2007). As mutações mais comumente observadas são transversões TA:AT ou transições TA:CG (JUSTICE et al., 1999; SHIBUYA; MORIMOTO, 1993).

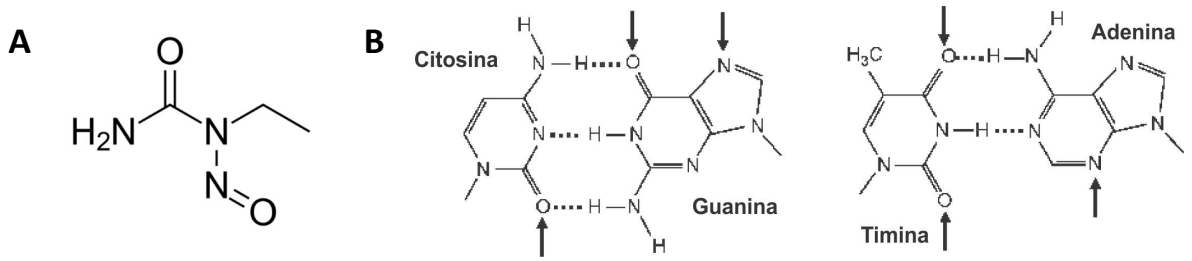

Figura 3.2 O agente mutagênico N-etil-N-nitrosouréia (ENU). Estrutura química do ENU (fórmula química $C_3H_7N_3O_2$) (**A**) e átomos alvos nas bases do DNA indicados pelas setas (adaptado de KOELSCH; KINDLER-RÖHRBORN, 2009).

Várias estratégias foram desenvolvidas para a utilização de ENU em genética direta, basicamente envolvendo o tratamento de espermatogônias com doses controladas de ENU causando mutações pontuais em células germinativas. Os machos tratados com ENU passam por um período de descanso e são então cruzados com fêmeas não tratadas por uma ou duas gerações originando proles mutantes (**Figura 3.3**). Alterações fenotípicas visíveis, de herança dominante ou recessiva, são recuperadas para estabelecimento de testes para a seleção e

caracterização dos candidatos para a mutação causal do fenótipo observado (MASSIRONI et al., 2006).

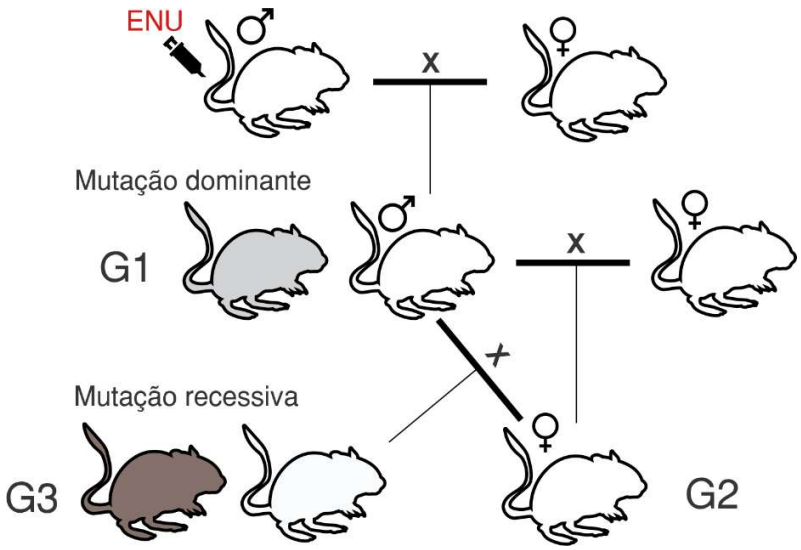

Figura 3.3 Estratégia de varredura de genética direta de mutagênese por ENU. Camundongos BALB/c tratados com ENU são acasalados com várias fêmeas BALB/c e sua progênie (G1) é observada para detecção de fenótipos de herança dominante. Os machos G1 são então acasalados com fêmeas BALB/c produzindo a geração G2. O acasalamento de fêmeas G2 com seu pai G1 dá origem a progênie G3, cujos fenótipos mutantes são de herança recessiva. Adaptado de (MASSIRONI et al., 2006).

A natureza aleatória desse tipo de mutagênese pode gerar fenótipos semelhantes a doenças humanas e acentuar determinadas vias de sinalização (SIMON et al., 2015). Isso significa que a função de novos genes pode ser descoberta sem que nenhuma anotação esteja disponível, já que nenhum gene em particular é um alvo específico da técnica.

Várias linhagens podem ser utilizadas, mas a linhagem mais utilizada para varreduras de mutagênese por ENU é a C57BL6/J, por causa principalmente da fertilidade elevada mesmo após o tratamento por altas doses de ENU (JUSTICE et al., 1999). Empregar a mesma linhagem em uma varredura significa reduzir a variação no *background* genético, facilitando a detecção de fenótipos variantes. Empregar mais de uma linhagem nas varreduras dificulta a distinção entre um provável fenótipo gerado pela mutagênese e entre fenótipos gerados pela mistura entre linhagens diferentes (GONDO, 2010).

A estratégia de indução de mutações pontuais por ENU aliada a uma seleção fenotípica consistente constitui uma ferramenta poderosa para identificação de mutações responsáveis por fenótipos complexos (ARNOLD et al., 2012). Considerando os estudos que correlacionaram as mutações identificadas aos

fenótipos observados, cerca de 75% dos fenótipos observados por mutagênese por ENU foram causadas por SNVs (*single nucleotide variants*) em exons (JUSTICE et al., 1999). Aproximadamente 63% destas são mutações do tipo *missense*, 26% causaram *splicing* anormal, 10% são mutações *nonsense* e aproximadamente 1% causaram mutações que converte um códon de parada em códon que codifica um aminoácido (NOVEROSKE; WEBER; JUSTICE, 2000).

3.1.3 Mapeamento das mutações

Um dos maiores gargalos para o estudo de mutantes induzidos por ENU é a identificação da mutação pelo mapeamento genético fino utilizando marcadores polimórficos seguido pelo sequenciamento (pelo método Sanger) de exon por exon de um ou mais genes na região mapeada. De maneira clássica, a localização – ou mapeamento – de *loci* afetados por uma mutação é baseada na detecção do desequilíbrio de ligação depois de uma varredura genômica com marcadores moleculares. Esse mapeamento requer uma série de cruzamentos, onde o fenótipo deverá segregar com o maior número possível de polimorfismos detectáveis (BEIER; HERRON, 2004). Isso significa, por definição utilizar cruzamentos linhagens diferentes do *background* utilizado nos mutantes. A escolha de linhagens isogênicas distantes da linhagem isogênica onde a mutação foi induzida melhora a distribuição e densidade dos polimorfismos de todos os tipos ao longo do genoma (BECK et al., 2000).

No caso de fenótipos cuja herança é recessiva, os cruzamentos para mapeamento das mutações consistem basicamente em cruzamentos entre mutantes homozigotos com a linhagem distante. A progênie é então cruzada com mutantes heterozigotos, sendo realizados *intercrossing* ou *bakcrossing* com mutantes da linhagem de origem da mutação heterozigotos, dando origem a geração F2. Todos os camundongos que apresentarem fenótipo em F2 são considerados para o mapeamento com marcadores moleculares (MASSIRONI et al., 2006). Esse mapeamento consiste no uso de mais de 50 marcadores moleculares microssatélites polimórficos entre as duas linhagens utilizadas nos cruzamentos de mapeamento. Pelo menos 50 amostras de DNA da geração F2 que apresentam fenótipo mutante são utilizados para a primeira varredura, que visa a localização cromossomal das mutações. Uma vez que é encontrado o desequilíbrio de ligação mais marcadores do mesmo cromossomo são utilizados para confirmar e refinar a localizar a mutação – procedimento que é conhecido

como mapeamento de alta-resolução (MASSIRONI et al., 2006). Após o mapeamento fino é utilizada a clonagem posicional de vários segmentos de DNA-alvo para sequenciamento Sanger. Dependendo do tamanho da região mapeada, que pode variar de poucas kilobases até o nível de megabases – são utilizados vários pares de oligonucleotídeos visando amplificar subdivisões da região-alvo em produtos de amplificação de até 1000 kbp, que são sequenciados individualmente (KILE; HILTON, 2005).

Além do mapeamento tradicional com microssatélites, painéis de hibridização (microarranjos de DNA) têm sido utilizados para varreduras em conjuntos de SNPs de exomas, apesar de serem restritos (MORAN et al., 2006; SUN et al., 2012). Esse processo teve muito sucesso no passado mas é lento e requer trabalho intensivo e também envolve assumir correlações diretas entre a causa genética e o fenótipo (MORESCO; LI; BEUTLER, 2013).

Com o uso de técnicas modernas de sequenciamento esse gargalo diminuiu consideravelmente, de forma a aumentar a chance, a rapidez e a confiabilidade na identificação das mutações candidatas. Considerando que aproximadamente 40-75% das mutações induzidas por ENU e causadoras de fenótipo foram identificadas em exons, alguns grupos de pesquisa vêm identificando mutações induzidas por ENU utilizando análises por sequenciamento do exoma (WES, do inglês, *Whole Exome Sequencing*) (ENDERS et al., 2012; FAIRFIELD et al., 2011, 2015; SUN et al., 2012). Ao considerar o sequenciamento completo do genoma (WGS) cerca de 90 Gb de dados de sequenciamento são necessários para alcançar cerca de 30X de cobertura média, enquanto apenas 3 Gb são necessários para alcançar quase 75X, quase o dobro da cobertura média, em se tratando de exomas.

As tecnologias de NGS e também os algoritmos de identificação de SNVs possuem uma taxa de erro e de falsos-positivos elevada em comparação com o sequenciamento Sanger. Além disso, exons não anotados, promotores e *enhancers* não são consideradas em WES, apesar do reconhecimento cada vez maior dessas regiões em doenças e manifestações fenotípicas (ZHANG; LUPSKI, 2015). Variantes estruturais como inserções, deleções e translocações também não são identificadas com precisão em WES (WARR et al., 2015), embora também possam ser originadas por processos mutagênicos induzidos ou espontâneos (MORESCO; LI; BEUTLER, 2013).

3.1.4 Características gerais dos mutantes BALB/c selecionados induzidos por ENU

A estratégia de mutagênese por ENU foi o objetivo do projeto "O CAMUNDONGO COMO ORGANISMO MODELO – MAPEAMENTO DE MUTAÇÕES INDUZIDAS POR ETIL-NITROSOURÉIA" apoiado pela FAPESP (Projetos 00/06963-5 e 03/04531-9) coordenado pela Dra. Silvia Massironi. O projeto deu origem a 11 camundongos mutantes com fenótipos diversos (MASSIRONI et al., 2006), gerados pelo esquema de varredura (**Figura 3.3**) em camundongos BALB/c e mantidos em *background* BALB/c. O mapeamento dessas mutações foi iniciado pelos métodos tradicionais utilizando-se microssatélites (RHODES et al., 1998) por mapeamento meiótico convencional e foram identificados, no mínimo, os cromossomos onde as mutações se localizavam. Em alguns casos, foi possível realizar a identificação dos genes candidatos posicionais determinados por sequenciamento pelo método de Sanger. Esse método foi utilizado, por exemplo, para o sequenciamento da mutação Anêmico, cujo gene causal codifica a hemoglobina, que tem poucos exons. Para outras mutações os tamanhos do genoma contendo genes candidatos são muito maiores, o que dificulta e encarece essa estratégia (JUSTICE et al., 1999).

A **Tabela 3.1** apresenta dados de 7 mutantes, cujas mutações não foram identificadas pelos métodos convencionais, e seus controles, estão sendo estudados neste projeto. Observa-se que os nomes dos mutantes se referem ao principal fenótipo observado, uma vez que o gene mutado ainda não era conhecido.

Tabela 3.1 Camundongos mutantes e linhagens selvagens selecionados para o projeto.

Mutação	Tipo de Herança	Background	Principal fenótipo observado	Cromossomo[a]	Posição (cM) [a]	Grupo Colaborador
ataxico 1	recessivo	BALB/c	Roda em círculos	10	25-40	ICB/UNIFESP
bate palmas	recessivo	BALB/c	Não sustenta as pernas traseira na natação	15	19-37	ICB/FMVZ
careca	recessivo	BALB/c	Pelagem rala por toda a vida	7	70-78	ICB/FMVZ/UFMG
cruza pernas	recessivo	BALB/c	Cruza pernas traseira quando segurado pela cauda	11	27-58	ICB/FMVZ
equilíbrio	recessivo	BALB/c	Não se equilibra no rota rod. Desorganização da camada de Purkinje	17	1-10	ICB/UNIFESP e FMVZ
fraqueza	recessivo	BALB/c	Perda progressiva da coordenação motora	1	11-81	ICB/UFMG
Sacudidor	dominante	BALB/c	Sacode a cabeça continuamente	15	32-57	ICB/UFMG
C57BL/6	-	-	-	-	-	Biotério Imunologia ICB
BALB/c	-	-	-	-	-	Biotério Imunologia ICB

[a] De acordo com mapeamento por microssatélites realizado pelo grupo da Dra. Silvia Massironi.

3.1.5. Resumo das características individuais dos mutantes

Atáxico-1: Os camundongos mutantes *Atáxico-1* (*atxrec1*) possuem como traço fenotípico principal o movimento estereotipado de rodar em círculos. Esse fenótipo possui herança mendeliana recessiva, cuja mutação foi mapeada por microssatélites no cromossomo 10 entre 25 e 40 cM. Além de movimentos estereotipados, os camundongos *atxrec1* apresentam tremores, disbasia, movimento involuntário da cabeça, ausência de respostas de reflexo variadas, discreto atraso na abertura das pálpebras e infertilidade em machos da ordem de 95%. Esses animais são estudados pelo grupo da Profa. Marimélia Porcionatto, do Departamento de Bioquímica, da Escola Paulista de Medicina da UNIFESP (Massironi; Mori, comunicação pessoal).

Bate palmas: O mutante batepalmas (*bapa)* (**Figura 3.4B**) é mantido em *background* BALB/c e apresenta alterações fenotípicas de herança recessiva marcantes como: movimento repetitivo das patas dianteiras quando suspenso pela cauda (MASSIRONI et al., 2006); incapacidade de nadar quando colocado em água e deficiência em testes comportamentais de orientação espacial (Massironi; Mori, comunicação pessoal). Análises comportamentais recentes indicam uma possível deficiência motora, com menores níveis de ansiedade sem perda de memória (OLIVEIRA, 2017). Atualmente o camundongo *bapa* é modelo de estudo do grupo da Profa. Cláudia Mori, do Departamento de Patologia da Faculdade de Medicina Veterinária e Zootecnia da USP.

Careca: A mutação careca (*carc*) induz anomalias no pelo da primeira pelagem – que é esparsa durante toda a vida do camundongo, especialmente em redor dos olhos, membros e parte ventral. Logo após o desmame os camundongos perdem toda sua pelagem, depois disso a pelagem reaparece sempre com falhas. Análises histológicas mostram um aumento da camada de pele devido ao alto número de folículos em fase anágena (MASSIRONI et al., 2006). A mutação *carc* foi mapeada no cromossomo 7, provavelmente entre 70 e 78 cM – porém preferimos considerar que a mutação estaria restrita a todo o cromossomo 7. Esse mutante está sendo estudado pelo grupo da Profa. Ana Lúcia Godard do Depto. de Biologia Geral, do Instituto de Ciências Biológicas, da UFMG.

Cruza pernas: Camundongos cruza pernas (*crup*) (**Figura 3.4D**) foram identificados pelo fenótipo característico do cruzamento das patas traseiras entre si

quando suspensos pela cauda (MASSIRONI et al., 2006). Esse fenótipo é acentuado com o envelhecimento, sendo que após seis a oito meses de idade os camundongos se enrolam, juntando as patas traseiras e dianteiras. É uma mutação recessiva e foi localizada, através do mapeamento por microssatélites no cromossomo 11 entre 27 e 58 cM (MASSIRONI, comunicação pessoal). É estudado pelo grupo da Profa. Cláudia Mori, da FMVZ-USP.

Equilíbrio: A mutação *eqlb* (**Figura 3.4C**) é caracterizada pelo fenótipo de coordenação motora anormal, verificado pela incapacidade dos camundongos *eqlb* se manterem no aparato rota-rod, sugerindo algum defeito no controle do equilíbrio. Os camundongos *eqlb* apresentam proliferação de precursores neuronais no cerebelo durante o desenvolvimento pós-natal acompanhada da desorganização da camada de Purkinje (Mazzonetto, dados não publicados). Esse mutante é estudado pelo grupo da Profa. Marimélia Porcionatto, da UNIFESP e pelo grupo da Profa. Cláudia Mori, da FMVZ-USP.

Fraqueza: O mutante fraqueza (*frqz*) (**Figura 3.4A**) apresenta perda progressiva da coordenação motora, que fica evidente à partir da segunda semana de vida e raramente sobrevivem à quarta semana de vida (Massironi, comunicação pessoal). Cortes histológicos dos músculos mostraram núcleos corados, indicando processo inflamatório e morte celular. Nervos e gânglios intra-espinais apresentam degeneração acompanhados por neurônios medulares anucleados, indicando também morte celular (SILVA SILVA; GODARD, 2012). A mutação foi mapeada previamente na região entre 11 e 81 cM do cromossomo 1 (MASSIRONI et al., 2006) e esse mutante está sendo estudado pelo grupo da Profa. Ana Lúcia Godard da UFMG.

Sacudidor: Única mutação dominante do grupo estudado, camundongos *Sacc* apresentam contínuo movimento de balanço da cabeça mas sem apresentar o comportamento de andar em círculos. A intensidade do movimento da cabeça aumenta quando o animal é perturbado. Não são surdos - quando estimulados por um pulso agudo sonoro geram o reflexo de Preyer (MASSIRONI et al., 2006). A mutação foi mapeada por microssatélites no cromossomo 15 entre as regiões 32 e 57 cM. Também estão sendo estudados pelo grupo da Profa. Ana Lúcia Godard da UFMG.

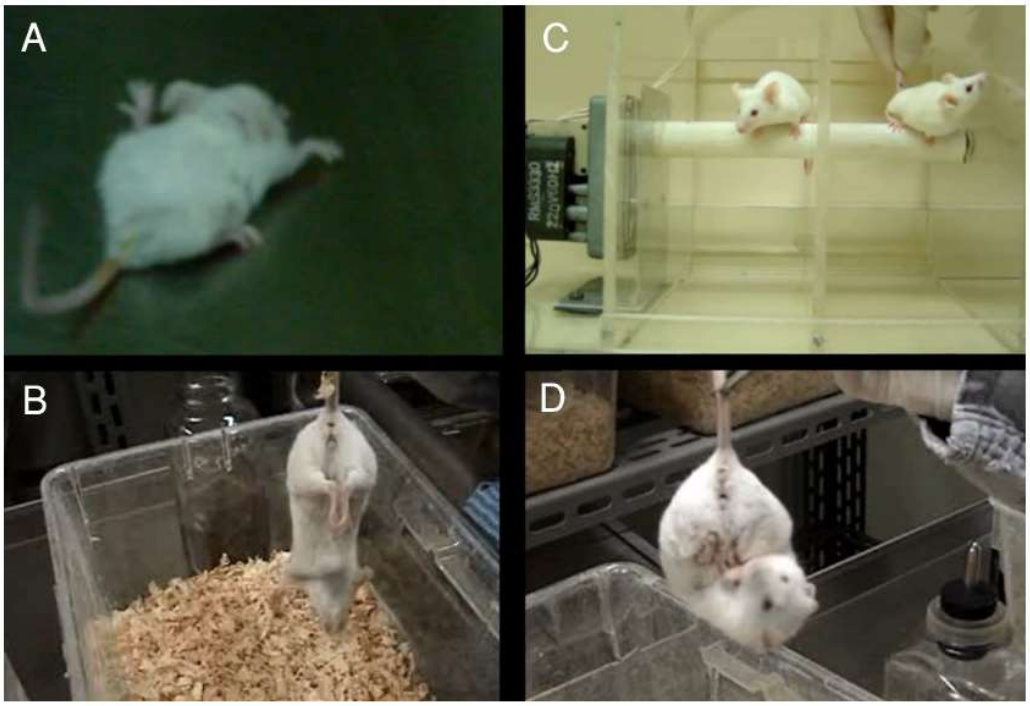

Figura 3.4 Fenótipos de alguns mutantes induzidos por ENU. O camundongo **fraqueza** (A) apresenta severo comprometimento motor pós-natal enquanto o camundongo **batepalmas** (B) realiza seu movimento estereotipado com os membros inferiores e posteriores. O camundongo **equilíbrio** (C) não se sustenta no rotarod enquanto o selvagem a sua esquerda não apresenta nenhuma dificuldade no teste. O camundongo **cruzapernas** (D), quando erguido pela cauda, se curva de maneira ventral cruzando os membros inferiores e superiores. Fotografias cedidas pela Dra. Massironi.

3.2 Material e Métodos

A metodologia completa utilizada para seleção das mutações causativas induzidas por ENU está resumida na **Figura 3.5** e todos os passos para análise dos dados são detalhados a seguir.

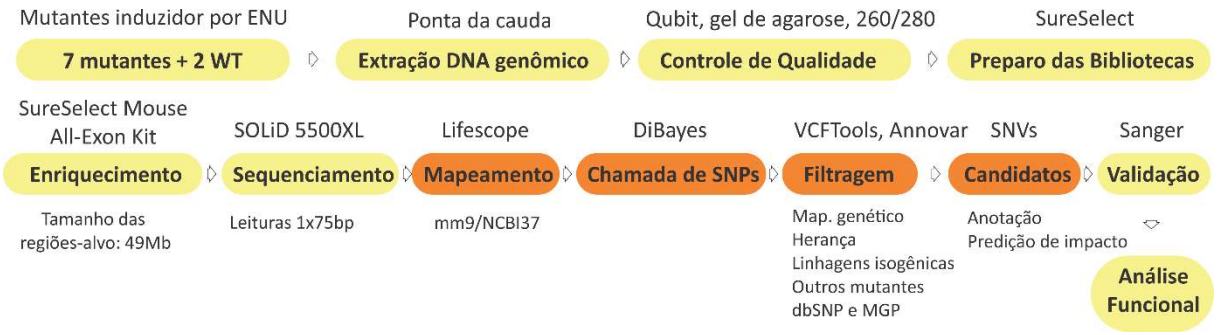

Figura 3.5 Resumo da metodologia de sequenciamento NGS e filtragem de SNVs candidatos. As caixas claras indicam experimentos em bancada enquanto as caixas escuras indicam etapas de experimentos *in silico.*

3.2.1 Mapeamento das leituras e chamada de SNPs

O mapeamento em *colorspace* das leituras das 7 amostras de mutantes e das duas linhagens isogênicas foi realizado pela ferramenta LifeScope 2.1 (ThermoFisher Scientific) utilizando o genoma de referência mm9 (NCBI37/mm9) com os módulos presentes na *pipeline* de análise de ressequenciamento de regiões-alvo - *Targeted Resequencing Analyses*. Como parâmetros-base para essa análise foram considerados apenas alinhamentos primários e o mapeamento no exoma foi considerado apenas se pelo menos uma base tiver sido alinhada nas regiões-alvo. As coordenadas das sondas de enriquecimento para as regiões-alvo foram obtidas através do fornecedor do kit SureSelect Mouse All-Exon para o genoma mm9. As estatísticas de mapeamento foram obtidas pelo módulo *BamStats* com coordenadas cobertas (*covered*) das sondas de enriquecimento do kit SureSelect Mouse All-Exon kit (S0276129), originalmente desenhadas no mesmo genoma usado como referência.

A chamada de SNPs foi realizada pelo algoritmo diBayes, componente do LifeScope, utilizando a estringência padrão média para chamada dos SNPs. Apenas SNPs com proporção de leituras válidas - *mapping QV* maior que 8 – por total de leituras maior que 65% e bases com valores mínimos de qualidade de igual a 26 foram considerados, sem requerimento do SNP ser identificado em ambas as fitas. O número mínimo de posição de início de leituras mapeadas utilizado foi maior que 2 para variantes em heterozigose e homozigose, com proporção mínima igual a 0,15 para o alelo alterado. Esse conjunto de parâmetros forma o tipo de chamada com estringência média, que equilibra o número de SNPs com um número médio de falsos-positivos. A lista gerada pela chamada de SNPs foi chamada de lista de SNPs brutos.

Considerando que a ferramenta diBayes utiliza uma abordagem bayesiana para coberturas menores que 30% e uma abordagem frequentista para coberturas maiores que 30%, temos que a proporção alélica para a classificação como heterozigoto situa-se geralmente entre 15% e 65% dos reads alterados. De maneira similar ao algoritmo bayesiano, o algoritmo frequentista considerado fontes de erro no cálculo da probabilidade para uma dada chamada heterozigota, se a probabilidade excede um valor limite. Estima-se que a sensitividade, para a razão alélica de 70:30 é de 95,73% e a especificidade é de 100%, com taxa de chamada de falsos positivos de $6,4 \times 10^{-5}$ (TANG et al., 2008).

3.2.2 Filtragem e seleção de candidatos

As análises de comparação e filtragem das variantes foram realizadas através da ferramenta VCFTools (DANECEK et al., 2011). Os filtros aplicados em cada conjunto de SNPs brutos correspondente a cada mutante foram: filtro de mapeamento, onde somente as variantes encontradas dentro do intervalo de mapeamento cromossômico prévio realizado foi considerado; filtro de zigosidade – que considera SNPs homozigotos somente para mutantes com herança recessiva; filtro de exclusividade em relação a todos os SNPs brutos dos controles BALB/cICBI, C57BL/6ICBI, mutantes não relacionados, dbSNP132 e SNPs listados no Mouse Genomes Project - REL-1211 e REL-1505, realizado de forma posicional; filtro de anotação, onde somente SNVs em regiões exônicas com mutações não sinônimas e em regiões de *splicing* (até dois nucleotídeos da borda do exon) eram selecionados.

Os filtros de mapeamento, zigosidade e exclusividade foram aplicados utilizando parâmetros da ferramenta VCFTools (DANECEK et al., 2011). Algumas etapas do filtro de exclusividade e todas as etapas do filtro de anotação foram realizadas pela ferramenta ANNOVAR (WANG; LI; HAKONARSON, 2010) utilizando o banco de dados UCSC RefSeq/mm9.

3.2.3 Predição de impacto das mutações candidatas

As análises de impacto das variantes foram realizadas através das ferramentas PolyPhen2 (ADZHUBEI; JORDAN; SUNYAEV, 2013), SIFT (KUMAR; HENIKOFF; NG, 2009), PROVEAN (CHOI; CHAN, 2015) e SpliceMan (LIM; FAIRBROTHER, 2012). As ferramentas SIFT e PROVEAN foram utilizadas através de implementações *web* na plataforma Variant Effect Predictor (VEP-ENSEMBL)[23], como valores de *cutoff* de predição deletéria iguais a -2.5 e 0.05, respectivamente. Para a análise pela ferramenta PolyPhen-2 foi utilizada uma implementação local da ferramenta em um servidor dedicado com a configuração dos bancos de dados para sequências de camundongos[24], sendo as coordenadas anotadas em função dos indicadores UNIPROT.

3.2.4 Validação das mutações por sequenciamento Sanger

---

[23] Variant Effect Predictor (VEP-ENSEMBL): http://www.ensembl.org/info/docs/tools/vep/index.html
[24] Polyphen-2 Standalone configuration protocol: http://genetics.bwh.harvard.edu/pph2/dokuwiki/docs

A validação dos SNVs candidatos foi realizada por amplificação por PCR de regiões de 300-800 pb ao redor de cada SNV detectado com pares de oligonucleotídeos específicos desenhados para flanquear as mutações (**Tabela 3.2**). Para cada ensaio de validação foram utilizadas amostras de DNA genômico de camundongos, cujos exomas não foram sequenciados previamente, consistindo de quatro amostras controle (C57BL/6ICBI, BALB/cICBI, A/J e um outro mutante não relacionado) e uma amostra do mutante cujo SNV candidato foi detectado. Os produtos de PCR foram avaliados quanto ao tamanho esperado e purificados através de eletroforese E-Gel 2% (ThermoFisher Scientific), submetidos à reação de sequenciamento com o uso do kit BigDye 3.1 (ThermoFisher Scientific) com a adição dos dois oligonucleotídeos *forward* e *reverse* em reações separadas e então enviados para sequenciamento Sanger na plataforma ABI 3130XL (ThermoFisher Scientific) disponibilizada pela Central Analítica do Instituto de Química (IQ-USP).

Tabela 3.2 Sequência dos oligonucleotídeos utilizados para validação dos SNVs por sequenciamento Sanger

Mutante	Gene candidato	Oligonucleotídeo (5'-3')
batepalmas	Mll2_F	TGCTAGCAAACATCGGACTG
	Mll2_R	TGGGTCCCTTCCATCACTTA
careca	Cyp2b9_F	GGGATGTGACAATCCAAAGG
	Cyp2b9_R	CCCAATTAGCGGGCTAAGAAGTAG
careca	Vwa3a_F	GAGTGACATTGATATGTTCACTGG
	Vwa3a_R	GTTCTATCCATAGCCTACTGAAGG
careca	Cars_F	CATACTGAAGAGCAGACTCCATAGTG
	Cars_R	AGCCTCAGGTGCTCCACTTA
cruzapernas	Heatr6_F	CTGCCCAATCATCAGCTCTCC
	Heatr6_R	AAAGAGAGGAGGCTGAAAAGG
cruzapernas	Slfn9_F	CTGCATTGCGTACCACTTGT
	Slfn9_R	CAGGGAAGACAATCATAGCCATG
cruzapernas	Taf15_F	CAGTCAAGGCTATGGACAAACACC
	Taf15_R	GTTGTCCCTGGTTATTGTATGATTG
cruzapernas	Slfn1_F	AGGGATCACACTTGGGACAG
	Slfn1_R	CTAAGACATGAGGAGCTTGATCC
fraqueza	Klhl12_F	CTGTTGTAAGTTTGAGACCAACCTG
	Klhl12_R	CTTGTTCTGGGTCCTTGCATCTG
fraqueza	Cntnap5a_F	TTCCCCTACAGGCGTCATAG
	Cntnap5a_R	CTCGTTTGCATTCAGTGGCTGTG
fraqueza	Dst_F	CCACCAAAGGAGATGGAGAAATC
	Dst_R	CGTTCTCAGAGCATGAGGATTGAG
Sacudidor	Celsr1_F	GCTTACCAGAGGAGCAGACG
	Celsr1_R	GATTTCTACATTGAGCCCACGTC
Sacudidor	Tubgcp6_F	AAAGGCATCCAACTTTCAGG
	Tubgcp6_R	CAGGGGCCTACTGACAGAGA

Para cada SNV foram produzidas no mínimo 5 leituras e no máximo 10 leituras, correspondentes aos sequenciamentos *forward* e *reverse* das amostras provenientes do DNA genômico das 5 amostras selecionadas (4 controles e 1 mutante). As sequências foram analisadas quanto a qualidade utilizando a ferramenta Chromas (Technelysium) e os alinhamentos globais das sequências foram realizados pela ferramenta Clustal Omega (SIEVERS et al., 2014).

3.2.5 Extração de histonas e Western blot

Camundongos mutantes bate palmas, fraqueza e BALB/cICBI selvagens com idades de 30 dias e 1 ano foram utilizados para a retirada de tecidos do baço e cérebro. Os tecidos foram macerados em nitrogênio líquido, pesados e usados para o protocolo de extração de histonas com o kit *Histone Extraction Kit* (Abcam, Cambridge, EUA).

A extração consumiu cerca de 50 mg de cada tipo de tecido e consistiu na incubação com tampão de pré-lise e remoção do sobrenadante por centrifugação. O protocolo de extração consiste na incubação do precipitado em três volumes de tampão de lise por 30 minutos em gelo. A solução é então centrifugada por 5 minutos a 12,000 RPM. Adicionou-se 0,3 volumes de tampão de balanceamento com DTT. As proteínas extraídas foram submetidas a eletroforese em SDS-PAGE em gel de poliacrilamida 8% e quantificadas pelo kit Pierce BCA Protein Assay kit (Thermo Fisher) para as análises de *western blot,* descritas a seguir.

Após a separação das proteínas em gel deu-se a transferência das proteínas para a membrana de nitrocelulose. A membrana foi bloqueada com solução de bloqueio com leite (5% leite desnatado em PBS e 0,1% Tween-20) ou com BSA (5% BSA em PBS e 0,1% Tween-20) por 1 hora a temperatura ambiente. Os seguintes anticorpos primários foram utilizados para incubação *overnight* a 4°C com agitação: Anti-H3 (1:5000), Anti-H3K4me1 (1:1000), Anti-H3K4me2 (1:1500) e Anti-H3K4me3 (1:1000). Após esse período a membrana foi lavada três vezes por 5 minutos em solução de lavagem PBS + 0,1 % Tween-20 e depois incubada com o anticorpo secundário anti-rabbit HRP (1:5000) por uma hora com agitação. A membrana foi incubada com a solução Lunianta Forte Western HRP (Merck Millipore, MA, EUA) e as bandas detectadas por quimioluminescência por fotodocumentador (ChemiDoc, BioRad, Hercules, CA, EUA). As bandas foram quantificadas pelo *software* ImageJ.

3.3 Resultados e Discussão

3.3.1 Sequenciamento dos camundongos mutantes e busca por SNV candidatos

Cerca de 92% das leituras foram alinhadas no genoma de referência e 79,54% dessas leituras, em média, foram mapeadas em regiões alvo de exons (**Tabela 3.3**). Em média, apenas 4,6% das regiões alvo não tiveram cobertura e 91,35% das bases-alvo foram cobertas pelo menos 5X. A cobertura média do exoma das amostras variou de 37,79X (mutante atáxico-1) a 156,9X (mutante fraqueza*)*, com média de cobertura de todas as amostras de 77,08X (**Tabela 3.3**).

O número médio de SNPs totais chamados para os indivíduos mutantes ENU foi de aproximadamente 76.500 SNPs. Para as amostras das linhagens isogênicas, usando os mesmos parâmetros de chamada usado para as amostras dos mutantes, o número de SNPs foi de cerca de 78.700 (BALB/cICBI) e 2.900 SNPs (C57BL/6ICBI) (**Tabela 3.3**). Isso é concordante se considerarmos que o genoma de referência é majoritariamente C57BL/6J e que as amostras tiveram coberturas médias do exoma diferentes.

A estratégia de filtragem de SNVs para localização dos candidatos a mutações causais, conforme descrita em (**3.2.2**), teve suporte na premissa que os filtros diminuiriam a quantidade de mutações candidatas para validação. De fato, conforme descrito na **Tabela 3.4,** houve uma diminuição gradativa da quantidade de SNPs a cada etapa de filtragem. A premissa de que a mutação deveria ser não-sinônima ou estar localizada em sítios exônicos ou de *splicing* também diminuiu consideravelmente o número de mutações candidatas (**Tabela 3.4**) e consequentemente o futuro esforço de validação.

Tabela 3.3 Estatísticas de alinhamento das amostras sequenciadas.

		Linhagem isogênica		Mutantes induzidos por ENU						
	Média	**C57BL/6ICBI**	**BALB/cICBI**	**ataxico-1**	**bate palmas**	**careca**	**cruzapernas**	**equilibrio**	**fraqueza**	**Sacudidor**
Num. leituras	105327953	140456591	173961821	55471554	84590152	55471554	64304657	62364613	217138559	94192080
Num. leituras mapeadas	97389838	131198511	161084247	50978662	77678233	50978662	59775994	57499486	200547575	86767171
% leituras mapeadas	92,36	93,4	92,6	91,90	91,83	91,90	92,96	92,2	92,4	92,12
Leituras mapeadas nos alvos	61888930	82555255	97162199	30488426	48295647	48802221	35902213	35786749	125597730	52409931
% leituras mapeadas nos alvos	79,54%	80,39%	78,83%	77,84%	81,29%	78,29%	78,03%	81,28%	81,95%	77,95%
Leituras mapeadas fora dos alvos [1]	15599434	20140757	26094818	8678319	11115018	13533859	10108127	8241085	27657849	14825072
% leituras mapeadas fora dos alvos[1]	20,46%	19,61%	21,17%	22,16%	18,71%	21,71%	21,97%	18,72%	18,05%	22,05%
Número de regiões-alvo sem cobertura	3790	3306	3245	4360	3816	4062	4083	4259	3144	3832
Porcentagem de bases-alvo sem cobertura	2280639 (4,6%)	1932195 (3,9%)	1944767 (3,9%)	2634910 (5,3%)	2333519 (4,7%)	2371193 (4,8%)	2488797 (5,0%)	2664606 (5,4%)	1884665 (3.8%)	2271103 (4,6%)
Porcentagem de bases-alvo cobertas >= 1X:	95,43%	96,09%	96,06%	94,66%	95,28%	95,20%	94.96%	94,60%	96,18%	95,40%
Porcentagem de bases-alvo cobertas >= 5X:	91,35%	93,38%	93,49%	88,24%	90,74%	90,97%	89.32%	88,61%	93,97%	91,43%
Porcentagem de bases-alvo cobertas >= 10X:	86,44%	90,54%	91,02%	79,43%	85,09%	85,80%	81.80%	80,86%	91,99%	86,75%
Porcentagem de bases-alvo cobertas >= 20X:	75,67%	84,24%	85,73%	60,55%	72,39%	73,99%	65.18%	64,41%	88,00%	76,01%
Média de cobertura nas regiões-alvo	77,08	102,1	121,52	37,79	60,49	60,32	44,75	44,66	156,9	64,63
Número de total de SNPs chamados	76522[2]	2906	78786	75434	77075	77803	76110	73209	79276	76745

[1] Foram consideradas fora dos alvos também as regiões próximas aos exons-alvo - frequentemente associadas a *splicing.*
[2] Média de SNPs totais chamados por mutante.

De fato, a aplicação dos filtros de mapeamento e dos filtros de exclusividade foram muito eficientes na filtragem dos SNVs reduzindo para no máximo 5 candidatos (*careca*) por mutante (**Tabela 3.4**). Essa redução, promovida pela estratégia de filtragem, corresponde a consideráveis 0,06 % a 0,8 % do total de SNVs detectados nas regiões mapeadas.

Tabela 3.4 Número de SNVs nas etapas de filtragens.

Mutante	*Tipo de herança*	*Região mapeada (chr:cM)* [a]	*Total de variantes na região mapeada*	*Homozigotas na região mapeada*	*Exclusivas* [b]	*Não presentes no dbSNP* [c]	*Não-sinônimas ou em sítios de* splicing [d]	*Únicas* [e]
ataxico-1	Recessiva	10: 25-40	*121*	*78*	*8*	*7*	*2*	*1*
bate palmas	Recessiva	15: 19-37	*1668*	*1525*	*87*	*18*	*3*	*1*
careca	Recessiva	*7*	*5815*	*5301*	*336*	*81*	*9*	*5*
cruzapernas	Recessiva	11	*4851*	*4386*	*313*	*76*	*15*	*3*
equilibrio	Recessiva	17: 1-10	*212*	*185*	*16*	*7*	*1*	*1*
fraqueza	Recessiva	1: 11-81	*4345*	*3591*	*244*	*49*	*4*	*3*
Sacudidor	Dominante	15:40-60	*509*	-	*49*	*34*	*11*	*1*

[a] Unidades de recombinação média (cM) médias macho-fêmea COX (http://cgd.jax.org/mousemapconverter/).
[b] SNVs exclusivos em relação aos camundongos isogênicos C57BL/6ICBI e BALB/cICBI.
[c] dbSNP versão 132 (dbSNP132, 26/09/2010)
[d] Foram considerados sítios de splicing até dois nucleotídeos das bordas de exons.
[e] SNVs únicos em relação aos outros mutantes sequenciados e em relação às linhagens sequenciadas pelo Mouse Genomes Project (Sanger) REL-1211 e REL-1505.

Foi possível obter pelo menos um candidato para todos os mutantes sequenciados (**Tabela 3.5**), com coberturas locais variando de 10X a 185X. Quatro mutantes (ataxico-1, bate palmas, equilíbrio e Sacudidor), que equivalem a mais da metade de todos os mutantes sequenciados, tiveram apenas um único SNV encontrado, segundo os critérios de filtragem utilizados. Os outros três mutantes – careca, cruza pernas e fraquezas – tiveram um número maior de mutações candidatas encontradas, entre 3 e 5 SNVs candidatos (**Tabela 3.5**).

Tabela 3.5 Genes candidatos causativos para cada mutante induzido por ENU.

	Coordenadas (NCBI37/mm9) [a]	***Posição (cM)***	***Cobertura local***	***Genes***	***Variante no transcrito***
ataxico-1	chr10:g.59779820C>A	24.273	10X	Cdh23	*NM_001252635:c.G7165T:p.E2389X*
bate palmas	chr15:g.98691505T>C	49.263	40X	Kmt2d	*NM_001033276:c.A3865G:p.T1289A*
careca	chr7:g.26995160G>T	6.609	116X	Cyp2b9	*NM_010000:c.G1333T:p.E445X*
	chr7:g.29626295A>G	9.217	22X	Lgals4	*NM_010706:c.A803G:p.Y268C*
	chr7:g. 30133683G>T	10.118	14X	Sipa1l3	*NM_001081028:c.C3798A:p.N1266K*
	chr7:g.127929977A>G	61.391	112X	Vwa3a	*NM_177697:c.A2027G:p.E676G*
	chr7:g.150756020A>G	82.381	61X	Cars	*NM_001252593:c.T1430C:p.F477S*
cruzapernas	chr11:g.82934641C>A	45.345	25X	Slfn1	*NM_011407:c.C80A:p.T27N*
	chr11:g.83298375G>A	45.406	90X	Taf15	*NM_027427:c.G163A:p.G55S*
	chr11:g.83589258C>T	45.702	21X	Heatr6	*NM_145432:c.C2354T:p.S785F*
equilibrio	chr17:g. 3695328T>A	0.039	115X	Nox3	*NM_198958:c.A250T:p.N84Y*
fraqueza	chr1:g.34213755T>A	10.948	185X	Dst	*NM_001276764:c.1679+2T>A* *NM_133833:c.1145+2T>A* [b]
	chr1:g.118477109C>T	48.391	88X	Cntnap5a	*NM_001077425:c.C3773T:p.T1258I*
	chr1:g.136372369T>C	55.977	134X	Klhl12	*NM_153128:c.T722C:p.F241S*
Sacudidor	chr15:g.85861082T>C	35.091	91X	Celsr1	*NM_009886:c.A3119G:p.D1040G*

[a] *Nomenclatura de acordo com as recomendações da HGVS (*http://varnomen.hgvs.org*)*
[b] *Considerando transcrito mais longo.*
[c] Há pelo outras duas variantes de splicing anotadas para o gene *Dst.*

3.3.2 Validação das mutações candidatas e frequência de falsos-positivos

A estratégia de validação consistiu na amplificação por PCR de 15 regiões específicas ao redor das mutações detectadas (**Tabela 3.5**) correspondentes a cada SNV detectado. O tamanho estimado dos produtos de amplificação variou entre 191 a 967 pb. O processo de validação consistiu no sequenciamento de produtos de PCR oriundos de amostras de DNA genômico do mutante e de 4 controles (C57BL/6J, BALB/cICBI, A/J e um mutante ENU aleatório) e da amostra mutante cujo SNV candidato foi detectado pelo sequenciamento do exoma (descrito em **3.2.4**).

Os SNVs candidatos dos mutantes foram submetidos ao processo de validação por sequenciamento Sanger em relação a presença da mutação, zigosidade e ausência em todos os controles. Dois SNVs (Cdh23:c.G7165T:p.E2389X e Nox3:c.A250T:p.N84Y) foram revalidados de forma independente pelo grupo da Prof. Marimélia Porcionatto (UNIFESP). Somente dois SNVs, do mutante careca, não puderam ser submetidos ao processo de validação (Sipa1|3:c.C3798A:p.N1266K e Lgals4:c.A803G:p.Y268C), devido a problemas na amplificação dos produtos de PCR.

Duas mutações, que foram excluídas da última etapa de filtragem (*Tubgcp6* e *Slfn9*) por terem sido encontradas no banco de dados do MGP, também foram submetidas ao processo de validação como controles para estimativa da taxa de falsos-positivos na chamada dos SNVs. Uma mutação (Tubgcp6:c.G5138A:p.R1713H), detectada em heterozigose e com baixa cobertura (9X) não foi validada em relação a presença da mutação e zigosidade. A mutação (Slfn9:c.G2038A:p.G680R) foi validada em relação a presença da mutação mas não em termos de zigosidade, sendo detectada em heterozigose no mutante (**Tabela 3.6**).

Tabela 3.6 Mutações candidatas selecionadas para validação por Sanger.

Coordenadas (NCBI37/mm9)	***Coordenadas (GRCm38/mm10)[a]***	***Cobertura***	**Het/Hom[b]**	***Gene***	***Valid. Sanger (mutação)***	***Valid. Sanger (zigosidade)***
chr15:g88931095C>T	*chr15:g.89100665C>T*	*9X*	Het	*Tubgcp6*	Não	Não
chr10:g.59779820C>A	chr10:g.60317072C>A	*10X*	Hom	*Cdh23*	Sim	Sim
chr11:g.83589258C>T	chr11:g.83775756C>T	*21X*	Hom	*Heatr6*	Sim	Sim
chr11:g.82934641C>A	chr11:g.83121139C>A	*25X*	Hom	*Slfn1*	Sim	Sim
chr11:g. 82795373C>T	chr11:g.82981871C>T	*25X*	Hom	*Slfn9*	Sim	Não
chr15:g.98691505T>C	chr15:g.98861074T>C	*40X*	Hom	*Kmt2d*	Sim	Sim
chr7:g.150756020A>G	chr7:g.143570115A>G	*61X*	Hom	*Cars*	Sim	Sim
chr1:g.118477109C>T	chr1:g.116580532C>T	*88X*	Hom	*Cntnap5a*	Sim	Sim
chr11:g.83298375G>A	chr11:g.83484873G>A	*90X*	Hom	*Taf15*	Sim	Sim
chr15:g.85861082T>C	chr15:g.86030652T>C	*91X*	Het	*Celsr1*	Sim	Sim
chr7:g.127929977A>G	chr7:g.120786463A>G	*112X*	Hom	*Vwa3a*	Sim	Sim
chr17:g. 3695328T>A	chr17:g.3695328T>A	*115X*	Hom	*Nox3*	Sim	Sim
chr7:g.26995160G>T	chr7:g.26210141G>T	*116X*	Hom	*Cyp2b9*	Sim	Sim
chr1:g.136372369T>C	chr1:g.134475792T>C	*134X*	Hom	*Klhl12*	Sim	Sim
chr1:g.34213755T>A	chr1:g.34156910T>A	*185X*	Hom	*Dst*	Sim	Sim
chr7:g.30133683G>T	chr7:g.29348664G>T	*14X*	Hom	*Sipa1l3*	-	-
chr7:g.29626295A>G	chr7:g.28841276A>G	*22X*	Hom	*Lgals4*	-	-

[a] Coordenadas geradas a partir das coordendas NCBI37/mm9 pela ferramenta LiftOver (UCSC Genome Browser).
[b] Het=heterozigotos (MAF=0.35-0.65), Hom=homozigotos.

A taxa de acurácia em relação à detecção da mutação e zigosidade no processo de validação, considerando apenas SNVs filtrados e candidatos, foi de 100% (12 SNVs). Se considerarmos apenas a detecção da mutação em todas as variantes, sem considerar a zigosidade, a taxa de validação é de 93,3% (14 SNVs) e se considerarmos todos os SNVs submetidos a validação independente se foram filtrados ou não essa taxa é de 86,7 % (**Figura 3.6**). A taxa de falsos positivos estimada considerando apenas a detecção da mutação foi de apenas 6,7%, correspondente ao único SNV não

validado (Tubgcp6:c.G5138A:p.R1713H), do mutante Sacudidor, que não foi selecionado por todos as etapas de filtragem mas escolhido para testar a acurácia em baixa cobertura (9X).

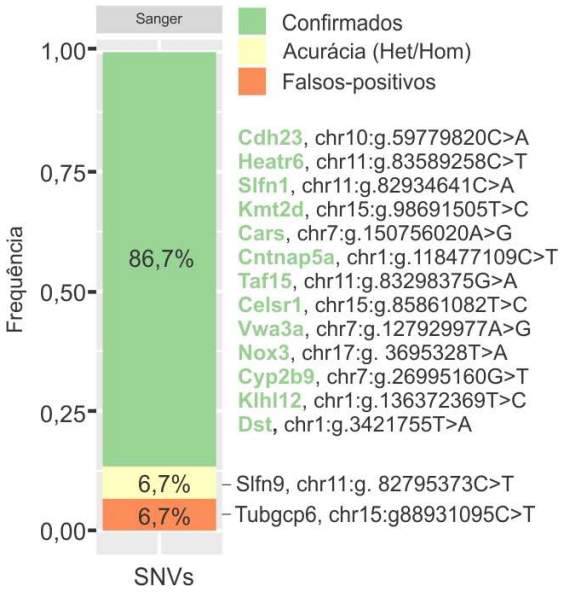

Figura 3.6 Validação por sequenciamento Sanger dos SNVs detectados. Os SNVs estão representados conforme as coordenadas do genoma NCBI37/mm9. SNVs confirmados foram definidos como validados completamente. A acurácia Het/Hom é definida pela validação da mutação mas não em relação à zigosidade. Falsos-positivos são SNVs não foram validados tanto com relação à própria mutação e também em relação à zigosidade.

3.3.3 Análise do impacto dos SNVs nos principais candidatos dos mutantes ENU.

A predição do impacto do SNV no produto dos principais genes candidatos foi realizada pelos algoritmos PROVEAN, SIFT e Polyphen-2 (**Tabela 3.7**). Para o gene *dst* foi utilizado o algoritmo SpliceMan, que avalia o impacto de SNVs em sítios de *splicing.* Todos os mutantes possuem pelo menos um SNV que implica em algum tipo de dano para o respectivo produto gênico, predito pelas ferramentas utilizadas ou considerando o impacto de mutações *nonsense,* como as detectadas para o gene *cdh23* (ataxico-1) e *cyp2b9* (careca*)* (**Tabela 3.7**). A única exceção é a mutação encontrada no mutante batepalmas, no gene *kmt2d*, que apesar de não implicar em dano predito pelas ferramentas de análise de impacto, essa mutação pode afetar modificações pós-traducionais na proteína Kmt2d (**3.3.4**).

Em relação ao mutante careca, temos que todas as mutações encontradas foram preditas como potencialmente impactantes na função do produto gênico, por pelo

menos dois algoritmos de predição (**Tabela 3.6**). O mutante cruzapernas apresenta apenas uma mutação predita como deletéria (Heatr6:p.S785F) entre os três candidatos encontrados. A única mutação candidata encontrada para o camundongo equilíbrio também é predita como deletéria para o produto gênico (Nox3:p.N84Y), enquanto para o mutante fraqueza temos dois candidatos com impacto predito (Dst:c.1679+2T>A e Klhl12: p.F241S). Finalmente para o mutante Sacudidor apenas uma mutação foi encontrada (Celsr1:pD1040G), sendo a mesma predita como deletéria pelas três ferramentas de predição de impacto (**Tabela 3.7)**.

Tabela 3.7 Predição do impacto das mutações nos principais genes candidatos

	Candidatos[a]	***Provean***	***SIFT***	***Polyphen-2***	***SpliceMan***	***Validação por Sanger***
ataxico-1	Cdh23:p.E2389X	-	-	-	-	Sim
bate palmas	Kmt2d:p.T1289A	Neutral	Tolerated	Unknown	-	Sim
careca	Cyp2b9:p.E445X	-	-	-	-	Sim
	Lgals4:p.Y268C	Deleterious	Tolerated	Prob. damaging	-	ND
	Sipa1l3:p.N1266K	Deleterious	Damaging	Prob. damaging	-	ND
	Vwa3a:p.E676G	Deleterious	Damaging	Prob. damaging	-	Sim
	Cars:p.F477S	Deleterious	Damaging	Prob. damaging	-	Sim
cruzapernas	Slfn1:p.T27N	Neutral	Tolerated	Benign	-	Sim
	Taf15:p.G55S	Neutral	Unknown	Benign	-	Sim
	Heatr6:p.S785F	Deleterious	Damaging	Prob. damaging	-	Sim
equilibrio	Nox3:p.N84Y	Deleterious	Damaging	Prob. damaging	-	Sim
fraqueza	Dst:c.1679+2T>A	-	-	-	72%	Sim
	Cntnap5a:p.T1258I	Neutral	Tolerated	Benign	-	Sim
	Klhl12: p.F241S	Neutral	Damaging	Benign	-	Sim
Sacudidor	Celsr1:p.D1040G	Deleterious	Damaging	Prob. damaging	-	Sim

3.3.4 O camundongo bate palmas possui uma mutação não sinônima no gene kmt2d*.*

A mutação detectada no mutante *bapa* é uma troca T/C no transcrito (c.A3865G:p.T1289A) localizada no exon 13 (**Tabela 3.5** e **Tabela 3.6**). O SNV foi o único detectado pela estratégia de filtragem (**Figura 3.7A**) com cobertura local de 40X e validado por sequenciamento Sanger com amostras extraídas de linhagens isogênicas, de um mutante não relacionado e de um outro indivíduo mutante *bapa* (**Figura 3.7B**), corroborando a exclusividade da troca e a forte evidência da presença do alelo em homozigose na população mutante. O SNV encontrado (*kmt2d*:c.A3865G:p.T1289A) implica em uma troca na região N-terminal da proteína, antes do início do segundo grupo de domínios PHD sendo uma troca não-sinônima de um resíduo de treonina (T)

para um resíduo de alanina (A) (**Figura 3.7C**). O gene *kmt2d* de camundongo possui 55 exons, sendo que o transcrito primário possui 39 kbp e a proteína 5588 resíduos aminoácidos e está localizado no cromossomo 15 (GRCm38/mm10).

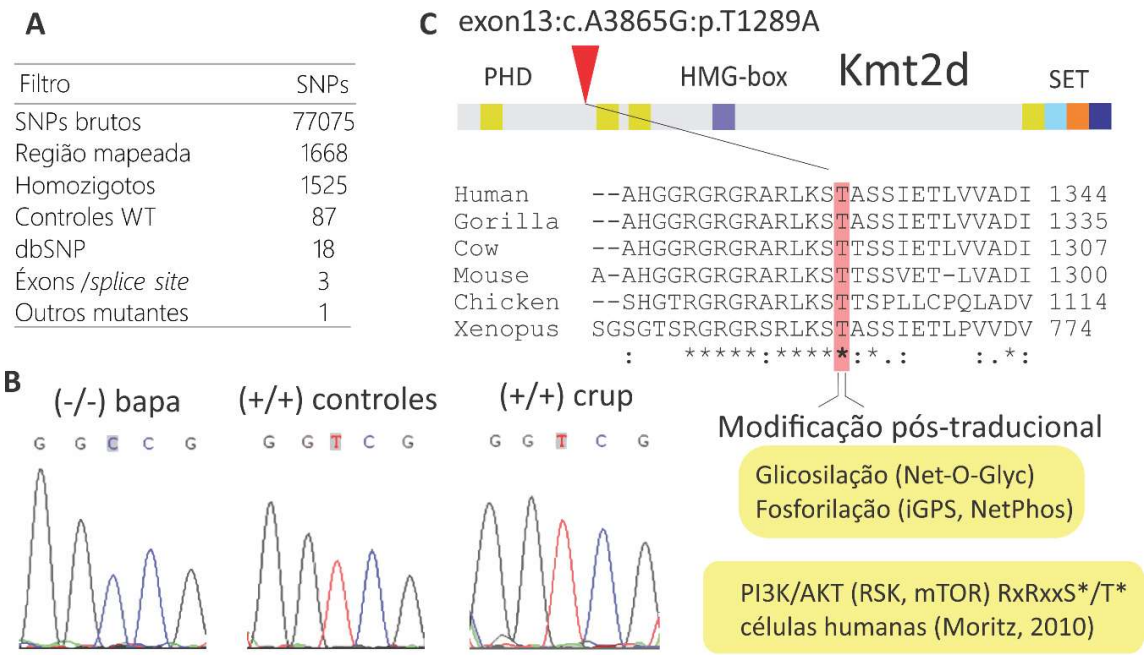

Figura 3.7 Um SNV não-sinônimo foi encontrado no gene que codifica uma Kmt2d H3K4 metiltransferase (**A**) Seleção de SNVs através de etapas de filtragem. (**B**) Validação SNV por sequenciamento Sanger. As amostras controles correspondem a C57BL/6ICBI, BALB/cICBI e A/JICBI. (**C**) O candidato SNV encontrado no exon 13 do *kmt2d* (previamente conhecido como MLL2 / 4) gene implica em uma mudança não-sinônima de um resíduo de treonina para alanina na proteína Kmt2d . Embora este resíduo esteja localizado fora de domínios proteicos conhecidos, ele é conservado entre vertebrados e pode ser um alvo para modificações pós-translacionais, como previsto por ferramentas *in silico*. Além disso, o homólogo humano foi encontrado Kmt2d fosforilado *in vivo* no mesmo resíduo de treonina devido à quinase de sinalização de PI3K /AKT (Moritz , 2010).

O impacto da troca foi avaliado por três diferentes algoritmos de predição de impacto: Polyphen-2, SIFT e PROVEAN, que avaliam principalmente o impacto da troca na estrutura secundária, em informações filogenéticas e nas características de estruturas terciárias semelhantes. Embora as ferramentas usem diferentes parâmetros na classificação do impacto, as predições têm geralmente elevada concordância (MARTELOTTO et al., 2014). Não foi possível realizar a predição pela ferramenta Polyphen-2, devido a problemas com proteínas não globulares (~ 5% das sequencias depositadas no UniProtKB) como fatores de transcrição e ligases de RNA/DNA, segundo a documentação da ferramenta (ADZHUBEI et al., 2012). A predição realizada

pela ferramenta PROVEAN indicou que a troca é neutra, com base no *score* igual a -0,25, sendo que apenas trocas com *scores* maiores que -2,5 são consideradas deletérias (CHOI et al., 2012). A ferramenta SIFT indicou que a troca é tolerada por um *score* de 0,095, em que as trocas com *scores* menores que 0,05 são consideradas danosas (*damaging*) (KUMAR et al., 2010) (**Tabela 3.7**).

Um alinhamento múltiplo das sequências de prováveis homólogos a Kmt2d em mamíferos, aves e anfíbios foi realizado e o resíduo T é conservado em todos as sequências analisadas (**Figura 3.7C**). Os resíduos de treonina podem ser alvo de fosforilação serina-treonina ou mesmo de glicosilação. Os softwares de predição Net-O-Glyc, iGPS e NetPhos indicaram a possibilidade de fosforilação do resíduo T ou mesmo a glicosilação do resíduo, indicando que uma possível modificação pós-traducional pode ocorrer no resíduo, que seria impedida pela presença do alelo mutante (**Figura 3.7C**). Além disso, a região é condizente com a sequência consenso de fosforilação em humanos PI3K/AKT (RSK, mTOR) RxRxxS*/T*.

Dada a função da proteína Kmt2d na metilação de histonas H3K4 investigamos o balanço de metilação dessas histonas em tecidos do camundongo mutante bate palmas. Consideramos dois tecidos – cérebro e baço – e o anticorpo Anti-H3K4me1, específico para resíduos lisina-4 da histona H3 que estão monometilados, como indícios de desbalanço de regulação epigenética mediada por metilação (**Figura 3.8**). Resultados preliminares indicam que a metilação H3K4 não está corretamente balanceada no camundongo bate palmas, que possui mais histonas H3 mono-metiladas no resíduo K4, pelo menos em amostras de proteínas do baço (**Figura 3.8**). No cérebro, há menor monometilação no camundongo controle, independentemente da idade. No baço, com 30 dias, observamos uma menor monometilação em relação ao controle, enquanto em 1 ano não há diferença nesse tecido entre o camundongo mutante o controle (**Figura 3.8**).

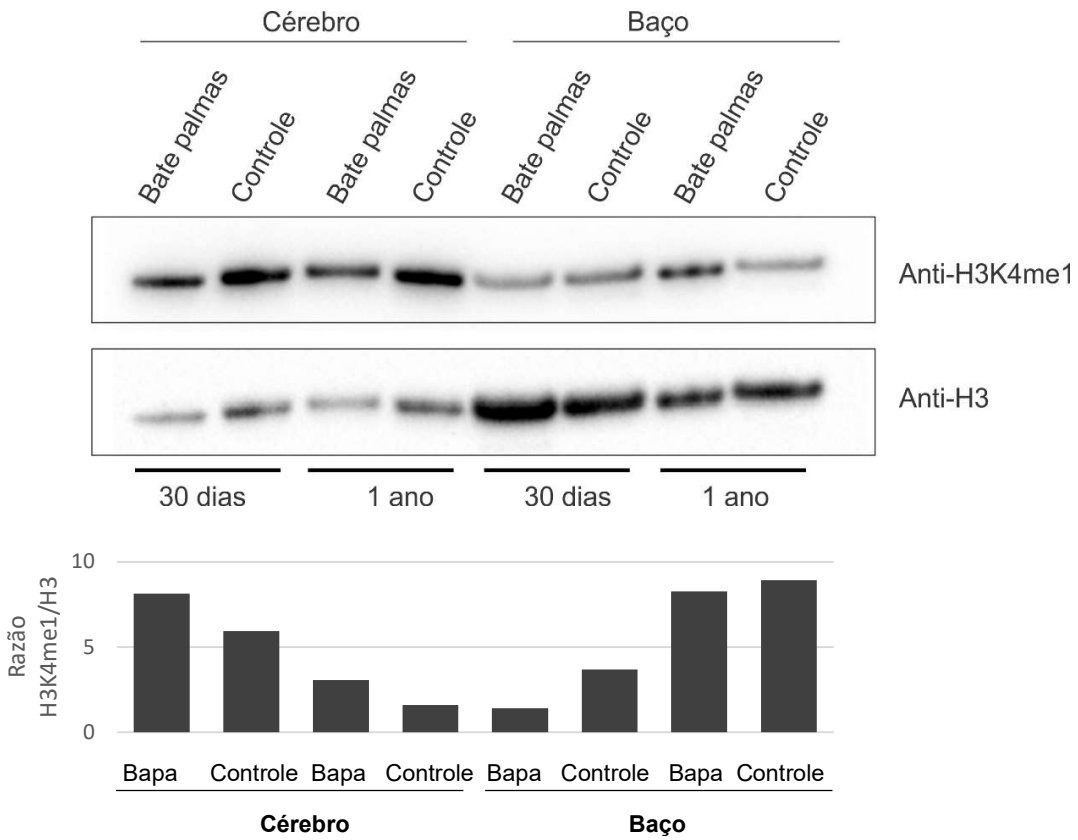

Figura 3.8 Avaliação preliminar da monometilação de histonas H3 como marcador de regulação epigenética. Análise de metilação de histonas por western blot com anti- H3 e anti-H3K4me1. Camundongos fraqueza foram utilizados como controle.

3.3.5 O mutante careca possui pelo menos três SNVs candidatos causadores do fenótipo

Foram encontrados um total de 5 SNVs candidatos, sendo que três SNVs candidatos causais para o mutante careca puderam ser validados (**Figura 3.9A**). Todos os 5 SNVs foram preditos como potencialmente impactantes (*damaging/deleterious*) e embora os genes *vwa3a* e *cars* estejam envolvidos em processos de divisão celular, o SNV encontrado no gene *cyp2b9* é o mais promissor, visto que ele acarreta na inclusão de um códon de parada prematuro (**Figura 3.9B**). Todos esses três SNVs foram validados por sequenciamento Sanger (**Figura 3.9C**), porém não foi possível realizar os ensaios de validação para os candidatos Lgal4 e Sipa1|3.

O produto do gene *cyp2b9* é uma NADPH oxidase (aldeído desidrogenase) envolvida na biossíntese de DHT (di-hidrotestosterona) e T (testosterona), composto por um único domínio P450 conservado (INGELMAN-SUNDBERG et al., 2013).

Portanto, uma provável perda de função devido ao códon de parada prematura no gene *cyp2b9* pode implicar em uma desordem da síntese hormonal de T ou DHT. A aquisição do códon de parada prematuro implica em uma perda de pelo menos 45 aminoácidos (ou 10% da proteína). Além disso, a presença de códons de parada prematuros próximos a junções de exons pode implicar em um fenômeno de NMD (*non-sense mediated RNA decay*), que implica na degradação completa de RNAs mensageiros sinalizada pela interferência entre o processo de *splicing* e tradução (SCHWEINGRUBER et al., 2013).

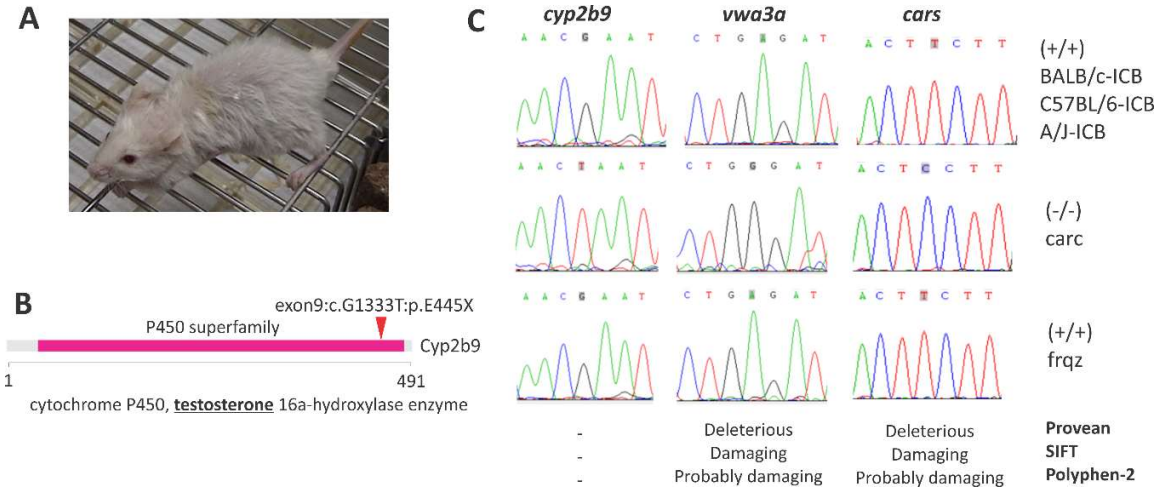

Figura 3.9 Três SNV candidatos foram selecionados para mutações causativas no mutante carc. (**A**) Ambos os sexos masculino e feminino de Carc apresentam perda progressiva dos pêlos. (**B**) Uma mutação de ganho de parada foi encontrada no gene cyp2b9 , que codifica uma enzima descrita por estar envolvida na biossíntese de testosterona. (**C**) Os SNV selecionados nos genes *cyp2b9, vwa3a* e *cars* foram validados por sequenciamento Sanger e foram previstos como sendo prejudiciais por ferramentas *in silico* .

3.3.6 O mutante fraqueza possui um SNV em um sítio de splicing do gene dst

Foi possível identificar, através do presente estudo, três SNVs candidatos que foram confirmados por sequenciamento Sanger. Porém, apenas o SNV encontrado no gene *dst (dystonin/bpag1*) foi predito como potencialmente danoso à função do gene, pela ferramenta SpliceMan (**Tabela 3.7**). A mutação Klhl12: p.F241S foi predita como deletéria apenas pela ferramenta SIFT, sendo predita como benigna pelas outras ferramentas.

O gene *dst (dystonin/Bpag1*) codifica para uma proteína chamada de distonina (*dystonin*), um membro da família das plaquinas. Esse SNV, situado a dois nucleotídeos da extremidade 3' do exon 11, foi predito como danoso ao processo de *splicing* pelo

software SpliceMan. A extremidade 5' do intron de 3.7 kbp a ser removido por *splicing* possui uma sequência GU (GT no DNA) amplamente conservada, chamada de sítio doador do *splicing.* O exon 12 é comum a todas as isoformas da Dst, sendo que somente a isoforma 4 não possui o exon 11. Curiosamente, a isoforma 4, também conhecida como distonina variante "e", é a isoforma epitelial, que não está associada até o momento com o fenótipo dos camundongos que possuem perda de função no gene distonina, cujo fenótipo é chamado de distonia musculorum (*dt*) (FERRIER et al., 2014). São conhecidos vários alelos do gene *dst* cujos fenótipos se caracterizam principalmente por desordens psicomotoras progressivas. Um dos mais conhecidos é o camundongo Dst^{Tg4}, que apresenta uma inserção no gene *Dst* e não expressa as isoformas 1 e 2 mas ainda expressa a isoforma 3 (POOL et al., 2005). Considerando a similaridade entre o fenótipo do mutante fraqueza e os camundongos dt fica reforçada a hipótese de que o SNV candidato possa afetar o mecanismo de *splicing* do gene dst e levar à diminuição da transcrição de uma ou mais isoformas do gene.

A caracterização do fenótipo (**Figura 3.10A**) junto com a identificação do SNV (**Figura 3.10B** e **3.10C**) direcionou nossa hipótese de perturbação do *splicing* no gene *dst* nos camundongos *frqz* (**Figura 3.10C** e **Figura 3.10D**). Fenômenos como o *exon skipping* ou retenção de intron podem estar acontecendo, afetando pelo menos as isoformas neuronais e musculares da Dst. Como a isoforma 4 epitelial é menor e seu ínicio é praticamente a partir do exon 12, provavelmente ela não é afetada pela presença do SNV. Um conjunto de oligonucleotídeos foi desenhado para estudos detalhados de RT-PCR direcionados aos exons 10, 11 e 12 e também ao intron para a detecção de uma eventual perturbação do *splicing* desse gene. Esse tipo de estudo pode auxiliar no entendimento da relação dos mecanismos de *splicing* associados à especificidade e papel das isoformas de Dst na caracterização do fenótipo dos camundongos *dt.*

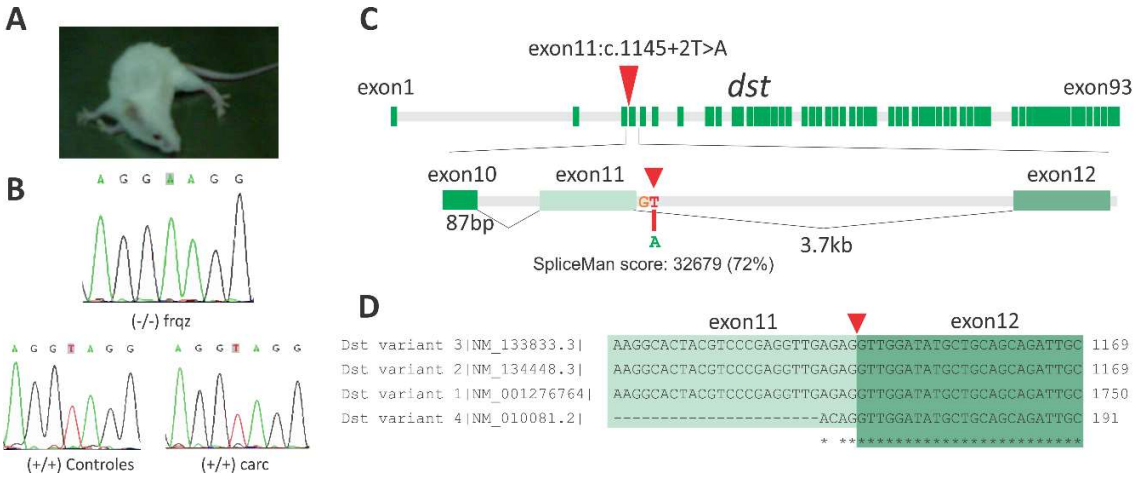

Figura 3.10 O *Splicing* de variantes de transcritos de distonina pode ser afetado por uma SNV vizinho ao exon 11 nos camundongos frqz. (**A**) Camundongo frqz apresentando perda progressiva da coordenação motora. (**B**) Validação do SNV no gene *dst* por seqüenciamento Sanger, amostras controle correspondem a C57BL/6ICBI, BALB/cICBI e A/JICBI. (**C**) Estrutura gênica do gene *dst* e a mutação encontrada, a duas bases da borda do exon 11. (**D**) O alinhamento múltiplo das isoformas Dst na junção do exon 11 e 12 revela que apenas uma isoforma não possui o exon 11 (variante 4) e as outras três isoformas possuem os dois exons.

3.3.7 Mutantes atáxico, equilíbrio, cruzapernas e Sacudidor

Os mutantes atáxico-1 e equilíbrio são estudados pelo grupo da Prof. Marimélia Porcionatto, da UNIFESP-EPM. Em ambos os casos, encontramos SNVs candidatos que foram confirmados por sequenciamento Sanger, nos genes *cdh23* e *nox3*, respectivamente. Diversos estudos funcionais estão sendo realizados e parecem confirmar a impacto das alterações no fenótipo dos camundongos mutantes pelo grupo na UNIFESP-EPM. O mutante equilíbrio, cuja mutação no gene *nox3* que codifica para uma NADPH oxidase 3, possui aumento da proliferação de células precursoras granulares (GCPs) e mudanças na expressão de genes envolvidos no controle de proliferação celular. As células GCPs do mutante também produzem uma maior quantidade de espécies reativas de oxigênio e expressão aumentada de genes alvos da sinalização por SHH (MAZONETTO et al., submetido).

No caso do mutante cruzapernas foi encontrado mais de um SNV candidato com potencial de impacto na função do produto gênico e estão sendo estudados pelas Dra. Silvia Massironi e pela Dra. Cláudia Mori. O mutante apresentou alterações psicomotoras e sensoriais, com baixo índice reprodutivo – condizente com a deficiência motora desses animais. A maioria dos SNVs foram validados, porém mais estudos são necessários para a definição de abordagens funcionais que possibilitem a

caracterização e a associação da presença do SNV com o fenótipo dos camundongos cruzapernas. Consideramos que os SNVs nos genes *taf15* e *heatr6* são os candidatos mais fortes para esse mutante.

Finalmente, o mutante Sacudidor também apresentou somente um candidato, no gene *celsr1,* cuja mutação foi predita como patogênica.O gene *celsr1* codifica para uma caderina da família de proteínas G acopladas de adesão e controla processos de regulação dependente de ácido retinóico envolvidas com neurogênese. Camundongos mutantes com perda de função no gene *celsr1* possuem anormalidades comportamentais e microcefalia, que estão associadas com o comprometimento de células progenitoras neurais no desenvolvimento do córtex cerebral (BOUCHERIE et al., 2017).

3.3.8 Padrão global das mutações únicas encontradas nos mutantes, potencialmente induzidas por ENU.

Para estabelecer um eventual padrão de mutação associado a mutagênese por ENU, foram selecionados apenas os SNVs exclusivos de cada amostra em comparação à todas as outras. Essa abordagem pressupõe que essas mutações únicas (ou exclusivas) têm menor probabilidade de constituírem polimorfismos e maior probabilidade de terem sido induzidas por ENU, sendo chamadas de mutações incidentais. Porém, não foi possível detectar nenhum padrão óbvio de mutações no contexto de trinucleotídeos (**Figura 3.11A**), mesmo em relação ao controle BALB/c. Dessa forma, pressupondo que o padrão de mutações do controle BALB/c deveria ser diferente das outras amostras, foi construído um modelo de 4 assinaturas definidas pelo método NMD. As 4 assinaturas encontradas S1, S2, S3 e S4 (**Figura 3.11B**) podem ser divididas em três grupos: um grupo formado apenas pela assinatura S4, que constitui um padrão de mutações basal sem proeminência óbvia de nenhuma troca e/ou trinucleotídeo, um grupo formado pela assinatura S3, em que há maior proporção de trocas C>T e T>C e um último grupo formado pelas assinaturas S1 e S2, com uma maior proporção de C>T, T>C e também T>A, com uma proporção considerável da troca T>G no contexto GTG (**Figura 3.11B**).

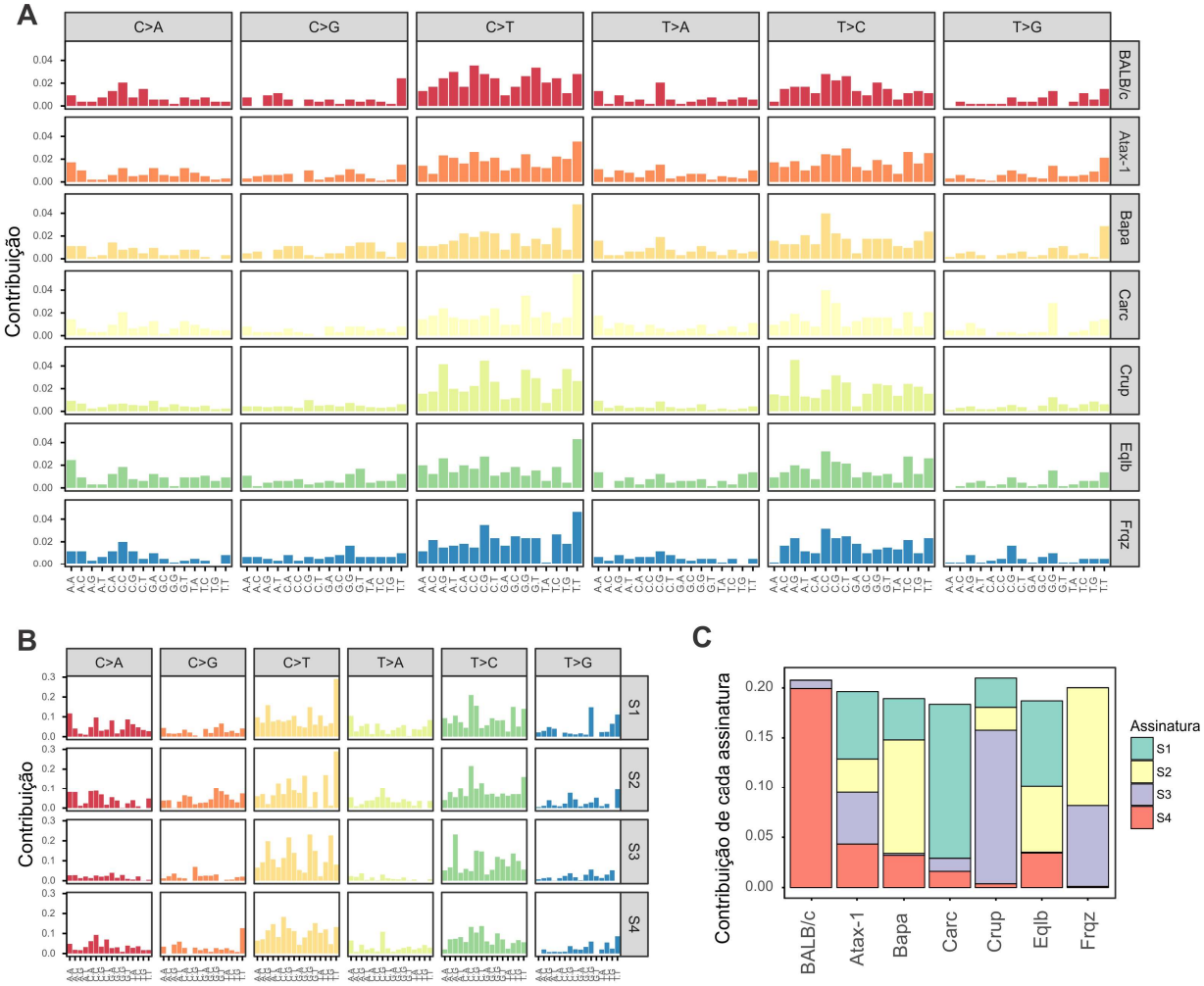

Figura 3.11 Padrão de mutações incidentais em trinucleotídeos nas amostras dos mutantes e do controle BALB/c. (**A**) Espectro de mutações únicas encontradas em cada uma das amostras. (**B**) Modelo de 4 assinaturas definidas pelo método NMD. (**C**) Contribuição das assinaturas S1, S2, S3 e S4 nas amostras dos mutantes e do controle BALB/c.

Considerando o modelo de 4 assinaturas proposto, foi possível mensurar a participação de cada uma das assinaturas em cada uma das amostras (**Figura 3.11C**). A amostra controle BALB/c é composta majoritariamente da assinatura S4, pouco representada em cada uma das amostras dos mutantes. Os mutantes possuem contribuições variáveis das assinaturas S1, S2 e S3, sendo que a maioria deles apresentam uma maior contribuição das assinaturas S1 e S2 (**Figura 3.11C**), que podem estar associadas com um eventual padrão de mutagênese por ENU.

3.4 Discussão

O uso de modelos murinos para doenças humanas é essencial para o entendimento de evoluções patológicas e possíveis tratamentos (ERMANN; GLIMCHER, 2012; GONDO, 2010; JUSTICE; DHILLON, 2016). Nesse sentido, a mutagênese por ENU em camundongos é uma ferramenta robusta para a obtenção de alelos causadores de fenótipos similares a doenças humanas e consequentemente melhores modelos experimentais murinos (BULL et al., 2013; NOLAN; HUGILL; COX, 2002). O presente trabalho estabelece uma metodologia estratégica para detecção de variantes candidatas usando dados de sequenciamento NGS de exomas completos, cuja estratégia se mostrou eficiente na detecção das variantes ao diminuir consideravelmente o esforço para a identificação e validação de candidatos para mutações causativas em todos os camundongos mutantes induzidos por ENU.

A utilização da plataforma SOLiD 5500XL aliada ao preparo de bibliotecas do kit SureSelect e análise de dados pela plataforma LifeScope se mostrou superior, em vários aspectos como maior número de exons cobertos, menor quantidade de exons não cobertos, cobertura local e média de cobertura nos exons maior, a estudos semelhantes com a mesma plataforma (TANISAWA et al., 2013). Se compararmos com a plataforma Illumina, temos que a análise apresentada foi comparável em termos de proporção de leituras mapeadas e cobertura local nos exons, além de superior em termos de enriquecimento e média de cobertura das regiões-alvo (FAIRFIELD et al., 2011).

Consideramos que grande parte do sucesso da estratégia de filtragem foi a sustentação do suporte dado pelas seguintes premissas: a região mapeada previamente por microssatélites; tipo de herança do fenótipo; exclusividade da mutação em relação ao *background* genético (BALB/c) e ao banco de polimorfismos dbSNP e, finalmente, a variante causadora implica em uma troca não-sinônima ou deve estar situada em sítios de *splicing.* Para todos os mutantes tínhamos informação sobre o padrão de herança e pelo menos a informação, proveniente de mapeamento com microssatélites (**Tabela 3.1**), sobre qual cromossomo estaria localizada a mutação causadora (MASSIRONI et al., 2006). De fato, estratégias semelhantes foram desenvolvidas (FAIRFIELD et al., 2011) e vem sendo aprimoradas (FAIRFIELD et al.,

2015; SIMON et al., 2015), utilizando as mesmas premissas do presente estudo. Devido a esse motivo, preferimos não utilizar um filtro de cobertura, mesmo que isso implicasse no aumento de falsos positivos (**3.5.1**). Apesar de não utilizarmos nenhum filtro de cobertura mínima para as variantes, foi obtida uma proporção muito baixa de falsos-positivos (**3.3.2**) indicando que não há uma correlação clara entre cobertura e confiabilidade. O sucesso na validação por sequenciamento Sanger, que ficou entre 100 e 86,7%, (**3.3.2** e **Figura 3.6**) valida o procedimento experimental adotado, desde o enriquecimento dos exons, sequenciamento, mapeamento, chamada de variantes e todo o processo de filtragem (**Figura 3.5**).

Após validação, os SNVs passam por uma etapa importante – não excludente – que é a avaliação do impacto do SNV na função do gene afetado. Algoritmos de predição de impacto têm sido muito utilizados para a avaliação do impacto de mutações *missense,* e, indiretamente, na correlação do impacto com manifestações fenotípicas (ADZHUBEI; JORDAN; SUNYAEV, 2013; CHOI; CHAN, 2015; KUMAR; HENIKOFF; NG, 2009). O desenvolvimento de protocolos criteriosos de avaliação de impacto, que levam em consideração não só ferramentas de predição de impacto, mas também estudos funcionais e populacionais, é de extrema importância para a identificação do gene responsável pelo fenótipo também em doenças humanas. O guia mais utilizado para avaliação do impacto de variantes em humanos foi desenvolvido pela ACMG - *American College of Medical Genetics and Genomics* (RICHARDS et al., 2015). As principais variantes, encontradas neste estudo (**Tabela 3.7**), associadas aos mutantes podem ser classificadas como patogênicas de impacto forte ou muito forte segundo os critérios da ACMG, confirmando o sucesso do sequenciamento e análise dos dados.

Todas essas informações em conjunto indicam a alta probabilidade de que os SNVs detectados foram induzidos por ENU e prejudiquem a função dos genes afetados, consistindo de fato como a principal causa dos fenótipos observados. No geral, como discutido em (MORESCO; LI; BEUTLER, 2013), esse tipo de abordagem em genética direta têm tido sucesso em revelar a função de genes e no estabelecimento de modelos animais cada vez mais sofisticados. Porém, há uma influência considerável do *background* e também de mutações incidentais nas manifestações fenotípicas e o uso de ferramentas de edição genômica têm sido

consideradas como a validação definitiva na associação de variantes a fenótipos (SIMON et al., 2015).

3.4.1 Mutante bate palmas e a variante no gene kmt2d

Análises do sequenciamento do exoma do mutante indicaram um único SNV candidato localizado no gene *kmt2d*, conhecido anteriormente como *mll2 ou mll2-alr* (BÖGERSHAUSEN; WOLLNIK, 2013). O gene *kmt2d* é uma metiltransferase lisina (K) específica, cuja função primária é a metilação (mono, di ou principalmente trimetilação) do resíduo K4 em histonas H3, também conhecida como metilação H3K4 (RUTHENBURG; ALLIS; WYSOCKA, 2007). Esse tipo de metilação, principalmente a trimetilação H3K4 por Kmt2d, está associada com a modificação de histonas nas regiões 5' e o consequente aumento dos níveis de transcrição gênica de virtualmente todos os genes ativos (RUTHENBURG; ALLIS; WYSOCKA, 2007). O principal domínio responsável pela ação de metiltransferase é o domínio SET, que está também está presente no grupo de genes homólogos *trithorax* de *Drosophila melanogaster,* importantes para o desenvolvimento embrionário e envolvidos na regulação do padrão de genes *hox* (GLASER et al., 2006). A função dos componentes da família Kmt2d em camundongos parece não ser redundante e pode estar relacionada à formação de diferentes complexos proteicos, semelhantes à estruturação do complexo COMPASS em *Saccharomyces cerevisae,* relacionados ao balanço entre diferentes tipos de metilação e acetilação de histonas e também à diferente localização da expressão desses genes (EISSENBERG; SHILATIFARD, 2010). Os principais membros desse complexo incluem a proteína KDM6A, Menin, UTX e a porção CTD da RNA polimerase II (EISSENBERG; SHILATIFARD, 2010). O gene *km2td* também parece estar envolvido com o desenvolvimento de câncer e regulação da diferenciação de células embrionárias para tecido cardíaco (GUO et al., 2012).

Recentemente, com o advento de estudos amplos de NGS, houve a correlação entre a presença de mutações no gene *kmt2d* em humanos como a causa primária da síndrome de Kabuki (NGUYEN et al., 2011). A síndrome de Kabuki é uma doença congênita infantil rara, que afeta cerca de 1 em 32,000 nascimentos e é caracterizada por um espectro amplo de sintomas, que inclui anomalias cranofaciais e incapacidade

intelectual, muitas vezes confundida com o espectro autista (BÖGERSHAUSEN; WOLLNIK, 2013). A maioria das mutações encontradas nos pacientes acarreta em perda de função do gene *kmt2d* e estão localizadas na porção C-terminal, na região do domínio SET (BÖGERSHAUSEN; WOLLNIK, 2013; NGUYEN et al., 2011). A localização da mutação também parece influenciar diretamente o tipo de sintoma/anomalia encontrada nos pacientes, salientando uma interessante associação genótipo-fenotípica que pode estar relacionada diretamente com a função molecular do produto gênico (MAKRYTHANASIS et al., 2013).

Um estudo de 2014 (BJORNSSON et al., 2014) caracterizou um modelo murino com perda de função do gene *kmt2d* em heterozigose, chamado de Kmt2d$^{+/\beta Geo}$. Esse camundongo apresenta problemas de aprendizado e memória, além de malformações no hipocampo, características muito semelhantes às encontradas nos pacientes com síndrome de Kabuki, além da consequente diferença no padrão de metilação H3K4 global. Além disso, o estudo demonstrou que a administração da droga histona-deacetilante (HDAc) AR-42, usada em testes clínicos para tratamento de câncer de próstata (BUSH et al., 2012), foi capaz de reverter o fenótipo neurológico-comportamental dos camundongos mutantes adultos. A causa dessa reversão pode estar fundamentada na recuperação do equilíbrio da regulação gênica dada pela relação metilação/acetilação das histonas H3 (BJORNSSON et al., 2014).

Embora o diagnóstico da síndrome de Kabuki seja estabelecido somente por aspectos clínicos, mutações nos genes *kmt2d* ou *kd6ma* são atualmente consideradas como a base genética da síndrome (BÖGERSHAUSEN; WOLLNIK, 2013). Os estudos genéticos marcantes dessa associação foram publicados em 2011 (NGUYEN et al., 2011), com sequenciamento de grupos numerosos de pacientes diagnosticados com a síndrome. A maioria das variantes de Kmt2d identificadas nos indivíduos eram dominantes e truncavam a proteína antes da tradução do domínio SET, indicando a perda de função metiltransferase associada a esse domínio. Outras mutações puderam ser identificadas em outros pacientes, inclusive no Brasil (KOKITSU-NAKATA et al., 2012). Embora a maioria das mutações seja encontrada na porção C-terminal algumas mutações na região N-terminal também foram identificadas (BÖGERSHAUSEN; WOLLNIK, 2013). Da mesma forma, o modelo murino introduzido em 2014

(BJORNSSON et al., 2014) também possui um mutação que trunca a proteína antes do domínio SET. A mutação encontrada no mutante *bapa* pode constituir um modelo interessante para o estudo da função do gene *kmt2d* e possibilitar a investigação de uma cascata de sinalização por fosforilação associada a modificações epigenéticas. Essa modificação pode afetar a função metiltransferase de Kmt2d bem como também afetar as interações proteína-proteína e/ou sua localização intracelular. Portanto, considerando que a síndrome de Kabuki possa constituir um espectro de alterações fenotípicas, o camundongo *bapa* pode constituir um interessante modelo para o estudo da função do gene *kmt2d* e da própria síndrome de Kabuki.

O resíduo-alvo da mutação encontrada em Kmt2d (**Figura 3.7**) foi encontrado fosforilado em experimentos de fosfoproteômica em células humanas (MORITZ et al., 2010). A inibição de FLT3 (tirosina-quinase) afeta a fosforilação de Kmt2d em humanos (BEAUSOLEIL et al., 2011) portanto modificações pós-traducionais podem ser importantes na regulação da função, localização ou interação proteína-proteína de Kmt2d. De fato, o camundongo bate palmas aparentemente apresenta um desbalanço no padrão de metilação H3K4 (**Figura 3.8**), dando indícios fortes do impacto da mutação na proteína Kmt2d.

3.4.2 Mutante careca e a candidatos encontrados

O mutante careca (*carc*) é caracterizado por anormalidades no pelo, desde o surgimento da primeira pelagem. Os camundongos são viáveis e férteis, mas exibem alterações pronunciadas em toda pelagem, embora com vibrissas normais. O pelo é escasso durante toda a vida, especialmente em torno dos olhos e nas pernas e barriga. Entre 45 e 60 dias, a maioria dos camundongos perdem toda sua pelagem e sua pelagem apresenta falhas durante toda a vida. A análise histológica da pele de um adulto mostrou aumento na espessura da pele devido a um grande número de folículos em fase anágena, que é a fase de crescimento ativa quando a fibra de cabelo é produzida. Em camundongos controle BALB/c a maioria dos folículos encontram-se em fase telógena, conhecida como fase de repouso (Massironi, comunicação oral). Sendo assim o mutante careca é caracterizado pela perda progressiva de pelo em ambos os

sexos, com uma alteração significativa do ciclo dos folículos pilosos que se acumulam em fase anágena.

As análises de mapeamento por microssatélites indicaram uma região pequena, de cerca de 5 Mb, entre os marcadores D7Mit105 e D7Mit304 na parte proximal do cromossomo 7 (70-78 cM), porém nenhum candidato foi encontrado nessa região. Dessa forma, decidimos estender a busca em todo o cromossomo 7, o que poderia aumentar o número de candidatos para a mutação causativa do fenótipo. De fato, vários candidatos foram encontrados, mas consideramos que a mutação no gene *cars,* pela proximidade da região candidata e a mutação do gene *cyp2b9,* dado um provável fenômeno de *nonsense mediated mRNA decay* (NMD) e o envolvimento do gene na via de metabolismo da testosterona, são os candidatos mais fortes para esse mutante.

Canonicamente o fenômeno de NMD ocorre quando o códon de parada prematuro está situado de 50 a 55 nucleotídeos da borda do exon mais próximo e está diretamente envolvido na manifestação de doenças, como a fibrose cística (LINDE et al., 2007). O códon de parada prematuro do gene *cyp2b9* do mutante careca está a cerca de 40 nucleotídeos da borda mais próxima do exon, porém há relatos de NMD mesmo em casos em que o códon de parada prematura não obedece a regra dos 50-55 nucleotídeos de distância da borda, como observado no colágeno (FANG et al., 2013). Além disso, foi demonstrado que Cyp2b9 é alvo de regulação por microRNAs (INGELMAN-SUNDBERG et al., 2013) e que essas moléculas podem mediar o processo de NMD (ZHAO et al., 2014).

Em mamíferos, um dos mais importantes controles hormonais do crescimento de pelos em mamíferos são os hormônios andrógenos, cujos membros mais importantes são a testosterona (T) e a di-hidrotestosterona (DHT). Ambos os hormônios são secretados nos homens e nas mulheres, sendo que a di-hidrotestosterona (DHT) é considerada mais potente que a testosterona devido à sua alta afinidade por receptores andrógenos (INUI; ITAMI, 2013).

3.4.3 Mutante fraqueza e a variante no gene dst

Em relação ao camundongo fraqueza temos que proteínas dessa família estão associadas diretamente à manutenção da morfologia celular e são necessárias em

diversos processos fundamentais de diferenciação (SONNENBERG; LIEM, 2007). Mutações em genes que codificam plaquinas levam a degeneração neuronal, crescimento anormal de axônios e fragilidade de tecidos (LEUNG; GREEN; LIEM, 2002). Um dos principais modelos para o estudo de doenças severas relacionados ao movimento são os camundongos chamados de *dystonia musculorum* (*dt*). Atualmente existem vários camundongos *dt*, com mutações espontâneas ou induzidas, que são caracterizados pela perda progressiva, geralmente a partir de duas semanas, da coordenação dos membros e pela postura anormal das patas e tronco (POOL et al., 2005). A doença progride agressivamente e geralmente os camundongos não sobrevivem à terceira semana e morrem sem causa definida (STANLEY et al., 1988). O gene *dst* codifica isoformas epiteliais, neuronais e musculares da proteína distonina e somente a perda de função das isoformas neuronais e musculares estão associadas com o fenótipo dos camundongos *dt* (LEUNG; GREEN; LIEM, 2002). A proteína distonina é expressa durante todo o desenvolvimento do camundongo predominantemente em neurônios craniais, em gânglios espinais sensoriais e também no sistema motor extrapiramidal, cerebelo e em neurônios motores (LEUNG; GREEN; LIEM, 2002).

Os camundongos *dt* possuem características patológicas, como a degeneração dos neurônios sensoriais, anomalias nucleares e no retículo endoplasmático e desorganização das redes de microtúbulos no citoesqueleto aliadas ao fenótipo de perda de coordenação progressiva (BERNIER et al., 1995). Por esse motivo são muito usados como modelo no estudo de doenças neurodegenerativas humanas, como a distonia e a neuropatia hereditária sensorial e autonômica do tipo VI (FERRIER et al., 2014). O fenótipo descrito para os camundongos *dt* é muito similar ao camundongo *frqz,* reforçando a relação causal com a mutação encontrada no gene *dst* (**Figura 3.10**). Ensaios funcionais de RT-PCR visando a elucidação da provável consequência da mutação no *splicing* das isoformas de Dst, em tecidos onde as isoformas são expressas – como cérebro e cerebelo – serão essenciais para a confirmação do efeito do SNV e o estabelecimento do mutante fraqueza como modelo de estudo de síndromes de perda de coordenação progressiva.

3.4.4 Padrão global das mutações potencialmente induzidas por ENU

Embora a mutagênese por ENU seja considerada de certa forma aleatória, as mutações mais comumente observadas são transversões TA:AT ou transições TA:CG , observadas por ensaios *in vitro* e *in vivo* de mutagênese (JUSTICE et al., 1999; SHIBUYA; MORIMOTO, 1993). Não há um estudo abrangente, em escala genômica, do impacto da mutagênese por ENU no DNA. Assim, selecionamos mutações únicas de cada mutante em relação às outras amostras sequenciadas, de forma a selecionar mutações incidentais com maior chance de terem sido originadas por ENU. Em nosso estudo não foi possível detectar um maior número dessas mutações quando consideramos as mutações únicas e exclusivas de cada mutante. Porém, quando utilizamos uma abordagem de procura de assinaturas (**Figura 3.11**) foram encontradas pelo menos dois tipos de assinaturas (S1 e S2) com alta representatividade nos mutantes e que não foi encontrada no controle BALB/c (**Figura 3.11B** e **Figura 3.11C**). Essas assinaturas constituem uma maior proporção de transversões T>A e transições T>C, embora sutis. Acreditamos que as várias gerações de criação desses animais pode ter eliminado boa parte das mutações induzidas inicialmente.

3.5 Referências

ADZHUBEI, I.; JORDAN, D.; SUNYAEV, S. Predicting Functional Effect of Human Missense Mutations Using PolyPhen-2. **Current Protocols in Human Genetics**, v. 2, 2013.

ANTONARAKIS, S. E.; BECKMANN, J. S. Focus on Monogenic Disorders. **Nature reviews. Genetics**, v. 7, n. April, p. 277–282, 2006.

ARNOLD, C. N. et al. Rapid identification of a disease allele in mouse through whole genome sequencing and bulk segregation analysis. **Genetics**, v. 187, n. 3, p. 633–41, mar. 2011.

ARNOLD, C. N. et al. ENU-induced phenovariance in mice: inferences from 587 mutations. **BMC research notes**, v. 5, n. 1, p. 577, jan. 2012.

BEAUSOLEIL, S. et al. Survey of Activated FLT3 Signaling in Leukemia. **PloS one**, v. 6, n. 4, p. 1–10, 2011.

BECK, J. A et al. Genealogies of mouse inbred strains. **Nature genetics**, v. 24, n. 1, p.

23–5, jan. 2000.

BEIER, D. R.; HERRON, B. J. Genetic mapping and ENU mutagenesis. **Genetica**, v. 122, n. 1, p. 65–69, 2004.

BJORNSSON, H. T. et al. Histone deacetylase inhibition rescues structural and functional brain deficits in a mouse model of Kabuki syndrome. **Science translational medicine**, v. 135, 2014.

BÖGERSHAUSEN, N.; WOLLNIK, B. Unmasking Kabuki syndrome. **Clinical Genetics**, v. 83, n. 3, p. 201–211, 2013.

BOLES, M. K. et al. Discovery of candidate disease genes in ENU-induced mouse mutants by large-scale sequencing, including a splice-site mutation in nucleoredoxin. **PLoS Genetics**, v. 5, n. 12, 2009.

BOUCHERIE. C.; BOUTIN, C.; JOSSIN, Y., SCHAKMAN, O.; GOFFINET, A.M.; RIS, L.; GAILLY, P.; TISSIR, F. Neural progenitor fate decisions defects, cortical hypoplasia and behavioral impairment in Celsr1-deficient mice. **Mol Psychiatry**, In Press, 2017.

BULL, K. R. et al. Unlocking the bottleneck in forward genetics using whole-genome sequencing and identity by descent to isolate causative mutations. **PLoS genetics**, v. 9, n. 1, p. e1003219, jan. 2013.

CHOI, Y.; CHAN, A. P. PROVEAN web server: A tool to predict the functional effect of amino acid substitutions and indels. **Bioinformatics**, v. 31, n. 16, p. 2745–2747, 2015.

COOPER, D. N. et al. On the sequence-directed nature of human gene mutation The role of genomic architecture and the local DNA sequence environment in mediating gene mutations underlying human inherited disease. **Human Mutation**, v. 32, n. 10, p. 1075–1099, 2011.

CORDES, S. P. N -Ethyl- N -Nitrosourea Mutagenesis : Boarding the Mouse Mutant Express. **Microbiology and molecular biology reviews**, v. 69, n. 3, p. 426–439, 2005.

DANECEK, P. et al. The variant call format and VCFtools. **Bioinformatics (Oxford, England)**, v. 27, n. 15, p. 2156–8, 1 ago. 2011.

EISSENBERG, J. C.; SHILATIFARD, A. Histone H3 lysine 4 H3K4 methylation in development and differentiation. **Developmental Biology**, v. 339, n. 2, p. 240–249, 2010.

ENDERS, A. et al. Finding new immune regulatory genes by ENU mutagenesis. **Journal of Translational Medicine**, v. 10, n. Suppl 3, p. I6, 2012.

ERMANN, J.; GLIMCHER, L. H. After GWAS: mice to the rescue? **Current opinion in immunology**, v. 24, n. 5, p. 564–70, out. 2012.

FAIRFIELD, H. et al. Mutation discovery in mice by whole exome sequencing. **Genome biology**, v. 12, n. 9, p. R86, 2011.

FAIRFIELD, H. et al. Exome sequencing reveals pathogenic mutations in 91 strains of mice with Mendelian disorders. **Genome Research**, v. 25, p. 948–957, 2015.

FANG, Y. et al. Nonsense-mediated mRNA decay of collagen -emerging complexity in RNA surveillance mechanisms. **Journal of cell science**, v. 126, n. Pt 12, p. 2551–60, 2013.

FARRELL, A. et al. Whole genome profiling of spontaneous and chemically induced mutations in Toxoplasma gondii. **BMC genomics**, v. 15, n. 1, p. 354, 2014.

FERRIER, A. et al. Transgenic expression of neuronal dystonin isoform 2 partially rescues the disease phenotype of the dystonia musculorum mouse model of hereditary sensory autonomic neuropathy VI. **Human Molecular Genetics**, v. 23, n. 10, p. 2694–2710, 2014.

GLASER, S. et al. Multiple epigenetic maintenance factors implicated by the loss of Mll2 in mouse development. **Development (Cambridge, England)**, v. 133, n. 8, p. 1423–32, abr. 2006.

GONDO, Y. Now and future of mouse mutagenesis for human disease models. **Journal of Genetics and Genomics**, v. 37, n. 9, p. 559–572, 2010.

GUÉNET, J. L. The mouse genome. **Genome Research**, v. 15, p. 1729–1740, 2005.

GUO, C. et al. Global identification of MLL2-targeted loci reveals MLL2's role in diverse signaling pathways. **Proceedings of the National Academy of Sciences of the United States of America**, 23 out. 2012.

HERRON, B. J. et al. Efficient generation and mapping of recessive developmental mutations using ENU mutagenesis. **Nature genetics**, v. 30, n. february, p. 185–189, 2002.

INGELMAN-SUNDBERG, M. et al. Special Section on Epigenetic Regulation of Drug Metabolizing Enzymes and Transporters — Symposium Report Potential Role of Epigenetic Mechanisms in the Regulation of Drug Metabolism and Transport. **Drug Metabolism and Disposition**, p. 1725–1731, 2013.

INUI, S.; ITAMI, S. Androgen actions on the human hair follicle: perspectives. **Experimental Dermatology**, v. 22, n. 3, p. 168–171, 2013.

JUSTICE, M. J. et al. Mouse ENU Mutagenesis. **Human Molecular Genetics**, v. 8, n. 10, p. 1955–1963, 1999.

JUSTICE, M. J.; DHILLON, P. Using the mouse to model human disease : increasing validity and reproducibility. **Disease Models & Mechanisms**, v. 9, p. 101–103, 2016.

KILE, B. T.; HILTON, D. J. The art and design of genetic screens: mouse. **Nature Reviews Genetics**, v. 6, n. 7, p. 557–567, 2005.

KOELSCH, B.U.; KINDLER-RÖHRBORN, A. Neuro-oncogenesis induced by nitroso compounds in rodents and strain-specific genetic modifiers of predisposition. In: **CNS Cancer: Models, Markers, Prognostic Factors, Targets, and Therapeutic Approaches.** Ed: Erwin G. Van Meir. Springer Science & Business Media, 2009.

KOKITSU-NAKATA, N. M. et al. Analysis of MLL2 gene in the first Brazilian family with Kabuki syndrome. **American journal of medical genetics. Part A**, ago. 2012.

KUMAR, P.; HENIKOFF, S.; NG, P. C. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. **Nature protocols**, v. 4, n. 8, p. 1073–1082, 2009.

LIM, K. H.; FAIRBROTHER, W. G. Spliceman — a computational web server that predicts sequence variations in pre-mRNA splicing. **Bioinformatics**, v. 28, n. 7, p. 1031–1032, 2012.

LINDE, L. et al. Nonsense-mediated mRNA decay affects nonsense transcript levels and governs response of cystic fibrosis patients to gentamicin. **Journal of Clinical Investigation**, v. 117, n. 3, p. 683–692, 2007.

MAKRYTHANASIS, P. et al. MLL2 mutation detection in 86 patients with Kabuki syndrome: a genotype-phenotype study. **Clinical genetics**, v. 84, n. 6, p. 539–45, dez. 2013.

MARTELOTTO, L. G. et al. Benchmarking mutation effect prediction algorithms using functionally validated cancer-related missense mutations. **Genome Biology**, v. 15, n. 10, p. 484, 2014.

MASSIRONI, S. M. G. et al. Inducing mutations in the mouse genome with the chemical mutagen ethylnitrosourea. **Brazilian journal of medical and biological research**, v. 39, n. 9, p. 1217–26, set. 2006.

MORAN, J. L. et al. Utilization of a whole genome SNP panel for efficient genetic mapping in the mouse. **Genome research**, p. 436–440, 2006.

MORESCO, E. M. Y.; LI, X.; BEUTLER, B. Going forward with genetics: recent technological advances and forward genetics in mice. **The American journal of pathology**, v. 182, n. 5, p. 1462–73, maio 2013.

MORITZ, A. et al. Akt-RSK-S6 kinase signaling networks activated by oncogenic receptor tyrosine kinases. **Science signaling**, v. 3, n. 136, p. ra64, jan. 2010.

NGUYEN, N. et al. Random mutagenesis of the mouse genome : a strategy for discovering gene function and the molecular basis of disease. **Gastrointestinal and Liver Physiology**, v. 300, 2011.

NOLAN, P. M.; HUGILL, A.; COX, R. D. ENU mutagenesis in the mouse: Application to human genetic disease. **Briefings in Functional Genomics and Proteomics**, v. 1, n. 3, p. 278–289, 2002.

NOVEROSKE, J. K.; WEBER, J. S.; JUSTICE, M. J. The mutagenic action of N-ethyl-N-nitrosourea in the mouse. **Mammalian Genome**, v. 483, p. 478–483, 2000.

OLIVEIRA, N. S. **Caracterização fenotípica do camundongo mutante bate palmas induzido pelo agente mutagênico químico ENU (N- Ethyl- N- Nitrosourea) como potencial modelo para a síndrome de Kabuki**. [s.l.] Universidade de São Paulo, 2017.

OLIVER, P. L.; DAVIES, K. E. New insights into behaviour using mouse ENU mutagenesis. **Human Molecular Genetics**, v. 21, n. 1, p. 72–81, 2012.

PATTON, E. E.; ZON, L. I. The art and design of genetic screens: zebrafish. **Nature Reviews Genetics**, v. 2, n. December, p. 956–966, 2001.

POOL, M.; BOUDREAU LARIVIÈRE, C.; BERNIER G.; YOUNG, K.G; KOTHARY, R. Genetic alterations at the Bpag1 locus in dt mice and their impact on transcript expression. **Mamm. Genome.** n. 16 p. 909–917, 2005.

RHODES, M. et al. A High-Resolution Microsatellite Map of the Mouse Genome. **Genome research**, p. 531–542, 1998.

RICHARDS, S. 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**, v. 17, n. 5, p. 405–423, 2015.

RUTHENBURG, A. J.; ALLIS, C. D.; WYSOCKA, J. Methylation of lysine 4 on histone H3: intricacy of writing and reading a single epigenetic mark. **Molecular cell**, v. 25, n. 1, p. 15–30, 12 jan. 2007.

SCHWEINGRUBER, C. et al. Nonsense-mediated mRNA decay - Mechanisms of substrate mRNA recognition and degradation in mammalian cells. **Biochimica et Biophysica Acta - Gene Regulatory Mechanisms**, v. 1829, n. 6–7, p. 612–623, 2013.

SHIBUYA, T.; MORIMOTO, K. A review of the genotoxicity of 1-ethyl-1-nitrosourea. **Mutation research**, v. 297, p. 3–38, 1993.

SIEVERS, F. et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. **Molecular Systems Biology**, v. 7, n. 1, p. 539–539, 2014.

SILVA E SILVA, D.A; GODARD, A.L.B. Mapeamento genético e estudo de genes candidatos para um modelo animal de doença neuromuscular. Dissertação de Mestrado, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, 2012.

SIMON, M. M. et al. Current strategies for mutation detection in phenotype-driven screens utilising next generation sequencing. **Mammalian Genome**, v. 26, n. 9, p. 486–500, 2015.

SUN, M. et al. Multiplex Chromosomal Exome Sequencing Accelerates Identification of ENU-Induced Mutations in the Mouse. **G3 (Bethesda, Md.)**, v. 2, n. 1, p. 143–50, jan. 2012.

TAKAHASI, K. R.; SAKURABA, Y.; GONDO, Y. Mutational pattern and frequency of induced nucleotide changes in mouse ENU mutagenesis. **BMC molecular biology**, v. 8, p. 52, jan. 2007.

TANG, S. et al. DiBayes: A SNP detection algorithm for next-generation dibase sequencing.http://www3.appliedbiosystems.com/cms/groups/mcb_marketing/documents/generaldocuments/cms_057817.pdf, 2008

TANISAWA, K. et al. Exome sequencing of senescence-accelerated mice (SAM) reveals deleterious mutations in degenerative disease-causing genes. **BMC genomics**, v. 14, p. 248, jan. 2013.

WANG, K.; LI, M.; HAKONARSON, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. **Nucleic acids research**, v. 38, n. 16, p. e164, set. 2010.

WARR, A. et al. Exome Sequencing : Current and Future Perspectives. **G3 (Bethesda, Md.)**, v. 5, n. August, p. 1543–1550, 2015.

ZHANG, F.; LUPSKI, J. R. Non-coding genetic variants in human disease. **Human Molecular Genetics**, v. 24, n. July, p. 102–110, 2015.

ZHAO, Y. et al. MicroRNA-mediated repression of nonsense mRNAs. **eLife**, p. e03032, 2014.

CAPÍTULO 4 - DESENVOLVIMENTO DE UMA FERRAMENTA PARA ANÁLISE DO PADRÃO DE MUTAÇÕES PONTUAIS

CAPÍTULO 4 - DESENVOLVIMENTO DE UMA FERRAMENTA PARA ANÁLISE DO PADRÃO DE MUTAÇÕES PONTUAIS

4.1 Introdução

O genoma é um alvo constante de alterações vindas de metabólitos celulares endógenos, agentes mutagênicos externos e da própria instabilidade química espontânea do DNA. Dessa forma, mecanismos moleculares de reparo de danos ao DNA se desenvolveram ao longo da evolução para corrigir esses danos de maneira rápida e eficaz. Porém, a própria vida depende de um equilíbrio fino e dinâmico da diversidade promovida pela mutação. Na maioria das células somáticas, o acúmulo de mutações é danoso para o ser vivo, levando a processos associados à morte celular, envelhecimento e câncer (MENCK; MUNFORD, 2014). Porém, em alguns tipos de células somáticas, como células do sistema imune, o acúmulo de mutações em *loci* específicos gera diversidade de receptores e anticorpos para reconhecer patógenos, por exemplo (revisado por TENG; PAPAVASILIOU, 2007). Em células germinativas, o acúmulo de mutações também pode ser danoso, gerando alterações que podem levar a síndromes hereditárias (ANTONARAKIS; BECKMANN, 2006), mas ao mesmo tempo a mutação é considerada nesse contexto a fonte primária de diversidade necessária para processos adaptativos dos seres vivos.

Esse equilíbrio fino entre mudança e permanência é um dos segredos para a manutenção da vida. Entender esse balanço envolvendo a natureza dupla da mutação é essencial para compreender importantes mecanismos biológicos e solucionar importantes desafios na saúde humana, ecologia, conservação ambiental e na agricultura. Antes da era de genômica em larga escala, a maioria dos estudos de mutagênese era restrito a pequenas sequências e poucos genes, por exemplo *TP53,* o gene que codifica a proteína p53 (GLAZKO; MILANESI; ROGOZIN, 1998; ROGOZIN et al., 2003). A genômica tem contribuído muito para o entendimento dessas questões, com o uso de tecnologias de sequenciamento cada vez mais baratas e rápidas (KOBOLDT et al., 2013).

A maioria das análises em larga escala de identificação de mutações é oriunda de análises de ressequenciamento, que consiste no alinhamento de leituras à um genoma referência e a posterior chamada de variações. Esse tipo de análise é

normalmente utilizado para a identificação de trocas de um único nucleotídeo, que podem estar associadas a patologias específicas. Portanto, o objetivo principal é quase sempre identificar o impacto de variantes raras no produto gênico associado, e por consequência, associá-lo ao fenótipo (DO; KATHIRESAN; ABECASIS, 2012). Dependendo do estudo em questão, no entanto, a avaliação do padrão de mutação global pode ser muito relevante. Um exemplo é a identificação de padrões de mutações somáticas em câncer (ALEXANDROV et al., 2013a, 2013b). Foram descobertas assinaturas mutagênicas específicas para vários tipos de câncer, como o câncer de mama (NIK-ZAINAL et al., 2012), através da caracterização das trocas e da sequência contexto das mutações pontuais utilizando banco de dados de mutações somáticas de vários pacientes. Embora algumas dessas assinaturas sejam específicas e possam ser úteis em termos de diagnóstico, elas indicam alguns padrões que podem estar associados à processo de reparo específicos e a regiões enriquecidas em mutações (ALEXANDROV et al., 2013b). A identificação desses padrões pode revelar as bases genéticas de transformação maligna e indicar formas de tratamento específicas e eficientes.

Os processos mutagênicos estão historicamente muito associados ao estudo do câncer, porém estudos de análise do padrão de mutações em outras situações e outros organismos são muito importantes. Dentre esses estudos destacamos, por exemplo, a avaliação do impacto global de agentes mutagênicos no genoma (POON et al., 2014), seleção e aquisição de resistência a medicamentos por bactérias (LÁZÁR et al., 2014), toxicogenômica (BESARATINIA et al., 2012) e estudos ecológicos de adaptação em populações (STAPLEY et al., 2010). A ferramenta *woland* se encaixa justamente nesse contexto, fornecendo um tipo de análise voltado para a identificação de padrões de mutações em diferentes tipos de dados de ressequenciamento, independentemente do organismo e plataforma de obtenção dos dados, e que ofereça diferentes análises do processo e impacto das mutações pontuais no objeto de estudo.

4.2 Implementação

Woland é uma ferramenta multiplataforma em Perl e R para análise de padrões de mutação pontuais em um determinado conjunto de dados de SNVs. *Woland* é capaz

de: quantificar o tipo de cada troca de nucleotídeos; identificar regiões enriquecidas com mutações (*hotspots*); extrair as sequências contexto de cada SNV; quantificar os motivos (ou *motifs*) de agentes mutagênicos e identificar a presença de viés de fita (*strand bias*) em relação a esses motivos. *Woland* é compatível com dados de ressequenciamento de genoma total (*whole-genome*) ou de enriquecimento (*targeted-resequencing*), tal como exomas. O objetivo de *Woland* é fornecer ao usuário informações simplificadas sobre prováveis padrões de mutação relacionados ao tipo de mutação pontual, posicionamento e mutações originadas por agentes mutagênicos em uma determinada amostra.

Woland está disponível para download pelo GitHub[25], juntamente com instruções simplificadas de instalação e um manual[26] completo de utilização.

4.3 Métodos

O ressequenciamento convencional, sem nenhum tipo de preparo de amostra especial, não permite a identificação das mutações pontuais em relação à fita senso (+ ou 5'-3') ou anti-senso (- ou 3'-5'). As *pipelines* de chamadas de SNP, assinalam arbitrariamente o alelo referência e o alelo alterado sempre na fita senso. Dessa forma, se uma mutação ocorre na fita anti-senso ela será reparada e replicada de forma a alterar, de forma complementar, a fita senso. Portanto, *woland* ignora *a priori* a informação de localização de uma mutação pontual em relação fita senso ou anti-senso. Os SNPs chamados podem ser fornecidos através de uma tabela de SNVs em formato .VCF (*Variant Calling Format),* que agrupa diversas informações como zigosidade e qualidade para cada SNV detectado. Para interpretar biologicamente esses dados são utilizados etapas de filtragem e anotação de acordo com o objetivo do experimento.

---

[25] *Woland* – GitHub: http://www.github.com/tiagoantonio
[26] *Woland* – ReadTheDocs: http://woland.readthedocs.io/en/latest/

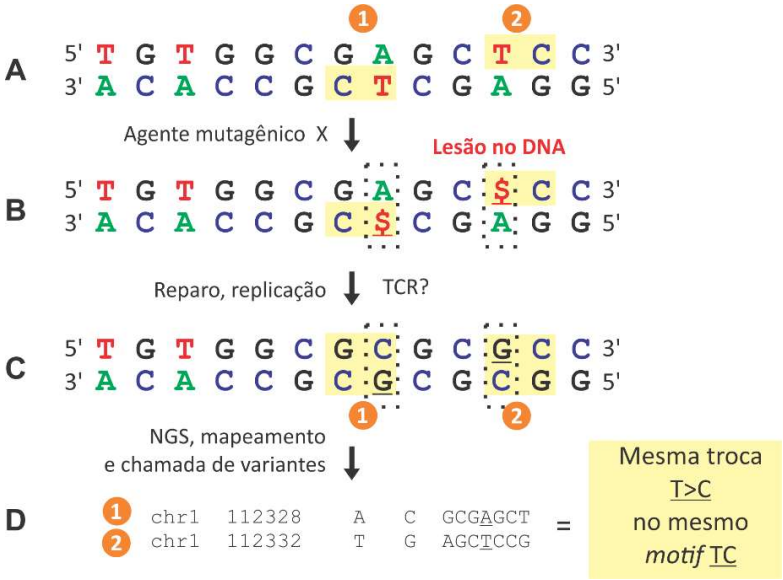

Figura 4.1 Como as bases que sofreram mutação são detectadas por NGS. (**A**) Um agente mutagênico X altera uma timina quando é seguida por uma citosina 3'. Dois *motifs* TC (**1** e **2**) estão destacados na sequência. (**B**) O agente mutagênico X altera a estrutura química da timina induzindo uma lesão no DNA. (**C**) Bases alteradas são submetidas ao reparo de DNA e ao processo de replicação e a timina sofre uma mutação para guanina nos dois casos. (**D**) O sequenciamento NGS seguido pelo mapeamento e chamada de *variants* fornece as trocas baseados na fita 5'3', o que significa que a troca T>C ocorreu no mesmo *motif* TC.

Woland não realiza mapeamento de leituras, chamada de SNPs, anotação ou filtragem e, portanto, são consideradas como uma etapa de pré-análise (**Figura 4.2**). O arquivo de entrada primário para *woland* é um arquivo tabular dos SNVs, que pode ser gerado diretamente pela ferramenta de anotação ANNOVAR (arquivo *variant_function*) ou formatado manualmente pelo usuário (arquivo TAB). O usuário também deve fornecer como *inputs* secundários: um arquivo de perfil cromossômico; o genoma referência em formato FASTA e o arquivo de anotação do genoma referência em formato RefSeq. O arquivo de SNVs (*variant_function* ou TAB) é utilizado para cinco processos automáticos: contagem dos tipos de trocas; janela deslizantes para *hotspots*; extração das sequências contexto; busca por assinaturas mutagênicas e concordância de fita (**Figura 4.2**). Cada um desses processos será discutido a seguir.

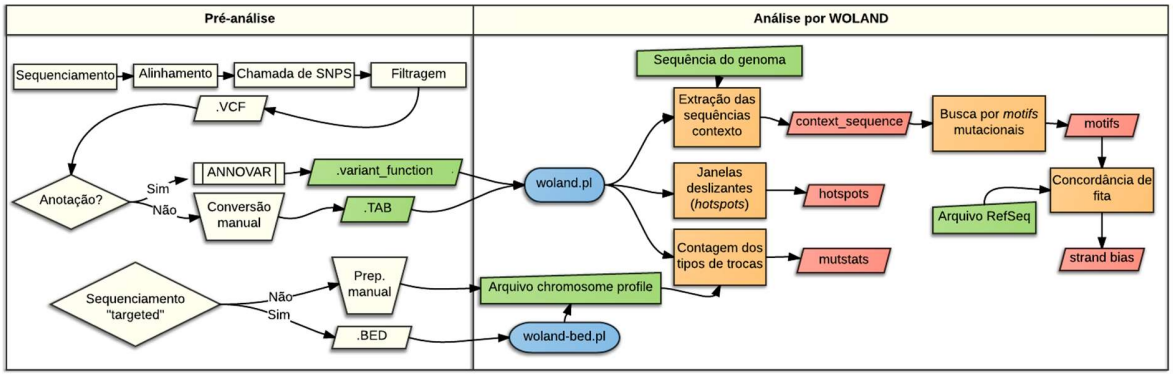

Figura 4.2 Fluxograma do funcionamento da ferramenta woland.

4.3.1 Contagem dos tipos de trocas.

Esse processo realiza a importação da lista de SNVs no formato de dados tabulados e conta em termos absolutos e de frequência os tipos de troca da base nitrogenada da amostra. As trocas comumente são classificadas em seis diferente grupos: C:G>A:T, C:G>G:C, C:G>T:A, T:A>A:T, T:A>C:G e T:A>G:C. Em woland a notação dos tipos de trocas foram simplificadas para C>A, C>G, C>T, T>A, T>C e T>G, tendo como referência pirimidina da base de Watson-Crick mutada. Além disso, as trocas também são agrupadas em transições e transversões.

O *input* secundário de perfil de cromossomos (*chromosome profile*) é um arquivo tabular com o tamanho das sequências de cada cromossomo. Esse arquivo pode ser editado manualmente, de acordo com o genoma referência, ou gerado a partir de um arquivo BED (*targeted-resequencing)*. *Woland* pode converter um arquivo de coordenadas BED em um arquivo de sumário da soma do total de regiões alvo para cada cromossomo, criando um perfil da soma do tamanho total das regiões-alvo. O *input* de perfil de cromossomos é utilizado para os cálculos de frequência das trocas e para o cálculo da quantidade de mutações associada a cada cromossomo. O arquivo de saída desse processo é um arquivo tabular chamado de *mutstats.* Esse arquivo agrupa todos os resultados de classificação, contagem e frequência apresentados anteriormente para cada amostra analisada.

4.3.2 Janelas deslizantes para identificação de potenciais hotspots.

O objetivo desse processo é identificar regiões enriquecidas com mutações pontuais. Para isso, é realizada a contagem das mutações ao redor de cada linha do arquivo de entrada, independentemente do tipo de troca encontrado. A janela ao redor do SNV é definida por N nucleotídeos que flanqueiam a coordenada do SNV, pelo usuário. A contagem é feita com base em um intervalo (N+x) e (N-x), considerando chrZ:x a coordenada para um determinado SNV. Cada mutação pontual é contada pelo menos uma vez, sem considerar o tipo de troca do nucleotídeo. O arquivo de saída desse processo é um arquivo tabular, chamado de *hotspots,* em que cada SNV é assinalado o número de mutações encontradas dentro do intervalo (ou janela) definido pelo usuário. Os *hotspots* podem ser considerados sinônimos para o termo *kaetegis* (ALEXANDROV et al., 2013a), dado para o fenômeno de hipermutação localizada (*foci*), que parece ser comum em alguns tipos de câncer.

4.3.3 Extração das sequências-contexto de cada SNV.

Esse é um processo simples para extração da sequência contexto de cada SNV baseado no genoma referência fornecido. A sequência contexto é definida por três nucleotídeos ao redor de cada SNV, em um intervalo (3+x) e (3-x) considerando o SNV chrZ:x. A sequência extraída inclui apenas o alelo referência do SNV. O arquivo de saída é um arquivo FASTA *context_sequence_logo, em que as* sequências são fornecidas com o identificador >chrZ_x. Esse arquivo pode ser utilizado em ferramentas como o Weblogo para análise de frequência de nucleotídeos ao redor de cada mutação. Um segundo arquivo de saída, chamado de *context_sequence_motif* é gerado se o usuário forneceu os SNPs anotados. Esse arquivo é utilizado para a busca de motivos referentes a agentes mutagênicos pelo processo a seguir.

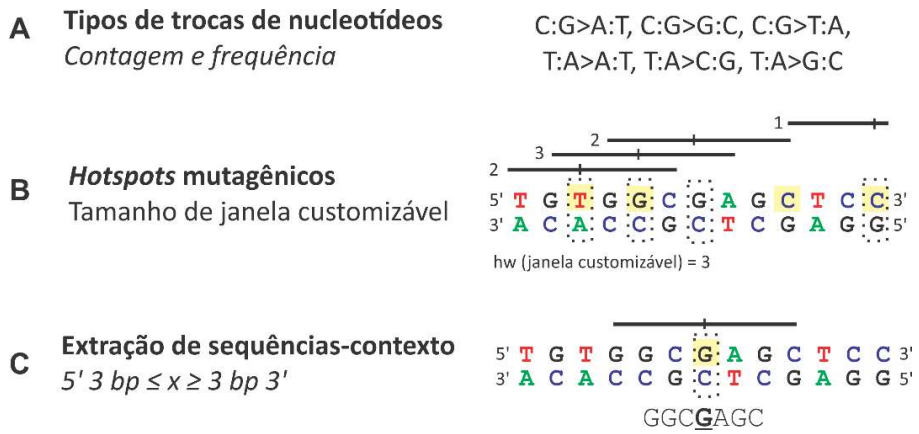

Figura 4.3 Métodos usados pela ferramenta woland. (**A**) Tipos de trocas de nucleotídeos. (**B**) Busca por regiões enriquecidas em mutações (*hotspots).* (**C**) Extração das sequências-contexto a partir do genoma de referência, que serão utilizados para a busca de *motifs* associados a agentes mutagênicos.

4.3.4 Busca por motivos referentes a agentes mutagênicos.

Alguns agentes mutagênicos possuem assinaturas canônicas definidas por sequências de nucleotídeos, e também preferência quanto ao tipo de base mutada (**Tabela 4.1**). O processo de busca por motivos referentes a agentes mutagênicos utiliza a sequências-contexto de cada SNV para associar as assinaturas às mutações detectadas em uma determinada amostra. Porém, novamente não é possível afirmar em que fita de fato ocorreu a lesão. Por isso, são consideradas também as sequências reverso-complementares de cada assinatura. Cada sequência-contexto oriunda do SNV pode ser associada a mais de um motivo canônico. Dessa forma, além do número de cada tipo de motivo encontrado, é realizada uma normalização pelo número total de SNVs.

A saída desse processo consiste em dois arquivos tabulares, com informações resumidas de contagem brutas e normalizas das assinaturas e o tipo de motivo associada a cada SNV.

Tabela 4.1. Painel de agentes mutagênicos e motivos utilizadas por woland.

Agente mutagênico	*Motivo*	*Sequências consideradas*	*Referências*
Agentes alquilantes do tipo SN1	RG	nnAGnn; nnGGnn; nnnCTnn; nnnCCnn	(ROGOZIN et al., 2003)
Hotspots de erros da DNA polimerase η	WA	nnAAnnn; nnTAnnn; nnnTTnn; nnnTAnn	(ROGOZIN et al., 2003)
8-oxoguanina	RGR	nnAGAnn; nnGGGnn; nnAGGnn; nnGGAnn; nnTCTnn; nnCCCnn; nnCCTnn; nnTCCnn	(ROGOZIN et al., 2003)
Luz UV (fago lambda)	YY	nnTCnn; nnTCnnn; nnnCTnn; nnCTnnn; nnTTnnn; nnnTTnn; nnCCnnn; nnnCCnn; nnnGAnn; nnnAAnn; nnAAnnn; nnnGGnn; nnGGnnn	(ROGOZIN et al., 2003)
Luz UV solar	YCG	nnTCGnn; nnnTCGn; nTCGnnn; nnCCGnn; nnnCCGn; nCCGnnn. nnCGAnn. nnnCGAn. nCGAnnn; nnCGGnn; nnnCGGn; nCGGnnn	(IKEHATA et al., 2014)
Fotoprodutos de Pirimidina-pirimidona (6-4)	YTCA	TTCA; CTCA; TGAA; TGAG	(ROGOZIN et al., 2003)
ENU	SWS	nnCAGnn; nnGACnn; nnGAGnn; nnCACnn; nnCTGnn. nnGTCnn; nnGTGnn; nnCTCnn	(ARNOLD et al., 2012; BARBARIC; WELLS, 2007)

R= A ou G; W= A ou T; Y= T ou C; S= G ou C

4.3.5 Concordância de fita (motivos com assimetria de fita).

Alguns estudos sugerem que o reparo de lesões mutagênicas é mais eficiente quando é acoplado à transcrição, no processo chamado de reparo acoplado à transcrição. Embora não seja possível identificar em que fita do DNA a mutação pontual ocorreu, podemos identificar em que fita do DNA o motivo foi encontrado considerando a assinatura canônica 5' – 3'. Além disso, podemos determinar, com as informações de anotação do genoma referência, qual fita é transcrita onde o motivo foi encontrado. Assim, para cada motivo associado a um determinado SNV temos:

$$SC = TS - MS$$

MS (*Motif Score*): Se o motivo mutagênico encontrado estiver localizado na fita *plus* MS=1 e se estiver localizado na fita *minus* MS=0 (**Figura 4.4A**)

TS (*Transcript Score*): Medida da preferência da fita transcrita (**Figura 4.4B**), de acordo com a equação:

$$\frac{1}{n}\sum_{i=1}^{n} T_i$$

Em que n = número de transcritos e para cada transcrito é dado o valor T=0, se a fita transcrita é a fita *minus* ou o valor T=1, se a fita é a *plus*. Valores de TS próximos ou iguais a 0 indicam que a transcrição ocorre na fita *minus* e valores próximos ou iguais a 1 indicam que a transcrição ocorre na fita *plus.*

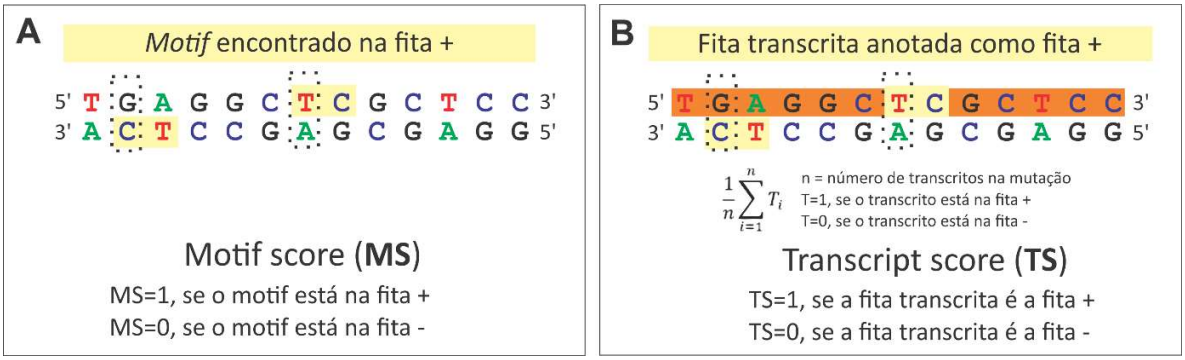

Figura 4.4 Concordância entre *motifs* e fitas transcritas. Esquema para cálculo do *motif score* (MS) (**A**) e *transcript score* (TS) (**B**).

SC (*Strand Score*): Cálculo da concordância entre a fita transcrita e a fita do motivo mutagênico associado a cada SNV. Valores de SC iguais a 0 indicam que há concordância entre a fita transcrita e o motivo e valores diferentes de 0 indicam discordância. SC<0 indicam que o motivo foi encontrado na fita *plus* e o transcrito corresponde majoritariamente à fita *minus* e valores de SC>0 indicam que o motivo foi encontrado na fita *minus* e o transcrito corresponde majoritariamente à fita *plus*.

4.4 Resultados

Os outputs originados por *woland* permitem uma série de tipos de análises. A seguir demonstramos exemplos de resultados originados por cada um dos *outputs* de *woland*, utilizando dados do TGCA (*The Cancer Genome Atlas*) provenientes de um total de 822 pacientes. Amostras sequenciadas de três diferentes tipos de câncer foram utilizadas: 306 amostras de pacientes com HNSC (*Head-Neck Squamous*

Cell Carcinoma), 176 amostras de LUSC (*Lung Squamous Cell Carcinoma*) e 340 amostras de pacientes com SKCM (*Skin Cutaneous Melanoma*). Os resultados apresentados aqui são utilizados apenas como um exemplo prático do uso de woland.

4.4.1 Contagem dos tipos de trocas.

A contagem do tipo de trocas de nucleotídeos, fornecida como *output* pelo *woland-mutstats*, pode ser utilizada para comparação de frequências e número absoluto de mutações. Considerando o número total de mutações podemos avaliar a frequência média de mutação em cada cromossomo, assim como a taxa de mutação média global (**Figura 4.5**). Pela análise da figura, podemos dizer que amostras de SKCM possuem, em média, taxa global de mutação superiores às amostras LUSC e HNSC.

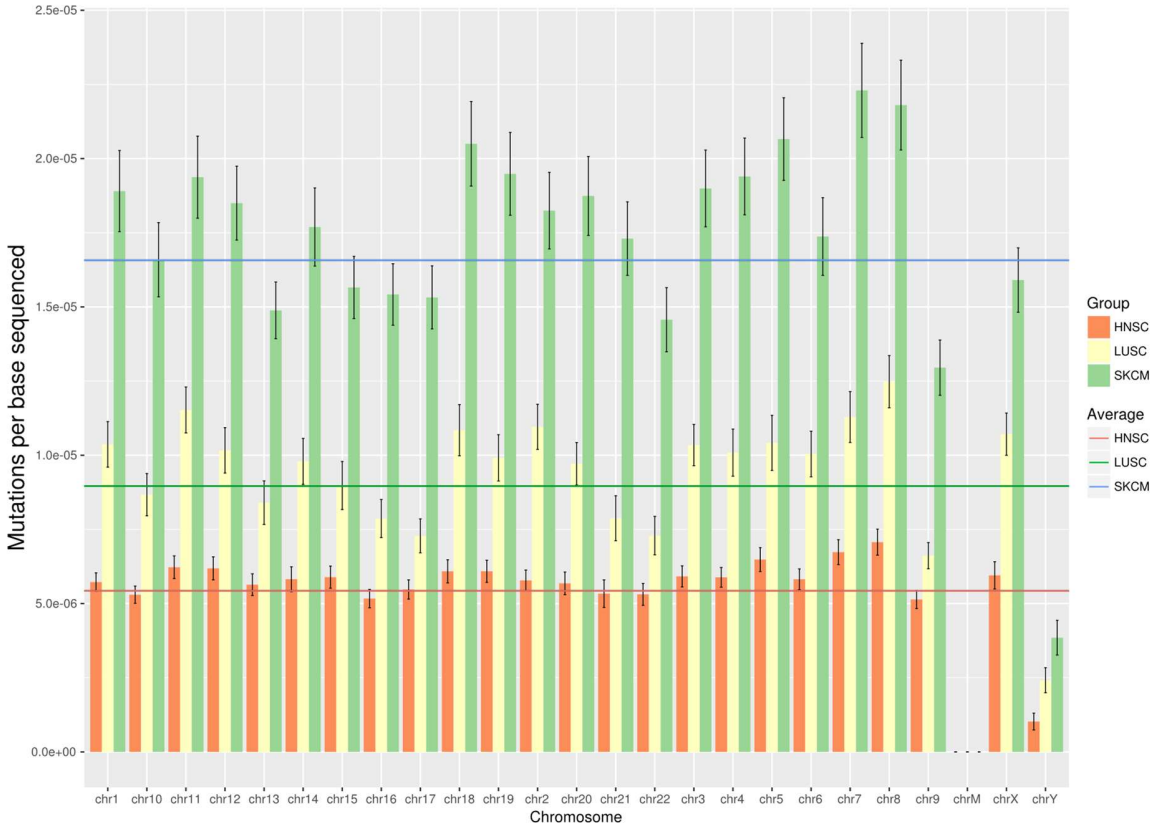

Figura 4.5 Frequência de mutações por base sequenciada em cada cromossomo. As barras indicam a média de mutações por base sequenciada de cada grupo de câncer (HNSC, LUSC e SKCM). As linhas indicam a taxa de mutação média global de cada grupo. A barra de erro indica o desvio-padrão.

Um dos aspectos globais a serem analisados em termos de padrão de mutação é a razão transições e transversões. Woland fornece uma saída gráfica interessante que calcula a frequência de transições e transversões para cada amostra em cada grupo experimental. A **Figura 4.6** demonstra que a maioria das amostras do grupo LUSC possuem mais transversões do que transições em relação as outras amostras enquanto o grupo SCKM possui uma relação de transições/tranversões muito maior em relação aos outros grupos experimentais (**Figura 4.6**).

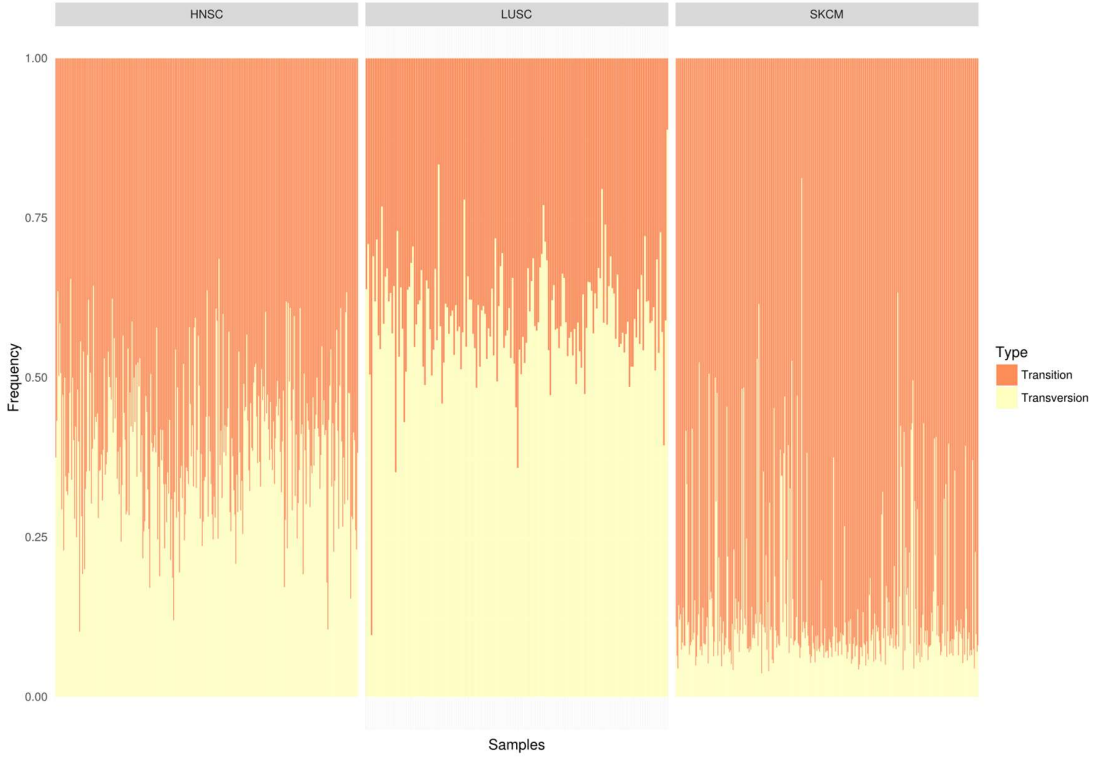

Figura 4.6 Frequência de transições e transversões das amostras em cada grupo experimental. As barras laranjas indicam a frequência de transições e as barras amarelas indicam a frequência de transversões de cada amostra (paciente).

Os tipos de trocas de nucleotídeos possíveis (T>A, T>C, T>G, C>G, C>A e C>T) também podem ser avaliados pelo número total de trocas em cada amostra e pela frequência de cada troca em relação ao número total de trocas em uma mesma amostra (**Figura 4.7**). Demonstramos que a frequência de trocas C>T por amostra é em média muito maior em SKCM do que nos outros tipos de câncer analisados, condizente com o tipo de mutação induzida por luz UV, componente da luz solar. Além disso trocas C>A são muito mais frequentes em amostras de LUSC do que em outros tipos de câncer, o

que indica que esse tipo de câncer apresenta majoritariamente esse tipo de troca, que pode ser causada por lesões do tipo 8-oxoguanina (**Figura 4.7**). Pacientes com HNSC não apresentam nenhum tipo de troca majoritária em média, sendo as trocas T>C e C>G iguais ou maiores em frequência aos outros tipos de câncer. Todos os tipos de câncer apresentaram frequência igual para a troca T>G (**Figura 4.7**).

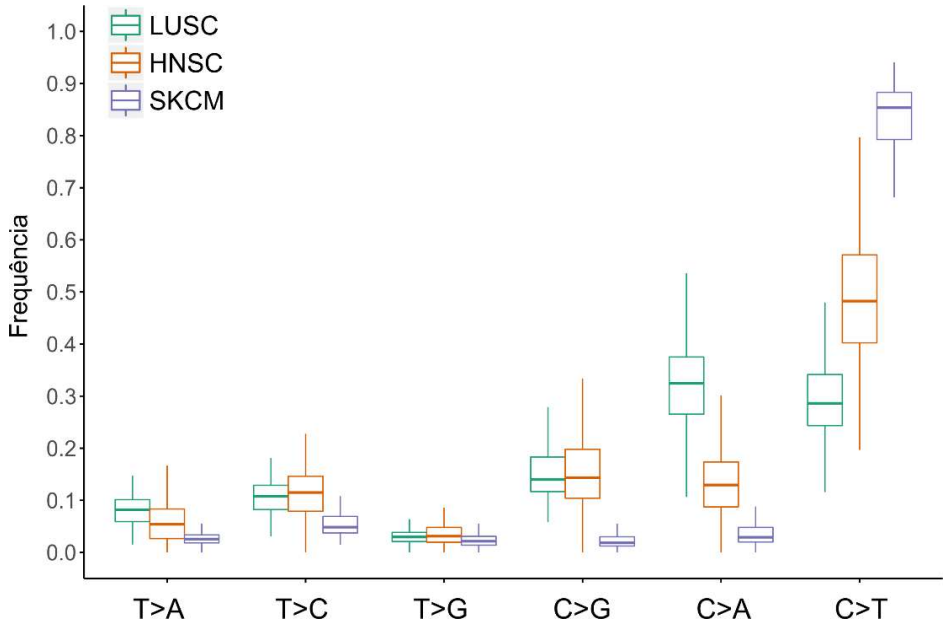

Figura 4.7 Exemplo de análise do padrão do tipo de trocas. Frequência de cada tipo de troca em relação ao total de mutações pontuais detectadas para os tipos de câncer LUSC, HNSC e SKCM. As barras de erro indicam o desvio padrão.

4.4.2 Identificação de motivos associados à agentes mutagênicos.

A identificação e quantificação da presença de motivos (*motifs*) canônicos de agentes mutagênicos pode ser utilizada para avaliar a ação desses agentes em um determinado padrão de mutação (**Figura 4.8**).

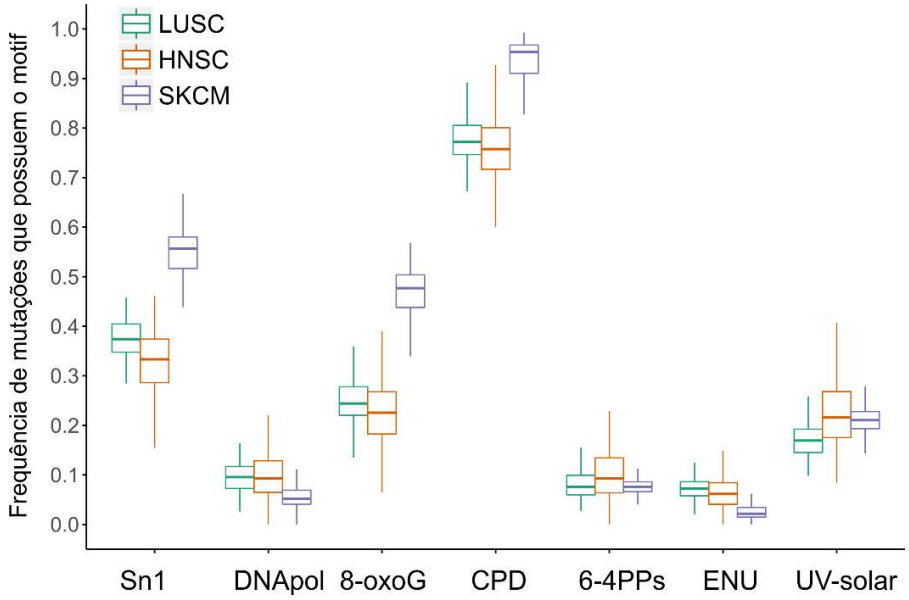

Figura 4.8 Exemplo de análise de assinaturas de agentes mutagênicos. Frequência de mutações que possuem determinado *motif* associado a agentes mutagênicos. Uma mesma mutação pode estar associada a mais de um *motif,* devido a similaridades na sequência-contexto utilizada pela busca. Frequências iguais a 1 indicam que todos as mutações possuem o motivo, enquanto frequências iguais a 0,5 indicam que metade das mutações possuem o motivo indicado.

Em nosso exemplo houve um aumento no número de motivos por mutação detectado de lesões do CPD em amostras de SKCM, confirmando o envolvimento da luz UV nesse tipo de câncer. Motivos associados com 8-oxoG também são maiores em SKCM, reforçando o envolvimento de mutagênese ambiental por luz UV nesse tipo de câncer (**Figura 4.8**).

4.4.3 Uso de janelas deslizantes para identificação de potenciais hotspots.

Woland também permite a alteração do tamanho das janelas deslizantes de acordo com o tipo de *hotspot* esperado. Como exemplo, utilizaremos um teste realizado com um conjunto de dados proveniente de um estudo de mutagênese por UVA em células humanas (CESTARI, 2017). Utilizando uma janela deslizante de 1000 pb, cujo objetivo era identificar principalmente *hotspots* em exons, foi possível identificar *hotspots* no locus HLA, situado no cromossomo 6 humano (**Figura 4.9A**). Uma parte desse locus, que concentrava a maior parte das mutações, foi analisada via *Genome Browser* (**Figura 4.9B**) sendo possível identificar que as mutações se concentravam em regiões com alto conteúdo GC com sobreposição a ilhas CpG. As mutações também

estavam distribuídas em vários genes HLA e não foi observada nenhuma correlação com genes altamente transcritos.

A identificação de regiões enriquecidas com mutações pode ser realizada de forma a identificar *hotspots* que acumulem, por exemplo, mais que 5 mutações em uma janela de 100 pb. Isso pode levar a identificar genes e funções que podem estar sendo alvo de algum tipo de seleção ou processo específico de mutagênese. O uso de ferramentas auxiliares, como o *Genome Browser,* pode agregar uma imensa quantidade de informação à essas regiões como por exemplo: conteúdo GC, repetições, ilhas CpG, metilação de DNA, acetilação de histonas, etc. Além disso, a identificação também pode ser realizada de forma a caracterizar não somente regiões, mas classes de genes mais frequentemente mutados.

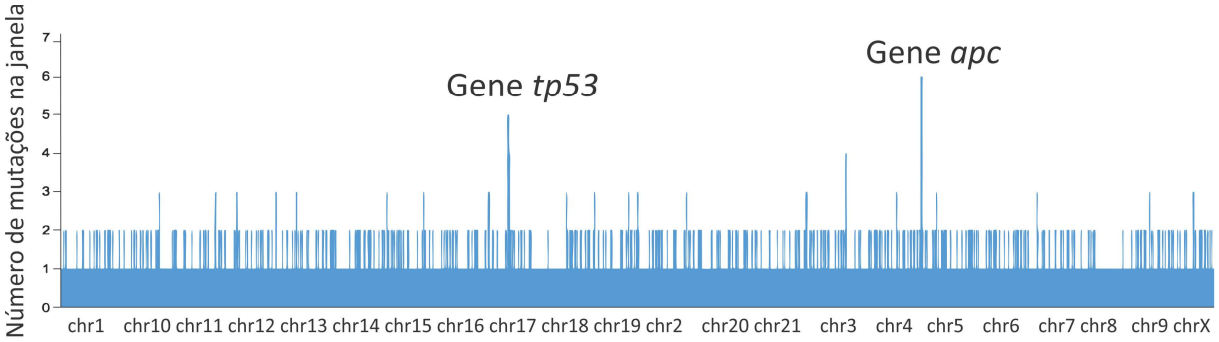

Figura 4.9 Exemplo de análise do padrão de *hotspots*. Número de *hotspots* calculados utilizando uma janela deslizante de *hotspot* de 100 pb ao redor de cada SNV, em um conjunto de dados obtido do TCGA relativo a tumores colorretais. Duas regiões enriquecidas, localizadas nos genes *tp53* e *apc,* puderam ser identificadas, e esses genes codificam para supressores de tumor diretamente envolvidos no desenvolvimento do câncer colorretal.

4.4.4 Cálculo de viés associado à transcrição dos motivos mutagênicos.

Apesar de não ser possível identificar a fita cujo nucleotídeo foi mutado utilizando apenas a informação da troca de base, é possível calcular o número de assinaturas encontradas em cada uma das fitas do DNA juntamente com a identificação da fita transcrita. Valores de MS iguais a 0 indicam concordância entre a fita da assinatura e a fita transcrita e valores diferentes de 0 indicam discordância. O tratamento parece indicar um aumento no número de mutações associadas à luz UV concordantes com a fita transcrita em ambas as amostras (**Figura 4.10A** e **4.10B**). Porém, dado que mutações com MS menores ou maiores que 0 são consideradas discordantes, um

cálculo mais apurado é a relação concordância/discordância (**Figura 4.10C**). Nesse caso o tratamento parece aumentar o número de assinaturas UV concordantes com a fita transcrita. Isso significa que o reparo acoplado à transcrição pode estar sendo responsável pela redução de mutações, visto que a fita molde para a transcrição é oposta à fita onde houve a mutação.

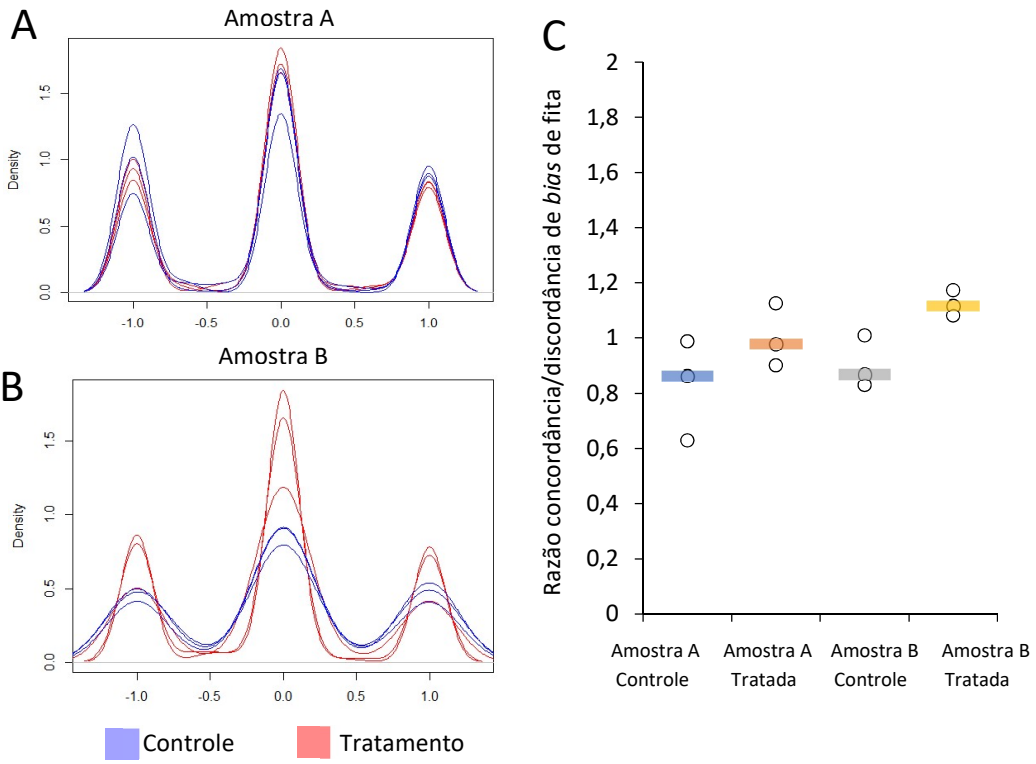

Figura 4.10 Exemplo de análise de assimetria de fita em mutações pontuais com assinaturas de luz UV. (**A**) Gráfico de densidade gaussiana do MS (*motif score*) com largura de banda igual a 0,1101 da amostra A, controle e tratada com janela. (**B**) Gráfico de densidade gaussiana do MS da amostra B, controle e tratada. (C) Razão da concordância e discordância do número de assinaturas com valores de MS, em que MS=0 foi considerado como concordante e MS ≠ 0 como discordantes. Os destaques indicam a mediana de cada conjunto de dados.

4.4.5 *Análise de performance computacional*

Para avaliar a performance computacional da ferramenta woland, foram utilizados dois conjuntos de dados de tamanhos diferentes. Um conjunto grande, chamado de *big dataset* e um conjunto menor, chamado de *small dataset*. Também foram utilizadas duas situações diferentes em termos computacionais para os testes, um servidor multiusuário com 80 *cores* de processamento e 512 Gb de memória RAM e

um computador *ARM-based* de baixo custo *Raspberry Pi3* com 4 *cores* e 1 Gb de RAM (**Figura 4.11**).

Para conjuntos grandes de dados a utilização da ferramenta woland no servidor de alta performance é inquestionavelmente mais adequada, sendo que em menos de 2 horas foram analisadas 1424 amostras e 645866 SNPs (**Figura 4.11A**). Em comparação, o processamento no computador de baixo custo *ARM-based* levou mais de 24 horas. No entanto, considerando o *small dataset,* o computador de baixo custo finalizou a análise de 60 amostras e 256 SNPs em menos de 7 horas enquanto o servidor de alto desempenho levou um pouco menos de 2 horas (**Figura 4.11B**). De fato, quanto maior o número de amostras maior é o tempo de processamento pela ferramenta woland, sendo que o número de SNPs tem pouca influência no tempo de análise (dados não mostrados).

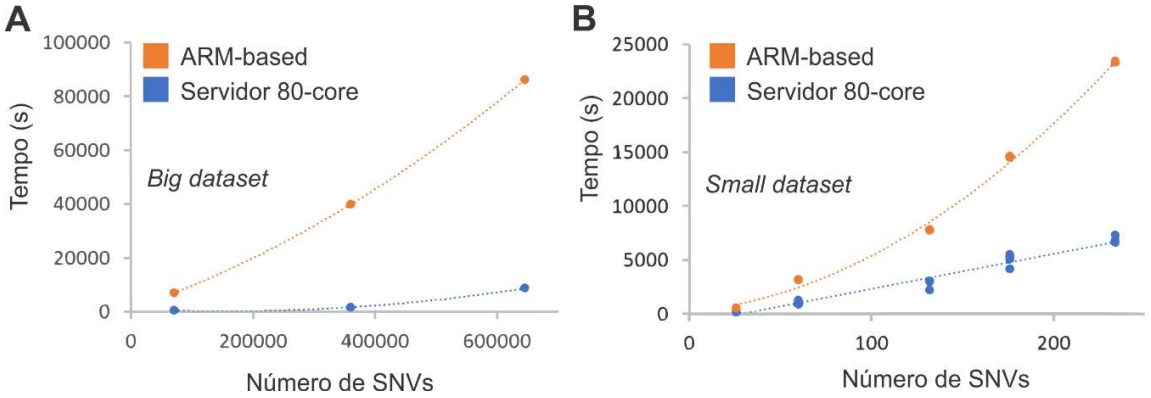

Figura 4.11 Análise de performance computacional da ferramenta woland. Desempenho utilizando o *big dataset* (**A**) e *small dataset* (**B**) da ferramenta woland implementada em um servidor de alto desempenho compartilhado (Servidor 8-core) e em um computador de baixo custo (ARM-based).

4.5 Conclusões

O surgimento de tecnologias de sequenciamento de DNA na última década pode ser considerado uma revolução no estudo de genomas individuais e até mesmo de células isoladas. No contexto desse era emergente de informação genética individualizada – genômica de precisão – e *big data,* é importante desenvolver métodos e ferramentas que permitam testes confiáveis sobre hipóteses relevantes. Apresentamos, nesse **Capítulo 4**, a ferramenta multiplataforma woland que permite analisar padrões de mutação de dados de ressequenciamento de qualquer organismo ou célula. A principal aplicação da ferramenta *woland* é auxiliar no estudo do impacto

de agentes mutagênicos endógenos e exógenos no genoma dos organismos, constituindo uma ferramenta para identificação de mecanismos mutagênicos moleculares. A identificação de padrões específicos associados a agentes mutagênicos em amostras ambientais também é uma aplicação em potencial para a ferramenta *woland*. Foi demostrado que a utilização de woland em microplacas de baixo custo, como o Raspberry Pi, é possível, permitindo a utilização em situações de pesquisa como ecotoxicogenômica, onde os recursos são limitados. Em resumo, acreditamos que a ferramenta woland pode ajudar a esclarecer a complexidade crescente que os estudos de mutagênese em escala genômica apresentam nos dias de hoje.

4.6 Referências

ALEXANDROV, L. B. et al. Signatures of mutational processes in human cancer. **Nature**, v. 500, n. 7463, p. 415–421, 2013a.

ALEXANDROV, L. B. et al. Deciphering Signatures of Mutational Processes Operative in Human Cancer. **Cell Reports**, v. 3, n. 1, p. 246–259, 2013b.

ANTONARAKIS, S. E.; BECKMANN, J. S. Focus on Monogenic Disorders. **Nature reviews. Genetics**, v. 7, n. April, p. 277–282, 2006.

ARNOLD, C. N. et al. ENU-induced phenovariance in mice: inferences from 587 mutations. **BMC research notes**, v. 5, n. 1, p. 577, jan. 2012.

BARBARIC, I.; WELLS, S. Spectrum of ENU-induced mutations in phenotype-driven and gene-driven screens in the mouse. **Molecular Mutagenesis**, v. 142, n. December 2006, p. 124–142, 2007.

BESARATINIA, A. et al. A high-throughput next-generation sequencing-based method for detecting the mutational fingerprint of carcinogens. **Nucleic Acids Research**, v. 40, n. 15, p. e116–e116, 2012.

DO, R.; KATHIRESAN, S.; ABECASIS, G. R. Exome sequencing and complex disease: Practical aspects of rare variant association studies. **Human Molecular Genetics**, v. 21, n. R1, p. 1–9, 2012.

GLAZKO, G. B.; MILANESI, L.; ROGOZIN, I. B. The subclass approach for mutational spectrum analysis: application of the SEM algorithm. **Journal of theoretical biology**, v. 192, n. 4, p. 475–487, 1998.

IKEHATA, H. et al. Remarkable induction of UV-signature mutations at the 3'-cytosine of dipyrimidine sites except at 5'-TCG-3' in the UVB-exposed skin epidermis of xeroderma pigmentosum variant model mice. **DNA Repair**, v. 22, p. 112–122, 2014.

KOBOLDT, D. C. et al. Review The Next-Generation Sequencing Revolution and Its Impact on Genomics. **Cell**, v. 155, n. 1, p. 27–38, 2013.

LÁZÁR, V. et al. Genome-wide analysis captures the determinants of the antibiotic cross-resistance interaction network. **Nat Commun**, v. 5, 2014.

MENCK, C. F.; MUNFORD, V. DNA repair diseases: what do they tell us about cancer and aging? **Genetics and Molecular Biology**, v. 37, n. 1, p. 220–233, 2014.

NIK-ZAINAL, S. et al. Mutational processes molding the genomes of 21 breast cancers. **Cell**, v. 149, p. 979–993, 2012.

POON, S. et al. Mutation signatures of carcinogen exposure: genome-wide detection and new opportunities for cancer prevention. **Genome Medicine**, v. 6, n. 3, p. 24, 2014.

ROGOZIN, I. B. et al. Computational analysis of mutation spectra. **Briefings in bioinformatics**, v. 4, n. 3, p. 210–227, 2003.

STAPLEY, J. et al. Adaptation genomics: the next generation. **Trends in ecology & evolution**, v. 25, n. 12, p. 705–12, dez. 2010.

TENG, G.; PAPAVASILIOU, F. N. Immunoglobulin Somatic Hypermutation. **Annu Rev Genet**, v. 41, p. 107–120, 2007.

5 CONSIDERAÇÕES FINAIS

O surgimento e o estabelecimento das tecnologias de sequenciamento NGS possibilitaram um salto qualitativo e quantitativo enorme em genômica. Porém, um dos grandes desafios é justamente a interpretação dos dados em um contexto biológico. Dessa forma, o presente trabalho teve como objetivo utilizar o sequenciamento de exomas para a caracterização de variantes genéticas em exomas de camundongos, desenvolvendo estratégias robustas para análise e interpretação desses dados.

Primeiramente, estabelecemos uma estratégia de sequenciamento com otimização de custos que possibilitou a obtenção de uma quantidade de leituras suficiente para a obtenção de cobertura superior a 70X, em média, do exoma de duas linhagens isogênicas mantidas pelo ICB, e de 7 camundongos mutantes. Para as linhagens isogênicas, foi descartada a possiblidade de contaminação cruzada com outras linhagens isogênicas. Além disso, o sequenciamento do exoma revelou que a linhagem C57BL/6ICBI é mais próxima da linhagem C57BL/6NJ, proveniente do NIH, e não da sublinhagem C57BL/6J, proveniente da JAX. Já a linhagem BALB/c é bem próxima a sublinhagem BALB/cJ. Os SNVs novos detectados nas sublinhagens do ICB constituem uma valiosa fonte de informação para toda a comunidade científica usuária dos camundongos do ICB.

Em relação aos camundongos mutantes induzidos por ENU, foi desenvolvida uma estratégia de filtragem para seleção de variantes potencialmente causadoras dos fenótipos observados. Essa estratégia possibilitou a seleção de candidatos para todos os mutantes avaliados, com alta taxa de validação por sequenciamento Sanger. Análises de impacto funcional dessas variantes indicaram interessantes genes candidatos, principalmente para os mutantes bate palmas, equilíbrio, atáxico-1, fraqueza e careca.

Em resumo, acreditamos que o trabalho traga contribuições valiosas tanto para os usuários das linhagens isogênicas do ICB bem como incentivar a continuidade de estudos funcionais para os mutantes induzidos por ENU, pelos grupos colaboradores. Além disso, também acreditamos que o desenvolvimento de uma ferramenta de bioinformática aberta de análise do padrão de mutações possa também contribuir para o entendimento e para a formulação de novas hipóteses de processos de mutagênese.

APÊNDICE

6.1 Tecnologias de sequenciamento NGS

As tecnologias NGS mais utilizadas compartilham algumas características entre si, apesar do emprego das tecnologias em cada plataforma variar consideravelmente (RIEBER et al., 2013). Todas as tecnologias buscam uma diminuição crescente no custo por base sequenciada, velocidade no sequenciamento, grande quantidade e qualidade de leituras produzidas (MARDIS, 2008). As diferenças mais marcantes entre as plataformas são principalmente o tamanho das leituras (*reads)* produzidas: *small reads* - em geral leituras menores que 100 pb - ou *long reads* - maiores que 100 pb e a codificação das leituras produzidas, codificadas em *basespace* (A,T,C ou G) ou *colorspace* - seguindo o esquema *2-base encoding* (MARDIS, 2013).

Tecnologias NGS convencionais compartilham alguns princípios básicos: os ácidos nucléicos de alto peso molecular extraídos e purificados são fragmentados de forma aleatória; bibliotecas de DNA são produzidas com os fragmentos e adaptadores artificiais; as bibliotecas são amplificadas de forma clonal em uma superfície sólida; o sequenciamento dos nucleotídeos é realizado etapa por etapa e de forma massiva e paralela, isto é, vários clones de DNA da biblioteca são sequenciados ao mesmo tempo; finalmente, a informação é organizada fornecendo sequências dos fragmentos de DNA de cada clone, chamadas de leituras - adaptado de (DE SOUZA; IENNE, 2018). A grande maioria das plataformas de sequenciamento NGS, por sua vez, consiste basicamente de três unidades fundamentais: uma unidade fluídica automatizada - responsável pelo fluxo de reagentes, uma unidade de captação de informação - normalmente um sistema de detecção, como um microscópio de fluorescência - e uma unidade de processamento, que converte os dados captados em informação posicional de bases nitrogenadas em forma de sequências, denominadas leituras - adaptado de (DE SOUZA; IENNE, 2018).

Normalmente plataformas que produzem *small reads* geram um volume grande de dados em comparação com plataformas que geram *long reads*. Naturalmente, com o estabelecimento e o uso das tecnologias, as plataformas foram se ajustando a aplicações específicas (MARDIS, 2008, 2013). Projetos que priorizem o mapeamento de variações genômicas simples (SNPs, INDELs, CNVs, etc.) geralmente são

realizados em plataformas de *small reads,* visto que há a disponibilidade de um genoma referência para alinhamento das leituras. Análises quantitativas e qualitativas de transcriptoma, quando há um genoma referência disponível, também são adequadas a esse tipo de sequenciamento. Análises de ressequenciamento são geralmente mais simples do que montagens sem referência, dependendo dos objetivos a serem alcançados (DE SOUZA; IENNE, 2018). Já as plataformas de *long reads* são também adequadas para projetos de sequenciamento de genomas (ou transcriptomas) cuja análise consiste basicamente de uma montagem de novo, sem a utilização de um genoma referência (RIEBER et al., 2013). Todas essas aplicações são dependentes diretamente das características genômicas e do delineamento experimental proposto.

As maiores diferenças entre as plataformas NGS convencionais se dão na forma de geração do sinal de distinção das bases, ou seja, na própria química de sequenciamento em si. A primeira química de sequenciamento NGS surgiu em 2004, chamada de pirossequenciamento e utilizados nas plataformas 454 da *Roche*. Considera-se que o sequenciamento individual de moléculas, cujas tecnologias mais promissoras são desenvolvidas pelas empresas Pacific Biosciences e Oxford Nanopore Technologies, é o próximo e revolucionário passo da genômica (MARDIS, 2017). Atualmente a química mais utilizada é o sequenciamento por síntese - presente em todas as plataformas *Illumina* – cujos custos são muito baixos em relação a qualidade das bases produzidas. Porém, uma das químicas que se destacaram entre os anos de 2007 e 2015 foi o sequenciamento por ligação, utilizado nas plataformas SOLiD da A*pplied Biosystems* – mais tarde incorporada pela *Life Technologies* – cuja acurácia de detecção podia chegar a 99,9 % (BARBA; CZOSNEK; HADIDI, 2013).

6.1.1 A química de sequenciamento por ligação (SOLiD)

Nesse tipo de sequenciamento NGS, a amplificação clonal das bibliotecas se dá por PCR em emulsão, gerando esferas com um clone único amplificado, que são depositadas covalentemente em uma lâmina para o sequenciamento (**Figura 6.1**).

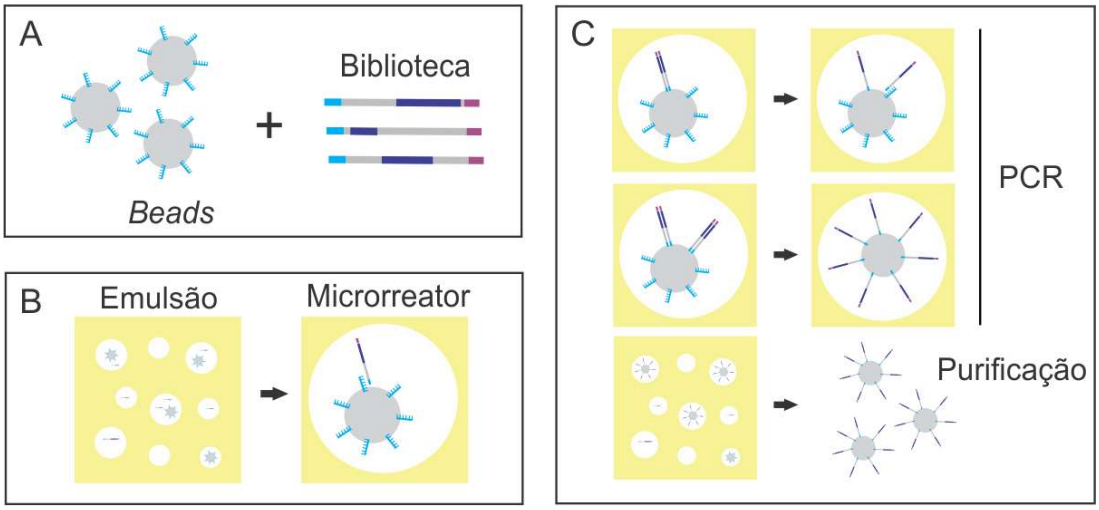

Figura 6.1 Etapas da PCR em emulsão para sequenciamento na plataforma SOLiD 5500XL. As *beads*, que contêm adaptadores ligados de forma covalente em sua superfície, são misturadas com a bibliotecas DNA (**A**). O preparo da emulsão consiste em misturar, em proporções corretas, as *beads,* bibliotecas e a solução de emulsão, de forma a formar microrreatores que devem conter, de forma preferencial, apenas uma *bead* e um fragmento de DNA da biblioteca (**B**). A emulsão é então submetida a uma reação de PCR para amplificação dos fragmentos dentro dos microrreatores, de forma a produzir *beads* com várias cópias do fragmento aderidas em sua superfície. A emulsão é então quebrada e as *beads* são purificadas da emulsão (**C**). Adaptado de (DE SOUZA; IENNE, 2018).

A detecção do sinal correspondente a cada nucleotídeo é obtida através de uma química específica, chamada sequenciamento por ligação (**Figura 6.2**). Esse método utiliza as características da enzima DNA ligase, e não uma DNA polimerase, como a maioria dos métodos de sequenciamento. O método é baseado na especificidade da ligase no pareamento de bases das fitas do DNA e na utilização de sondas marcadas com fluoróforos. Cada sonda possui um par de bases (di-bases) cuja ordem codifica para um fluoróforo específico, seguido de três nucleotídeos randômicos e três inosinas, sendo que, à última inosina está ligado o fluoróforo. Ao todo há 16 tipos possíveis de di-bases, cada quatro tipos correspondendo a um dos quatro fluoróforos disponíveis. Cada etapa de ligação corresponde à exposição da lâmina às sondas contendo os fluoróforos e aos oligonucleotídeos compostos por “n” bases complementares ao adaptador conectado às esferas (DE SOUZA; IENNE, 2018; MARDIS, 2008).

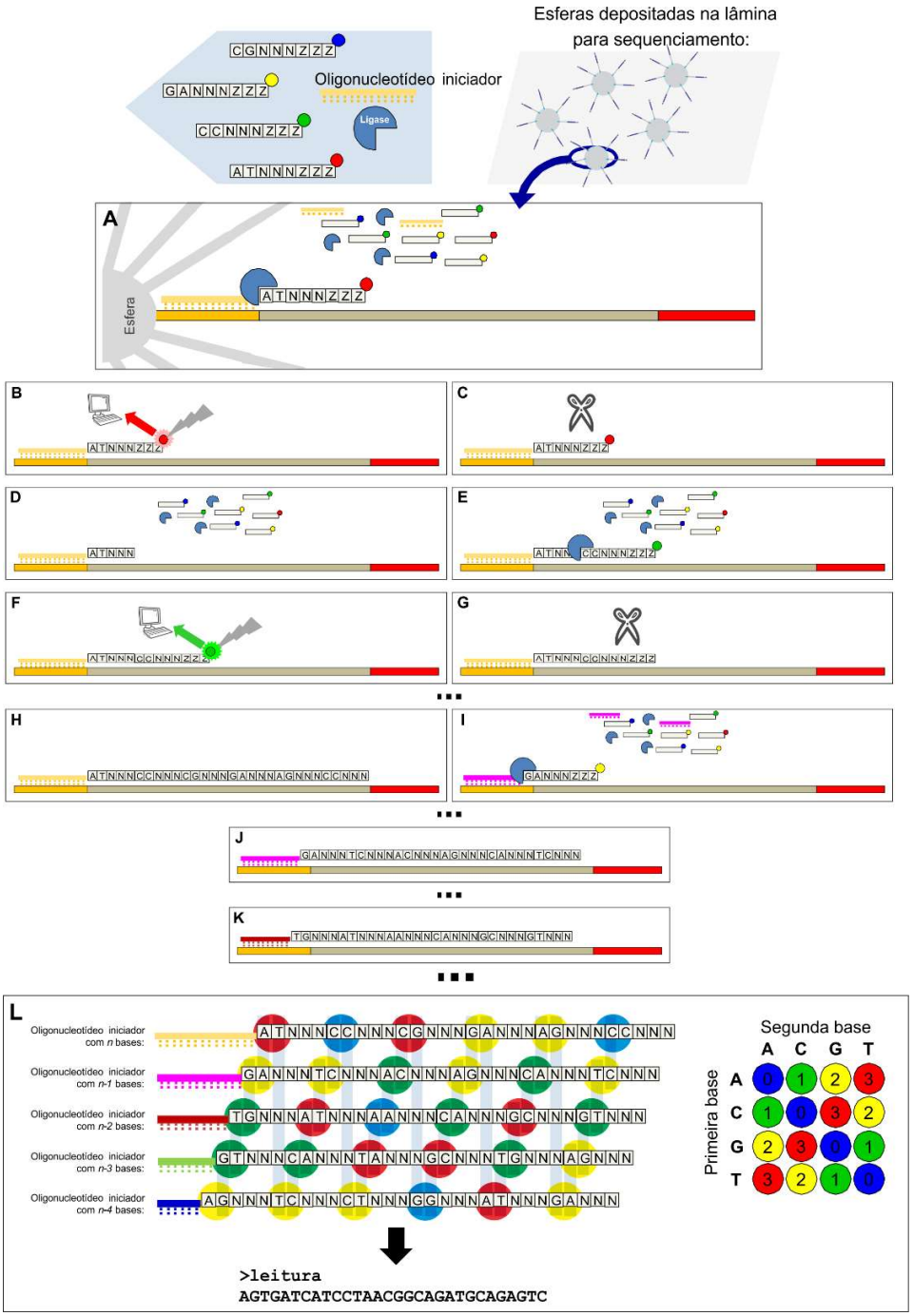

Figura 6.2 Química de sequenciamento por ligação. (**A**) A reação de sequenciamento acontece em cada esfera originada do processo de amplificação clonal pela adição dos reagentes necessários. (**B**) Detecção da fluorescência emitida pela primeira reação de ligação e (**C**) clivagem das três inosinas da extremidade 5' da sonda, liberando para uma nova etapa de ligação. (**D**) Disponibilização das sondas e continuidade do processo de incorporação (**E**) pela ligase, detecção (**F**) e clivagem (**G**). Os ciclos de ligação se repetem por um número determinado de ciclos (**H**) e uma nova etapa se inicia com a hibridização com um novo tipo de oligonucleotídeos (**I**). O uso de diversos tipos de oligonucleotídeos a cada etapa de ligação (**J** e **K**), permite o sequenciamento de todo o fragmento de DNA através da obtenção das leituras interpretadas pelo código de 16 cores (**L**). Fonte: (DE SOUZA; IENNE, 2018).

Após o anelamento do oligonucleotídeo, a DNA ligase é adicionada e fixará a sonda que se anelar especificamente às cinco primeiras posições do fragmento de DNA. A lâmina é exposta ao laser com os filtros para leitura de cada um dos quatro fluoróforos e passa por um processo de clivagem e lavagem das inosinas da posição 5' das sondas. Como consequência, o fluoróforo é removido da sequência, deixando uma extremidade 5' fosfato pronta para uma nova etapa de ligação, com uma nova sonda. Ao final de várias ligações, as sondas são removidas e um novo ciclo é reiniciado com um oligonucleotídeo de tamanho “n-1”, permitindo a leitura da base adjacente à leitura anterior e novamente uma série de ligações de sondas com os fragmentos nas esferas. Para a produção de leituras de 75 pares de bases são necessários cinco ciclos com 15 ligações cada, por exemplo (DE SOUZA; IENNE, 2018; MARDIS, 2008).

Esse método de sequenciamento possui características singulares que permitem que a taxa de erro por base seja inferior a um erro para cada mil pares de bases. Cada base é sequenciada virtualmente duas vezes nesse tipo de sequenciamento. Todavia, é o método de sequenciamento que possui um dos menores tamanhos de leitura. A alta acuidade do sequenciamento por ligação também depende que o mapeamento das leituras seja realizado utilizando um código diferente de bases, chamado de *colorspace* - codificação por cores, convertidas a números correspondentes a di-bases (DE SOUZA; IENNE, 2018; MARDIS, 2008).

6.2 Métodos Suplementares

6.2.1 Extração de DNA genômico da ponta da cauda de camundongos

Testes iniciais de extração de amostra provenientes de camundongos C57BL6/ICBI do baço e da ponta da cauda geraram quantidades similares de aproximadamente 3,6 ng/µL e razões 260/280 maiores que 1,8, 260/270 iguais a 1,2 e 260/230 maiores que 1,9. Apesar dos indicadores de qualidade semelhantes as amostras provenientes do baço apresentaram, na análise por eletroforese em gel de agarose, bandas de arrastes (dados não mostrados) que poderiam estar ligadas a uma provável degradação ou provável contaminação com RNA – moléculas que interferem na a captura de exons. No entanto, amostras de DNA genômico extraídos a partir de tecido da ponta da cauda (**Figura 6.3**) não mostraram sinal aparente de degradação

e/ou contaminação por RNA e dessa forma foram selecionados para as etapas posteriores de preparo de bibliotecas.

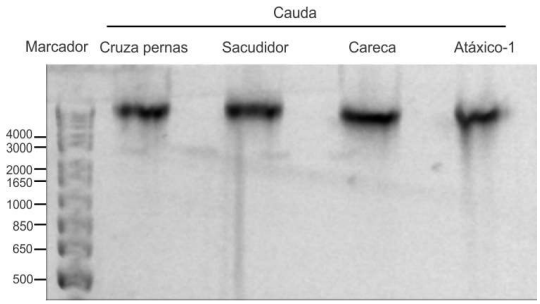

Figura 6.3 - DNA genômico de amostras representativas extraídas da ponta da cauda. Eletroforese em gel de agarose (2%) de DNA genômico de alto peso molecular (>4000 pb) extraído da ponta da cauda de 4 amostras representativas (Cruza Pernas, Sacudidor, Careca e Atáxico-1). Marcador 1Kb Plus DNA Lader (Thermo Fisher).

As quantificações por fluorescência das amostras de DNA genômica variaram de 167 a 1020 ng/µL, com todos os indicadores de qualidade por absorbância adequados. Não foi detectada, no DNA genômico extraído da ponta da cauda, indícios de degradação aparente de DNA nem contaminação aparente por RNA, indicadores que poderiam afetar o processo de hibridização e enriquecimento.

6.2.2 Produção das bibliotecas e PCR em emulsão

Um total de 12 bibliotecas foram produzidas com a utilização de 12 diferentes adaptadores, referentes ao *barcodes* BC1 a BC12. A amostra do mutante fraqueza, escolhida ao acaso, foram duplicadas (BC7 e BC10) juntamente com amostras dos camundongos isogênicos C57BL/6ICBI (BC9 e BC12) e BALB/cICBI (BC8 e BC11). Após o preparo, as bibliotecas foram quantificadas e analisadas quanto ao padrão de distribuição dos fragmentos. A **Figura 6.4** retrata resultados obtidos de uma amostra representativa da amostra que corresponde ao BC14, indicando que não há indícios de dímeros de oligonucleotídeos de tamanho 10 a 100 pb (**Figura 6.4A e 6.4B**) e que a grande parte dos fragmentos, cerca de 82%, tem um tamanho médio de 281 pb (**Figura 6.4B**).

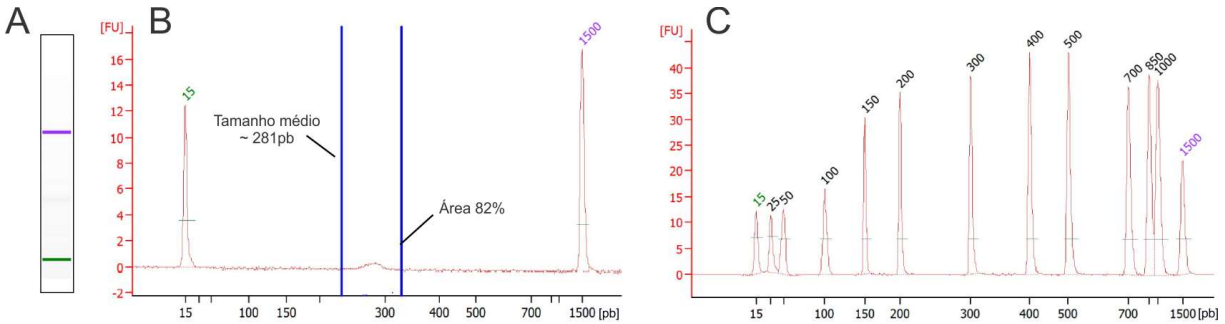

Figura 6.4 - Dispersão dos fragmentos de DNA de uma biblioteca representativa. Eletroforese microfluídica da amostra BC14 (fraqueza). (A) Gel virtual esquemático do padrão de dispersão dos fragmentos da amostra BC14. (B) Quantificação em unidades de fluorescência (FU), com os limites de cálculo de tamanho em azul. (C) Dispersão dos fragmentos do peso molecular de 15 pb a 1500 pb utilizado como controle da corrida e para correspondência entre o tempo de corrida e tamanho em pares de bases.

Todas as bibliotecas analisadas foram submetidas ao mesmo tipo de análise for eletroforese microfluídica demonstrada na **Figura 6.4** e também submetidas a uma nova quantificação de DNA por fluorescência. Em média, a concentração de DNA das bibliotecas produzidas foi igual a 5,96 ng/µL ± 3,41 e o tamanho médio calculado por eletroforese microfluídica foi de 289,33 pb ± 6,98. O tamanho médio esperado para as bibliotecas construídas com o kit utilizado é de cerca de 300 pb, indicando o sucesso na fragmentação e ligação dos adaptadores de sequenciamento e *barcodes* (**Tabela 6.1**).

Portanto, o tamanho médio das bibliotecas variou apenas cerca de 2,4% permitindo bons parâmetros para uma PCR em emulsão eficiente sem amplificação preferencial de nenhuma amostra em particular (**Tabela 6.1**). A produção das bibliotecas foi realizada com alta qualidade para todas as amostras desejadas, em condições ideais para o preparo do *pool* de todas as amostras em um único tubo, necessário para as etapas seguintes de PCR de emulsão e sequenciamento.

Tabela 6.1 - Quantificação e tamanho médio dos fragmentos de cada biblioteca produzida e seus respectivos barcodes.

Amostra	*Barcode*	DNA (ng/µL)	Tamanho médio (pb)
Atáxico-1	BC1	5,9	298
Careca	BC2	8,6	301
Sacudidor	BC3	7,6	295
Cruza pernas	BC4	13,0	286
Bate palmas	BC5	3,0	285
Equilíbrio	BC6	1,7	289
Fraqueza	BC7 e BC10	5,2	285
BALB/cICBI	BC8 e BC11	4,9	284
C57BL/6ICBI	BC9 e BC12	3,7	281

As bibliotecas produzidas foram misturadas em quantidades equimolares, de acordo com a quantificação obtida e o tamanho médio dos fragmentos pela relação:

$$\text{pmol DNA} = \mu\text{g DNA}\left(\frac{pmol}{660\text{ pg}}\right)\left(\frac{10^6}{1\ \mu g}\right)\left(\frac{1}{N}\right)$$

Em que N é o número de nucleotídeos e (600 pg/pmol) é a média do peso molecular de um par de nucleotídeos.

O *pool* das bibliotecas misturadas de forma equimolar foi novamente quantificado e cerca de 48,5 ng foi utilizado para a PCR de emulsão na plataforma EZBeads. A plataforma é dividida em três etapas distintas de preparo da emulsão (*Emulsifier*), amplificação (*Amplifier*) e enriquecimento de *beads* P2. Para cada rodada de PCR de emulsão foram utilizados kits E120 e E80, que diferem no rendimento do número total de *beads* produzidas. O rendimento esperado para os *kits* E120 e E80 é de aproximadamente 2,2 bilhões e 1 bilhão de *beads*, respectivamente. O limite máximo de deposição nas *flowcells,* supondo um limite de 250.000 *beads* por painel é de aproximadamente 300 milhões de *beads*. Sendo assim cada kit E120 é suficiente para 6 lanes e cada kit E80 é suficiente teoricamente para 3 *lanes*.

Portanto, a correta quantificação das *beads* com capacidade de ligação às lanes (P2) é crucial para o sucesso da deposição e sequenciamento. No total foram realizados 5 ciclos de PCR de emulsão, totalizando três kits E120 e dois kits E80, cujas

beads foram utilizadas para uma corrida de controle de qualidade (Workflow Analysis) e para 11 lanes divididas em três corridas de sequenciamento.

6.2.3 Sequenciamento das bibliotecas e leituras produzidas

A quantidade disponível de *beads* a serem depositadas na corrida foi escolhida a partir da análise de WFA. Cada *flowcell* possui 6 *lanes* independentes, que podem ser sequenciadas para aplicações diferentes em uma mesma corrida. Foram sequenciadas 11 lanes divididas em corridas diferentes, que foram sequenciadas em três ocasiões de corridas distintas, com compartilhamento de *lanes* visando a máxima economia de reagentes e custo de operação da plataforma de sequenciamento.

Como primeira etapa para avaliação da qualidade das leituras produzidas, é realizada a avaliação dos relatórios de sequenciamento. As leituras em *colorspace* são produzidas preferencialmente de uma forma diferente da análise de leituras em *basespace,* como as obtidas pelas plataformas Illumina. A qualidade dos dados gerados é medida principalmente pelas métricas obtidas do gráfico *Satay* e da classificação das *beads* detectáveis em melhores beads (*best beads)* e boas *beads* (*good beads*), como mostrado para uma *lane* na **Figura 6.5**. A organização esquemática de uma *bead* é descrita na **Figura 6.5A**.

As principais métricas avaliadas pelo gráfico Satay (**Figura 6.5B**) são: o número total de *beads* P2, ou seja, aquelas *beads* que possuem um DNA *template* e, portanto, irão gerar dados utilizáveis; a razão ruído/sinal (N2S), que é um número indicador de clonalidade da amostra. Um número baixo indica que a maioria das *beads* são monoclonais; A porcentagem nos eixos (% nos eixos) indica a frequência da quantidade de *beads* que geram dados fluorimétricos que caem em até 10% de cada canal de cor. *Beads* que geram sinais puros, ou seja, mais próximos aos eixos, indicam um alto teor de monoclonalidade; A mediana de *beads* P2 por painel (Mediana P2/painel) indica a quantidade de beads depositadas por painel de detecção. O limite de *beads* a serem depositadas gira em torno de 200.000 *beads* P2 por painel. Todas essas métricas são mostradas para o primer F3 da *lane* 04 (**Figura 6.5B**), indicando bons níveis de detecção dos canais fluorimétricos e quantidade de *beads* monoclonais.

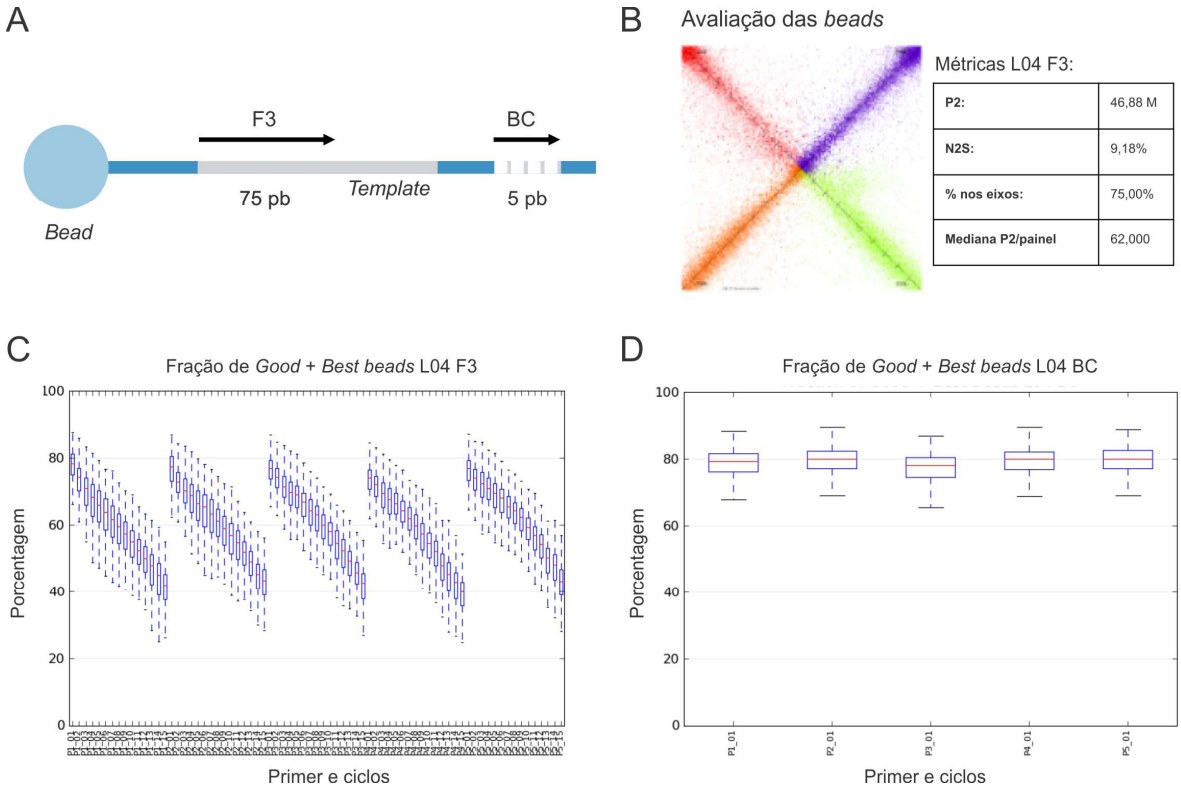

Figura 6.5 - Parâmetros gerais de qualidade das leituras produzidas pela plataforma SOLiD 5500xl para uma *lane*. (A) Desenho esquemático da organização de uma *bead* e as respectivas *tags*. (B) Gráfico *Satay* de avaliação das *beads* em relação aos canais de detecção de fluorescência. (C) Porcentagem das frações *good* e *best beads* da *tag* F3 (75 pb) a cada primer e ciclo de sequenciamento, totalizando 15 *primers* e 5 ciclos de sequenciamento. (C) Porcentagem das frações *good* e *best beads* da tag BC (5 pb) de cada um dos 5 primers de um único ciclo.

Em média, são necessários de 10 a 12 dias para a finalização do sequenciamento de cada *flowcell* em modo 1x75 pb. As leituras F3 são produzidas em um total de 75 ciclos de ligação divididos em 5 rodadas com 5 diferentes oligonucleotídeos de sequenciamento n-1 (**Figura 6.5C**) e as leituras correspondentes aos barcodes são produzidas por 5 ciclos de ligação em uma única rodada do oligonucleotídeo (**Figura 6.5D**). Algumas métricas de corrida são avaliadas em tempo real e permitem que ciclos ou etapas de ligação possam ser interrompidas ou repetidas conforme a opção pelo usuário. Nas três corridas realizadas não foi necessária a interrupção ou repetição de nenhuma etapa, sendo as corridas checadas pelo menos 4 vezes ao dia.

6.3 Referências

BARBA, M.; CZOSNEK, H.; HADIDI, A. Historical perspective, development and applications of next-generation sequencing in plant virology. **Viruses**, v. 6, n. 1, p. 106–136, 2013.

DE SOUZA, T. A.; IENNE, S. **Capítulo 11: Sequenciamento de DNA**. In: Biologia Molecular, Editor: Nancy Rebouças, submetido. Sao Paulo: Editora Atheneu, 2018.

MARDIS, E. R. Next-Generation DNA Sequencing Methods. **Annual Reviews Genomics and Human Genetics**, v. 9, p. 387–402, 2008.

MARDIS, E. R. Next-Generation Sequencing Platforms. **Annual Reviews Genomics and Human Genetics**, v. 6, p. 287–303, 2013.

MARDIS, E. R. DNA sequencing technologies: 2006–2016. **Nature Protocols**, v. 12, n. 2, p. 213–218, 2017.

RIEBER, N. et al. Coverage bias and sensitivity of variant calling for four whole-genome sequencing technologies. **PloS One**, v. 8, n. 6, p. e66621, jan. 2013.